Development Of High Throughput Epigenomic Profiling Technologies And Their Application To Twin Based DNA Methylation Studies

By

Zachary Aaron Kaminsky

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Institute of Medical Science
University of Toronto

© Copyright by Zachary Aaron Kaminsky 2009
Thesis Abstract

Epigenetic studies hold the promise of addressing some of the fundamental questions of human biology including development, cell differentiation, and the aetiological mechanisms of complex disease. Over the last years, several new large scale high throughput technologies have been developed to allow genome wide profiling of epigenetic signals such as DNA methylation and histone modifications. Two of such technologies were developed in our laboratory enabling a genome wide microarray based profiling of DNA methylation signatures and a high throughput method for the site specific interrogation of the density of methylated cytosine. Using these techniques, we identified a DNA methylation difference in the 3’UTR of the DLX1 gene with potentially functional implications to discordance in risk taking behavior in a single pair of MZ twins. We modeled a power analysis on the effect size of the detected difference and determined that approximately 6~25 discordant twin pairs will be adequate to yield 80% power across the entire 12 K CpG island microarray platform using our epigenomic microarray profiling technique. We performed a DNA methylome analysis of MZ twins in white blood cells (WBC), buccal epithelial cells, and gut (rectum) biopsies (N=57 pairs in total) using 12K CpG island microarrays providing the basis for the first annotation of epigenetic metastability of ~6,000 unique genomic regions in MZ twins. We performed a classical twin study on DNA methylation differences in WBC and buccal epithelial cells from 39 pairs of MZ twins to 40 pairs of DZ twins. DZ co-twins exhibited significantly higher epigenetic difference compared to the MZ co-twins in buccal cells (p=1.2x10^{-294}). While such higher epigenetic discordance in DZ twins can result from DNA sequence differences, our in silico SNP analyses and comparison of methylomes in inbred vs. outbred mice favour the hypothesis that this is due to epigenomic
differences in the zygotes. This study suggests that molecular mechanisms of heritability may not be limited to DNA sequence differences.
Table of Contents

Thesis Abstract................................................................................................................................ ii
List of Tables ............................................................................................................................... viii
List of Figures ........................................................................................................................... ix
List of Appendices .................................................................................................................... x
List of Abbreviations ............................................................................................................... xi
Acknowledgments.................................................................................................................... xiii
Thesis Introduction ................................................................................................................... 1
  Epigenetics .............................................................................................................................. 1
  Epigenetic signals and biological development ..................................................................... 3
  Epigenetic metastability ......................................................................................................... 4
  Epigenomic profiling technologies ....................................................................................... 5
  Site specific DNA methylation profiling .............................................................................. 6
  Epigenomic microarray technologies................................................................................... 6
  The classical twin design ....................................................................................................... 8
  Epigenetics and the classical twin design ............................................................................ 10
  Epigenetic inheritance ........................................................................................................... 13
Thesis Objectives.................................................................................................................... 18
Chapter 1................................................................................................................................... 0
  Microarray-based DNA Methylation Profiling: Technology and Applications .................... 0
  Summary ............................................................................................................................... 21
  Introduction .......................................................................................................................... 22
  Results .................................................................................................................................. 23
    Enrichment of the unmethylated fraction of genomic DNA ................................................ 23
    Microarray design............................................................................................................. 28
List of Tables

Table 1.1. Enzymes that generate protruding ends in the restriction fragments, which are complementary to the adaptors U-CG1, TA-1 and AATT-1
Table 1.2. Distribution of the detected unmethylated sites in respect to the known genes as defined by the combined set of RefSeq and UCSC Known Genes for each brain DNA sample (M17-M25) and the merged map.
Table 4.1. GO analysis of loci with high MZ co-twin epigenetic similarity
Table 4.2. GO analysis of loci with low MZ co-twin epigenetic similarity
Table 4.3. Sodium bisulfite treated loci and primers
List of Figures

Figure 1.1. Schematic outline of the microarray-based method for identification of DNA methylation differences and DNA polymorphisms in genomic DNA

Figure 1.2. Selective enrichment of restriction fragments with the universal adaptor U-CG1

Figure 1.3. *Comt* Microarray Design

Figure 1.4. Combined methylation- and SNP-analysis on a CpG island microarray

Figure 1.5. Reproducibility and sensitivity of the method

Figure 1.6. Applications of the epigenetic profiling technology

Figure 1.7. Examples of applications using a CpG island microarray

Figure 1.8. Profiles of unmethylated sites in three loci on human chromosomes 21 & 22

Figure 1.9. Genomic views showing unmethylated regions on chromosomes 21 and 22

Figure 2.1. Bisulfite modified COMT promoter region and SNaPshot primers

Figure 2.2. Single SNP oligonucleotide templates

Figure 2.3. Multiple SNP oligonucleotide templates

Figure 2.4. Galectin1 and Humanin SNaPshot primers

Figure 2.5. SNaPshot primers accurately measure DNA methylation

Figure 2.6. Multiplexed SNaPshot reaction output

Figure 2.7. Primer mismatch induced bias

Figure 2.8. Correction of mismatch induced bias

Figure 2.9. SNaPshot vs. cloning and sequencing data

Figure 2.10. Multiple peaks

Figure 3.1. The Toronto Gambling Task displays and contingencies

Figure 3.2. MMPI-2 scores for the “law” twin and the “war” twin

Figure 3.3. Gambling performance for the twins as well as control subjects (n = 11)

Figure 3.4. Relative DNA methylation profiles of the “war” twin vs. “law” twin

Figure 3.5. Power vs. technical replicate hybridization number and sample size (N)

Figure 3.6. Power vs. effect size and sample size (N)

Figure 4.1. Biological vs. technical variation

Figure 4.2. Correlations between microarray and sodium bisulfite sequencing data

Figure 4.3. Pyrosequencing correlations as a function of distance

Figure 4.4. Karyogram of MZ co-twin epigenetic similarity in WBCs

Figure 4.5. Raw binding intensities of MC and DC MZ twin hybridizations

Figure 4.6. MZ and DZ ICC distributions in buccal cells

Figure 4.7. Karyogram of MZ$_{ICC}$-DZ$_{ICC}$ values in buccal cells of DC MZ twins

Figure 4.8. Technical variation volcano plots of HpaII and MspI based enrichments

Figure 4.9. Distributions of inbred and outbred epigenetic variation

Figure A.1. Karyogram of MZ co-twin epigenetic similarity in buccal cells

Figure A.2. Karyogram of MZ co-twin epigenetic similarity in gut

Figure A.3. Karyogram of MZ$_{ICC}$-DZ$_{ICC}$ values in WBCs

Figure A.4. Karyogram of MZ$_{ICC}$-DZ$_{ICC}$ values in buccal cells of MC MZ twins
Appendix 1. Supplementary Figures
List of Abbreviations

3’ untranslated region (3’ UTR)
Applied Biosystems (ABI)
Armadillo repeat gene deleted in VCFS (ARVCF)
CIq tumor necrosis factor-related protein 8 precursor (CIQTNF8)
Canadian Institutes for Health and Research (CIHR)
Catechol-o-methyltransferase gene (COMT)
Chromatin immunoprecipitation (ChIP)
Comprehensive high-throughput arrays for relative methylation (CHARM)
CpG island (CGI)
D4 receptor (D4DR)
Differentially methylated region (DMR)
Distal-less Homeobox 1 gene (DLX1)
Dizygotic (DZ)
DNA adenine methyltransferases (DAM)
DNA methyltransferases (DNMT)
Dopamine receptor (D4DR)
False discovery rate (FDR)
Family-wise error rate (FWER)
Gene Ontology (GO)
Genome wide association studies (GWAS)
Genomic DNA (gDNA)
Heterozygosity quotient (HQ)
Histone 3 (H3)
Histone 3-lysine 4 (H3-K4)
Histone acetyltransferases (HAT)
Histone deacetylases (HDAC)
Histone methyltransferases (HMT)
Hypothalamic pituitary adrenal axis (HPA)
In vitro fertilization (IVF)
Intraclass correlation coefficient (ICC)
Long interspersed nuclear element (LINE)
Long tandem repeat (LTR)
Methylated cytosine (\textsuperscript{m}C)
Methylation binding domain (MBD)
Methylation pattern error rates (MPER)
Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE)
Methyl-CpG binding protein 2 (MECP2)
Methyl-DNA immunoprecipitation (MeDIP)
Methylene tetrahydrofolate reductase (MTHFR)
Monozygotic (MZ)
Multiphasic Personality Inventory-2 (MMPI-2)
Neuropeptide Y (NPY)
Online Mendelian Inheritance in Man (OMIM)
Ontario Mental Health Foundation (OMHF)
Peptidylprolyl isomerase E-like (PPIEL)
Polycomb group complexes (PcG)
Polymerase chain reaction (PCR)
Position effect variegation (PEV)
S-adenosyl methionine (SAM)
SET and RING associated (SRA)
Short interfering RNA (siRNA)
Short interspersed nuclear element (SINE)
Single nucleotide polymorphism (SNP)
SWItch/Sucrose NonFermentable (SWI/SNF)
Thioredoxin reductase 2 gene (TXNRD2)
Wechsler Abbreviated Scale of Intelligence (WASI)
Wechsler Adult Intelligence Scale (WAIS-I)
White blood cells (WBC)
Acknowledgments

I would like to thank all those who contributed to the completion of my PhD thesis. First of all, I would like to thank my supervisor, Art Petronis, who inspired me to pursue my degree in the first place and whose mentorship and faith in my abilities over the years has allowed me to succeed. I would also like to thank John Vincent and Rosanna Weksberg, my advisory committee members for their direction over the years. Additionally, I would like to thank Jim Kennedy, who always made time for me despite his busy schedule. I would also like to thank James Flanagan and Jon Mill for our numerous discussions that contributed to my scientific growth. Thanks to those who contributed to the data and analysis presented within including Abbas Assasadazeh, Sigrid Zeigler, Carolyn Ptak, Gabriel Oh, Jon Mill, Allan McRae, Peter Visscher, Grant Montgomery, Carl Virtanen, Neil Winegarden, Jill Cheng, Thomas Gingeras, Thomas Tang, Axel Schumacher, Philipp Kapranov, Sun-Chong Wang, Albert Wong, and Laura Feldcamp. I would also like to thank our collaborators, Anthony Feinstein, Jonas Halfvarson, Curt Tysk, Darlene Floden, and Nick Martin, for providing us with samples to study. I would like to acknowledge the Canadian Institutes of Health Research who funded my work with a Canadian Graduate Scholarship, Doctoral Award. Finally, I would like to thank my wife, Sharah Mar, for her support and confidence in me over the years.
Thesis Introduction

Epigenetics

Epigenetics refers to regulation of various genomic functions controlled by partially stable modifications of DNA and histones [1]. The epigenetic information is encoded in two types of synergistically acting covalent modifications: DNA methylation and chromatin protein modification [2]. In mammals, DNA methylation occurs most commonly on cytosines that are directly followed by guanine, forming what is known as a CpG dinucleotide. Clusters of CpG dinucleotides are referred to as CpG islands [3]. Methylated cytosine ($^{\text{met}}$C) is often referred to as the 5\textsuperscript{th} base of the genetic code; however its function is related to transcriptional control as opposed to DNA sequence based coding. There are numerous examples demonstrating that transcription factor binding affinity is directly limited by the presence of methylation at the binding sites [4, 5]. The density of $^{\text{met}}$C in a gene regulatory region also contributes to gene activity with a large number of genes exhibiting an inverse correlation between the degree of methylation and the level of gene expression [6, 7]. DNA methylation can regulate genomic functioning not only in terms of gene expression but also in the suppression of repetitive DNA sequences [8], and the formation of architecturally functional chromatin structures such as centromeric regions [9].

Methylation of DNA is mediated by proteins called DNA methyltransferases (DNMTs), four of which have been identified thus far[10]. These include DNMT1, DNMT2, DNMT3a and DNMT3b. DNMTL is believed to facilitate the activity of DNMT3a and 3b. DNMT1 is primarily the maintenance DNA methyltransferase, responsible for methylating hemimethylated daughter strands during DNA replication (\textit{ibid}). DNMT3a and 3b are believed to be responsible for \textit{de novo} DNA methylation events [11]. The functions of DNMT2 are not well characterized.

Passive DNA demethylation can be achieved via the binding of transcription factors during DNA replication[12] as well as through nuclear exclusion of oocyte specific DNA methyltransferase, DNMT1o, during the demethylation that occurs post fertilization in mammals [13]. The existence of a DNA demethylase capable of active DNA demethylation is a controversial subject as none of the proposed mechanisms or implicated genes have been replicated or have affected mouse viability in knock out experiments [12]. Recently, DNMT3a and 3b have been implicated in DNA demethylase activity [14]. The authors observed periodic and strand specific
methylation and demethylation of the PS2 gene promoter in response to oestrogen activation in cultured human breast cancer cells. Periods of demethylation coincided with chromatin immunoprecipitation of DNMT3a and 3b. The authors proposed a mechanism whereby methylcytosine is deaminated by the DNMTs in the absence of S-adenosyl methionine (SAM), resulting in a mismatch that is subsequently excised and repaired with non-methylated cytosine by DNA glycosylases (ibid). Unlike the other proposed DNA demethylases, knockout of DNMT3a causes embryonic lethality in female mice and impairs spermatogenesis in males [15]; however, the deleterious effects may be a reflection of this gene’s methyltransferase activity, not demethylase activity. Like the other implicated DNA demethylases, further study will be necessary to confirm its role in human development.

DNA modification acts in concert with alterations in chromatin structure that occur through acetylation, methylation, phosphorylation, ubiquitination, and sumoylation of various histone amino acid residues including lysine, arginine, and serine [16-21]. Histone acetylation and methylation are the most heavily studied modifications of the ‘histone code’, as it has been aptly named [2]. Histone acetylation mediated by histone acetyltransferases (HAT) and histone deacetylases (HDAC) unravel and compact chromatin, respectively, through acetylation and deacetylation of lysine residues on histone H4, respectively [22, 23]. Acetylated lysine residues are recognized by proteins containing a conserved bromodomain that subsequently recruit components of the basal transcription machinery [24]; however, alterations in histone acetylation status are believed to be a short term mechanism [25]. Methylation occurs at lysine and arginine residues via histone methyltransferases (HMT) [26]. Methylation of Histone 3-lysine 4 (H3-K4) is associated with euchromatin (a loosely packed chromatin state) where as dimethylation of lysine 9 (H3-K9) and trimethylation of lysine 27 (H3-K27) is associated with heterochromatin (a tightly packed chromatin state) [24]. Histone methylation is not restricted to lysine residues and may also occur at arginine residues at H3-R2, H3-R17, H3-R26, and H4-R3 [26]. The functions of such modifications are only just beginning to be understood [27]. The epigenetic regulation of the histone code is believed to be a dynamic process, switching between ‘on’ and ‘off’ states during transcriptional regulation [28]. Conversely, DNA methylation is believed to represent a more permanent mark denoting silenced genomic regions.

There is still much that is unknown about the scope and function of all aspects of the histone code and its synergistic interactions with DNA methylation; however, there is evidence for interaction between these two epigenetic signatures. Methylation of DNA recruits proteins
containing a methylation binding domain (MBD), termed MBD proteins[29]. The most well characterized MBD protein is methyl-CpG binding protein 2 (MECP2), a protein that binds methylated DNA and recruits HDAC complexes and ATP dependent chromatin remodeling complexes such as the SWItch/Sucrose NonFermentable (SWI/SNF) complex [30, 31]. These complexes deacetylate histones and compact chromatin, respectively, conferring a silenced chromatin state complementing the DNA methylation status [32].

Epigenetic information encoded in histone modifications may also direct DNA methylation patterning through interaction with the polycomb group proteins (PcGs), a heavily studied protein family implicated in maintaining the tissue specific epigenetic profiles as well as maintaining the pluripotency of embryonic stem cells [33]. PCGs bind genes involved in developmental pathways both in *Drosophila* and mice as evidenced by chromatin immunoprecipitation microarray (Chip on Chip) experiments binding a high proportion of developmental transcription factors upon differentiation in murine embryonic stem cells [34, 35]. PCGs in turn can direct DNMTs, modulating DNA methylation in specific genomic regions [36] [37] [38].

Other mechanisms for directing and maintaining epigenetic silencing exist through recruitment of HDAC and DNMT complexes by proteins recognizing and being directed by the sequences of non-coding and short interfering RNAs (siRNAs) in yeast, *Drosophila*, and plants [39, 40]. To date, the extent of the roles of non-coding RNA in transcription and genome stability is not well understood [41] although these mechanisms have been implicated in X chromosome inactivation[42] and silencing of repetitive elements[43].

**Epigenetic signals and biological development**

Epigenetic signals are necessary for the proper regulation and functioning of the genome [1], with epigenetic mutations, or epimutations, having the potential to be as harmful to an organism as genetic mutations. Knockout mice with homozygous deletions of DNMT1 exhibit embryonic lethality. In addition to regulation of gene activity [4, 44-48] epigenetic factors may affect DNA mutability [49] and genetic recombination [50]. Epigenetic patterns are established in a tissue specific manner and are believed responsible for establishing and maintaining the cellular identity of the >200 cell types in the human body[51-54]. DNA methylation controls lineage commitment in hematopoietic cells, with aberrant methylation observed in B-cell lymphoma cell
lines [55]. These studies and such examples highlight the importance of the epigenetic code to a properly functioning genome [44, 56].

**Epigenetic metastability**

The epigenetic status of genes and genomes is far more dynamic in comparison to the DNA sequence and is subject to changes under the influence of developmental programs, in the presence of internal or external environmental epigenetic modifiers, or simply as a result of stochastic processes relating to maintenance of epigenetic factors.

Numerous lines of evidence suggest that DNA methylation undergoes a stochastic rearrangement referred to as metastability. Cell culture models of higher eukaryotic systems have demonstrated that metastability can result from the relatively low fidelity of the DNA methylation maintenance enzymes, such as DNMT1, as compared to that of the DNA repair machinery. DNMT1 has a preference for binding hemimethylated CpGs in a manner dependent on the sequence content. Acting alone, DNMT1 has a 30-fold affinity for long CpG stretches and 5 fold for randomly dispersed CpGs [57]. Further protein interaction is required for a more stable transmission of DNA methylation during replication. DNMT1 is recruited to replicated DNA by the ubiquitinated NP95 protein [57], which preferentially binds hemimethylated DNA via a SET and RING associated (SRA) domain. NP95 then sequesters the N-terminal domain of DNMT1, which subsequently methylates the replicating daughter strand (ibid). Tissue culture experiments have identified a range of maintenance DNA methylation fidelity from 97% to 99.9% as well as an additional fluctuation of 3-5% per mitosis in the form of de novo methylation [58]. In mice, DNMT1 was found to methylate hemimethylated double stranded DNA with a fidelity of ~95% [59]. This translates to a difference of roughly 3 orders of magnitude lower mitotic fidelity of epigenetic patterns as compared to the DNA sequence (10^-6 and 10^-3 for DNA sequences and DNA modification, respectively)[58]. It is clear that a portion of DNA methylation signals will be lost or gained through thousands of replication events resulting in a DNA sequence independent drift of epigenetic signals.

Another classic example of stochastic epigenetic regulation is position effect variegation (PEV). PEV was first observed in 1930 in Drosophila when the expression of the White gene responsible for white or red eye color manifested a mosaic pattern [60]. Such mosaic expression was dependent on chromosomal rearrangements placing the White gene proximal to epigenetically silenced heterochromatic regions, resulting in variable silencing [60, 61]. Models
for epigenetic silencing include a heterochromatic spreading in \textit{cis} as well as effects mediated in \textit{trans}, possibly through spatial interactions and chromatin folding [60]. Importantly, maintenance of the heterochromatic state of silenced genes is transmitted in a metastable state through multiple cell divisions [60]. PEV highlights that the borders of epigenetically silenced regions of the chromatin are not fixed and can affect the silenced status of neighboring regions.

Effects similar to PEV may be caused by aberrant regulation of repetitive elements, repetitive DNA code of retroviral origin that comprises approximately 45\% of the human genome [62]. Repetitive elements are often classified as long and short interspersed nuclear elements (LINEs and SINEs) and long tandem repeats (LTRs) [63]. Epigenetic silencing via DNA and histone methylation occurs at repetitive elements, which serves to silence their expression and inhibit their retrotranspoable capability[62, 64]. Imperfect silencing of retrotransposons during embryonic epigenetic reprogramming has been suggested to result in mosaic expression of proximal genes [65].

**Epigenomic profiling technologies**

The past two decades have seen a dramatic increase in the available technologies for epigenetic profiling, both at individual loci and at the genome wide level, have led to a promising beginning to the profiling of the epigenome. Many of the new methods reflect an assimilation of existing high throughput genome scanning technologies such as microarrays following a refitting to meet the complexities of epigenetic studies. The primary epigenetic technologies can be broken down into two categories. The first method is specific to DNA methylation and involves the chemical treatment with sodium bisulfite, which deaminates all un-methylated cytosines to uracil while methyl cytosine remains protected. This procedure produces sequence polymorphisms through subsequent PCR amplification that can be detected with a variety of new techniques developed in the last years, including the manuscript in chapter 2. These techniques allow a site specific quantification of DNA methylation levels to within 5\% with the resolution of 1 bp, allowing for the identification of epimutations of individual CpGs with possible functional relevance. Such techniques are invaluable for understanding the functional consequences of DNA methylation changes at individual genes. The resolution of sodium bisulfite modification makes it the ‘gold standard’ method of DNA methylation quantification; however, its applications lack the scope of the second method, the genome wide technologies.
The second involves the segregation of the desired components of the genome, either with antibodies specific to a chromatin modification or DNA methylation or through the selective cutting of methylation sensitive restriction enzymes that will only cut at specific non-methylated consensus sequences. This is then followed by the identification of the isolated sequences through hybridization to microarrays or sequencing techniques. Such techniques allow a great increase in the scope of epigenetic studies, often achieving a genome wide interrogation but usually lacking in resolution.

**Site specific DNA methylation profiling**

The use of epigenomic microarray technology may identify regions of DNA methylation difference of interest in the studies in which it is employed; however, in order to approach a functional understanding of such differences, a more detailed investigation of DNA methylation status at individual CpGs is necessary. While microarray based techniques will deliver a fold difference indicative of levels of methylation relative to other samples, such as in a twin vs. a cotwin or a case vs. a control comparison, these measures do not address the actual quantitative percentage (from 0% to 100%) of DNA methylation in a region. Conversely, high throughput sodium bisulfite modification techniques allow a site specific quantification of DNA methylation levels to within 5% with the resolution of 1 bp. Such methods include pyrosequencing [66] and methylation sensitive single nucleotide primer extension (Ms-SNuPE) [67-71]. Each technology will be reviewed in more detail in the manuscript in Chapter 2; however, a key facet of each technique is the necessity to anneal a primer within the region of interest. In epigenetically interesting CpG rich regions such as CpG islands, the DNA methylation status of CpGs within the region may introduce polymorphisms at this positions following sodium bisulfite modification. Up until recently, the effects of polymorphisms within the primer annealing regions were unknown but seemingly ignored[69-71]; however, the work published in chapter 2 demonstrates the potential bias introduced by such polymorphisms. It further presents a method to correct this, allowing a non-biased measurement of DNA methylation levels. The solution presented in chapter 2 can be applied to the pyrosequencing technologies as well.

**Epigenomic microarray technologies**

The most heavily used epigenomic technologies involve the interrogation of tens of thousands of genomic regions using microarrays, followed by an in depth interrogation of specific targets
identified through sodium bisulfite modification methods. The microarray based studies are limited to the array platform employed, of which there are a number available with varying resolutions, such as spotted oligo arrays ranging from 12,000 loci of about 1kb in length such as the human CpG island microarray [72] to tiling arrays with ~40 million probes spaced at 35 bp intervals covering the entire non-repetitive genome (http://www.affymetrix.com/index.affx). Microarray probes are often designed to highlight key functional areas, like CpG islands, gene promoters, exons, and 3’UTRs, amongst others. In this way, these platforms fulfill a form of candidate gene approach, but over tens or hundreds of thousands of regions. Probe sequences on microarray platforms represent the 4 bp code DNA sequence as opposed to the degenerated 3 bp code compatible with sodium bisulfite modified sequences. One technique has combined sodium bisulfite modified gene products with Illumina GoldenGate SNP genotyping microarrays to quantitatively measure DNA methylation [73], but this technology is limited to this panel of candidate CpG sites.

The vast majority of alternate techniques, including MeDIP, HELP, McrBC, CHARM, and the method outlined in chapter 1 for which there is no acronym, distinguish a quantitative level of DNA methylation from hybridization to the unmodified genomic DNA sequence. These methods allow a segregation of the genomic DNA that is either methylated or unmethylated, depending on the technique, after which comparative microarray hybridization identifies relative DNA methylation differences between sample cohorts of interest. MeDIP is an immunoprecipitation technique with an enzyme specific to methylated cytosine [74]. The other techniques use restriction enzymes that cut based on cytosine methylation status, followed by adaptor ligation and PCR enrichment. HELP and our method use Hpa II methylation sensitive restriction enzymes to enrich the unmethylated fraction of genomic DNA, while McrBC and CHARM employ a similar but reverse strategy, using the MCRBC enzyme to cut and enrich only methylated cytosine, with CHARM followed by a genome smoothing algorithm that increases data accuracy [75]. Additionally, while CHARM, MCRBC, and our method are compatible with the standard common reference, balanced block, and randomized hybridization designs, the hybridization design employed with HELP is limited, requiring a parallel enrichment produced by methylation insensitive isoschizomers (MspI). The detected fold differences between the HpaII and MspI digestion are meant to control for single nucleotide polymorphisms (SNPs) within the restriction sites; however, some evidence suggests that the enrichment dynamics may vary between the two enzymes, ultimately producing inaccuracies [75]. This design limits the
comparability of the detected DNA methylation levels between samples, as there is no common reference between samples. The HELP assay does highlight a potential flaw in the other enzyme based methodologies, namely, that SNPs within restriction enzyme consensus sites will inhibit cutting in the same way as a DNA methylation difference, potentially causing false positive epigenetic differences that are in fact, genetic differences.

A recent study compared the accuracy of the HELP, MCRBC, MeDIP and CHARM techniques against a panel of ~1466 CpG DNA methylation patterns tested by sodium bisulfite modification followed by hybridization onto Illumina Golden Gate SNP genotyping arrays [75]. MeDIP was found to be unreliable when compared to the reference panel, whereas the MCRBC and HELP techniques performed best in regions where DNA methylation differences were within 200–600 and 700–1200 bp of each other, respectively. The CHARM technique requires specific custom built microarrays in order to apply the smoothing algorithm, which consists of an averaging of neighboring DNA methylation measurements for a given region [75]. Unfortunately, the method used in our laboratory was not evaluated against such a large panel of CpGs; however, this method was validated at select loci by sodium bisulfite sequencing in a number of studies [67, 76-78].

Epigenetics has the potential to address a number of the yet unsolved mysteries that are hallmarks of complex non-Mendelian biology that cannot be explained by the traditional genetic paradigm. These include twin discordance, sex effects, parental origin effect, fluctuating course, presence of familial and sporadic cases, and late age at onset of disease, amongst others. The second half of my thesis is dedicated to the application of the developed technologies to epigenomic studies of twins.

The classical twin design

A comparison of phenotypic concordance in monozygotic (MZ) to dizygotic (DZ) twins has long been one of the most elegant systems through which to infer the influence of inherited factors on a trait. In the traditional model, the DNA sequence represents the sole inherited factor while all other influences are attributed to environmental factors. MZ twinning occurs when a single fertilized egg produces two embryos, while DZ twins result from two eggs being fertilized by two separate sperm [79]. MZ twins share approximately 100% DNA sequence identity while DZ twins on average share 50% of all segregating DNA polymorphisms [80-82]. In the traditional
model, classical twin studies divide the population variance for a given trait into the following components, where:

Differences between DZ pairs = G/2 + E

Differences between MZ pairs = E

where G represents the genetic contribution or DNA sequence based factors and E represents the environmental factors contributing to the phenotype. Algebraically solving for the genetic influence within the population for a trait gives:

G = 2*(DZ difference – MZ difference).

This level of genetic influence has been termed ‘Heritability’ (H) [81, 83]. It should be noted that in the brain, instability of repetitive elements have been linked to behavioral variation both in vitro and in vivo in the vole [84]. Additionally, variation of repeat length has been linked to anticipation in neurological diseases in humans such as Huntington’s disease and fragile X mental retardation, amongst numerous others [85]. Importantly, such repeat instabilities appears to occur in the developing embryo [86, 87]; thus, not all genetic variation resulting in phenotypic change in the brain is inherited genetic variation. Finally, repeat instability has been observed to result in twin discordance such that MZ twin brain tissue may vary phenotypically as a result of genetic discordance at repetitive elements [88]. Despite these complexities, for quantitative traits, the correlation between twin groups serves as the most common measure of co-twin similarity, or inversely, co-twin difference. The standard correlation measure employed in classical twin designs is the intraclass correlation coefficient (ICC) [89]. The ICC allows for a correlation between twin groups independent of an a priori segregation of each twin of a pair into one group or another. A distinct advantage of this approach is that it brings the variance of each group to a common statistical base, that is, within a range of -1 to 1. This allows heritability estimates produced by twin comparisons to be equivalent to those measures produced by other methods of estimating the genetic influence on a trait; such as, between vs. within family comparisons [89]. Advantages of the classical twin design over family based comparisons are that age and environmental experiences of DZ twins are more tightly controlled than in comparisons of non-twin siblings [81, 82]. Including only DZ twin pairs with shared environments and that are of the same gender are critical assumptions of the classical twin design. When using ICCs, the equation for heritability becomes:
H= 2*(ICC_{MZ}-ICC_{DZ})

Based on such models, the proportion of genetic and environmental influences contributing to a phenotype can be calculated based on such assumptions and have subsequently directed the research efforts to elucidating the specific causes implicated in such studies. Performing a search on the NCBI’s Online Mendelian Inheritance in Man (OMIM) site under “twin” and “disease” yields over 250 entries where genetic factors have been implicated by classical twin studies. For example, classical twin studies on one of the most common psychiatric disorders, major depressive disorder, determined a heritability range from 0.36 to 0.7 [90-95]. An evaluation of 36 MZ and 53 DZ twins found panic disorder to be twice as frequent in MZ twins than DZ twins [96] suggesting a heritability of 1 or that the disease is caused purely by genetic factors. Calculation of heritability estimates based on proband wise concordance for bipolar disorder from 55 MZ and 54 DZ twins from the Danish twin registry is 0.94 [97] with recent studies in alternative populations reaching the same conclusions [98]. An analysis of five twin studies in schizophrenia all reported heritability values ranging from 0.80 to 0.85 [99]. The field of behavioral genetics has employed the classical twin design to fuel the debate regarding ‘Nature’ and ‘Nurture’s’ influence on human behavior [100]. A Pubmed search on the terms “twin studies” and “behavior” yields over 1330 citations. Some recently investigated traits include children’s food preferences [101], beverage intake [102], antisocial behavior [103], and sexual excitation in men [104]. The results and interpretation of such modern twin studies have fueled the search for the implicated genetic and environmental factors; however, a closer inspection of the assumptions of the classical twin design shows that the design may not be as elegant as it seems.

