Bioinformatic and biological analysis of DNA methylation in the human genome

Martin Ingi Sigurðsson, cand. med.

Thesis for the degree of Philosophiae Doctor

University of Iceland

Faculty of Medicine

School of Health Sciences

June 2011

Bioinformatic and biological analysis of DNA methylation in the human genome

Martin Ingi Sigurðsson, cand. med.

Thesis for the degree of Philosophiae Doctor

Supervisor:

Jón Jóhannes Jónsson, cand. med., Ph.D.

Co-instructor:

Hans Tómas Björnsson, cand. med., Ph.D.

Doctoral committee:

Albert V. Smith, Ph.D.
Eiríkur Steingrímsson, Ph.D.
Bernhard Ö. Pálsson, Ph.D.

University of Iceland
School of Health Sciences
Faculty of Medicine
June 2011

Lífupplýsingafræðileg og sameindalíffræðileg greining á eiginleikum DNA-metýlunar í erfðamengi mannsins

Martin Ingi Sigurðsson, cand. med.

Ritgerð til doktorsgráðu

Umsjónarkennari:

Jón Jóhannes Jónsson, cand. med., Ph.D.

Meðleiðbeinandi:

Hans Tómas Björnsson, cand. med., Ph.D.

Doktorsnefnd:

Albert V. Smith, Ph.D.
Eiríkur Steingrímsson, Ph.D.
Bernhard Ö. Pálsson, Ph.D.

Háskóli Íslands
Heilbrigðisvísindasvið
Læknadeild
Júní 2011

ISBN 978-9935-9024-5-0

Printing by Háskólaprent ehf.
Reykjavík, Iceland 2011

ÁGRIP

Utangenaerfðir (e. epigenetics) fjalla um upplýsingar tengdar erfðaefninu sem erfast við frumuskiptingu án þess að vera hluti af DNA röðinni sjálfri. Metýlun á DNA er mest rannsakaða utangenamerkið. Hún kemur við sögu í óvirkjun X litningsins, vörnum gegn stökklum, stýringu á vefjasérhæfðri tjáningu gena og tjáningu genagreyptra gena. Breytt DNA-metýlun er talin hluti af meingerð margra algengra sjúkdóma, þar á meðal krabbameins. Markmið doktorsverkefnisis var að beita lífupplýsinga- og líffræðilegum aðferðum til að auka skilning á dreifingu og hlutverki DNA-metýlunar í erfðamengi mannsins.

Til að auðvelda túlkun á mælingum á heildarmetýlun erfðaefnis var gerð lífupplýsingafræðileg greining á eiginleikum skerðiensíma sem unnt er að nota til slíkra mælinga. Sýnt var fram á að heildarmetýlun erfðaefnis og metýlun í stýriröðum ýmissa gena breyttist með tíma í tveimur mismunandi langsniðsþýðum frá Íslandi og Bandaríkjunum. Heildarmetýlun breyttist um meira en 10% hjá 29% einstaklinga í íslenska þýðinu (P<0.001). Breytingin var í báðar áttir, DNA-metýlun jókst hjá hluta en minnkaði hjá hluta einstaklinganna milli mælipunktanna tveggja. Í þýðinu frá Bandaríkjunum varð sambærileg breyting á heildarmetýlun, og reyndist varðveisla metýlmynstursins fjölskyldulægur eiginleiki (h^2=0.99, P<0.001).

Þar sem erfitt er að afla sýna úr kímlínu mannsins, var þróað lífupplýsingafræðilegt merki fyrir metýlun kímlínunnar sem byggir á kortlagningu metýltengdra eins basapara erfðabreytileika (mSNP). Merkið var nýtt til að sýna fram á jákvæða fylgni milli DNA-metýlunar kímlínunnar og endurröðunar erfðaefnis (r=0.622, P<10^{-15}) sem hélst þó leiðrétt væri fyrir bjagandi breytum (r=0.172, P<10^{-15}). Merkið var einnig nýtt til að kanna samband milli metýlunar kímlínunnar og mismunandi undirfjölskyldna stökkla í erfðamengi mannsins. Í ljós kom neikvæð

fylgni milli mSNP merkisins og *Alu* undirfjölskyldunnar leiðrétt fyrir bjagandi breytum (r=-0.14, -0.16, -0.16, -0.20 fyrir 125, 250, 500 og 1000 kb erfðamengisglugga). Fylgnimynstrið milli mSNP merkisins og L1 undirfjölskyldunnar var breytilegt eftir gluggastærðum (r=-0.01, -0.01, -0.01, -0.17 fyrir 125, 250, 500 og 1000 kb erfðamengisglugga). Loks var aðferðum kerfislíffræði beitt til að kanna áhrif breyttrar tjáningar genagreyptra gena á efnaskipti mannsins. Mesta breyting í líkani af efnaskiptum varð þegar hermt var eftir breyttri tjáningu *ATP10A* gensins. Áhrif breyttrar tjáningar á efnaskipti studdu ekki tilgátu Haigs um samkeppni genagreyptra gena frá föður og genagreyptra gena frá móður, því niðurstöður einungis 50% genanna fylgdu forspá tilgátunnar (P=1.0).

Niðurstöður verkefnisins renna stoðum undir þá forsendu utangenaerfðafræðilegs líkans af meingerð algengra sjúkdóma að utangenamerki breytist með aldri (Grein II). Þá benda niðurstöður verkefnisins til þess að DNA-metýlun kímlínunnar sé tengd endurröðun erfðaefnis (Grein III). DNA-metýlun er ósennilega hluti af varnarkerfi gegn *Alu* stökklum, en hún kann að vera þáttur af vörnum gegn L1 stökklum (Grein IV). Niðurstöður kerfislíffræðilegrar greiningar á áhrifum genegreyptra gena á efnaskipti studdu ekki tilgátu Haigs (Grein V). Aðferðir sem þróaðar voru nýtast til að túlka niðurstöður mælinga á heildarmetýlun erfðaefnis með skerðiensímum (Grein I), til að kortleggja DNA-metýlun kímlínu mannsins (Grein III) og til að kanna magnbundin áhrif tjáningar gena á efnaskipti mannsins með aðferðum kerfislíffræði (Grein V).

Lykilorð: Utangenaerfðir, DNA-metýlun, eins basa erfðabreytileiki, endurröðun, stökklar, kerfislíffræði

ABSTRACT

Epigenetics is the study of DNA related information heritable through meiosis and mitosis that does not include the DNA code itself. DNA methylation is currently the most studied epigenetic mark. It is a part of the inactivation of the X chromosome, defense against transposable elements and the control of tissue-specific gene expression and the expression of imprinted genes. Changes in DNA methylation are thought to be involved in the pathogenesis of many common diseases, including cancer. The aim of the Ph.D. project was to apply bioinformatic and biological methods to further the understanding of the properties of DNA methylation in the human genome.

To assist interpreting results from global methylation assays, a bioinformatics analysis of the properties of methylation-sensitive restriction endonucleases suitable for such measurements was performed. Intra-individual changes in DNA methylation over time were demonstrated in longitudinal samples from two populations from Iceland and USA. Global methylation changed by more than 10% for 29% of the individuals in the Icelandic population ($P<0.001$). The change was bi-directional; DNA methylation decreased for a part of the population but increased for another part of the population. The global methylation similarly changed in the USA population, and there was a familial clustering of conservation of methylation ($h^2=0.99$, $P<0.001$).

Since sampling the human germline is difficult, a genome-wide bioinformatic surrogate marker for germline methylation utilizing methylation-associated single base pair polymorphism (mSNP) was developed. It was used to demonstrate a positive correlation between germline methylation and homologous recombination ($r=0.622$, $P<10^{-15}$) that remained significant after correcting for confounding variables ($r=0.172$, $P<10^{-15}$). The marker was then used to explore the relationship

between germline methylation and subfamilies of transposable elements in the human genome. After correcting for confounding variables, a negative correlation was found between the mSNP marker and *Alu* subfamily (r=-0.14, -0.16, -0.16, -0.20 for 125, 250, 500 and 1000 kb genome windows) The correlation pattern between the mSNP marker and the L1 subfamily varied with window size (r=-0.01, -0.01, -0.01, -0.17 for 125, 250, 500 and 1000 kb genome windows). Finally methods of systems biology were used to study the metabolic effects of different expression level of imprinted genes. The greatest perturbation in the metabolic reconstruction occurred when differential expression of the *ATP10A* gene was simulated. The simulated effects of differential expression on metabolism did not support Haig's parental intergenome conflict theory, since only 50% of the genes followed its predictions (P=1.0).

The results of the thesis support one prerequisite of an epigenetic model of common disease pathogenesis, i.e. that epigenetic marks change with time (Paper II). The results of the project suggest that DNA methylation of the germline is associated with homologous recombination (Paper III). The results indicate that DNA methylation is unlikely a part of a defense system against *Alu* elements, although it might participate in a defense system against L1 elements (Paper IV). The systems biology analysis of the metabolic effect of imprinted genes are not supportive of Haig's theorem (Paper V). Methods developed in this project can be used to interpret global DNA methylation analysis measurements with restriction endonucleases, to map the DNA methylation of the human germline and test dosage sensitivity of human metabolic genes.

Key words: Epigenetics, DNA methylation, single nucleotide polymorphism, transposon-derived repeats, systems biology.

ACKNOWLEDGEMENTS

The work presented in this thesis was mostly carried out in Dr. Jón Jóhannes Jónsson's lab in the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Iceland. Analysis of age-related changes in DNA methylation was done in Dr. Andrew P. Feinberg's lab at Johns Hopkins University in Baltimore, USA. The network analysis of metabolic imprinted genes was partially done in Dr. Bernhard Ö. Pálsson's lab at University of California in San Diego, USA.

Firstly, I would like to express gratitude toward my instructors. I would like thank Dr. Jón Jóhannes Jónsson for excellent supervision and valuable lessons in critical thinking and scientific writing. I would also like to thank Dr. Hans Tómas Björnsson for turning a curious medical student into a scientist by endless support and encouragement. I am extremely grateful to Hans for his contribution to my scientific carrier.

I am also grateful to the members of my Ph.D. Committee, Dr. Albert Vernon Smith, Dr. Eiríkur Steingrímsson and Dr. Bernhard Ö. Pálsson. I've been fortunate enough to collaborate with all of my committee members on various projects and each has shared with me valuable experience and enthusiasm for good science.

Several colleagues and friends from Iceland and USA are thanked for collaboration and stimulating discussions. Especially I thank Dr. Andrew P. Feinberg, Dr. Neema Jamshidi from the Pálsson lab and Guðný Eiríksdóttir and Dr. Vilmundur Guðnason from Hjartavernd.

Furthermore, I am grateful to my family and friends for invaluable support during this work. Especially I would like to thank my mother, Emilía, who has been very supportive and shared with me several lessons from her successful carrier.

Finally I would like to thank my wonderful wife and best friend,

Anna. In addition to her support and contribution to my work, she has given me many precious moments during my off-work hours.

This work was funded by the University of Iceland Research Fund, the Icelandic Student Innovation Fund, The Landspítali University Hospital Science Fund and the Memorial Fund of Bergþóra Magnúsdóttir and Jakob Bjarnason.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ORIGINAL PAPERS

I. **Sigurdsson MI**, Bjornsson HT, Jonsson JJ. (2011). Analysis of sequence specificity of methylation-sensitive restriction endonucleases in the human genome. *[Manuscript]*.

II. Bjornsson HT†, **Sigurdsson MI**†, Fallin DM†, Irizarry RA, Aspelund T, Cui H, Yu W, Rongione MA, Ekström TJ, Harris TB, Launer LJ, Eiriksdottir G, Leppert MF, Sapienza C, Gudnason V, Feinberg AP. (2008). Intra-individual change over time in DNA methylation with familial clustering. *JAMA, 299*, 2877-83. †Authors contributed equally to the work.

III. **Sigurdsson MI**, Smith AV, Bjornsson HT, Jonsson JJ. (2009). HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Research, 19*, 581-9.

IV. **Sigurdsson MI**, Smith AV, Bjornsson HT, Jonsson JJ. (2011). Correlation Between Transposon-Derived Repeats and Markers of Germline Methylation illustrate divergent Epigenomic Responses to the major classes. *[Manuscript]*.

V. **Sigurdsson MI**, Jamshidi N, Jonsson JJ, Palsson BO. (2009). Genome-scale network analysis of imprinted human metabolic genes. *Epigenetics 4*, 43-6.

CLARIFICATION OF CONTRIBUTION

Paper I

I conceived of the study along with coauthors. I wrote and tested computer algorithms to process data. I performed statistical analysis of the data and drew the figures. I wrote the manuscript draft and participated in its revision.

Paper II

I participated in setting up and optimizing the LUMA protocol and performed a bioinformatics analysis of its properties. I performed LUMA measurements of both cohorts in the study. I participated in data analysis and revision of the manuscript.

Paper III

I conceived of the study along with coauthors. I wrote and tested computer algorithms to process data. I performed statistical analysis of the data and drew the figures. I wrote the manuscript draft and participated in its revision.

Paper IV

I conceived of the study along with coauthors. I wrote and tested computer algorithms to process data. I performed statistical analysis of the data and drew the figures. I wrote the manuscript draft and participated in its revision.

Paper V

I conceived of the study along with coauthors. I performed phenotype simulations, analyzed the results and drew the figures. I wrote the manuscript draft and participated in its revision.

ABBREVIATIONS

AGES	Age, Gene/Environment Susceptibility Reykjavik Study
CDGE	Common Disease genetic epidemiology in the context of both Genetic and Epigenetic variation
cM	centiMorgan
COBRA	Constraint-based reconstruction and analysis
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTPs	Deoxynucleotide triphosphate
DSBR	Double-strand break repair
DZ	Dizygous
ESC	Embryonic stem cell
FBA	Flux balance analysis
FVA	Flux variability analysis
HEP	Human Epigenome Project
HPLC	High-performance liquid chromatography
IAP	Intracisternal A-particle
ICF	Immunodeficiency-centromeric instability-facial anomalies
kb	Kilobase; 1,000 bases
LINE	Long interspersed nucleotide element
LOI	Loss of imprinting
LTR	Long terminal repeat
LUMA	Luminometric methylation assay
Mb	Megabase; 1,000,000 bases
mRNA	Messenger RNA
mSNP	Methylation-associated SNP
MZ	Monozygous
NAHR	Non-allelic homologous recombination

ORF	Open reading frame
PCGT	Polycomb group target
PCR	Polymerase chain reaction
piRNA	PIWI interacting RNA
PIWI	P-element induced wimpy testis
RNA	Ribonucleic acid
SDSA	Synthesis-dependent strand-annealing
SINE	Short interspersed nucleotide element
siRNA	Short interfering RNA
SNP	Single nucleotide polymorphism
TDR	Transposon-derived repeat

1 INTRODUCTION

The focus of this Ph.D. thesis is several aspects of DNA methylation in the human genome. The body of the work presented involves the development and application of bioinformatic and biological methods to quantify DNA methylation in the germline and somatic tissues and to test its conservation and correlation with genetic variables such as recombination and the amount of repetitive sequences.

1.1 Epigenetics

The concept of epigenetics was originally established by Conrad Waddington in 1942 (Waddington, 1942). He proposed that environmental stimulus could be converted into an internal genetic factor by "canalization of development" (Waddington, 1942), explaining how complex phenotypes could form from interaction between genes and environment. Epigenetics focuses on DNA related information, heritable through both meiosis and mitosis, that does not involve the DNA sequence itself. Recently, an operational consensus definition of epigenetics was established (Berger *et al.*, 2009). The official definition of epigenetic trait reads: "An epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" (Berger *et al.*, 2009).

Epigenetic marks (epigenetic maintainers) maintain the heritable chromosome changes (Berger *et al.*, 2009). Currently, the best known marks are DNA methylation, post-translational histone modifications and nucleosome positioning. These epigenetic marks affect gene expression (Jones *et al.*, 1998; Razin & Riggs, 1980) and have tissue-specific patterns (Eckhardt *et al.*, 2006) underlying tissue-specific gene expression of genes (Musco & Peterson, 2008). The various epigenetic marks interact resulting in a complex epigenetic machinery (Ikegami *et al.*, 2009; Ng & Bird, 1999). Disruption of this

machinery underlies the pathogenesis of many human diseases, such as cancer (Esteller, 2008).

1.2 DNA methylation

1.2.1 DNA methylation in the human genome

DNA methylation in the human genome is a covalent addition of a methyl group to the fifth carbon of the cytosine base (Fig. 1). In humans the fraction of methylated cytosine is 0.76-1% depending on tissues, corresponding to 4-5% of all cytosine bases (Ehrlich *et al.*, 1982). Approximately 70-80% of all CpGs in the human genome are methylated (Bird, 2002). However, embryonic stem cells seem to have a substantial amount of non-CpG methylation (Lister *et al.*, 2009; Ramsahoye *et al.*, 2000). The first whole-genome single nucleotide methylation analysis of two human cell lines indicated that while 99.98% of all cytosine methylation was within the CpG dinucleotide in a fetal fibroblast cell line, 24,5% of cytosine methylation in an embryonic stem cell (ESC) line was within non-CpG cytosines (Lister *et al.*, 2009). The majority of non-CpG methylation was within the cytosine base of the CWG (W=A or T) trinucleotide. After differentiation of the ESC line, cytosine methylation in non-CpGs disappeared and induction of the fibroblast cell line into a pluripotent state resulted in the reappearance of cytosine methylation of non-CpGs (Lister *et al.*, 2009). This indicates that cytosine methylation of non-CpGs might be a part of a specific pluripotent cell mechanism.

Several methods are available to assess DNA methylation at individual cytosines. The most popular method involves treating the DNA sample with bisulfite. This converts unmethylated cytosine into uracil while methylated cytosine remains intact (Zilberman & Henikoff, 2007). Following this, the amount of conversion and thereby methylation is measured by performing either polymerase chain reaction (PCR) with targeted primers for methylated

and unmethylated DNA (bisulfite PCR) (Fraga & Esteller, 2002), hybridization of the sample onto microarray or direct sequencing (bisulfite sequencing) (Zilberman & Henikoff, 2007). Another popular method, methylated DNA immunoprecipitation (MeDIP), involves isolating methylated DNA with antibody specific for methylated DNA (Sørensen & Collas, 2009). After immunoprecipitation, the methylated and unmethylated part of the sample can be differentially labeled and applied to microarray (MeDIP-chip) or sequenced (MeDIP-seq) (Sørensen & Collas, 2009). Several methods are also available to measure global methylation of DNA. These include assays based on High-performance liquid chromatography (HPLC) (Armstrong *et al.*, 2010) and mass spectrometry (Rocha *et al.*, 2010), in addition to methods measuring the amount of cut by methylation-sensitive restriction endonucleases (Karimi *et al.*, 2006a). It was not clear how many suitable endonuclease pairs were available for such measurements prior to the work described in this thesis. Furthermore, knowing the frequencies of the restriction endonucleases within subsets of the genome can aid in the interpretation of global DNA methylation measurements and comparison with results by other methods. In paper I, the sequence specificity of restriction endonucleases potentially suitable for global methylation analysis in the human genome was analyzed. In particular two restriction endonuclease target sites suitable for global CpG dinucleotide methylation measurement and one target site suitable for global CWG trinucleotide methylation measurement were studied.

1.2.2 The role of DNA methylation in mammalian genomes

DNA methylation was proposed as a form of cellular memory in 1975 by two independent researchers (Holliday & Pugh, 1975; Riggs, 1975). They predicted the discovery of eukaryotic methyltransferases capable of maintaining methylation through cell division by high affinity for hemimethylated DNA

(Riggs, 1975). This enzyme, DNA methyltransferase I, was described eight years later (Bestor & Ingram, 1983). It belongs to the family of DNA methyltransferases (Dnmts) that establish and maintain DNA methylation in mammalian genomes (Cheng & Blumenthal, 2008) (Figure 1). Members of the Dnmt3 subfamily are responsible for *de novo* methylation (Okano *et al.*, 1999).

DNA methylation has several roles in mammalian genomes. It is a part of the control of tissue-specific regulation of gene expression. Early evidence includes the silencing of the *Aprt* gene by methylation following insertion into mouse cells (Stein *et al.*, 1982), and the expression of silenced genes on the X chromosome when cell cultures were treated with a methyltransferase inhibitor (5-azacytitine) (Venolia *et al.*, 1982). CpG islands are long stretches of DNA with a high observed/expected ratio of CpGs, generally thought to be result from over-representation of non-methylated CpG dinucleotides that are less susceptible to mutations. They represent 0.7% of the human genome (Gardiner-Garden & Frommer, 1987). Nearly all housekeeping genes and the majority of genes with tissue-specific expression contain one or more CpG islands in their promoter region (Cross & Bird, 1995). The methylation of the CpGs in these promoter regions correlates inversely with promoter activity (Weber *et al.*,

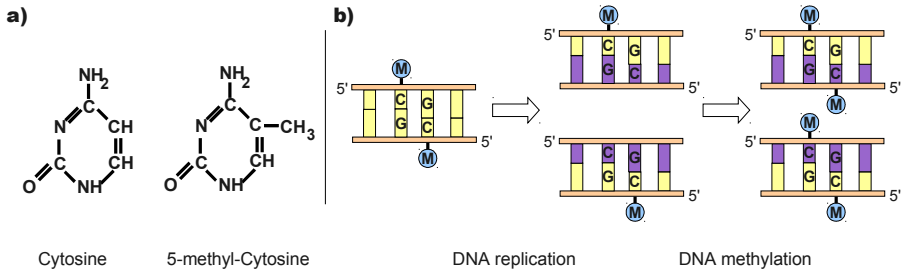


Figure 1: DNA methylation in the human genome

a) Structure of cytosine and methylated cytosine. b) The dyad symmetry of the CpG dinucleotide, ensuring accurate maintenance methylation during cell division. Drawing based on Reik *et al.* (2001).

2007) and gene expression (Song *et al.*, 2005).

An important role for DNA methylation in mammals is its part in the maintenance of the inactive state of one of the female X chromosomes. The inactivation occurs by production of non-coding RNA by a single copy of the *XIST* gene that coats the entire chromosome (Brown *et al.*, 1991). This is followed by histone tail modifications (Cohen *et al.*, 2005) and later DNA methylation of gene bodies that maintains stable inactivation (Hellman & Chess, 2007; Venolia *et al.*, 1982).

In mammals, DNA methylation is vital for genome stability, as shown by various mouse mutants of the methyltransferase genes. Homozygous mice for disruptive mutations in the *Dnmt1* gene die at the embryonic stage (Li *et al.*, 1992). The offspring of female mice with conditional mutations of the *DNMT3a* die *in utero* and have lost methylation of long terminal repeats and allele specific methylation of normally maternally silenced loci (Kaneda *et al.*, 2004). Mice with mutations in the *Dnmt3l* gene similarly have no methylation of long terminal repeats and germ cell meiosis is arrested (Bourc'his & Bestor, 2004). Patients with mutations in the C terminal DNA methyltransferase domain of *DNMT3B* gene on chromosome 20q develop immunodeficiency-centromeric instability-facial anomalies (ICF) syndrome (Xu *et al.*, 1999). They have hypomethylation of satellite repeats, affecting the centromere stability of chromosomes 1, 9 and 16 (Okano *et al.*, 1999).

Changes in DNA methylation and other epigenetic marks are involved in the pathogenesis of many human diseases. Feinberg and Vogelstein reported differences in the global methylation pattern between cancers and their healthy tissue counterparts in 1983 (Feinberg & Vogelstein, 1983). Since then, alterations in both DNA methylation and histone modifications have been

discovered in many cancer types and several treatments targeting these changes have emerged (Esteller, 2008). Epigenetic modifications have also been correlated with several non-malignant diseases, such as systemic lupus erythematosus (Javierre *et al.*, 2010).

An epigenetic and genetic model of disease pathogenesis has been suggested to explain several aspects of complex human disease (reviewed in chapter 1.4.1). One prerequisite of the model involves acquired changes in the pattern of epigenetic marks over time. However, limited data supporting this prerequisite existed at the time of the initiation of the Ph.D. project. The longitudinal measurement of acquired changes in the pattern of epigenetic marks in two cohorts was the focus of Paper II. DNA methylation was selected as the epigenetic mark since it only requires isolated DNA and not whole cells. We measured intra-individual changes in both global and local patterns of DNA methylation over 11-16 years. One of the cohorts included samples from 21 families with up to three generations. This cohort was used to assess the familial correlation of DNA methylation conservation.

1.2.3 DNA methylation patterns in the germline

An accurate reprogramming of germline DNA methylation patterns, including the differentially methylated regions of imprinted genes, is vital for normal development (Weaver *et al.*, 2009). In mice following fertilization, a genome-wide demethylation occurs (Fig.2). The male genome is actively demethylated by demethylating enzymes in a single cell cycle while the maternal genome undergoes passive demethylation over a few cell cycles (Santos *et al.*, 2002; Weaver *et al.*, 2009). However, imprinted regions are spared, and they keep their parent-of-origin specific methylation patterns. Several *cis* and *trans* acting

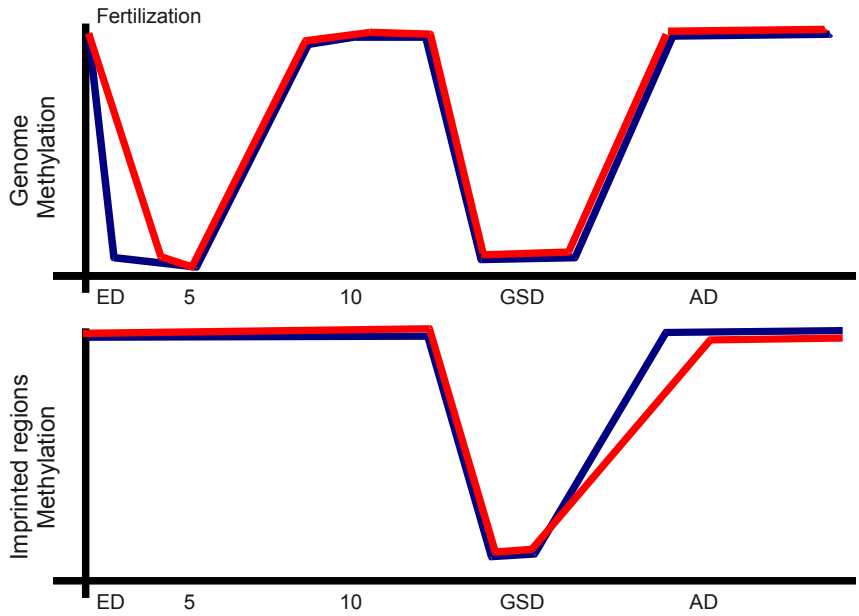


Figure 2: Reprogramming of the germline methylation pattern during embryogenesis.

Shown is the reprogramming, both genome wide (top) and at imprinted loci (bottom) for both maternal (red) and paternal (blue) genomes during embryogenesis. Genome-wide demethylation occurs at ED (embryonic day) 5 sparing imprinted loci. At gonadal sex differentiation (GSD) the formation of primordial germ cells, DNA methylation patterns in the germline are reset to establish transfer of correct information on parental origin to the offspring. This continues into adult development (AD). Drawing based on Reik *et al*. (2001), Jirtle *et al*. (2007) and Weaver *et al*. (2009).

proteins participating in this protection have been proposed (Weaver *et al.*, 2009). Furthermore, the maintenance of methylation is carried out by a cooperation of two isoforms of the Dnmt1 methyltransferase (Hirasawa *et al.*, 2008).

After the initial cell division and following the demethylation wave, a population of primordial germ cell has been produced and sex specification and gonadal cell differentiation is initiated (Weaver *et al.*, 2009). Thereafter, the germline undergoes a second wave of demethylation, completed at embryonic day 12-13 (Reik & Walter, 2001). This time, however, the imprinted regions are also demethylated (Hajkova *et al.*, 2002). Following this process, methylation patterns of imprinted genes are established by *de novo* methylation to secure correct transfer of parent-of-origin information (Reik & Walter, 2001). The experimental measurement of DNA methylation in the human germline is limited by sampling difficulties. The only readily available sample of the human germline is sperm, the final product of the male germline. Samples of the female germline and the earlier stages of the male germline are much more difficult to obtain. In paper III, we therefore took a novel approach and developed a bioinformatic marker based on hypermutability of methylated cytosine (reviewed in chapter 1.5) that represents germline methylation. This marker was subsequently used to test its correlation with homologous recombination (Paper III) and the correlation with various classes of repetitive elements in the human genome (Paper IV).

1.2.4 DNA methylation and imprinted genes.

Imprinted genes have allele-specific expression patterns based on parental origin of the allele in at least one tissue. DNA methylation maintains the stable expression patterns of imprinted genes in mammalian genomes (Reik & Walter, 2001). Alternation in the expression of several imprinted genes, that can result

from changes in methylation of their control elements, can lead to severe disease. Several suggestions of the role of imprinted genes exist. Haig's parental conflict theory suggests differential metabolic effects of paternally and maternally imprinted genes (reviewed in chapter 1.8.2). In Paper V, we simulated the metabolic effects of differential expression of imprinted genes *in silico* by applying methods of systems biology on a recent reconstruction of the human metabolic network.

1.3 Histone modifications and nucleosome positioning

Several epigenetic marks other than DNA methylation exist, although an in depth review of them is outside the scope of this thesis. The nature of their effects on gene expression and relationship with other cell mechanisms is currently under active research that has been greatly advanced by progress in immunoprecipation methods.

Post-translational modifications of histones are not all epigenetic in origin, since some of them are not stable under cell division (Berger, 2007; Berger *et al.*, 2009). Histone modifying proteins modify the amino- and carboxy terminal of the histone tails changing their properties to alter the access of transcriptional factors to the DNA (Berger, 2007). Examples include histone lysine acetylation that increases transcriptional activity (Soutoglou *et al.*, 2000) and lysine sumoylation that repress transcription (Verger *et al.*, 2003). Other histone modifications display a more complex relationship with DNA transcription, suggesting that they are just a part of the complex transcriptional machinery (Berger, 2007).

Patterns of DNA methylation and post-translational histone modifications correlate (Lister *et al.*, 2009). An example is the binding of the MeCP2 CpG binding protein (Fuks *et al.*, 2003) to methylated DNA and

subsequent recruitment of histone deacetylase to repress transcription (Jones *et al.*, 1998).

Epigenetic and genetic factors also participate in establishing epigenetic memory. Most of those are factors affecting the stability of the nucleosome, a 146 base pair (bp) long stretch of the DNA strand wrapped tightly around a histone protein octamer (Henikoff, 2008). Proteins involved in chromatin assembly locate nucleosomes at positions interfering with transcription (Henikoff, 2008). The sequence preference of nucleosome positioning is not random (Kaplan *et al.*, 2009) and has been suggested to be an epigenetic phenomena (Segal & Widom, 2009).

1.4 Age-associated changes in methylation and their potential role in complex disease pathogenesis (Paper II)

The pathogenesis of many common human diseases is a complex relationship between genetic and environmental factors. Several models explaining this relationship have been proposed (Bjornsson *et al.*, 2004; Jiang *et al.*, 2004). One of those is the CDGE (Common Disease genetic epidemiology in the context of both Genetic and Epigenetic variation) model (Bjornsson *et al.*, 2004). It supplements a purely genetic model of disease pathogenesis by adding epigenetic variability layers that can be pathogenic, either by themselves or in combination with genetic variants. An important prerequisite of the model is changes in patterns of epigenetic marks over time. Testing acquired intra-individual changes in epigenetic marks was the focus of the work presented in Paper II. Previously, this had mostly been studied in cross-sectional cohorts. We used DNA methylation as our epigenetic mark as it only requires DNA rather than whole cells required for histone modifications. Changes in global methylation and gene promoter specific methylation over 11-16 years was measured in two longitudinal cohorts with more than 100 participants each.

1.4.1 Epigenetic model of complex human disease

A classical genetic model of disease pathogenesis suggests that it results from genetic disruptions, but the phenotype might be modified by the environment. Examples of diseases adhering to this models are Mendelian disorders such as phenylketonuria (Scriver, 2007). However, the vast majority of human diseases have complex genetics and do not segregate as simple monogenic traits. This is demonstrated by low disease concordance rates for monozygotic twins, such as 40% for diabetes mellitus (Knip *et al.*, 2005). Furthermore, the majority of genome-wide association studies have discovered genome variants that commonly explain only a fraction of the cases and the presence of the variant found is neither necessary nor sufficient for establishment of disease (Manolio, 2010). Several factors could contribute to the pathogenesis of diseases with complex genetics, such as numerous genetic loci contributing to disease pathogenesis and their interaction, different or variable penetrance of mutations, acquired somatic mutations or gene/environment interaction.

The CDGE (common disease genetic epidemiology in the context of both genetic and epigenetic variation) model adds an epigenetic layer of information to supplement the genetic model of complex human disease (Bjornsson *et al.*, 2004). According to the model, genetic and epigenetic variation can interact differently to establish a pathogenic phenotype. A sequence variant contributing to disease can either be independent from epigenetic effects, or epigenetic modifications can modify its penetrance. Furthermore an epigenetic variability can either contribute independently to a disease phenotype or be affected by genetic variants. Several sources of epigenetic variability have been suggested, such as individual environment, parental environment, stochastic changes and age-dependent degeneration of epigenetic marks (Bjornsson *et al.*, 2004).

A number of observations support this. Folic acid is a key component of single carbon metabolism necessary for maintenance of DNA methylation. Hyperhomocysteinemia is a significant risk factor for cardiovascular disease (Homocysteine Studies Collaboration, 2002) and Alzheimer's disease (Seshadri *et al.*, 2002). Uremic individuals with hyperhomocysteinemia have abnormal patterns of DNA methylation corrected by folic acid supplementation (Ingrosso *et al.*, 2003). Neural tube defects are both a relatively common and severe type of birth defects. Recently, global levels of DNA methylation in the brain were found to be significantly lower in human fetuses with neural tube defects compared control fetuses, and the level of hypomethylation correlated with the severity of the defect (Chen *et al.*, 2010). Homozygous mice for mutations in the *Dnmt3B* gene are unable to perform adequate *de novo* methylation, and their offspring suffer from neural tube defects (Juriloff & Harris, 2000). Supplementation of folic acid during pregnancy or food fortification with folic acid has greatly reduced the risk of neural tube defects (Obican *et al.*, 2010; Zeisel, 2009). Maternal exposure to famine during pregnancy has been shown to stably affect DNA methylation of the *IGF2* gene in the offspring (Heijmans *et al.*, 2008). Furthermore, studies of mice with the A^{vy} allele of the *Agouti* gene have revealed that epigenetic patterns can be transferred between generations (Morgan *et al.*, 1999), and that the parental environment can affect the pattern of epigenetic marks in the offspring (Waterland & Jirtle, 2003). Collectively, this evidence supports the suggestion that environmental factors can affect the establishment and maintenance of epigenetic marks and affect disease pathogenesis.

Since the publication of the CDGE model, vast amount of supporting evidence has been discovered, in addition to results from Paper II. Stochastic changes in DNA methylation in cancer compared to healthy tissue have been

known for many years (Feinberg & Vogelstein, 1983) and epigenetic mechanisms are now thought to contribute to cancer pathogenesis (Esteller, 2008). Monozygous twins discordant for systemic lupus erythematosus were recently found to differ markedly in DNA methylation patterns for 49 genes, and the greatest differences were in genes relevant for autoimmune disease (Javierre *et al.*, 2010). A recent revision of a genome-wide data on genetic variants associated with type II diabetes found a new association correlating with adjacent methylation when the sequence variants were reviewed in light of parental origin (Kong *et al.*, 2009). This suggested that imprinted and epigenetic mechanisms might be involved in the pathogenesis of this common disease.

1.4.2 Epigenetic marks and aging

Acquired changes in epigenetic marks have been suggested to be a part of the age-dependent onset of many human diseases (Feinberg, 2004; Feinberg, 2007). At the time of publication of the CDGE model, few studies on changes in epigenetic marks with aging existed. In an inbred mouse experimental system, female X chromosome inactivation, maintained by DNA methylation, decreased with increasing age of the mice (Bennett-Baker *et al.*, 2003). Furthermore, the expression of the inactive allele of two imprinted genes (*Atp7a* and *Igf2*) was found to increase with aging (Bennett-Baker *et al.*, 2003). In humans, the results from studies of the possible decay of X chromosome inactivation with age were contradicting (Busque *et al.*, 1996; Racchi *et al.*, 1998). The methylation of a CpG island in the promoter region of the estrogen receptor (*ER*) was found to increase with age in colonic mucosa of 39 healthy individuals (Issa *et al.*, 1994). The same CpG island was hypermethylated in 45 colorectal cancers (Issa *et al.*, 1994). Similarly, the methylation of the *IGF2* promoter in 34 individuals aged 8-90 years was found to increase with age.

Furthermore, the methylation of both alleles of the differentially methylated region within the *IGF2* promoter increased with age in the 25 individuals suitable for testing (Issa *et al.*, 1996). The methylation of the non-methylated allele approached the methylation of the methylated allele with increased age of the subjects. Hypermethylation of the promoter region of the *IGF2* gene was also found in various tumors, including colorectal tumors and premalignant adenomas of the colon (Issa *et al.*, 1996). However, out of six other genes involved in the pathogenesis of colorectal cancer, only two (*MYOD, N33*) were found to have similar age-related promoter hypermethylation in colonic mucosa, while the methylation of the other genes *(p16, THBS1, HIC-1, CALCA)* did not change significantly (Ahuja *et al.*, 1998). This indicates that age-associated changes in epigenetic marks might be site-specific.

A more comprehensive analysis compared the methylation of 1.9 million CpGs in a cross-sectional sample of old and young individuals (mean age 68 vs. 26 years). The average DNA methylation changed significantly in two out of five tissues tested (Eckhardt *et al.*, 2006). The changes were in CD4+ lymphocytes and dermal fibroblasts, while no change was observed in liver, heart muscle and skeletal muscle (Eckhardt *et al.*, 2006). The authors concluded that DNA methylation is likely to be more stable than previously thought (Eckhardt *et al.*, 2006).

A recent cross-sectional study measuring the methylation of more than 27,000 CpGs in promoters of the human genome in whole blood from 93 individuals found 213 sites with age-associated hypermethylation and 147 sites with age-associated hypomethylation (Rakyan *et al.*, 2010). These findings were replicated using both monocytes and T-cells of peripheral blood in an independent cohort. Furthermore, the age-associated hypermethylated sites were found to be located within bivalent chromatin domains (containing

epigenetically activating and repressing factors) and regions hypermethylated in multiple adult-onset cancers (Rakyan *et al.*, 2010).

The results from cross-sectional studies on the effect of age on epigenetic marks can be difficult to interpret, as several possibly confounding genetic and environmental variables remain unaccounted for. Furthermore, if changes in epigenetic marks are not all uni-directional (e.g. if DNA methylation could either increase or decrease), then they are likely missed with a cross-sectional approach. No changes in X inactivation patterns were found when they were compared in two samples from 133 individuals sampled with a 13-21 year interval (Sandovici *et al.*, 2004). Similarly, the methylation of the region controlling the expression of the *IGF2* gene did not change significantly with time (Sandovici *et al.*, 2003).

Fraga *et al*. studied differences in epigenetic marks between 40 pairs of monozygous (MZ) twins of various age and environmental backgrounds (Fraga *et al.*, 2005). For the female twins, 13 out of 16 pairs had the same X chromosome methylation pattern. However, a total of 35% (14 out of 40 MZ pairs) differed significantly in both global cytosine methylation levels, global H3 acetylation levels and global H4 acetylation levels (Fraga *et al.*, 2005). Furthermore, while the youngest MZ twin pairs were epigenetically indistinguishable, the older pairs differed more. Remarkably, those twin pairs who reported having spent less of their life together and/or those who differed markedly in their medical history demonstrated the greatest differences in these global markers of epigenetic marks. This suggests that different amount of shared environment might explain the observed differences in the pattern of epigenetic marks. The sites that differed markedly between MZ twins were enriched in *Alu* elements and single-copy genes. These differences were also observed in gene expression levels (Fraga *et al.*, 2005).

Two other twin studies supporting these results have recently emerged. Probing the methylation of approximately 6000 DNA loci in 59 twin pairs, monozygous (MZ) twins were found to be more epigenetically similar than the dizygous (DZ) twins (Kaminsky *et al.*, 2009). This suggests a genetic contribution to the conservation of epigenetic marks. Additionally, information on mono or dichorionic status of 20 MZ twin pairs was available. In dichorionic MZ twins, blastocyst separation of the twins occurs within four days of fertilization whereas in monochorionic MZ, the separation occurs later. Interestingly, monochorionic MZ twins were more epigenetically similar than dichorionic twins (Kaminsky *et al.*, 2009). This suggests that in addition to identical DNA sequence, different phenotypes of MZ twins might arise due to differences in epigenetic marks as early as by blastocyst separation (Kaminsky *et al.*, 2009). In a recent study of young MZ and DZ twins sampled at 5 year intervals, the methylation of all three genes tested changed significantly between the sample points (Wong *et al.*, 2010). Furthermore, the change in all three genes was bi-directional (i.e. some individuals lost methylation while others gained methylation) (Wong *et al.*, 2010). However, the changes in MZ twins were not significantly different from DZ twins, suggesting that common environment rather a genetic variability caused the observed change (Wong *et al.*, 2010).

Several other studies on the conservation of epigenetic marks have been published recently. In 718 subjects (age 55-92) sampled repeatedly over 8 years, DNA methylation of the *Alu* elements was found to decline steadily (0.089% decrease in cytosine methylation of *Alu* elements per year). In contrast, the methylation L1 repeats, the other major repeat subfamily tested, did not change significantly (Bollati *et al.*, 2009). A microarray and restriction endocuclease-based methylation profiling of colonic mucosa compared the methylation of

promoter regions in 3627 genes between young (<12 months) and old (>12 months) inbred mice (Maegawa *et al.*, 2010). The study found that 774 (21%) of the gene promoters showed increased methylation in the older mice while 466 (13%) showed decreased methylation. Out of those, 11 regions that revealed clear change in methylation in small intestine were examined in other tissues. This revealed variable levels of change in methylation between young and old mice (Maegawa *et al.*, 2010). Furthermore, the expression of four genes demonstrating promoter hypermethylation with age, one gene demonstrating promoter hypomethylation with age and two genes with no changes with age was compared between younger and older mice. The expression changes corresponded to the methylation changes (Maegawa *et al.*, 2010).

1.5 Hypermutability of methylated cytosine and its use for bioinformatic assays of germline methylation (Paper III)

Due to limited availability of samples, direct measurements of DNA methylation in the human genome can be difficult. The germline tissue usually available for biologic analysis is sperm, the final product of the male germline. Therefore, alternative assays of the human germline methylation are useful. In Paper III, we developed and validated a novel bioinformatic marker of human germline methylation, the methylation-associated SNP (mSNP). This relied on using the density of mutations due to the hypermutable methylated cytosine, as a surrogate measurement of germline methylation. Subsequently, the marker was applied to test suggested correlation with homologous recombination and subfamily-specific density of repeated elements.

1.5.1 The hypermutability of methylated cytosine

A spontaneous deamination of unmethylated cytosine forms uracil and an uracil-guanine mismatch (Lindahl, 1974) (Fig. 3a). This mutation is readily repaired by uracil DNA glycosylase, which efficiently removes the uracil base.

Deamination of methylated cytosine causes a transition from cytosine to thymine (Fig. 3a). This transition results in a mismatch between thymine and the guanine on the complementary DNA strand (Cooper & Youssoufian, 1988; Coulondre *et al.*, 1978)(Fig 3b). A thymine DNA glycosylase enzyme (MBD4), which removes thymine from G/T mismatches, exists in mammals (Hendrich *et al.*, 1999). Mice homozygous for mutations in the *Mbd4* gene had significantly higher accumulation of C→T mismatches with age compared to controls (Millar *et al.*, 2002).

The G/T mismatch repair system seems to be much less efficient in repairing the point mutation correctly than the G/U mismatch repair system. An analysis of the repair of G/T mismatches introduced into mammalian cells revealed that although 99% of the mismatches were corrected, the thymine base was removed (correct repair) in only 92% of the cases, while the guanine base

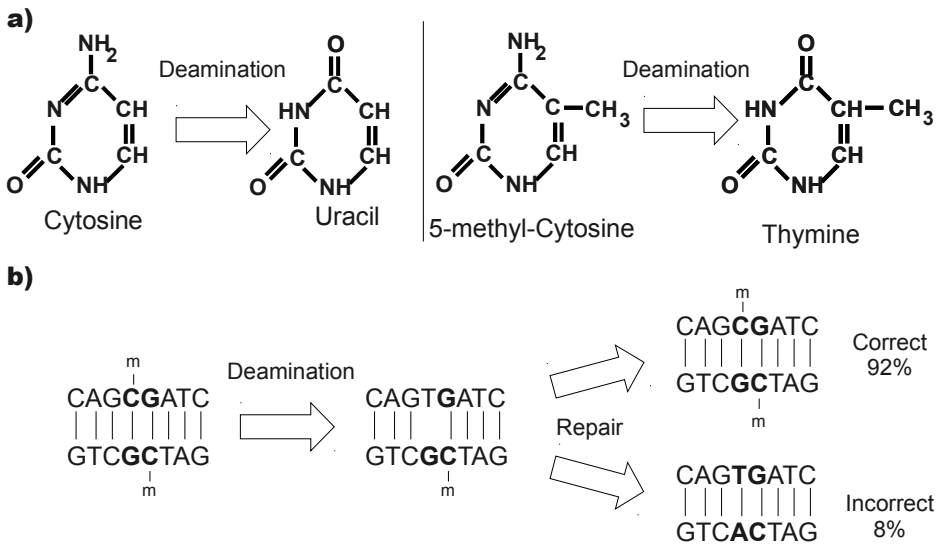


Figure 3: Instability of methylated cytosine in the human genome.

a) Deamination of cytosine results in an uracil base readily identified by the DNA repair system while deamination of methylated cytosine results in thymine. b) Following deamination of methylated cytosine, the G/T mismatch is repaired less efficiently, resulting in a high frequency of C→T transitions within the CpG dinucleotide.

was removed (incorrect repair) in 8% of the cases (Brown & Jiricny, 1987) (Fig 3b).

As a result, methylated cytosine is hypermutable. Almost a third of all point mutations are C→T transitions within the CpG dinucleotide (Cooper & Youssoufian, 1988). This is likely the cause for a great under-representation of the CpG dinucleotide in the human genome (Bird, 1980; Josse *et al.*, 1961).

The mutation rate of the male germline is substantially higher than the mutation rate of the female germline (Huang *et al.*, 1997). The proposed mechanism behind this is that mutations predominantly occur via errors in replication, and that the male germline undergoes more cell divisions than the female germ line. However, if the mechanism behind C→T mutations in the methylated CpG dinucleotide is predominantly deamination, then the male bias in the CpG dinucleotides should be substantially less than in non-CpG dinucleotides. This was verified when the mutation spectra obtained by comparing the human and chimpanzee genomes was compared between the X chromosome and the autosomes (Taylor *et al.*, 2006). The male bias (α) was ~6-7 for the non-CpGs, similar to the ratio of male/female germ line divisions. In contrast, the male bias in CpG dinucleotides was only ~2-3. The male bias was similar for non-CpGs and CpGs within CpG islands (Taylor *et al.*, 2006). This is indicative of an alternative mechanism explaining mutations within the CpG dinucleotides.

1.5.2 Bioinformatic markers of germline methylation

In the germline, an incorrect repair of a C→T point mutation of a methylated cytosine is inherited to the progeny. Accumulation of such germline mutations in populations results in C/T (or G/A) single nucleotide polymorphisms (SNP) or the depletion of CpG dinucleotides. These mutations and CpG depletion

patterns can be used to create bioinformatic surrogate markers of germline methylation.

Several researchers have used measurements of CpG density as a surrogate marker for germline methylation. Xing *et al.* used the ratio of CpG vs. non-CpG substitutions in *Alu* elements to estimate the germline methylation of elements of different age (Xing *et al.*, 2004). Kim *et al.* calculated the ratio of observed versus expected frequency of the CpG dinucleotides in several families of repetitive elements (Kim *et al.*, 2007). The CpG depletion should be inversely correlated to the overall germline methylation of the elements. They found that most repetitive elements had substantial CpG depletion apart from those in close proximity to CpG islands. In addition, CpG depletion was greater for two *Alu* element when their orientation was inverted (Kim *et al.*, 2007).

The density of C/T and G/A mutations within human genes was correlated against the density of adjacent repetitive elements to infer their germline methylation (Bjornsson *et al.*, 2006). In paper III of the Ph.D. thesis, a germline methylation marker based on the density of C/T and G/A mutations within the CpG dinucleotide was developed. Xie *et al.* similarly counted clusters of C/T and G/A SNPs in the dbSNP database and used the cluster density as a surrogate marker for germline methylation. Subsequently they counted the number of the cluster adjacent to several subfamilies of repetitive elements and genes to infer their germline methylation (Xie *et al.*, 2009).

1.6 Homologous recombination and its genetic and epigenetic aspect in the human genome (Paper III)

Homologous recombination is the exchange of chromosomal parts by homologous chromosomes at meiosis or mitosis. Homologous recombination in mitosis is a fundamental part of the DNA repair system used to repair double-stranded breaks, lesions that form during DNA replication and due to DNA

damaging agents (Moynahan & Jasin, 2010). Along with genetic mutation, homologous recombination in meiosis is responsible for the majority of genetic variation, thus being an essential part of evolution. Meiotic recombination rate is unevenly distributed around the genome. Recombination hot spots, short areas with high recombination rate, exist in most species. Although several sequence features affect regional recombination rate they do not explain all variation in the rate of recombination. An epigenetic contribution to recombinational rate variability has been suggested. In Paper III, the density of the mSNP bioinformatic marker of germline methylation was correlated with regional levels of homologous recombination.

1.6.1 Homologous recombination

In 1911, Thomas Hunt Morgan suggested that crossing over, or recombination, explained why traits that were thought to be linked occasionally separated (Lobo & Shaw, 2008). Twenty years later, Creighton and McClintock reported meiotic cross over in maize (Creighton & McClintock, 1931) and Stern reported mitotic cross over in Drosophila (Coe & Kass, 2005; Stern, 1931).

The dominant mechanistic model of homologous recombination is the Szostak double-strand break repair (DSBR) model (Szostak *et al.*, 1983), a modification of the original recombination model posed by Holliday (Holliday, 1964). It involves a double-stranded break (DSB) of a single non-sister chromatid followed by a resection of the 5' strand (Fig. 4a and 4b) (Szostak *et al.*, 1983). The 3' overhanging end invades a non-sister chromatid, and DNA synthesis is initiated. In the DSBR, double Holiday junctions are formed as the loop anneals to the homologous chromatid. Two different strand cuts result either in a no cross-over or cross-over molecule (Fig. 4d and 4e) (Szostak *et al.*, 1983). Following this, either DSBR or synthesis-dependent strand annealing (SDSA) occurs (McMahill *et al.*, 2007). In SDSA, the synthesized strands re-

anneals to the original strand, resulting in no crossover (Fig. 4f) (McMahill *et al.*, 2007).

Support for the mechanistic models of recombination in addition to the discovery of the proteins involved in the recombination machinery is largely based on yeast studies (Krogh & Symington, 2004). The SPO11 protein is a key protein that interacts with other proteins to form double-stranded breaks that initiate meiotic recombination (Keeney, 2001). Large recombination nodules including the RAD51 and DMC1 proteins then catalyze the recombination reaction (Handel & Schimenti, 2010). The part of the meiotic break sites that

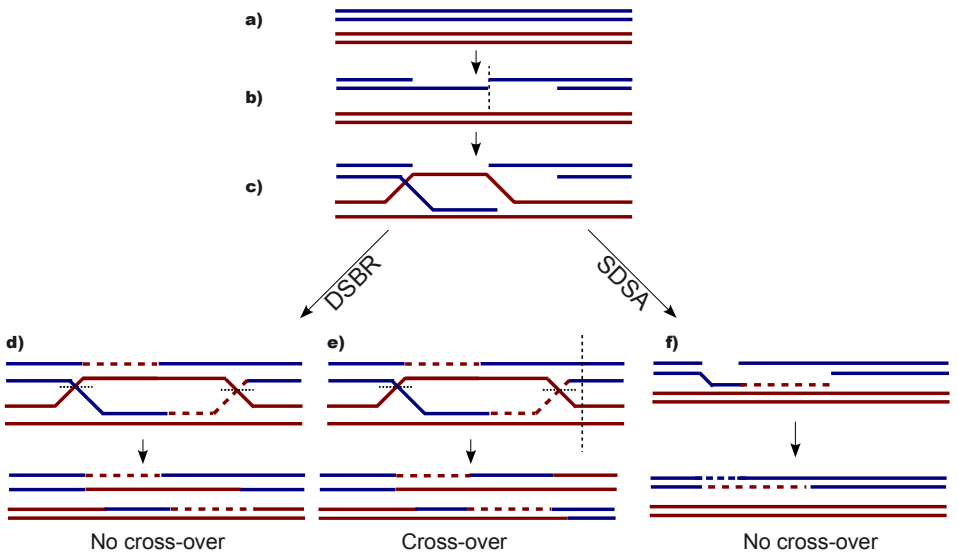


Figure 4: Homologous recombination.

a) Both DNA strands of homologous chromosomes are shown. b) A double-stranded break is initiated followed by a resection of the 5' strand, leaving 3' overhangs. c) The 3' overhang invades a non-sister chromatid and forms a Holliday junction. Two pathways can then be followed, double-strand break repair (DSBR) or synthesis-dependent strand annealing (SDSA) pathway. In DSBR, double Holiday junctions are formed and their resolution and ligation can either result in d) no cross-over or e) crossover molecule. Sites of cleavage and ligation are pictured with broken lines. f) Alternatively, synthesis-dependent strand annealing yields no crossover molecules. Drawn based on Hawley & Walker (2002), Szostak *et al*. (1983) and McMahill *et al*. (2007).

become cross-overs are marked by the MLH1 and MLH3 proteins (Handel & Schimenti, 2010).

Homologous recombination at meiosis must involve a mechanism for accurate separation of sister vs. non-sister chromatids, as recombination between sister chromatids does not generate genetic variation. This mechanism is probably not based on differences in the DNA sequence itself, as inbred experimental animals exhibit efficient recombination (Paigen & Petkov, 2010). DNA methylation (hemimethylation) has been suggested to be a part of this mechanism, but evidence is lacking (Paigen & Petkov, 2010).

Non-allelic homologous recombination (NAHR) affects genomic stability by causing duplications, deletions and inversions (Sasaki *et al.*, 2010). NAHR is the molecular mechanism for many human genomic disorders, such as Charcot-Marie-Tooth disease 1A, Sotos syndrome, congenital adrenal hyperplasia and diGeorge syndrome (Sasaki *et al.*, 2010).

1.6.2 Cross-over interference

Genomic cross-overs are not spaced equally, and there seems to be a minimum distance between the events. This mechanism is called cross-over interference and its molecular basis is largely unknown (Berchowitz & Copenhaver, 2010). Several models have been proposed. According to the mechanical stress model, cross-over stress is relieved in a linear fashion from the breakpoint location after a cross-over event (Kleckner *et al.*, 2004). The counting model suggests that between each cross-over event, a fixed number of non-cross-over events must occur (Foss & Stahl, 1995). The third model, the polymerization model, suggests that cross-over precursor molecules are spread on the chromosome independent of each other, with an even probability of initiating crossover (King & Mortimer, 1990). A binding of the cross-over machinery and crossover

results in inhibition or removal of adjacent cross-over precursor molecules (polymers). However, the molecules are yet to be identified. Suggestions for a potential interference molecule include a protein or an epigenetic factor such as DNA methylation or histone tail modifications (Berchowitz & Copenhaver, 2010).

1.6.3 Recombination in the human genome at the chromosomal level

With improved methodology, several genome-scale recombination maps of the human genome have shed light on the non-random distribution of recombinational events in the human genome. The Marshfield recombination map was created by mapping >8000 short repeat polymorphisms in eight families (Broman *et al.*, 1998). It has a resolution of ~3 cM (centiMorgan, 1% likelihood of cross-over in a single generation) and reveals both individual and sex-specific differences in recombination rates (Broman *et al.*, 1998). The deCODE recombination map was created by mapping >5000 polymorphic microsattelite markers in 146 families (1,257 meioses), increasing the resolution of the map to approximately 0.6 cM (Kong *et al.*, 2002). It confirmed higher recombination rate of females. A multiple linear regression model of sex-averaged recombination rates in 3 megabase (Mb) resolution suggested that the polyA/polyT (multiple A/T bases) ratio, CpG counts and GC ratio were the most significant predictor variables (Kong *et al.*, 2002). Further analysis in 5 and 10 Mb windows revealed that additionally the $W_{n>9}$ (W: A or T, n>9 : more than nine bases in a row) ratio and $R_{n>10}$ (R: A or G) ratio, distance from centromere and chromosome length significantly contributed to the recombination rate (Jensen-Seaman *et al.*, 2004).

1.6.4 Recombinational hot spots

Recombinational hot spots, i.e. areas with a high recombination frequency, have been found in all eukaryotes, although the recombination rate and the size of the

hot spots vary between species (Nishant & Rao, 2006). Hot spots of recombination were first discovered in the lambda phage (McMilin *et al.*, 1974), and have been known in humans for more than twenty years (Chakravarti *et al.*, 1984; Chakravarti *et al.*, 1986). Using laborious polymerase chain reaction (PCR)-based techniques of assaying recombination in sperm samples, more than thirty recombinational hot spots have been discovered in the human genome (Jeffreys & May, 2004; Jeffreys & Neumann, 2005; Jeffreys *et al.*, 2001; Jeffreys *et al.*, 2005; Paigen & Petkov, 2010). These hot spots are 1-2 kilobase (kb) long areas of high recombinational activity and some are located at regions that benefit from large scale genomic shuffling such as the human major histocompatibility complex (Jeffreys *et al.*, 2001).

The presence of population-based fine-scale SNP genotyping increased the discovery rate of recombination hot spots greatly. SNPs form haplotype blocks, areas of non-random association between adjacent SNPs (where the SNPs are in linkage disequilibrium) (Wall & Pritchard, 2003). These haplotype blocks are generally 5-100 kb long. The junctions of those haplotype blocks correspond to recombination hot spots (Paigen & Petkov, 2010). Recombination hot spots found with linkage disequilibrium maps include experimentally discovered hot spots (McVean *et al.*, 2004), and mapped hot spots have been experimentally verified (Webb *et al.*, 2008).

Using a fine resolution (~1 kb) recombination map including more than 25,000 recombination hot spots, it has been shown that approximately 80% of recombination occurs in 10-20% of the sequence (Myers *et al.*, 2005). This suggests that recombination hot spots mediate the majority of recombination activity. Furthermore, several DNA sequence features correlating with recombination have been discovered using these recombination maps. On the scale of 16-1024 kb, only GC content has a consistent positive correlation with

recombination rate (Myers *et al.*, 2006). However, in a some window sizes, exon density (128-512 kb), repeats (16 kb) and CpG density (16 kb) have a negative correlation with the recombination rate (Myers *et al.*, 2006). On the smallest scale (1-8 kb), several DNA motifs have been discovered to have a correlation with regional recombination rate (Myers *et al.*, 2006). The seven nucleotide motif CCTCCCT was found to be significantly enriched within recombinational hot spots (Myers *et al.*, 2005). Its occurrence within two repetitive elements (THE1A and THE1B) resulted in 60% likelihood that the repetitive elements were included in a recombination hot spot, although it explained only a fraction of hot spot activity (Myers *et al.*, 2005). Mild over-representation of CT and GA rich repeats, but under-representation of GC rich repeats, TA rich repeats and certain L1 elements was found near hot spots of recombination (Myers *et al.*, 2005). Recently, this analysis was repeated using a much larger SNP data set from the HapMap II project (Myers *et al.*, 2008). This suggested that a 13-mer DNA motif, CCNCCNTNNCCNC was within about 40% of all human recombination hot spots. The new motif includes the previous 7-mer and is found in several of the hot spots previously located by sperm typing (Myers *et al.*, 2008).

1.6.5 Epigenetic aspect of recombination

Although several features of the DNA sequence contribute to recombination rates, the DNA sequence in *cis* does not provide the whole explanation for recombinational variability. The recombination rate for hot spots of recombination has been shown to differ about 50-fold despite identical DNA sequence in the 15 kb region studied (Neumann & Jeffreys, 2006). Even though the sequence of the chimpanzee genome differs only about 1% from the human genome, almost no recombination hot spots are shared between the two species (Winckler *et al.*, 2005). This indicates that DNA sequence in *cis* is not the sole

determinant of recombination rate. Additionally, regions adjacent to imprinted genes have a high recombination rate of imprinted genes (Lercher & Hurst, 2003; Pàldi *et al.*, 1995; Robinson & Lalande, 1995; Sandovici *et al.*, 2006). Therefore, several authors had suggested that unexplained features of recombination rate might partly be due to differences in epigenetic marks either locally or in distal elements (Neumann & Jeffreys, 2006; Sandovici *et al.*, 2006; Winckler *et al.*, 2005). However, DNA sequence features in *trans* cannot be ruled out.

A regulator of recombinational activity on the mouse chromosome 1 was recently found on chromosome 17 in humans (Myers *et al.*, 2010; Parvanov *et al.*, 2009). This regulator, PRDM9 (PR domain containing 9), is a protein with three domains; 1) a protein-protein binding domain on the N-teminus, 2) a central domain with a histone H3K4 trimethylating properties and 3) a terminal zinc finger domain (Parvanov *et al.*, 2010). The regulator is expressed in male and female meiosis and its knockout results in abnormal meiosis (Parvanov *et al.*, 2010). The zinc binding domain was found to bind to the DNA motifs (CCNCCNTNNCCNC that includes the CCTCCCT motif) previously discovered to be enriched in recombination hot spots (Baudat *et al.*, 2010). Several *PRDM9* alleles coding for different zinc fingers exist (Berg *et al.*, 2010). These different zinc fingers have variable binding affinity to the CCNCCNTNNCCNC motifs (Baudat *et al.*, 2010; Berg *et al.*, 2010). Individuals with different alleles of the *PRDM9* gene have different recombinational activity of hot spots either containing or lacking the 13 bp DNA motif (Berg *et al.*, 2010). Therefore the protein might bind to other sequences. These findings suggest that global genomic recombinational system acting in *trans* exists. The PRDM9 protein has an epigenetic aspect (at least via histone H3K4 trimethylation) (Parvanov *et al.*, 2010), although its binding to

DNA is sequence-specific.

1.7 Repetitive elements in the human genome and defense against their harmful activity (Paper IV)

Approximately 45 % of the human genome is comprised of transposon derived repeats (TDRs) from several families (Lander *et al.*, 2001). Their activity can result in mutations such as insertions, deletions and inversions (Belancio *et al.*, 2008). A global defense system based on DNA methylation has been suggested (Yoder *et al.*, 1997). Any defense system against harmful TDR activity ought to be active in the germline. If a global defense system based on methylation is operative in the germline, the germline methylation landscape should be greatly shaped by the TDR landscape despite its overall hypomethylation. As genome-wide methods measuring DNA methylation generally lack power in TDR-rich areas, our mSNP marker might be suitable for testing this relationship. In Paper IV, we tested the correlation between our mSNP marker of germline methylation and regional density of different TDR subfamilies, with special emphasis on those families with active elements.

1.7.1 Repetitive elements in the human genome

Only a few TDRs in the human genome are currently capable of retrotransposition within the genome, The currently active TDR belong to two TDR families. A total of 21% of the human genome consists of approximately 850,000 long interspersed nucleotide elements (LINE) (Lander *et al.*, 2001). The 6 kb full length element contains two open reading frames (ORFs); the first ORF codes for a protein with chaperon activity and the second one codes for an endonuclease and a reverse transcriptase necessary for transposition (Belancio *et al.*, 2008). The majority of LINE elements belong to the L1 subfamily (17% of the human genome), the largest family currently capable of retrotransposition (Dombroski *et al.*, 1991). L1 elements frequently reside within AT rich

sequences. Recently mobilized L1 elements are, however, located within sequences with a higher GC ratio than the fixed L1 elements (Boissinot *et al.*, 2004). The 20 kb regions adjacent to *de novo* L1 insertions have the genome average GC ratio (Gasior *et al.*, 2007). This is suggestive of a non-biased insertion followed by post-insertional selection (Gasior *et al.*, 2007). The estimated insertion frequency in two recent studies was one insertion in every 95-270 births (Ewing & Kazazian, 2010b) and one insertion every 108 births (Huang *et al.*, 2010).

The second major TDR family is the family of short interspersed nucleotide elements (SINE), representing approximately 13% of the human genome. SINEs are the most numerous TDRs, with approximately 1,500,000 copies of 100-400 bp length present in the human genome (Lander *et al.*, 2001). The major subfamily is the primate-specific *Alu* subfamily (10.6% of the human genome), approximately 300 bp long elements transcribed by RNA polymerase III (Belancio *et al.*, 2008). As SINEs do not code for proteins, the currently active *Alu* subfamily of SINEs is thought to be dependent on the LINE ORF2 retrotransposition machinery for transposition (Dewannieux *et al.*, 2003). However, *Alu* elements are preferentially located within GC- and gene rich regions of the genome (Lander *et al.*, 2001), indicating differential post-insertional selection for *Alu* elements compared to L1 elements (Gasior *et al.*, 2007). Several effects of *Alu* elements suggest that their inclusion might potentially be beneficial to the host genome. *Alu* elements are a part of the regulation of mRNA transcription (Ponicsan *et al.*, 2010). After heat shock, the transcription of Alu is increased (Liu *et al.*, 1995). The transcribed *Alu* RNA binds to polymerase II at repressed genes, and mediates transcriptional repression (Mariner *et al.*, 2008). Also B1, the murine *Alu* analogous element, has been found to be a boundary element regulating transcription of the growth

hormone during organogenesis (Lunyak *et al.*, 2007), suggesting that these elements might participate in specific gene transcription regulation. *Alu* RNA elements included in mRNA can also affect splicing of flanking exons (*Alu* exonization), thereby participating in alternative splicing (Lev-Maor *et al.*, 2008; Lev-Maor *et al.*, 2003; Sorek *et al.*, 2002).

1.7.2 The effects of TDRs on genome stability

The retrotransposition of active TDR can affect the stability of the host genome. Insertions of TDRs are estimated to account for about 0.2% of all human mutations resulting in disease (Kazazian, 1999). This is probably an underestimation given the limitations of PCR based methods to detect TDR insertions and deletions. A substantial threat to genome stability also results from the effects of TDRs on repair mechanisms relying on sequence homology, such as the repair of double-stranded breaks using homologous recombination (Hedges & Deininger, 2007). Currently, 0.17% of all human genetic disease is estimated to result from non-homologous *Alu/Alu* recombination (Callinan & Batzer, 2006). Further genomic rearrangements associated with TDRs involve deletions of adjacent genomic material following a retrotransposition of a TDR, the formation of microsatellites and introduction of double-stranded breaks (Arcot *et al.*, 1995; Callinan *et al.*, 2005). Since methylation can spread from TDRs to adjacent sequences, TDRs could also mediate epigenetic modification affecting gene transcription in metastable epialleles (Morgan *et al.*, 1999).

1.7.3 Genome defense systems against TDR activity

Several species have defense systems against TDR that involve DNA methylation. For example, the repeat-induced point mutation process in *Neurospora crassa* introduces C→T mutations into duplicated sequences, and the adjacent sequences are heavily methylated (Galagan & Selker, 2004). The methylation might mediate epigenetic silencing or render the repeats

hypermutable (Galagan & Selker, 2004). Similarly, transposons are the primary target of RNA-directed DNA methylation in *Arabidopsis thaliana*, where methylation mediates scilencing (Matzke *et al.*, 2007).

Given the abundance of TDRs in the human genome and their destructive potential, the observed number of deleterious events is low. The "Genome Host Defense" hypothesis proposed by Yoder *et al*. suggests that DNA methylation might be a key mediator in genome defense against harmful TDR effects (Yoder *et al.*, 1997). The model was originally supported by the large proportion of DNA methylation located within TDRs in somatic tissues and the generally inhibitory effects of DNA methylation on transcription (Yoder *et al.*, 1997). Later, methylation has been demonstrated to control L1 transcription in vitro (Hata & Sakaki, 1997). In addition, a deletion of the *Dnmt3L* (DNA (cytosine-5-methyltransferase 3-like) gene in mice results in both loss of L1 and LTR methylation as well as a corresponding increase in their transcription (Bourc'his & Bestor, 2004). The Genome Host Defense hypothesis has been challenged. A defense system against TDRs ought to be critical in the germline. However, several TDRs are actively transcribed and hypomethylated in the germline, arguing against a global defense system (Bird, 1997). Also, the most striking examples of DNA methylation as a host defense system are limited to L1 elements and the IAP (Intracisternal A-particle) element in mouse (Zamudio & Bourc'his, 2010).

Several other genome defense mechanisms against TDR activity have been described (Zamudio & Bourc'his, 2010). RNA editing enzymes alter the properties of RNA transcripts from TDRs, thereby stopping their potentially harmful activity (Zamudio & Bourc'his, 2010). These enzymes include the APOBEC3 family that inhibit TDR retrotransposition by deamination of cytosine into uracil (Belancio *et al.*, 2008). Some of the family members

(APOBEC3A and B) inhibit transposition of both L1 and *Alu* elements (Bogerd *et al.*, 2006) while others (APOBEC3G) are selective for inhibition of *Alu* retrotransposition (Hulme *et al.*, 2007). Short interfering RNA (siRNA) molecules containing L1 sequences (and other TDRs) are produced by oocytes (Tam *et al.*, 2008), although their importance in TDR defense is unknown (Zamudio & Bourc'his, 2010).

Recently, a germline-specific defense system against TDR activity based on small RNA molecules has been described in the human and mouse. Members of the Argonaute protein family, the PIWI (P-element induced wimpy testis) proteins, form the basis of this defense system (Zamudio & Bourc'his, 2010). Mice homozygous for mutations in the PIWI subfamily protein genes (*Mili, Miwi2*) are sterile and do not destroy L1 and IAP transcripts (Aravin & Bourc'his, 2008; Carmell *et al.*, 2007), suggesting their importance for germline stability via suppression of TDRs. The PIWI proteins bind to piRNA (PIWI-interacting RNA) elements. These elements are 24-30 nt long RNA products that are found in clusters on most human and mouse chromosomes (Girard *et al.*, 2006). Their location is not dependent on repeat or gene density (Girard *et al.*, 2006).

After the global demethylation phase, TDRs are be transcribed. The mRNA elements are efficiently cleaved into piRNAs by the PIWI proteins (Zamudio & Bourc'his, 2010). Potentially, all mRNA can be cleaved. However, both the sense and antisense transcript of TDRs is transcribed, resulting in a multiplication cycle (ping-pong mechanism) increasing their relative abundance compared to mRNAs from genes. In particular, LINE and long terminal repeats (LTR) are in abundance during this period and are subsequently dominant in the piRNA transcripts (Aravin & Bourc'his, 2008). In addition to the destruction of TDR transcription by their cleavage, a complex including PIWI protein bound

to TDR piRNAs moves into the nucleus (Aravin *et al.*, 2009) where the piRNAs direct methylation of their corresponding TDR elements by recruiting Dnmt3L / Dnmt3A methyl-transferase proteins by a currently unknown mechanism (Zamudio & Bourc'his, 2010). This mechanism thereby links a protein- and an RNA-based TDR defense mechanism and DNA methylation, albeit mostly for LTR and L1 elements.

1.8 Imprinted genes and their metabolic effects (Paper V)

An important function of DNA methylation in the human genome is the maintenance of parent-of origin specific expression pattern of imprinted genes. To date, around 65 genes are known to be imprinted in at least one human tissue although they are presumed to be more common. Loss of imprinting of imprinted genes is associated with the appearance of several human diseases. Several hypotheses on the physiological function of imprinted genes exist. Amongst those is the Haig's parental conflict hypothesis of differential metabolic effects of maternally and paternally imprinted genes.

Discovery of imprinted genes has traditionally been based on mapping loci associated with diseases with a distinct pattern of heritage. In paper V, we applied methods of systems biology on an *in silica* reconstruction of the human metabolic network to simulate the effects of differences in expression of imprinted genes. This can be used to test Haig's hypothesis. Additionally, if a particular systemic response results from expression perturbations of imprinted genes, this might be used to predict imprinted genes.