**Epigenetics and the classical twin design**

The intensive study of complex traits and disease over the last several decades has been facing significant difficulties in identification of both molecular genetic and specific environmental factors, which warrants re-thinking the fundamental principles of human biology and their interpretation.

**MZ twin discordance**

One hallmark of complex non-Mendelian disease is the observation of MZ twin discordance.
Proband-wise MZ concordance for major depression is 31% for male- and 48% for female- MZ twins [105], 62% -79% in bipolar disorder[97] , and 41%-65% in schizophrenia [99] . Traditionally, the levels of discordance are attributed to differences in environmental effects between twins. As these environmental influences must differ between the co-twins, this environmental influence has been termed non-shared environment [106]; however, experimental evidence does not support a strong role for non-shared environment in phenotypic outcome.

A number of studies have taken advantage of adoption registries and measured phenotypic similarity in MZ twins that were reared together as compared to those reared apart. The most famous of these studies, the Minnesota study of twins reared apart tracked > 200 twin pairs longitudinally from 1970 -1990 and measured a number of behavioral characteristics, temperament, leisure time, occupation, and social interests [107]. Independent of the environmental influences over the course of these 20 years, the MZ co-twins in both groups were remarkably similar in the correlations for these measures [108]. While perhaps the largest of its kind, the Minnesota study is not the only adoption study to address the issue of non-shared environment. A review of studies from 1920 to 1987 demonstrated that MZ twin concordance rates for schizophrenia were ~46%, independent of whether twins were reared together or apart [109]. Twin discordance in atopic disease did not vary in measures of asthma, rhinitis, skin-test response, and serum IgE levels between twins reared together and apart [110]. Levels of type A-like behavior was evaluated in a large Swedish Adoption/Twin study cohort including 229 and 160 MZ twins reared apart and together, respectively, where no evidence for non-shared environment was found [111]. In addition to humans, a considerable degree of phenotypic variability has been observed in inbred and cloned animals, which contain minimal genetic variation, even after strictly controlling for environmental variation[112-115]. It is apparent from these examples that for a number of phenotypes, environmental influence alone is insufficient to explain the observed levels of discordance between genetically identical organisms.

The metastable nature of epigenetic signals on one hand and their primary role in determining various phenotypes on the other makes them an ideal candidate to account for phenotypic differences in genetically identical organisms including identical twins. A number of studies have identified epigenetic differences in MZ twins. In an investigation of skin fibroblasts from 5 female MZ twin pairs discordant for Beckwith Wiedemann syndrome, an imprinting defect was identified at the *KCNQ1OT1* gene of 11p15 only in the affected cases[116]. The authors suggested this locus may be vulnerable to DNA methylation maintenance aberrations occurring
during preimplantation and that the epimutation itself may predispose to the twining event (ibid).

A large scale investigation of greater than 300 twins demonstrated substantial variation of DNA methylation levels a differentially methylated region (DMR) associated with the H19/IgF2 locus[117]. DNA methylation patterns of pairs of MZ twins discordant for schizophrenia investigated in the promoter region of the DRD2 gene demonstrated that disease affected individuals were more epigenetically similar to each other than to unaffected co-twins [118]. After failing to identify sequence differences between a pair of MZ twins discordant for caudal duplication, a sodium bisulfite modification based investigation revealed differential methylation of a CpG island functioning as a promoter in the AXIN1 gene and was believed responsible for discordance [119]. This methylation discordance was subsequently identified in affected singletons. A methylation sensitive restriction screening method was applied to a pair of MZ twins discordant for bipolar disorder. Sodium bisulfite modification based analysis through pyrosequencing revealed an altered DNA methylation pattern at the peptidylprolyl isomerase E-like (PPIEL) gene, which correlated with expression differences between the twins [120]. Fraga et al. performed the first large scale investigation of epigenetic differences in twins identifying 35% of twins (N=40) displaying significant global DNA methylation differences and histones H3 and H4 acetylation differences using an HPLC based detection method [121]. Twins displaying large global epigenetic differences corresponded to global gene transcription differences as measured by mRNA hybridization to Affymetrix Human U133 Plus 2.0 gene chips. Using a restriction enzyme based enrichment technique, the authors enriched the methylated fraction of genomic DNA, cloned it into plasmid vectors and sequenced, identifying similar percentages of DNA methylation difference corresponding to the twin pairs identified by global results. Hybridization of these enriched products from pairs of variable co-twins to metaphase chromatin demonstrated a higher density of epigenetic variation in the telomeres and several gene rich regions (ibid). A metastable drift of epigenetic factors could account for the observations of twin discordance in complex traits. A high resolution genome wide epigenetic profiling effort in MZ twins as performed in chapter 4 would significantly benefit our understanding of genomic regions displaying epigenetic metastability.

A number of studies have shown that epigenetic factors themselves can be subject to environmental influence. In a study performed by Weaver et al, it was observed that the arched back nursing and grooming behavior of rat mothers resulted in a DNA methylation change at a critical CpG located within the NGFI-A transcription factor binding site within the promoter of
the glucocorticoid receptor gene [122]. This was identified in the hypothalamus, directly affected transcription factor binding, and ultimately mediated hypothalamic-pituitary-adrenal (HPA) axis based responses to stress. The DNA methylation change was also associated with increased histone acetylation, which was reversed with infusion of an HDAC inhibitor, abolishing the phenotype (ibid). Through dietary supplementation of vitamin B12, folic acid, and betaine in early inbred mouse diet, Waterland et al., observed an increase in DNA methylation at the agouti $A^{vy}$ locus associated with phenotypic differences in the agouti inbred mouse strain [123]. These studies highlight a direction functional link between an environmental influence and phenotype mediated by epigenetic mechanisms. In twin studies demonstrating evidence for phenotypic variation induced by non-shared environment, epigenetic changes may reflect such influences in a molecular record. Epigenetics may therefore provide a tangible and measurable molecular substrate that is indicative environmental influence that can affect discordant phenotypes in genetically identical organisms.

**Epigenetic inheritance**

The second problem with traditional heritability studies is that, while twin studies implicate genetic factors in causing various complex non-Mendelian diseases, to date, strikingly few genetic mutations have been identified to account for disease manifestation. In the past 20 years, there has been a major effort into the development of experimental and computational strategies for the identification of molecular causes of human complex disease. Unlike simple Mendelian disorders, for which the cloning of disease genes has become a routine procedure, identification of the molecular genetic basis of complex disease represents a major challenge to human biologists. Although the genes for some rare early onset and familial cases of complex diseases (e.g. colon cancer, breast cancer, and Alzheimer's disease) have been identified, the overwhelming proportion of non-Mendelian pathology remains unexplained. Research in complex non-Mendelian diseases is reaching an unprecedented level of sophistication and scale: from building massive databases of DNA sequence variants (single nucleotide polymorphisms, SNPs; HapMap, haplotype maps) to screening thousands of polymorphisms in thousands of affected individuals and controls. Why is it that complex diseases are so resistant to the discovery of the aetiological factors?

Some argue that the lack of replicated findings is a reflection of the heterogeneity of the samples. The concept of the endophenotype arose in an attempt to reduce the heterogeneity of diseases
with broad diagnostic criteria by identifying sub-phenotypes possibly indicative of a more homogeneous form disease with common aetiological factors [124]. Heritability studies have been performed on endophenotypes of schizophrenia [125], depression [126] and nicotine dependence [127], amongst numerous others. A recent meta analysis of endophenotype association studies determines, however, that the effect sizes produced in these studies are no larger than those produced by associating based on disease diagnosis [128].

Another approach has been to increase the sample sizes of traditional genetic association techniques to upwards of ~2000 cases and controls in order to perform genome wide association studies (GWAS). In such studies, dense microarray based SNP genotyping platforms allow a screening of all haplotype blocks in the genome through the investigation of TAGsnps in linkage disequilibrium with surrounding SNPs. The use of genome wide association studies have provided successful identification of genes associated with some diseases including Crohn’s disease and type II diabetes and is a promising start to understanding the genetic architecture for some disorders [129]. While studies of this kind can identify regions conferring risk to disease; however, the functional implications of the risk genotypes are not always clear [130]. For example, risk genotypes identified in type II diabetes and Crohn’s disease can account for only 3% and 10% of the phenotypic variance of these diseases, respectively [130].

While most of the complex non-Mendelian diseases under investigation are believed to be multifactorial, resulting from complex interactions of numerous gene pathways and environmental stimuli, the inability to define common genetic variants of strong effect despite increasingly large sample sizes and reduced heterogeneity calls into question the targets under investigation. Have classical twin studies been pointing to the right targets or are there other factors that should be considered?

Evidence for the existence of such factors has been produced in an elegant experiment performed by Gartner and Baunack [113, 131]. Through the splitting of murine blastocysts, the authors were able to create inbred MZ twin mice that originated from the same germ cell, which they could then compare to inbred polyzygotic mice fertilized by separate germ cells. This scenario is akin to the means by which MZ and DZ twins are fertilized, respectively. After strictly controlling for environment and employing a classical twin design analysis, approximately ~75% of the variance in body weight was determined to result from a yet unidentified “third component”, independent of genetic and environmental factors. This experiment created a
system where the only difference between the zygosity groups, was the status of contributing germ cells \textit{(ibid)}. Therefore, the third component identified in these experiments is an inherited factor independent of the DNA sequence and environment. These results suggest that perhaps the majority of classical twin designs are pointing to molecular factors beyond the DNA sequence and environment that are influencing phenotypic outcome, namely, epigenetic factors.

In mammals it is conventionally believed that there is no passage of epigenetic information from parent to offspring generations due to the massive epigenetic rearrangements occurring during gametogenesis necessary for development and critical for re-establishing parent specific genomic imprints [21, 132-135]. These include a global demethylation of DNA accompanied by massive histone modification rearrangement [136]. Epigenetic reprogramming of the germline cells means that passage of non-genetic information in mammals is distinctive from traditional neo-Lamarckian inheritance, which postulates that any adaptive changes acquired during the life of the organism are transmitted to the offspring [137]. Such a scenario is more similar to the inheritance of epigenetic factors in plants, as plant germline cells are derived from the somatic tissues of the mature organisms and erasure of epigenetic information in these cells is less extensive [137]. In mammals, any adaptive or ‘soft’ inheritance of this sort is contingent on an environmental influence that affects the epigenetic status of the germline. However, even a passage of epimutations occurring in the germline spontaneously from stochastic rearrangements has the potential to influence phenotype and will influence the results of classical twin studies.

There are a number of examples in humans and mice in which epimutations in the parent germline result in molecular and phenotypic changes in the offspring. In the case of some epigenetic patterns that are inherited transgenerationally, such as in the \textit{agouti A}\textsuperscript{vy} and \textit{Axin}\textsuperscript{Fu} genes in mice [138, 139], a hypomethylated state of retrotransposons upstream of these genes results in ectopic expression and penetrance of the phenotype. For the \textit{agouti} phenotype, the hypomethylated state results in mice with a yellow coat color and is also associated with obesity[137]. Loss of epigenetic transposon silencing at the \textit{Axin}\textsuperscript{Fu} locus results in a kinked tail[137]. Both the \textit{agouti A}\textsuperscript{vy} and \textit{Axin}\textsuperscript{Fu} genes exhibit a degree of parent of origin inheritance and it is hypothesized that differential treatment of DNA methylation states between the male and female germline during embryogenesis could result in varying resistance of these retrotransposons to epigenetic silencing [138]. Although there is no conclusive evidence of transgenerational epigenetic inheritance in humans, an epimutation was identified in the tumor
suppressor gene, *MLH1*, in spermatozoa of individuals with colorectal cancer [140], which may reflect an escape from reprogramming during gametogenesis.

A number of recent studies supply intriguing possibilities as to the nature and identity of the molecular substrates facilitating a carryover of epigenetic information through the male germline that are capable of contributing to the epigenotype of the developing embryo. During spermatogenesis in the male germline, ~85% of histones are replaced by protamines, achieving a tightly packed chromatin conformation while histones H2A, H2AX, H2AZ, H2B, H3.1, H3.3, CenH3 and H4 remain incorporated in the mature sperm [141, 142]. The remaining 15% of core histones are sufficient to package the entire coding sequence of the genome and have been proposed to mediate a passage of meiotic information to subsequent generations independent of genomic sequence context [142]. Male germline mediated passage of histone H3 variants, as identified in humans [141] and mice [143], have been suggested to mark genomic imprints and to protect them from the global demethylation events occurring after fertilization [142].

There is growing evidence that histone modifications conferred to the developing embryo help to direct the behavior of DNA methylation reprogramming that occurs in early embryogenesis. After fertilization, the maternal and paternal chromatin remains distinguishable with the maternal and paternal germlines retaining primarily s-phase H3 and non-s-phase H3.3 variants, respectively [143]. The male germline histone content appears to be linked with levels of DNA methylation in the paternal zygotic genome. Injection of round spermatids that have yet to undergo protamine condensation into oocytes results in significantly higher DNA re-methylation after global DNA de-methylation in the fertilized embryo [144]. Treatment of round spermatid infused zygotes with HDAC inhibitor trichostatin A led to decreased DNA methylation further suggesting a functional link between histone code content and paternal DNA methylation and gene expression [144]. Finally, a recent interrogation of DNA methylation status in the gametes, zygotes, and blastocysts of mice carrying the inherited *agouti A* methylation phenotype demonstrated that DNA methylation is not the inherited mark associated with maternal transmission, despite its eventual presence in the mature organism [145]. Interestingly, parent of origin specific inheritance of the phenotype was abolished in mice that where haploinsufficient for a protein that recognizes the suppressive H3-K27 methylation mark (*ibid*). Taken together, these experiments suggest the histone content transferred through fertilization by the paternal genome to the zygote may be critical for establishing the gene expression necessary for proper
development. This interpretation is consistent with observations that men with abnormal histone to protamine ratios suffer from infertility problems [146-148].

There is a mounting body of evidence demonstrating that epigenetic signals can be transmitted across generations. While there are striking examples of non-genetic regulation of the phenotype of offspring generations, such as the *agouti* A<br>mice reviewed above, evidence for such phenomenon in humans remains poorly investigated. A passage of epigenetic signals from parent to offspring generations will have implications to classical twin studies and the study of complex non-Mendelian disease.
Thesis Objectives

Chapter 1

1.) Epigenetic studies have the potential to help uncover the molecular basis of complex traits; however, the available technologies are limited to locus specific analyses and therefore limit the scope of epigenetic studies. Our objective was to create a high throughput technique capable of highly parallel interrogation of methylation status at thousands of DNA loci.

Chapter 2

1.) The use of Ms-SNuPE based reactions for the measurement of DNA methylation levels at single CpG positions has been employed in a number of studies; however, the effect of designing extension primers to anneal in regions with potentially polymorphic CpGs is unknown. The objective was to optimize Ms-SNuPE reaction conditions for partially complementary extension primers and estimate possible bias the accuracy of the quantified DNA methylation level.

Chapter 3

1.) Given the potential for epigenetic differences to account for phenotypic differences, the ability to map such locus-specific epigenetic differences becomes of primary interest. To achieve such a goal, an estimate of adequate sample and effect sizes must be obtained to ascertain the statistical power necessary to find epimutations. We have employed our epigenetic microarray profiling technique on a single pair of MZ twins discordant for personality traits including stress response. The objective is to estimate the power of the technique as a function of sample size relative to the population variance.

Chapter 4
1.) Epigenetic metastability of DNA methylation could result in a drift of epigenetic patterns over time in genetically identical organisms such as MZ twins. While a number of studies have identified epigenetic differences indicative of this phenomenon in MZ twins in a limited context, epigenetic metastability has not been investigated on an epigenome wide level. The **first objective** was to estimate and map detectable DNA methylation differences between MZ co-twins in WBC, buccal epithelial cells, and gut tissues.

2.) With the exception of genomically imprinted loci, there is little evidence for the passage of epigenetic information through the germline. Such a passage of epigenetic information could have profound consequences to inherited phenotype and may provide an explanation for complex modes of inheritance observed in non-Mendelian diseases. To date, no classical twin studies comparing DNA methylation levels of MZ to DZ twins on a genome wide level have been performed. The **second objective** was to compare the degree of DNA methylation variation between age and sex matched MZ and DZ twins. This was tested in buccal epithelial tissue and WBCs.

3.) Evidence for a greater DNA methylation variability in the DZ twin group could be the result of two variables in the above design, namely, a passage of epigenetic differences through the germline and/or a contribution of DNA sequence variants in the DZ twin group to epigenetic variability. The **third objective** was to perform a comparison of DNA methylation variation in genetically identical (inbred) mice to that of genetically non-identical (outbred) mice to enable an estimation of the effect of DNA sequence variants on genome wide epigenetic profiles.
Chapter 1

Microarray-based DNA Methylation Profiling: Technology and Applications

Axel Schumacher¹, Philipp Kapranov², Zachary Kaminsky¹, James Flanagan¹, Abbas Assadzadeh¹, Patrick Yau³, Carl Virtanen³, Neil Winegarden³, Jill Cheng², Thomas Gingeras², & Arturas Petronis¹,⁵

Originally Published in Nucleic Acids Research

¹The Krembil Family Epigenetics Laboratory, Centre for Addiction and Mental Health, 250 College St, Toronto, ON, Canada M5T 1R8

²Affymetrix, Santa Clara, USA

³The Microarray Centre, The University Health Network, 200 Elizabeth St., Toronto, ON, Canada M5G 2C4

⁵To whom correspondence should be addressed: The Krembil Family Epigenetics Laboratory, Rm 28, Centre for Addiction and Mental Health, 250 College St, Toronto, ON, Canada M4T 1R8

Phone: +1-416-5358501-4880

Fax: +1-416-979-4666

e-mail: Arturas_Petronis@camh.net

Contributions: I contributed a considerable effort necessary for the completion of this manuscript. I helped to develop the technique itself through numerous experiments including the optimization of conditions in all steps of the enrichment of the unmethylated fraction of genomic DNA. I performed experiments evaluating the effects of McrBC digestion on method specificity and performed comparative hybridizations between the enriched unmethylated and methylated fractions. I performed numerous hybridizations to evaluate replicability of the method, schizophrenia vs. matched control hybridizations, and cross tissue comparisons. I performed a portion of the data analysis and helped to validate the microarray technology using sodium bisulfite modification.
Summary

This work is dedicated to the development of a technology for the unbiased, high-throughput DNA methylation profiling of large genomic regions. In this method, unmethylated and hypermethylated DNA fractions are enriched using a series of treatments with methylation sensitive restriction enzymes, and interrogated on microarrays specifically designed for epigenomic studies. In this study we have investigated various aspects of the technology including its replicability, informativeness, sensitivity, and optimal PCR conditions using microarrays containing oligonucleotides representing 100 kb of genomic DNA derived from the chromosome 22 **COMT** region in addition to 12,192 element CpG island microarrays. Several new aspects of methylation profiling are provided, including the parallel identification of confounding effects of DNA sequence variation, the description of the principles of microarray design for epigenomic studies, and the optimal choice of methylation sensitive restriction enzymes. We also demonstrate the advantages of using the unmethylated DNA fraction vs. the methylated one, which substantially improve the chances of detecting DNA methylation differences. We applied this methodology for fine-mapping of methylation patterns of chromosomes 21 and 22 in eight individuals using tiling microarrays consisting of over 340,000 oligonucleotide probe-pairs.
Introduction

Over the last decade the field of DNA methylation has grown dramatically and become one of the most dynamic and rapidly developing branches of molecular biology. The methyl group at the 5th-position of the cytosine pyrimidine ring, that is present in about 80% of CpG-dinucleotides in the human genome, can be of major functional significance and is regarded as the ‘fifth base’ of the genome [149]. DNA methylation, along with histone modifications (acetylation, methylation, phosphorylation, etc), are referred to as epigenetic phenomena that control various genomic functions without a change in nucleotide sequence [1]. Such functions include meiotic and mitotic recombination, replication, control of “parasitic” DNA elements, establishing and maintenance of gene expression profiles, X chromosome inactivation, and have a putative role in developmental programming and cell differentiation [140, 150-152].

Aberrations in epigenetic regulation, or ‘epimutations’, cause several paediatric syndromes (Prader-Willi [OMIM #176270], Angelman [OMIM #105830], Beckwith-Wiedemann [OMIM #130650], and Rett [OMIM #312750]) [153] and may also predispose to cancer [154].

Our understanding of the peculiarities of DNA methylation in the human genome is still very superficial. Based on the review of available publications, our estimate is that less than 0.1% of the genome has been subjected to a detailed DNA modification analysis. The recently completed Human Genome sequencing project did not attempt to differentiate between methylated and unmethylated cytosines. To some extent our understanding of the dynamic state of genome-wide DNA methylation has been hampered by the lack of high throughput technologies that would interrogate DNA methylation profiles over large genomic regions. A gold standard technique in DNA methylation studies, the bisulfite modification-based fine mapping of \textsuperscript{m}C [155], although precise, is very labour intensive and in most cases limited to short DNA fragments, often less than a kilobase.

The advent of microarray technologies that enabled the interrogation of a large number of DNA/RNA fragments in a highly parallel fashion has opened new opportunities for epigenetic studies [156]. A number of microarray-based technologies used for epigenetic analyses are already available [157-169]. However, all of these methods have some limitations, which renders them unsuitable for some experimental setups. Additionally, many technological parameters, such as the influence of DNA sequence variations, amplification conditions, and sensitivity of the methods have not been investigated before. Here we present a detailed analysis of various
parameters of epigenetic profiling and provide a substantially improved microarray-based high throughput technology for DNA methylation profiling of DNA regions that span from hundreds of kilobases to megabases. Eventually, this technology will be applied to the entire human genome, as exemplified by the methylation mapping of chromosomes 21 and 22.

Results

Enrichment of the unmethylated fraction of genomic DNA

The strategy for enrichment of unmethylated portions of the genome is presented in Figure 1. Genomic DNA is digested with methylation-sensitive restriction enzymes (Fig.1, middle panel). Whereas methylated restriction sites remain unaltered, the sites containing unmethylated CpGs are cleaved by the enzymes, and DNA fragments with 5’-CpG protruding ends are generated. The proportion of interrogated CpG sites depends on the methylation sensitive restriction enzymes used for the restriction of DNA. Based on our analysis of the CpG dinucleotides within the sites of methylation sensitive restriction enzymes across several megabases of human genomic DNA, the combination of three enzymes, $Hpa$II, $Hin6$I, and $Aci$I, should interrogate ~32% of all CpG dinucleotides in mammalian DNA (Tab.1). The addition of two other relatively inexpensive methylation-sensitive CpG-overhang generating enzymes, $HpyCH4IV$ and $Hin1$I, would theoretically increase the proportion of interrogated CpGs to ~41%. Depending on the microarray-type, in our experiments we usually use either a single enzyme or a ‘cocktail’ of up to three restriction enzymes. The application of a set of enzymes might be disadvantageous for the analysis of GC-rich regions as such a strategy would produce restriction fragments too short for an efficient hybridization. In the latter case, it is advisable to use a smaller number of restriction enzymes. Based on our experimental results and computer-based analysis of 100 randomly selected CpG islands, the most suitable restriction enzymes are $Hin6$I and $Hpa$II, followed by $Aci$I, and $Hin1$I (Tab. 1). In contrast, for regular DNA sequences, double- or triple-digest combinations of $Aci$I, $Hpa$II, $HpyCH4IV$ and $Hin6$I are recommended.
Figure 1.1. Schematic outline of the microarray-based method for identification of DNA methylation differences and DNA polymorphisms in genomic DNA

*Left panel:* Analysis of DNA sequence variation. Middle panel: The main strategy of the method is based on enrichment of unmethylated DNA fragments. DNA samples are cleaved by methylation-sensitive restriction endonucleases, and the resulting DNA fragments are then selectively enriched by adaptor-specific aminoallyl-PCR’s, labeled, and hybridized to microarrays. *Right panel:* Alternative procedure to enrich the hypermethylated DNA fraction.
<table>
<thead>
<tr>
<th>Enzymes</th>
<th>% coverage of CpGs in λ DNA</th>
<th>% coverage of CpGs in human gDNA</th>
<th># of fragments (per kb) in CpG islands*</th>
<th># of fragments (per kb) in non-CpG islands*</th>
</tr>
</thead>
<tbody>
<tr>
<td>HpaII (BsiSI)</td>
<td>10.5%</td>
<td>8.6%</td>
<td>3.98</td>
<td>1.18</td>
</tr>
<tr>
<td>Hin6I (HinP1I)</td>
<td>6.9%</td>
<td>6.4%</td>
<td>3.98</td>
<td>0.61</td>
</tr>
<tr>
<td>AciI (SsiI)</td>
<td>16.6%</td>
<td>17.4%</td>
<td>3.23</td>
<td>1.79</td>
</tr>
<tr>
<td>Hin1I (Acyl, BsaH)</td>
<td>0.1%</td>
<td>2.0%</td>
<td>1.92</td>
<td>0.11</td>
</tr>
<tr>
<td>HpyCH4IV</td>
<td>4.6%</td>
<td>6.6%</td>
<td>1.31</td>
<td>1.08</td>
</tr>
<tr>
<td>Bsu15I (CiaI, BspDI)</td>
<td>0.5%</td>
<td>0.2%</td>
<td>&lt;0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>NarI (MlyI)</td>
<td>&lt; 0.1%</td>
<td>0.6%</td>
<td>1.08</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Bsp119I (AsulI, CbiI)</td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.11</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>BstBI (FspII)</td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.11</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Psp1406I (AclI, Pspl)</td>
<td>0.2%</td>
<td>0.3%</td>
<td>&lt;0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>XmnI (Accl)</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.19</td>
<td>0.34</td>
</tr>
<tr>
<td>Tasi</td>
<td>na</td>
<td>0.80</td>
<td>2.88</td>
<td></td>
</tr>
<tr>
<td>Csp6I</td>
<td></td>
<td>2.23</td>
<td>1.41</td>
<td></td>
</tr>
<tr>
<td>Msel</td>
<td></td>
<td>0.80</td>
<td>2.88</td>
<td></td>
</tr>
<tr>
<td>Bfal</td>
<td></td>
<td>1.56</td>
<td>1.55</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.1. Enzymes that generate protruding ends in the restriction fragments, which are complementary to the adaptors U-CG1, TA-1 and AATT-1**

Asterisk (*) indicates the number of 50 bp – 1.5 kb long (‘informative’) fragments, derived from several Mbp of randomly selected CpG island and non-CpG island sequences on chromosomes 1, 2, 4, 5, 6, 9, 17, 19 and 20; bold numbers represent the most informative enzymes; na = not applicable.

After the digestion of genomic DNA, the double-stranded adaptor U-CG1 is ligated to the CpG-
overhangs. At this point, it is expected that most of the relatively short (<1.5kb) and amplifiable DNA fragments derive from the unmethylated DNA regions. To some extent, the length of the amplified fragments depends on the primer annealing temperature of the PCR reaction (Fig. 1.2A). Some ligation fragments, however, may still contain methylated cytosines. A proportion of such fragments can be eliminated by treatment with \( \textit{McrBC} \), which cleaves DNA containing \( \text{met}C \) and will not act upon unmethylated DNA. \( \textit{McrBC} \) restriction sites consist of two half-sites of the form \( \text{(G/A)}^{\text{met}}C \), which can be separated by up to 3 kb [170, 171]. Hence, as can be seen in Fig. 1.2B, a proportion of DNA fragments with two or more \( \text{(G/A)}^{\text{met}}C \) within the restriction fragment are cleaved and therefore deleted from the subsequent enrichment steps. The remaining pool of unmethylated DNA fragments is then enriched by aminoallyl-PCR amplification that uses primers complementary to the adaptor U-CG1. One important advantage of using protruding ends in the adaptor ligation step is that degraded genomic DNA fragments (which are common in human post-mortem tissues) will not be ligated and amplified, and therefore will not interfere with DNA methylation analysis.

Figure 1.2. Selective enrichment of restriction fragments with the universal adaptor U-CG1
A: A DNA smear between 50 bp - 2.0 kb in a standard adaptor-PCR indicates an efficient ligation and amplification. The size of the amplification products varies with the annealing temperature used for PCR. B: Scatter-plot that shows a comparison of ligation products treated with McrBC vs. the untreated sample on the COMT array. McrBC treated fragments that contained at least two methylated cytosines were cleaved and could not be amplified in the following adaptor-PCR, resulting in reduced signal intensities in the Cy5 channel. C: Co-hybridization of enriched unmethylated and hypermethylated fragments derived from the same DNA source to a CpG island microarray. A large portion of amplicons is present only in one of the enriched fractions. Although the hypermethylated fraction hybridised to ~75% of the microarray spots, based on our DNA sequence analysis, only a small fraction of them provide epigenetic information in comparison to the unmethylated fraction.