1.8.1 Imprinted genes

Imprinted genes have an allele-specific expression pattern based on parental origin. To date, there is an experimental evidence of imprinting for 64-68 genes in the human genome in at least one tissue (www.geneimprint.com and

www.igc.otago.ac.nz, accessed 2010/09/18). This is likely an underestimation, and sequence pattern algorithms applied to the human genome have recently suggested that 156 additional genes are likely to be imprinted (Luedi *et al.*, 2007). Furthermore, novel analysis of five single nucleotide polymorphisms (SNPs) associated with cancer and type II diabetes demonstrated a parent-of-origin based association with several phenotypes (Kong *et al.*, 2009). This suggests that imprinted genes are more common than originally thought.

The majority of known imprinted genes cluster on several chromosomes. This might reflect a knowledge bias, since uniparental disomies that involve chromosomes with clusters of imprinted genes are likely to result in more severe phenotypes and thus more likely to be discovered. The imprinting clusters often contain imprinting control elements that affect the expression of one or more imprinted genes simultaneously (Reik & Walter, 2001). Two examples of such regulation elements lie within the 11p15.5 and the 15q11 imprinting clusters in the human genome (Verona *et al.*, 2003). The cluster of chromosome 11 includes the paternally expressed *IGF2* gene and the maternally expressed *H19* gene. Uniparental disomy of the region can cause a loss of imprinting (LOI) of the *IGF2* gene resulting in Beckwith-Wiedemann syndrome (Maher & Reik, 2000). Loss of imprinting of the *IGF2* gene has been shown in colon cancer (Cui *et al.*, 1998) and individuals with a family or personal history of colon cancer have a significantly higher odds ratio of having LOI of *IGF2* in colon mucosa (Cui *et al.*, 2003).

An interesting but unexplained phenomena of imprinting clusters is their increased frequency of homologous recombination. A sex-specific pattern of recombination was observed near the imprinting clusters of chromosomes 11 (Pàldi *et al.*, 1995) and 15 (Pàldi *et al.*, 1995; Robinson & Lalande, 1995) including both male-specific and female-specific recombination hot spots. This

has been expanded with larger data sets of imprinted genes and more dense recombination information. Using pedigree-based recombinational mapping a higher female recombination rate was found adjacent to imprinted genes (Lercher & Hurst, 2003). High-density recombination rate analysis using linkage disequilibrium data also found a higher recombination rater in regions containing imprinted genes (Sandovici *et al.*, 2006).

Parent-of-origin expression of the majority of imprinted genes is stably maintained with differential methylation of the parental alleles, exemplified by the effects of mutations in methyltransferases on the expression of imprinted genes (Bourc'his & Bestor, 2004; Kaneda *et al.*, 2004). The majority of imprinted genes contain differentially methylated regions (DMRs) (Reik & Walter, 2001). For example, the methylation of a DMR between the *IGF2* and *H19* genes controls their expression. Methylation of the paternal allele results in the expression of *IGF2* gene enhanced by sequences downstream of the *H19* gene. The maternal allele has an insulator factor bound to the unmethylated DMR box. The binding of the insulator to the DMR results in blockage of *IGF2* expression and subsequent enhancement of *H19* expression by the enhancer (Hark *et al.*, 2000).

1.8.2 The metabolic effects of imprinted genes

The parental intergenome conflict theory was proposed by Haig in 1991 to explain the physiologic implications of imprinted genes. It suggests that imprinted genes influence growth based on parental origin; paternally and maternally expressed genes increase and decrease pre- and postnatal growth, respectively (Moore & Haig, 1991). This was originally supported by the observed phenotypes from disruption of two imprinted genes in mice (Haig & Graham, 1991). With more clinical phenotypes from uniparental disomies recognized, an analysis of the phenotypes in light of Haig's theory found that

only 7 out of 15 imprinted had a metabolic profile as predicted by the parental conflict theory (Tycko & Morison, 2002). However, the metabolic phenotypes resulting from changed expression of imprinted genes is only known for a handful of genes, so the available experimental evidence is limited. Therefore alternative methods must be sought, such as simulating the metabolic effects of changed expression of imprinted genes using methods of systems biology.

1.8.3 A primer on systems biology and reconstruction of the human metabolism

Although a formal definition of systems biology is lacking, systems biology is generally described as a multidisciplinary field involving description and analysis of molecular components of a biological system and their interaction (Palsson, 2009b; Palsson, 2006). It involves studying complex biological information by applying computational and mathematical methods. Three of the currently most commonly applied aspects of systems biology methods include: a) the analysis of complex high throughput data (such as microarrays, and high throughput sequencing data), b) reconstruction and analysis of computational models for complex biological systems (such as metabolism and transcriptional regulation) and c) predicting and engineering of biological systems based on the reconstructed biological systems (Kitano, 2002). Since metabolism is one of the best studied systems and data generated from decades of research exists, metabolic systems biology has been at the forefront of methodological development within systems biology (Palsson, 2009a).

Reconstruction of a metabolic network can follow either top-down or bottom-up approach. The top-down approach involves a computational analysis of a high throughput data (genomic sequence, gene transcription sequence etc.) to determine the most statistically likely connections between components of the network. The bottom-up approach is a laborious process that involves

manually surveying existing literature on components within the network and their connectivity. Knowledge gaps are highlighted by this approach, but they can be automatically filled after reconstruction to provide a more complete reconstruction (Palsson, 2009b; Palsson, 2006; Thiele & Palsson, 2010).

A reconstructed metabolic network involves a mathematical representation of each reaction included in the model. The collection of reactions is termed the stoichiometric matrix **S**. Other information matrices can be linked to this network (such as information on enzyme isozymes, gene-protein information, metabolite matrix, etc.). To convert the reconstruction into a functioning model, a set of constraints must be applied to each reaction in **S**. These constraints include the conservation of energy and mass, pH and temperature. In addition, the flux of molecules through each reaction and its reversibility is constrained by applying flux range limits (Palsson, 2009b; Palsson, 2006).

Environmental and genetic perturbations can be simulated using the mathematical form of a reconstructed metabolic network. Methods of linear algebra are applied to obtain a solution space of flux vectors **v** fulfilling the criteria of $\mathbf{S} \cdot \mathbf{v} = 0$ under assumption of a steady state (i.e. no metabolites are accumulated or depleted) (Palsson, 2009b). A range of methods have been developed to assay the model properties and perform various simulations (Price *et al.*, 2004), including the analysis of gaps, phenotype simulation, analysis of metabolic network evolution and bioengineering (Oberhardt *et al.*, 2009). These methods have been included in the COBRA (Constraint-based reconstruction and analysis) toolbox that allow the simple and efficient study of metabolic models on a personal computer (Becker *et al.*, 2007).

Three reconstructions of the human metabolism exist. Two

reconstructions, the HumanCyc (Romero *et al.*, 2005) and the Edinburgh Human Metabolic network (Hao *et al.*, 2010; Ma *et al.*, 2007) were built using an automated top-down approach. The Human Recon 1 metabolic reconstruction, in contrast, is a bottom-up approach (Duarte *et al.*, 2007). It was compiled by six researchers in an iterative manner, assigning each reaction within the network a confidence score and applying rigorous quality controls to the input data. Gaps were analyzed and filled to produce a functional model of human metabolism. Before completion, the resulting model was able to correctly simulate 288 core metabolic functions of human metabolism (such as creation of all non-essential amino acids from essential amino acids) (Duarte *et al.*, 2007). The human Recon 1 model accounts for 1496 genes coding for 2004 metabolic proteins. It has 1510 transport and exchange reactions in addition to 2233 biochemical reactions operating in eight cellular compartments (Duarte *et al.*, 2007).

Several applications have been published since the release of Recon 1. Expression profiling data has been applied to Recon 1 to produce tailored metabolic networks for ten human tissues (Shlomi *et al.*, 2008). Mapping human diseases onto Recon 1 revealed functional relationships, possibly explaining disease co-morbidity (Lee *et al.*, 2008). Recently, gene homology data was used to create a functional reconstruction of mouse metabolism based on Recon 1. The resulting model was used to address the phenotype prediction properties of compartmentalized metabolic models of complex organisms (Sigurdsson *et al.*, 2010).

2 AIMS

The general aim of the thesis was the development and application of biological and bioinformatic assays to study several aspects of DNA methylation in the human genome. The specific aims were:

1. To analyze the sequence specificity of restriction endonucleases suitable for measurements of whole-genome methylation (Paper I).
2. To use both whole-genome methylation assay and site-specific methylation to measure the longitudinal change in somatic DNA methylation in two populations, and a familial component of the conservation of DNA methylation (Paper II). This might support acquired changes in epigenetic marks, an important prerequisite for an epigenetic model of the pathogenesis of complex diseases.
3. To create and validate a novel bioinformatic assay of human germline DNA methylation (Paper III).
4. To apply the bioinformatic assay to study the relationship between:
   (a) The germline DNA methylation and homologous recombination (Paper III).
   (b) The link between germline DNA methylation and transposable elements in the human genome (Paper IV).
5. To develop and apply methods of systems biology to study the biological role of imprinted genes in the human genome (Paper V).

3 MATERIALS AND METHODS

3.1 Analysis of methylation sensitive restriction endonucleases suitable for whole-genome methylation analysis in the human genome (Paper I)

A list of candidate restriction endonucleases suitable for whole-genome methylation analysis in the human genome was created by two means: 1) From McClelland *et al*. (McClelland *et al.*, 1994) and 2) from the REBASE database of restriction endonucleases (Roberts *et al.*, 2010) (Accessed 2010/09/08). Criteria for inclusion into study were: i) two available isoschizomeric restriction endonucleases differing in 5-methylcytosine sensitivity; ii) Target sequence includes either CG, representing CpG methylation in all cell types or CWG (W=A or T), representing the recently discovered non-CpG methylation in embryonic stem cells (Lister *et al.*, 2009).

The properties of the target sequences selected were analyzed by writing software (Restrictionsearch 1.0 and CpGsearch 1.0) in the JAVA programming language searching for the target sequences of the restriction endonuclease pairs used, in addition to the CpG dinucleotide and CWG trinucleotide. All programs were validated by comparing the output file from programs to a manual search for the target sequences and CpG in a 500 bp modified DNA sequence. Program runs counting the target sequence frequency in 500 kilobase (kb) windows for the entire human genome (NCBI36, hg 18, March 2006) were performed on a personal computer. Data tables on CpG islands (Gardiner-Garden & Frommer, 1987), exons (Hsu *et al.*, 2006), and repeated elements (Jurka *et al.*, 2005) were downloaded from the UCSC table browser (Kent *et al.*, 2002) in the NCBI36/hg18 version. The target sequence frequency for all endonucleases was determined in each data set using modified programs for a genome-wide analysis.

3.2 Change in somatic DNA methylation over time (Paper II)

3.2.1 Samples

Initial analysis tested the effect in a cross-sectional and longitudinal manner. For the cross-sectional study, DNA isolated from whole blood of 84 individuals participating in AGES (Age, Gene/Environment Susceptibility Reykjavik study) (Harris *et al.*, 2007). The sample included 35 individuals with type II diabetes, a common acquired metabolic disorder, and 49 control individuals. For the longitudinal study, 111 individuals from the AGES cohort were chosen for measurement of whole-genome methylation. Of those, 61 individuals were chosen based on the amount of DNA available from two different time points, in addition to 50 individuals alive with a diagnosis of cancer at adulthood. This allowed us to compare methylation changes in a cohort with a disease that presents at late age with controls.

For replication of the longitudinal result in an independent cohort, DNA from whole blood from the Salt Lake City CEPH pedigrees (680 individuals from 48 three generation families) (Sandovici *et al.*, 2003) were used. Of those, 126 individuals from 21 families donated a second blood sample and had sufficient DNA from both time points for the analysis. Seven individuals were sampled repeatedly (2-4 times) over 30 days to test the short-term stability of DNA methylation in stored DNA from whole blood. All individuals signed an informed consent form, and the study was approved by the Icelandic National Bioethics committee (FS-04-001), the Icelandic Data Protection Authority (2005/497) and the institutional review boards of the Johns Hopkins Bloomberg School of Public Health and the University of Utah.

White blood cells were isolated from whole blood using the buffy coat method. DNA was extracted from the total extract of white blood cells using phenol-chloroform extraction.

3.2.2 The LUMA assay for analysis of global methylation

The luminometric methylation assay (LUMA) is a whole-genome methylation assay. It is based on quantifying the cut of isoschizomeric restriction endonucleases differing only in methylation sensitivity (Karimi *et al.*, 2006b) (Fig. 5). The *Msp*I/*Hpa*II restriction endonucleases were used in the measurements. They both cleave the target sequence CCGG, leaving a GC 3' overhang. However, *Hpa*II is sensitive to methylation at the fifth carbon of the second cytosine whereas *Msp*I is insensitive to methylation. As a control for DNA input and restriction activity, a double digestion with *Eco*RI was done. *Eco*RI cleaves at the GAATTC sequence, leaving TTAA 3' overhang.

DNA was quantified using fluorescence measurements of PicoGreen

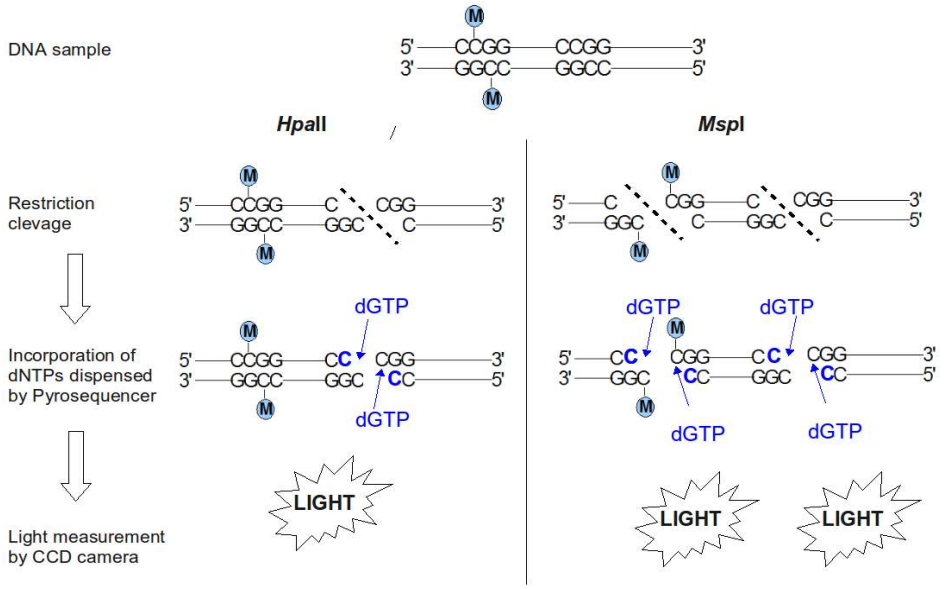


Figure 5: Principles of LUMA.

Following restriction by isoschizomer endonucleases differing in methylation sensitivity, pyrosequencing is performed to quantify the endonuclease cut. Incorporation of dNTPs matching the overhanging ends results in the generation of light by the Pyrosequencing enzyme system. The amount of light corresponds to the amount of overhanging ends. The difference in light between the isoschizomeres is proportional to the overall methylation of the restriction target sequence.

(Invitrogen, USA) reagent binding to double-stranded DNA. Following this, 1 µg of DNA was divided into two 0.2 mL PCR reaction tubes and a double endonuclease digestion at 37° C for four hours performed with either *Hpa*II / *Eco*RI (tube A) or *Msp*I / *Eco*RI (tube B) (New England Biolabs, USA) using the Tango buffer (Fermentas, USA) for optimal reaction conditions for all endonucleases. A total of 5 IU of each endonuclease was used in a total volume of 10 µL. Following the digestion, the amount of overhangs (corresponding to amount of DNA cut) was quantified by Pyrosequencing (Biotage, Sweden) (Fig. 5). The product from the restriction endonuclease reaction was mixed with 12 µL Annealing buffer (Biotage, Sweden) and 3 µL put in each well of a 96 well Pyrosequencing plate and placed in a Pyrosequencer loaded with the enzyme reaction mix (containing DNA polymerase and Luciferase) and deoxynucleotide triphosphates (dNTPs). The base dispensation sequence was modified from the original LUMA protocol as follows. The base dispensation order which the Pyrosequencer used was GTGTGTCACACATGTGTGTGTG. Using this order, the first six dispensations of guanine dNTP and thymine dNTP will bind to degraded DNA and not overhangs left by the restriction endonucleases. The next six dispensations will bind to the overhangs in addition to degraded DNA. The light peaks formed at dispensation number 13 *(Eco*RI, $v_{13,A}$ and $v_{13,B}$) and 14 (*Msp*I / *Hpa*II, $v_{14,A}$ and $v_{14,B}$) were collected and used for analysis.

For each sample, the ratio of *Hpa*II and *Msp*I digestion was calculated using the absolute measured luminometric values from the Pyrosequencer for tubes A (*Hpa*II / *Eco*RI digestion) and B (*Msp*I / *Eco*RI digestion) using the formula:

$$r = \frac{v_{14,A} / v_{13,A}}{v_{14,B} / v_{13,B}}$$

The *r* value was converted into absolute percentage methylation using a standard curve. The curve was created by methylating lambda phage DNA with *Sss*I CpG methylase (New England Biolabs, USA) and mixing 100% with 0% methylated lambda phage DNA in various proportions of methylation (0%, 25%, 50%, 75% and 100%). The LUMA measurement of each sample was done in triplicate. The results were used to create a standard curve for conversion of measurement of endonuclease cut into whole-genome methylation percentage. For a further validation of the LUMA assay, DNA isolated from the Het116 and DKO Dnmt1 knockout cell lines was used as controls for the whole-genome luminometric methylation assay (LUMA), as the DKO cell line has significantly lower global DNA methylation compared to the Het116 parent cell line (Rhee *et al.*, 2002). Each sample was measured in triplicate.

3.2.3 Bisulfite microarray analysis of individual genes

A subset of 41 individuals was chosen for a microarray assessment of methylation changes within gene promoters between the two time points. Selection was based on results from the global methylation analysis (17, 5 and 19 individuals with the greatest loss, least change or greatest gain of methylation, respectively). A total of 0.5 µg of DNA was bisulfite-treated with the EZ DNA methylation kit (Zymo Research, USA). Bisulfite treatment of DNA converts unmethylated DNA into uracil while methylated DNA remains unchanged. The converted DNA was applied to the Illumina GoldenGate Methylation Solution plate, using the Cancer Panel I platform (Illumina, USA). The microarray determines the methylation status of 1505 CpGs selected from the promoter regions of 807 genes in the human genome. It has been validated by both direct bisulfite sequencing and methylation-specific PCR (Bibikova *et al.*, 2006).

3.2.4 Statistical analysis

Intra-individual changes in methylation over time in the Icelandic data set was assessed by permutation. For each individual, six methylation measurements were selected randomly (three replications of two time points) and the difference between the two time points calculated. The distribution of 10,000 permutations performed was then compared to the observed distribution of methylation difference. Furthermore, the ratio of variance in methylation across all six measurements over the variance within each time point were calculated, and compared between the permutations and the observed values. Permutation testing was performed using the SAS statistical software (version 9.1).

Heritability estimate for methylation difference was estimated by calculating the change between time 2 and time 1 adjusted for time 1 values. Residual values at time 2 were used for maximum likelihood estimate of heritability. This was done with variance components models in the ASSOC program of the SAGE statistical package (version 5.2.0).

For microarray analysis, readings from different samples were quantile-normalized after pooling all raw data. The normalized values were then separated again and log ratios of intensities for methylated/unmethylated (red(cy5)/green(cy3)) calculated. Significance testing for changes in methylation across time points for those individuals of interest was done with *t*-testing of the log ratios. However absolute differences in percentage methylation were reported, as they are more easily interpreted than the logarithmic transformation.

3.3 A bioinformatic assay of human germline DNA methylation and its correlation with homologous recombination and TDR subfamilies (Paper III and IV)

3.3.1 Definitions

A methylation-associated SNP (mSNP) was defined as any C/T (corresponding to a C-T polymorphism on the read strand) or G/A (corresponding to a C/T polymorphism on the opposite strand) polymorphism with a 3' guanine base (i.e. within a CpG dinucleotide). In any given database of SNPs, the mSNP subset therefore includes all possible methylation-associated mutations occurring within the CpG dinucleotides. To increase specificity of the mSNP method in the genome-wide associations, an additional criteria was used to define $mSNP_{GENOME}$ as all mSNPs within the HapMap data set with evidence that the ancestral allele was either C or G requiring that the transition causing the SNP was a C→T transition. This should theoretically increase the likelihood that it was due to methylation.

To correct for possible confounders in the germline methylation map, a methylation index (MI) for each 500 kilobase window of the human genome was plotted. MI was defined as:

$$MI = \frac{N_{mSNP}}{N_{CpG} \cdot N_{SNP}}$$

Where N_{mSNP} is the number of mSNP in the window, N_{CpG} is the number of CpG dinucleotides in the window, and N_{SNP} is the total number of SNPs in the window. The numerator reflects the observed number of mSNPs while the denominator reflects a size directly proportional to the expected number of mSNPs. The window size (500 kb) was chosen so sufficient number of SNPs were available for each window. This reduces the error in the estimation of methylation based on mSNP counts. The benefit of using the mSNP approach

over e.g. observed/expected CpG ratio is that the mSNP approach does not rely on the evolutionary conservation of methylation. However, the approach inherently inherently risks losing methylation variability on a smaller scale given its more limited resolution.

3.3.2 Data sets

The entire second release (July 2006) of the non-redundant genotype data set for all 22 autosomal chromosomes was downloaded from the International HapMap Consortium (www.hapmap.org) (HapMap Consortium, 2007). To test if the mSNP subset was subject to positive selection, the integrated haplotype score (www.haplotter.uchicago.edu/selection, downloaded on 2009/01/01) (Voight *et al.*, 2006) was compared between the mSNP and non-mSNP subset. Furthermore, a derived alleles data set was used to determine ancestral SNPs based on comparison with the chimpanzee and macaque genomes (Thomas *et al.*, 2007).

The entire ENCODE (ENCyclopedia of DNA elements) data set for 10 500 kb human genome regions was downloaded (www.hapmap.org/downloads/encode1.html, accessed 2008/07/11) (The ENCODE Project Consortium, 2004).

For both the genome-wide and the ENCODE data sets, all four populations included were pooled after searching for mSNP, and redundant polymorphisms erased prior to further analysis. The analyzed data set therefore contained a single copy of every mSNP available in at least one population.

The entire human genome sequence (NCBI build 35, UCSC hg 17) in addition to the ENCODE regions (NCBI build 34, UCSC hg 16) was downloaded from the UCSC genome browser. Information on recombination rate and recombination hot spots was downloaded from the HapMap website

(www.hapmap.org). Data tables on CpG islands (Gardiner-Garden & Frommer, 1987), exons (Hsu *et al.*, 2006), and repeated elements (Jurka *et al.*, 2005) were downloaded from the UCSC table browser (Kent *et al.*, 2002) in the appropriate genome release.

A list of experimentally imprinted genes was downloaded from the GeneImprint website (www.geneimprint.org, accessed at 2009/03/01). A list of computationally predicted imprinted genes was downloaded from Luedi *et al*. (Luedi *et al.*, 2007). A list of housekeeping genes was downloaded from Eisenberg *et al*. (Eisenberg & Levanon, 2003).

The entire human epigenome data set (HEP) containing bisulfite sequencing results of 2524 amplicons from three chromosomes in 12 different tissues was downloaded from the HEP web site (Eckhardt *et al.*, 2006).

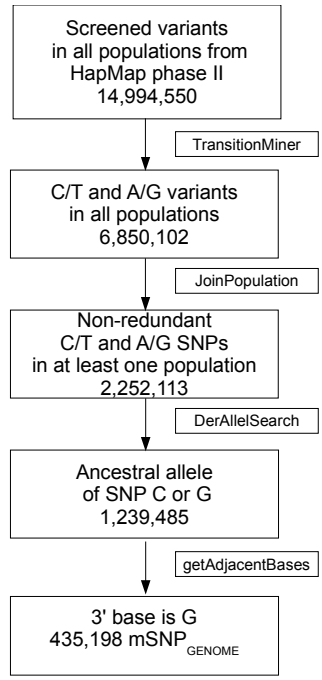


Figure 6: Data flow in the creation of the mSNP data set.

Large boxes describe SNP statistics, small boxes indicate programs used for data processing.

3.3.3 Programs and data flow

All data processing programs were written in JAVA programming language (Sun, USA) and thoroughly tested prior to usage by comparing automatic and manual data processing. Data processing using the scripts were performed either on a personal computer or a cluster computer. A local CENSOR server (version 4.2) (Jurka *et al.*, 2005) was set up on a cluster computer to search for

repeats within submitted sequences. Table 1 describes the various features of the major data processing programs used (several versions were created of each version to fit various data sets). Figure 6 describes the data flow from the raw data to the finalized $mSNP_{GENOME}$.

Table 1: Name and description of major programs written for data handling.

SNP programs	
Name	**Function**
Transition-Miner	Finds C/T and G/A SNPs
JoinPopu-lations	Joins all HapMap populations and erases redundancies
Prepare-MEtMutMatrix	Counts mSNPs in genome windows
GetAdjacent-Bases	Locates bases next to SNP or genomic location
getiHScore	Gets integrated haplotype score for SNPs of choice
SNPsearcher	Searches for SNPs adjacent to genomic location
Genome sequence/tables programs	
Name	**Function**
prepare-repeatsmatrix	Counts repeats of any subfamily in genome windows
prepare-CpGisland-matrix	Counts CpG islands in genome windows
prepareGene-DensityMatrix	Counts exon bases in genome windows
Recombination-ratesWA	Calculates recombination rate of given location in genome
ExonSplitter	Splits exons to give a single copy of each exon in the genome
ChromoBand	Estimates Giemsa banding of genome windows
CpA/C/G/T-search	Counts dinucleotides in genome windows

Data handling and statistical programs	
Name	**Function**
TableSplitter	Splits any UCSC table according to chromosome
Genome-Splitter	Splits genome sequence into tiny fragment to speed computing
Create-Statistical-Matrix	Combines searches for genomic features into single matrix, performs lognormal transformation, removes windows with sequence gaps.
TableRepeats-Splitter	Split repeats table based on repeat subfamilies
ResultMatrix	Creates results matrix from Censor searches of repeats
pullout-Low/Average/High	Selects amplicons from Human Epigenome Project data set with certain average methylation values
MethyMap	Data visualization program
RandomGene	Creates a random set of genes.
RandomPoly	Creates a random set of SNPs
OutlierSelector	Selects a subgroup of the microarray data representing outlier methylation values.
CreateCensor-OutputFile	Creates a file for repeats analysis by the CENSOR server
MonteCarlo-Simulation	Randomly divides the TDR searches for HEP amplicons into two groups 10,000 times and calculates TDR statistics between the two groups. Creates 10,000 random samplings of amplicons into two groups and calculates TDR statistics for the randomly created groups.

3.3.4 Statistical analysis

Data on mSNP counts, SNP counts, GC ratio, CpG density and gene density in addition to the density of different TDR subfamilies was transformed with Box-Cox lognormal transformation prior to multiple linear regression analysis. For genome-wide association between mSNPs and sequence features, correlations were done in four different window sizes (125 kb, 250 kb, 500 kb and 1000 kb). For the ENCODE regions, two different window sizes were used (25 kb and 50 kb). The genome was divided into non-overlapping windows of the appropriate size. For each window, programs counting mSNP, total number of SNPs, CpG dinucleotides, GC ratio, observed/expected CpG ratio,number of bases within exons, DNA motifs and number of bases within TDR subfamilies were written. These counts were then used to calculate the exon and TDR proportion in each window. Windows containing sequencing gaps were removed prior to analysis.

Single and multiple correlations were done with Spearman's ranked correlation, since all variables failed tests of normality. Multiple linear regression of recombination rate, proportion of recombination hot spot and various TDR subfamilies as a function of other assayed genomic and epigenomic variables (mSNPs density and germline tumor methylation) was performed choosing predictor variables based on results from single/multiple correlations, available literature and an automatic stepwise backward method. To compare the contribution from each predictor variable to the model, the standardized β was used. It reports the number of standard deviations that the outcome variable will change as a result of one standard deviation change in the predictor variable.

The average CpG methylation was calculated for each amplicon. Average methylation between the group of amplicons within recombination hot spots and the group of amplicons not within recombination hot spots were

compared with Welch's *t*-test. To compare searches for repeats flanking hyper- and hypomethylated amplicons (>80% and <20% methylation, respectively) the amplicons were randomly split into two groups 10,000 times, using an in-house Monte Carlo simulation program. For each randomization, the difference in TDR proportions were calculated between the two groups and the observed difference compared against the distribution of values from randomized data to estimate a *P* value.

A *P* value less than 0.05 was considered statistically significant. To correct for multiple testing when n tests were performed, a Bonferroni adjusted *P* value of 0.05/n was considered statistically significant. Several of the genomic values are inter-correlated, such as GC ratio and CpG density. Therefore, the Bonferroni adjustment is likely too stringent, resulting in a greater likelihood that a false null hypothesis is not rejected (Type II error). A part of the multiple linear regression was done in SPSS version 15, all other statistical analysis and figure preparation was done in the R statistical package, versions 2.5-2.11.

3.4 Systems biology approach to study the function of imprinted genes in humans (Paper V)

3.4.1 Definitions and data preparation

Experimentally verified imprinted genes were listed from two public databases (www.geneimprint.com and www.otago.ax.nz/IGC, accessed 2008/08/26). Computationally predicted imprinted genes were from Luedi *et al*. (Luedi *et al.*, 2007). Both lists were crossed against the list of 1,496 metabolic genes in the Recon 1 reconstruction of human metabolism (Duarte *et al.*, 2007).

3.4.2 Setup of human metabolic network model

The human metabolic network reconstruction, Recon 1, was used allowing all internal and external metabolic reactions. However, the only unconstrained

cellular uptake was of vital amino acids, vital fatty acids, glucose, molecular oxygen, protons, sulfate and phosphate. The model was set up to optimize for cellular biomass (Sheikh *et al.*, 2005).

3.4.3 Flux balance and flux variability analysis

A mathematical form of a reconstructed network (such as the human metabolic network) including constraints on each reaction (representing maximal and minimal flux and reversibility of the reaction) can be used for network properties calculations. Flux balance analysis (FBA) calculates a set of fluxes in a metabolic network in a mathematical form that maximizes a given biological objective (such as biomass) (Orth *et al.*, 2010). Several derived applications use FBA to analyze metabolite flow through a model. Flux variability analysis (FVA) uses quadratic or linear programming methods to find for each reaction in the reconstruction the absolute maximum and minimum flux values that result in the optimal solution of the objective function selected (Mahadevan & Schilling, 2003). It therefore gives a boundary box of the solution space of fluxes resulting in the optimal solution for each reaction in the model (Fig. 7e). The FVA solutions of a particular metabolic network can be compared for two different conditions (for example wild type vs. gene knockout) to study the change in solution space boundaries for the pertubation.

Both FBA and FVA have been successfully applied to prediction of metabolic phenotypes in microbes (Feist & Palsson, 2008). For example, FBA analysis has an 85% success in predicting lethality from knockouts of 555 genes (Shlomi *et al.*, 2005). Data on the predictive capability of FBA and FVA in models of multicellular organisms is limited. However, FVA analysis of a reconstruction of mouse metabolism based on the human Recon 1 indicated that out of 17 genes with experimental data and prediction of essentiality by FVA, 14 were indeed essential (Sigurdsson *et al.*, 2010). Also, FVA correctly

predicted metabolic changes resulting from mutations in the LPL gene (Sigurdsson *et al.*, 2010). No genes involved in the human Recon 1 had a clear anabolic or catabolic phenotype assigned to them in the OMIM database.

For analysis of the metabolic effects of imprinted genes, a FVA with no additional constraints was first evaluated to get maximum (F_{max}) and minimum (F_{min}) flux resulting in the optimized objective function (biomass). Three differential conditions were created for each imprinted gene, simulating three epigenotypes (Fig. 7a). For the normal epigenotype (single allele expression epigenotype II), flux through all reactions coded by the imprinted gene were set to ½ F_{max}. For the simulation of neither allele expression (epigenotype I), the flux through all reactions coded by the imprinted gene were set to zero. For the simulation of biallelic expression, the flux through all reactions coded by the imprinted gene was set to F_{max}.

For each imprinted gene simulated, FVA was performed for all three epigenotype simulations. The FVA results from each abnormal epigenotype (I and III) were then compared against the FVA result from the normal epigenotype (II). Each of the 3,311 reaction was assigned a status of increased flux capacity, decreased flux capacity and unchanged flux capacity (Fig 7e). Increased and decreased flux capacity correspond to increased and decreased flux of molecules through the reaction, respectively.

For each of the 97 subsections of human metabolism in Recon 1, the number of reactions with decreased and increased flux capacity were counted. The counts were compared against even probability of increased and decreased flux capacity using a single value Chi-Square test with one degree of freedom. Those sections with a significant difference in reactions with increased and decreased flux capacity given multiple testing ($P< 0.05/97$) were considered

contributing to phenotype. As FVA on the whole human metabolic reconstruction is very computationally demanding, permutation maintaining autocorrelation structure was not possible to assess significance.

3.4.4 Computer runs and statistical analysis

Model setup and calculations were done in MATLAB (Mathworks Inc.) with the COBRA toolbox (Becker *et al.*, 2007) using the Tomlab Optimization Environment linear solver (Tomlab Inc.). Output processing was done using a customized JAVA program. Statistics and charts were done with the R statistical package, version 2.5.1.

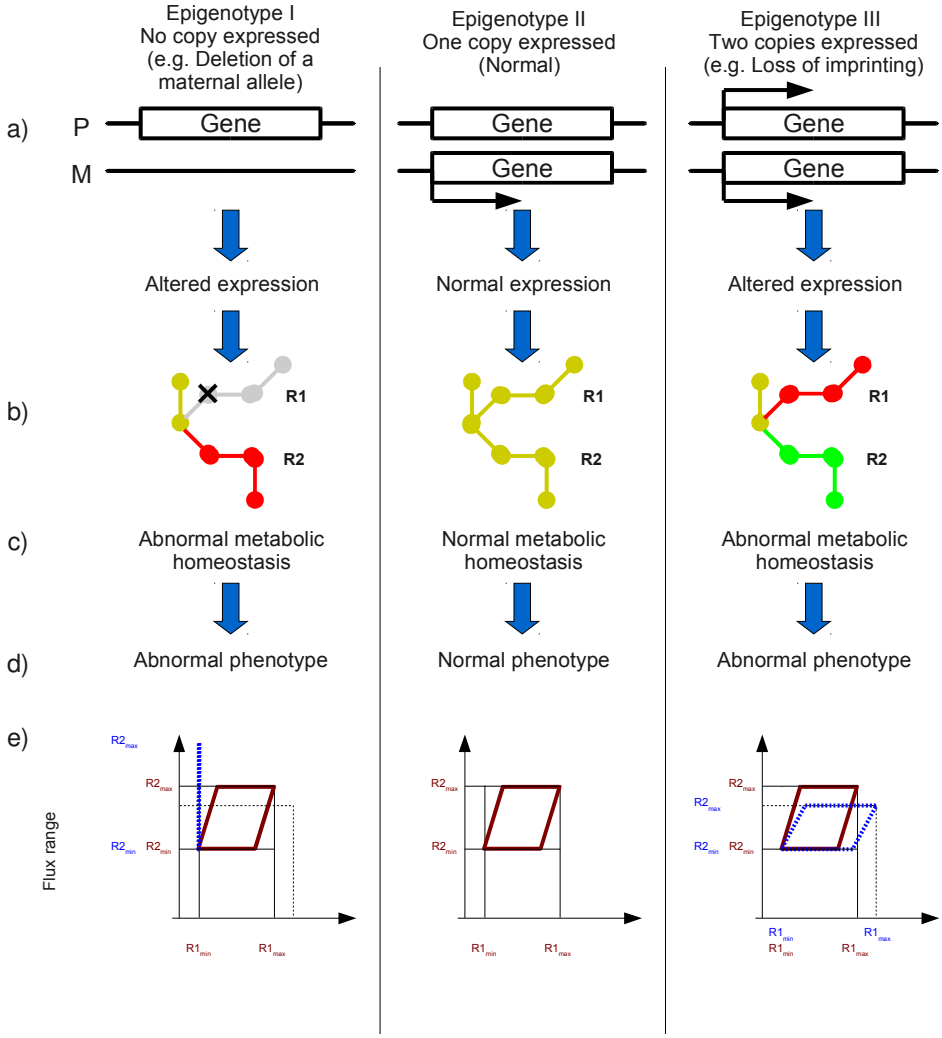


Figure 7: In silico epigenotype simulation and FVA analysis.

a) Two abnormal epigenotypes (no expression (epigenotype I) and biallelic expression (epigenotype III)) are compared against single copy expression for an imprinted gene. b) The expression pattern is translated into constraints set on the metabolic network. c) and d) The network homeostasis is calculated and compared to predict phenotypes originating from the abnormal epigenotypes. e) Flux variability analysis (FVA) determines the boundaries of flux ranges for each reaction in the metabolic network that result in the optimal solution for the objective value (here: biomass). For epigenotype II, fluxes through reactions R1 and R2 pictured in b) are within the parallelogram (red). Epigenotype I results in an increased flux capacity through R2 but zero flux capacity through R1, depicted by the blue line. Epigenotype III results in increased flux capacity for R1 but decreased flux capacity for R2 (blue).