Most previous microarray-based epigenetic studies target hypermethylated DNA sequences [161, 163, 172, 173]; however, interrogation of the unmethylated fraction is significantly more informative. For example, the 100 kb region of chromosome 22 interrogated by our COMT oligonucleotide array (TXNRD2-COMT-ARVCF region, see the Microarray Design section below), contains 2,193 methylatable cytosines. Enrichment of the unmethylated fraction can generate up to 401 amplicons of sufficient size (50 bp - 1.5 kb), each representing the methylation status of at least one cytosine. In contrast, the combination of MseI (+BsuI, to remove unmethylated fragments), the most frequently used enzymes for enrichment of the hypermethylated fraction [161, 163, 172, 173], would produce 227 amplicons. Seventy-seven amplicons would either contain no CpG dinucleotides or would be too short to stringently hybridize to a microarray. Of the remaining 150 fragments, 144 contain multiple CpGs; hence, they are not fully informative since a single unmethylated BsuI restriction site would eliminate the entire fragment from the eventual amplification. Overall, only six of the 2,193 methylatable cytosines are truly informative, and none of these CpG dinucleotides are targeted by BsuI. Computer-based analysis of 50 randomly selected CpG island sequences revealed that the unmethylated fraction derived from HpaII cleavage results in ~22 times more fragments (19.9 fragments/kb) of the suitable size range (50 bp - 1.5 kb) than the hypermethylated fraction (0.9 fragments/kb) using MseI.

Nevertheless, analysis of the hypermethylated DNA fraction may also add some new information to the methylation profiles, especially in the case of hypermethylated CpG islands or when the overall level of methylation in the genome is low (e.g. in insects). Thus, we developed an additional, modified method to previously published methods of enrichment of methylated sequences to complement our data from the unmethylated fraction (Fig.1, right panel). This enrichment method relies on cleavage with the 4-basepair frequent cutters TasI (AATT↓) and/or
Csp6I (G↓TAC). Alternatively, BfaI or MseI can be used in combination with the Csp6I-specific adaptor. All four enzymes produce DNA fragments in mammalian genomes of an average length 400 bp - 750 bp. The recognition sequences of TasI and Csp6I are infrequent within GC-rich regions, leaving most CpG-islands intact. The analysis of 50 randomly selected CpG islands and several megabases of different chromosomes revealed that Csp6I would produce more informative fragments in CpG islands than a digest with MseI, whereas TasI and MseI produce informative fragments preferentially in DNA regions outside of CpG islands (Tab.1). After ligation to the AATT- and TA-overhang specific adaptors “AATT-1” and “TA-1”, the un- and hypo-methylated ligation products are eliminated from the reaction by cleavage with a cocktail of methylation-sensitive restriction enzymes such as HpaII, HhaI (Hin6I), HpyCH4IV, Hin1I and AciI. Compared to a single digestion with BstUI [163], a cocktail of restriction enzymes will delete a higher percentage of unmethylated sequences from the DNA fraction. The remaining pool of mostly hypermethylated DNA fragments is subsequently enriched by the aminoallyl-PCR amplification as described for the unmethylated fraction, and then hybridized to a microarray (Fig. 1.2C).

Microarray design

Various aspects of the microarray-based DNA modification profiling were investigated on the oligonucleotide-microarray that interrogates ~100 kb fragment on 22q11.2 (Fig. 1.3A). In addition to the catechol-O-methyltransferase (COMT, [MIM 116790]), this chromosomal region contains also the gene encoding the thioredoxin reductase 3 gene (TXNRD2, [MIM 606448]) and the armadillo repeat gene deleted in velocardiofacial syndrome (ARVCF, [MIM 602269]). For maximal informativeness, it is necessary to design oligonucleotides according to the restriction sites of the methylation sensitive endonucleases used for the treatment of genomic DNA (Fig. 1.3B). For the COMT array, 384 oligonucleotides were designed, each 50 nucleotides long, representing every restriction fragment flanked by HpaII, Hin6I, and AciI restriction sites. In addition, control DNA fragments containing λ phage, pBR322, ΦX174, pUC57, and Arabidopsis sequences were spotted on the array (see Methods). Additionally, we used 12,192 element containing CpG island- and high-density chromosome 21/22- microarrays (see Methods).
Figure 1.3. Comt Microarray Design

A: Structure and GC-content of the chromosomal region on human chromosome 22q11.2 that spans the catechol-o-methyltransferase gene (COMT), the thioredoxin reductase 2 gene (TXNRD2), and the armadillo repeat gene deleted in VCFS (ARVCF). B: To determine the methylation profile of the 100kb TXNRD2-COMT-ARVCF region, 384 oligonucleotides (50-mers, black horizontal bars) were designed based on the restriction sites for the methylation-sensitive endonucleases, HpaII, Hin6I and AciI (additional alternative enzymes are HpyCH4IV or Hin1I). Depending on the methylation status of the CpG-dinucleotides several combinations of amplicons (grey horizontal bars) can potentially hybridize to the oligonucleotides. C: Typical hybridization patterns on the oligonucleotide-microarray. As discussed in the Results section, the complexity and informativeness of the hybridization signals increases with increasing number of methylation-sensitive restriction enzymes.

Detection of confounding effects of DNA sequence variation

Since restriction enzymes are used in the enrichment of differentially modified DNA fractions, DNA sequence variation may simulate epigenetic differences. However, until now, microarray methods used in epigenetic studies have not been differentiating between real DNA methylation differences and single nucleotide polymorphisms (SNPs) within the restriction sites of the applied restriction enzymes. This problem applies to some extent also to the $\text{mC}$ antibody-based strategy[168], which does not differentiate unmethylated CpG and TpG dinucleotides. In order to exclude the impact of DNA sequence variation, two approaches are suggested. One is to check the available SNP databases in order to identify the DNA sequence variation within the
restriction sites of the enzymes used. For example, our 100 kb COMT array contains a total of 273 SNPs (SNPper, http://snpper.chip.org/bio/snpper-enter), of which 101 (37%) reside within CpG dinucleotides and 55 (20%) are located within the restriction site of the four main enzymes used to interrogate methylation patterns, HpaII, Hin6I, AciI, and HpyCH4IV. The majority of these CpG-SNPs were located in AciI and HpaII restriction sites, with Hin6I and HpyCh4IV sites containing fewer polymorphisms (data not shown). Using database specific information, the probability of a SNP influencing a given microarray result can be taken into account and results interpreted accordingly.

The second approach to differentiate the DNA sequence effects from the genuine epigenetic differences consists of performing an identical microarray experiment on the same DNA sample that has been stripped of all methylated cytosines. Our protocol utilizes the Phi29 DNA polymerase to amplify whole genomic DNA, which creates a copy of the genome with all methylated cytosines replaced by unmethylated cytosines. Amplified DNA samples are then subjected to the same steps as depicted in Fig. 1.1 and hybridized on the microarrays. In this experiment all of the outliers must be a result of DNA sequence variations within the restriction sites of the enzymes used. These data can then be plotted against the DNA methylation data, which are assayed in parallel (Fig. 1.4). In six experiments that used amplified genomic DNA, the number of SNP-based outliers (threshold log-ratio <-0.3, >0.3) ranged from 272 to 741 (432 ± 165, mean ± SD), or 2.2% - 6.1% of 12,192 CpG islands. Out of these, 72 to 234 (120 ± 66, mean ± SD) were initially identified as DNA methylation differences in microarray experiments using the unmethylated fraction derived from the triple-digest with HpaII, AciI and Hin6I. From the CpG island array studies, our estimate is that 10% to 30% of the outliers detected in DNA methylation experiment were due to DNA sequence variation.
Figure 1.4. Combined methylation- and SNP-analysis on a CpG island microarray

The data of two separate hybridizations of brain DNA samples derived from two individuals are plotted against each other. The y-axis contains the data derived from a methylation analysis (triple-cleavage with \textit{HpaII}, \textit{Hin6I}, and \textit{AciI}), whereas the x-axis contains the SNP-data derived from the hybridization of the same DNA samples, which were subjected to the entire genome amplification prior to cleavage by the methylation-sensitive restriction enzymes (see Methods). Significant outliers (log-ratio < -0.3, > 0.3, 2–fold difference) can be classified into four clusters (S = SNPs, M = DNA methylation differences), enabling the differentiation of epigenetic differences and nucleotide polymorphisms between the test-samples. Amp = Whole-genome amplified sample.

Reproducibility

To test the reproducibility of the method, a genomic DNA sample was split and subjected to the procedure of enrichment of the unmethylated fraction. The resulting amplification products were labeled with Cy5 and Cy3 and then co-hybridized on the \textit{COMT} array, which contains probes that flank the \textit{HpaII}, \textit{Hin6I} and \textit{AciI} restriction fragment around the \textit{COMT} gene. The Cy3 and Cy5 hybridization intensities exhibited very similar values ($R^2=0.997$; Fig. 1.5A). Analogous experiments, including switch dye hybridizations, were repeated several times also with the CpG island arrays and in all cases were highly reproducible ($R^2>0.97$).
Figure 1.5. Reproducibility and sensitivity of the method

A: A COMT microarray scatter plot representing two sets of amplification products derived from the same DNA source but produced at different time-points by different researchers. The high correlation coefficient of signal intensities demonstrates a high reproducibility of the method. B: Influence of the PCR cycle number. Scatter plot diagrams show hybridization signal intensities of the unmethylated fraction that was amplified using 20 PCR cycles (Cy3 channel) and 30 cycles (Cy5 channel). Amplification products of each PCR were co-hybridized to the COMT microarray that contained oligonucleotides representing single copy sequences (black circles), partially repetitive sequences (grey squares; 15-99 copiesgenome) and highly repetitive DNA fragments (white squares; >100 copies/genome), such as ALU and LINE repeats. C: Scatter plot representing the unmethylated fraction of human genomic DNA ‘spiked’ with different amounts of control DNA. The test samples were hybridized to the COMT array and contained either a 16-fold excess of λDNA (16 genome equivalents [GE] vs. 1 GE; 10 fragments) or a 16-fold excess of pBR322 (128 GE vs. 8 GE; 2 fragments), respectively. The amplicons of the spiked DNA (representing unmethylated DNA) can be easily distinguished as outliers; whereas the signals representing genomic DNA are located close to the regression line. Median signal intensities of different length oligonucleotides (40-50 bases) that target a specific HpaII restriction fragment in λ DNA reveal that the length of spotted sequences directly influences the spot intensity and therefore the sensitivity of the microarray. D: Sensitivity of the CpG-island microarray hybridization. 2 • g of control amplicon was labeled with Cy5 and co-hybridized with 2 • g (0% difference), 1.9 ug (5% difference), 1.8 ug (10% difference), 1.5 ug (25% difference) or 1.0 ug (50% difference) of Cy3-labeled amplicon. For each hybridization to a COMT array, the regression lines represent the overall intensity that mimics methylation differences over the entire sample. The decrease of amount of DNA is reflected in the angle of the regression lines, which deviated by 5%-7% from the expected values.

Another critical factor in the amplification of unmethylated or hypermethylated DNA fragments is to ensure that no sequence specific biases are introduced. The rate of amplification of repetitive sequences generally declines faster than that of less abundant fragments in the later cycles of PCR [174]. With increasing amplification cycles, repetitive DNA strands reach relatively high concentration and begin re-annealing to each other during the steps below the DNA melting temperature. To avoid this, a two-temperature PCR that uses a combined high-temperature elongation-annealing step was applied. A series of experiments were performed investigating how the number of PCR cycles would affect the hybridization patterns. As can be seen in figure 5B, the relative intensities of the hybridization signals of both single copy sequences and repetitive DNA fragments, were similar in the range of 20 to 30 amplification cycles (R²=0.991). Only when increasing the cycle numbers beyond 40 cycles was a biased amplification of some DNA sequences observed (data not shown).
Sensitivity

To test if differentially represented DNA fragments in two different DNA samples can be detected by this method, human genomic DNA was ‘spiked’ with unmethylated heterologous DNA, \(\lambda\) phage and pBR322 plasmid (Fig. 1.5C). The amount of \(\lambda\) and pBR322 corresponded to the increasing numbers of human genomic equivalents (1 GE of ‘spike’ DNA equals 16.28 pg \(\lambda/\mu\)g gDNA and 1.45 pg/\(\mu\)g gDNA of pBR322, respectively). Hence, each of the experiments compared the intensities generated by 1 GE of \(\lambda\) plus 128 GE of pBR322 (Y axis) versus 16 GE of \(\lambda\) plus 8 GE of pBR322 (X axis). While the plotted signal intensities of the human genomic DNA sequences are positioned on or close to the regression line, the \(\lambda\) and pBR322 fragments were identified as outliers. The average signal intensity ratio of \(\lambda\) oligonucleotides was 15.4, which is very close to the ratio of spiked DNA (16:1). The intensity values for pBR322 were not as linear and exhibited a 6.5 – 10 fold difference (expected the same ratio of 1:16), most likely due to saturation effects of the hybridization.

In order to determine the sensitivity of the hybridization per se, a control amplicon DNA was compared to itself but by decreasing the amounts of DNA by 5%, 10%, 25%, and 50%. On the global level, the regression lines \([y=f(x)]\) reflected reproducible differences of the amount of amplicon DNA used in the hybridization and varied by 5%-7% from the expected values (Fig. 1.5D). Individual sites exhibited a lower accuracy, which depended on the signal intensity, i.e. the stronger the signal, the closer the observed spot intensity was to the expected one. The rate of false outliers (log-ratio < -0.3; > 0.3; 2-fold difference) was on average 3%. Usually, replication of microarray experiments reduced the degree of aberration (log-ratio < -0.3; > 0.3) below 2% for all types of microarrays.

Examples of DNA methylation profiles

Identification of DNA modification differences is provided in a series of examples below. The COMT oligonucleotide array was used to identify DNA methylation changes in a brain tumour (Fig. 1.6A). In contrast to the pair of control brain DNA samples, where hybridization signals are close to the regression line (indicating similar DNA methylation patterns), a visible proportion of the hybridization signals originating from the unmethylated DNA fraction of the brain tumour deviates from the regression line. More subtle changes in DNA methylation patterns have been identified when post-mortem brain tissues of healthy individuals were compared with the same
tissues from schizophrenia patients (Schumacher, Petronis et al; work in progress; representative example is shown in Fig. 1.6B). The differences of the cancer and psychosis studies show that diseases other than cancer may reveal more subtle epigenetic differences, and therefore, the informativeness and sensitivity of the epigenetic profiling method is of critical importance.

Figure 1.6. Applications of the epigenetic profiling technology

A: Changes of methylation profiles at TXNRD2-COMT-ARVCF in a brain tumor. The data from two different microarrays experiments are superimposed over each other. The analysis of two post-mortem brain samples (black dots) reveals no major difference in methylation levels, whereas the signal intensities vary significantly in the brain tumor (orange dots) when compared to the normal brain. B: Comparison of DNA methylation profiles using a CpG island microarray in brain tissue of a healthy control and a schizophrenia patient.

Another application of the technology includes epigenetic profiling of different tissues. One example of tissue specific effects is shown using the CpG island microarrays that contain 12,192 CpG island clones of whom 8,025 represent unique sequences. CpG islands tend to be found in many promoter sequences and their methylation has profound effects on gene silencing in mammalian genomes. The scatter plot shows two distinct spot areas, which represent predominantly unmethylated fragments in placenta (yellow spots) and brain (orange spots) respectively (Fig. 1.7A). About 11% of the CpG island-fragments exhibited 2-fold or more signal intensity difference between the two tissues. Some of the strongest brain-specific signals could be identified for CpG islands associated with neuronal genes such as DPYSL5, FABP7, DIRAS2, GRIN3A, SLC24A3, and DSCAML1, whereas strong placenta-specific outliers were associated with genes expressed in placenta, such as PCM1, CCND1, HA-1, and ADAMTS11. Overall, analysis revealed that brain DNA harboured notably more unmethylated CpG islands than placenta DNA.
Figure 1.7. Examples of applications using a CpG island microarray

A: Hybridization of the unmethylated fraction of placenta DNA and post-mortem brain DNA to a CpG island array. Two pools of CpG island elements could be identified, which display extensively different methylation levels between these tissues (note: some of the identified differences could be due to DNA sequence variation). B: To validate the identified methylation differences, several CpG islands were subjected to bisulfite modification based mapping of methylated cytosines as exemplified for CpG island clones 22_B_12 (promoter region of Galectin-1) and 52_C_03 (promoter region of a brain-specific transcript, CR606704). The top sequence shows the reverse strand (-) of the original restriction sites, the bottom sequence displays the bisulfite-modified DNA. For each bisulfite-modified CpG-island, 8 to 10 clones were sequenced per tissue. Sequence 52_C_03 revealed several fully methylated CpG’s in placenta, which were unmethylated in brain. In contrast, clone 22_B_12 showed subtler methylation differences (15%-100%), depending on the position of CpG-dinucleotide. C: Methylation patterns of clones 22B_12 and 52_C_03 derived from bisulfite sequencing of 10-12 clones per tissue. The yellow boxes indicate CpG dinucleotides that are shown in the sequenced graph (Fig. 1.7B).

Verification of detected methylation differences

Several loci that displayed methylation differences in our experiments were selected for verification by the sodium bisulfite modification mapping of methylated cytosines (see Material and Methods). The technique is based on the reaction of genomic DNA with sodium bisulfite under conditions such that cytosine is deaminated to uracil but 5-methylcytosine remains unaltered. In the sequencing of amplified products, all uracil and thymine residues are detected
as thymine and only $\text{methC}$ residues remain as cytosine. The sites for the methylation-sensitive restriction enzymes used in our experiments showed the expected methylation difference across the DNA samples, as exemplified for CpG island clones located in the promoter region of galectin-1 and in the promoter region of a brain-specific transcript CR606704 (Fig. 1.7B and 7C).

**Chromosome-wide mapping of DNA methylation differences**

Analysis of the unmethylated fraction from brain specific DNA of 8 adults using a chromosome 21/22 tiling array detected 488 to 747 unmethylated sites per sample (Table 1.2). This number increased to 977 in a merged map, showing that many sites were common between different individuals. The vast majority of the sites (~90%) were positioned outside of the 5' ends and 5' flanking regions of the genes consistent with abundant transcriptional activity and a significant fraction of transcription factor binding sites found outside of known annotations [175-177]. The unmethylated sites outside of the 5' ends of known genes were about equally distributed between sites residing within introns of known genes and outside of the gene boundaries. Interestingly, while some genes, like *BCR*, showed a large number of sites inside the gene boundaries, some loci, like *C21orf55* spanning ~150 kb, were essentially devoid of internal unmethylated sites and in some cases, such as the *SIM2* locus, the unmethylated sites were limited to the first intron (Fig. 1.8A-C). Such intragenic methylation may inhibit inappropriate transcriptional initiation at cryptic sites [178] or may serve as regulators of alternate transcripts as can be seen for *SIM2*. In the case of the *BCR* gene, which contains several alternative transcripts (Fig. 1.8A), CpG methylation along the whole gene may influence the transcriptional control of several transcripts within this genomic region. Overall, unmethylated sites detected in this study cover ~0.47 Mbp or ~4% of the 12 Mbp of non-repetitive sequences of chromosomes 21 and 22 interrogated in the combined map of all 8 individuals with an average of 0.28 Mbp or 2.3% in any given individual. Maps of the methylation patterns (average value of the eight tested individuals) of the q-arms of chromosome 21 and 22 are shown in Fig. 1.9A-B. Detailed maps of all individuals for chromosome 21 and 22, linked to the UCSC Genome Browser (http://genome.ucsc.edu) are also available on our web-based methylation database (see Web Resources). A comparison of the hypomethylation tracks with data from the Affymetrix transcriptome project [177, 179] indicate that many of the unmethylated chromosomal regions overlap with mapped transcriptional active regions (Fig. 1.9A-C, bottom tracks). These DNA methylation data complement existing studies
on transcriptional activity and histone modifications on human chromosomes 21 and 22 [180]. We found that in the majority of cases, specific histone modification patterns reported by Bernstein et al. for the human hepatoma cell line HepG2 overlapped notably with the observed DNA methylation patterns. An example is shown in Fig. 1.9C for the PEX26 gene that is ubiquitously transcribed in most tissues. The gene harbors an extensively unmethylated CpG rich region in its’ promoter. The comparison of the different epigenetic profiles of both studies shows that the same genomic region was also highly acetylated at Lysine 9 and 14 of histone 3 (H3), accompanied with H3 di- and trimethylation of Lysine 4. A comparison of histone modification tracks and our hypomethylation patterns for the q-arms of chromosome 21 and 22 revealed that H3 acetylation and Lys4 methylation usually correlated with unmethylated CpGs.

<table>
<thead>
<tr>
<th>Individual</th>
<th>3'-flanking</th>
<th>3'flanking-3'flanking</th>
<th>5'-flanking</th>
<th>5'flanking-3'flanking</th>
<th>5'ter</th>
<th>Distal</th>
<th>Internal</th>
<th>Total</th>
<th>Site coverage, bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>#M17 chr21/22</td>
<td>13/12</td>
<td>2/16</td>
<td>8/20</td>
<td>2/4</td>
<td>10/20</td>
<td>64/122</td>
<td>98/97</td>
<td>488</td>
<td>64943/134730</td>
</tr>
<tr>
<td>%Total</td>
<td>5.1</td>
<td>3.7</td>
<td>5.7</td>
<td>1.2</td>
<td>6.1</td>
<td>38.1</td>
<td>40.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#M18 chr21/22</td>
<td>17/22</td>
<td>9/15</td>
<td>13/29</td>
<td>3/3</td>
<td>16/28</td>
<td>95/191</td>
<td>134/152</td>
<td>727</td>
<td>98456/236797</td>
</tr>
<tr>
<td>%Total</td>
<td>5.4</td>
<td>3.3</td>
<td>5.8</td>
<td>0.8</td>
<td>6.1</td>
<td>39.3</td>
<td>39.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#M19 chr21/22</td>
<td>15/24</td>
<td>11/14</td>
<td>12/27</td>
<td>2/5</td>
<td>14/21</td>
<td>86/173</td>
<td>119/130</td>
<td>653</td>
<td>88290/221721</td>
</tr>
<tr>
<td>%Total</td>
<td>6.0</td>
<td>3.8</td>
<td>6.0</td>
<td>1.1</td>
<td>5.4</td>
<td>39.7</td>
<td>38.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#M21 chr21/22</td>
<td>20/24</td>
<td>12/18</td>
<td>15/29</td>
<td>2/5</td>
<td>14/22</td>
<td>102/184</td>
<td>143/157</td>
<td>747</td>
<td>109595/252347</td>
</tr>
<tr>
<td>%Total</td>
<td>5.9</td>
<td>4.0</td>
<td>5.9</td>
<td>0.9</td>
<td>4.8</td>
<td>38.3</td>
<td>40.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#M22 chr21/22</td>
<td>18/20</td>
<td>8/17</td>
<td>9/29</td>
<td>3/6</td>
<td>15/24</td>
<td>86/169</td>
<td>127/143</td>
<td>674</td>
<td>87604/213453</td>
</tr>
<tr>
<td>%Total</td>
<td>5.6</td>
<td>3.7</td>
<td>5.6</td>
<td>1.3</td>
<td>5.8</td>
<td>37.8</td>
<td>40.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr21/22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>#M23 chr21/22</td>
<td>12/15</td>
<td>4/13</td>
<td>10/25</td>
<td>2/3</td>
<td>10/21</td>
<td>68/150</td>
<td>101/111</td>
<td>545</td>
<td>70912/163322</td>
</tr>
<tr>
<td>%Total</td>
<td>5.0</td>
<td>3.1</td>
<td>6.4</td>
<td>0.9</td>
<td>5.7</td>
<td>40.0</td>
<td>38.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#M24 chr21/22</td>
<td>14/18</td>
<td>5/12</td>
<td>7/20</td>
<td>4/3</td>
<td>10/20</td>
<td>61/158</td>
<td>88/107</td>
<td>527</td>
<td>655639/187229</td>
</tr>
<tr>
<td>%Total</td>
<td>6.1</td>
<td>3.2</td>
<td>5.1</td>
<td>1.3</td>
<td>5.7</td>
<td>41.6</td>
<td>37.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#M25 chr21/22</td>
<td>17/15</td>
<td>7/13</td>
<td>10/18</td>
<td>3/3</td>
<td>9/22</td>
<td>65/171</td>
<td>102/97</td>
<td>552</td>
<td>69937/171073</td>
</tr>
<tr>
<td>%Total</td>
<td>5.8</td>
<td>3.6</td>
<td>5.1</td>
<td>1.1</td>
<td>5.6</td>
<td>42.8</td>
<td>36.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merged chr21/22</td>
<td>26/28</td>
<td>13/22</td>
<td>19/36</td>
<td>4/9</td>
<td>19/34</td>
<td>142/237</td>
<td>187/201</td>
<td>977</td>
<td>152148/314374</td>
</tr>
<tr>
<td>%Total</td>
<td>5.5</td>
<td>3.6</td>
<td>5.6</td>
<td>1.3</td>
<td>5.4</td>
<td>38.8</td>
<td>39.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.2. Distribution of the detected unmethylated sites in respect to the known genes as defined by the combined set of RefSeq and UCSC Known Genes for each brain DNA sample (M17-M25) and the merged map.

“5’ter” or “3’ter” refers to a 5’ or 3’ terminal site internal and within 1kb of a gene boundary “5’flanking” or “3’flanking” refers to a site outside and within 5kb of a gene boundary; “internal” refers to an intronic site and “distal” refers to an intergenic site outside of the -5kb/+1kb boundaries. A site can also be both 5’ and 3’ flanking in a gene rich regions and referred as “5’flanking-3’flanking”.
Figure 1.8. Profiles of unmethylated sites in three loci on human chromosomes 21 & 22 (501 bp window, see Methods):

BCR (A), C21orf55 (B), and SIM2 (C) for human brain DNA (average of 8 individuals). The graphs are based on p values for each individual interrogation that show the significance of the enrichment in the unmethylated fraction vs total genomic DNA. The p values were converted to the \(-10\log_{10}\) scale, such that, for example, p value of \(10^{-4}\) becomes 40. The vertical axes are adjusted to represent probes in the 40 to 120 range (p values of \(10^{-4}\) to \(10^{-12}\)), thus only probes that pass p < \(10^{-4}\) threshold are shown. Enlarged is a part of the chr 22q11.21 region (181 bp window), spanning the BCR gene that contains the site of breakpoints found in the generation of the two alternative forms of the Philadelphia chromosome translocation observed in chronic myeloid leukemia and acute lymphocytic leukemia. C = genomic DNA control.
Figure 1.9. Genomic views showing unmethylated regions on chromosomes 21 and 22

A,B: The top tracks (dark red) in the two chromosomal graphs shows the average amount of hypomethylation in the brain cortex of eight adult individuals. Also displayed are known genes (dark blue) and CpG islands (green). The
bottom tracks display transcriptome data derived from 11 different tissues from the *Affymetrix* transcriptome phase 2 study [179]. The track is coloured blue in areas that are thought to be transcribed at a statistically significant level. Regions that have a significant homology to other chromosomal regions or that overlap putative pseudogenes are colored in lighter shades of blue. All other regions of the track are colored brown. C: Enlarged is a part of chromosome 22q11.21, containing the peroxisome biogenesis factor 26 (*PEX26*, MIM 608666) that shows correlation between histone modifications and unmethylated DNA in its promoter region. Abnormal transcription of *PEX26* is associated with the two peroxisome biogenesis disorders neonatal adrenoleukodystrophy (NALD; OMIM #202370), Zellweger syndrome (ZS; OMIM #214100) and also the infantile Refsum disease (IRD; OMIM #266500). The top three tracks represent histone modification data for H3 Lys4 dimethylation (orange bar), H3 Lys4 trimethylation (blue bar) and H3 Lys9/14 acetylation (yellow bar) [180]. Underneath are the tracks for the average methylation patterns (unmethylated sites) observed in brain and the individual methylation patterns of all tested individuals (dark red). It is noteworthy that methylation patterns exhibit some interindividual differences (indicated by arrows).

**Discussion**

The microarray based technology for DNA modification analysis enables a highly parallel screening of numerous restriction fragments that represent DNA methylation profiles over large segments of genomic DNA. Building on the principles described in earlier publications [11-23] our method addresses a series of critical issues and exhibits several advantages. An earlier method [164] used a fractionation in a sucrose gradient to enrich the unmethylated DNA fraction, which requires a large amount of DNA template and is rather imprecise in terms of the upper limit of the fragments that are subjected to hybridization. Other microarray methods for DNA methylation analysis can be categorized into three main classes which are based on: I) identification of bisulfite induced C$\rightarrow$ T transitions [157-159, 181, 182], II) cleavage of genomic DNA by methylation-sensitive restriction enzymes and III) immunocapturing with antibodies against methylated cytosine. In the bisulfite arrays, each tested CpG is represented by a pair of either C(G) or T(A), containing oligonucleotides that measure the C(G)/T(A) ratio in the bisulfite treated DNA (corresponding to $^{\text{mel}}$C/C in the native DNA). Although informative and precise, the microarray can contain only a limited number of oligonucleotides because treatment with bisulfite degenerates the four-nucleotide code, resulting in the loss of specificity of a large portion of the genome. For example, after bisulfite treatment all of the possible 16 permutations of a four base sequence containing unmethylated C and T (CCCC, CTCT, CCCT, CCTT, TCTC, TTTC, TTTT, etc…) will become identical TTTT. The bisulfite method is also laborious and cannot be easily applied to profile a large set of samples. Furthermore, it is difficult to design suitable oligonucleotides that would exhibit similar melting temperatures since the specificity of
base discrimination varies considerably [158]. In our approach, the arrays can contain an almost unlimited number of oligonucleotides: from individual genes to entire chromosomes represented by millions of oligonucleotides on glass chips. Whole genome tiling arrays are already available for Arabidopsis thaliana and E.coli, and will soon be available for the entire human genome.

Restriction enzyme based methods are used for enrichment of the hypermethylation fraction of genomic DNA [160, 161, 163, 172] or enrichment of the unmethylated fraction, either with McrBC [169], or alternatively with the rare cutter NotI. The majority of these methods rely on the enrichment and detection of hypermethylated DNA, and have thus far been predominantly used for the identification of abnormally methylated CpG islands in malignant cells [161-163, 172]. Although this strategy seems to be useful for detecting major epigenetic changes in some regions of the genome, the overall proportion of the interrogated CpG sites is substantially lower in comparison to the approach based on the analysis of the unmethylated fraction. As shown in the Results section, interrogation of the unmethylated fraction of genomic DNA may be up to several hundred-fold more efficient in comparison to the hypermethylated fraction scenario. Furthermore, since unmethylated cytosines are much more abundant than methylated cytosine (depending on the tissue, 70%-90% of cytosines are methylated), analysis of this smaller unmethylated fraction of genomic DNA is more sensitive to detect subtle changes. For example, an increase of 10% from the normal density of $^\text{met}$C would result in a 100% (from 20% to 10%) difference in the unmethylated fraction, but only a 12% (from 80% to 90%) difference in the hypermethylated fraction of genomic DNA. The unmethylated fraction was used in some approaches of the class II microarray methods, for instance by using the methylation-specific McrBC enzyme [169]. In that protocol McrBC is used to deplete the hypermethylated fraction. However, the remaining unmethylated DNA fragments (> 1kb) have to be gel-purified, requiring huge amounts of starting material. Additionally, by using McrBC, it becomes impossible to differentiate between dense and sparse methylation within a relatively short DNA fragment. For example, the 2 kb human COMT promoter region, which contains 27 McrBC target sites, can be cut to shorter than 1 kb fragments, both in the cases when there are two (7%) or 27 (100%) methylated McrBC sites. Additionally, the enrichment of the unmethylated fraction by McrBC is less informative, since the enzyme cannot differentiate between unmethylated and polymorphic CpG sites (NpG). Whilst the enrichment of the hypomethylated fraction with McrBC can be very useful in studying plant methylation patterns, more precise methods are needed for the detailed analysis of human tissues, especially when only limited amounts of DNA can be acquired.
Another method of enriching the hypomethylated fraction uses the rare cutter NotI (5’-GCGGCCGC-3’)[165-167]. However, NotI-sites are not well represented in the genome and will only provide a very superficial overview of genomic methylation patterns. An alternative to these methods is the use of antibodies specific for methylated cytosines (MeDIP[168]). In this method, antibodies are used to immunocapture methylated genomic fragments. However, this approach requires large amounts of genomic DNA (> 8 µg) and also relies on the enrichment of the less informative hypermethylated fraction of the genome.

We addressed another important issue: the interference of DNA polymorphisms that may simulate DNA modification differences across individuals. Data from the SNP consortium indicate that roughly every 360th nucleotide in the human genome represents a SNP. In humans approximately 2.16 million SNPs are detectable in CpG dinucleotides, and such CpG SNPs are 6.7-fold more abundant than expected [183]. Depending on the restriction enzyme combination, our CpG island array-based studies demonstrated that 10%-30% of all outliers that were originally detected as methylation differences contained SNPs (Fig. 1.4). Information on the SNPs within the restriction sites of the enzymes used for the enrichment of the unmethylated or hypermethylated fractions is helpful in differentiating the epigenetic variations from the DNA sequence ones.

Another advantage of PCR-based methylation profiling methods is the ability to work with limited DNA resources. Although our basic protocol requires about 500 ng of genomic DNA, the amount of template DNA can be significantly lower. In our recent experiments, methylation patterns at the COMT region generated from a relatively small number of Jurkat tissue culture cells (up to 500 cells, or 3 ng) did not reveal any significant differences compared to the methylation patterns generated from a substantially larger number of cells from the same tissue.

There are also some limitations of the technology described above. The methylation sensitive restriction enzymes do not interrogate every cytosine, and with the current design, more than half of CpG sites remain uninterrogated. This may be critical when the phenotypic outcomes are determined by a methylation change at an isolated cytosine that is not within the restriction site of a methylation sensitive restriction enzyme. This problem may be partially overcome by the application of the same arrays to the CpG specific immuno-precipitation technique (MeDIP) [168] in addition to histone modification analysis through chromatin immunoprecipitation (ChIP) technology, which identifies DNA sequences associated with modified histones [156]. DNA and
histone modifications seem to be inter-dependent, and consequently the possibility of a combined approach that interrogates both DNA methylation and chromatin modification in parallel might be a productive approach to the fine mapping of epigenetic changes.

In summary, the microarray–based identification of unmethylated cytosines is a high throughput approach for profiling of DNA methylation patterns. The ability to analyze minute amounts of DNA may enable the epigenetic screening of DNA in plasma, serum or other body fluids as well as in prenatal diagnostics. Although all the examples provided in this work investigated human DNA, the same strategies can be used for epigenetic analyses of numerous other species. It is evident that epigenetic profiling should be performed in a systematic, unbiased fashion and not limited to the traditionally preferable regions such as CpG islands. Outside of CpG islands, numerous other genomic loci exist that may be sites for important epigenetic modification, including enhancers, imprinting control elements [184] or the regions that encode regulatory RNA elements.

Our technology, in combination with existing epigenetic profiling methods, can be of significant benefit to the identification of inter-individual variation, the identification of epigenetic changes during tissue differentiation and between species, and the understanding of the epigenetic effects of various environmental factors. Of particular interest is the application of high throughput DNA methylation analyses to addressing the molecular basis of various non-Mendelian irregularities of complex diseases, such as discordance of monozygotic twins, remissions and relapses of a disease, parent of origin- and sex- effects, tissue- and site- specificity [185]. Further technological developments may include building the high resolution oligonucleotide–based microarrays for the entire human genome, improving the enrichment strategies through application of more specialized methylation sensitive restriction enzymes, and substantial reduction of the initial template DNA up to the amount of a haploid or diploid genome. All these developments may provide the basis for identification of the methylation profile of the entire genome in a single cell, one of the “quantum leaps” in the post-genomic biology [186].

**Material and Methods**

**Microarray fabrication and data processing**

*COMT* and CpG island microarrays were printed on Corning CMT-GAPSII slides (Corning Life Sciences, Acton, Ma) using a VersArray ChipWriter Pro System (Bio-Rad Laboratories,
Hercules, CA). For the COMT array, we designed 384 oligonucleotides (Operon/Qiagen, US), each 50 bases long, representing every restriction fragment flanked by HpaII, Hin6I, and AciI restriction sites. In addition, control DNA fragments containing λ phage, pBR322, ΦX174, and pUC57 sequences were spotted on the slide. Each oligonucleotide was diluted to a 25 µM solution and spotted four times to give a total of 1,536 elements. In addition, 192 blank spots consisted of SSC buffer and 48 spots contained Arabidopsis clones. The human CpG island array contains 12,192 sequenced CpG island clones derived from a CpG island library that was originally created with MeCP2 DNA binding columns [72, 187].

Hybridized arrays were scanned on a GenePix 4000A scanner (Axon Instruments, Union City/CA, USA) and analyzed using the GenePix 6.0 software. The GenePix PMT voltage for Cy3 and Cy5 channels were balanced with the histogram feature of the scanner software to ensure a similar dynamic range for the two channels. Final scans were taken at 10µm resolution, and images for each channel were saved as separate 16-bit TIFF files. The emission signals for each channel were determined by subtracting the local background from its corresponding median average intensity. These raw data were either exported into a custom Excel spreadsheet for subsequent data analysis or directly imported into the Acuity 4.0 software (Axon Instruments). The resulting datasets were normalized for the normalization features (spike-DNAs) and for signal intensity (Lowess normalization).

Profiling of unmethylated sites in the brain tissue of 8 adults was carried out using a tiling array spanning ~12 Mb of non-repetitive sequence of the distal ~1/3 of chromosome 21 and ~1/3 of the proximal portion of chromosome 22 with probes spaced on average every 35 bp center-to-center [177]. The genomic DNA from these individuals was cut with HpaII and Hin6I, amplified and hybridized to the microarray as previously described [176, 177]. Total genomic DNA was used as a control. Unmethylated sites were defined using a two-step analysis approach similar to the one used to determine transcription factor binding sites in the ChIP-chip assay [176]. First, a smoothing-window Wilcoxon approach was applied to generate a p-value graph for each individual where probe signal from the enriched fraction was compared to the total genomic DNA in a one-sided upper paired test. The window used in this report was 501 bp. Second, three thresholds were applied to determine the boundaries of the unmethylated site: a) an individual probe threshold of p<10^{-4} to determine if a probe is significantly enriched in the unmethylated fraction compared to the control total genomic DNA; b) the maximum distance between the two
positive probes set to 250 bp and c) the minimal size of a site set to 1bp. The graphs can be downloaded from the internet (see Web Resources). All coordinates and annotation analysis was done on the April 2003 version of the genome.

**Methylation-sensitive digestion of genomic DNA (gDNA)**

Prior to treatment with restriction enzymes, gDNA was supplemented with “spike”-DNAs (different concentrations of λ and Arabidopsis fragments), which were used as controls for signal normalization. For enrichment of the unmethylated fraction, depending on the number of CpG dinucleotides to be interrogated, several combinations of methylation-sensitive enzymes, HpaII, Hin6I, AcI and HpyCH4IV, were used. Genomic DNA was cleaved with a cocktail of these enzymes (10U/µl in 2xY+/Tango buffer, Fermentas Life Sciences/Lithuania) for 8h at 37°C. For enrichment of the methylated fraction, genomic DNA was cleaved by TasI or Csp6I (10U/µl in G⁺-buffer, Fermentas) for 8h at 65°C (TasI) or at 37°C (Csp6I). After the restriction reaction, TasI was inhibited by 0.5 M EDTA.

**Adaptor-Ligation**

For the ligation step, genomic DNA was supplemented with 8 GE MspI-cleaved pBR322 plasmid (1 GE = 1.45 pg/ 1 µg gDNA), which was used as control for a potential ligation bias. The ends of the cleaved DNA fragments were ligated to the unphosphorylated adaptors. Our adaptors contained a sequence-specific protruding end, a non-target homologous core sequence, a specific antisense-overhang that prevents tandem repeat formation and blunt-end ligation, a ‘disruptor’ sequence that interrupts the original restriction sites after ligation, a non-palindromic Alw26I (BsmAI) restriction site that enables the blunt-end cleavage of the adaptor from the target sequences (e.g. for library enrichment) and a non-5’-complementary end. The CpG-overhang specific universal adaptor “U-CG1” for the unmethylated DNA fraction ligates to DNA fragments generated by 11 CpG-methylation-sensitive restriction enzymes HpaII, Hin6I (Hinp1I), HpyCH4IV, Bsu15I (ClaI, BspDI), AcI (SsiI), Psp1406I (AciI), Bsp119I (AsuII), HinII (AcyI, BsaHI), Xm1I (AccI), NarI, BstBI (FspII) and also TaqI and MspI, which are not affected by methylation of the internal cytosine. The adaptor represents the annealing product of the two primers:

U-CG1a: 5’-CGTGGAGACTGACTACCAGAT-3’
U-CG1b: 5’-AGTTACATCTGGTAGTCAGTCTCCA-3’

The AATT-overhang specific adaptor “AATT-1” for the methylated DNA fraction fits to DNA ends produced by the restriction enzyme TasI (TspEl), whereas the “TA-1” adaptor fits to ends produced by Csp6I, BfaI or MseI respectively:

AATT-1a: 5’-AATTGAGACTGACTACCAGAT-3’;
AATT-1b: 5’-AGTTACATCTGGTAGTCAGTCTC-3’;
TA-1a: 5’-TATGAGACTGACTACCAGAT-3’
TA-1b: 5’-AGTTACATCTGGTAGTCAGTCTCA-3’

All adapters were prepared by mixing equimolar amounts of the primer pairs, incubating the mixture at 80°C for 5 min, and then cooling it down to 4°C with 1°C/min. The double-stranded adaptors [200 pmol/µl] were added at 0.1 pmol per enzyme for each ng of the cleaved DNA (e.g. 0.3 pmol/ng in a triple-digest HpaII/Hin6I/AciI). The ligation-mixture with 400 ng template DNA was supplemented with 2 µl of 10x ligation buffer (Fermentas), 1 µl ATP [10mM], and water to 18 µl. The reaction was started in a thermal-cycler at 45°C for 10 min, chilled on ice and 2 µl T4 ligase (Fermentas) was added. The ligation reaction was carried out at 22°C for 18h, followed by a heat-inactivation step at 65°C for 5 min. The mixture was then cooled down to room temperature with 1°C/min and stored at 4°C for subsequent procedures.
PCR

To control for a potential PCR bias, the DNA mixture was supplemented with 2 GE ΦX174 plasmid (1 GE = 1.8 pg of ΦX174 corresponding to 1 µg gDNA) that was cut with HpyCH4IV and ligated to the adaptor. PCR amplifications were conducted for up to 25 cycles. A standard aminoallyl-PCR mixture included 400 ng of the ligate, 40 µl of 10x reaction-buffer (Sigma), 42 µl MgCl₂ [25mM], 3 µl aminoallyl-dNTP Mix [containing 15 mM aminoallyl-dUTP, 10 mM dTTP and 25 mM each dCTP, dGTP and dATP], 200 pmol primer (U-CG1a, AATT-1b or TA-1b, respectively), 3 µl Taq enzyme (5 U/µl, NEB) and water to a final volume of 400 µl.

Array hybridizations

Each microarray slide was prehybridized with a mixture consisting of DIG Easy Hyb (Roche Diagnostics), 25 µg/ml tRNA and 200 µg/ml BSA. The printed area was covered with the prehybridization mixture under a coverslip for 1h at 45°C. The microarray slides were then washed in two changes of water for 2 min at 45°C, followed by two wash-steps at room temperature and a final wash-step in isopropanol for 1 min. The slides were immediately blown dry with pressurized air and stored for hybridization. The hybridization mixtures were then pipetted onto the arrays and covered with Sigma Hybri-slips. The microarrays were placed in hybridization chambers (Corning Microarray Technologies, New York, USA) and incubated on a level surface for 16 h at 42°C for the COMT-arrays and 44 to 52°C for the CpG island microarrays in a covered water bath. The coverslips were removed by immersion of the arrays in a wash solution containing 2x SSC and 0.5% SDS (washing buffer I). The array was washed twice for 15 min at 42-52°C in washing buffer I (low stringency), followed by two wash-steps in washing buffer II (0.5x SSC, 0.5% SDS), followed by 2 min of incubation in water. The slides were then rinsed quickly in isopropanol and finally dried with pressurized air. The hybridization method used for the chromosome 21 and 22 tiling arrays was described before [176, 177].

Whole genome amplification

Genomic DNA was amplified using the GenomiPhi Kit (Amersham Biosciences) according to the manufacturer’s protocol. Briefly, 10 ng of gDNA (1 µl) was mixed with 9 µl of sample buffer, denatured at 95 °C for 3 min, cooled on ice and then added to 9 µl of reaction buffer and 1 µl of Phi29 DNA polymerase. The reaction was incubated at 30°C for 16 h and then
inactivated at 65 °C for 10 min.

**Bisulfite sequencing**

The methylation status of a number of CpG islands were analysed by direct sequencing of sodium bisulphite modified genomic DNA [155]. Genomic DNA samples were subjected to bisulfite modification using a standard protocol [188]. The primer sequences, PCR conditions and cloning methods are provided in the Supplement section.

**Genomic DNA**

Genomic DNA from all tissues was purified using standard laboratory methods (Phenol/Chloroform or Qiagen Blood and Cell DNA Midi columns). To avoid cross reactivity of amine groups with the aminoallyl-labeling procedure, DNA samples were stored in 0.5 M POPSO buffer (pH 8.0) instead of Tris-EDTA. Male placental DNA was purchased from Sigma and the *post-mortem* brain samples were provided by the Stanley Medical Research Institute. All parts of the study were approved by the CAMH review/ethics board.
Chapter 2

Single Nucleotide Extension Technology for Quantitative Site Specific Evaluation of metC/C in GC-Rich Regions

Zachary A. Kaminsky, Abbas Assadzadeh, James Flanagan, Arturas Petronis

The Krembil Family Epigenetics Laboratory
Centre for Addiction and Mental Health, and University of Toronto

Originally Published in *Nucleic Acids Research*

\(^1\)To whom correspondence should be addressed:

The Krembil Family Epigenetics Laboratory, Rm 28
Centre for Addiction and Mental Health
250 College St, Toronto, ON,
Canada M5T 1R8
Phone: +1-416-5358501-4880
Fax: +1-416-979-4666
e-mail: arturas_petronis@camh.net

Note: This work was not supported by the Applied Biosystems. There is no other conflict of interest.

Contributions: For this manuscript, I performed all laboratory experiments, experimental design, data analysis, and wrote the paper. Select experimental data were replicated by the other authors.
Abstract

The development and use of high throughput technologies for detailed mapping of methylated cytosines ($^\text{met}$C) is becoming of increasing importance for the expanding field of epigenetics. The single nucleotide primer extension reaction used for genotyping of single nucleotide polymorphisms has been recently adapted to interrogate the bisulfite modification induced ‘quantitative’ C/T polymorphism that corresponds to $^\text{met}$C/C in the native DNA. In this study, we explored the opportunity to investigate C/T (and G/A) ratios using the Applied Biosystems (ABI) SNaPshot technology. The main effort of this study was dedicated to addressing the complexities in the analysis of DNA methylation in GC-rich regions where interrogation of the target cytosine can be confounded by variable degrees of methylation in other cytosines (resulting in variable C/T or G/A ratios after treatment with bisulfite) in the annealing site of the interrogating primer. In our studies, the mismatches of the SNaPshot primer with the target DNA sequence resulted in a biasing effect of up to 70% while these effects decreased as the location of the polymorphic site moved upstream of the target cytosine. We demonstrated that the biasing effect can be corrected with the SNaPshot primers containing degenerative C/T and G/A nucleotides. A series of experiments using various permutations of quantitative C/T and G/A polymorphisms at various positions of the target DNA sequence demonstrated that SNaPshot is able to accurately report cytosine methylation levels with less than 5% average standard deviation from the true values. Given the relative simplicity of the method and the possibility to multiplex C/T and G/A interrogations, the SNaPshot approach may become a useful tool for large scale mapping of $^\text{met}$C.

Keywords: DNA Methylation, bisulfite modification, SNP, single nucleotide primer extension
Introduction

Technological advancement in DNA methylation analysis is an important and ongoing endeavor of epigenetic research. The gold standard technique for the fine mapping of methylated cytosines ($^{\text{methyl}}$C) utilizes bisulfite modification, which converts unmethylated cytosines into uracils, while the $^{\text{methyl}}$C remain unchanged. Following polymerase chain reaction (PCR), uracils are replaced by thymines (T), allowing differentiation between what was $^{\text{methyl}}$C and C by examining the proportions of C to T present in the PCR product at positions of interest. However, because DNA from different cells exhibit various degrees of epigenetic heterogeneity, it is necessary to clone the target sequences of the bisulfite treated DNA and sequence numerous clones (in the literature the number of clones varies from 3 to over 65 [118, 189]). Therefore, although precise, the method is very labor intensive and as a rule is limited to the analysis of relatively short (<1kb) DNA fragments. In mammals methylation targets are cytosines in CpG dinucleotides [190] and in some cases CpNpG [191] and therefore sequencing the entirety of the clone is not necessarily efficient as only a small fraction of the nucleotides are of interest. Because of this, a method that could efficiently analyze methylatable cytosines only is highly desirable. Since bisulfite modification induces DNA sequence changes (“induced DNA polymorphism”), the objective to identify $^{\text{methyl}}$C and C, which become C and T (or G and A in the complementary strand) after bisulfite modification, is quite similar to that of genotyping single nucleotide polymorphisms (SNPs). The difference, however, is that SNP genotyping generates three discreet outcomes (heterozygotes and two alternative homozygotes), while C/T (G/A in the complementary strand) ratios may exhibit a high degree of variation from 0%/100% to 100%/0% and anywhere in between. There also have been successful attempts to develop a protocol for SNP allele frequency estimation in DNA pools where allele frequency may vary between 100% and 0% (e.g. [192]. Among the myriad of techniques for the detection and measurement of SNPs, a frequently used approach is the single nucleotide primer extension reaction [193]. Several studies have provided evidence that methylation sensitive single nucleotide primer extension-based SNP genotyping (Ms-SNuPe) can be performed in a quantitative manner and thus could be useful for evaluation of $^{\text{methyl}}$C/C ratios [69, 70, 192, 194-199]. However, several important issues remained to be addressed. Firstly, can the mapping of $^{\text{methyl}}$C/C be performed in an automated high throughput fashion? Secondly, what are the effects of neighboring cytosines when their methylation status is unknown a priori? The latter is a typical situation for CpG islands, regions
of primary interest in DNA methylation studies. Until recently, many such regions have been investigated using Ms-SNuPe by avoiding CpG dinucleotides in the annealing site of the interrogating primer [189], which is a marked limitation of the technology. Recent attempts have been made to overcome the problem of Ms-SNuPe primer mismatch effects in order to interrogate CpG sites independent of sequence context, including GC–rich regions, using MALDI mass spectrometry [200]. While it is intuitive that primer binding mismatches may affect results, the effects of such mismatches have not been analyzed in previous studies.

As the need for high throughput experimental designs to investigate methylation profiles increases, so it seems that the techniques employed to accurately investigate them require increasingly specific equipment platforms that may not be available at all facilities. At present, the platforms that are capable of investigating complex methylation patterns unhindered by CpG rich regions are mass spectrometers, pyrosequencing machines, and a software analysis program of direct sequencing called ESME [66, 200-204], all of which are not very common in standard molecular biology laboratories. It was therefore one of our goals to successfully employ a SNaPshot technique that is both cost effective and easy to perform on an alternate equipment platform and therefore widen the availability of high throughput Ms-SNuPe reactions to those facilities with access to electrophoresis machines, such as the Applied Biosystems (ABI) Avant 3100 genetic analyzer. The main effort was dedicated to interrogation of methylated status in the GC-rich regions where the density of methylatable cytosines is high. After bisulfite modification such methylated and unmethylated cytosines will result in numerous ‘quantitative’ C/T and G/A polymorphisms, which may compromise binding of the interrogating primer and distort SNaPshot results. We have examined the effect of mismatches in the primer binding region and found a position dependant biasing effect. More importantly, however, we have demonstrated that SNaPshot primers containing degenerative C/T or G/A bases at potential mismatches can accurately quantify methylation at a given target and may produce multiple peak patterns that are indicative of methylation differences upstream of that target.

The analyses performed in this study can be stratified into three categories. The first category consisted of experiments to operationalize and test the ability of multiplexed SNaPshot primers to quantitatively measure C/T (G/A) ratios in the target sequences where there were no methylatable cytosines and therefore no possible primer mismatches with the bisulfite-induced variable C/T (G/A) in the SNaPshot primer binding region. In the second category, the more
complex scenarios were explored, such as when a SNaPshot primer binding region contains such quantitative C/T (G/A) polymorphisms, and strategies to reliably measure target C/T (G/A) ratios were developed. Finally, the third category experiments were dedicated to verification of the adapted SNaPshot approach on various oligonucleotides and bisulfite modified DNA targets.

**Materials and Methods**

**DNA sequence targets for SNaPshot interrogation**

Three types of DNA targets were generated to address various issues surrounding the adaptation of the SNaPshot approach for the quantitative evaluation of C/T (G/A) ratios:

1. A fragment of the promoter region of the gene encoding human catecholamine –o-methyltransferase (*COMT*) was amplified as follows: 10x PCR Buffer, 2mM MgCl₂, 2.5 mM dNTP, 1M Betaine, 0.4 mM primers, and 1 U of Taq polymerase (New England Biolabs), primers: `comtF1 5'-agaccacaggtgcagtcagcacag-3'` and `comtR1 5'-caccctatcccagtgttccaccta-3'` at 95°C - 5 min, (94°C -1min 61°C -1.5 min, 72°C -1min) 30 cycles, and 72°C -5min.

CCGG and GCGC sites of the amplicon were subsequently methylated using *M-HpaII* and *M-HhaI*, respectively, in two separate fractions. The third fraction of the amplicon was left unmethylated. Both methylated and unmethylated DNA samples were then subjected to bisulfite modification [188]. Briefly, DNA was boiled for 5 minutes, cooled on ice, and denatured for 15 minutes at 50°C after adding 4 µl of fresh 2M NaOH in a total reaction volume of 25 µl. Two volumes of 2% LMP agarose in distilled water was added and 10 µl aliquots of this solution was pipetted into cold mineral oil and placed immediately back into dry ice to create beads. The mineral oil was removed and a solution comprised of 1.9g sodium metabisulfite in 2.5 ml H₂O, 720 µl of 2M NaOH, and 500 µl of 1mM hydroquinone was added. Samples were incubated on ice for 30 minutes followed by incubation at 50°C for 3.5 hours. The agarose beads were washed 4x for15 minutes with 1 ml TE, 2x for 15 minutes with 0.2 M NaOH, 3x for10 minutes with 1 ml TE, and 2x for 15 minutes with H₂O. This was followed by semi-nested PCR using identical reaction and cycling conditions as above with semi-nested primers: `BisF1 5’-gaagggggtatttggtgtagaa-3’`, `BisF2 5’-gatttttgagtaagaggtagataag-3’` and `BisR1 5’-aacaacctacactacccaa-3’`. 
C (\textsuperscript{methyl}C in the amplicon) containing templates were mixed with the T (unmethylated C in the amplicon) containing fraction to create a standard curve from 100% to 0% of C signal in increments of 5%: 100%C: 0%T; 95%C: 5%T; 90%C: 10%T… 0%C: 100%T. This was done for those templates containing C at \textit{M-HpaII} sites and \textit{M-HhaI} sites separately. The \textit{M-HpaII} sites were interrogated with three forward primers, while three primers for the \textit{M-HhaI} sites were added as negative controls. In a similar way, \textit{M-HhaI} sites were interrogated with all six forward primers (three of which were for the \textit{M-HpaII} sites as negative controls) in one run and the three reverse primers in a second run (Figure 2.1). The interrogating primers were designed to have a T\textsubscript{m} close to 50°C to allow for similar annealing dynamics in the multiplexed reaction. To vary the length of the primers, non-complementary tails were designed on the 5’- end of each primer by repeating the sequence GACT [shown in brackets]: at least two sets of the GACT for oligos with the total length under 40 nucleotides and by one set of GACT for those above 40 nucleotides. The interrogating SNaPshot primers were:

\begin{itemize}
  \item[a)] 5’-agtaagattagattaagagt-3’; 5’-[gact]_1gatatttttaggatat-3’, and 5’-[gact]_0ttatggtgtggttat-3’ for the \textit{HpaII} sites;
  \item[b)] 5’- [gact]_4gatattttggttattttg-3’; 5’- [gact]_6ttttgatattttttggttt-3’, and 5’- [gact]_7agttttatatattttgtag-3’ for the \textit{Hha I} sites (direct primers);
  \item[c)] 5’- ccacaataaatatccac-3’; 5’-[gact]_2tataaacaacaaacacaaac-3’; and 5’-[gact]_3acactacaaaaatataaaaaac-3’ for the remaining three \textit{Hha I} sites (reverse primers).
\end{itemize}

2. In order to investigate the effects of quantitative G/A and C/T polymorphisms in the SNaPshot primer binding region, sets of oligonucleotides containing variable C/T and G/A were synthesized. Five polymorphic positions were investigated: -2, -5, -10, -15, and -18 (A\_2/G\_2, A\_5/G\_5, T\_10/C\_10, A\_15/G\_15, T\_18/C\_18) upstream of the nucleotide that was interrogated (“target” nucleotide, N\_target)(Figure 2.2). SNaPshot primers in the experiment were named according to the polymorphic site, while the DNA template itself was named by the nucleotide in the polymorphic position and also the target nucleotide. Therefore, the primer T\_2 is fully complementary to the templates A\_2 A\_target and A\_2 G\_target, but not complementary at the -2 position to the templates G\_2 A\_target and G\_2 G\_target. It is evident that the T\_2 primer will preferentially bind to (and interrogate) the DNA sequence that contains A\_2, in comparison to
G at the upstream position. The degree of such bias, however, is unknown as is the impact of the location of the mismatch proximal to the target nucleotide. To elucidate the degree of bias, DNA templates containing an upstream polymorphism, e.g. G and A, were added in equal amounts which resulted in a polymorphic G/A site at the –2 position. This template mix was tested in two different primer scenarios: first with primer T, then with primer C. All other polymorphic sites at positions –5, -10, -15, and –18 were analyzed in the same way. For the mismatch bias correction, numerous DNA template combinations with varying percentage of polymorphic nucleotides in the primer binding site were tested using different primer combinations (see Results).

3. To verify that degenerative SNaPshot primers are able to interrogate numerous polymorphic C/T and A/G containing DNA sequence targets, two types of DNA templates were used:

   a) Six oligonucleotide templates were synthesized with quantitative G/A polymorphisms in different positions of the SNaPshot primer annealing region (Figure 2.3). The target nucleotide A/G proportions were synthesized to be 50%:50% in each template while the upstream A/G ratios were synthesized according to Figure 2.3. SNaPshot primers contained a 50%:50% proportion of C/T at degenerative positions corresponding to polymorphic positions in the templates.

   b) Two human genomic DNA samples from brain and placenta were bisulfite modified and subjected to both SNaPshot interrogation and to cloning plus sequencing-based measurement of methylated C density. The two selected CpG island regions were identified as exhibiting DNA methylation differences according to our microarray-based DNA methylation profiling (Schumacher, Petronis, et al. unpublished). These regions are located 28 and 276 bp upstream of known genes coding for LGALS1 (lectin, galactoside-binding, soluble, 1), otherwise referred to as galecin 1, and humanin, respectively. Three CpG positions were selected for each CpG island and will be refered to as gal1, gal2, and gal3 for galecin 1 and hum1, hum2, and hum3 for humanin (Figure 2.4).
Figure 2.1. Bisulfite modified COMT promoter region and SNaPshot primers

A continuous sequence of DNA representing the bisulfite modified COMT promoter region amplified as a template for the SNaPshot multiplexing experiments. Only the top strand of DNA is depicted along with the positions of all SNaPshot primers used. Polymorphic target sites created by bisulfite modifications of M-HpaII and M-HhaI methylated amplicons are marked in bold capital letters. Primers overlapping M-HhaI sites (bottom right) were designed with T at the overlapping CpG positions. Forward SNaPshot primers are boxed above the sequence while reverse SNaPshot primers are boxed in red below the sequence. The non-binding GACT repeat tails (placed on the 3' end of some primers) are denoted by a number, the purpose of which is to vary the primer length in order to distinguish them in the ABI 3100 Genetic Analyzer.
Figure 2.2. Single SNP oligonucleotide templates

Oligonucleotide templates and complementary primers were synthesized to test the effects of C/T and G/A polymorphisms upstream of the target nucleotide at positions -2, -5, -10, -15 and -18 bp. Sequences are depicted for C/T and G/A polymorphisms at positions –2, -10, and –18; however, primers and templates were also tested for positions –5 and –10. Templates are named for the complementary strand to which the SNaPshot primer binds, the two nucleotides representing the polymorphic position (signified by the number) and the target nucleotide, respectively. SNaPshot primers are named according to the nucleotide complementary to the upstream polymorphism.
Figure 2.3. Multiple SNP oligonucleotide templates

A list of oligonucleotide templates synthesized to contain between 1-4 polymorphic positions in the SNaPshot primer binding region and respective primers containing degenerative bases at positions corresponding to those polymorphisms. Percentages of nucleotides synthesized into the templates are depicted while SNaPshot primers were designed with a 50%:50% proportion of C/T at all polymorphic positions.