4 RESULTS

4.1 Analysis of the sequence specificity of methylation sensitive restriction endonucleases suitable for global methylation analysis (Paper I)

To aid in selection of restriction endonucleases suitable for global methylation analysis and the interpretation of such assays, a bioinformatic analysis of the sequence specificity of the target sequences of endonucleases suitable for such analysis was done. This involved counting the number of different target sequences in the human genome to determine if the frequency was sufficient to give consistent signal in global methylation assays such as LUMA, and to test representation of the target sequences in subsets of the genome. The ideal target site is sufficiently frequent, not over-represented in genome repeats and over-represented or neutral in CpG islands and genes.

Results from database and literature searches for appropriate target sequences suitable for global methylation analysis are shown in Table 2. Expected number of cuts is estimated by the probability of each target sequence, given the estimated nucleotide frequency in the human genome (A=0.3, C=0.2, G=0.2 and T=0.3). The observed frequency is the counted frequency of the target sequence in the human genome (hg18, UCSC36).

Out of the nine endonuclease pairs studied, three had an observed target sequence frequency in the human genome high enough to be efficient in genome-scale methylation analysis. The target sequences of those pairs are CCGG and GCGC for assays of CpG methylation and CCWGG (W=A or T) for assays of non-CpG methylation. The CCGG target sequence is the target of the *Hpa*II/*Msp*I endonuclease pair traditionally used in the LUMA assay. The CCGG and GCGC target sequences represent 8.1% and 5.9% of all CpG dinucleotides in the human genome and the CCWGG target sequence represents 8.5% of all CpWpG trinucleotides in the human genome. However, it is

possible that the endonucleases with target sites that are less frequent might be usable in more sensitive assays than currently known.

Table 2: Methylation sensitive restriction endonucleases sequences chosen for the study.

Sequence	***Cut by***	***Blocked by***	***Expected frequency of target sequence (per Mb)***	***Observed frequency of target sequence (per Mb)***	***Obs/Exp***
CCGG	*Msp*I	*Hpa*II	1600	802	0.50
ACCGGT	*Csp*AI	*Age*I	144	19	0.13
CCCGGG	*Xma*I	*Sma*I	64	131	2.05
TCCGGA	*Acc*III	*Bso*MII	144	33	0.23
TTCGAA	*Asu*II	*Bst*BI	324	37	0.11
TCGCGA	*Nru*I	*Spo*I	144	5	0.04
GCGC	*Hha*I	*Cfo*I	1600	578	0.36
ACCWGGT	*Mab*I	*Sex*AI	86	103	1.19
CCWGG	*Ajn*I	*Psp*6I	960	3425	3.57

The CpG dinucleotide is greatly underrepresented in the genome (Bird, 1980). The observed/expected value for the CpG dinucleotide in the sequenced human genome was 0.25 (data not shown). The observed/expected values for the restriction sites GCGC (0.36) and CCGG (0.50) were also low. This might be due to germline methylation and subsequent hypermutability of the target sequence. The high observed/expected value of the CCWGG target sequence might indicate hypomethylation of this sequence in the human germline. However, there were also notable discrepancies between the observed/expected values of several sequences with the same core sequence, such as CCGG / ACCGGT (0.50 vs 0.13) and CCCGGG / TCCGGA (2.05 vs. 0.23). This cannot be explained by differences in cytosine methylation, so alternative explanations for different observed/expected ratios, such as differential selection, cannot be ruled out.

Table 3 demonstrates the average target sequence frequency (f_{seq}) within

various subsets of the genome, normalized by the average genome frequency (f_{genome}). For most target sequences, there was a slight over-representation in repeats, especially in SINE elements. Figure 8 shows the distribution of relative frequencies of the three most interesting target sequences chosen for further analysis (CCGG, GCGC and CCWGG) within gene-related sequences. The distribution was similar for all target sequences, but the skewed distribution was less for the CCWGG target sequence compared to the CCGG and GCGC target sequences. The CCGG target sequence was slightly overrepresented in repeats, especially SINE. However, it was greatly over-represented in both CpG islands, gene promoter regions and gene exons. The GCGC target sequence had similar distribution but more over-representation in CpG islands, gene promoter regions and gene exons (Table 3)(Fig. 8). The CCWGG target sequence had a slight over-representation in CpG islands, gene promoters and exons (Table 3) (Fig. 8).

Many chromosomes demonstrated enrichment for the target sequences near the chromosome ends. Figure 9 shows the distribution of the three most interesting target sequences in chromosome 16 as an example. Appendix I contains chromosome frequency images for the three target sequences chosen for further analysis.

Therefore, using the CCGG and GCGC target sites for endonuclease-based assays of global CpG methylation analysis is probably feasible, as the target sites are relatively frequent, and are over-represented in intersecting subsets of the genome, such as CpG islands and genes while not being very over-represented in repetitive elements. Similarly, the CCWGG target site is probably the most useful target site for assessing global CWG methylation.

Table 3: Relative frequencies of target sequences of methylation sensitive restriction endonucleases studied in repeats and gene-related sequences.

Sequence	***Relative frequency in repeats***				***Relative frequency in gene-related sequences***		
	% within repeats	**All repeats** f_{seq}/f_{genome}	**LINE** f_{seq}/f_{genome}	**SINE** f_{seq}/f_{genome}	**CpG islands** f_{seq}/f_{genome}	**Promoters** f_{seq}/f_{genome}	**Exons** f_{seq}/f_{genome}
CCGG	53%	1.2	0.3	2.0	11.7	5.7	3.9
ACCGGT	46%	0.8	0.6	0.6	4.7	2.9	2.6
CCCGGG	62%	1.6	0.1	3.1	11.3	5.5	3.0
TCCGGA	43%	0.9	0.5	1.1	8.1	4.6	4.3
TTCGAA	43%	0.9	0.8	0.9	1.7	1.8	1.9
TCGCGA	46%	1.1	0.3	1.8	20.7	12.4	5.0
GCGC	55%	1.3	0.3	2.3	16.4	8.3	4.5
CCWGG	56%	1.4	0.5	2.2	2.0	1.5	1.5
ACCWGGT	45%	0.9	0.7	0.9	1.3	1.1	1.6

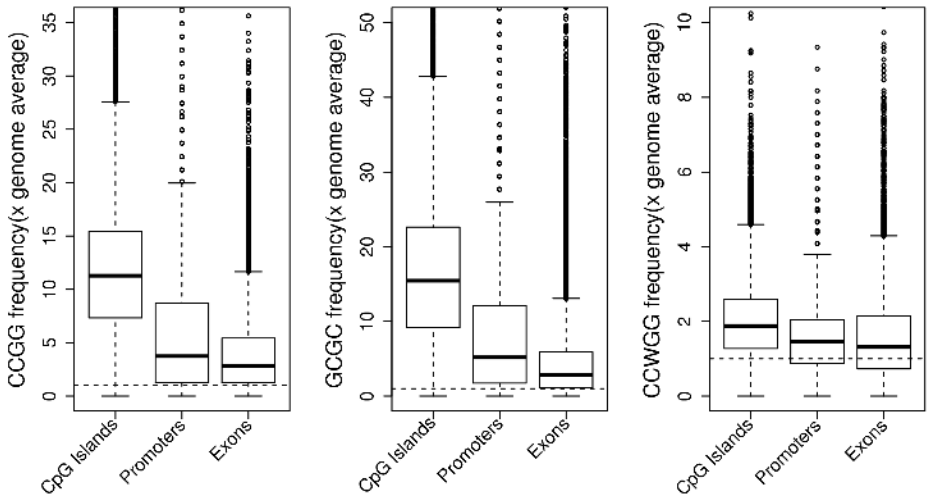


Figure 8: Relative frequency distribution of CCGG, GCGC and CCWGG target sequences within gene-related sequences.

Shown are the representation in CpG islands, promoters and exons times the genomic average representation. Broken line shows average genome representation of each target sequence.

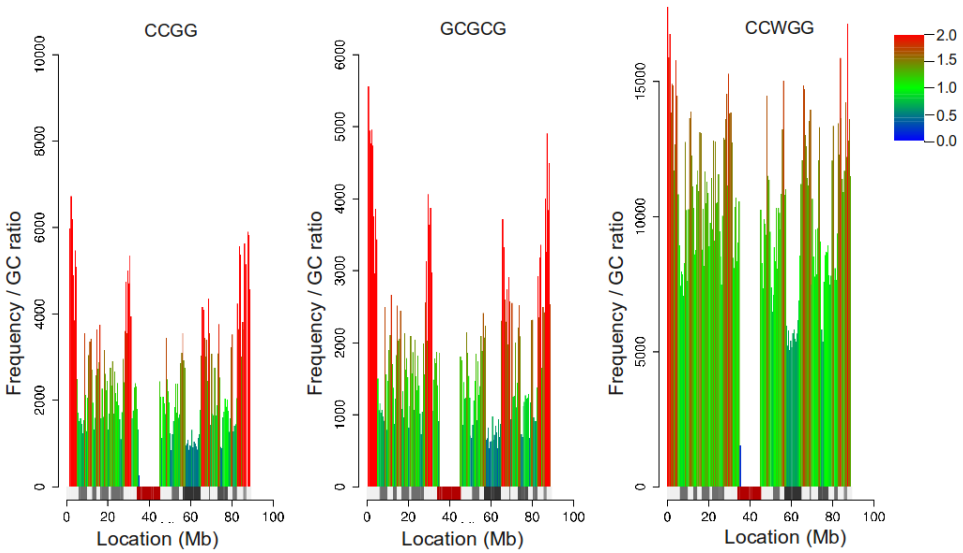


Figure 9: Relative frequency distribution for CCGG, GCGC and CCWGG in 500 kb windows within chromosome 16.

As an example, the frequency in chromosome 16 is shown, normalized for GC ratio in the window. Height of the bars indicates the absolute frequency, while the color of the bars indicates the relative frequency compared to the genome average. The relative frequency ranges from zero (blue) to two times (red) the genome average.

4.2 Intra-individual change over time in DNA methylation with familiar clustering (Paper II)

To compare intra-individual change in DNA methylation over time, we decided to measure both changes in global DNA methylation and site specific changes in promoter regions of 807 genes. For the global DNA methylation analysis, we used the recently developed LUMA assay. Prior to its usage, the assay was thoroughly validated and optimized.

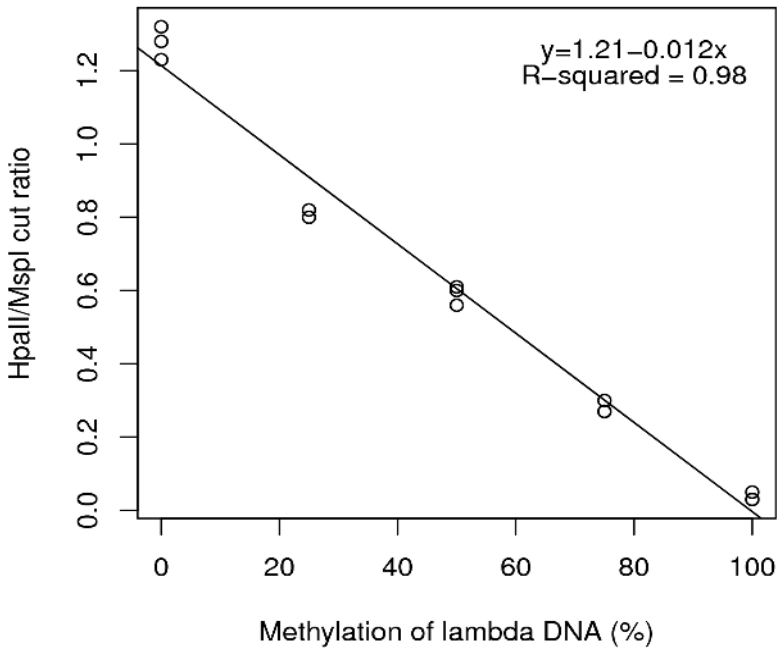


Figure 10: A standard curve demonstrating linearity of LUMA.

The curve was subsequently used to convert measurements of *Hpa*II/*Msp*I endonuclease cut into absolute percentage of *Hpa*II/*Msp*I target sequence methylation.

4.2.1 Properties of the LUMA assay

A standard curve for LUMA was done by mixing 100% and 0% methylated lambda phage DNA and performing LUMA measurements of each mixture in triplicate. The curve demonstrated linearity of the LUMA method over a wide range of methylation values (Fig. 10). Furthermore, the curve was used for conversion of *r* values representing ratios of *Hpa*II/*Msp*I endonuclease cuts normalized by *Eco*RI endonuclease cuts into an absolute methylation percentage in the following studies.

As a biological control, the methylation of DNA methyltransferase I (Dnmt1) double knockout cell line was compared against the methylation of its parent cell line (HCT116). Previously, a double knockout of the Dnmt1 gene was found to be necessary to significantly affect global DNA methylation levels (Rhee *et al.*, 2002). The findings were confirmed using LUMA. The average *Hpa*II/*Msp*I target sequence methylation of the HCT116 line was 82% while the average methylation of the DKO cell line was 31% (P=0.001). The variance of the assay was established by performing three separate enzymatic digestions of 25 samples and measuring methylation with LUMA. This revealed an average variance of 2% (data not shown). To test the stability of methylation in samples from whole blood, seven individuals were sampled 2-4 times over 30 days (Fig. 11). No significant change in *Hpa*II/*Msp*I target sequence methylation was found in DNA from total blood or the dominating cell type in buffy white coat layer used for DNA isolation of all study samples. Also, 18 measurements from nine individuals demonstrating the most extreme differences were repeated one year after the original measurement confirming the initial measurements (data not shown).

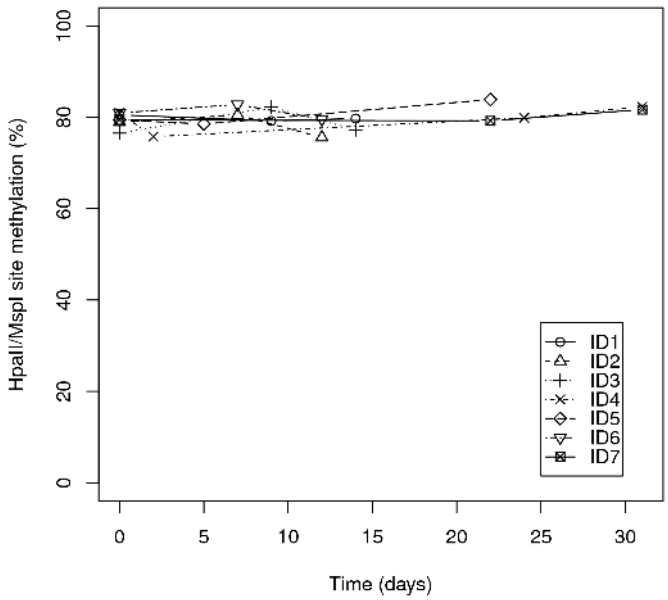


Figure 11: Short-term stability of methylation of HpaII/MspI target sequence in peripheral blood measured by LUMA.

Seven individuals (ID1-ID7) were sampled repeatedly 2-4 times over 30 days to test short-term stability of LUMA measurements. Each measurement was done in triplicate.

4.2.2 Cross-sectional analysis of changes in global DNA methylation over time in an Icelandic cohort

First, we tested changes in global DNA methylation in a cross-sectional cohort. This was both done to test the LUMA assay and to compare LUMA measurements of global methylation changes by birth year to results from other studies, that had generally found no change with age using cross-sectional cohorts. LUMA measurements of *Hpa*II/*Msp*I target sequence methylation in a cohort of 84 Icelanders born between 1940-1949 revealed no linear trend for methylation based on the year of birth (ANOVA test, P=0.96)(Fig. 12). No obvious biological explanation was found for outliers within the group, who had consistent triplicate measures of their methylation value. Additionally, half

of the cohort was with type II diabetes, an acquired metabolic disease whose incidence increases with age. The LUMA measurements of *Hpa*II/*Msp*I target sequence methylation did not differ significantly between the group with diabetes and the control group (*t*-test, *P*=0.46) (Fig. 13).

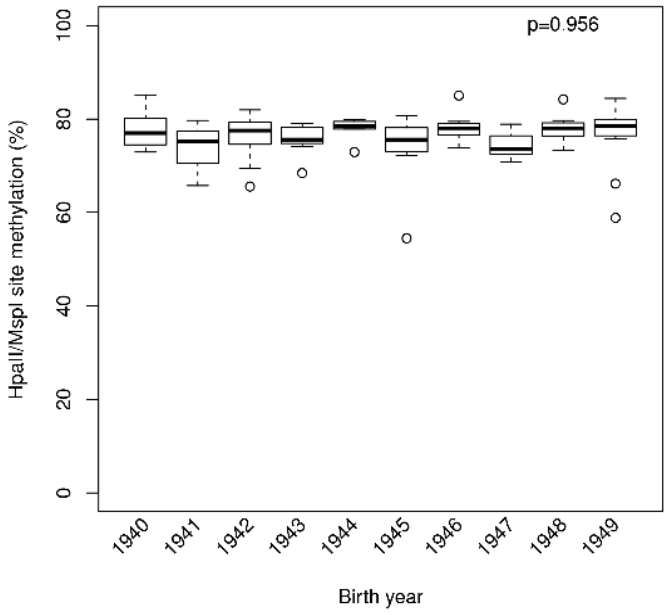


Figure 12: Cross-sectional analysis of changes in methylation with age.

A total of 84 individuals born between 1940-1949 were measured, each in triplicate. There was no association between global methylation measured by LUMA and year of birth.

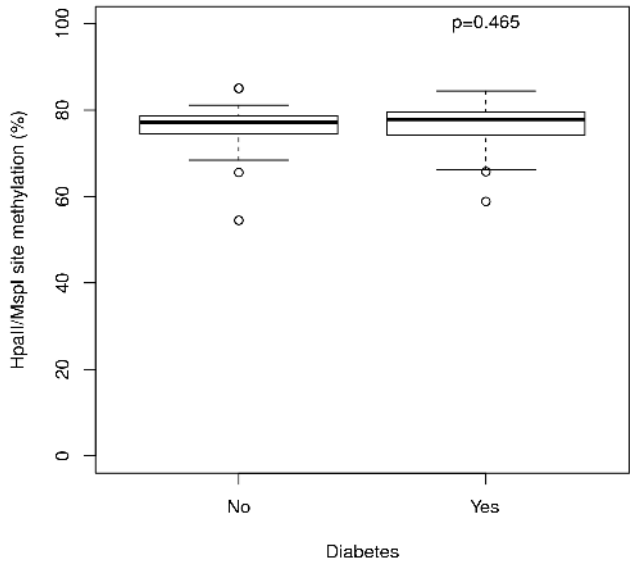


Figure 13: DNA methylation by diabetes status.

There difference in global DNA methylation measured by LUMA was not significant between individuals with type II diabetes(n=35) and control individuals (n=49).

4.2.3 Longitudinal analysis of changes in global DNA methylation over time in the Icelandic cohort

Since changes could not be demonstrated in a cross-sectional cohort, we tested intra-individual changes in global methylation in a longitudinal manner, using two samples from each participant sampled with a considerable time interval. A total of 111 individuals from the Icelandic longitudinal study population had valid measurements in triplicate from both time points. The average time between sampling was 11 years. The mean inter-individual difference in methylation over an average of 11 years was zero. However, individuals demonstrated intra-individual changes with time. Figure 14 shows the change in absolute *Hpa*II/*Msp*I methylation between the two time points for all individuals. Notably, the change was bi-directional; some individuals gained methylation between sampling while other individuals lost methylation. A total of 70 individuals (63%) had an absolute change in *Hpa*II/*Msp*I methylation of at least 5% between the two measurements, a total of 33 individuals (30%) had an absolute change of at least 10% and a total of 9 individuals (8%) had an absolute change of 20% or more between the two measurements. The observed differences did not occur in 10,000 permutations of the data when the two time points were randomly created for each individual, using all six measurements, simulating no change in methylation between the two time points (Fig. 14, gray area). Furthermore, the observed ratio of variation within each individual and variation within each time point ($R=Var_{between}/Var_{within}=11.23$) did not occur during the 10,000 permutations ($P<0.001$ of no change over time).

The LUMA measurements did not correlate with measurements of inflammation markers (C-reactive proteins, absolute and differential white blood cell count, erythrocyte sedimentation rate). Age or length of storage for the samples was not correlated with the results.

Half of the longitudinal population had a diagnosis of adult-onset cancer. As DNA methylation is involved in the pathogenesis of many adult-onset cancers, the global methylation changes were compared between individuals with and without lifetime diagnosis of cancer. There was no difference in the methylation change between the two time points for individuals diagnosed with cancer compared to individuals not diagnosed with cancer (Fig. 15).

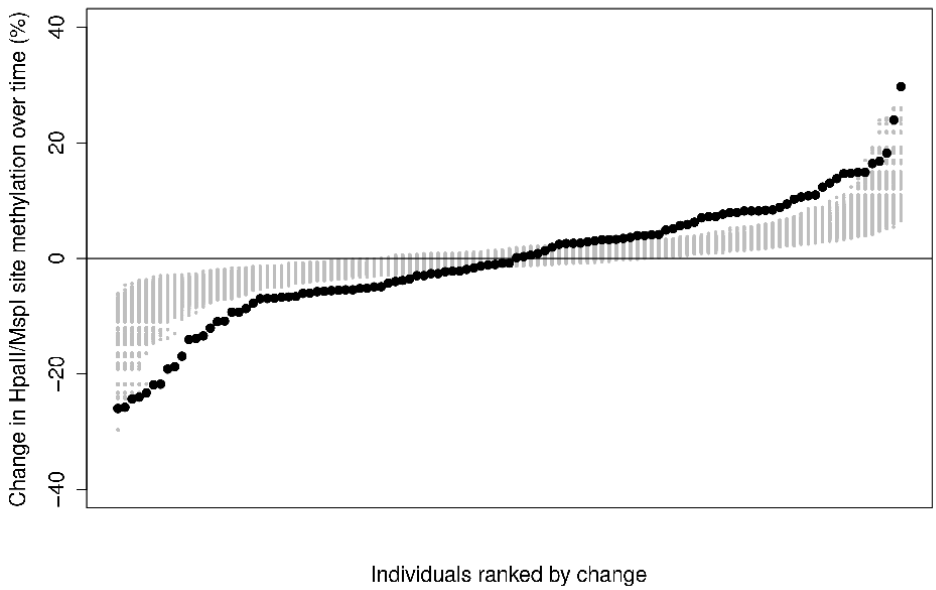


Figure 14: Longitudinal results for the Icelandic population.

Difference in methylation measured by the LUMA assay between the two time points for each individual is shown (black points). The gray area represents the null hypothesis of no effects with time, created by 10,000 permutations of the data.

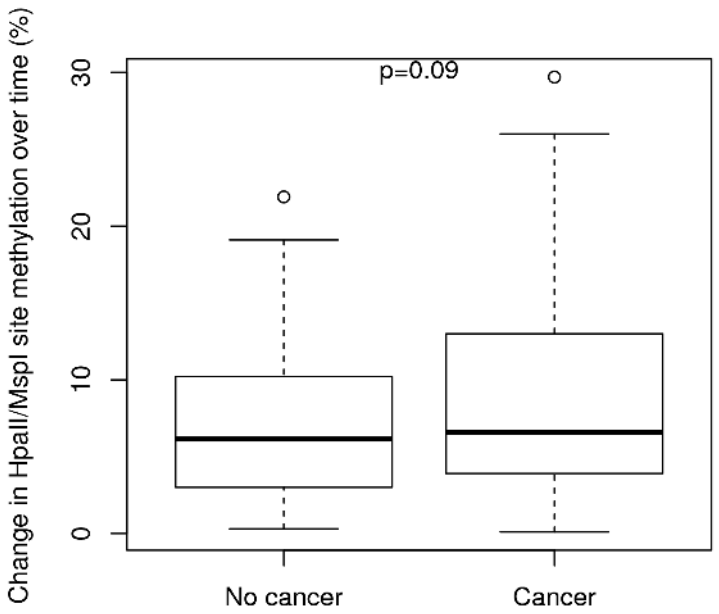


Figure 15: Methylation changes by cancer status.

Shown is the absolute difference in methylation measured between the two time points for individuals with cancer of any kind (n=50) or individuals without cancer (n=61).

4.2.4 Genome-wide changes in DNA methylation over time in Utah cohort and heritability analysis of the changes

In the Utah cohort, a total of 126 individuals had valid measurements from two time points with an average of 16 year interval. Figure 16 shows the change in absolute *Hpa*II/*Msp*I methylation between the two time points for all individuals. The mean inter-individual difference in methylation over an average of 16 years was zero like in the Icelandic population. Similar to the results from the Iceland cohort, the effect was found to be bi-directional. A total of 50 individuals (40%) had an absolute change in *Hpa*II/*Msp*I methylation of at least 5% between the two measurements, a total of 23 individuals (18%) had an absolute change of at least 10% and a total of 13 individuals (10%) had an

absolute change of 20% or more between the two measurements.

The Utah population represented individuals from up to three generations of 21 families. Most families had two adult generations (average age at sampling was 17 and 32 years for time 1 and 2 respectively). This indicates that the household was not shared for a substantial amount of time in the sampling interval. A tight clustering was seen for many families, both within those losing and gaining methylation over time (Fig. 17). After adjusting the value at time 2 for the value at time 1, the residual value was used as a phenotype for heritability analysis. The heritability analysis, done with the ASSOC program in SAGE, estimates the familial correlation in a model of a continuous trait (h^2). There was a high heritability estimate (h^2=0.99, P<0.001) that remained high (h^2=0.74, P=0.003) even after removing the family with the most extreme values (family 21). The heritability analysis indicates a significant

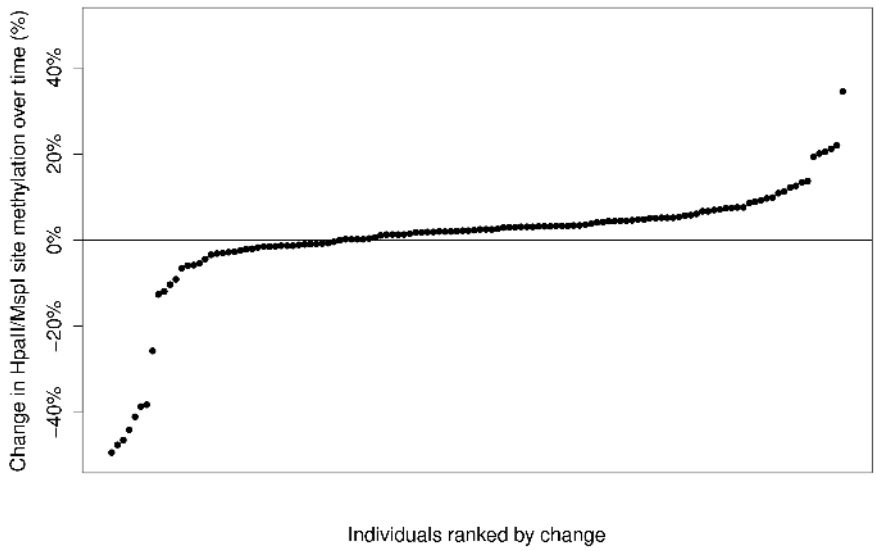


Figure 16: Longitudinal results for the Utah population.

Difference in methylation measured by the LUMA assay between the two time points for each individual is shown (black points).

genetic component in the maintenance of methylation or shared environmental factors between family members.

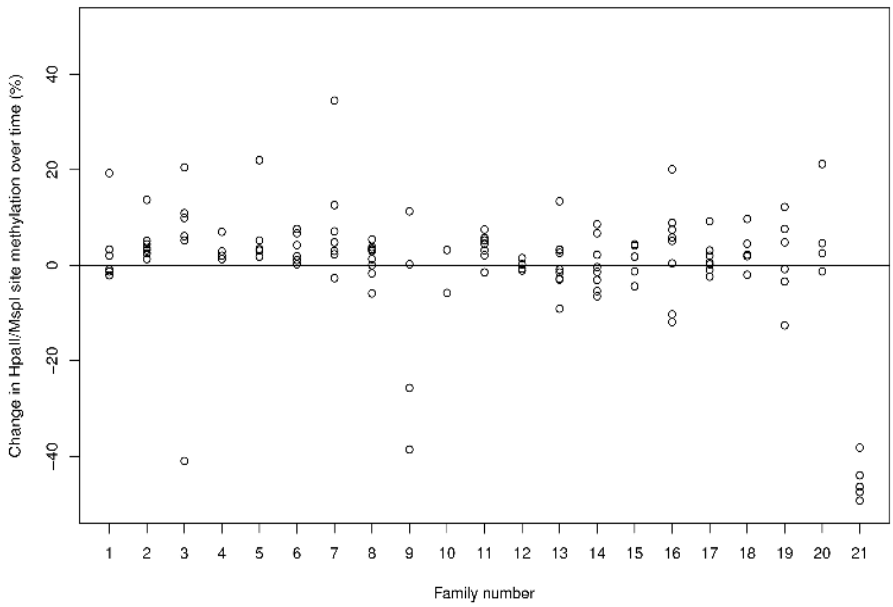


Figure 17: Longitudinal methylation changes by families.

Shown is the change in *Hpa*II/*Msp*I sequence methylation measured by LUMA between the two time points in the Utah cohort categorized by families.

4.2.5 Site-specific longitudinal DNA methylation changes in a subset of individuals from both cohorts

Given the results on intra-individual bidirectional changes in global methylation, we next tested if these changes could also be observed at specific sites in the genome. A subset of 41 individuals from the Icelandic and Utah cohort were studied further by applying bisulfite treated DNA onto Golden Gate methylation microarray probing the methylation of 1505 CpG dinucleotides in 807 genes. The samples used were from the 17 and 19 individuals demonstrating the greatest loss and gain of methylation in addition to five individuals with the least change between the two time points based on the

LUMA measurements.

In general, trends in methylation changes assessed with the microarray followed the global changes found by LUMA. There was a tight clustering in family 21, the same family that demonstrated the most extreme changes and clustering in the LUMA assay. All five family members lost global methylation between the two time points. Similarly, out of the 50 CpG methylation probes with the greatest change between the two time points, 49 of them indicated loss of methylation ($P<0.001$ against even odds of gain/loss of methylation) (Table 4). There was a slight enrichment for imprinted genes within the list of genes with most changes (5/50 vs. 28/807, $P=0.047$). Out of the 50 CpGs with the greatest differences across all individuals, 13 were shared with family 21. Several genes involved in immunological mechanisms were included in the list of genes with large changes in CpG methylation between the two time points (Table 5). Unfortunately, no phenotype information was available for the Utah cohort, so the effects of the methylation change on phenotype are unknown.

Table 4: Fifty genes with the greatest change in methylation over time for five members of family 21.

No.	***Gene symbol***	***Chr***	***Δm***	***P value***	***No.***	***Gene symbol***	***Chr***	***Δm***	***P value***
1	***PWCR1***[i]	**15**	**-0.723**	**<0.001**	26	*G6PD*	X	-0.177	0.005
2	*IL1B*	2	-0.42	0.001	27	*FMR1*	X	-0.177	0.023
3	*KCNK4*	11	-0.402	0.001	28	*BCAP31*	X	-0.174	0.002
4	***AIM2***	**1**	**-0.372**	**<0.001**	29	*SNRPN*[i]	15	-0.172	0.001
5	*PI3*	20	-0.306	0.003	30	*BAX*	19	-0.17	0.001
6	*CSF3R*	1	-0.301	0.01	31	*SYK*	9	-0.169	0.013
7	*GLA*	X	-0.274	0.001	32	*GLA*	X	-0.169	0.007
8	*PLA2G2A*	1	-0.259	0.007	33	*VBP1*	X	-0.168	0.006
9	*NOTCH4*	6	-0.255	0.001	34	*IL10*	1	-0.168	0.055
10	*TRPM5*[i]	11	-0.251	0.001	35	*LMO2*	11	-0.163	0.076
11	*HDAC6*	X	-0.247	0.003	36	*MPL*	1	-0.162	0.012
12	***GFAP***	**17**	**-0.246**	**<0.001**	37	*TRIP6*	7	-0.162	0.037
13	*HOXA5*	7	-0.242	0.011	38	*IRAK1*	X	-0.16	0.068
14	*PTK6*	20	-0.226	0.014	39	*VBP1*	X	-0.158	0.001
15	*G6PD*	X	-0.21	0.017	40	*BIRC4*	X	-0.155	0.027
16	*ELK1*	X	-0.205	0.005	41	*SLC22A18*[i]	11	-0.154	0.016
17	*G6PD*	X	-0.204	0.041	42	*LCN2*	9	-0.152	0.007
18	*ERCC3*	2	-0.202	0.063	43	*SLC22A2*[i]	6	-0.152	0.002
19	*LMO2*	11	-0.201	0.021	44	*IL16*	15	-0.151	0.012
20	*CSF2*	5	-0.2	0.001	45	*SNCG*	10	-0.142	0.11
21	*LIF*	22	0.2	0.006	46	*LCN2*	9	-0.142	0.11
22	*ELK1*	X	-0.195	0.003	47	*DNASE1L1*	X	-0.142	0.014
23	*PLG*	6	-0.19	0.001	48	*EMR3*	19	-0.138	0.004
24	*ARAF*	X	-0.18	0.004	49	*ELK1*	X	-0.138	0.059
25	*DKC1*	X	-0.18	0.19	50	*DNASE1L1*	X	-0.135	0.076

[a)]Fractional difference in DNA methylation between the two time points is shown (Δm). Negative values indicate loss of DNA methylation between the two time points. Imprinted genes are marked with [i]. *P*-values are unadjusted, but values reaching significance after Bonferroni correction ($P<0.05/807$) are typeset in bold. Chr-chromosome.

Table 5: A list of genes with greatest difference in methylation over time in all individuals. Shown are the 13 genes on the list that also revealed the greatest difference over time individuals of family 21.

Gene symbol	***Gene function***
AIM2[im]	Interferon gamma inducible transcript
CSF3R[im]	Colony stimulating factor 3 receptor
HOXA5	Hox gene
PTK6	Protein tyrosine kinase
ERCC3	Helicase with excision-repair functions
LMO2[im]	Role in erythropoiesis and in T-cell leukemogenesis
SYK[im]	Spleen-tyrosine kinase
IL10[im]	Cytokine
BIRC4	Apoptosis inhibitor
IL16[im]	Cytokine
LCN2[im]	Protein associated with neutrophil gelatinase
TRIP6	Regulates lysophosphatidic acid induced cell migration
EMR3[im]	Myeloid-myeloid interactions during immune and inflammatory responses

[a)]Genes involved in immunological mechanisms are marked with [im].

4.3 Development of a surrogate marker for germline methylation (Paper III)

4.3.1 Development and properties of methylation-associated SNP (mSNP) markers

Since the human germline is subject to sampling difficulties, alternative approaches of assessing germline methylation might be sought. One possibility is to develop bioinformatic markers. This was done in Paper III.

To create a bioinformatic marker of germline methylation using the hypermutability of methylated cytosine, a large well validated data set of SNPs in the human genome is needed. We decided to use the HapMap database of human SNPs as the creation of the dataset aimed at typing at least one SNP every 5 kb that increases the homogeneous distribution of SNPs, and information on population frequency and ancestral allele is readily available.

In the second phase HapMap database, 2,252,113 non-redundant C/T or G/A SNPs were found within the autosomal chromosomes. Of those 763,035 were within a CpG dinucleotide. Using a derived allele data set (indicating the ancestral allele for the majority of HapMap SNPs), 1,239,485 C/T or G/A non-redundant SNPs were located in the autosomal chromosomes. Of these 434,198 (35%) were within CpG dinucleotide and the ancestral allele was either C or G (indicating C→T or A→G mutation). These therefore met the criteria for $mSNP_{GENOME}$. The average ± standard deviation $mSNP_{GENOME}$ density was 79 ± 39 $mSNP_{GENOME}$ per 500 kb of sequence.

In the 10 Mb of sequence within the ENCODE regions, 9809 non-redundant C/T or G/A were found. A total of 2987 SNPs were within the CpG dinucleotide, meeting the criteria of $mSNP_{ENCODE}$. The derived allele criteria was not applied to the ENCODE data set to maximize the resolution of these regions. The average ± standard deviation $mSNP_{ENCODE}$ density was 299±123

$mSNP_{ENCODE}$ per 500 kb of sequence.

Most substitutions in DNA sequences are caused by genetic drift and not selection and have negligible effects on fitness (Kimura, 1991). SNPs are generally functionally neutral and represent the local mutation frequency. Fixation or elimination is dependent on population size but independent from SNP type. Therefore methylation-associated SNPs, mSNPs, are assumed to reflect the mutation rates of methylated cytosines and not to be more affected by selection than non-mSNPs.

To test if the subset of mSNPs undergoing recent selection was larger than the subset of non-mSNPs, the integrated haplotype score (iHS) was compared between the two subsets. The iHS score is a measurement of the ratio of haplotype decay in the 1 MB region adjacent to a SNP between the ancestral and the derived SNP allele (Voight *et al.*, 2006). It is a normally distributed parameter with a mean of 0 and a standard deviation of 1. An iHS score of 0 indicates no selection whereas an iHS score of more than 2.5 indicates recent selection.

Table 6 shows the population statistics for the iHS parameter for all populations. Both the mSNPs and non-mSNPs subsections had a mean iHS score around 0 and standard deviation of 1. Furthermore, approximately 1 % of all SNPs in all subsets had absolute iHS scores of more than 2.5 indicating recent selection. The mean iHS score did dot differ significantly between the mSNP and the non-mSNP subsets. The number of SNPs with an iHS absolute value of more than 2.5 either did not differ significantly between mSNP or non-mSNP or the non-mSNP subset had a slightly higher number of SNPs with high iHS scores. This indicates that the mSNP subset is not subject to more selection than the non-mSNP subset. As an example, Figure 18 shows the distribution of

iHS scores for mSNPs and non-mSNPs for the east Asian population, the combined Japanese and Chinese populations.

Table 6: iHS summary statistics for mSNP and non-mSNP subsets of all populations within HapMap.

	Non-mSNP					***mSNP***				
Pop.		**iHS**		**|iHS|>2.5**			**iHS**		**|iHS|>2.5**	
	N	**Mean**	**SD**	**n**	**%**	**N**	**Mean**	**SD**	**n**	**%**
ASI	749,791	-0.066	1.005	8,494	1.1	235,639	-0.008	0.993	2,521	1.1
CEU	931,439	-0.071	1.001	10,353	1.1	281,911	0.002	0.995	3,173	1.1
YRI	1,216,014	-0.052	1.005	15,481	1.3	313,138	0.019	0.992	3,811	1.2

[a)] Mean and SD for iHS values were similar for the mSNP and non-mSNP subset for all populations. Similar ratio of SNPs within the mSNP and non-mSNP subsets had absolute iHS values over 2.5. ASI: East Asian Population, CEU: Northern and Western European Population, YRI: Youruba Population.

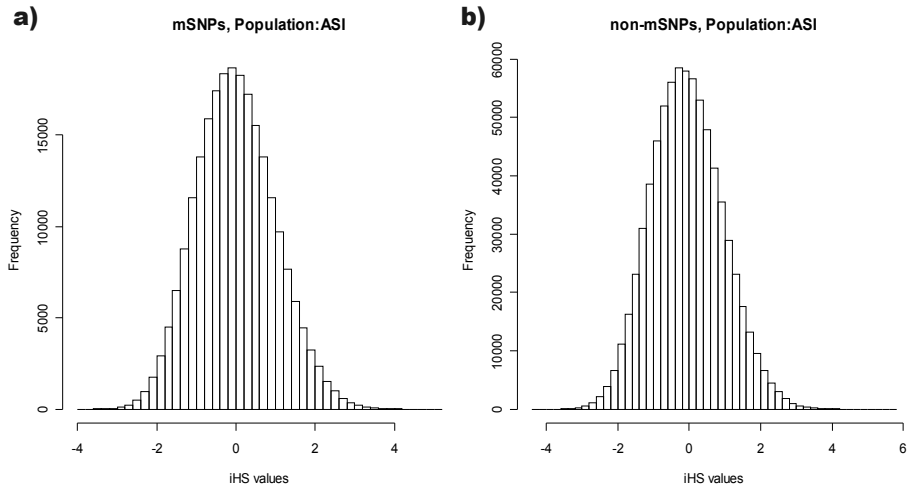


Figure 18: Distribution of integrated haplotype scores (iHS) for mSNPs and non-mSNPs in the East Asian population.

An iHS score of 0 indicates no signs of recent selection whereas iHS of more than 2.5 (or less than -2.5) indicates recent selection.

Several methods were used to validate the assumption that mSNP counts represented the hypermutability of methylated cytosine bases and could therefore be used as a surrogate marker for germline methylation.

A total of 26,635,559 out of 548,370,281 (4.9%) cytosines are within a CpG dinucleotide. If cytosine dinucleotides were not hypermutable, 4.9% of all C/T polymorphisms should be within CpG dinucleotides. Therefore, the estimated number of C/T polymorphisms within CpG dinucleotides would be 60,205 and 476 C/T for the genomic and ENCODE data sets, respectively. However, the observed numbers (434,198 and 2987) demonstrate a 7.2 and 6.3 fold over-representation, respectively ($P<10^{-15}$ for both). The ratio is in a similar range to the observed five to sixfold mutation frequency of CpG sites in the human genome due to methylation (Zhao & Zhang, 2006). This suggests that the mSNP data sets represent hypermutability of methylated cytosine within the CpG dinucleotide.

The observed mSNP counts were used to estimate the ratio of mSNPs in both data sets likely to represent methylation. For the genomic data set:

$$Ratio_{trueMeth} = \frac{mSNP_{obs} - mSNP_{\exp}}{mSNP_{obs}} = \frac{434{,}198 - 60{,}205}{434{,}198} = 0.86$$

For the ENCODE data set:

$$Ratio_{trueMeth} = \frac{mSNP_{obs} - mSNP_{\exp}}{mSNP_{obs}} = \frac{2{,}987 - 476}{2{,}987} = 0.84$$

Therefore, approximately 84-86% of the SNPs defined as mSNP are estimated to result from the hypermutability of methylated cytosines.

It is likely that the assignment of ancestral base at CpG sites is more error-prone than at non-CpG sites due to the hypermutability of methylated CpGs. Nonetheless, there were significantly more C/T or G/A SNPs within the CpG dinucleotide with C (or G) as the ancestral allele compared with T (or A) as the ancestral allele (443,657 vs. 333,606, ratio 1.32, $P<10^{-15}$ against even probability). However, the large number numbers of C/T or G/A SNPs with T

(or A) as the ancestral allele indicates that many of those might represent methylation as well. Table 7 compares absolute counts of SNPs in the derived allele data set between C/T and G/T SNPs, both within and not within the CpG dinucleotide. There was an increased relative proportion of C/T SNPs within the CpG dinucleotide and with C as the ancestral allele (0.54 vs. 0.24 for C vs. G as the ancestral allele). This suggests that the derived allele data set represents the known hypermutability of methylated DNA.

Table 7: Number of SNPs in the derived allele data set within the CpG and CpH (H=A, C, T) dinucleotide.

SNP type	***Count***	***SNP type***	***Count***	***Ratio***
(C*/T)pG	434,198	(C*/T)pH	805,207	0.54
(G*/T)pG	54,550	(G*/T)pH	227,845	0.24
(T*/C)pG	324,960	(T*/C)pH	677,057	0.48
(T*/G)pG	76,920	(T*/G)pH	180,319	0.43

[a)]Two types of SNPs, C/T and G/T were compared. The derived allele is marked with (*). According to the model, (C*/T)pG SNPs are informative of methylation based on hypermutability of methylated cytosine.

4.3.2 A genome-wide map of germline methylation

For mapping purposes, a methylation index (MI) was calculated as explained in chapter 3.3.1 for all autosomal chromosomes in a 500 kb resolution. There was a positive correlation between adjacent windows (r=0.362, $P<10^{-15}$) (Data not shown), indicating that MI defines a genome feature extending over at least 500 kb. The MI had a strong negative correlation (Fig. 19) with the number of bases within CpG islands (r=-0.483, $P<10^{-15}$), consistent with the known hypomethylation of these sequences.

Figure 20 shows a plot of germline methylation in human chromosomes 1-22 using the MI as a surrogate marker. These plots were used in a specially made data viewer (MethyMap) to generate a visual representation of the MI

marker along with several sequence features for hypothesis generation. The lowest mean MI was within chromosome 19 (Fig. 21). While this chromosome has the highest proportion of CpG islands (Grimwood *et al.*, 2004), it is also notable for the lowest density of recombination hot spots (Myers *et al.*, 2005).

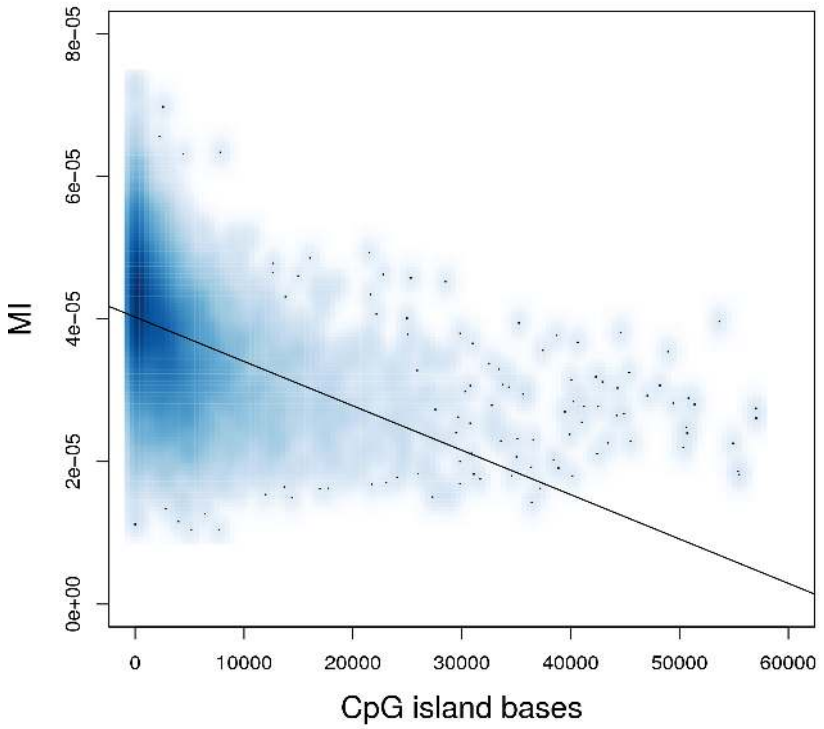


Figure 19: Correlation between the methylation index (MI), and bases in CpG islands.

Shown is the a smoothed scatterplot of MI and number of bases within CpG islands in 500 kb genome-wide windows.

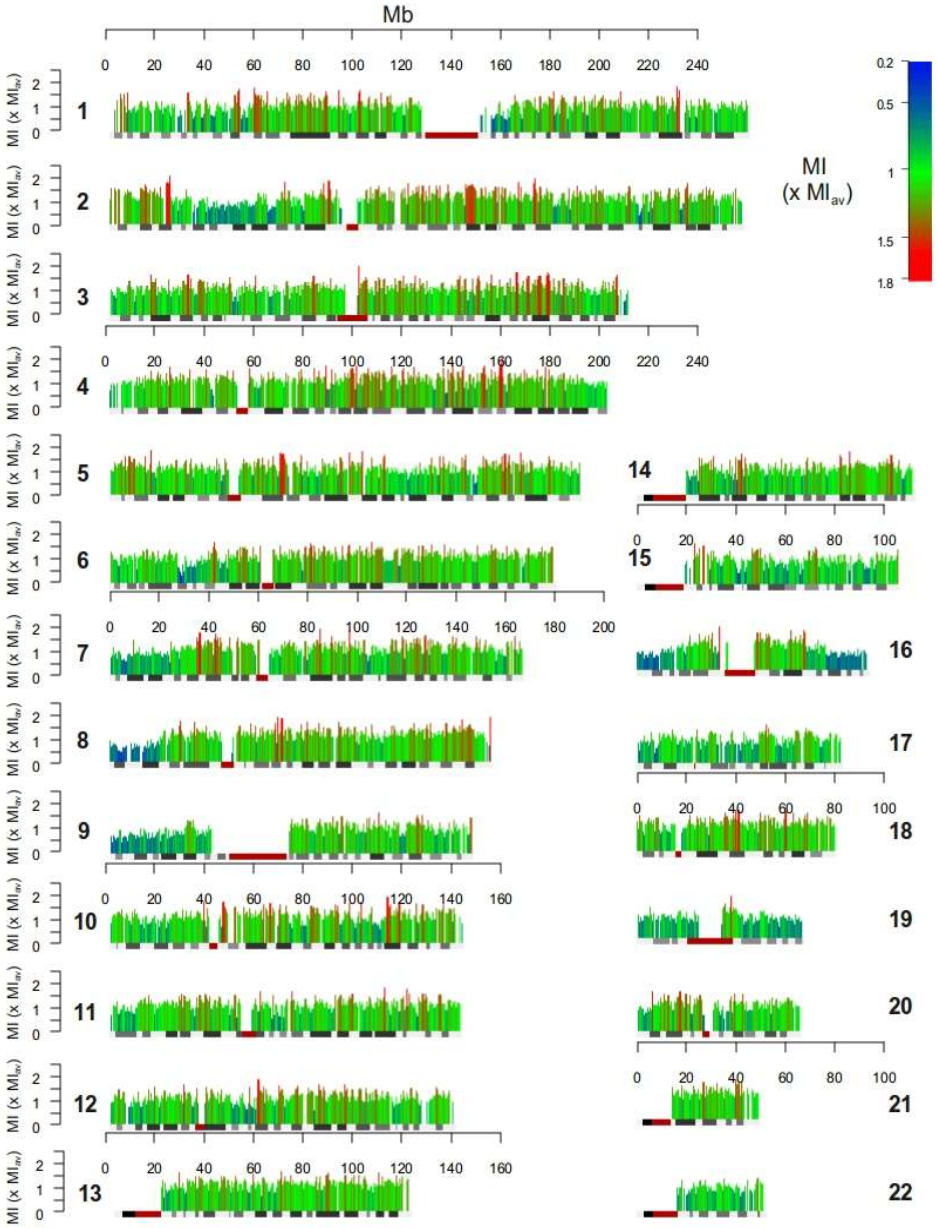


Figure 20: A genome-wide map of the methylation index (MI), a bioinformatic surrogate marker of germline methylation.

MI was calculated in a 500 kb resolution where sufficient data existed. Y axis values and color indicate MI normalized by the average genomic MI (MI_{AV}).The MI of dark chromosomal bands was significantly higher than the MI of light chromosomal band ($3.7 \cdot 10^{-5}$ vs $3.9 \cdot 10^{-5}$, $P<10^{-15}$).

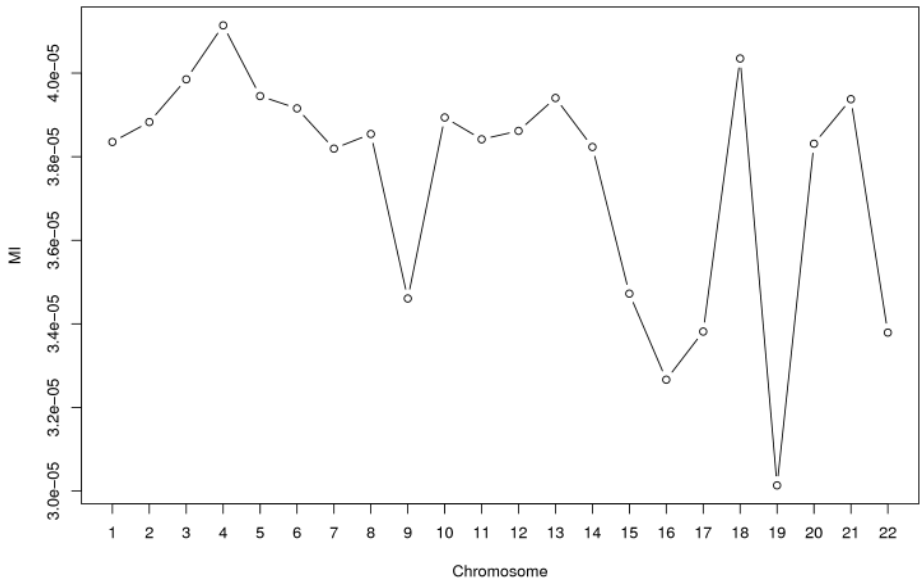


Figure 21: Average methylation index (MI) of the 22 human autosomes.

4.4 Correlation between mSNP density and meiotic homologous recombination in the human genome (Paper III)

4.4.1 Genome-wide correlation between the mSNP marker of germline methylation and regional recombination

The mSNP marker developed in Paper III might be suitable to support our suggestion that epigenetic mechanisms such as germline methylation might be involved in the mechanism behind homologous recombination. This was therefore pursued in Paper III, initially by correlating the density of the mSNP marker and regional levels of homologous recombination and density of recombination hot spots.

The absolute correlation between two markers of recombination activity in the window-based approach (recombination rate of window and number of bases within recombination hot spots) was high and positive in all four window sizes, i.e. 0.725, 0.747, 0.782 and 0.822 in 125, 250, 500 and 1000 kb windows,

respectively ($P<10^{-15}$ for all observations). Both recombination rate and number of bases within recombination hot spots were used in the further analysis. Furthermore, to allow correction for multiple confounders, mSNPs were used as a marker of germline methylation in a multiple linear model holding the other components of the MI constant (density of CpG and SNPs).

There was a significant positive absolute correlation between the number of $mSNP_{GENOME}$ per window and recombination rate in all window sizes tested (Fig. 22 and Table 8). Furthermore, there was a significant positive absolute correlation between the number of $mSNP_{GENOME}$ and the number of bases within recombination hot spots in all window sizes tested (Fig. 22 and Table 8).

Confounding variables were selected on model properties and available literature on homologous recombination. Sequence factors known to affect

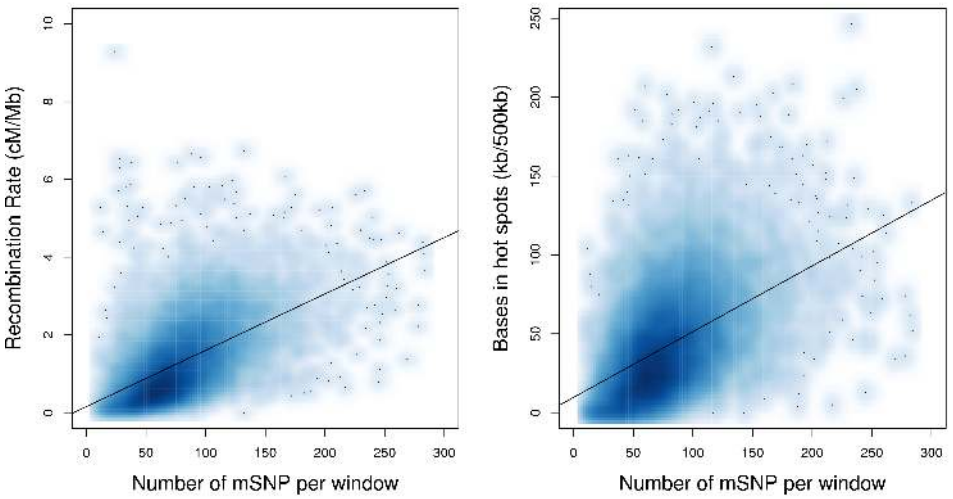


Figure 22: Correlation between mSNP and measurements of recombination.

Smoothed scatterplot of mSNP and measurements of recombination. Shown is correlation in 500 kb resolution between the mSNP marker and recombination rate ($r=0.622$, $P<10^{-15}$) and bases within recombination hot spots per window ($r=0.508$, $P<10^{-15}$).

recombination at window sizes of 125-1000 kb are GC content, repeats and exons (Myers *et al.*, 2006; Smith *et al.*, 2005). The 7 bp DNA motif did not correlate with recombination activity in window sizes greater than 8 kb (Myers *et al.*, 2006). There was a significant positive correlation between the density of the 13 bp motif (CNCCNNTNCCNCC) and both recombination rate and density of recombination hot spots. The methylation marker is also sensitive to SNP density and CpG density. Therefore, confounding variables included were GC ratio and the density of CpG dinucleotides, exons, repeats, the 13 bp DNA motif and SNPs.

Table 8: Absolute correlation between mSNP and other sequence features and either recombination rate or bases within recombination hot spots.

	Recombination Rate				***Recombination Hot spots***			
	125 kb	**250 kb**	**500 kb**	**1000 kb**	**125 kb**	**250 kb**	**500 kb**	**1000 kb**
mSNP density	0.502	0.564	0.622	0.641	0.377	0.45	0.508	0.556
SNP density	0.357	0.358	0.355	0.349	0.245	0.249	0.239	0.231
Repeats density	-0.255	-0.294	-0.332	-0.348	-0.178	-0.216	-0.255	-0.276
Exon density	NS	NS	0.07	0.112	NS	0.049	0.104	0.142
GC ratio	0.346	0.37	0.39	0.41	0.274	0.321	0.365	0.399
CpG density	0.27	0.308	0.353	0.402	0.212	0.267	0.333	0.389
DNA motif	0.201	0.236	0.273	0.314	0.175	0.221	0.277	0.318

[a)]Shown is the Spearman's correlation coefficient (r) of all significant (P<0.0001) correlations for four window sizes tested. NS=not significant. DNA motif is CCNCCNTNNCCNC (Myers *et al.*, 2008).

A positive partial correlation correcting for the above confounders showed a significant albeit attenuated correlation between $mSNP_{GENOME}$ and recombination rate in all window sizes tested (r=0.099-0.142, P<0.0001 for window sizes 125-1000 kb). Similarly, a positive partial correlation between $mSNP_{GENOME}$ and number of bases within recombination hot spots was observed in all window sizes tested (r=0.088-0.157, P<0.0001 for window sizes 125-1000 kb).

4.4.2 Genome-wide multiple linear regression model of homologous recombination

Multiple linear modeling allows estimation of the relative contributions from each variable while holding the other variables constant. This was used to follow up the results found by correlation. A multiple linear model was pursued using either recombination rate or bases of recombination hot spots as a response variable and mSNP in addition to confounding variables (SNP density, repeats density, exon density, GC ratio and CpG density). The $mSNP_{GENOME}$ was the fourth strongest predictor variable for recombination rate in 250 kb and 500 kb window sizes (Table 9). The variability proportion of the recombination rate explained by the linear model (R^2) was 0.337-0.523 (Table 9). Similarly, $mSNP_{GENOME}$ was the strongest predictor for bases within recombination hot spots in 250 kb and 500 kb window sizes, respectively (Table 9). The variability proportion of bases within recombination hot spots explained by the model (R^2) was 0.199-0.372 depending on window size (Table 9). The 13mer DNA motif CCNCCNTNNCCNC had a significant positive contribution to the model of recombination rate for 125 kb windows and recombination hot spots for 125 kb, 250 kb and 500 kb windows. Analysis of non-log-transformed data gave similar results, but the R^2 values of the models were lower (Data not shown).

Table 9: Multiple linear regression model of recombination rate or bases within recombination hot spots as a function of mSNP and sequence features.

	Recombination Rate				***Recombination Hot Spots***			
	125 kb	250 kb	500 kb	1000 kb	125 kb	250 kb	500 kb	1000 kb
mSNP density	0.149	0.148	0.182	0.116	0.218	0.227	0.283	0.204
SNP density	0.269	0.328	0.353	0.441	0.139	0.183	0.171	0.266
Repeat density	-0.147	-0.144	-0.141	-0.147	-0.065	-0.095	-0..110	-0.092
Exon density	-0.138	-0.179	-0.155	-0.151	-0.106	-0.109	-0.098	-0.121
GC ratio	1.030	0.232	0.229	0.203	0.470	0.157	0.112	-0.153*
CpG density	-0.934	0.269	0.351	0.540	2.652	0.125	0.186	0.358
DNA motif	0.113	0.016*	-0.046*	-0.117*	0.072	0.066	0.071	-0.023*
Model R^2	0.337	0.406	0.470	0.523	0.199	0.267	0.336	0.372

[a)]The standardized β (shown) is the number of standard deviations that the outcome variable will change as a result of one standard deviation change in the predictor variable. All variables have $P<0.0001$ except those marked with *. RM=variable was removed in modeling. DNA motif is CCNCCNTNNCCNC (Myers *et al.*, 2008).

The mSNP approach was also compared to using the ratio of observed vs. expected CpGs (CpG O/E). This is an alternative germline methylation marker. Changing the CpG count to CpG O/E in the linear models did not increase their R^2 ratio, and $mSNP_{GENOME}$ was stronger predictor of recombination rate than CpG O/E ratio (Table 10). The mSNP count was the third the strongest predictor of recombination rate in 250 kb and 500 kb window sizes (Table 10). The variability proportion of the recombination rate explained by the linear model (R^2) was 0.386-0.462 (Table 10). Similarly, $mSNP_{GENOME}$ was the strongest predictor for bases within recombination hot spots in 250 kb and 500 kb window sizes, respectively (Table 10). The variability proportion of bases within recombination hot spots explained by the model (R^2) was 0.189-0.245 depending on window size (Table 10).

Table 10: Multiple linear regression model of recombination rate or bases within recombination hot spots as a function of observed/expected CpG ratio (O/E) and sequence features.

	Recombination Rate				***Recombination Hot Spots***			
	125 kb	250 kb	500 kb	1000 kb	125 kb	250 kb	500 kb	1000 kb
mSNP density	0.097	0.129	0.140	0.187	5.210	3.917	2.030	0.129*
SNP density	0.441	0.382	0.401	0.347	7.825	4.000	1.484	1.024
Repeat density	-0.138	-0.132	-0.123	-0.121	-3.863	-1.562	-0.703	-0..201
Exon density	-0.136	-0.100	-0.050	-0.014*	-3.340	-0.847	-0.183*	-0.319*
GC ratio	0.971	0.397	0.429	0.267	19.012	2.511	0.178*	0.292*
CpG O/E	-0.071	0.059	0.057*	0.181	-0.410*	-0.069*	0.022*	0.229
DNA motif	0.001*	-0.026*	-0.065*	-0.061*	1.261	0.729*	0.866	0.265*
Model R^2	0.393	0.389	0.406	0.473	0.189	0.222	0.241	0.245

[a)]The standardized β (shown) is the number of standard deviations that the outcome variable will change as a result of one standard deviation change in the predictor variable. All variables have $P<0.0001$ except those marked with *. RM=variable was removed in modeling. DNA motif is CCNCCNTNNCCNC (Myers *et al.*, 2008).

4.4.3 High-resolution correlation between the mSNP marker of germline methylation and regional homologous recombination

The ENCODE regions comprise a detailed haplotype analysis of 5 Mb of human genome sequence. They include a threefold higher resolution of SNP information. This data set was used to study the relationship between $mSNP_{ENCODE}$ and recombination in a 25 and 50 kb window resolution.

There was a significant positive absolute correlation between the number of $mSNP_{ENCODE}$ per window and recombination rate in both window sizes tested (Fig. 23 and Table 11). The partial correlation between recombination rate and $mSNP_{ENCODE}$, after correcting for the same confounding variables as in the genome-wide approach, was significant and positive in both window sizes (25 kb: r=0.335; 50 kb: r=0.445, $P<0.0001$ for both window sizes). Similarly, the partial correlation between bases within recombination hot spots and $mSNP_{ENCODE}$ was significant and positive for both window sizes (25

kb: r=0.211, *P*=0.003; 50 kb: r=0.209, *P*=0.042).

Table 11: Absolute correlation in the ENCODE regions between mSNP and sequence features and recombination rate.

	Recombination Rate				***Recombination Hot Spots***			
	25 kb		**50 kb**		**25 kb**		**50 kb**	
	r	***P value***	**r**	***P value***	**r**	***P value***	**r**	***P value***
mSNP density	0.319	<0.0001	0.301	0.002	0.136	0.055	0.116	0.252
SNP density	-0.084	0.235	0.027	0.79	-0.109	0.123	-0.121	0.229
Repeat density	-0.05	0.481	-0.136	0.177	-0.074	0.297	-0.019	0.851
Exon density	0.061	0.392	0.045	0.656	0.162	0.022	0.161	0.11
GC ratio	0.363	<0.0001	0.211	0.035	0.268	<0.0001	0.267	0.007
CpG density	0.266	<0.0001	0.172	0.087	0.278	<0.0001	0.282	0.004

[a)]Shown is the Spearman's correlation coefficient (r) and *P* value for both window sizes tested.

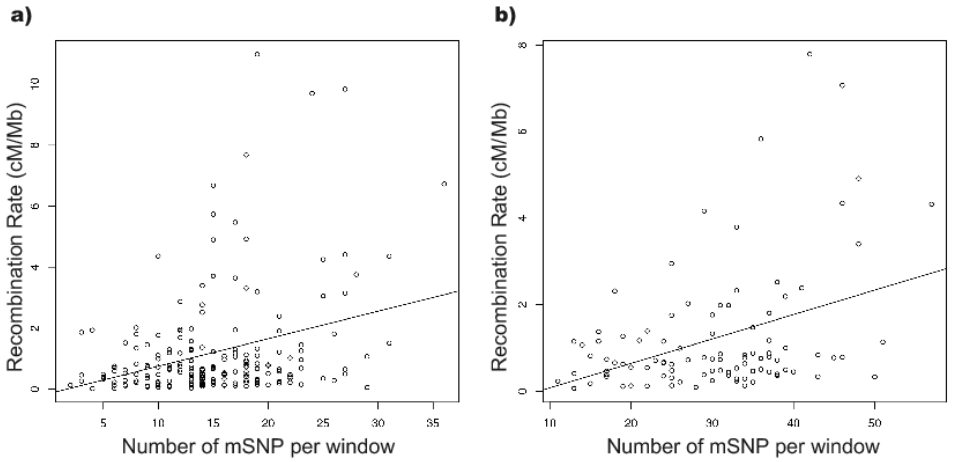


Figure 23: Correlation between mSNP and recombination rate in the ENCODE regions.

Shown is the correlation in window sizes of a) 25 kb (r=0.319 P<0.0001 and b) 50 kb (r=0.301 P=0.002).

4.4.4 Multiple linear regression model of homologous recombination in the ENCODE regions

A multiple linear model of the recombination rate as a function of $mSNP_{ENCODE}$ and sequence features for the ENCODE regions revealed that $mSNP_{ENCODE}$ was the strongest predictor of recombination for both window sizes (Table 12). A

linear model of bases within recombination hot spots lacked power because 150 out of 200 25 kb windows and 60 out of 100 50 kb windows did not have any recombination hot spot.

Table 12: Multiple linear regression model of recombination rate as a function of mSNP and sequence features for the ENCODE regions.

	25 kb		***50 kb***	
	β	***P value***	**β**	***P value***
mSNP density	0.49	<0.0001	0.661	<0.0001
SNP density	-0.293	<0.0001	-0.462	<0.0001
Repeat density	RM	RM	RM	RM
Exon density	-0.165	0.043	-0.214	0.06
GC ratio	RM	RM	RM	RM
CpG density	0.298	<0.0001	0.301	0.013
Model R^2	0.235		0.394	

4.5 Relationship between recombination and germline methylation in a biological data set (Paper III)

After establishing a positive correlation between levels of the mSNP marker and recombination, we sought to show the same relationship in a biological data set.

The Human Epigenome Project has released a data set containing the methylation status of 2524 amplicons in three different human chromosomes for 12 tissues. Each amplicon contains on average 16 CpG and the average amplicon length is 411 bp (Eckhardt *et al.*, 2006). Methylation was determined by bisulfite sequencing, the current gold standard of methylation analysis (Eckhardt *et al.*, 2006).

The data set from sperm, the final product of the male germline, was used. For each amplicon, the methylation was averaged for all CpGs. Then, the amplicons were sorted based on whether they were within a recombination hot

spots (n=219) or not (n=1745).

After averaging, the distribution was still bimodal (Fig 24a), suggesting that each amplicon was either hypomethylated or hypermethylated. The density of amplicons in the hypomethylated range of methylation was higher for the amplicons not within hot spots of recombination (Fig 24a). The average methylation of amplicons within hot spots of recombination was significantly higher than the average methylation of amplicons not within hot spots (Fig 24b).

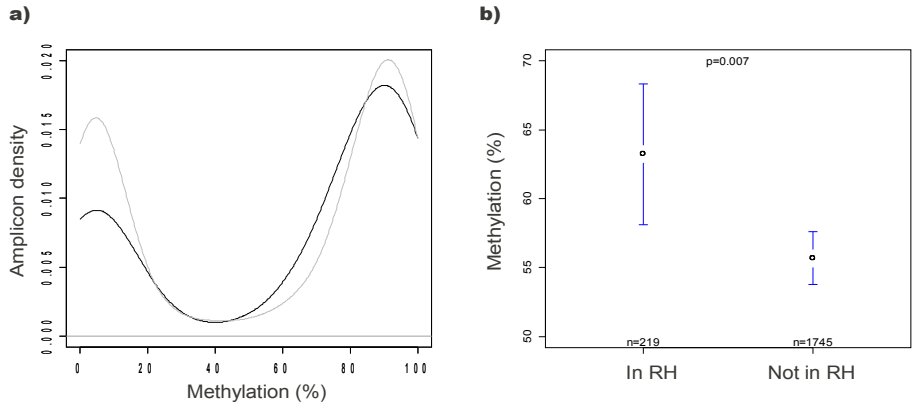


Figure 24: Methylation in sperm within or not within recombination hot spots.

a) Density plot, showing the distribution of average methylation for amplicons located within hot spots of recombination (black) or not within hot spots of recombination (gray). Density was estimated with the Gaussian smoothing kernel; b) The average methylation of amplicons within hot spots of recombination (RH) was significantly higher than the average methylation of amplicons not within hot spots of recombination (0.632 vs. 0.557, *P*=0.007) Blue bars indicate 95% confidence interval.

4.6 The density of mSNPs adjacent to imprinted genes (Unpublished)

Imprinted genes are hot spots of recombination (Sandovici *et al.*, 2006). Therefore, the mSNP counts might be higher adjacent to imprinted genes compared to non-imprinted genes, given the positive correlation between mSNPs and regional recombination rates. This was tested by comparing mSNP counts flanking either imprinted or non-imprinted genes.

The mSNP density was calculated in 125, 250, 500 and 1000 kb flanks around the center of imprinted genes. Two sets of genes were tested, genes with experimental verified imprinting status and genes computationally predicted to be imprinted.

There was a significantly higher $mSNP_{GENOME}$ density adjacent to genes with experimentally verified imprinting status in all flank sizes compared with randomly chosen genes (Fig. 25). Similarly, there was a significantly higher $mSNP_{GENOME}$ density adjacent to experimentally predicted imprinted genes compared with randomly chosen genes in three out of four flank sizes (125, 250, 500 kb) (Fig. 26). In contrast, there was no significant difference in $mSNP_{GENOME}$ density adjacent to genes expressed in all somatic cells under all conditions (housekeeping genes) compared with randomly chosen genes (data not shown).

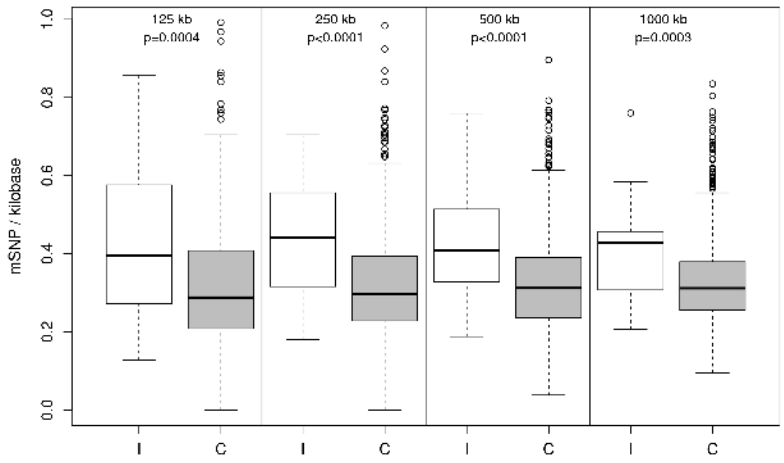


Figure 25: mSNP density of sequences flanking experimentally verified imprinted vs. random genes.