Figure 2.4. Galectin1 and Humanin SNaPshot primers
SNaPshot primers for the galectin1 (A) and humanin (B) genes that were identified as being differentially methylated between placenta and brain tissue.

Bisulfite modification reactions were performed as described above. Target sequences were amplified using fully nested PCR. PCR conditions were as follows: 10x PCR Buffer (Sigma), 2mM MgCl₂, 2.5 mM dNTP, 0.4 mM primers, and 1 U of Taq polymerase. The first PCR was performed using primers, for galectin 1: 22_f1 5'-gtagaatgtaattttggtagaataat-3' plus 22_r1 5'-ctcaaccatctctcaaccac-3'; and for humanin: 52_f1 5'-agtggtaggttaagttagttaggtat-3' plus 52_r1 5'-aaccaacaaaaacacacac-3'. The second (nested) PCR used primers, for galectin 1: 22_f2: 5'-gttattgggggyggagtt-3' plus 22_r2 5'-acttataaacctaactcatcaactat-3', and for humanin: 52_f2: 5'-aattagatttttgagtttaaatccacataac-3' plus 52_r225'-aaccaacaaaaacacacac-3' site. Two successive rounds of touch down PCR was used with cycling conditions consisting of 95°C- 3 min, [94°C -1min 60 °C–30 sec (minus 1 °C/cycle), 72 °C –40 sec] 10 cycles, [94°C –1 min, 50 °C–30 sec, 72°C –40 sec ]-30 cycles, 72°C-5 min. The sequences of the interrogating SNaPshot primers were: gal1 5'-gttattggggyggagtt -3'; gal2 5'-[gact]2 gaggatgttttyggttagg-3'; and gal3 5'-[gact]4 gatygggtygggtagtt-3'. Primers for humanin were: hum1 5'-acagttyggattttgagtttaaatccacataac-3'; hum2 5'-aaccaacaaaaacacacac-3'; and hum3 5'-ygagggtagagggagagggagag-3'. Amplicons were cloned into the pDrive plasmid (Qiagen) that were used for transformation of DH5-α competent cells. Individual colonies were grown at 37°C for 15 hours followed by plasmid purification using the Qiagen Spin Miniprep kit. Sequencing of 12-15 plasmid inserts per template was carried out with M13 reverse primer using ABI Big Dye Terminator kit 3.1. Six CpG positions were investigated using SNaPshot primers individually and five were multiplexed in two groups, (gal1, gal2, and gal3) and (hum1 and hum3). The differences in length of degenerative primers to be multiplexed were in accordance with the specifications recommended by the manufacturer (ABI). All SNaPshot experiments on bisulfite modified DNA were repeated in quadruplicate.
SNaPshot

SNaPshot reactions were carried out using 0.01pmol - 0.4pmol of DNA template, 2 µM primers in the final reaction, 5 µl of SNaPshot master mix (ABI), and water to a final reaction volume of 10 µl. Reactions were carried out in the ABI 9700 Thermocycler at 96°C for 10 sec, 50°C for 5 sec, and 60°C for 30 sec, and repeated for a total of 25 cycles. Following cycling, samples were treated with 3 U of calf intestine phosphatase at 37°C for 1 hr and heat inactivated at 72°C for 15 min. In an ABI optical plate, 9 µl HI-DI formamide, 0.5 µl of Genescan 120 LIZ size standard (ABI), and 0.5 µl of the reaction were combined, denatured at 95°C for 5 min, and immediately placed on ice for 2 min. Samples were then loaded on the ABI Avante 3100 Genetic Analyzer for analysis. To determine which target produced which signal in a multiplexed reaction, peak position was correlated with the length of the primer designed for a given target. The X-axis reports the relative length of the primers compared to the loaded size marker. Therefore, shorter SNaPshot primers will produce peaks closer to the origin on the X-axis while longer ones will produce peaks farther from it. It should be noted that a C and T peak incorporated by the same primer should be offset by at least 0.5 nucleotide spacing because primers with different ddNTP compositions travel at different rates through the polymer matrix in the ABI machine. The precise peak height was determined by the Genescan 6.0 analysis software, which provides a specific data output below the graph of peaks that includes the relative size in bp, peak height, and peak area for each primer. The data output corresponding to any given peak can be determined by clicking on a peak, which will in turn highlight the corresponding data cell below. The percentage of C at a given position was determined by the formula: \( C\% = \frac{(C_i)}{(C_i + T_i)} \times 100 \), where \( C_i \) and \( T_i \) stand for the peak height of C and T signal, respectively. When interrogating the reverse strand, G and A peaks will be produced in place of C and T peaks. The same formula can be used in this case by substituting \( G_i \) and \( A_i \) for \( C_i \) and \( T_i \), respectively. Peak area should not be used as diffusion over large distances in the polymer matrix may affect the consistency of results. In cases where multiple peaks for a single target were observed, cumulative intensities for all C peaks and all T peaks were substituted into \( C_i \) and \( T_i \), respectively, in the above formula.

Results

1. Simple DNA templates: multiplexed SNaPshot reactions
The maximum number of SNaPshot primers that would fit on the COMT amplicon in the forward direction without competing for annealing sites was 6 (Figure 2.1). The remaining three primers were designed for the reverse strand and run in a separate reaction (ibid.). To test the multiplexing capability, we created a standard curve of decreasing C signal - increasing T signal from 100% to 0% in increments of 5% (100%C: 0%T) (95%C; 5%T) (90%C: 10%T)…(0%C: 100%T) in each of the 6 sites by mixing methylated (C) and unmethylated (T) templates (Figure 2.5). In the SNaPshot reaction, the primers bind to their complementary strand and incorporate a single fluorescent ddNTP. Primers are then separated and measured proportionally in the ABI 3100 Genetic Analyzer. Consequently, those primers designed in the forward and reverse directions produce C/T and G/A peaks, respectively. Methylated HpaII site- or methylated HhaI site- derived templates were separately tested with all six forward SNaPshot primers. Three sets of experiments were performed in triplicate for templates comprised of i) three HpaII sites and ii) three HhaI sites interrogated by forward primers, and iii) three HhaI sites interrogated by reverse primers. The results showed that all primers produced mean C/T and G/A (for the reverse primers) ratios reflecting the expected values (the range of standard deviations [SD]: 2.7%-4.6%) (Figure 2.6).
A graph of the average methylation values reported by nine primers interrogating control templates created to contain only C or T at the CpG islands of interest. Each dilution series was tested in triplicate for each primer so that each data point is an average of 27 experiments. Templates with all CpG islands of interest containing C were diluted and mixed in increments of 5% with templates containing T at all CpG being interrogated to test the ability of the primers to accurately measure the amount of methylation.

![Graph of average methylation values](image)

**Figure 2.6. Multiplexed SNpShot reaction output**

Data output from the ABI Avante 3100 combining 60% bisulfite treated \( M\-Hpai \) methylated template and 40% unmethylated template. The peak heights show 60% C to 40% T signal for those peaks methylated with \( M\-Hpai \) (peak pairs 1, 2, and 4). Peaks 3, 5, and 6, representing \( M\-Hha \) sites, show no C signal and hence no methylation. Peak order is determined by primer size.

### 2. Complex DNA templates: SNpShot Primer Mismatch Bias and Bias Correction

#### Mismatch bias

One of the key objectives of the study was to investigate the suitability of the primer extension technique for the investigation of DNA methylation patterns when dealing with CpG dinucleotides (that can be methylated to different proportions) in close proximity to the target cytosine. Depending on their \( a\ priori \) unknown methylation status, such CpGs may generate a wide variety of C/T or G/A upstream polymorphisms that may lead to one or more mismatches between the DNA template and the primer and thus result in incorrect measurement of C/T or G/A ratios at the target site. Templates listed in 1. 2 were mixed in equal amounts to produce a 50%:50% ratio of the target and upstream position. For example, the mix of \( G_NG_{target} \) and \( A_NA_{target} \) templates was tested with each complementary primer, \( C_N \) and \( T_N \), individually. The degree to which a base mismatch could bias the reflected proportion of target G/A decreased
when it moved farther upstream in the 5’ direction (Figure 2.7). The $G_2/A_2$ mismatch biased the results between 40% and 65% while the $G_5/A_4$ mismatch resulted in 25%-70% bias. The average for the $T_{-10}/C_{-10}$, $G_{-15}/A_{-15}$, and $T_{-18}/C_{-18}$ mismatches was 35%, 23%, and 9%, respectively (ibid.)

![Figure 2.7. Primer mismatch induced bias](image)

**Figure 2.7. Primer mismatch induced bias**

A graphical representation of the percentage that a mismatch in the primer binding region at various positions upstream from the target CG can affect the reading. There is a correlation between the proximity of a primer mismatch to the target and the degree to which the resulting $\text{meC/C}$ reading will be inaccurate. Zero baseline represents 0% biasing effect.

**Mismatch bias correction**

Equal template mixes used to determine the degree of mismatch bias in the above experiment, e.g. $G_NG_{\text{target}}$ and $A_NA_{\text{target}}$, were now interrogated with a 50% mix of SNaPshot primers complementary to each template. In all cases, the expected 50% proportion of represented targets was observed. An additional set of experiments was performed to test if such an equal representation of primers could accurately identify known amounts of target A/G nucleotides independent of the upstream polymorphisms that cause bias. A series of dilutions of oligonucleotide templates were prepared to interrogate $G_{\text{target}}/A_{\text{target}}$ in decrements of 25% from 100% G to 0% G (1.8). For each dilution series, the $-2$ and $-5$ positions ($G_{-2}/A_{-2}$ and $G_{-5}/A_{-5}$)
polymorphisms also varied. Each dilution series was tested in duplicate with each of the following five scenarios of upstream polymorphism concentration: 1) 100%G_N vs. 0% A_N; 2) 75%G_N vs. 25% A_N; 3) 50% G_N vs. 50% A_N; 4) 25% G_N vs. 25% A_N; and 5) 0% G_N vs. 100% A_N. A 50%:50% primer mix, primers C_2/T_2 (50%:50%) and primers C_5/T_5 (50%:50%), was able to accurately quantify the amounts of the target present in each of the above scenarios (SD 2.7%-5.6%) (Figure 2.8A-B). To test the effects of multiple upstream mismatches, two sets of experiments were performed in triplicate using randomly selected combinations of templates with all 5 upstream polymorphisms (1. 8C). These dilution series were tested with all 10 primers represented in equal amounts, accounting for each template polymorphism (primers C_2, T_2, C_5, T_5, G_10, A_10, C_15, T_15, A_18, G_18 (each at 10%). Data output reflected the known amounts of target (SD 0.79%-5.0%). The presence of primers in the reaction mixture for which there was no complementary template did not bias results. For all experiments run without an equal representation of all primers, curves were drastically biased and results did not reflect the known amounts of target nucleotides (Supplementary Figure 2.1).
Figure 2.8. Correction of mismatch induced bias

A) Data points produced by 25% increments of $G_{\text{target}}/A_{\text{target}}$ templates while varying the percentage of polymorphic G/A 2 bp upstream from the target nucleotide. B) Data points produced by 25% increments $G_{\text{target}}/A_{\text{target}}$ templates while varying the percentage of polymorphic G/A 5 bp upstream from the target nucleotide. Results for each data point in the –2 (A) and –5 (B) permutations represent an average of 10 experiments. C) Data points produced by diluting the $G_{\text{target}}/A_{\text{target}}$ 25% increments while varying the percentage of polymorphic nucleotides in the amounts shown to the right.
3. C/T (G/A) interrogations in complex oligonucleotide templates and bisulfite modified DNA

Degenerative primer experiments on oligonucleotide templates

To test the ability of primers designed with C/T and G/A degenerative bases at polymorphic positions to interrogate a given target, six oligonucleotide templates and corresponding primers were designed as shown in Figure 2.3. All templates were 50%:50% of A and G (A₅ and G₅) at the target position and variable upstream proportions (0.1-0.9; 0.2-0.8; 0.3-0.7; 0.4-0.6; etc) of C/T and G/A ranging from one to four positions. Primers had a 50%:50% mix of C/T (C₅ and T₅) at positions corresponding to any polymorphic positions in the template. All corresponding SNaPshot primers with degenerative bases at polymorphic positions accurately reported equal intensities of fluorescent C and fluorescent T that correspond the 50%:50% of G and A in templates (SD = 0.22% - 3.5%) (data not shown).

Degenerative primer experiments on bisulfite modified DNA

The SNaPshot approach was used to identify DNA methylation differences in several genes that exhibit tissue (brain and placenta) specific differential epigenetic modification (Schumacher, Petronis, et al., unpublished). Such target sequences have also been subjected to cloning of bisulfite modified DNA and direct sequencing of individuals clones was performed on at least 12 clones. Five CpG positions were investigated using SNaPshot primers in multiplexed primer groups (gal2, gal3, and gal1) and (hum1 and hum3). Primer hum2 was not multiplexed because it was designed as a reverse primer and exhibited a high degree of sequence homology with primers hum1 and hum3. All SNaPshot experiments on bisulfite modified DNA were repeated in quadruplicate. In all experiments, an identical peak pattern was observed and the proportions of all peaks remained identical between runs for each sample. The methylation status of target positions determined by SNaPshot and by sequencing individual clones was similar within 5% (SD 1.0%–3.16%) (Figures 1.9 and 1.10A).
Figure 2.9. SNaPshot vs. cloning and sequencing data

Graphical representation of methylation profiles quantified by sequencing of at least 12 clones of bisulfite modified genomic DNA and SNaPshot for six CpG dinucleotide positions in bisulfite modified DNA amplified from brain tissue (A) and placenta tissue (B).
Figure 2.10. Multiple peaks

A) SNaPshot results on bisulfite modified DNA from brain interrogated with primers gal1, gal2, and gal3. B) The peak pattern identified with gal1 SnPshot primer only. C) A depiction of how the proportion of multiple peaks in the scenario of a single upstream polymorphism were indicative of the methylation profile of the upstream CpG (Y position) measured by gal1 and verified by sequencing of bisulfite modified genomic DNA. D) SNaPshot peaks resultant from an interrogation of the Gal1 upstream CpG (Y) using primer sequence: 5’- TTGGGGGTATTGGGG-3’.
Some of the experiments revealed multiple C or T peaks. Such patterns were observed only when SNaPshot primers with degenerative positions were used, which indicates that this is due to subtle differences in electrophoretic mobility of primers containing nucleotide differences. If two primer variants, e.g. T_N and C_N, both incorporate a fluorescent ddTTP at the target position, two T peaks may be observed. Results for those primers producing multiple peaks were calculated by incorporating all C signals for a given primer cumulatively and comparing them to the cumulative T signals. Calculating C/T ratios in this way increased the resolution of the method and decreased the standard deviation between replicates on average by 0.7%.

An interesting example of multiple peaks was observed from SNaPshot primer gal1, which generated two T peaks, the ratio of which was consistently 22.5%/77.5% (Figure 2.10 B, C). This finding suggested that the only degenerative nucleotide (Y) in the primer corresponded to a differentially modified cytosine in the genomic DNA. Investigation of 15 clones of bisulfite treated DNA from this tissue identified that this CpG position was in fact differentially methylated at a frequency of 13%/87% \textsuperscript{met}C/C. A dedicated SNaPshot primer was synthesized to directly interrogate the CpG position located within the primer gal1 binding region and reported a C/T ratio of 15%/85% (Figure 2.10D). Additionally, the primer 1/template 1 interrogation (Figure 2.3) produced a set of two C peaks, template 1 being synthesized with a single degenerative position displaying 20% C and 80% T. In this case, the proportion of the smaller peak of the C peaks to the cumulative C signal ranged from 19%-23% C between replicates. This suggests that in addition to the target CpG, sometimes a SNaPshot primer containing one degenerative base can quantitatively measure \textsuperscript{met}C/C proportion at the upstream CpG. Primers with more than one degenerative base, however, have more than two possible variants, making identification of the presence of upstream methylation differences possible, though not quantifiable.

**Discussion**

In order to expedite mapping of methylated cytosines, we have operationalized the Applied Biosystems method known as SNaPshot to quantitatively discern C/T and G/A proportions in the bisulfite modified DNA sequence, which reflects \textsuperscript{met}C/C ratios in the native DNA. In this method, the primer extension reaction is performed using fluorescent 2’-3’-dideoxynucleotide triphosphates that extend the 3’- end of an interrogating primer designed to bind exactly to the target site. The primers, now labeled by specific fluorescent dyes, are measured by an
electrophoresis platform and indicate the ratios of the polymorphic C/T and G/A nucleotides at the target position. The mean C/T and G/A ratios generated by all primers displayed a standard deviation range of 2.7%-4.6% with the known C/T and G/A values falling within this range. This standard deviation range is similar to other studies in which the SNaPshot method displayed a range between 0% and 3.7% in measuring C/T ratios [199].

Accurate measures of bisulfite induced C/T (G/A) polymorphisms, however, strongly depend on a perfect match of the interrogating primer and the DNA sequence. When methylatable cytosines are present in the primer annealing region, bisulfite modification will generate quantitative C/T (or G/A) polymorphisms that may lead to partial or full mismatch with the primer. According to our data, a mismatched nucleotide is capable of biasing the results up to 70%, depending on the position of the mismatch. In an unknown system where the upstream polymorphisms could vary by any degree, as in the case of the unknown \textit{a priori} methylation status of bisulfite treated CpG dinucleotides, it becomes increasingly important to have reliable interrogative capability. A relatively large degree of bias was observed by one group using the SNaPshot approach to interrogate a single CpG that had previously been shown to vary in methylation status between tumor and non-tumor tissue [199, 205]. Our interpretation is that the observed bias could have been produced by a mismatch in the SNaPshot primer binding region. This might have occurred if primers were designed to bind template DNA containing a methylatable CpG or CpNpG, or in the case of incomplete bisulfite conversion upstream of the target nucleotide. Another possibility is the fact that the authors used peak area instead of peak height for their analysis. While this is less likely to produce the levels of bias observed, ABI technical support warns that peak area readings may vary as a result of primer diffusion in the ABI polymer matrix. Consistently with other published work [189, 195] we detected that primer design in GC- rich areas is of primary importance to address the biasing effects of neighboring polymorphisms. According to our data, primers containing up to five degenerative bases at the positions of the putative mismatches can accurately interrogate a given target. Equal mixture of all possible primers complementary to any of the possible polymorphisms is able to accurately interrogate and reflect the actual percentages of target nucleotides, independently of the quantitative upstream polymorphisms.

It is important to note that degenerative primers may exhibit subtle differential mobility in the electrophoresis platform, which results in multiple SNaPshot peaks. For example, if two primer variants, one with an upstream C and one with an upstream T, both incorporate a fluorescent T at
the target position, two T peaks may be observed. This is consistent with the finding that multiple peaks were only observed on templates containing upstream methylation differences (although not all templates containing upstream differences exhibited multiple peaks). We detected that accurate results were obtained from combining together all C intensities belonging to the same target site and comparing them to the cumulative T intensities. That is, while the migration behaviour of primer variants may split a given C or T target signal, the cumulative C and T proportions represented in the data output remain the same relative to each other. Absence of multiple peaks may indicate that there is no sequence variation upstream of the target nucleotide or that the migration dynamics of the moiety of primers is similar.

The presence of multiple peaks may add informativeness to the method as SNaPshot primer variants may reflect C/T or G/A variation in binding regions of the bisulfite treated DNA templates. In some cases (e.g. gal1, see the Results and Figure 2.10), the methylation status of a CpG located upstream of the target CpG was identified from a multiple peak pattern. In such cases, only two primer variants are present and a maximum of two peaks can be produced from either or both the C or T signal resultant from upstream methylation. The nucleotide differences between primer variants causes i) a preferential binding to the most complementary template and ii) a spatial segregation of these primer variants in the electrophoretic matrix. Therefore, the percentage of methylation appears to be proportional to the difference of the two multiple peaks. Because multiple peaks may be small it becomes important to take the signal intensities of the overall reaction into consideration. According to the ABI machine specifications, a peak intensity ranging between 50 and 2000 is optimal for accurate readings. Those peaks with an intensity lower than 50 cannot be differentiated from the background and will not be recognized by the machine. Peaks with very high signal intensities have the danger of pulling up background that could be confused for multiple peaks.

When there is more than one degenerative base in the SNaPshot primer, it is not trivial to assign multiple peaks to a specific CpG in the primer annealing region. In order to avoid possible overlaps of peaks belonging to different primers, one possibility is to increase the size difference of those primers designed. If the result is still in question, re-running a given SNaPshot primer individually on the same template can be recommended. In general, since SNaPshot primers with degenerative bases can identify the presence of \textsuperscript{met}C/C in the primer binding site, DNA methylation analysis can be optimized by a two tiered approach. The first tier would use a single
degenerative SNaPshot primer to scan high CpG density areas for possible complex peak (i.e. differential DNA methylation) patterns. The second round of SNaPshot primers may then be designed to interrogate specific CpG dinucleotides implicated in the first round of SNaPshot experiments.

In comparison to other methods used to quantitatively estimate C/T (G/A) ratios, the SNaPshot approach exhibits several advantages. Adding non-binding “tails” to the 5’-ends of multiple SNaPshot primers changes their lengths relative to each other and results in a spatial segregation during electrophoresis. This multiplexing technique enables multiple CpG sites to be interrogated in a single reaction. In our experiments, six primer extension reactions were multiplexed without any evidence for compromising the quality of interrogations. Given longer amplicons, the SNaPshot method can potentially accommodate 10 primers according to the manufacturer’s specifications [206, 207]. It should be noted that another approach has avoided the issues of reading multiple primer variants produced by primers using degenerative bases by incorporating tagged bases in the interrogating primer. This allows degenerative bases to be removed, thereby creating a core sequence of uniform size, which is in turn measured by MALDI mass spectrometry [200]. Major differences between the SNaPshot and the MALDI mass spectrometry approaches involve the generation of a 4 nucleotide core sequence to avoid confounding signal and the use of B-cyanoethyl phosphoramidite nucleotide tags as opposed to fluorescent ddNTPs. While the two assays are very similar at the early stages, advantages of the electrophoresis platform lie in a greater multiplexing capability and no restrictions regarding incorporation of a degenerative base within 4 nucleotides of the target position. Additionally, the production of multiple peaks in the electrophoresis platform adds informativeness over a larger DNA area using fewer primers, which translates into a lower cost.

While the multiple peaks highlighted in this method allow for a two tiered investigation of larger CpG rich regions of DNA, other technologies such as pyrosequencing [66, 201, 202] and the newly developed ESME approach [203] enable identification of methylation profiles of multiple CpG dinucleotides as well as the surrounding sequence in a single reaction. The situation in pyrosequencing, however, would be optimal when interrogating a number of CpG dinucleotides within close proximity, while the SNaPshot method can interrogate any CpG for which a primer has been designed without spatial constraint. The detection limit of methylation differences in the ESME approach is 20% or greater, while pyrosequencing and SNaPshot can provide accurate
measures of methylation within approximately 5%. While the pyrosequencing and the ESME approaches offer attractive features, the added informativeness, ease, and affordability of the modified SNaPshot approach makes it a convenient alternative for those laboratories with an electrophoresis platform.
Chapter 3

Epigenetics of Personality Traits: An Illustrative Study of Identical Twins Discordant for Risk Taking Behavior

Zachary Kaminsky¹,⁴, Arturas Petronis¹,⁴, Sun-Chong Wang¹,⁶, Brian Levine²,⁴, Omar Ghaffar³,⁴, Darlene Floden⁵, Anthony Feinstein ³,⁴*

Originally Published in Twin Research and Human Genetics

¹ The Krembil Family Epigenetics Laboratory, Centre for Addiction and Mental Health, Toronto, Ontario
² Rotman Research Institute, Toronto, Ontario
³ Sunnybrook Health Sciences Centre, Toronto, Ontario
⁴ University of Toronto, Toronto, Ontario
⁵ Cleveland Clinic, Cleveland, Ohio
⁶ Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan

*Corresponding Author:
Anthony Feinstein MPhil., PhD., FRCPC
Professor, Department of Psychiatry
University of Toronto
Sunnybrook Health Sciences Centre
2075 Bayview Avenue
Toronto, Ontario
Canada M4N 3M5

Tel: 416.480-4216
Fax: 416.480-4613
email: antfeinstein@aol.com

Running Head: Motivation and epigenetics of monozygotic twins

Contributions: For this manuscript, I performed all epigenetics related laboratory experiments, experimental design, data analysis, and wrote the paper for sections pertaining to epigenetics. Psychological testing was performed by Anthony Feinstein and colleagues, while zygosity testing was performed by Proactive Genetics INC.
Abstract

DNA methylation differences between identical twins can account for phenotypic twin discordance of behavioral traits and diseases. High throughput epigenomic microarray profiling can be a strategy of choice for identification of epigenetic differences in phenotypically different monozygotic (MZ) twins. Epigenomic profiling of a pair of MZ twins with quantified measures of psychometric discordance identified several DNA methylation differences; select differences may have developmental and behavioral implications and are consistent with the contrasting psychometric profiles of the twins. In particular, differential methylation of CpG islands proximal to the homeobox DLX1 gene could modulate stress responses and risk taking behavior and deserve further attention as a potential marker of aversion to danger. The epigenetic difference detected at DLX1 of ~1.2 fold change was used to evaluate experimental design issues such as the required numbers of technical replicates. It also enabled us to estimate the power this technique would have to detect an epigenetic difference related to phenotype given a range of 1-50 twin pairs. We saw that, use of epigenomic microarray profiling in a relatively small number (15-25) of phenotypically discordant twin pairs has sufficient power to detect 1.2 fold epigenetic changes.

Keywords: Epigenetics, DNA methylation, Monozygotic Twins, War Journalism, Risk Taking Behavior
**Introduction**

Phenotypic differences between identical (monozygotic, MZ) twins has been poorly explained by measurable environmental discordance. Various degrees of phenotypic differences have been observed in all traits ranging from normal behavior and normal traits to manifestation of complex disease [80]. MZ twins arise from a single fertilization and later split in early embryogenesis to develop into two distinct human beings. Because of their common gametic origin, MZ twins are identical at the DNA sequence level and yet they often display phenotypic differences later in life [80]. Traditionally, such twin discordance is attributed to environmental influences acting differently on each twin [208]. A number of epidemiological studies in the past years questioned this assertion by measuring phenotypic variation in MZ twins raised in the same environment vs. those raised apart and found measurable environmental factors insufficient to explain observed differences, thus bringing into question the underlying mechanisms of MZ twin discordance [107].

In attempts to address this question, recent attention has returned to molecular studies. Looking beyond the identical DNA sequence in MZ twins, numerous differences in epigenetic patterns between twins have been identified, some of which are believed to result in the observed discordant phenotypes [117-119, 121, 209, 210]. Epigenetic signals are molecular signatures that control various aspects of genome organization, including gene transcription. Epigenetic modifications consist of methylation of cytosines as well as modifications of histones including methylation, phosphorylation, acetylation, and ubiquitination [17-21]. In mammals, DNA methylation occurs most commonly where cytosine is directly followed by guanine, forming a CpG dinucleotide. Clusters of CpG dinucleotides are referred to as CpG islands [3]. Many CpG islands across the genome are located proximal to gene promoters and their DNA methylation status can affect levels of gene transcription through modulation of chromatin conformation and sequestration of components of the basal transcription machinery [6, 7]. DNA methylation has been shown to restrict the access of transcription factors and limit gene expression [4, 5]. Conversely, unmethylated DNA is associated with an open chromatin conformation, allowing for access of DNA binding elements and transcriptional activation [2, 22, 211, 212]. In addition to the promoter, the 3’ untranslated region (3’ UTR) of a gene can also be an important regulator of expression levels and can be directly affected by DNA methylation [213, 214]. DNA methylation
differences may arise between twins as a result of a drifting of epigenetic signals over time as well as through random inactivation of one of the two x chromosomes in females. DNA methylation differences between twins may therefore translate into the expression of phenotypic differences through varying levels of gene transcription.

Since epigenetic factors in MZ twins may be the underpinning molecular mechanism to explain phenotypic differences between identical co-twins, it is critical to map epigenetic differences between twins to enable discovery of epigenetic changes, which may be associated with discordance for complex traits. With this in mind, our primary objective was to employ a novel epigenomic microarray profiling strategy on a pair of normal MZ twins. The key question we attempted to address was whether high throughput microarray–based epigenetic profiling can reliably identify DNA methylation differences in normal twins, and if so, what size and how many epigenetic differences can be detected in twins differing for normal psychological traits. The co-twins investigated for this purpose exhibited major differences for various psychometric measures of risk taking behavior. Our secondary objective was to investigate any links between the identified differences and quantified phenotypic differences and to estimate the power of the technique for identifying possible etiological epigenetic differences in larger populations of twins discordant for normal behavioral traits. In fulfilling these objectives, the study provides an illustrative example of experimental design and power considerations when performing epigenomic microarray profiling in discordant MZ twins.