Shown are results from the four flank sizes studied. *P* values are from t-tests comparing the mSNP density (mSNP/kb) flanking either imprinted genes (n=50) (I, white) or a set of 500 genes randomly chosen from the human genome (C, gray).

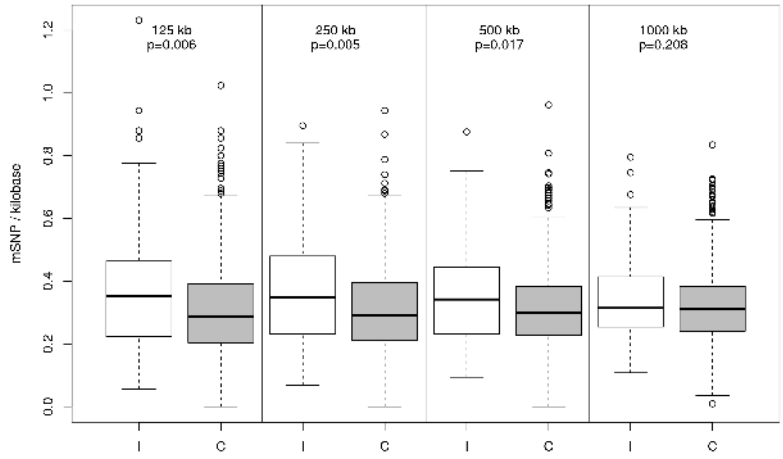


Figure 26: mSNP density of sequences flanking computationally predicted imprinted vs. random genes.

Shown are results from the four flank sizes studied. *P* values are from t-tests comparing the mSNP density (mSNP/kb) flanking either imprinted genes (n=113) (I, white) or a set of 500 genes randomly chosen from the human genome (C, gray).

4.7 Correlation between mSNP and TDR subfamilies in the human genome (Paper IV)

If a global TDR defense system based on methylation is active in the germline, then patterns of germline methylation should be shaped by the TDR pattern, even if the germline is overall hypomethylated. This might be especially evident for TDR subfamilies including active elements. The mSNP marker is suitable for testing if a relationship between germline methylation and subfamilies of TDR exists, prior to experimental verification and causal determination of such relationship. This can be done by correlation study followed by linear regression analysis of the major subfamilies as a function of variables confounding variables and the mSNP marker density in a similar manner as done in the previous chapter (Paper III).

4.7.1 Genome-wide correlation between the mSNP marker and TDR subfamilies

The correlation between $mSNP_{GENOME}$ density and the proportion of TDRs and TDR subfamilies was tested in a similar manner as the homologous recombination followed by a partial correlation correcting for confounders (GC ratio, recombination rate, gene density, SNP density, CpG dinucleotide density and repeat density). This was supplemented by multiple linear modeling of the major TDR subfamilies. The study was mainly focused on the two TDR subfamilies that currently have active elements in the human genome (L1 of the LINE and *Alu* of the SINE). In addition, a subset of about 11,000 repeat elements (*Alu*, L1 and SVA elements) that are differentially present in the human and chimpanzee genome, indicating that they include active elements, was used (Nichol & Pearson, 2002). Those elements were collectively termed active elements.

There was a strong negative correlation, both absolute and partial,

between the $mSNP_{GENOME}$ density and proportion of TDRs in all window sizes tested (Table 13). This was mostly explained by a strong negative correlation between $mSNP_{GENOME}$ density and proportion of *Alu* elements in all window sizes (Table 13). In contrast, there was a significant negative correlation between $mSNP_{GENOME}$ density and proportion of L1 elements only in the largest window size tested (1000 kb). In other window sizes any observed absolute correlation was eliminated by correcting for confounders (Table 13).

Correlation patterns for long terminal repeats (LTR) was similar to L1 element patterns in all window sizes. There was a significant positive correlation between the $mSNP_{GENOME}$ marker and proportion of simple repeats in all window sizes tested (Table 13). These results validate the approach as the methylation of simple repeats is critical for genome stability, at least in somatic cells (Nichol & Pearson, 2002). The proportion of active elements had a significant negative correlation with the $mSNP_{GENOME}$ marker in all window sizes similar to the *Alu* elements (Table 13).

Table 13: Genome-wide correlation between TDR families and the mSNP marker of germline methylation.

	125 kb		*250 kb*		*500 kb*		*1000 kb*	
	Absolute	**Partial**	**Absolute**	**Partial**	**Absolute**	**Partial**	**Absolute**	**Partial**
Total	-0.29****	-0.15****	-0.32****	-0.18****	-0.35****	-0.19***	-0.44****	-0.39****
LINE	-0.20****	-0.01	-0.23****	-0.01	-0.28****	-0.02	-0.43****	-0.18***
L1	-0.21****	-0.01	-0.25****	-0.01	-0.29****	-0.01	-0.45****	-0.17***
L2	0.09***	-0.02*	0.09***	-0.02	0.09***	-0.03	0.10**	0.03
SINE	0.02	-0.18****	0.04**	-0.21****	0.08**	-0.22****	0.20***	-0.18***
MIR	0.30****	0	0.32****	-0.01	0.34****	-0.01	0.37****	0.06
AluY	-0.16****	-0.14****	-0.16****	-0.16****	-0.12***	-0.16***	-0.03	-0.20***
AluJ	-0.06***	-0.12****	-0.04**	-0.15****	0	-0.15***	0.12**	-0.12**
AluS	-0.04**	-0.14****	-0.02	-0.17****	0.03	-0.17***	0.14***	-0.18***
LTR	0.02	-0.01	0	0	-0.02	0.03	-0.20***	-0.14***
DNA transposons	0.04**	0.01	0.05**	0.02	0.07**	0.01	0.08*	0.02
Simple repeats	0.18****	0.11****	0.22****	0.16****	0.26****	0.18***	0.25***	0.14***
Active elements	-0.10***	-0.15****	-0.08***	-0.18****	-0.03	-0.18***	0.09**	-0.21***

[a)]Shown are results for all window sizes tested, both absolute correlation and partial correlation. In partial correlations correction was made for GC ratio, recombination rate and density of SNPs, CpG dinucleotides and exons. Number indicates Spearman's correlation coefficients, and levels of statistical significance are marked with asterisks; $P<10^{-50}$ (****), $P<10^{-10}$ (***), $P<10^{-5}$(**) and $P<0.002$ (*). Given multiple testing $P<0.002$ is the lowest level of significant correlation

4.7.2 Multiple linear regression of the density of major TDR subfamilies

A multiple linear regression model of *Alu* element proportions as a function of $mSNP_{GENOME}$ and confounding variables was pursued. There was a consistent negative contribution from $mSNP_{GENOME}$ to the model in all window sizes, with $mSNP_{GENOME}$ being the second strongest predictor of *Alu* element proportions in all windows tested (Table 14). The amount of variability explained by the models (R^2) was 0.530-0.769 based on window size (Table 14).

Table 14: Multiple linear regression model of the Alu elements proportion as a function of mSNP and sequence features for the whole genome.

Alu	***1000 kb***		***500 kb***		***250 kb***		***125 kb***	
	β	***P* value**	**β**	***P* value**	**β**	***P* value**	**β**	***P* value**
mSNP density	-0.28	$< 10^{-15}$	-0.33	$< 10^{-15}$	-0.26	$< 10^{-15}$	-0.22	$< 10^{-15}$
SNP density	0.11	$3 \cdot 10^{-10}$	0.20	$< 10^{-15}$	0.16	$< 10^{-15}$	0.13	$< 10^{-15}$
Exon density	0.13	$< 10^{-15}$	0.14	$< 10^{-15}$	0.13	$< 10^{-15}$	0.12	$< 10^{-15}$
GC ratio	-0.17	$7 \cdot 10^{-10}$	-0.21	$< 10^{-15}$	-0.21	$< 10^{-15}$	0.04	$1 \cdot 10^{-7}$
CpG density	0.98	$< 10^{-15}$	1.00	$< 10^{-15}$	0.90	$< 10^{-15}$	0.63	$< 10^{-15}$
Recombination Rate	-0.10	$7 \cdot 10^{-13}$	-0.10	$< 10^{-15}$	-0.10	$< 10^{-15}$	-0.11	$< 10^{-15}$
Non-*Alu* elements proportion	-0.18	$< 10^{-15}$	-0.12	$< 10^{-15}$	-0.13	$< 10^{-15}$	-0.12	$< 10^{-15}$
Model R^2		0.769		0.720		0.643		0.530

[a)]Shown is the standardized β value and corresponding *P* value. The β is the number of standard deviations that the outcome variable will change as a result of one standard deviation change in the predictor variable.

A multiple linear regression model of L1 elements revealed a significant negative contribution from $mSNP_{GENOME}$ density to the model in the 1000 kb windows, where $mSNP_{GENOME}$ density was the third strongest predictor of L1 elements proportion. For 125 kb windows, $mSNP_{GENOME}$ density was the sixth strongest predictor and the contribution was insignificant in other window sizes (Table 15). The amount of variability explained by the models (R^2) was 0.384-0.607 based on window size (Table 15).

Table 15: Multiple linear regression model of the L1 elements proportion as a function of mSNP and sequence features for the whole genome.

L1	***1000 kb***		***500 kb***		***250 kb***		***125 kb***	
	β	***P* value**	**β**	***P* value**	**β**	***P* value**	**β**	***P* value**
mSNP density	-0.24	$< 10^{-15}$	-0.02	0.45	-0.03	0.07	-0.10	$< 10^{-15}$
SNP density	-0.14	$6 \cdot 10^{-11}$	-0.01	0.77	-0.01	0.50	0.03	0.002
Exon density	-0.03	0.048	-0.04	0.002	-0.08	$< 10^{-15}$	-0.10	$< 10^{-15}$
GC ratio	-0.01	0.863	-0.12	$1 \cdot 10^{-15}$	-0.04	0.02	0.42	$< 10^{-15}$
CpG density	-0.32	$2 \cdot 10^{-15}$	-0.41	$< 10^{-15}$	-0.45	$< 10^{-15}$	-0.64	$< 10^{-15}$
Recombination Rate	-0.13	$9 \cdot 10^{-13}$	-0.12	$< 10^{-15}$	-0.11	$< 10^{-15}$	-0.12	$< 10^{-15}$
Non-L1 elements proportion	-0.30	$< 10^{-15}$	-0.20	$< 10^{-15}$	-0.17	$< 10^{-15}$	-0.15	$< 10^{-15}$
Model R^2		0.607		0.534		0.462		0.384

Results for the active elements were similar to the results for *Alu* elements. There was a significant negative contribution from $mSNP_{GENOME}$ density to the model of active elements. For all window sizes, $mSNP_{GENOME}$ was the third strongest predictor of active element proportions (Table 16). The amount of variability explained by the models (R^2) was 0.483-0.730 based on window size (Table 16).

Neither analysis with a non-log-transformed data nor exchanging exon density for gene density changed the magnitude or statistical significance of any results (data not shown).

Table 16: Multiple linear regression model of the active elements proportion as a function of to mSNP and sequence features for the whole genome.

Active elements	***1000 kb***		***500 kb***		***250 kb***		***125 kb***	
	β	***P* value**	**β**	***P* value**	**β**	***P* value**	**β**	***P* value**
mSNP density	-0.29	$< 10^{-15}$	-0.36	$< 10^{-15}$	-0.29	$< 10^{-15}$	-0.26	$< 10^{-15}$
SNP density	0.10	$1 \cdot 10^{-7}$	0.21	$< 10^{-15}$	0.18	$< 10^{-15}$	0.15	$< 10^{-15}$
Exon density	0.10	$2 \cdot 10^{-15}$	0.10	$< 10^{-15}$	0.10	$< 10^{-15}$	0.09	$< 10^{-15}$
GC ratio	-0.29	$< 10^{-15}$	-0.29	$< 10^{-15}$	-0.30	$< 10^{-15}$	-0.10	$< 10^{-15}$
CpG density	1.19	$< 10^{-15}$	1.20	$< 10^{-15}$	1.09	$< 10^{-15}$	0.81	$< 10^{-15}$
Recombination Rate	-0.10	$1 \cdot 10^{-12}$	-0.10	$< 10^{-15}$	-0.11	$< 10^{-15}$	-0.12	$< 10^{-15}$
Non-Active elements proportion	-0.01	0.315	0.46	$9 \cdot 10^{-6}$	0.05	$6 \cdot 10^{-11}$	0.08	$< 10^{-15}$
Model R^2		0.730		0.668		0.578		0.483

4.8 Analysis of TDR subfamilies flanking differentially methylated regions in a biological data set (Paper IV)

To study the relationship of TDR subfamilies and methylation on a smaller scale in an independent data set, we used the HEP dataset. This data set includes the results from bisulfite sequencing of a small section of the genome for methylation analysis in a number of tissues.

The 2,524 amplicons from the HEP sperm methylation data set were divided into two groups; hypermethylated amplicons (>80% methylation, 1001 amplicons) and hypomethylated amplicons (<20% methylation, 701 amplicons). The methylation criteria was based on the initial analysis of the data of the HEP group (Eckhardt *et al.*, 2006). Four different flank lengths (3, 5, 10 and 15 kb) were extracted from the human sequence for each amplicon and submitted to the CENSOR server searching the sequence for TDRs. Results of the TDR analysis were then compared between the hypermethylated and hypomethylated amplicon group. To estimate a *P* value, the amplicons were randomly divided into two groups including 1001 and 701 amplicons, to

simulate no effects of methylation on flanking TDRs. The observed difference in TDR subfamilies were compared against the simulated difference and an *P* value estimated for each TDR subfamily.

The proportion of *Alu* elements flanking hypermethylated amplicons was significantly lower than the proportion flanking hypomethylated amplicons (Fig. 27a, Table 17). There was a significantly higher proportion of L1 elements flanking hypermethylated amplicons compared to hypomethylated amplicons only in 3 and 5 kb flanks (Fig. 27b, Table 17). For the other two flank sizes tested (10 and 15 kb) the differences were not significant. Results for LTR elements had a similar trend as the results for L1 elements, but they were not statistically significant (Table 17).

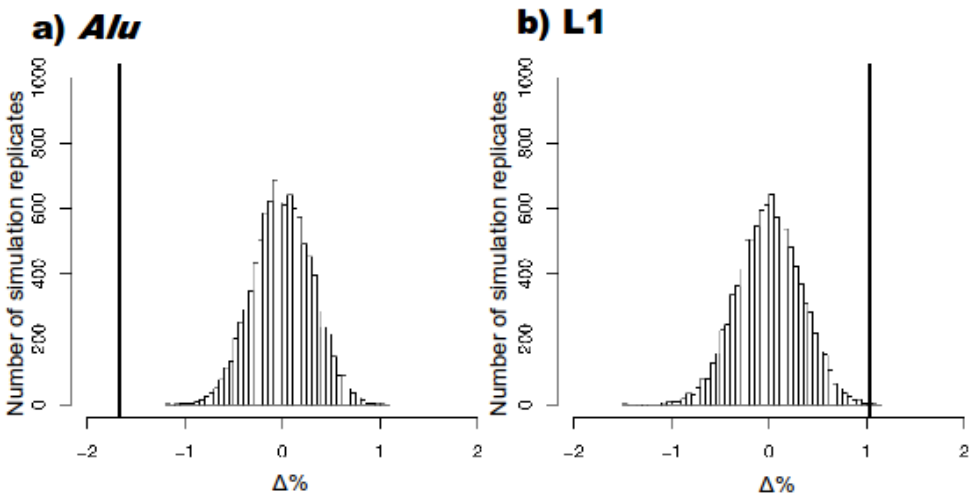


Figure 27: The difference ($\Delta\% = \%_{hypermethylated} - \%_{hypomethylated}$) between absolute proportion of repeats flanking hypermethylated and hypomethylated amplicons (vertical line) against the distribution of 10,000 permutations of the data regardless of methylation (bell curve). Shown is the difference for a) Alu and b) L1 repeat families.

The observed difference (Δ%) (vertical line, -1.7% and 1.0% for *Alu* and L1, respectively) was compared against the distribution of Δ% when the data was sampled randomly into two groups regardless of methylation (bell curve). The observed differences occurred 0 and 4 times in 10,000 simulations, giving estimated *P* values of <0.0001 and 0.0004 for *Alu* and L1 elements.

Table 17: Proportion of TDR subfamilies in 3-15 kb flanking hypermethylated and hypomethylated amplicons from the HEP data set.

	3 kb flanks						***5 kb flanks***					
	Hyperm.		**Hypom.**				**Hyperm.**		**Hypom.**			
Family	**kb**	**%**	**kb**	**%**	**Δ %**	***P* value**	**bp**	**%**	**bp**	**%**	**Δ %**	***P* value**
Total	1,183	19.9	764	18.6	1.4	0.0098	1,860	18.6	1,281	18.4	0.2	0.6862
LINE	331	5.6	167	4.1	1.5	<0.0001	568	5.7	303	4.3	1.3	0.0001
L1	230	3.9	102	2.0	1.4	<0.0001	406	4.1	212	3.0	1.0	0.0004
L2	88	1.5	58	1.4	0.1	0.3496	137	1.4	79	1.1	0.2	0.0253
SINE	502	8.5	383	9.3	-0.8	0.0126	708	7.1	619	8.9	-1.8	<0.0001
Alu	414	7.0	327	7.9	-1.0	0.0041	608	6.1	540	7.7	-1.7	<0.0001
MIR	72	1.2	42	1.0	0.2	0.0106	73	0.7	46	0.7	0.1	0.0829
LTR	193	3.2	106	2.6	0.7	0.0061	323	3.2	198	2.8	0.4	0.0514
DNA transp.	76	1.3	53	1.3	0.0	0.4625	116	1.2	71	1.0	0.1	0.1331
Simple rep.	81	1.4	54	1.3	0.1	0.284	144	1.4	88	1.3	0.2	0.0186
Analyzed bases	5,934		4,116				10,010		6,970			
	10 kb flanks						**15 kb flanks**					
	Hyperm.		**Hypom.**				**Hyperm.**		**Hypom.**			
	kb	**%**	**kb**	**%**	**Δ %**	***P* value**	**bp**	**%**	**bp**	**%**	**Δ %**	***P* value**
Total	3,190	15.9	2,357	16.8	-0.9	0.0176	4,333	14.4	3,236	15.4	-1.0	0.038
LINE	1,074	5.4	641	4.6	0.8	0.0051	1,535	5.1	939	4.5	0.6	0.0642
L1	840	4.2	492	3.5	0.7	0.0113	1,246	4.2	749	3.6	0.6	0.0785
L2	189	0.9	125	0.9	0.1	0.2362	231	0.8	154	0.7	0.0	0.3458
SINE	1,084	5.4	1,038	7.4	-2.0	<0.0001	1,398	4.7	1,365	6.5	-1.8	<0.0001
Alu	946	4.7	915	6.5	-1.8	<0.0001	1,220	4.1	1,212	5.8	-1.7	<0.0001
MIR	85	0.4	49	0.3	0.1	0.0049	91	0.3	54	0.3	0.0	0.0801
LTR	552	2.8	375	2.7	0.1	0.3254	697	2.3	502	2.4	-0.1	0.3986
DNA transp.	162	0.8	115	0.8	0.0	0.4425	197	0.7	137	0.7	0.0	0.4849
Simple rep.	310	1.5	186	1.3	0.2	0.0001	495	1.6	289	1.4	0.3	0.0008
Analyzed bases	20,020		14,020				30,030		21,030			

Shown are both absolute numbers (kb) and percentages (%). The absolute difference between hypermethylated (Hyperm.) and hypomethylated (Hypom.) amplicons was compared against 10,000 permutations of the data to estimate a *P* value.

4.9 A network analysis of the metabolic effects of human imprinted genes (Paper V)

DNA methylation is involved in the control of the expression of imprinted genes. Haig's hypothesis of anabolic effects of paternally imprinted genes and catabolic effects of maternally imprinted genes on metabolism has limited support. The available data is only based on a handful of imprinted genes. We were interested in applying methods of metabolic systems biology to test the effects of expression changes of imprinted genes on human metabolism. Additionally, these methods might be used to test dosage sensitivity of other genes, and even to suggest novel imprinted genes. Unfortunately, no positive control was found for either anabolic or catabolic phenotype. However, a mouse metabolic reconstruction highly homologous to the human reconstruction has been successfully used to predict both gene essentiality and softer phenotypes (Sigurdsson *et al.*, 2010).

A list of experimentally verified and computationally predicted imprinted genes was crossed against a list of 1496 genes in the reconstructed human metabolic network. A total of three experimentally confirmed imprinted genes (*ATP10A, SLC22A2,* and *SLC4A2*) and six computationally predicted imprinted genes were found (*CYP1B1, FUCA1, GPT1, NDUFA4, PPAP2C* and *SLC4A2*). Allowing cellular uptake of minimum medium (essential amino acids, essential fatty acids, glucose, oxygen, sulphate, phosphate and vitamins), three epigenotypes were simulated as described in Materials and Methods (Chapter 3.4.2). The simulations tested expression of no copy, one copy (wild type) and two copies of each gene. Flux variability analysis (FVA) of each model was then performed and the results from each FVA compared against the wild type FVA. The absolute number of reactions with increased or decreased flux capacity compared to the wild type were then counted. Counts within 97

subsections of metabolism (Table 18) were then compared to suggest a resulting phenotype.

The average number of subsections of metabolism with a significant change was 15 when no expression was simulated, compared to 16 subsections changing when expression of both alleles was simulated (Table 19).

In general, the results were lopsided and consistent for most of the simulated genes. Generally, simulating no expression caused increased flux capacity within many metabolism subgroups and simulating the expression of two copies resulted in decreased flux capacity within many subgroups (Table 19). The gene with the greatest metabolic perturbation resulting from simulating expression changes was *ATP10A,* the only gene of the simulated genes with a known clinical phenotype from abnormal expression. Simulation of no expression resulted in significant changes within 27 metabolic subsections, and simulation of the expression of both alleles resulted in significant changes within 29 subsections (Table 19). Simulation of no gene expression of six genes (*CYP1B1, GPT1, PPAP2C, SLC22A2, SLC22A3,*and *FUCA1*) revealed increased flux capacity for metabolic subsections within structural carbohydrates (such as keratan sulfate and N-glycan pathways) and lipid metabolism, while simulation of both allele expression revealed an opposite flux pattern. This might be consistent with an anabolic and a catabolic phenotype resulting from no expression and expression of both alleles, respectively. Simulation results of no expression of the *NDUFA4* gene revealed decreased flux capacity for lipid and structural carbohydrate metabolism while simulation of both allele expression revealed increased flux capacity for structural carbohydrate metabolism but decreased flux capacity for lipid metabolism. The overall effect was therefore hard to predict. Epigenotype simulation for *SLC4A2* suggested negligible effects of the epigenotypes on

metabolism.

The loss of expression simulation results for *ATP10A* are shown in Figure 28. No subsections had decreased flux capacity, whereas 27 had increased flux capacity compared to wild type. Of those, 10 subsections were within lipid metabolism and eight within carbohydrate metabolism, including structural carbohydrates such as keratan sulfate. In that perspective it is likely that an anabolic phenotype resulting in obesity might result from no expression of the *ATP10A.* This is supported from several literature sources on the phenotype. A maternal deletion of the homologous gene in mice results in increased body fat (Dhar *et al.*, 2000). The mouse model of Angelman syndrome (including loss of expression of *ATP10A*) has an obese phenotype (Cattanach *et al.*, 1997). Furthermore, a subset of patients with Angelman syndrome have an obese phenotype resembling Prader-Willi syndrome, and this phenotype is suggested to result from absent ATP10A expression (Gillessen-Kaesbach *et al.*, 1999; Meguro *et al.*, 2001).

When simulation results were reviewed in light of Haig's parental conflict theory, the metabolic profiles from the simulation of four genes (*ATP10A, GPT1, NDUFA4, PPAP2C*) fitted the theory while the profiles from four other genes (*SLC22A2, SLC22A3, CYP1B1, FUCA1*) did not (P=1.0).

Table 18: Metabolic subgroups in Recon 1

ID	Name	ID	Name	ID	Name	ID	Name
1	Alanine and Aspartate Metabolism	26	Fatty Acid Metabolism	51	Miscellaneous	76	Steroid Metabolism
2	Alkaloid biosynthesis II	27	Fatty acid oxidation	52	N-Glycan Biosynthesis	77	Stilbene, coumarine and lignin biosynthesis
3	Aminosugar Metabolism	28	Fatty acid oxidation, peroxisome	53	N-Glycan Degradation	78	Taurine and hypotaurine metabolism
4	Arginine and Proline Metabolism	29	Folate Metabolism	54	NAD Metabolism	79	Tetrahydrobiopterin
5	Ascorbate and Aldarate Metabolism	30	Fructose and Mannose Metabolism	55	Nucleic acid degradation	80	Thiamine Metabolism
6	beta-Alanine metabolism	31	Galactose metabolism	56	Nucleotide Sugar Metabolism	81	Transport, Endoplasmic Reticular
7	Bile Acid Biosynthesis	32	Glutamate metabolism	57	Nucleotides	82	Transport, Endoplasmic Reticular
8	Biotin Metabolism	33	Glutathione Metabolism	58	O-Glycan Biosynthesis	83	Transport, Extracellular
9	Blood Group Biosynthesis	34	Glycerophospholipid Metabolism	59	Others	84	Transport, Golgi Apparatus
10	Butanoate Metabolism	35	Glycine, Serine, and Threonine Metabolism	60	Oxidative Phosphorylation	85	Transport, Lysosomal
11	C5-Branched dibasic acid metabolism	36	Glycolysis/Gluconeogenesis	61	Pentose and Glucuronate Interconversions	86	Transport, Mitochondrial
12	Carnitine shuttle	37	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	62	Pentose Phosphate Pathway	87	Transport, Nuclear
13	Cholesterol Metabolism	38	Glyoxylate and Dicarboxylate Metabolism	63	Phenylalanine metabolism	88	Transport, Peroxisomal
14	Chondroitin / heparan sulfate biosynthesis	39	Heme Biosynthesis	64	Propanoate Metabolism	89	Triacylglycerol Synthesis
15	Chondroitin sulfate degradation	40	Heme Degradation	65	Purine Catabolism	90	Tryptophan metabolism
16	Citric Acid Cycle	41	Heparan sulfate degradation	66	Pyrimidine Biosynthesis	91	Tyr, Phe, Trp Biosynthesis
17	CoA Biosynthesis	42	Histidine Metabolism	67	Pyrimidine Catabolism	92	Tyrosine metabolism
18	CoA Catabolism	43	Hyaluronan Metabolism	68	Pyruvate Metabolism	93	Ubiquinone Biosynthesis
19	CYP Metabolism	44	IMP Biosynthesis	69	R Group Synthesis	94	Urea cycle/amino group metabolism
20	Cysteine Metabolism	45	Inositol Phosphate Metabolism	70	Riboflavin Metabolism	95	Valine, Leucine, and Isoleucine Metabolism
21	D-alanine metabolism	46	Keratan sulfate biosynthesis	71	ROS Detoxification	96	Vitamin A Metabolism
22	D-arg and D-orn metabolism	47	Keratan sulfate degradation	72	Salvage Pathway	97	Vitamin B12 Metabolism
23	Eicosanoid Metabolism	48	Limonene and pinene degradation	73	Selenoamino acid metabolism	98	Vitamin B6 Metabolism
24	Fatty acid activation	49	Lysine Metabolism	74	Sphingolipid Metabolism	99	Vitamin D
25	Fatty acid elongation	50	Methionine Metabolism	75	Starch and Sucrose Metabolism		

Table 19: Results from epigenotype simulation of the nine imprinted genes found.

Gene symbol	***A***	***IK***	***Epigenotype I, no copy expressed***		***Epigenotype III, both copies expressed***	
			Decreased flux	**Increased flux**	**Decreased flux**	**Increased flux**
ATP10A	M	E	None	4, 8, 12, 13, 25, 26, 27, 28, 29, 30, 33, 34, 35, 38, 45, 46, 47, 52, 53, 57, 58, 61, 68, 69, 75, 76, 85	7, 8, 9, 12, 13, 23, 25, 27, 28, 29, 33, 34, 38, 44, 45, 46, 47, 52, 53, 57, 58, 61, 68, 69, 74, 75, 76, 85, 95	95
SLC22A2	P	E	7, 74, 9	68, 26, 25, 69, 34, 12, 57, 29, 13, 76, 45, 46, 47, 52	86, 27, 34, 29, 25, 69, 13, 7, 57, 76, 9, 45, 12, 52, 46, 74, 47	None
SLC22A3	P	E	7, 74, 9	68, 26, 25, 69, 34, 12, 57, 29, 13, 76, 45, 46, 47, 52	86, 27, 34, 29, 25, 69, 13, 7, 57, 76, 9, 45, 12, 52, 46, 74, 47	None
CYP1B1	P	C	None	52, 47, 46, 45, 57, 76, 12, 34, 29, 7, 69, 25	52, 47, 74, 46, 12, 45, 9, 57, 76, 7, 69, 34, 25, 29, 27	None
FUCA1	P	C	47	68, 27, 26, 35, 25, 69, 34, 57, 29, 13, 76, 12, 45, 52	68, 27, 57, 29, 35, 69, 34, 25, 13, 76, 7, 52, 45, 12	47
GPT1	M	C	None	35, 86, 69, 25, 34, 29, 92, 7, 57, 12, 76, 45, 9, 46, 74, 47, 52	27, 35, 29, 25, 69, 34, 13, 7, 57, 76, 9, 45, 12, 46, 74, 47, 52	None
NDUFA4	P	C	26, 35, 76, 29, 34, 13, 57, 45, 46, 47, 52	25, 12	27, 76, 13, 74, 69, 25, 7, 9, 12, 26	95, 29, 45, 46, 47, 52
PPAP2C	M	C	None	85, 25, 69, 74, 35, 29, 34, 12, 13, 76, 57, 45, 46, 47, 52	27, 35, 29, 7, 57, 34, 25, 69, 13, 12, 76, 45, 52, 46, 47	9, 74
SLC4A2	M	C	46	52	46	52

[a)]Shown is the expressed allele (A; M=maternal, P=paternal), the imprinting information (IK; E=experimentally verified, C=computationally predicted) and the ID number of metabolic subsections (Table 18) with a decreased or increased flux capacity.

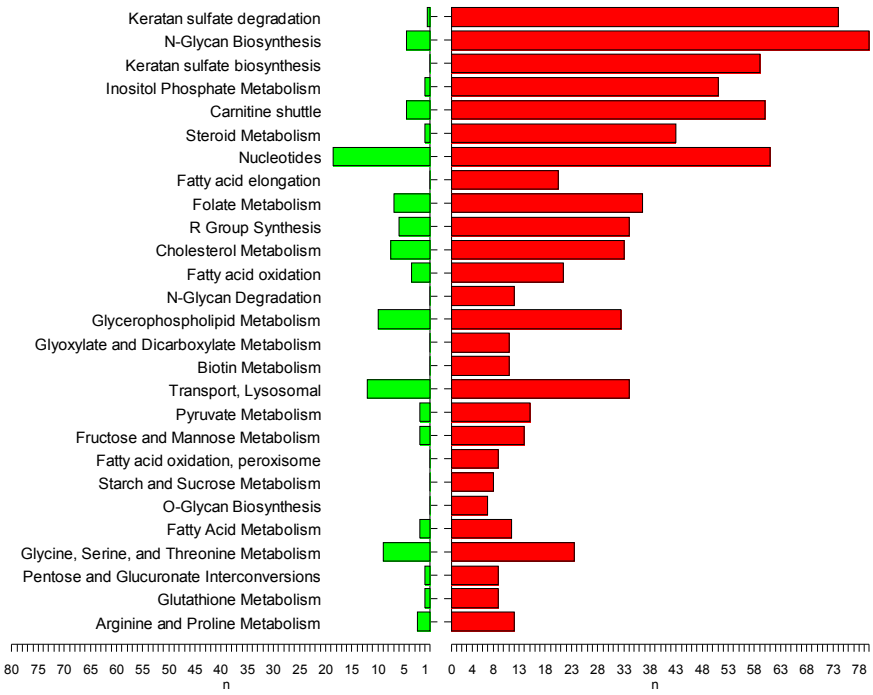


Figure 28: Simulation results for no expression of ATP10A gene.

Absolute number of reactions with increased flux capacity (red) and decreased flux capacity (green) compared to wild type. In total, 27 metabolic subsections changed significantly and are shown in the figure.

5 DISCUSSION

5.1 Summary of results

In this Ph.D. thesis, several aspects of DNA methylation in the human genome were explored. In paper I, the target sequence properties of restriction endonucleases suitable for global methylation analysis in the human genome were analyzed. The sequence specificity properties of two target sequences suitable for global CpG methylation analysis and one target sequence suitable for global CWG methylation analysis were described. The results aid in interpretation of measurements of global methylation with restriction endonucleases. In Paper II, longitudinal changes in both global and gene promoter methylation in two cohorts were measured. The observed change in both cohorts was bi-directional; while some individuals lost methylation between measurements others gained methylation. A familial component in the conservation of methylation was also observed. In paper III, a novel surrogate bioinformatic marker for germline methylation in the human genome was developed. After validation of the marker, a positive correlation between homologous recombination and the marker was found. In paper IV, the marker was used to test the correlation between germline methylation and subfamilies of TDRs, in particular subfamilies with active elements. This revealed a negative correlation between DNA methylation and the *Alu* subfamily, but a variable correlation for the L1 subfamily based on resolution. In paper V, the metabolic properties of imprinted genes were evaluated with the methods of systems biology. Simulations using reconstruction of the human metabolism predicted that the metabolic effects of the imprinted genes whose expression was simulated in the study were generally lopsided and consistent. Out of the nine simulated genes, the single gene with a known clinical phenotype resulted in the greatest metabolic perturbation.

5.2 Sequence specificity of restriction endonucleases suitable for global methylation analysis in the human genome (Paper I)

Changes in global methylation, especially in repetitive elements, are cardinal features of tumorigenesis (Esteller, 2008). They have recently been shown to predict clinical outcomes in patients with acute myelogenous leukemia (Deneberg *et al.*, 2010). Global methylation changes might therefore become a tumor marker suitable for classifying and staging selected cancers. Methods for quantifying global methylation could subsequently become clinically useful, provided they are accurate, robust, reproducible, and scalable to process a large amount of samples with reasonable resources.

Several assays for global methylation analysis are available (Fraga & Esteller, 2002). Global methylation assays based on restriction endonucleases, especially when coupled with accurate and fast quantification methods such as pyrosequencing, are feasible for high-throughput and reproducible results (Karimi *et al.*, 2006a; Karimi *et al.*, 2006b).

When interpreting the results from restriction endonuclease assays and comparing with results from other methods, knowledge of the target sequence specificity of the restriction endonuclease pair used is helpful. This was important for understanding the results on the longitudinal change in global methylation that was measured in Paper II.

In Paper I, the commercially available restriction endonuclease pairs available for global methylation analysis were systematically reviewed and their target sequence frequencies in various subsets of the human genome estimated. As expected, the target sequence that has generally been used for evaluating methylation with restriction based endonucleases (CCGG, exemplified by the *Hpa*II/*Msp*I isoschizomers and used in LUMA), had several attractive features. These include a high frequency of the target sequence in the human genome and

an over-representation in interesting subsets of the genome such as promoters, exons and CpG islands. Approximately 53% of the target sequences are within repeats, with an over-representation in LINE repeats and an underrepresentation in SINE repeats.

The analysis also identified a second target sequence, (GCGC, exemplified by the *Hha*I/*Cfo*I isoschizomers) that has several attractive properties, including a similar overall genome frequency as the CCGG target sequence. In addition, the GCGC sequence has more over-representation in several interesting subsets of the human genome such as promoter regions and exons. A double digest by restriction endonucleases targeting both CCGG and GCGC sequences probes the methylation of approximately 14% of all CpGs in the human genome.

The discovery of non-CpG methylation in the CWG in embryonic stem cells, where 25% of all methylated cytosines reside within CWG trinucleotides (Lister *et al.*, 2009), calls for development of experimental methods for its analysis. One target sequence, CCWGG, seems suitable for global methylation analysis of the CWG trinucleotide. It seems to be fairly homogeneously represented within repeat sequences and gene-related sequences. However, it shares the observed over-representation of the target sequence near chromosome ends with the other target sequences mapped. Recently, endonucleases targeting CCWGG were used in a modified LUMA protocol to probe non-CpG methylation in myocytes in various stages of differentiation (Barrès *et al.*, 2009).

The results from Paper I aid in selecting restriction endonucleases suitable for global methylation analysis, and interpreting the results of such assays. Further experiments should compare experimentally measurements of

global CpG methylation by endonucleases targeting the two suggested target sequences, and determine the feasibility of using them together. The target sequence including the CWG trinucleotide can be used in a global CWG methylation assay to further the understanding of this recently described stem cell mechanism. Finally the programs written in Paper I could be used to analyze the frequencies of other target sequences in other organisms.

5.3 Intra-individual changes in DNA methylation with time and assessment of familial clustering (Paper II)

An important prerequisite of the epigenetic model of complex human disease is the suggestion of acquired change in epigenetic marks (Bjornsson *et al.*, 2004). At the time of the publication of the model, limited data existed supporting this prerequisite. Most results were based on cross-sectional populations (Issa *et al.*, 1994; Issa *et al.*, 1996) and only a handful of CpG sites (Issa *et al.*, 1994; Issa *et al.*, 1996; Sandovici *et al.*, 2003) or levels of X inactivation (Busque *et al.*, 1996; Racchi *et al.*, 1998; Sandovici *et al.*, 2004). The most comprehensive study available on epigenetic changes with age was based on 40 MZ twins, indicating significant changes in global DNA methylation and patterns of histone modifications (Fraga *et al.*, 2005).

The results from Paper II substantially added to the picture of age-associated changes of epigenetic marks. The study was based on longitudinal data sets of DNA extracted from blood sampled at two different time points with an 11-16 year interval. Individuals came from two different cohorts of 111 individuals from Iceland and 126 individuals from Utah. For each individual, global methylation was compared between the two time points by measuring the methylation of the CCGG target sequence with LUMA. This revealed more than 5% change in global methylation in 40% and 63% of the population from the Utah and Iceland cohorts, respectively. The methylation changed by more

than 10% in 10% and 30% of the population from the Utah and Icelandic cohorts, respectively. One explanation of the discrepancies between the populations could be the fact that the Icelandic population was significantly older than the Utah population. If true, this would suggest an increasing rate of acquired changes in DNA methylation with age.

Interestingly, the results were bi-directional; whereas some individuals gained methylation others lost methylation between the two time points. This finding has been replicated by other researchers, both in mice (Maegawa *et al.*, 2010) and humans (Wong *et al.*, 2010). The observed bi-directionality of the changes could explain why a large cross-sectional study of human methylation failed to demonstrate any changes, as the average change in a cross-sectional cohort bidirectional change can become zero (Eckhardt *et al.*, 2006).

Since the Utah samples were from families with samples available for up to three generations, conservation of methylation within families could be tested. This revealed a high correlation of conservation of methylation within members of the same family. This could be due to genetic factors or shared environment. An example of an environmental factor that might be shared in family members could be availability of factors critical for maintenance of single carbon metabolism, such as folic acid. A genetic component to the conservation of methylation is extremely interesting. Candidate genes could be one of the human methyltransferase genes, in particular the *DNMT1* gene responsible for maintenance methylation. An interesting follow-up study would be to combine measurements of conservation of DNA methylation and typing of polymorphisms near or within the genes participating in maintenance of methylation.

A microarray based experiment probing the methylation of 807 CpG

dinucleotides within gene promoters established the same trends as found with the global methylation analysis. This argues against suggestions that acquired changes in DNA methylation only occur in non-coding regions. There were more imprinted and immunological genes within the genes with the greatest methylation changes than expected by chance. The enrichment of imprinted genes is interesting in light of the metabolic perturbations resulting from changes in expression of imprinted genes presented in Paper V. Perhaps changes in promoter methylation of the dosage-sensitive imprinted metabolic genes could contribute to acquired metabolic disease, such as diabetes mellitus and obesity.

The results support an important prerequisite of a model of complex disease pathogenesis based on genetic and epigenetic mechanisms, that epigenetic marks undergo acquired changes (Bjornsson *et al.*, 2004). These changes might in part explain the late onset of many common diseases, such as cancer. The model is consistent with other well known mechanisms contributing to acquired disease (including cancer pathogenesis), such as acquired somatic mutations. For example, epigenetic mechanisms that might explain the late onset of many types of cancer might involve both increased expression of oncogenes (perhaps modified by loss of inhibition by methylation) and silencing of tumor suppressor genes (perhaps modified by methylation of their promoter regions).

The longitudinal results have since publication been supported by several other studies. A recent study comparing the methylation of 3627 genes in young and old inbred mice found a bidirectional change in methylation with age with approximately 21% and 13% of the genes with increased and decreased promoter methylation, respectively (Maegawa *et al.*, 2010). Changes in expression of all genes tested were correlated with the methylation changes

in their promoters (Maegawa *et al.*, 2010). Also, DNA extracted from the blood of MZ twins sampled at two time points revealed bidirectional changes in promoter regions of three genes between the two time points (Wong *et al.*, 2010). In a cohort of 718 subjects sampled twice over 8 years, methylation of *Alu* elements declined significantly between the measurements, while L1 elements methylation was unchanged (Bollati *et al.*, 2009). The results were, however, not bidirectional between individuals. The LUMA assay targets repeats of most subfamilies and non-repeats. As presented in Paper I, the target site of the restriction endonucleases used for our LUMA assay is under-represented in the SINE family (that include *Alu* elements) and over-represented in the LINE family (that include L1 elements). Therefore, methods targeting only specific subfamilies of repeats might give different results than those obtained with LUMA measurements.

Further studies on the age-associated change in epigenetic marks should focus on establishing if these patterns can be replicated in other tissues, given their availability. Furthermore, a possible link between tissue-specific changes in methylation with time and disease progression in the same organs should be sought. An example of diseases suitable for such studies would be e.g. adult onset diabetes mellitus, immunological diseases such as rheumatoid arthritis and various cancers that demonstrate increased incidence with age.

5.4 Development of a surrogate marker for germline methylation (Paper III)

A novel marker of germline methylation, the mSNP marker, was developed and tested in Paper III. The marker is based on searching in large genome variation databases for mutations that stem from the hypermutability of methylated cytosine, and use the density of these mutations as a surrogate marker for germline methylation. The mSNP marker was validated by demonstrating that

the mutation databases used had mutation spectra representing the hypermutability of methylated cytosine, Furthermore, there was an inverse correlation between the marker and the density of CpG islands, that are normally spared of cytosine methylation. Compared with using the observed/expected CpG ratio as a marker of germline methylation, the mSNP approach is not dependent on the conservation of methylation. In contrast, the observed/expected CpG ratio can be applied in a higher resolution to capture variation in methylation on smaller scales.

Several limitations to the approach must be mentioned. Inevitably not all germline methylation is represented by mSNPs. Also, an estimated 16-17% of the SNPs defined as mSNPs do not represent germline methylation. The marker also cannot differentiate methylation patterns between the male and female germline. The resolution of the information on germline methylation is furthermore dependent on SNP density.

The usage of bioinformatic germline methylation surrogate markers as a hypothesis generating tool can be very beneficial. Future work might increase the resolution of the germline methylation surrogate markers by using the emerging large scale databases of genomic variation, such as from the 1000 genomes project (Durbin *et al.*, 2010).

5.5 A positive correlation between the mSNP marker and regional homologous recombination in the human genome (Paper III)

In paper III, a positive correlation between the mSNP marker and regional rates of homologous recombination was found, both at a genome-wide scale of 125-1000 kb resolution and in a limited region with sufficient data for 25-50 kb resolution. Furthermore, reanalyzing methylation data from sperm revealed that the methylation of sequences within hot spots of recombination was

significantly higher than sequences not within hot spots. The marker was more informative than the observed/expected CpG ratio in the same size range. However, the observed/expected CpG ratio is likely to be more useful in higher resolution.

The results increase the understanding of the causes of discrepancy between regional DNA sequence and recombination rate. Examples of this are variable recombination rate despite identical sequence (Neumann & Jeffreys, 2006) and different locations of hot spots of recombination in human and chimpanzee despite a high sequence homology (Winckler *et al.*, 2005). They support earlier notions that epigenetic mechanisms might be involved in the control of recombination rate (Neumann & Jeffreys, 2006; Sandovici *et al.*, 2006). Furthermore, since the publication of our results, the discovery of the PRDM9 protein and its role in recombination has further illuminated the control mechanism influencing recombination rate. The protein has three domains including a domain with histone methylating properties and a zinc finger domain that binds to the DNA motif previously found to be enriched in hot spots of recombination (Parvanov *et al.*, 2010). Variations of the protein affect recombination rate (Berg *et al.*, 2010; Kong *et al.*, 2010). The discovery of this protein suggests an epigenetic aspect to the control of homologous recombination, similar to our observation of a regional epigenetic mechanism that correlates with recombination. Currently no known link between DNA methylation and PRDM9 activity exists, but testing such link would be of an considerable interest.

Although a cause-and effect relationship cannot be determined based on correlation data, several testable hypotheses can be generated based on the results from Paper III. Two suggested models could explain the results, and they are not mutually exclusive. Perhaps the preferred sites of recombination in the

genome are be marked by methylation. Alternatively, sites that have recently undergone homologous recombination might be secondarily marked by methylation. A second recombination in the same meiosis close to the first recombination might erase the potential benefit of the first event. The molecular mechanism behind cross-over interference is largely unknown. Perhaps DNA methylation might be a part of the molecular mechanism mediating cross-over interference (Berchowitz & Copenhaver, 2010). Additionally, methylation might suppress non-homologous recombination and enhance homologous recombination. This is exemplified by chromosomal instability resulting from mutations in the *DNMT3b* methyltransferase gene (Xu *et al.*, 1999). In paper II, a correlation in the conservation of DNA methylation patterns within families was found. Perhaps similar methylation patterns between the parental genomes minimize the likelihood of harmful non-homologous recombination? In this context, the positive correlation between kinship and fertility of human couples, with a peak in reproductive success at the level of third and fourth cousins is an interesting observation (Helgason *et al.*, 2008). This level of kinship might maximize the positive effects of homologous recombination on fertility against the harmful effects from inbreeding.

Under the assumption that epigenetic marks change with time in the germline following a similar general pattern as observed in somatic cells and described in Paper II, several hypotheses can be generated. For example, recombination rate could change with increasing age, following changes in methylation patterns of the germ line. This is interesting in light of observations of a positive correlation between recombination rate and maternal age (Kong *et al.*, 2004). Furthermore, the recombination rate was an independent predictor of family size and the effect increased with maternal age (Kong *et al.*, 2004). This effect might be mediated by changes in germline methylation patterns similar to

those demonstrated in somatic cells in Paper II.

If the relationship between DNA methylation and somatic (mitotic) recombination is similar to the relationship between germline DNA methylation and meiotic homologous recombination, an interesting mechanism of acquired disease can be suggested based on our observed changes in DNA methylation over time (Paper II). Perhaps acquired changes in somatic DNA methylation might interfere with repair mechanisms based on somatic homologous recombination (Moynahan & Jasin, 2010). In contrast, if DNA methylation follows repair by meiotic homologous recombination, the methylation of repaired sites might affect expression of nearby genes. This interference might participate in the pathogenesis of diseases with increased age-associated incidence, such as cancer. This might be tested by comparing acquired methylation changes in a well defined cancer patient cohort. The cancers initially selected could be those previously associated with disruptions in somatic recombination, such as breast cancer.

Future experiments should try to determine the cause-and-effect of relationship between DNA methylation and homologous recombination. Several experiments can be suggested to test this. Comparing the recombination rate of a methyltransferase deficient mouse mutant with wild type, decreased homologous recombination rate would be suggestive of methylation preceding recombination. Using male sperm, the final product of the male germline, direct measurements of recombination rate of recombination hot spots could be correlated with the methylation of the adjacent sequence. In addition to an experimental verification of the bioinformatics result, the strength of the relationship might suggest the underlying cause-and effect relationship. A significant and strong positive correlation would suggest that methylation precedes recombination. A weak or no relationship might be suggestive of

methylation following recombination, as the methylation of the relatively few recombinant molecules (<1%) would not cause a significant correlation. The methylation of single recombinant molecules could furthermore be compared with non-recombinant molecules (e.g. by single molecule PCR). A higher methylation of recombinant molecules compared to non-recombinant molecules would be suggestive of methylation following recombination, while similar methylation of recombinant and non-recombinant molecules might be suggestive of methylation preceding recombination or no relationship. Ideally, these experiments should also compare the methylation and recombination rate of individuals with differential *PRDM9* alleles to generate hypotheses regarding the relationship between the PRDM9 protein, homologous recombination and DNA methylation.

5.6 Methylation-based defense systems against TDR activity in the human genome (Paper IV)

In Paper IV, the mSNP germline methylation marker was applied to study the subfamily specific relationship between TDRs and germline methylation of adjacent sequences. The host defense system proposed by Yoder *et al*. suggests that DNA methylation is a cornerstone of a global defense system against genome instability caused by TDR insertions (Yoder *et al.*, 1997). Given that the methylation of TDRs affects the methylation landscape of adjacent sequences (Jähner & Jaenisch, 1985), the genome host defense theory predicts a positive correlation between major subgroups of TDRs and germline methylation. This should hold even if the germline is overall hypomethylated.

However, after controlling for confounders, there was a consistent negative correlation between regional proportion of *Alu* elements and the mSNP marker. The same results were found using bisulfite sequencing results from sperm. The results were less clear for L1 elements, the other TDR subfamily

with active elements. There were more L1 elements flanking hypermethylated regions when the relationship was studied with small flank sizes (3-5 kb). However, the relationship diminished with increasing window size and the correlation became negative in 1000 kb window resolution.

These results suggest that DNA methylation is unlikely to be a global defense mechanism against the *Alu* TDR subfamily. This is in line with the recent discoveries of alternative defense systems targeted at *Alu* elements, such as the APOBEC3G family (Hulme *et al.*, 2007). However, a DNA methylation-based TDR defense system might exist for L1 elements based on the results presented in Paper IV. This system might influence methylation of a few kilobases of adjacent sequence, according to our results. Recently a TDR defense system based on piRNA elements linked to proteins from the PIWI family was discovered in the human germline (Zamudio & Bourc'his, 2010). The PIWI-piRNA complexes recruit DNA methyltransferases to L1 and LTR elements and mediate their methylation in the germline via unknown mechanisms (Aravin & Bourc'his, 2008).

The positive correlation between L1 elements and DNA methylation in the flanking 3-5 kb observed in Paper IV supports a PIWI-piRNA defense system targeting L1 elements and mediating their methylation. Further studies of the relationship between DNA methylation and TDR defense should focus on the PIWI-piRNA pathway and its effects on the germline methylation of L1 elements. The mSNP marker could even be used to generate hypotheses regarding this mechanism. With the emerging knowledge of the high insertion frequency of L1 elements in the human genome (Ewing & Kazazian, 2010a; Ewing & Kazazian, 2010b), further understanding of defense mechanisms against these elements is critical.

The results can be used to generate hypotheses as to the nature of the relationship between TDR subfamilies and DNA methylation of the germline, although a causal relationship cannot be established based on correlation data. According to the most parsimonious model explaining all correlation results between the germline methylation marker and the TDR subfamilies including active elements presented in Paper IV, a preferential insertion of the two major repeat subfamilies (*Alu* and L1) into relatively hypomethylated regions of the human genome is suggested. This is supported by similar insertional preferences of both families (Gasior *et al.*, 2007). The fates of the two families are proposed to be different. According to the model, L1 elements might undergo post-insertional methylation, perhaps by a defense system such as the PIWI-piRNA system. Alternatively, there could be a negative selection against L1 elements insertion or a positive selection for *Alu* elements insertion into gene-rich and hypomethylated regions. The negative effects of insertional mutagenesis could be counteracted by positive effects of *Alu* insertions such as exonization (Lev-Maor *et al.*, 2008) or positive effects on gene expression (Eller *et al.*, 2007). As genome-scale sequencing becomes more accurate and cost-effective, testing this model should become feasible. With sufficient sequencing resolution, novel L1 and *Alu* germline insertions could be found by comparing sperm and whole blood sequencing of the same individual. This could determine the adjacent sequence of novel TDR insertions. Additionally, a whole genome bisulfite sequencing of the germ cells could also highlight the methylation landscape adjacent to TDR and novel TDR insertions.

The recombination results from Paper III and the TDR results from Paper IV can also be interpreted together. Non-homologous recombination of *Alu* elements is often harmful to the host genome (Callinan & Batzer, 2006). If DNA methylation marks sites of the germline suitable for recombination, it is

unlikely that *Alu* elements are heavily methylated in the germline, since this might result in increased non-homologous recombination. If DNA methylation comes after a recombination (either homologous or non-homologous), then the low degree of *Alu* methylation might be indicative of an alternative defense mechanism against non-allelic recombination.

5.7 Simulating the metabolic effects of human imprinted genes (Paper V)

The biology of imprinted genes was visited in Paper V. The second motive for the project was to provide insight into the application of systems biology methods to study the genotype-phenotype relationship for large multicellular organisms. Such applications are highly dependent on the quality of the reconstruction and the data used to generate the model. Methods of predicting phenotypes are also being further developed as knowledge of the field is gained.

The monoallelic expression of imprinted genes is stably maintained by differential methylation of alleles based on parental origin. Using methods of systems biology on the recently compiled genome-scale reconstruction of the human metabolic network, both loss of expression and biallelic expression of metabolic imprinted genes were simulated. The simulation results were interpreted based on the imprinting status of the gene and in light of Haig's parental intergenome conflict theory that suggests that maternally imprinted genes repress growth whereas paternally imprinted genes increase growth (Moore & Haig, 1991). Several metabolic subsets were affected for each gene tested. In general, the simulated effects on the metabolic system were lopsided, with no expression resulting in increased flux capacity of many metabolic subsets, but increased expression resulting in decreased flux capacity of a large number of metabolic subsets. Of the genes tested, the simulation of loss of expression of the maternally imprinted gene *ATP10A* resulted in the greatest perturbation of the metabolic system, and the overall effect was likely anabolic.

This was in line with the obese phenotype observed mouse model with *ATP10A* knock-out (Cattanach *et al.*, 1997). It supports that a subgroup of patients with Angelman syndrome with an obese phenotype suffers from deletion of the maternal allele of *ATP10A* (Gillessen-Kaesbach *et al.*, 1999; Meguro *et al.*, 2001).

The simulation results presented in Paper V were, however, not supportive of Haig's intergenome conflict theory. The same number of simulated phenotypes were in accordance to the theory as expected by chance. The simulation results are similar to the results of systematical review of human and mouse phenotypes resulting from abnormal expression of imprinted genes (Tycko & Morison, 2002). The genes reviewed were not the same as tested in Paper V. The review found that only 7 out of 15 genes had phenotypes as predicted by the theory, a similar number as expected by chance (Tycko & Morison, 2002).

The lopsided effects of expression changes of imprinted genes might be indicative of their dosage sensitivity, and warrants further research. A prerequisite for this would be to compare the results of expression changes in non-imprinted genes with imprinted genes. If this is true, than the methods of system biology applied in Paper V might be used to test the gene dosage sensitivity of genes and suggest novel imprinted genes.

In summary, the work presented in this thesis aimed at developing several novel methods to assay human DNA methylation on a genome-wide scale. These assays were then applied to address interesting questions regarding the biology of DNA methylation in the human genome. The observation that changes occur over time both in global and site-specific methylation support an epigenetic model of disease pathogenesis. The changes in methylation might

contribute to the pathogenesis of common human diseases, such as cancer. The bidirectional changes might be explained by differences in environment, genetic factors, or both. An example might be genesis of cancer by loss of methylation and increased expression of oncogenes or increased methylation and decreased expression of tumor suppression genes. The positive relationship between a marker of germline methylation and homologous recombination suggests that methylation might be involved in the control of recombination. Currently, experimental evidence for an epigenetic aspect of homologous recombination involves a histone modification. Methylation might either contribute independently to recombination control or via recruitment of other epigenetic mechanism, such as histone tail modifications. Although DNA methylation is not be important in the defense against all TDRs according to the findings in the thesis, then it is likely to be important for defense against particular TDR families, such as the L1 subfamily. The bioinformatic assays developed can be applied to other data from the human or other genomes to further the understanding of this fundamental epigenetic mechanism and its relationship with other genomic and epigenomic variables.

6 REFERENCES

1. Ahuja N., Li Q., Mohan A.L., Baylin S.B. and Issa J.P. (1998). Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res.*, *58*, 5489-94.
2. Aravin A.A. and Bourc'his D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. *Genes Dev.*, *22*, 970-5.
3. Aravin A.A., van der Heijden G.W., Castañeda J., Vagin V.V., Hannon G.J. and Bortvin A. (2009). Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. *PLoS Genet.*, *5*, e1000764.
4. Arcot S.S., Wang Z., Weber J.L., Deininger P.L. and Batzer M.A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, *29*, 136-44.
5. Armstrong K.M., Bermingham E.N., Bassett S.A., Treloar B.P., Roy N.C. and Barnett M.P. (2010). Global DNA methylation measurement by HPLC using low amounts of DNA. *Biotechnol J*, *Adv. Publ*, 1-6.
6. Barrès R., Osler M.E., Yan J., Rune A., Fritz T., Caidahl K., et al. (2009). Non-CpG methylation of the PGC-1alpha promoter through DNMT3B controls mitochondrial density. *Cell Metab*, *10*, 189-98.
7. Baudat F., Buard J., Grey C., Fledel-Alon A., Ober C., Przeworski M., et al. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, *327*, 836-40.
8. Becker S.A., Feist A.M., Mo M.L., Hannum G., Palsson B.Ø. and Herrgard M.J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*, *2*, 727-38.
9. Belancio V.P., Hedges D.J. and Deininger P. (2008). Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res.*, *18*, 343-58.
10. Bennett-Baker P.E., Wilkowski J. and Burke D.T. (2003). Age-associated activation of epigenetically repressed genes in the mouse. *Genetics*, *165*, 2055-62.
11. Berchowitz L.E. and Copenhaver G.P. (2010). Genetic Interference: Dont Stand So Close to Me. *Curr. Genomics*, *11*, 91-102.
12. Berg I.L., Neumann R., Lam K.G., Sarbajna S., Odenthal-Hesse L., May C.A., et al. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat. Genet.*, *42*, 859-63.
13. Berger S.L. (2007). The complex language of chromatin regulation during

transcription. *Nature*, *447*, 407-12.

14. Berger S.L., Kouzarides T., Shiekhattar R. and Shilatifard A. (2009). An operational definition of epigenetics. *Genes Dev.*, *23*, 781-3.
15. Bestor T.H. and Ingram V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proc. Natl. Acad. Sci. U.S.A.*, *80*, 5559-63.
16. Bibikova M., Lin Z., Zhou L., Chudin E., Garcia E.W., Wu B., et al. (2006). High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.*, *16*, 383-93.
17. Bird A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.*, *16*, 6-21.
18. Bird A. (1997). Does DNA methylation control transposition of selfish elements in the germline?. *Trends Genet.*, *13*, 469-72.
19. Bird A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, *8*, 1499-504.
20. Bjornsson H.T., Ellingsen L.M. and Jonsson J.J. (2006). Transposon-derived repeats in the human genome and 5-methylcytosine-associated mutations in adjacent genes. *Gene*, *370*, 43-50.
21. Bjornsson H.T., Fallin M.D. and Feinberg A.P. (2004). An integrated epigenetic and genetic approach to common human disease. *Trends Genet.*, *20*, 350-8.
22. Bogerd H.P., Wiegand H.L., Hulme A.E., Garcia-Perez J.L., O'Shea K.S., Moran J.V., et al. (2006). Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc. Natl. Acad. Sci. U.S.A.*, *103*, 8780-5.
23. Boissinot S., Entezam A., Young L., Munson P.J. and Furano A.V. (2004). The insertional history of an active family of L1 retrotransposons in humans. *Genome Res.*, *14*, 1221-31.
24. Bollati V., Schwartz J., Wright R., Litonjua A., Tarantini L., Suh H., et al. (2009). Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech. Ageing Dev.*, *130*, 234-9.
25. Bourc'his D. and Bestor T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, *431*, 96-9.
26. Broman K.W., Murray J.C., Sheffield V.C., White R.L. and Weber J.L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, *63*, 861-9.
27. Brown C.J., Ballabio A., Rupert J.L., Lafreniere R.G., Grompe M., Tonlorenzi R., et al. (1991). A gene from the region of the human X

inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, *349*, 38-44.

28. Brown T.C. and Jiricny J. (1987). A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, *50*, 945-50.
29. Busque L., Mio R., Mattioli J., Brais E., Blais N., Lalonde Y., et al. (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood*, *88*, 59-65.
30. Callinan P.A. and Batzer M.A. (2006). Retrotransposable elements and human disease. *Genome Dyn*, *1*, 104-15.
31. Callinan P.A., Wang J., Herke S.W., Garber R.K., Liang P. and Batzer M.A. (2005). Alu retrotransposition-mediated deletion. *J. Mol. Biol.*, *348*, 791-800.
32. Carmell M.A., Girard A., van de Kant H.J.G., Bourc'his D., Bestor T.H., de Rooij D.G., et al. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell*, *12*, 503-14.
33. Cattanach B.M., Barr J.A., Beechey C.V., Martin J., Noebels J. and Jones J. (1997). A candidate model for Angelman syndrome in the mouse. *Mamm. Genome*, *8*, 472-8.
34. Chakravarti A., Buetow K.H., Antonarakis S.E., Waber P.G., Boehm C.D. and Kazazian H.H. (1984). Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.*, *36*, 1239-58.
35. Chakravarti A., Elbein S.C. and Permutt M.A. (1986). Evidence for increased recombination near the human insulin gene: implication for disease association studies. *Proc. Natl. Acad. Sci. U.S.A.*, *83*, 1045-9.
36. Chen X., Guo J., Lei Y., Zou J., Lu X., Bao Y., et al. (2010). Global DNA hypomethylation is associated with NTD-affected pregnancy: A case-control study. *Birth Defects Res. Part A Clin. Mol. Teratol.*, *88*, 575-81.
37. Cheng X. and Blumenthal R.M. (2008). Mammalian DNA methyltransferases: a structural perspective. *Structure*, *16*, 341-50.
38. Coe E. and Kass L.B. (2005). Proof of physical exchange of genes on the chromosomes. *Proc. Natl. Acad. Sci. U.S.A.*, *102*, 6641-6.
39. Cohen H.R., Royce-Tolland M.E., Worringer K.A. and Panning B. (2005). Chromatin modifications on the inactive X chromosome. *Prog. Mol. Subcell. Biol.*, *38*, 91-122.
40. Cooper D.N. and Youssoufian H. (1988). The CpG dinucleotide and human genetic disease. *Hum. Genet.*, *78*, 151-5.
41. Coulondre C., Miller J.H., Farabaugh P.J. and Gilbert W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, *274*, 775-80.

42. Creighton H.B. and McClintock B. (1931). A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proc. Natl. Acad. Sci. U.S.A.*, *17*, 492-7.

43. Cross S.H. and Bird A.P. (1995). CpG islands and genes. *Curr. Opin. Genet. Dev.*, *5*, 309-14.

44. Cui H., Cruz-Correa M., Giardiello F.M., Hutcheon D.F., Kafonek D.R., Brandenburg S., et al. (2003). Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science*, *299*, 1753-5.

45. Cui H., Horon I.L., Ohlsson R., Hamilton S.R. and Feinberg A.P. (1998). Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. *Nat. Med.*, *4*, 1276-80.

46. Deneberg S., Grövdal M., Karimi M., Jansson M., Nahi H., Corbacioglu A., et al. (2010). Gene-specific and global methylation patterns predict outcome in patients with acute myeloid leukemia. *Leukemia*, *24*, 932-41.

47. Dewannieux M., Esnault C. and Heidmann T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.*, *35*, 41-8.

48. Dhar M., Webb L.S., Smith L., Hauser L., Johnson D. and West D.B. (2000). A novel ATPase on mouse chromosome 7 is a candidate gene for increased body fat. *Physiol. Genomics*, *4*, 93-100.

49. Dombroski B.A., Mathias S.L., Nanthakumar E., Scott A.F. and Kazazian H.H.J. (1991). Isolation of an active human transposable element. *Science*, *254*, 1805-8.

50. Duarte N.C., Becker S.A., Jamshidi N., Thiele I., Mo M.L., Vo T.D., et al. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.*, *104*, 1777-82.

51. Durbin R.M., Abecasis G.R., Altshuler D.L., Auton A., Brooks L.D., Durbin R.M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*, 1061-73.

52. Eckhardt F., Lewin J., Cortese R., Rakyan V.K., Attwood J., Burger M., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, *38*, 1378-85.

53. Ehrlich M., Gama-Sosa M.A., Huang L.H., Midgett R.M., Kuo K.C., McCune R.A., et al. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.*, *10*, 2709-21.

54. Eisenberg E. and Levanon E.Y. (2003). Human housekeeping genes are compact. *Trends Genet.*, *19*, 362-5.

55. Eller C.D., Regelson M., Merriman B., Nelson S., Horvath S. and

Marahrens Y. (2007). Repetitive sequence environment distinguishes housekeeping genes. *Gene*, *390*, 153-65.

56. Esteller M. (2008). Epigenetics in cancer. *N. Engl. J. Med.*, *358*, 1148-59.
57. Ewing A.D. and Kazazian H.H. (2010a). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.*, *Adv. Publ.*, 1-17.
58. Ewing A.D. and Kazazian H.H.J. (2010b). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.*, *20*, 1262-70.
59. Feinberg A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, *447*, 433-40.
60. Feinberg A.P. (2004). The epigenetics of cancer etiology. *Semin. Cancer Biol.*, *14*, 427-32.
61. Feinberg A.P. and Vogelstein B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, *301*, 89-92.
62. Feist A.M. and Palsson B.Ø. (2008). The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.*, *26*, 659-67.
63. Foss E.J. and Stahl F.W. (1995). A test of a counting model for chiasma interference. *Genetics*, *139*, 1201-9.
64. Fraga M.F. and Esteller M. (2002). DNA methylation: a profile of methods and applications. *BioTechniques*, *33*, 632-49.
65. Fraga M.F., Ballestar E., Paz M.F., Ropero S., Setien F., Ballestar M.L., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U.S.A.*, *102*, 10604-9.
66. Fuks F., Hurd P.J., Wolf D., Nan X., Bird A.P. and Kouzarides T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J. Biol. Chem.*, *278*, 4035-40.
67. Galagan J.E. and Selker E.U. (2004). RIP: the evolutionary cost of genome defense. *Trends Genet.*, *20*, 417-23.
68. Gardiner-Garden M. and Frommer M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.*, *196*, 261-82.
69. Gasior S.L., Preston G., Hedges D.J., Gilbert N., Moran J.V. and Deininger P.L. (2007). Characterization of pre-insertion loci of de novo L1 insertions. *Gene*, *390*, 190-8.
70. Gillessen-Kaesbach G., Demuth S., Thiele H., Theile U., Lich C. and

Horsthemke B. (1999). A previously unrecognised phenotype characterised by obesity, muscular hypotonia, and ability to speak in patients with Angelman syndrome caused by an imprinting defect. *Eur. J. Hum. Genet.*, *7*, 638-44.

71. Girard A., Sachidanandam R., Hannon G.J. and Carmell M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, *442*, 199-202.
72. Grimwood J., Gordon L.A., Olsen A., Terry A., Schmutz J., Lamerdin J., et al. (2004). The DNA sequence and biology of human chromosome 19. *Nature*, *428*, 529-35.
73. Haig D. and Graham C. (1991). Genomic imprinting and the strange case of the insulin-like growth factor II receptor. *Cell*, *64*, 1045-6.
74. Hajkova P., Erhardt S., Lane N., Haaf T., El-Maarri O., Reik W., et al. (2002). Epigenetic reprogramming in mouse primordial germ cells. *Mech. Dev.*, *117*, 15-23.
75. Handel M.A. and Schimenti J.C. (2010). Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat. Rev. Genet.*, *11*, 124-36.
76. Hao T., Ma H., Zhao X. and Goryanin I. (2010). Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics*, *11*, 393.
77. HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*, 851-61.
78. Hark A.T., Schoenherr C.J., Katz D.J., Ingram R.S., Levorse J.M. and Tilghman S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, *405*, 486-9.
79. Harris T.B., Launer L.J., Eiriksdottir G., Kjartansson O., Jonsson P.V., Sigurdsson G., et al. (2007). Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am. J. Epidemiol.*, *165*, 1076-87.
80. Hata K. and Sakaki Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene*, *189*, 227-34.
81. Hedges D.J. and Deininger P.L. (2007). Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat. Res.*, *616*, 46-59.
82. Heijmans B.T., Tobi E.W., Stein A.D., Putter H., Blauw G.J., Susser E.S., et al. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U.S.A.*, *105*, 17046-9.
83. Helgason A., Pálsson S., Gudbjartsson D.F., Kristjánsson T. and Stefánsson K. (2008). An association between the kinship and fertility of human

couples. *Science*, *319*, 813-6.

84. Hellman A. and Chess A. (2007). Gene body-specific methylation on the active X chromosome. *Science*, *315*, 1141-3.
85. Hendrich B., Hardeland U., Ng H.H., Jiricny J. and Bird A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, *401*, 301-4.
86. Henikoff S. (2008). Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat. Rev. Genet.*, *9*, 15-26.
87. Hirasawa R., Chiba H., Kaneda M., Tajima S., Li E., Jaenisch R., et al. (2008). Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev.*, *22*, 1607-16.
88. Holliday R. (1964). A mechanism for gene conversion in fungi. *Genetics Research*, *5*, 282-304.
89. Holliday R. and Pugh J.E. (1975). DNA modification mechanisms and gene activity during development. *Science*, *187*, 226-32.
90. Homocysteine Studies Collaboration (2002). Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. *JAMA*, *288*, 2015-22.
91. Hsu F., Kent W.J., Clawson H., Kuhn R.M., Diekhans M. and Haussler D. (2006). The UCSC Known Genes. *Bioinformatics*, *22*, 1036-46.
92. Huang C.R.L., Schneider A.M., Lu Y., Niranjan T., Shen P., Robinson M.A., et al. (2010). Mobile interspersed repeats are major structural variants in the human genome. *Cell*, *141*, 1171-82.
93. Huang W., Chang B.H., Gu X., Hewett-Emmett D. and Li W. (1997). Sex differences in mutation rate in higher primates estimated from AMG intron sequences. *J. Mol. Evol.*, *44*, 463-5.
94. Hulme A.E., Bogerd H.P., Cullen B.R. and Moran J.V. (2007). Selective inhibition of Alu retrotransposition by APOBEC3G. *Gene*, *390*, 199-205.
95. Ikegami K., Ohgane J., Tanaka S., Yagi S. and Shiota K. (2009). Interplay between DNA methylation, histone modification and chromatin remodeling in stem cells and during development. *Int. J. Dev. Biol.*, *53*, 203-14.
96. Ingrosso D., Cimmino A., Perna A.F., Masella L., De Santo N.G., De Bonis M.L., et al. (2003). Folate treatment and unbalanced methylation and changes of allelic expression induced by hyperhomocysteinaemia in patients with uraemia. *Lancet*, *361*, 1693-9.
97. Issa J.P., Ottaviano Y.L., Celano P., Hamilton S.R., Davidson N.E. and Baylin S.B. (1994). Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat. Genet.*, *7*, 536-40.

98. Issa J.P., Vertino P.M., Boehm C.D., Newsham I.F. and Baylin S.B. (1996). Switch from monoallelic to biallelic human IGF2 promoter methylation during aging and carcinogenesis. *Proc. Natl. Acad. Sci. U.S.A.*, *93*, 11757-62.

99. Jähner D. and Jaenisch R. (1985). Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity. *Nature*, *315*, 594-7.

100. Javierre B.M., Fernandez A.F., Richter J., Al-Shahrour F., Martin-Subero J.I., Rodriguez-Ubreva J., et al. (2010). Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res.*, *20*, 170-9.

101. Jeffreys A.J. and May C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.*, *36*, 151-6.

102. Jeffreys A.J. and Neumann R. (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum. Mol. Genet.*, *14*, 2277-87.

103. Jeffreys A.J., Kauppi L. and Neumann R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, *29*, 217-22.

104. Jeffreys A.J., Neumann R., Panayi M., Myers S. and Donnelly P. (2005). Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.*, *37*, 601-6.