We undertook a detailed behavioral examination of two 49 year old female monozygotic twins, one of whom works as a war journalist and the other as an office manager in a law firm. They spent a close childhood in each other’s company. Their parents dressed them the same, ensuring they were essentially indistinguishable. They were bright students, but differed in their favorite subjects. One twin (who will later be referred to as the “war” twin) enjoyed languages and disliked domestic science, while the reverse applied to her twin (the “law” twin). At 17 years of age, the “war” twin left home setting in place a peripatetic existence. In the process she learned multiple languages. She eventually chose journalism as a career, gravitating to war zones where in a long and distinguished career she covered wars in Africa, the Middle East and the Balkans. Over the course of 20 years she was exposed to many life threatening situations, saw many people killed and wounded, and lost close colleagues. She married in her forties to a cameraman who also worked in war zones. She never had children and by her own admission never had the
maternal urge. She occasionally drinks in excess of nine units of alcohol per week, considered the upper limit of healthy drinking in a woman [215]. She does not smoke.

Her co-twin was bereft when her sister left home. She too thought of traveling but her choice of venue was more cautious, limited as it was to a single English speaking country. She married young, to a lawyer and soon had two children. She works part-time as a manager in a law office. She drinks three to four units of alcohol per week and does not smoke.

Neither twin has a history of psychiatric problems. Despite their geographical separation they remain close emotionally and meet as often as they can.

Methods

The twins underwent psychometric, genetic, and epigenetic testing.

Psychometric assessment

1. The Wechsler Abbreviated Scale of Intelligence (WASI)[216], a shortened form of the full Wechsler Adult Intelligence Scale (WAIS-III) gave an IQ score.

2. The Minnesota Multiphasic Personality Inventory (MMPI-II)[217] provided an index of personality attributes. The MMPI-2 is the leading commercially available clinical test of personality and psychopathology. It is composed of 567 true/false items that comprise 10 scales assessing clinical syndromes including anxiety, depression, and psychosis and personality traits such as coping style and patterns of interpersonal relationships. There are also scales for assessing the validity of the person's overall approach to the test. Numerous supplemental content scales are also available. Interpretation is based on analysis of the profile of scale elevations to provide an analysis of psychopathology and personality style.

3. The Toronto Gambling task (Floden and Stuss, 2004) is a computerized test that assesses the role of risk taking and impulsivity in decision making processes. On each trial, subjects are presented with a series of five gambles where the probability of ‘winning’ points systematically increases or, on half the trials, decreases. Subjects are free to select the gamble they prefer in order to win as many points as they can. However, gambles with a higher probability of ‘winning’ have a lower associated point value while gambles with a low probability of winning are linked with larger payoffs. Subjects with a risk-taking decision style consistently choose
gambles with low probabilities of obtaining a larger reward. This is differentiated from disinhibited responses through comparison of the presentation orders. Subjects with an impulsive or disinhibited decision style choose whichever type of gamble is presented first. Thus, they choose low probability gambles when they are presented in order of increasing probability (i.e., low to high) and choose high probability gambles when they are presented in decreasing order. Details of the full procedure are described in Figure 3.1.

The performances of the twins were compared to those from a group of 11 healthy controls matched for age and IQ. The control data had been published previously (Floden and Stuss2004).

4. The 28-item General Health Questionnaire [218] is a self report scale that measures psychological distress. It contains four subscales of 7 questions each that measure somatic symptoms, anxiety, social dysfunction and depression. The 4 point rating scale is scored 0-0-1-1 for each question, giving a range of scores from 0 through 28. By convention, scores $\geq 5$ are considered indicative of psychological distress.
Figure 3.1. The Toronto Gambling Task displays and contingencies. A schematic diagram of successive displays on one trial of the gambling task.

In the Low-High condition, presentation order moves from bottom to top – from low probabilities/large rewards to high probabilities/small rewards. In the High-Low condition, presentation order moves from top to bottom. Each display is present for 2 seconds or until the subject makes a response and the trial is terminated. Subjects complete 40 trials. Each trial is then followed by a blank screen for 2s (first 20 trials) or 10s (last 20 trials). Subjects initially received five practice trials of each presentation order to ensure comprehension of instructions.

In both conditions, up to five cards are presented face-down in a horizontal array. One of the five cards displays the word ‘WIN’ on its face whereas the other four cards are blank. Subjects are instructed that they can touch the screen at any time to select the gamble and “turn over” the cards present. If the WIN card is among the cards present at the response, the subject earns points. If the WIN card is absent at the response, no points are awarded. The position of the WIN card is random on each trial, meaning that the more cards on the screen, the higher the probability that the WIN card will be present. However, the point value of the WIN card is inversely related to the number of cards on the screen (i.e., more cards/higher probability = fewer points). The likelihood of finding the WIN card and its associated value are displayed on the screen at all times during the trial to minimize memory demands. Note that, as the probability of finding the target card increases, the point value decreases. Contingencies are shown adjacent to
each display ($P_{win} =$ the probability of finding the target card, $EV =$ expected value of the gamble across multiple trials). Unbeknownst to subjects, the first and last gambling options presented are slightly inferior to the middle options, which do not differ from each other in terms of expected value.

**Zygosity testing**

Genomic DNA was sent to a genetic testing company, Proactive Genetics Incorporated (http://www.proactivegenetics.com/), for zygosity testing.

**Epigenetic testing**

Epigenetic testing by microarray-based DNA methylation profiling was performed using the protocol described in by Schumacher et al [78]. Briefly, differences between twins were investigated by interrogating the enriched unmethylated fractions of total genomic DNA from the co-twins. Genomic DNA was extracted from peripheral blood cells using standard phenol chloroform techniques. Enzymatic digestion was performed with DNA methylation sensitive restriction enzyme HpaII (restriction site: CCGG). After digestion, DNA adaptors were ligated to the restriction fragments, which was followed by polymerase chain reaction (PCR) amplification using primers that were complementary to the adaptors. PCR conditions were adjusted in such a way that only fragments that were less than 1kb (i.e. short, digested, and therefore unmethylated) will amplify preferentially. The unmethylated fraction of genomic DNA is then end-labeled with Cy3 and Cy5 (GE Healthcare) dyes and subjected to hybridization at 42°C for 16 hours. All samples were interrogated on the human CpG island microarray consisting of 12,192 clones representative of numerous CpG island regulatory elements across the genome [72].

Microarray experiments were performed using a balanced block design. Separate enrichments of four genomic DNA aliquots per twin were produced in order to create a total of eight twin vs. co-twin hybridizations (biological variance group) that would stringently control for experimental variability. Hybridization signals in the biological variance group could be compared to seven self-self hybridizations (technical variance group) to determine whether epigenetic differences between the twins were detectable above the technical variance.

Epigenetic microarray profiling was performed on four dye swapped technical replicates on 9 pairs of normal MZ twins obtained from the Queensland Institute of Medical Research, Brisbane,
Australia in order to assess levels of DNA methylation variation in an alternate population of identical twins with no known phenotypic discordance.

Microarrays were scanned on an Axon 4000b scanner and analyzed using Genepix6.0 Software. Subsequent GPR files were subjected to ratio and print tip loess based normalization. Microarray data were trimmed on the microarray feature annotation, removing mitochondrial genes, translocation hot spots, and repetitive genomic regions. Fold change data as determined by log transformed loess M ratios were compared using a one way t test and subjected to correction for multiple testing by a Benjamini-Hochberg False Discovery Rate (FDR) test, a standard for microarray analysis. Gene IDs within 1 kb proximal to CpG islands were obtained from the microarray annotation data and cross referenced with the April 2007 build of the Gene Ontology Database (www.geneontology.org) to obtain gene ontology (GO) categories associated with each microarray locus. The average fold change for one identified locus (DLX1) was compared to the variation exhibited by the separate set of microarray data generated using the 9 sets of control MZ twins. A z-test was used to compare the absolute value of fold change to the distribution of absolute differences of the normal control MZ twins at this locus.

The spot wise standard deviation of co-twin DNA methylation difference across the 9 non-discordant MZ twin pairs was used to assess the biological DNA methylation variation per locus. The distribution of spot wise standard deviations was calculated from the mean values of 1, 2, and 4 dye swapped technical replicates per twin pair separately, to assess the influence of the number of technical replicates on biological variance. Subsequently, a power analysis was performed in R in a spot wise manner for each SD distribution to detect the proportion of loci per number of twin pairs (N) that would have 80% power to detect a range of fold changes (1.15, 1.2, 1.6) at an $\alpha$ level of 0.001 used to control the family-wise error rate (FWER) resultant from multiple testing. A more conservative Bonferroni corrected $\alpha$ level of $4.1 \times 10^{-6}$ was also used for power analysis for the 1.2 fold change. A fold change of 1.15 corresponds to the observed fold change threshold for technical variance (Figure 3.4) while the 1.6 fold change represents the maximum fold change observed of any FDR significant loci. The fold change of 1.2 was of particular interest as it corresponds to the observed fold change of an identified locus proximal to the distaless homeobox gene 1 (DLX1), which we speculate as having functional relevance to phenotypic differences in the “war”/”law” twin pair.
Results

Psychometric assessment

Based on the WASI, the twins had similar overall IQ’s (114 and 115; high average range) and both scored zero on the GHQ. However, differences emerged on the MMPI where the “war” twin’s personality profile was within normal limits, while her twin’s responses revealed a tendency to overreact to minor problems with anxiety and somatic (physical) symptoms.

The twins’ MMPI-2 profiles are presented in Figure 3.2. Analysis of the validity scales (L and K elevated) indicates that both twins approached the MMPI – 2 in a defensive manner, presenting themselves in a favorable light. Although these scores likely underestimate symptoms, this does not affect comparisons between the two twins as their validity score patterns were similar. If anything, the significant clinical scale elevations noted for the “law” twin over the “war” twin may be underestimated as the “law” twin was slightly more defensive.
Figure 3.2. MMPI-2 scores for the “law” twin (solid line, circles) and the “war” twin (dashed line, squares).

K-Corrected MMPI-2 scores are presented in the standard manner, with the validity scales (L, F, K) followed by the clinical scales.

The “law” twin’s profile of clinical scales included significant elevations on scales 3 (hysteria) and 7 (psychosomatic) a statistically infrequent profile found in less than 1% of the normative sample of women. This profile reflects a high degree of anxiety, tension, and discomfort. Individuals with this pattern tend to be agitated, lacking in self confidence, perfectionistic, and introverted. They often over-react to problems. The “law” twin’s profile also reflected a lack of insight into the psychological origins of her anxiety.

The “war” twin’s profile was normal; none of the clinical scales were elevated. It is notable that the “war” twin produced a relatively high score on scale 5, indicating rejection of traditional female roles, whereas her co-twin produced a relatively low score on this same scale, reflecting more traditionally feminine interests.

Additional sibling differences were present on the Toronto Gambling Task (Figure 3.3). The “war” twin preferred high risk gambles where there was a low probability of obtaining a high reward. In contrast, her sister showed risk-averse preferences, generally selecting conservative gambles. Both sisters were relatively extreme in comparison to a normal control group of comparable age (mean = 50.6, SD = 14.2) and estimated IQ(mean = 115.4, SD = 7.0).
Figure 3.3. Gambling performance for the twins as well as control subjects (n = 11)

Graph depicts the average gamble selection in both the Low-High (increasing probability) and High-Low (decreasing probability) presentation orders. Higher probabilities reflect conservative play whereas lower probabilities reflect riskier play. The “law” twin performed similar to other controls in the High-Low presentation order but was significantly more conservative in the Low-High presentation order (one-sample t(10) = 6.1, p < .001). In contrast, the war twin showed significantly more risk-taking than controls, regardless of presentation order (Low-High order, one-sample t(10) = 6.2, p < .001; High-Low order, one-sample t(10) = 6.9, p < .001).

Genetics

Monozygosity of the “war” and “law” twins was confirmed by Proactive Genetics Incorporated through genotyping of DNA markers D5S818, D13S317, D7S820, D16S539, vWA, TH01, TPOX, and CSF1PO.
The spotwise fold change exhibited by DNA vs. self hybridizations in the technical variance group was compared to that in the twin vs. co-twin hybridizations in the biological variance group across 12,148 microarray features, which excludes control and blank spots. The number of loci in the technical variance group with a mean fold change above a threshold of 1.15 was 1496, 591, 309, 258, and 160 for the spotwise average of 3, 4, 5, 6, 7 arrays, respectively. The number of loci in the biological variance group with mean fold change above this threshold was 1020, 644, 418, 486, 499, and 412 loci for 3-8 arrays, respectively. Above 4 technical replicates, the biological variance group consistently had 1.5% more loci above this threshold, on average (Figure 3.4).

The application of correction for multiple testing using FDR identified 38 loci that consistently exhibit statistically significant differences in DNA methylation levels in the biological variance group. No spots in the technical variance group survived correction for multiple testing, which was the case independent of the number of self-self hybridizations investigated. Only one locus from the technical variance group had a fold change greater than 1.15 ($\log_2(1.15)=0.2$) beyond a $P$ value of <0.001 before correction (Figure 3.4), and therefore an effect size of 1.15 fold change stands as a conservative experimental variance threshold for significantly different loci between twins.
Figure 3.4. Relative DNA methylation profiles of the “war” twin vs. “law” twin

A. Volcano plot of \( \log_2 \) transformed ratio information for each microarray element (X axis) vs. the P value of a paired t-test (Y axis) comparing DNA methylation differences in the co-twins. Data produced from the biological variance group is in green while technical variance data is in red. Data values with an FDR corrected P value below 0.05 are in blue including DLX1, represented by a blue triangle outlined in black.

Gene ontology classification of the most significant genes within 10 kb of the CpG islands revealed 23% involved in transcriptional regulation as defined by the April 2007 build of the Gene Ontology Database, including one with potential behavioral implications, namely DLX1 with a P value of 0.047 after correction for multiple testing with FDR. The fold change ratio for this locus was statistically higher than the variation observed in an alternative set of MZ twins (\( Z=3.5, P<0.0004 \)).

We also attempted to determine the proportion of the loci on the microarray in total that exhibit >80% power as a function of sample size. Additionally, we wanted to assess the power of the
method as a function of the number of technical replicates per twin pair performed. As expected, we observed that a larger number of technical replicates results in a smaller spot wise variance. Figure 3.5 depicts the proportion of loci achieving >80% power per sample size as a function of the SD distribution produced from the average of 1, 2, and 4 technical replicates per twin pair for an effect size of 1.2 fold change. Similar differences were observed for the other effect sizes (data not shown). In general, dye swapping is useful to eliminate spots resultant from dye bias and cross hybridization and so subsequent power analysis was performed with the spot wise SD distribution produced from two dye swapped technical replicates in order to represent the most economical yet conservative estimation of sample size requirements. The power analysis prognosticates that with 21, 14, and 6 twin pairs, >95% of the loci will have an 80% chance of detecting a true DNA methylation difference of 1.15, 1.2, and 1.6 fold between groups, respectively (Figure 3.6). When applying Bonferroni correction, $\alpha$ is reduced to 4.1 x10^-6 for a very conservative estimate of the false positive rate. The subsequent sample size in this case for an effect size of 1.2 to achieve >95% of loci with >80% power is 25 twin pairs. These power analyses assume that the biological variance in a group of twin pairs with a discordant phenotype will be similar to that of normal pairs.
Figure 3.5. Power vs. technical replicate hybridization number and sample size (N)

A plot representing the different proportions of loci on the microarray that will achieve >80% power per sample size when the log fold change across the 9 non-discordant twin pairs used to create a spot wise standard deviation distribution are resultant from an average of 1, 2, and 4 technical replicates. In all cases, the measured effect size was a fold change of 1.2.
Discussion

A number of consistent strands run through the twin’s phenotypes. The defining characteristics of the war twin’s life - dangerous career, married late to a man exposed to similar grave dangers, no children, drinks more alcohol than medically recommended, displays minimal anxiety and adopts high risk strategies on certain cognitive paradigms - fits well with a demographic and behavioral profile characteristic of war journalists in particular [219] and high sensation seekers in general [220]. Her twin, by contrast is the mirror opposite of all these factors and behavioral traits.
Of note was the finding that the “war” twin’s divergent patterns of responses on the Toronto Gambling task were high risk in keeping with a career choice that included working in zones of conflict. Yet despite the many life threatening events that she had had to confront, her scores on the GHQ did not reveal emotional distress, a result that matched her healthy personality profile on the MMPI-2. In contrast, her sister who had chosen the more predictable and safer work environment, while also showing no current symptoms of psychological distress on the GHQ, had an MMPI-2 result that revealed a propensity to develop anxiety and somatic complaints when confronting stress. Moreover, her gambling task performance was significantly more risk-averse than controls. Thus, in both twins it is possible to see a connection between career choices and psychometric characteristics.

While monozygotic twins are genetically identical at the DNA sequence level, a divergence of DNA methylation profiles occurring during development and over time could lead to dissimilar phenotypes. The Human CpG island microarray used to interrogate epigenetic variation between these twins does so at 12,192 loci representative primarily of CpG island regulatory elements. One of the microarray clones identified as differentially methylated between twins was located on chromosome 2q31.1 within the 3’ UTR of the distal-less Homeobox 1 gene (DLX1). When the mean fold change for DLX1 was compared to the general twin vs. co-twin standard deviation of this locus in an alternative set of 9 MZ twins without known discordant phenotypes, the differences identified between this study’s twin pair were significantly higher (P<0.0004). Epigenetic profiling was carried out in an identical manner in both cases, suggesting that large twin vs. co-twin DNA methylation differences at DLX1 in this twin pair have the potential to underlie the psychometric discordance measured.

The DLX1 gene encodes a transcription factor involved in the formation and maintenance of a distinct set of GABAergic interneurons [221, 222]. DLX1 derived neurons express neuropeptide Y (NPY), a peptide hormone that interacts with the hypothalamic pituitary adrenal axis (HPA), more commonly known as the stress center of the brain [222]. DLX1 expression is critical for NPY production as DLX1 knockout mice show a progressive loss of NPY most likely as a result of interneuronal loss [222]. Numerous animal model studies implicate NPY in modulating stress and anxiety as both NYP receptor antagonists and knockouts produce an anxiolytic effect [223]. This is likely because the stress response produced by the HPA is inhibited by GABA [224];
however, NPY release in turn inhibits the effects of GABA, effectively exciting stress response [225].

DNA methylation in the DLX1 3’UTR is likely to have important consequences in the regulation of this gene as, like many other homeobox containing genes, the region encoding DLX1 also codes for an overlapping antisense transcript [226, 227], which can modulate gene expression in varying ways [228]. Like any transcribed region, antisense transcripts can be modulated by DNA methylation status, as exemplified by the epigenetic control of the imprinted KvDMR antisense transcript, KCNQ1OT1. While we did not study the direct mechanisms by which the DLX1 3’ UTR methylation modulates gene expression, it is clear that the gene regulatory machinery sensitive to such methylation is in place at the DLX locus and could have implications for the downstream developmental pathways mediated by this gene. This could result in a reduced overall level of anxiety in the “war” twin as compared to the “law” twin, which is consistent with the risk taking behavior differences observed. While, peripheral blood should be relatively robust to environmental influence, it remains possible that the observed DNA methylation difference is a downstream effect of the different environments and lifestyles of these two individuals.

To our knowledge, this study represents the first use of microarray-based technology to identify epigenetic differences between identical twins. The technical variance group hybridizations highlight a biological detection threshold of a fold change of 1.15 for differences indicative of true DNA methylation differences. DLX1 represented the most significant identified epigenetic difference that appeared to have functional relevance to the measured phenotypic differences between these twins. While this locus was beyond the threshold of technical variance and thus represents a true methylation difference between twins, one set of MZ twins is certainly insufficient to make claims that a seemingly relevant epigenetic difference actually accounts for the observed phenotype. Therefore, we used DLX1 as an example of a detectable biological difference in order to estimate the required sample size of twin pairs with the same discordant phenotype that would be necessary to identify functionally relevant epigenetic changes of this effect size. The results demonstrate that, pari passo, were this twin set among a larger sample population of ~15-25 discordant twins with similar behavioral phenotypes, the technique would have 80% power to identify etiologically significant epigenetic differences.

One limitation in the interpretation of the power analysis is that there is a 36 year age difference between the war/law twin pair and the mean age of the cohort of MZ twins used to represent the
population DNA methylation variance. The HPLC-based analyses of density of methylated cytosines revealed a consistent age-dependent decrease of global methylation levels in human tissues [229]. A more recent study compared MZ co-twin DNA methylation differences in a young and aged cohort of twins and identified higher levels of DNA methylation variability in the aged cohort (Fraga et al. 2005). The authors hypothesized that DNA methylation patterns may drift over time, causing older MZ twins to be more epigenetically dissimilar than younger ones. Alternatively, a recent study by Heijmans et al. investigated DNA methylation variation at two imprinting control regions in both a young and aged cohort of MZ co-twins and determined that the observed variation was primarily the result of a heritable influence and that age attributed no effect (Heijmans et al. 2007). Of course, these disparate results could be attributed to differences in the specific genomic regions and twin populations investigated and to date, there is not a definitive answer regarding DNA methylation variation with age. However, it is important to note that such an epigenetic drift, as it was referred to by Fraga et al., could result in a larger population DNA methylation variance and would, in effect, decrease the power of the microarray technology used in this study in older twin cohorts.

Another important consideration when evaluating the informativeness of epigenetic markers for a given trait is that there could be tissue specific differences between the changes identified in peripheral blood and brain tissue where the gene products of these markers are known to function. That being said, it remains possible that for particular loci, early developmental epigenetic variation occurring prior to major tissue differentiation may be robust to change over time and could be measurable in peripheral tissues.

Researchers interested in performing epigenomic microarray profiling of this kind on discordant twin populations should also carefully consider the cost vs. power of performing technical replicate microarrays. Our experiments demonstrated that dye swapping was critical to eliminate false positive findings resultant from dye bias (data not shown). Therefore, performing two technical replicates per twin pair comparison is recommended. Beyond this, while averaging the values of more technical replicates reduces the spot wise standard deviation and thus increases power, it appears from Figure 3.5 that the relative power increase is not sufficient to justify the cost. For example, to detect a fold change of 1.2 with 80% power on >95% of microarray loci, an N of 14 and 12 twin pairs would be required for experiments performed with 2 and 4 dye swapped technical replicates, respectively. The total number of microarray hybridizations
required to achieve the same power in each case would therefore be 28 and 48, respectively, and thus it is obvious that performing 2 dye swapped technical replicates is sufficient to detect epigenetic differences without undue cost and effort.

Epigenetic markers cannot by themselves account for why one twin chose war journalism as a career while her co-twin opted for safer environs, for such complex decisions will always transcend genetic determinism. However, what the epigenetics may explain is the propensity of one twin to function well and without undue anxiety in highly dangerous situations. In turn, this trait may have influenced a career choice. Whether the same markers are to be found in others pursuing different hazardous occupations is not known, but these results suggest new avenues for research elucidating responses to danger and the ability of some to function well when confronted by risk.
Chapter 4

DNA Methylation Profiles in Monozygotic and Dizygotic Twins

Kaminsky ZA\textsuperscript{1,6}, Tang T\textsuperscript{1}, Wang SC\textsuperscript{1,8}, Ptak C\textsuperscript{1,6}, Oh G\textsuperscript{1,6}, Ziegler S\textsuperscript{1}, Wong A\textsuperscript{1,6}, Feldcamp LA\textsuperscript{1,6}, Virtanen C\textsuperscript{2}, Halfvarson J\textsuperscript{3,7}, Tysk C\textsuperscript{3,7}, McRae AF\textsuperscript{4}, Visscher PM\textsuperscript{4}, Montgomery GW\textsuperscript{4}, Gottesman II\textsuperscript{5}, Martin NG\textsuperscript{4}, Petronis A\textsuperscript{1,6\*}.

Originally Published in \textit{Nature Genetics}

1. Centre for Addiction and Mental Health, Toronto, Ontario, Canada
2. University Health Network Microarray Centre, Toronto, Ontario, Canada
3. Division of Gastroenterology, Department of Medicine, Örebro University Hospital, Örebro, Sweden
4. Queensland Institute of Medical Research, Brisbane, Australia
5. University of Minnesota, Minneapolis, Minnesota, USA
6. University of Toronto, Toronto, Ontario, Canada
7. School of Health and Medical Sciences, Örebro University, Örebro, Sweden
8. Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan

*Correspondence to:
Art Petronis
The Krembil Family Epigenetics Laboratory
Centre for Addiction and Mental Health
250 College Street
Toronto, Ontario M5T 1R8
Canada
Ph: 416 535 8501 ext 4880
Email: arturas_petronis@camh.net

Contributions: For this manuscript, I performed all human microarray related laboratory experiments, experimental design, and wrote the paper. I wrote the data analysis code in R and Perl for a majority of results in the paper and performed the analyses. Carl Virtanen was instrumental in the creation of the karyograms. The Gene Ontology analysis script in Bioconductor was provided by Thomas Tang. Consultation with bioinformaticians Thomas Tang, Sun-Chong Wang, and Allan McRae helped to direct the analyses performed. For animal related microarrays, I was responsible for experimental design and data analysis, and performed half of the microarray hybridizations. The other half of the animal experiment was performed by Carolyn Ptak under my direction. Animal sacrifice was performed by Laura Feldcamp under the direction of Albert Wong. I selected genomic regions for validation, designed pyrosequencing assays, performed the sodium bisulfite modification, and tested select loci with pyrosequencing. Under my direction, a portion of pyrosequencing and cloned sequencing was performed by Carolyn Ptak, Gabriel Oh, and Sigrid Ziegler.
Abstract

Comparisons of phenotypic similarities and differences in monozygotic (MZ) and dizygotic (DZ) twins have provided the basis for molecular genetic and epidemiological studies in human diseases and normal traits[81, 82]. In the wake of increasing evidence that epigenetic factors can contribute to phenotypic outcomes, we performed a DNA methylome analysis of MZ twins in white blood cells (WBC, N=19 pairs), buccal epithelial cells (N=20 pairs), and gut (rectum) biopsies (N=18 pairs) as well as WBC and buccal epithelial cells from DZ twins (N=20 pairs of each tissue) using 12K CpG island microarrays[72, 78]. The DNA methylation differences we detected provide the basis for the first annotation of epigenetic metastability of ~6,000 unique genomic regions in MZ twins. An intraclass correlation (ICC)-based comparison of matched MZ and DZ twins revealed that DZ co-twins exhibited significantly higher epigenetic difference over MZ co-twins in buccal cells (p=1.2x10^{-294}). While such higher epigenetic discordance in DZ twins can result from DNA sequence differences, our in silico SNP analyses and comparison of methylomes in inbred vs. outbred mice favour the hypothesis that this is due to epigenomic differences in the zygotes. This study suggests that molecular mechanisms of heritability may not be limited to DNA sequence differences.

Introduction

Twin research has been of fundamental importance in human studies for two main reasons. First, comparison of phenotypic concordance rates in MZ twins vs. DZ twins is a powerful strategy to estimate heritability. Second, phenotypic discordance in MZ co-twins has traditionally indicated roles for environmental factors. Countless twin studies have been performed over the last century on almost any trait imaginable but primarily on human disease[82]. Nearly universally, MZ twins exhibit various degrees of discordance, generally lower in comparison to discordance in DZ twins. These observations provided the basis for the current paradigm of human normal and morbid biology, which focuses on DNA sequence variation and environmental differences. In the last decade, evidence has been accumulating that epigenetic modifications of DNA and histones can play a primary role in phenotypic outcomes, including human disease[230]. DNA methylation exhibits only partial stability, which could be caused by a wide variety of factors, including developmental programs, environment, hormones, and stochastic events[231-234]. Such epigenetic metastability may result in substantial epigenetic differences across genetically
identical organisms[80]. Several studies have identified epigenetic differences, either at selected genes of MZ twins[117-119, 209] or the overall epigenome[121]. Despite this promising start, no locus-specific epigenome-wide studies have yet been performed to catalogue the extent of this phenomenon, and few have been performed in tissues other than peripheral blood cells.

Results and Discussion

In this study, we mapped DNA methylation differences in three types of tissues from MZ twins: white blood cells (WBC) (N=19 pairs), buccal epithelial cells (N=20 pairs), and gut (rectum) biopsies (N=18 pairs), by interrogation of the unmethylated genome on the 12K CpG island microarray[72]. We first ensured that the microarray technology identifies actual DNA methylation differences between MZ co-twins rather than artefactual differences due to technical variation. For this, 4 parallel enrichments of the unmethylated fraction of genomic WBC DNA were performed from the DNA stock of the same individual. DNA samples from 8 MZ twins (4 pairs) were compared against themselves (measuring technical variation) or the respective co-twin (measuring biological variation). The biological variation significantly exceeded the technical variation in all cases (P=1.4x10^{-238}, P=1.1x10^{-202}, P=2.1x10^{-7}, P=2.6 x10^{-39}) indicating that the detected MZ co-twin differences are genuine (Fig. 4.1). The technical variance ($\sigma^2$) was consistent between all self-self hybridizations, while the biological variance produced by the twin vs. co-twin hybridizations varied significantly between twin pairs (Fig. 4.1). For all MZ twins per tissue cohort, the mean absolute log fold change between MZ co-twins was significantly larger than that between technical replicates (WBC mean difference=0.013 ± 4.5.6x10^{-4}, P=3.6x10^{-173}, buccal mean difference=0.017 ± 6.8x10^{-4}, P=4.9x10^{-132}, gut mean difference=0.0053 ± 5.6x10^{-4}, P=1.02x10^{-14}), signifying biological variation was detectably higher than technical variation in all tissues. Furthermore, microarray validation performed by sodium bisulfite sequencing and pyrosequencing (Fig. 4.2 and Fig. 4.3) indicates that the microarray signals detected reflect the actual DNA methylation status in the tested samples. For WBC-based analyses, we also performed a spot wise correlation between cell sub-fraction counts and confirmed that differences observed in WBC samples were not resultant from cell sub-fraction differences.
Figure 4.1. Biological vs. technical variation

Volcano plots of 4 MZ twin vs. co-twin WBC DNA methylome comparisons (black) overlayed with 4 matched twin DNA vs. self comparisons (green) for each set of MZ twins. The x-axis represents the mean fold change across the 4 replicas. The y-axis represents the $-\log_{10}$ of the p value from a paired t test. Higher significance denotes a higher consistency between replicates. Significant variation in the spread of detected biological difference exists between twin pairs (Kruskal-Wallis $\chi^2 = 16.3$, df = 3, P=0.001) with a symmetrical large (A and B), symmetrical small (C), and asymmetrical (D) variation of the DNA methylome between co-twins. For each twin pair, a non-parametric Ansari-Bradley test demonstrated that levels of variance ($\sigma^2$) in the MZ twin - co-twin comparison were significantly larger than $\sigma^2$ in the self-self comparisons (twin set A: variance ratio= 2.91, P=1.4x10$^{-238}$; set B: 2.14, P=1.1x10$^{-202}$; set C: 1.12, P=2.1x10$^{-7}$; set D: 2.63, P=2.6 x10$^{-59}$). Levels of technical variation were not significantly different between groups (Kruskal-Wallis $\chi^2 = 1.81$, df = 3, P= 0.62).

Figure 4.2. Correlations between microarray and sodium bisulfite sequencing data

Microarray data were validated by sodium bisulfite modification based mapping of methylated cytosines in 18 CpGs at a locus displaying a range of co-twin variability, UHNhscpg0003195, which maps to the 3’ end of C1QTNF8. Over 1,300 clones representing 18 MZ twin pairs (36 clones on average per individual) were sequenced. Twin differences in the density of methylated cytosines revealed by bisulfite sequencing (x-axis) correlated significantly with the $\log_2$ DNA methylation differences produced by the microarray data (y-axis) (mean density across all 18...
CpGs \( R=0.65, P=0.0036 \) (A). Similarly, the density of methylated cytosines in the HpaII restriction site at the 9\(^{th}\) CpG position correlated significantly with the microarray data \( (R=0.58, P=0.01) \) (B). In both A and B, the x-axis values represent the sodium bisulfite based co-twin DNA methylation difference and y-axis values represent the \( \log_2 \) fold difference between co-twins generated by microarray data.

Figure 4.3. Pyrosequencing correlations as a function of distance

Bisulfite pyrosequencing of the total amplicon without cloning was performed at 5 loci showing a range of co-twin variation: UHNhscp0008483 (15 pairs of twins, \( \text{CA2} \) gene), UHNhscp0004390 (10 pairs, \( \text{RAX} \) gene), UHNhscp0004556 (19 pairs, \( \text{IL1A} \) gene), UHNhscp0000193 (18 pairs, \( \text{RNF110} \) gene), and UHNhscp0004262 (11 pairs, \( \text{DLX1} \) gene), in WBC DNA samples, which also positively correlated with the microarray data. A bar graph displaying the strength of correlation between \( \log_2 \) DNA methylation difference between co-twins in the bisulfite pyrosequenced loci compared to that of the microarray data. Correlations between the microarray data and HpaII position only are depicted in Red, while blue represents the correlation derived from the average methylation density over 5, 6, 4, 7 and 3 CpGs, respectively. Interrogated CpG sites located within the probe sequence (represented by a rectangle) showed the strongest correlation with microarray data. X-axis values (-141, -30, 221, 267, and 637) depict the position of the interrogated HpaII site relative to the 3’ end of each clone marked as zero. The y-axis depicts Pearson’s correlation \( (R) \) between microarray and pyrosequencing data. Similarly to Eckhardt et al.[235], we noticed that the strength of the correlation between microarray signal and bisulfite based-mapping of methylated cytosines located outside the probe sequence was a function of the distance of the interrogated CpG site from the probe.
In the microarray-based studies a large degree of MZ co-twin DNA methylation variation was detected in all tissues investigated. An intraclass correlation coefficient (ICC) measured MZ co-twin variation for each unique genomic region, where an ICC range from +1 to –1 denotes high to low epigenetic similarity between co-twins relative to the variation between unrelated pairs. For each tissue, we generated an ICC-based annotation of MZ co-twin DNA methylation variation across ~6,000 unique DNA loci (Fig. 4.4 – WBC; annotations for other tissues are available in Appendix 1 Fig A.1 and Fig A.2). Interestingly, DNA methylation profiles in the buccal epithelial cells from monochorionic MZ twins were significantly more variable within pairs than those from dichorionic MZ twins (median difference= 0.37 ± 0.0057, P<9.9x10^-324) (Fig. 4.5). Dichorionic MZ twins are believed to result from a splitting of the blastomere within the first four days following fertilization while monochorionic MZ twins arise after this point [236]. Chorionicity information was only available for the buccal and WBC samples; in WBC-based studies all MZ twins were dichorionic to avoid in utero twin blood transfusion effects. The varying degrees of epigenetic dissimilarity detected between dichorionic and monochorionic twins may reflect differences in epigenetic divergence among embryonic cells at the time the twin blastomeres separated.

Figure 4.4. Karyogram of MZ co-twin epigenetic similarity in WBCs
A chromosomal karyogram depicting levels of MZ co-twin similarity per interrogated locus in the WBC sample. Dark to light bars on the chromosomes represent chromosomal banding patterns as revealed by Giemsa staining, while red bars indicate regions of high microarray probe density. Bars to the right of each chromosome represent locus specific ICCs depicting levels of MZ co-twin epigenetic similarity. P values associated with the ICC statistic per locus were subjected to false discovery rate (FDR) correction for multiple testing. FDR corrected P values below the level of P<0.05 are depicted in green while those with greater P values are depicted in grey.

**Figure 4.5.** Raw binding intensities of MC and DC MZ twin hybridizations

Box plots of raw green (A) and red (B) signal intensities for 40 dichorionic (1) and 40 monochorionic (2) buccal MZ twin microarrays. Green and red center lines separate the two batches of samples. As monochorionic and dichorionic MZ twin buccal samples were performed in different batches, we wanted to evaluate if batch effects in sample
binding could be influencing this result. No batch effects are observed that could account for the significant differences in MZ co-twin epigenetic variation between dichorionic and monochorionic twins.

The spot-wise ICC values across the 5,919 loci that overlapped between data sets were compared between the buccal cells and WBC from the same set of DC MZ twins, and WBC samples from different individuals, by linear regression. A small but significant correlation was observed between WBC and buccal-derived ICCs from the same individuals ($R=0.046, P=4.08\times 10^{-4}$) but not between buccal cells and WBCs from unrelated individuals ($R=-0.0025, P=0.84$), which suggests that tissues in genetically identical individuals are more similar epigenetically as compared to those in unrelated individuals.

Using locus-specific DNA methylation information, we investigated whether the degree of co-twin epigenetic similarity is associated with functional genomic elements. In each tissue, we compared the distribution of ICCs of the CpG islands (CGIs) to that of all non-CGI loci. In an identical manner, the ICCs of promoters were compared to those of non-promoter loci in each tissue. In total, 6 tests were performed and p values were corrected for multiple testing using the Bonferroni method. Both CGIs and promoters were less epigenetically variable in WBC-derived (Wilcoxon Rank Sum test, mean$_{CGI}=0.43 \pm 0.0065$, mean$_{Non-CGI}=0.39 \pm 0.0053$, $P=1.5\times 10^{-4}$ and mean$_{Promoter}=0.43 \pm 0.0085$, mean$_{Non-Promoter}=0.4 \pm 0.0048$, $P=0.0077$; Bonferroni corrected $P=8.7 \times 10^{-4}$ and $P=0.047$, respectively). Promoters also showed a trend towards being less epigenetically variable in gut tissue (Wilcoxon Rank Sum test, mean$_{Promoter}=0.11 \pm 0.0065$, mean$_{Non-Promoter}=0.09 \pm 0.0037$, $P=0.057$; Bonferroni corrected $P=0.34$). No statistically significant differences in the degree of DNA methylation variation were detected in the buccal epithelial cells. The promoter and CGI probes were also subjected to the Gene Ontology (GO)-based analysis[237]. A majority of the identified GO categories associated with high ICCs (top 5$^{th}$ percentile), i.e. a high degree of co-twin DNA methylation similarity, had direct functional relevance to the tissue investigated; this was most striking in WBCs, where categories such as T cell proliferation (GO:004209) and activation of immune response (GO:000225) were identified. In buccal cells, the proteinaceous extracellular matrix (GO:0005578) and the metalloendopeptidase activity (GO:0004222) categories were identified; genes in these categories interact and are expressed in oral fibroblast cells[238, 239]. A portion of GO categories in gut appeared to be associated with regulation of cell proliferation (GO:0042127)
and epithelial to mesenchymal transition (GO:0001837), which is an intrinsic step of formation of the smooth muscle cells of the gut blood vessels[240]. Our observations are consistent with an earlier study[232] where the fidelity of CpG methylation patterns was twice as high in promoter as opposed to non-promoter regions. Taken together, greater epigenetic similarity between MZ co-twins at functionally important regions in comparison to the loci without clearly defined regulatory function suggests functional stratification of the epigenome.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>GO ID</th>
<th>Pvalue</th>
<th>OddsRatio</th>
<th>ExpCount</th>
<th>Count</th>
<th>Size</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBC</td>
<td>GO:0002274</td>
<td>0.0051</td>
<td>Inf</td>
<td>0.14341085</td>
<td>2</td>
<td>2</td>
<td>myeloid leukocyte activation</td>
</tr>
<tr>
<td>Promoters</td>
<td>GO:0042098</td>
<td>0.0277</td>
<td>13.277778</td>
<td>0.28682171</td>
<td>2</td>
<td>4</td>
<td>T cell proliferation</td>
</tr>
<tr>
<td></td>
<td>GO:0006909</td>
<td>0.044</td>
<td>8.8425926</td>
<td>0.35852713</td>
<td>2</td>
<td>5</td>
<td>phagocytosis</td>
</tr>
<tr>
<td></td>
<td>GO:0009615</td>
<td>0.044</td>
<td>8.8425926</td>
<td>0.35852713</td>
<td>2</td>
<td>5</td>
<td>response to virus</td>
</tr>
<tr>
<td></td>
<td>GO:0002253</td>
<td>0.044</td>
<td>8.8425926</td>
<td>0.35852713</td>
<td>2</td>
<td>5</td>
<td>activation of immune response</td>
</tr>
<tr>
<td>CGIs</td>
<td>GO:0030183</td>
<td>0.0314</td>
<td>12.325301</td>
<td>0.30630631</td>
<td>2</td>
<td>4</td>
<td>B cell differentiation</td>
</tr>
<tr>
<td></td>
<td>GO:0002253</td>
<td>0.0498</td>
<td>8.2088353</td>
<td>0.38288288</td>
<td>2</td>
<td>5</td>
<td>activation of immune response</td>
</tr>
<tr>
<td></td>
<td>GO:0002443</td>
<td>0.0498</td>
<td>8.2088353</td>
<td>0.38288288</td>
<td>2</td>
<td>5</td>
<td>leukocyte mediated immunity</td>
</tr>
<tr>
<td>Buccal</td>
<td>GO:0030518</td>
<td>0.0047</td>
<td>Inf</td>
<td>0.13835198</td>
<td>2</td>
<td>2</td>
<td>steroid hormone receptor signaling pathway</td>
</tr>
<tr>
<td>Promoters</td>
<td>GO:0019222</td>
<td>0.0066</td>
<td>2.0263338</td>
<td>15.8413021</td>
<td>25</td>
<td>229</td>
<td>regulation of metabolic process</td>
</tr>
<tr>
<td></td>
<td>GO:0006350</td>
<td>0.0183</td>
<td>1.8618926</td>
<td>14.4577823</td>
<td>22</td>
<td>209</td>
<td>transcription</td>
</tr>
<tr>
<td>Buccal</td>
<td>GO:0005578</td>
<td>0.0287</td>
<td>3.3724236</td>
<td>1.79654511</td>
<td>5</td>
<td>26</td>
<td>proteinaceous extracellular matrix</td>
</tr>
<tr>
<td>CGIs</td>
<td>GO:0004222</td>
<td>0.0324</td>
<td>5.3868243</td>
<td>0.73780488</td>
<td>3</td>
<td>11</td>
<td>metalloendopeptidase activity</td>
</tr>
<tr>
<td>Gut</td>
<td>GO:0042127</td>
<td>0.0365</td>
<td>3.8139535</td>
<td>1.31905465</td>
<td>4</td>
<td>19</td>
<td>regulation of cell proliferation</td>
</tr>
<tr>
<td>Promoters</td>
<td>GO:0001837</td>
<td>0.0138</td>
<td>27.388889</td>
<td>0.20923657</td>
<td>2</td>
<td>3</td>
<td>epithelial to mesenchymal transition</td>
</tr>
<tr>
<td></td>
<td>GO:0007179</td>
<td>0.0263</td>
<td>13.680556</td>
<td>0.27898209</td>
<td>2</td>
<td>4</td>
<td>transforming growth factor beta receptor signaling pathway</td>
</tr>
</tbody>
</table>

Table 4.1. GO analysis of loci with high MZ co-twin epigenetic similarity
Significantly over represented gene ontology categories in the positive 5th percentile of the ICC distribution of promoter and CGI associated loci in each tissue cohort.

In the bottom 5th percentile of ICC values representing the most epigenetically variable loci between MZ co-twins, GO categories in all tissues appeared to be associated with processes involved in cell division. Examples of significantly over-represented GO categories are regulation of progression through cell cycle (GO:0000002) in WBC, mitosis (GO:0000708) in buccal cells, and cell-cell adhesion (GO:0016337) in gut. One possible explanation for consistent epigenetic differences between MZ co-twins is that these differences are somehow related to the twinning process. One theory of MZ twinning is that populations of cells within the blastocyst recognize each other as different, forcing a separation of the inner cell mass[236]. Detected epigenetic differences with negative ICCs may therefore reflect an early developmental epigenetic discordance that increases the probability of twin formation in the first place.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>GO ID</th>
<th>Pval</th>
<th>OddsRatio</th>
<th>ExpCount</th>
<th>Count</th>
<th>Size</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBC</td>
<td>GO:0000074</td>
<td>0.0032</td>
<td>3.5637066</td>
<td>3.20930233</td>
<td>9</td>
<td>46</td>
<td>regulation of progression through cell cycle</td>
</tr>
<tr>
<td>Promoters</td>
<td>GO:0022402</td>
<td>0.019</td>
<td>2.4193548</td>
<td>4.88372093</td>
<td>10</td>
<td>70</td>
<td>cell cycle process</td>
</tr>
<tr>
<td>WBC</td>
<td>GO:0019882</td>
<td>0.0435</td>
<td>8.9004329</td>
<td>0.35585586</td>
<td>2</td>
<td>5</td>
<td>antigen processing and presentation</td>
</tr>
<tr>
<td>CGIs</td>
<td>GO:0045786</td>
<td>0.048</td>
<td>3.3833333</td>
<td>1.42342342</td>
<td>4</td>
<td>20</td>
<td>negative regulation of progression through cell cycle</td>
</tr>
<tr>
<td>Buccal</td>
<td>GO:0000279</td>
<td>0.0075</td>
<td>3.6943284</td>
<td>2.40895219</td>
<td>7</td>
<td>32</td>
<td>M phase</td>
</tr>
<tr>
<td>Promoters</td>
<td>GO:0007067</td>
<td>0.0398</td>
<td>3.0641822</td>
<td>1.95727365</td>
<td>5</td>
<td>26</td>
<td>mitosis</td>
</tr>
<tr>
<td></td>
<td>GO:0000776</td>
<td>0.0276</td>
<td>1.3308824</td>
<td>0.28659161</td>
<td>2</td>
<td>4</td>
<td>kinetochore</td>
</tr>
<tr>
<td></td>
<td>GO:0005876</td>
<td>0.0276</td>
<td>1.3308824</td>
<td>0.28659161</td>
<td>2</td>
<td>4</td>
<td>spindle microtubule</td>
</tr>
<tr>
<td></td>
<td>GO:0005768</td>
<td>0.0383</td>
<td>5.0317164</td>
<td>0.78812692</td>
<td>3</td>
<td>11</td>
<td>endosome</td>
</tr>
<tr>
<td></td>
<td>GO:0009922</td>
<td>0.0439</td>
<td>8.8627451</td>
<td>0.35823951</td>
<td>2</td>
<td>5</td>
<td>spindle pole</td>
</tr>
<tr>
<td>Buccal</td>
<td>GO:0051656</td>
<td>0.0143</td>
<td>26.767123</td>
<td>0.21367521</td>
<td>2</td>
<td>3</td>
<td>establishment of organelle localization</td>
</tr>
<tr>
<td>CGIs</td>
<td>GO:0006732</td>
<td>0.0151</td>
<td>5.4577778</td>
<td>1.03692762</td>
<td>4</td>
<td>13</td>
<td>coenzyme metabolic process</td>
</tr>
<tr>
<td>Promoters</td>
<td>GO:0019867</td>
<td>0.0313</td>
<td>12.489796</td>
<td>0.30676692</td>
<td>2</td>
<td>4</td>
<td>outer membrane</td>
</tr>
</tbody>
</table>
Table 4.2. GO analysis of loci with low MZ co-twin epigenetic similarity

| GO:0051186 | 0.0065 | 4.4351852 | 1.80395853 | 6 | 22 | cofactor metabolic process |
| CGIs: GO:0022610 | 0.0189 | 2.5604396 | 4.18190386 | 9 | 51 | biological adhesion |
| CGIs: GO:0016337 | 0.0285 | 3.4325681 | 1.80395853 | 5 | 22 | cell-cell adhesion |

Significantly over represented gene ontology categories in the negative 5th percentile of the ICC distribution of promoter and CGI associated loci in each tissue cohort.

Cases of DNA sequence variation in MZ twins have been documented[241], but these are uncommon and unlikely to account for even a fraction of the MZ co-twin differences identified in our experiments. Further epigenetic twin studies may include a more detailed annotation of epigenetic differences in MZ co-twins, a search for disease specific epigenetic changes in discordant MZ twins, and a dissection of environment-induced vs. stochastic epigenetic differences. Since MZ twins reared apart are generally quite similar to MZ twins reared together[108], we speculate that stochastic events in epigenetically determined phenotypic differences in MZ co-twins are much more important than environment.

While the first part of the twin study focused on epigenetic differences in MZ twins, we were also interested in comparisons of epigenetic similarities in MZ vs. DZ twins, the same design that has been used in heritability studies. Buccal epithelial cells from 20 sets of MZ co-twins (used in the first part of the study) exhibited a significantly lower degree of co-twin DNA methylation difference in comparison to 20 sets of DZ co-twins matched for age and sex (ICC\textsubscript{MZ} - ICC\textsubscript{DZ} = 0.15 \pm 0.0039, P=1.2x10^{-194}) (Fig. 4.6A). This result was the influence of the 10 sets of dichorionic MZ twins only(ICC\textsubscript{MZ} - ICC\textsubscript{DZ} = 0.35 \pm 0.0057, P<9.9x10^{-324}) (Fig. 4.6B) while the mean ICC of monochorionic MZ twins was close to 0 (Fig. 4.6C). In WBC from 19 sets of MZ twins (used in the first part of the study) and 20 sets of DZ twins matched for age, sex, and blood cell count (total WBC count, neutrophil, and lymphocyte fractions), MZ-DZ differences were much more subtle, but still significant (ICC\textsubscript{MZ} - ICC\textsubscript{DZ} = 0.0073 \pm 0.0034, P=0.044). It remains possible that the observed effect was diminished by our conservative efforts to bias against larger epigenetic MZ – DZ differences by selecting matched DZ twins with smaller co-twin cell sub-fraction differences as compared to the MZ twins. For buccal tissue, a locus specific annotation of ICC\textsubscript{MZ} - ICC\textsubscript{DZ} values representing dichorionic MZ co-twin similarity relative to DZ co-twin
similarity is provided (Fig. 4.7, and Fig. A.3 and A.4 in Appendix 1 for WBC and monochorionic Buccal samples).

**Figure 4.6. MZ and DZ ICC distributions in buccal cells**

ICC distributions in buccal epithelial cells of MZ and DZ twins. A) all MZ twins (N= 20 sets, red) and DZ twins (N=20 sets, blue); B) dichorionic MZ twins (N=10 sets, red) and matched DZ twins (N=10 sets, blue); C) monochorionic MZ buccal samples (N=10 sets, red) with matched DZ twins (N=10 sets, blue).

**Figure 4.7. Karyogram of MZ\textsubscript{ICC}–DZ\textsubscript{ICC} values in buccal cells of DC MZ twins**

A chromosomal karyogram depicting levels of dichorionic MZ co-twin similarity relative to DZ co-twin similarity per interrogated locus in the buccal sample. Blue bars to the right of each chromosome represent locus specific ICC\textsubscript{MZ}–ICC\textsubscript{DZ} values.
All techniques for enrichment of differentially methylated DNA sequences for microarray-based DNA methylation profiling can potentially be confounded by DNA sequence variation. In our experiments, SNPs within HpaII restriction sites may have caused enrichment differences, which resulted in larger variation in DZ twins. In addition, DNA sequence variants may influence the epigenetic status as in the literature there are several examples of DNA allele or haplotype association with specific epigenetic profiles[76, 117, 242].

Alternatively, DZ twins may exhibit more epigenetic differences than MZ twins because the former originate from different zygotes carrying two different epigenetic profiles while the latter develop from the same zygote, and therefore should possess similar epigenomes at the time of blastocyst splitting. While none of the below experiments unequivocally prove this second hypothesis, we favour the idea of these zygotic epigenetic effects for the following reasons.

First, epigenetic profiles are not fully determined by DNA sequence; if that were the case, MZ twins would exhibit no epigenetic differences. Therefore, the observed major, epigenome-wide differences in the buccal epithelial cells from MZ twins vs. DZ twins are highly unlikely to be caused exclusively by DNA sequence differences in DZ twins. Furthermore, ICC$_{MZ}$-ICC$_{DZ}$ differences were tissue specific, as the buccal epithelial cells from dichorionic MZ twins exhibited much larger MZ-DZ epigenetic differences in comparison to such difference in a subset of WBC, although both types of cells were obtained from the same individuals at the same time. As the DNA sequences should be identical (or nearly identical) between the tissues of the same organism, the tissue specific ICC$_{MZ}$-ICC$_{DZ}$ differences argue against DNA sequence as a major controlling factor of epigenetic profiles.

Second, to address the putative effects of differential digestion of polymorphic HpaII restriction sites in DZ twins we tried to perform a comparative HpaII vs. its isoschisomer, MspI, analysis as has been suggested in the HELP assay[243]; however, levels of technical variation produced in MspI-based experiments were dramatically larger than those of HpaII experiments (Ratio of HpaII/MspI variance= 0.37, P<9.9x10$^{-324}$) (Fig. 4.8). It is not surprising, given that digestion of genomic DNA with MspI generates at least an order of magnitude more short restriction fragments, which will cause different dynamics of the subsequent steps (adaptor ligation, PCR, hybridization) in comparison to the HpaII-based enrichment of the unmethylated DNA fraction. As a result, the two experiments were not directly comparable. Instead of the “wet” experiment,
we performed an *in silico* analysis whereby the SNP and allele frequency information available in the dbSNP ([http://www.ncbi.nlm.nih.gov/projects/SNP/](http://www.ncbi.nlm.nih.gov/projects/SNP/)) and HapMap ([http://www.hapmap.org/](http://www.hapmap.org/)) databases were obtained to calculate heterozygosity quotients that represent the probability that a given probe would have a restriction site disrupted by a SNP. From the 6,405 and 5,917 unique sequences within the WBC and buccal data sets, 109 and 98 loci containing HpaII SNPs were identified, respectively. For both data sets, there was no correlation of locus heterozygosity value with ICC$_{MZ}$-ICC$_{DZ}$ value (R=-0.0032, P=0.97, and R=0.024, P=0.81, for WBC and buccal cells, respectively). A similar analysis was performed to address the epigenetic effects of SNPs in *cis* by extending the interrogated region to include all SNPs within 1 kb proximal to and including the probe sequence. Again, correlation analysis of heterozygosity values at 1,369 (WBC) and 1,284 (buccal) SNP containing loci showed no correlation with ICC$_{MZ}$-ICC$_{DZ}$ value (R=-0.019, P=0.47, and R=0.033, P=0.23, respectively). These results are in agreement with a recent study that identified that only 0.16% of SNPs are associated with allele specific DNA methylation changes[244].

**Figure 4.8. Technical variation volcano plots of HpaII and MspI based enrichments**

Volcano plots measuring technical variation produced by HpaII (red) and MspI (blue) enrichments. Each plot is produced from 4 parallel self vs. self enrichments and hybridizations at 5,997 overlapping loci between the two data sets. MspI enriched samples produce significantly more technical variance ($\sigma^2$) than that of HpaII as measured by a non-parametric Ansari-Bradley test (Ratio of HpaII/MspI variance= 0.37, P<9.9x10$^{-324}$).
Third, we investigated if DNA variation may influence DNA methylation both in cis and in trans by methylome analysis of two strains of inbred (i.e. nearly genetically identical) mice as compared to two strains of outbred (i.e. genetically non-identical) mice. Mouse brains, which come from the ectoderm, the same primary germ layer as the buccal cells, investigated in humans were subjected to the 4.6K CpG island microarray-based DNA methylation profiling. First, we determined that the detected biological variance is significantly larger than technical variance in the mouse experiments ($P<9.9 \times 10^{-324}$). We then compared the spot-wise distribution of within sibship DNA methylation variation ($\sigma^2$) between inbred and outbred mice at 2,176 unique genomic regions and did not detect any significant difference (mean difference=$2.1 \times 10^{-5} \pm 3 \times 10^{-4}$, $P=0.68$) (Fig. 4.9). Although it is not completely clear to what extent mouse brain results can be extrapolated to human buccal cells, and DNA variation in the outbred mice is less than that of unrelated humans (based on the Welcome Trust study [http://www.well.ox.ac.uk/mouse/INBREDS], our estimate is that in general outbred mouse DNA heterozygosity is 2-4 times lower in comparison to unrelated humans), the impact of DNA polymorphisms on DNA methylation does not appear to be common.
Figure 4.9. Distributions of inbred and outbred epigenetic variation

The spot-wise distributions of the within sibship variance for both inbred (red) and outbred (blue) mice. A non-parametric comparison of the distributions with a paired Wilcoxon Signed Rank test did not identify any significant epigenetic difference between groups despite the genetic variation within the outbred group (Mean difference=2.1x10^{-5} ± 3x10^{-4}, P=0.68).

In the classical twin studies, greater phenotypic similarity among MZ twin pairs compared to DZ twins has been traditionally attributed to the degree of DNA sequence similarity. Our MZ and DZ twin methylome studies suggest that in addition to identical DNA, epigenetic similarity at the time of blastocyst splitting may also contribute to phenotypic similarities in MZ co-twins. By the same argument, DZ co-twins are more different from each other than MZ co-twins not only because they possess some DNA sequence differences but also because they originated from epigenomically different zygotes. Epigenomic inheritance may explain the “intangible variance”, the concept that originated from the observation that regular (polyzygotic) inbred mice were much more different from each other than the MZ inbred mice of the same strain[113]. In conjunction with such findings, our data suggests that the phenotypic effects of the individual epigenomes of each zygote could be substantial.

Methods

Twin sample

Three cohorts of twins representing various tissues were investigated. WBC of 19 dichorionic MZ and 20 DZ twin pairs matched for age, sex, and WBC count plus buccal epithelial cells from the 10 monochorionic MZ, 10 dichorionic MZ, and 20 DZ age and sex matched twin pairs were received from the Brisbane Adolescent Twin Study[245]. WBCs and buccal cells were obtained from the same individual for 10 dichorionic MZ and 10 DZ pairs. WBC samples were from twins 13.2 ± 1 yr old (mean ± SD) and consisted of 20 females and 18 males. Monochorionic and dichorionic buccal epithelial cells both consisted of 10 males and 10 females, aged 14 ± 0.77 yr old, and all 14 yrs old, respectively, all white Caucasians (mainly northern European ancestry). MZ and DZ twins in the WBC group were selected from several thousand sets of twins of the Australian Twin Registry using hematology report data. The percentage difference between cell sub-fraction counts for the whole WBC count, neutrophil, and lymphocyte counts did not exceed 10%. The mean percentage difference in selected DZ twins was smaller than that of MZ twins to
bias against the alternative hypothesis of more epigenetic variation in the DZ twin group. Zygosity was determined by comparisons of 9 microsatellite markers giving a probability of incorrect assignment of a DZ as an MZ of less than 0.0001. Gut biopsies from 18 pairs of MZ twins were received from a Swedish twin population with Inflammatory Bowel Disease, described previously[246]. Although all twin pairs had at least one twin affected with inflammatory bowel disease, we investigated biopsies from rectal mucosa, which were macroscopically not inflamed in any of the twins investigated.

**DNA methylation profiling**

The unmethylated fraction of genomic DNA was enriched using methylation sensitive restriction enzyme, HpaII[78] and interrogated on Human 12K CpG island microarrays[72]. Enrichment of the unmethylated genome of MZ and DZ twin pairs and hybridization to the microarrays was carried out in a randomized fashion. Two technical replicates were performed for each enrichment and hybridization, after which the log ratios per each replicate were averaged to produce one value per individual per locus. All samples were hybridized against a common reference (reference 1) with the exception of 9 MZ and 10 DZ pairs in WBC, which were originally hybridized against a different common reference (reference 2) and later transformed to match reference pattern 1. Transformation was achieved by first obtaining a spot-wise log ratio of reference 2 relative to reference 1 through a comparison of two dye swapped reference 1 vs. reference 2 hybridizations. Log ratios from the 9 MZ and 10 DZ pairs originally hybridized with reference 2 were multiplied by the log ratio values of reference 1 vs. reference 2 to obtain log ratio values relative to reference 2. This transformation was followed by between array normalization using the Limma package in Bioconductor. The reference pools were created by addition of equal quantities of the enriched unmethylated WBC DNA fraction from 10 MZ and 10 DZ pairs.

**Animal studies**

Genomic DNA was extracted using standard phenol and chloroform methods from whole brain tissue of four strains of mice: c57BL6 and FVB inbred strains and for CF-1–1 and CD-1 outbred strains; all obtained from Charles River Laboratories International, Inc (http://www.criver.com/). Three litters consisting of three male mice per litter were kept in uniform environments and sacrificed at post-natal day 43. Enrichment of the unmethylated fraction of genomic DNA and
creation of the common reference pool was carried out in an identical manner to the human reference design studies. The microarrays used were mouse 4.6K CpG island microarrays, all produced during a single printing at the Microarray facility of the University Health Network, Toronto. Hybridizations were carried out in batches of 18 microarrays consisting of one amplification set from one inbred and one outbred strain per day for a total of four hybridization days. Selection and order of hybridization was determined at random through sorting on a random number generator.

Data analysis

All microarrays were scanned on the Axon 4000A scanner and cross-referenced to annotated GAL files using Genepix 6.0 Software. Microarray GAL annotation was made available from the manufacturer and downloaded at www.microarrays.ca. Normalization procedures were carried out in Bioconductor using the Limma package. All arrays underwent log ratio-based normalization, background correction, print tip loess normalization and scale normalization between blocks. Low quality flagged loci identified by Genepix were removed. Microarray data was trimmed based on the annotation information such that spot IDs containing mitochondrial DNA, translocation hot spots and repetitive elements, and those located on the X and Y chromosomes were removed. After trimming and removal of flagged loci, 6,405 (WBC), 5,918 (buccal cells), and 5,941 (gut biopsies) unique DNA sequences in humans and 2,176 DNA sequences in mice were used for subsequent statistical analyses.

All statistical tests were performed in R (http://www.r-project.org/). Using an Anderson-Darling test from the nortest package, all distributions derived from microarray data rejected the null hypothesis of normality and were subsequently evaluated with non-parametric tests. All statistical tests performed were two tailed and a P<0.05 is considered significant. Unless otherwise specified ± denotes the standard error of the mean.

Test for association of epigenetic difference with cellular heterogeneity

WB cell counts were available for all twin blood samples, allowing us to investigate any association between twin pair wise variability and the fold difference of DNA methylation variability at each locus. A spot-wise correlation between the difference in log fold change value per twin pair and the log2 of the ratio of the cell count per twin pair was calculated with the Spearman method and subjected to correction for multiple testing using the qvalue package[247].
The three separate comparisons were performed on the cell fractions with the highest proportion of cells consisting of the whole white blood cell count, total neutrophil count, and total lymphocyte count.

**Biological and technical variation**

Levels of biological variation and technical variation for individual twin sets produced by twin vs. co-twin methylome comparisons and self vs. self methylome comparisons, respectively, were measured according to the variance ($\sigma^2$) over all ~6,000 loci. Non-parametric comparisons between matched biological and technical variation for all sets were carried out by the Ansari-Bradley test. Differences between the degrees of biological and technical variation in 4 MZ twin sets were evaluated with the Kruskal-Wallis test. Technical variation produced by MspI-based DNA enrichment was tested by 4 self vs. self hybridizations and compared to HpaII technical variation levels by the Ansari-Bradley test. For the common reference design data, we addressed the null hypothesis that the difference between co-twins was not significantly larger than that between replicate hybridizations. For each tissue, the median absolute value of the fold change difference between the two technical replicate enrichments/hybridizations performed per individual was determined and compared to that between co-twin hybridizations with a paired Wilcoxon Signed Rank test for MZ twins.

For animal data, assessment of technical variation was performed in the following way. For all mice, a spot-wise correlation between replicate hybridizations was produced at 2,176 unique genomic regions. To ensure that biological variation was detectably higher than technical variation, a Monte Carlo procedure was performed to test the null hypothesis that the spot-wise correlation between technical replicates would be higher than that produced from the random pairing of biological replicates from different mice. A simulated distribution was created by randomly shuffling the replicates and re-calculating a spot-wise correlation distribution for 10,000 permutations. For each permutation, the original distribution of technical replicate correlations was compared to each randomly created distribution with a paired Wilcoxon Signed Rank test. The proportion of times the correlation distribution of original technical replicates was higher than the randomly sorted distribution was tabulated and divided by the total number of permutations to obtain the quantile and relative P value.
Spot-wise epigenetic variation

A spot-wise ICC was calculated according to the one way consistency model using the irr package, designating co-twin pairs as a class. The ICC formula is $\text{ICC} = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_w}$. Here, $\sigma^2_b$ stands for the between pair variance and $\sigma^2_w$ represents the within pair variance of the specified class. As the ICC approaches 1, the co-twins are more similar to each other than unrelated twin pairs are to each other, whereas as they approach –1, the within co-twin difference across the group is consistently larger in comparison to unrelated twin pairs. Each unique DNA region investigated by the microarray was treated as an independent measurement. To address the null hypothesis that there are no differences in the amount of DNA methylation variability between MZ and DZ twins, the distributions of unique locus ICC between MZ and DZ twins in WBC cells were evaluated with a paired Wilcoxon Signed Rank test. For buccal epithelial cells, the same hypothesis for monochorionic and dichorionic twins was evaluated in a similar manner.

For inbred and outbred mice, separately, a spot-wise distribution of within sibship epigenetic variation was created by taking the average of the variance produced by the three mice per sibship. To address the null hypothesis that there are no differences in the levels of epigenetic variation between inbred and outbred mice, these spot-wise distributions were compared with a paired Wilcoxon Signed Rank test.

Cross tissue comparison

Ten WBC samples were obtained from the same individuals as the 10 DC MZ twins used in the buccal cell analysis. Separate spot wise ICC distributions were calculated for these 10 DC MZ twins in the WBC sample and from the remaining 9 unrelated DC MZ twin WBC samples. Each distribution was compared to the buccal cell derived ICC distribution at 5919 loci overlapping between datasets by linear regression.

Investigation of genomic element class

The list of microarray probes residing within CpG islands was obtained from the annotation data (www.microarrays.ca). A list of probes residing within 1 kb of gene promoters was created by cross referencing the chromosomal coordinates of each microarray probe with the genome locations of transcription start sites located within the Transcription Start Site database (http://dbtss.hgc.jp/) using an in house Perl algorithm. For each tissue cohort, the spot- wise ICC
distribution of probes residing within CpG islands was compared to non-CpG island probes with a Wilcoxon Rank Sum test: \( N_{\text{CGI}} = 2,542 \), \( N_{\text{non-CGI}} = 3,863 \) in WBC; \( N_{\text{CGI}} = 2,343 \), \( N_{\text{non-CGI}} = 3,575 \) in the buccal cells; \( N_{\text{CGI}} = 2,352 \), \( N_{\text{non-CGI}} = 3,590 \) in the gut. The same analysis was performed for promoter-associated loci: \( N_{\text{Promoter}} = 1,341 \), \( N_{\text{non-Promoter}} = 5,064 \) in WBC; \( N_{\text{Promoter}} = 1,248 \), \( N_{\text{non-Promoter}} = 4,670 \) in the buccal cells; and \( N_{\text{Promoter}} = 1,253 \), \( N_{\text{non-Promoter}} = 4,688 \) in the gut. P values were corrected for multiple testing using the Bonferroni method.

**Gene ontology analysis**

Over representation of gene ontology category within the top and bottom 5\(^{\text{th}}\) percentile of unique promoter loci was tested using the GOhyperG [237] function of the GOstats package in Bioconductor for WBC, buccal, and gut. The top and bottom 5\(^{\text{th}}\) percentile of unique CGI associated loci was interrogated in an identical manner for each tissue. GOhyperG does not correct for multiple testing. Mappings were based on data provided by: Gene Ontology (ftp://ftp.geneontology.org/pub/go/godatabase/archive/latest) on 2007/08.

**Validation of the microarray findings**

Validation of the microarray findings was carried out on using sodium bisulfite modification as performed previously in our laboratory[77]. Sodium bisulfite modification was followed by interrogation of specific CpG sites by pyrosequencing[248] or direct cloning and sequencing. PCR amplicon, pyrosequencing, and sequencing primers are provided in Table 4.3. PCR conditions included 0.5\(\mu M\) primers, 10 \(\mu l\) of Qiagen HotStar Taq Master Mix, and ddH\(_2\)O to a final reaction volume of 20 \(\mu l\). Cycling conditions were as follows: 95\(^{\circ}\)C -15 min, 40 cycles of 95\(^{\circ}\)C -30 sec, 50\(^{\circ}\)C -45 sec, 72\(^{\circ}\)C -30 sec, 72\(^{\circ}\)C -5 min, cool to 4\(^{\circ}\)C. PCR amplicons were pyrosequenced at EpigenDX Inc (http://www.epigendx.com). A representative CpG dense probe residing within the 3\(^{\prime}\) end of the Complement C1q tumor necrosis factor-related protein 8 precursor (\(C1QTNF8\)) gene containing 18 CpG positions in a 367 bp fragment was selected for in depth analysis by cloning and sequencing in WBC DNA from 18 twin pairs. On average, 1 \(\mu l\) of PCR amplicon was ligated into 50 \(ng\) of pGEMt easy plasmid vector (Promega) with 5\(\mu l\) of 2X Rapid Ligation buffer and 3 Weiss units of T4 DNA ligase in a 10 \(\mu l\) reaction volume, and incubated overnight at 4\(^{\circ}\)C. 2 \(\mu l\) of ligation product was transformed into 50 \(\mu l\) JM109 high efficiency competent cells and plated on LB agar plates containing 0.1 mg/ml ampicilin, 50 uM IPTG, and 80 ug/ml XGAL for white colony selection. For each individual, 36 clones were
grown overnight in 1 ml LB media, pelleted and sequenced at Functional Biosciences (http://www.functionalbio.com), after which the ratio of C to T was calculated at each CpG position per individual. The methylation difference at a CpG at position 9, located within a HpaII restriction site, as well as the mean methylation difference between co-twins was compared to the microarray log ratio differences by linear regression.

<table>
<thead>
<tr>
<th>Pyrosequenced Loci</th>
<th># Pairs</th>
<th>PCR Amplicon Primers</th>
<th>Pyrosequencing Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>UHNhscpg0004390</td>
<td>15</td>
<td>5'-ACACACTATTTGTGTAATTTTTTTAGTTTTTT-3'</td>
<td>5'-AAACCCACAACACA-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 5'-CTACTCATCAATAAAAAACCC-3'</td>
<td></td>
</tr>
<tr>
<td>UHNhscpg0008483</td>
<td>10</td>
<td>5'-GATTATGTTTTATTTGGGGAATCC-3'</td>
<td>5'-CAACTAAACAAAAACATCCC-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 5'-CAACTAAACAAAAACATCCC-3'</td>
<td></td>
</tr>
<tr>
<td>UHNhscpg0004556</td>
<td>19</td>
<td>5'-GGTTGGTAGATTTGAGTTTTA-3'</td>
<td>5'-GGTTGGTAGATTTAGTTAG-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R-B 5'-CAACTATACCATCTCTACTATTTAACC-3'</td>
<td></td>
</tr>
<tr>
<td>UHNhscpg0000193</td>
<td>18</td>
<td>5'-GGGAGGTTGTTGAGGGTATT-3'</td>
<td>5'-TCTACCCCTTTTCCCCATCTAAAA-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 5'-TCTACCCCTTTTCCCCATCTAAAA-3'</td>
<td></td>
</tr>
<tr>
<td>UHNhscpg0004262</td>
<td>11</td>
<td>5'-TAGGAATTAAAAAGGATGTTTGAAGAT-3'</td>
<td>5'-AAAACCTATACCTATCCCCCTAAAACCCACTCA-3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R 5'-AAAACCTATACCTATCCCCCTAAAACCCACTCA-3'</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequenced Loci</th>
<th># Pairs</th>
<th>PCR Amplicon Primers</th>
<th>Sequencing Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIQTNF8</td>
<td>18</td>
<td>5'-GGTTGGGGAATGTATAGGGATGTTTT-3'</td>
<td>M13 Reverse</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td>5'-AACCTCAACAAAAACCCACTACATCC-3'</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3. Sodium bisulfite treated loci and primers

Column 1: microarray probe IDs and for loci subjected to sodium bisulfite modification. Column 2: the number of twin pairs interrogated per locus. Column 3: primer orientation. “F” and “R” denote the forward and reverse primer sequence. “B” denotes the addition of a biotin modification for downstream pyrosequencing applications. Column 4: Primer sequences for amplifying the respective regions from post sodium bisulfite modified DNA for
pyrosequencing and cloning and sequencing strategies. Column 5: pyrosequencing and sequencing primers are provided in the far right.

**In silico SNP analysis**

SNP and allele frequencies were initially obtained from the October 2005 release of dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) and updated with information from the March 2007 release #22 of the HapMap (http://www.hapmap.org/) database. For each locus, a heterozygositiy quotient (HQ) was calculated for two scenarios. The first was for only those SNPs residing within HpaII positions and the second was for all SNPs residing within the probe sequence and 1kb upstream and downstream. An HQ was calculated by summing the quantity of 1 minus the sum of the squared allele frequencies for all SNPs located within the interrogated region. The relationship between HQ value and ICC\textsubscript{MZ}-ICC\textsubscript{DZ} difference was evaluated through linear regression.
Thesis Discussion

The work outlined in this thesis represents a progression from the development of epigenetic technologies to their use in addressing some fundamental questions of human biology.

This first effort involved the development of the high throughput epigenomic microarray technology in Chapter 1 and the site specific Ms-SNuPE technology in chapter 2. The microarray technology was among the first high throughput methods developed to interrogate epigenetic signatures and presented several advancements over previous methods. Before the advent of microarrays, the field of epigenetics was limited to the interrogation of individual loci. The epigenome wide studies enabled a better understanding of the functional role of DNA methylation in human development and disease. Using this method, our laboratory has completed the first epigenomic analysis of DNA methylation variation in the human germline [76], investigated the role of DNA methylation differences in major psychosis [77], and thoroughly investigated the role of epigenetic variation in genetically identical organisms as per the findings in subsequent chapters in this thesis.

The adaptation of the ABI SNaPshot technique allowed for a high throughput quantification of site specific quantification of methyl-cytosine density. Prior to this study, no other groups using MS-SNuPE for methylation analysis addressed the potential biasing effect of non-matching base pairs within the primer annealing region. This problem and its solution; namely, the incorporation of degenerative nucleotides at potentially polymorphic positions within the primer annealing site, have implications not only for assay design within MS-SNuPE based technologies but also for all techniques interrogating CpG rich regions with sodium bisulfite modification. This includes pyrosequencing, which has become one of the most heavily used techniques for evaluating methyl-cytosine content at individual loci.

Twin study design considerations

In the paper in chapter 3, we investigated the sample sizes necessary to achieve adequate interrogative power using our epigenomic microarray profiling technology. We modeled the effect size of an epigenetic difference associated with phenotypic discordance based on comparison of two behaviorally different twins. These twins were discordant for psychological measures of risk taking and stress response and the identified locus, DLX1, can be reasoned to
have functional relevance to the discordant phenotype. In light of the potential for epigenetic differences to account for phenotypic discordance between MZ twins, the ability to create efficient experimental designs to investigate such factors is imperative. The above study importantly demonstrates that studies investigating discordant twin populations with limited sample sizes have the power to detect functionally relevant epigenetic differences using our epigenomic microarray technology.

**MZ co-twin epigenetic variation**

Our findings in chapter 4 represent the first epigenome wide profiling effort demonstrating numerous DNA methylation differences in multiple tissues of MZ co-twins. These findings support the theory that epigenetic metastability may account for at least some phenotypic discordance in MZ twins and support the investigation of aetiological epigenetic factors in disease.

In our study, those loci representing CGIs and promoters were found to be less variable between MZ co-twins in the WBC cohort in comparison to non-CGI and promoter loci, respectively. This finding is in agreement with prior locus specific analyses. An analysis of *MAGE-A3* and *H19* genes in clonal mammary epithelial cells attempted to quantify levels of methylation pattern error rates (MPER) in various regulatory regions including promoters, CGIs, and differentially methylated regions (DMRs)[232]. MPERs were found to be higher in un-methylated regions, but two fold lower in functionally important promoter regions as opposed to non-promoter regions (*ibid*). Such a tight regulation of epigenetic metastability at functional elements suggests that the epigenetic control of these regions is under selective evolutionary pressure to ensure proper cellular differentiation and function. The buccal cell and gut cohorts did not replicate this result; however, promoter associated loci showed a trend towards being epigenetically more similar between co-twins before correction for multiple testing in gut. Gut tissue biopsies were sampled from adult twins that were discordant for intestinal inflammation associated with Crohn’s disease in 15 pairs and ulcerative colitis in 3 pairs. While we epigenetically profiled rectal tissue devoid of an inflamed phenotype, these twins may have exhibited molecular discordance that affected this analysis. In all tissues, however, GO analysis supported the hypothesis that metastability of DNA methylation signatures at genes functionally relevant to the tissue investigated will be more tightly controlled.
Our WBC data is also in agreement with high resolution gene expression studies measuring variations between MZ twins on the genome wide level using 12,000 element Affymetrix microarrays [249]. Notably, the authors identified a distinct lack of gene expression variation in housekeeping genes, genes known to be highly expressed in order to maintain the normal functions of the cell. The transcriptional control of housekeeping genes is linked to the DNA methylation status of neighbouring CpG islands[250]. Taken together, methylation of CGIs are important for the control of housekeeping genes, which in turn are important for maintaining cellular function. The higher degree of MZ co-twin epigenetic similarity detected at CGIs and promoters reflects a tightly regulated epigenetic control of these important functions.

The results of this analysis suggest that there are mechanisms in place regulating the amount of metastable epigenetic divergence between twins at functionally important loci and those genes being expressed.

**Epigenomics of monochorionic and dichorionic twinning**

In the buccal tissue, we observed that the MZ co-twin epigenetic variability was much higher in MC MZ twins as compared to DC MZ twins. Higher phenotypic discordance is generally observed in MC over DC MZ twins, most prominently in birth weight discordance and gestational outcomes [251, 252], which is often attributed to an unequal distribution of nutritional and placental resources reaching the two developing embryos because they share a placenta [253] [254]. The detected difference in epigenetic status may be affected by an unequal intrauterine environmental effects caused by these placental anastomoses *in utero.*

An alternate possibility is that the epigenetic state of the blastomere varies with a function of time, becoming more complex as the organism develops. DC and MC MZ twins arise from a splitting of the blastomere at different developmental stages and result in twins with either separate or shared placenta, respectively. If splitting occurs after ~4 days after fertilization, as in the case of MC MZ twins[236], the epigenetic status resulting in each twin will have unequal starting points from which epigenetic patterns further diverge. Under such a model the levels of epigenetic similarity between DC MZ co-twins might be expected to be larger than that of MC MZ co-twins because they will be more epigenetically similar at the time of splitting.
Our experiments provide the first measure of epigenetic co-twin variation as a function of twinning timing and may suggest that differences in levels of phenotypic discordance between MC and DC twins may be influenced by epigenetic factors. Discerning these influences is not possible from this experiment; however, if the epigenetic status is affected by the timing of twinning, this finding has implications to the results of classical twin studies. Additionally, this finding may have implications to the study in Chapter 3 where chorionicity of the investigated twin pair was not known. such that chorionicity based epigenetic changes may have affected the observed epigenetic differences between the twin pair. The existence of epigenetic divergence between MZ twins already had implications to the classical twin design in that all divergence between MZ twins, with the exception of non-random x chromosome inactivation induced mosaicism in females, is not necessarily driven by the external non-shared environmental influence. If the degree of epigenetic divergence is a factor of the timing of twinning, a portion of the phenotypic variance observed in the MZ twin group of classical twin studies may be contingent on the chorionicity of the twins analyzed, ultimately skewing the measured heritability.

**Comparison of epigenetic profiles in MZ and DZ twins**

**DNA sequence depended DNA methylation vs. epigenetic inheritance**

There are two interpretations to the finding that MZ twins are more epigenetically similar than DZ twins because the two groups differ both in the degree of their DNA sequence identity and also the methods of fertilization by which they arise. The first interpretation of the findings, therefore, is that the larger degree of epigenetic variation observed in DZ twins is a result of their DNA sequence differences. The second interpretation is that a larger degree of epigenetic variation in the DZ twin group is a reflection of a passage of a different epigenetic state from the parent generation and therefore that the difference observed is contingent on the epigenetic state of contributing germ cells.

We performed a number of experiments, the outcomes of which all favor the second interpretation. First, our in silico comparison of the HapMap SNP allele frequencies and DZ twin ICC data failed to identify any correlation between the SNP probability and the measured
epigenetic variability in the DZ twin groups in the WBC and Buccal tissue analyses. While a number of cases have been documented where a genetic difference is indicative of an allele specific DNA methylation difference [76, 242, 255, 256], the results of our analysis suggests that this phenomenon may not be common. This observation is confirmed by a recent genome wide analysis of allele specific methylation rates, which estimated that the frequency of such occurrences were as few as 0.16% [257]. Such an estimate would predict that less than ten loci detected as epigenetically variable in our DZ twin group are the result of DNA sequence directed epigenetic differences, a number too small to significantly affect our results.

Secondly, we failed to detect a difference in the genome wide DNA methylation variability between genetically identical and genetically non-identical mice. In addition to addressing the effects of SNPs in cis, this experiment also addresses any trans effects occurring as a result of DNA sequence differences and suggests that such effects are not large enough to significantly affect our results.

The inheritance of epigenetic signals may differ from genetic inheritance in a critical capacity, in that it is not necessarily transgenerational. For the DNA sequence, a portion of the genetic information conferred to the offspring will undoubtedly be present after meiosis and will be passed transgenerationally to subsequent generations. While our experiments support a passage of epigenetic signals from the parent to offspring generation, this phenomenon may be limited to a single generation. This is because the epigenetic signals undergo major rearrangement in the formation of the germ cells. During gametogenesis, there occurs a reprogramming of epigenetic patterns necessary for development and critical for re-establishing parent specific genomic imprints [132-135]. The DNA methylation patterns undergo a global erasure, coinciding with a replacement of a number of histone modifications [136]. These phenomenon are also believed necessary for the germ cells to confer totipotency upon fertilization [151]. For these reasons, it is unclear to what degree inherited epigenetic signals conferring phenotype are transmitted transgenerationally. The molecular underpinnings enabling the few documented cases of transgenerational epigenetic inheritance are not well understood. In the cases of the Agouti A^vy and Axin^Fu phenotypes, the propagation of the phenotype is carried out in a parent of origin specific manner, suggesting that the sex specific processes of germline epigenetic reprogramming may play a role. In some cases, an epigenetic modification affecting phenotype is propagated for only a few generations before eventually disappearing [258-260]. From these
limited examples, it is clear that the inheritance of epigenetic information is not carried out in a stable and Mendelian manner.

The hypothesis of passage of epigenetic information through the germline from the parent to offspring generation may have major implications to our understanding of human development and disease. To date, the detection of heritable factors from twin studies in disease has suggested that the inherited DNA sequence differences are the molecular substrates conferring heritable risk. This interpretation has fueled a decades long search for DNA risk factors with limited success. An inheritance of epigenetic signals suggests that the heritability detected in all classical twin studies may not be limited to the DNA sequence and supports the investigation of epigenetic misregulation in complex non-Mendelian disease with inherited predisposition. Epigenetic mutations, or epimutations, conferring risk may occur spontaneously during germ cell reprogramming and be passed to the offspring. In a recent study in our laboratory, we documented significant intra- and inter-individual differences in DNA methylation of the male germline using both locus specific- and epigenome wide- microarray techniques [76]. Major epigenetic variation was detected within samples as the overwhelming majority of sperm cells of the same individual exhibited unique DNA methylation profiles. The microarray analysis identified numerous DNA methylation variable positions in the germ cell genome. The largest degree of variation was detected within the promoter CpG islands and peri-centromeric satellites among the single copy DNA fragments and repetitive elements, respectively. This study suggests that particular genomic regions may be subject to variation in the germline, a portion of which may be misregulated to the point of conferring disease risk.

Environmental exposure during the parent’s life may alter the epigenetic patterns of the germline and be passed to subsequent generations. There are some cases where specific environmental influences have been documented to affect the epigenetic status of the germ cells. In a set of studies, Anway et al., subjected rats to antiandrogenic endocrine disruptor, vinclozolin, and observed a decreased spermatogenic capacity in F1-F4 generations as well as an altered DNA methylation state in the male germline that correlated with the observed phenotype [258-260]. These observations were concurrent with higher incidence of prostate disease, kidney disease, immune system abnormalities, testis abnormalities, and tumor development in these mice [260]. Outcrosses of F1 generation mice to wild types of the opposite sex demonstrated that the phenotype was transmitted through the male germline [259].
The findings reported within add support to the epigenetic theory of complex disease’. In this model, a predisposition to disease risk is inherited in a non-Mendelian fashion through the inheritance of an epigenetic state conferring risk. This ‘pre-epimutation’ may be insufficient to cause phenotypic irregularity in its inherited state; however, in combination with a metastable drift of epigenetic patterns over the course of development and maturation, this pre-epimutation may reach a critical threshold sufficient to result in disease onset. Conversely, metastable drift may result in the abolishment of the risk state. This model is consistent with the observations of age at onset effects, MZ twin discordance, sex transmission distortion ratio, parent of origin effects, and fluctuating course, all of which are hallmarks of complex Non-Mendelian disease. Our data support the existence of an inheritance and a metastable drift of epigenetic patterns and suggest that future investigations into elucidating the aetiological mechanisms of complex disease focus on epigenetic aberrations.

**Future Directions**

The findings of these studies provide many possible avenues for future research. Initially, the major results of these studies will need to be replicated in larger cohorts of MZ and DZ twins. An investigation of alternative epigenetic modifications such as histone protein modifications may also be performed in parallel to further elucidate the relationship between histone modifications and DNA methylation in relation to epigenetic inheritance and metastability in twins. In our studies, only three tissue types were profiled; however, it would be advantageous to increase the scope of interrogated tissues to better support the claims of tissue specific conservation of epigenetic patterns in genomic regions relevant to tissue function. Additionally, a comparison of regions of epigenetic similarity across tissues of the same and different primary germ layers may allow for a better understanding of the conservation and divergence of epigenetic patterns during major tissue differentiation. Finally, epigenomic profiling of MZ twins discordant for various diseases may yield insights into the aetiological mechanisms conferring risk to disease. Ideally, such studies would be performed in both the disease affected tissue and peripheral tissues including WBCs and germline samples, in order to address whether any observed epimutations are a cause or result of the disease state.
The above investigations should be carried out in an epigenome wide fashion, using more sophisticated techniques with higher resolution. Initially, these experiments should be performed using higher density microarrays such as the Affymetrix tiling arrays reviewed in the introduction capable of providing 35 bp resolution of the genome. However, even such high density platforms and techniques have limitations in terms of the inherent error in the technique, the resolution and scope of the platform, which often avoid potentially interesting regions of the genome such as repetitive elements and pericentromeric regions that are not on the arrays because they are difficult to sequence.

The recent development of high throughput sequencing technologies such as the Solexa, 454, and ABI SOLiD technologies may help to overcome the shortcomings of the current microarray based platforms for epigenetic profiling. Such techniques are just beginning to be applied to sodium bisulfite modified templates and will dramatically increase the scope and resolution of epigenetic twin studies. These technologies involve the isolation of the target sections of the genome (or in this case, either the sodium bisulfite modified genome or select sections isolated by immunoprecipitation or enzyme based enrichments) into short pieces followed by hybridization onto millions of probe sequences [261]. The number of probes is sufficient to allow the amplification of the entire genome in subsequent steps, after which, the sequences of each of the millions of probes is read in parallel by reading the emissions produced by base removal at each probe [261, 262]. Such advances will make epigenetic studies at both the genome wide scale possible while retaining high resolution of DNA methylation patterns of every base pair. These technologies are just beginning to be applied to epigenetic studies, investigating histone modifications [263] and allowing the first epigenome sequence reading of the *Arabidopsis Thaliana* genome[264], amongst others. Such new technologies will greatly improve our understanding of the role of epigenetic signals in human biology.
Appendix 1

Supplementary Notes

Correlation of MZ co-twin epigenetic variation with WB cell counts

The spot-wise correlation between twin pair loess M log ratio values and WB cell counts did not yield any significant loci after correction for multiple testing. The number of genes associated with loci showing an uncorrected significance value of P<0.001 in the whole WBC, neutrophil, and lymphocyte fractions were 6, 10, and 8, respectively. Of the genes associated with identified microarray probes beyond this threshold, 3 genes including the *EOMES*, *PDCD2*, and *PTPN9* genes that are related to immune system function[265-267]. While there is a possibility that these loci have surfaced by chance, the correlation between DNA methylation status and various immune system related genes suggests that some of the differences detected in this tissue could be a result of cellular sub-fraction differences between these twins. However, the proportion of seemingly relevant correlations is less that 0.04% of the total number of unique loci, which may be a testament to the effectiveness of matching WBC cellular sub-fractions prior to epigenomic profiling.
Figure A.1. Karyogram of MZ co-twin epigenetic similarity in buccal cells

A chromosomal karyogram depicting levels of MZ co-twin similarity per interrogated locus in the buccal sample. Black and grey bars on the chromosomes represent chromosomal banding patterns while red bars are indicative of regions of high microarray probe density. Bars to the right of each chromosome represent locus specific ICCs depicting levels of MZ co-twin epigenetic similarity. FDR corrected P values below the level of P<0.05 are depicted in green while those with greater P values are depicted in grey.
Figure A.2. Karyogram of MZ co-twin epigenetic similarity in gut

A chromosomal karyogram depicting levels of MZ co-twin similarity per interrogated locus in the gut sample. Black bars on the chromosomes represent chromosomal banding patterns while red bars are indicative of regions of high microarray probe density. Bars to the right of each chromosome represent locus specific ICCs depicting levels of MZ co-twin epigenetic similarity. Raw P values below the level of P<0.05 are depicted in green while those with greater P values are depicted in grey.
Figure A.3. Karyogram of $MZ_{ICC} - DZ_{ICC}$ values in WBCs

A chromosomal karyogram depicting levels of MZ co-twin similarity relative to DZ co-twin similarity per interrogated locus in the WBC sample. Blue bars to the right of each chromosome represent locus specific $ICC_{MZ} - ICC_{DZ}$ values.
Figure A.4. Karyogram of MZ\textsubscript{ICC}-DZ\textsubscript{ICC} values in buccal cells of MC MZ twins

A chromosomal karyogram depicting levels of MZ co-twin similarity relative to DZ co-twin similarity per interrogated locus in the monochorionic buccal sample. Blue bars to the right of each chromosome represent locus specific ICC\textsubscript{MZ}-ICC\textsubscript{DZ} values.
References