105. Jensen-Seaman M.I., Furey T.S., Payseur B.A., Lu Y., Roskin K.M., Chen C., et al. (2004). Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.*, *14*, 528-38.

106. Jiang Y., Bressler J. and Beaudet A.L. (2004). Epigenetics and human disease. *Annu Rev Genomics Hum Genet*, *5*, 479-510.

107. Jones P.L., Veenstra G.J., Wade P.A., Vermaak D., Kass S.U., Landsberger N., et al. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.*, *19*, 187-91.

108. Josse J., Kaiser A.D. and Kornberg A. (1961). Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.*, *236*, 864-75.

109. Juriloff D.M. and Harris M.J. (2000). Mouse models for neural tube closure defects. *Hum. Mol. Genet.*, *9*, 993-1000.

110. Jurka J., Kapitonov V.V., Pavlicek A., Klonowski P., Kohany O. and Walichiewicz J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, *110*, 462-7.

111. Kaminsky Z.A., Tang T., Wang S., Ptak C., Oh G.H.T., Wong A.H.C., et al. (2009). DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.*, *41*, 240-5.

112. Kaneda M., Okano M., Hata K., Sado T., Tsujimoto N., Li E., et al. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, *429*, 900-3.

113. Kaplan N., Moore I.K., Fondufe-Mittendorf Y., Gossett A.J., Tillo D., Field Y., et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, *458*, 362-6.

114. Karimi M., Johansson S. and Ekström T.J. (2006a). Using LUMA: a Luminometric-based assay for global DNA-methylation.. *Epigenetics : official journal of the DNA Methylation Society*, *1*, 45-8.

115. Karimi M., Johansson S., Stach D., Corcoran M., Grandér D., Schalling M., et al. (2006b). LUMA (LUminometric Methylation Assay)--a high throughput method to the analysis of genomic DNA methylation. *Exp. Cell Res.*, *312*, 1989-95.

116. Kazazian H.H.J. (1999). An estimated frequency of endogenous insertional mutations in humans. *Nat. Genet.*, *22*, 130.

117. Keeney S. (2001). Mechanism and control of meiotic recombination initiation. *Curr. Top. Dev. Biol.*, *52*, 1-53.

118. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., et al. (2002). The human genome browser at UCSC. *Genome Res.*, *12*, 996-1006.

119. Kim T., Chung Y., Rhyu M. and Jung M.H. (2007). Germline methylation patterns inferred from local nucleotide frequency of repetitive sequences in the human genome. *Mamm. Genome*, *18*, 277-85.

120. Kimura M. (1991). Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci. U.S.A.*, *88*, 5969-73.

121. King J.S. and Mortimer R.K. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics*, *126*, 1127-38.

122. Kitano H. (2002). Computational systems biology. *Nature*, *420*, 206-10.

123. Kleckner N., Zickler D., Jones G.H., Dekker J., Padmore R., Henle J., et al. (2004). A mechanical basis for chromosome function. *Proc. Natl. Acad. Sci. U.S.A.*, *101*, 12592-7.

124. Knip M., Veijola R., Virtanen S.M., Hyöty H., Vaarala O. and Akerblom H.K. (2005). Environmental triggers and determinants of type 1 diabetes.

Diabetes, *54 Suppl 2*, S125-36.

125. Kong A., Barnard J., Gudbjartsson D.F., Thorleifsson G., Jonsdottir G., Sigurdardottir S., et al. (2004). Recombination rate and reproductive success in humans. *Nat. Genet.*, *36*, 1203-6.

126. Kong A., Gudbjartsson D.F., Sainz J., Jonsdottir G.M., Gudjonsson S.A., Richardsson B., et al. (2002). A high-resolution recombination map of the human genome. *Nat. Genet.*, *31*, 241-7.

127. Kong A., Steinthorsdottir V., Masson G., Thorleifsson G., Sulem P., Besenbacher S., et al. (2009). Parental origin of sequence variants associated with complex diseases. *Nature*, *462*, 868-74.

128. Kong A., Thorleifsson G., Gudbjartsson D.F., Masson G., Sigurdsson A., Jonasdottir A., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, *467*, 1099-103.

129. Krogh B.O. and Symington L.S. (2004). Recombination proteins in yeast. *Annu. Rev. Genet.*, *38*, 233-71.

130. Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860-921.

131. Lee D., Park J., Kay K.A., Christakis N.A., Oltvai Z.N. and Barabási A. (2008). The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U.S.A.*, *105*, 9880-5.

132. Lercher M.J. and Hurst L.D. (2003). Imprinted chromosomal regions of the human genome have unusually high recombination rates. *Genetics*, *165*, 1629-32.

133. Lev-Maor G., Ram O., Kim E., Sela N., Goren A., Levanon E.Y., et al. (2008). Intronic Alus influence alternative splicing. *PLoS Genet.*, *4*, e1000204.

134. Lev-Maor G., Sorek R., Shomron N. and Ast G. (2003). The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science*, *300*, 1288-91.

135. Li E., Bestor T.H. and Jaenisch R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, *69*, 915-26.

136. Lindahl T. (1974). An N-glycosidase from Escherichia coli that releases free uracil from DNA containing deaminated cytosine residues. *Proc. Natl. Acad. Sci. U.S.A.*, *71*, 3649-53.

137. Lister R., Pelizzola M., Dowen R.H., Hawkins R.D., Hon G., Tonti-Filippini J., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*, 315-22.

138. Liu W.M., Chu W.M., Choudary P.V. and Schmid C.W. (1995). Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.*, *23*, 1758-65.

139. Lobo I. and Shaw K. (2008). Thomas Hunt Morgan, genetic recombination, and gene mapping. *Nature Education*, *1*, e1-e3.

140. Luedi P.P., Dietrich F.S., Weidman J.R., Bosko J.M., Jirtle R.L. and Hartemink A.J. (2007). Computational and experimental identification of novel human imprinted genes. *Genome Res.*, *17*, 1723-30.

141. Lunyak V.V., Prefontaine G.G., Núñez E., Cramer T., Ju B., Ohgi K.A., et al. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*, *317*, 248-51.

142. Ma H., Sorokin A., Mazein A., Selkov A., Selkov E., Demin O., et al. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.*, *3*, 135.

143. Maegawa S., Hinkal G., Kim H.S., Shen L., Zhang L., Zhang J., et al. (2010). Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res.*, *20*, 332-40.

144. Mahadevan R. and Schilling C.H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, *5*, 264-76.

145. Maher E.R. and Reik W. (2000). Beckwith-Wiedemann syndrome: imprinting in clusters revisited. *J. Clin. Invest.*, *105*, 247-52.

146. Manolio T.A. (2010). Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, *363*, 166-76.

147. Mariner P.D., Walters R.D., Espinoza C.A., Drullinger L.F., Wagner S.D., Kugel J.F., et al. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell*, *29*, 499-509.

148. Matzke M., Kanno T., Huettel B., Daxinger L. and Matzke A.J.M. (2007). Targets of RNA-directed DNA methylation. *Curr. Opin. Plant Biol.*, *10*, 512-9.

149. McClelland M., Nelson M. and Raschke E. (1994). Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. *Nucleic Acids Res.*, *22*, 3640-59.

150. McMahill M.S., Sham C.W. and Bishop D.K. (2007). Synthesis-dependent strand annealing in meiosis. *PLoS Biol.*, *5*, e299.

151. McMilin K.D., Stahl M.M. and Stahl F.W. (1974). Rec-mediated recombinational hot spot activity in bacteriophage lambda. I. Hot spot activity associated with spi-deletions and bio substitutions. *Genetics*, *77*,

409-23.

152. McVean G.A.T., Myers S.R., Hunt S., Deloukas P., Bentley D.R. and Donnelly P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, *304*, 581-4.

153. Meguro M., Kashiwagi A., Mitsuya K., Nakao M., Kondo I., Saitoh S., et al. (2001). A novel maternally expressed gene, ATP10C, encodes a putative aminophospholipid translocase associated with Angelman syndrome. *Nat. Genet.*, *28*, 19-20.

154. Millar C.B., Guy J., Sansom O.J., Selfridge J., MacDougall E., Hendrich B., et al. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science*, *297*, 403-5.

155. Moore T. and Haig D. (1991). Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet.*, *7*, 45-9.

156. Morgan H.D., Sutherland H.G., Martin D.I. and Whitelaw E. (1999). Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.*, *23*, 314-8.

157. Moynahan M.E. and Jasin M. (2010). Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat. Rev. Mol. Cell Biol.*, *11*, 196-207.

158. Musco G. and Peterson P. (2008). PHD finger of autoimmune regulator: an epigenetic link between the histone modifications and tissue-specific antigen expression in thymus. *Epigenetics*, *3*, 310-4.

159. Myers S., Bottolo L., Freeman C., McVean G. and Donnelly P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*, 321-4.

160. Myers S., Bowden R., Tumian A., Bontrop R.E., Freeman C., MacFie T.S., et al. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, *327*, 876-9.

161. Myers S., Freeman C., Auton A., Donnelly P. and McVean G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.*, *40*, 1124-9.

162. Myers S., Spencer C.C.A., Auton A., Bottolo L., Freeman C., Donnelly P., et al. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.*, *34*, 526-30.

163. Neumann R. and Jeffreys A.J. (2006). Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. *Hum. Mol. Genet.*, *15*, 1401-11.

164. Ng H.H. and Bird A. (1999). DNA methylation and chromatin

modification. *Curr. Opin. Genet. Dev.*, *9*, 158-63.

165. Nichol K. and Pearson C.E. (2002). CpG methylation modifies the genetic stability of cloned repeat sequences. *Genome Res.*, *12*, 1246-56.

166. Nishant K.T. and Rao M.R.S. (2006). Molecular features of meiotic recombination hot spots. *Bioessays*, *28*, 45-56.

167. Oberhardt M.A., Palsson B.Ø. and Papin J.A. (2009). Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.*, *5*, 320.

168. Obican S.G., Finnell R.H., Mills J.L., Shaw G.M. and Scialli A.R. (2010). Folic acid in early pregnancy: a public health success story. *FASEB J.*, *24*, 4167-74.

169. Okano M., Bell D.W., Haber D.A. and Li E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*, 247-57.

170. Orth J.D., Thiele I. and Palsson B.Ø. (2010). What is flux balance analysis?. *Nat. Biotechnol.*, *28*, 245-8.

171. Paigen K. and Petkov P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nat. Rev. Genet.*, *11*, 221-33.

172. Pàldi A., Gyapay G. and Jami J. (1995). Imprinted chromosomal regions of the human genome display sex-specific meiotic recombination frequencies. *Curr. Biol.*, *5*, 1030-5.

173. Palsson B. (2009a). Metabolic systems biology. *FEBS Lett.*, *583*, 3900-4.

174. Palsson B.O. (2009b). Metabolic systems biology: a constraint-based approach. *Encyclopedia of Complexity and Systems Science*, *13*, 5535-5552.

175. Palsson B.O. (2006). Systems Biology: Properties of Reconstructed Networks. Cambridge University Press.

176. Parvanov E.D., Ng S.H.S., Petkov P.M. and Paigen K. (2009). Trans-regulation of mouse meiotic recombination hotspots by Rcr1. *PLoS Biol.*, *7*, e36.

177. Parvanov E.D., Petkov P.M. and Paigen K. (2010). Prdm9 controls activation of mammalian recombination hotspots. *Science*, *327*, 835.

178. Ponicsan S.L., Kugel J.F. and Goodrich J.A. (2010). Genomic gems: SINE RNAs regulate mRNA production. *Curr. Opin. Genet. Dev.*, *20*, 149-55.

179. Price N.D., Reed J.L. and Palsson B.Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, *2*, 886-97.

180. Racchi O., Mangerini R., Rapezzi D., Rolfo M., Gaetani G.F. and Ferraris A.M. (1998). X chromosome inactivation patterns in normal females. *Blood*

Cells Mol. Dis., *24*, 439-47.

181. Rakyan V.K., Down T.A., Maslau S., Andrew T., Yang T., Beyan H., et al. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.*, *20*, 434-9.
182. Ramsahoye B.H., Biniszkiewicz D., Lyko F., Clark V., Bird A.P. and Jaenisch R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U.S.A.*, *97*, 5237-42.
183. Razin A. and Riggs A.D. (1980). DNA methylation and gene function. *Science*, *210*, 604-10.
184. Reik W. and Walter J. (2001). Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, *2*, 21-32.
185. Rhee I., Bachman K.E., Park B.H., Jair K., Yen R.C., Schuebel K.E., et al. (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, *416*, 552-6.
186. Riggs A.D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.*, *14*, 9-25.
187. Roberts R.J., Vincze T., Posfai J. and Macelis D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, *38*, D234-6.
188. Robinson W.P. and Lalande M. (1995). Sex-specific meiotic recombination in the Prader--Willi/Angelman syndrome imprinted region. *Hum. Mol. Genet.*, *4*, 801-6.
189. Rocha M.S., Castro R., Rivera I., Kok R.M., Smulders Y.M., Jakobs C., et al. (2010). Global DNA methylation: comparison of enzymatic- and non-enzymatic-based methods. *Clin. Chem. Lab. Med.*, *48*, 1793-8.
190. Romero P., Wagg J., Green M.L., Kaiser D., Krummenacker M. and Karp P.D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, *6*, R2.
191. Sandovici I., Kassovska-Bratinova S., Vaughan J.E., Stewart R., Leppert M. and Sapienza C. (2006). Human imprinted chromosomal regions are historical hot-spots of recombination. *PLoS Genet.*, *2*, e101.
192. Sandovici I., Leppert M., Hawk P.R., Suarez A., Linares Y. and Sapienza C. (2003). Familial aggregation of abnormal methylation of parental alleles at the IGF2/H19 and IGF2R differentially methylated regions. *Hum. Mol. Genet.*, *12*, 1569-78.
193. Sandovici I., Naumova A.K., Leppert M., Linares Y. and Sapienza C. (2004). A longitudinal study of X-inactivation ratio in human females. *Hum.*

Genet., *115*, 387-92.

194. Santos F., Hendrich B., Reik W. and Dean W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev. Biol.*, *241*, 172-82.

195. Sasaki M., Lange J. and Keeney S. (2010). Genome destabilization by homologous recombination in the germ line. *Nat. Rev. Mol. Cell Biol.*, *11*, 182-95.

196. Scriver C.R. (2007). The PAH gene, phenylketonuria, and a paradigm shift. *Hum. Mutat.*, *28*, 831-45.

197. Segal E. and Widom J. (2009). What controls nucleosome positions?. *Trends Genet.*, *25*, 335-43.

198. Seshadri S., Beiser A., Selhub J., Jacques P.F., Rosenberg I.H., D'Agostino R.B., et al. (2002). Plasma homocysteine as a risk factor for dementia and Alzheimer's disease. *N. Engl. J. Med.*, *346*, 476-83.

199. Sheikh K., Förster J. and Nielsen L.K. (2005). Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol. Prog.*, *21*, 112-21.

200. Shlomi T., Berkman O. and Ruppin E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U.S.A.*, *102*, 7695-700.

201. Shlomi T., Cabili M.N., Herrgård M.J., Palsson B.Ø. and Ruppin E. (2008). Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, *26*, 1003-10.

202. Sigurdsson M.I., Jamshidi N., Steingrimsson E., Thiele I. and Palsson B.O. (2010). A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst Biol*, *4*, 140.

203. Smith A.V., Thomas D.J., Munro H.M. and Abecasis G.R. (2005). Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.*, *15*, 1519-34.

204. Song F., Smith J.F., Kimura M.T., Morrow A.D., Matsuyama T., Nagase H., et al. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, *102*, 3336-41.

205. Sorek R., Ast G. and Graur D. (2002). Alu-containing exons are alternatively spliced. *Genome Res.*, *12*, 1060-7.

206. Sørensen A.L. and Collas P. (2009). Immunoprecipitation of methylated DNA. *Methods Mol. Biol.*, *567*, 249-62.

207. Soutoglou E., Katrakili N. and Talianidis I. (2000). Acetylation regulates

transcription factor activity at multiple levels. *Mol. Cell*, *5*, 745-51.

208. Stein R., Razin A. and Cedar H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc. Natl. Acad. Sci. U.S.A.*, *79*, 3418-22.

209. Stern C. (1931). Zytologisch-genetische untersuchungen alsbeweise fur die Morgansche theorie des faktoraustauschs. *Biol. Zentbl*, *51*, 547-87.

210. Szostak J.W., Orr-Weaver T.L., Rothstein R.J. and Stahl F.W. (1983). The double-strand-break repair model for recombination. *Cell*, *33*, 25-35.

211. Tam O.H., Aravin A.A., Stein P., Girard A., Murchison E.P., Cheloufi S., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, *453*, 534-8.

212. Taylor J., Tyekucheva S., Zody M., Chiaromonte F. and Makova K.D. (2006). Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol. Biol. Evol.*, *23*, 565-73.

213. The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*, 636-40.

214. Thiele I. and Palsson B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, *5*, 93-121.

215. Thomas D.J., Trumbower H., Kern A.D., Rhead B.L., Kuhn R.M., Haussler D., et al. (2007). Variation resources at UC Santa Cruz. *Nucleic Acids Res.*, *35*, D716-20.

216. Tycko B. and Morison I.M. (2002). Physiological functions of imprinted genes. *J. Cell. Physiol.*, *192*, 245-58.

217. Venolia L., Gartler S.M., Wassman E.R., Yen P., Mohandas T. and Shapiro L.J. (1982). Transformation with DNA from 5-azacytidine-reactivated X chromosomes. *Proc. Natl. Acad. Sci. U.S.A.*, *79*, 2352-4.

218. Verger A., Perdomo J. and Crossley M. (2003). Modification with SUMO. A role in transcriptional regulation. *EMBO Rep.*, *4*, 137-42.

219. Verona R.I., Mann M.R.W. and Bartolomei M.S. (2003). Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annu. Rev. Cell Dev. Biol.*, *19*, 237-59.

220. Voight B.F., Kudaravalli S., Wen X. and Pritchard J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.*, *4*, e72.

221. Waddington C. (1942). Canalization of Development and the Inheritance of Acquired Characters. *Nature*, *3811*, 563-566.

222. Wall J.D. and Pritchard J.K. (2003). Haplotype blocks and linkage

disequilibrium in the human genome. *Nat. Rev. Genet.*, *4*, 587-97.

223. Waterland R.A. and Jirtle R.L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol. Cell. Biol.*, *23*, 5293-300.

224. Weaver J.R., Susiarjo M. and Bartolomei M.S. (2009). Imprinting and epigenetic changes in the early embryo. *Mamm. Genome*, *20*, 532-43.

225. Webb A.J., Berg I.L. and Jeffreys A. (2008). Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc. Natl. Acad. Sci. U.S.A.*, *105*, 10471-6.

226. Weber M., Hellmann I., Stadler M.B., Ramos L., Pääbo S., Rebhan M., et al. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, *39*, 457-66.

227. Winckler W., Myers S.R., Richter D.J., Onofrio R.C., McDonald G.J., Bontrop R.E., et al. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, *308*, 107-11.

228. Wong C.C.Y., Caspi A., Williams B., Craig I.W., Houts R., Ambler A., et al. (2010). A longitudinal study of epigenetic variation in twins. *Epigenetics*, *5*, 516-526.

229. Xie H., Wang M., Bischof J., Bonaldo M.D.F. and Soares M.B. (2009). SNP-based prediction of the human germ cell methylation landscape. *Genomics*, *93*, 434-40.

230. Xing J., Hedges D.J., Han K., Wang H., Cordaux R. and Batzer M.A. (2004). Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J. Mol. Biol.*, *344*, 675-82.

231. Xu G.L., Bestor T.H., Bourc'his D., Hsieh C.L., Tommerup N., Bugge M., et al. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature*, *402*, 187-91.

232. Yoder J.A., Walsh C.P. and Bestor T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*, *13*, 335-40.

233. Zamudio N. and Bourc'his D. (2010). Transposable elements in the mammalian germline: a comfortable niche or a deadly trap?. *Heredity*, *105*, 92-104.

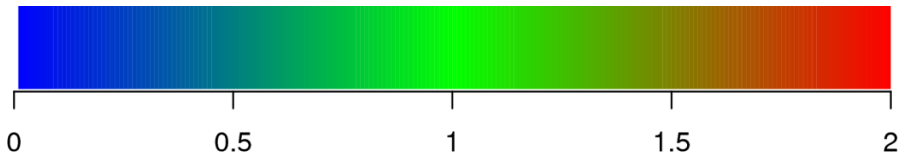
234. Zeisel S.H. (2009). Importance of methyl donors during reproduction. *Am. J. Clin. Nutr.*, *89*, 673S-7S.

235. Zhao Z. and Zhang F. (2006). Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene*, *366*, 316-24.

236. Zilberman D. and Henikoff S. (2007). Genome-wide analysis of DNA methylation patterns. *Development*, *134*, 3959-65.

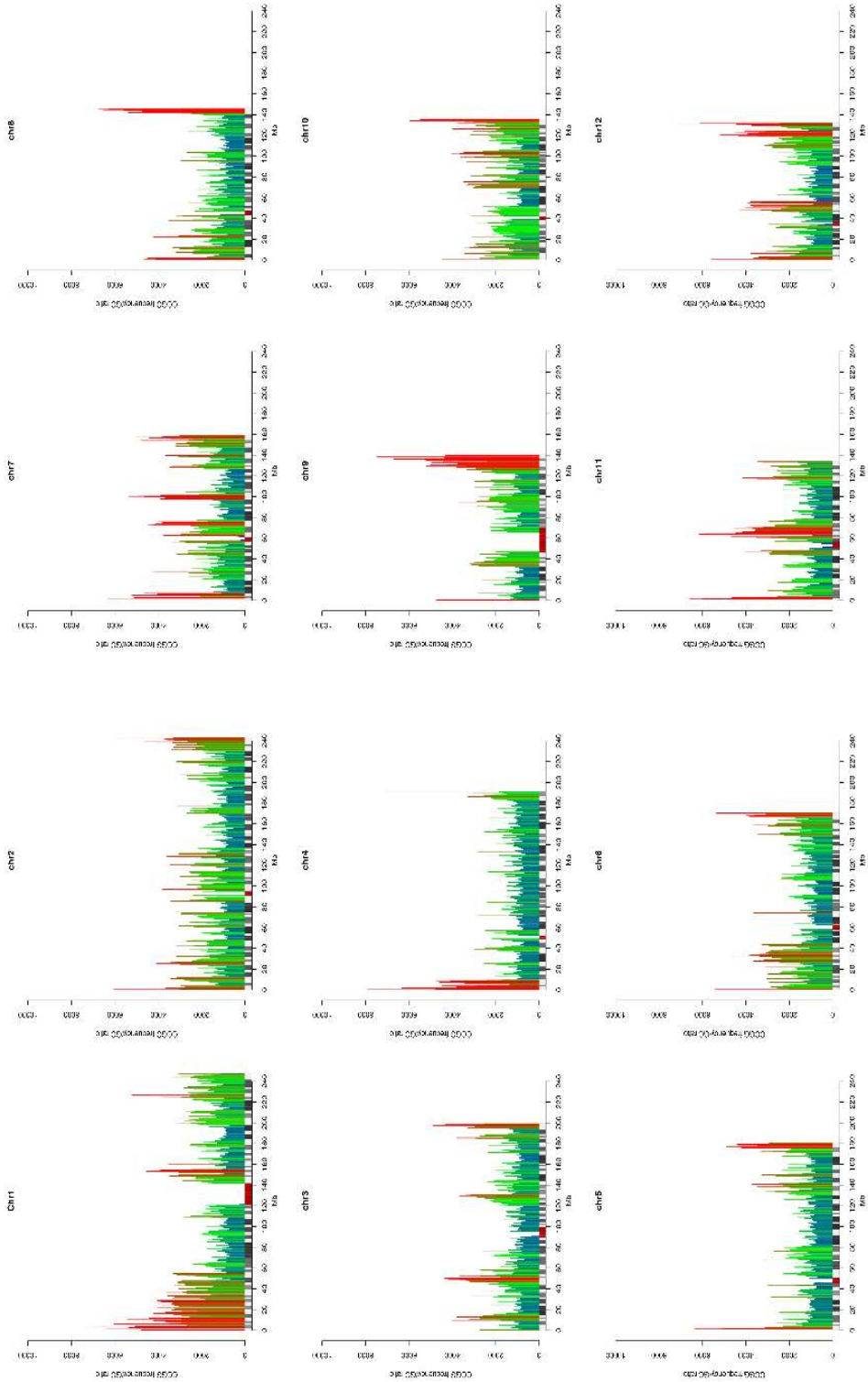
7 APPENDIX I

Relative frequency distribution for CCGG, GCGC and CCWGG sequences measured in 500 kb windows for the human genome. Shown is the frequency in each window divided by the GC ratio of the window. Colors represent the over-representation or under-representation compared to the genomic average frequency normalized by the genomic average GC ratio.

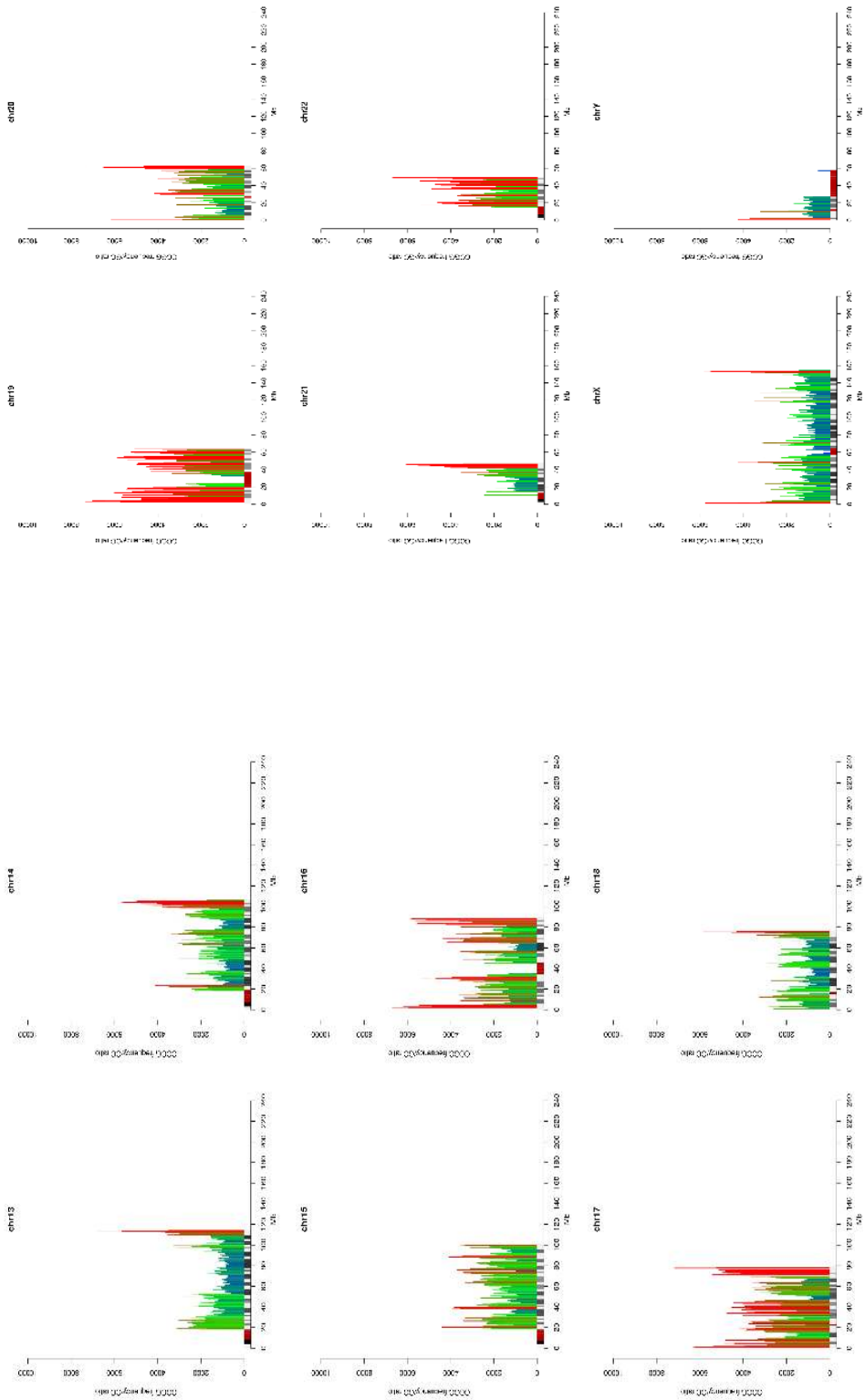


Color legend (Genomic average=1)

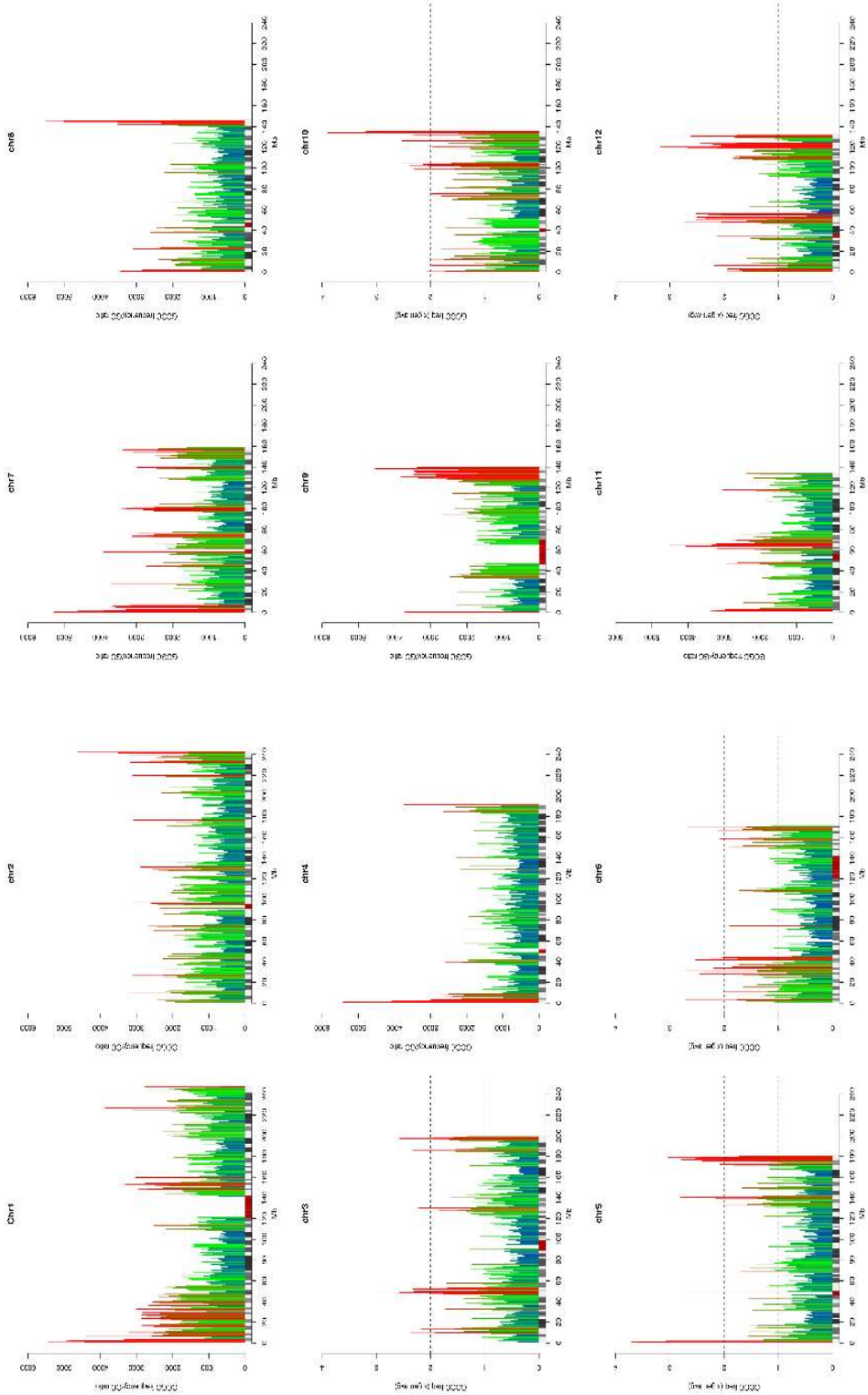
CCGG


chr8
chr10
chr12
chr7
chr9
chr11
chr2
chr4
chr6
Chr1
chr3
chr5
Mb

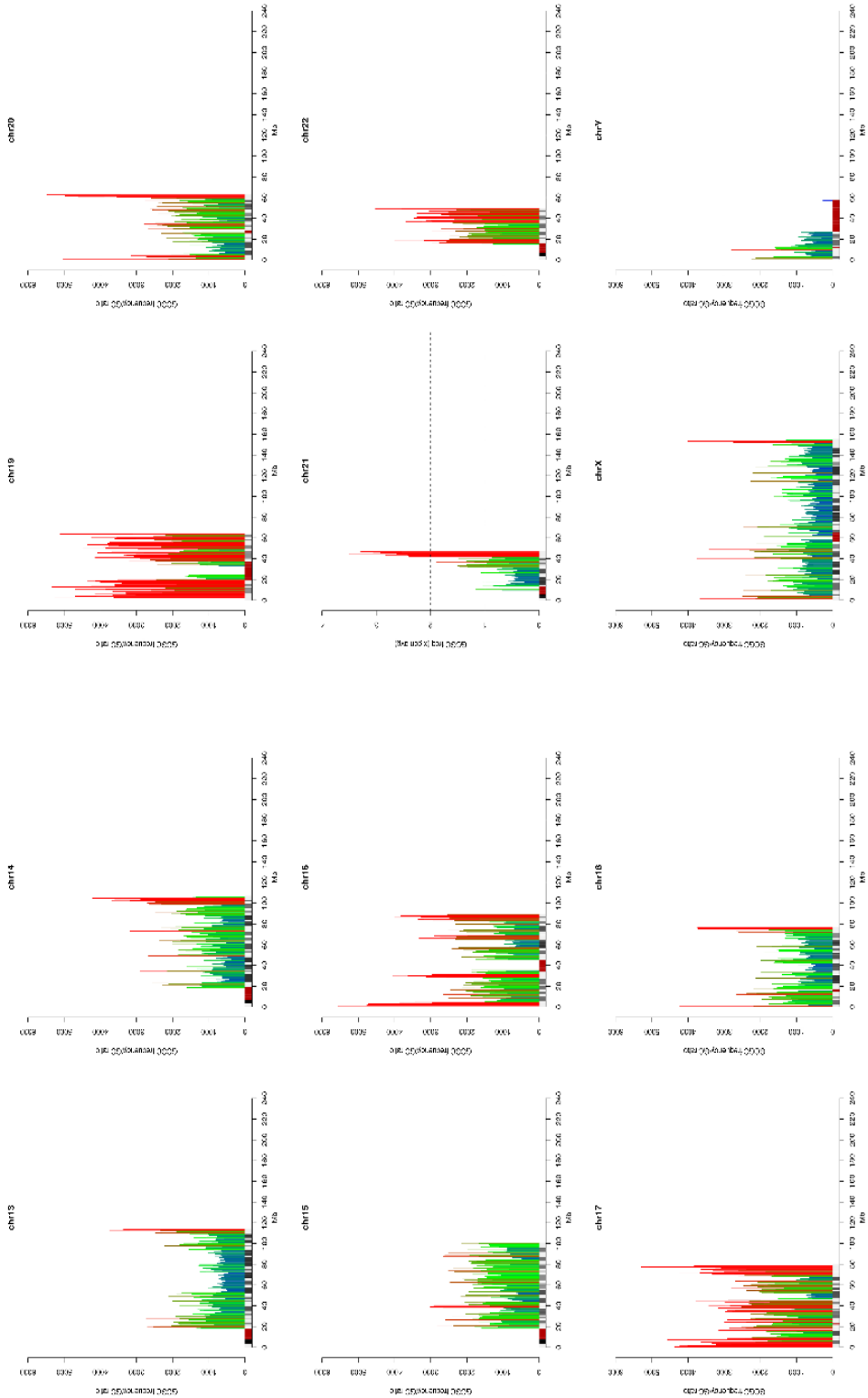
CCGG (cont)


chr20
chr22
chrY
chr19
chr21
chrX
chr14
chr16
chr18
chr13
chr15
chr17

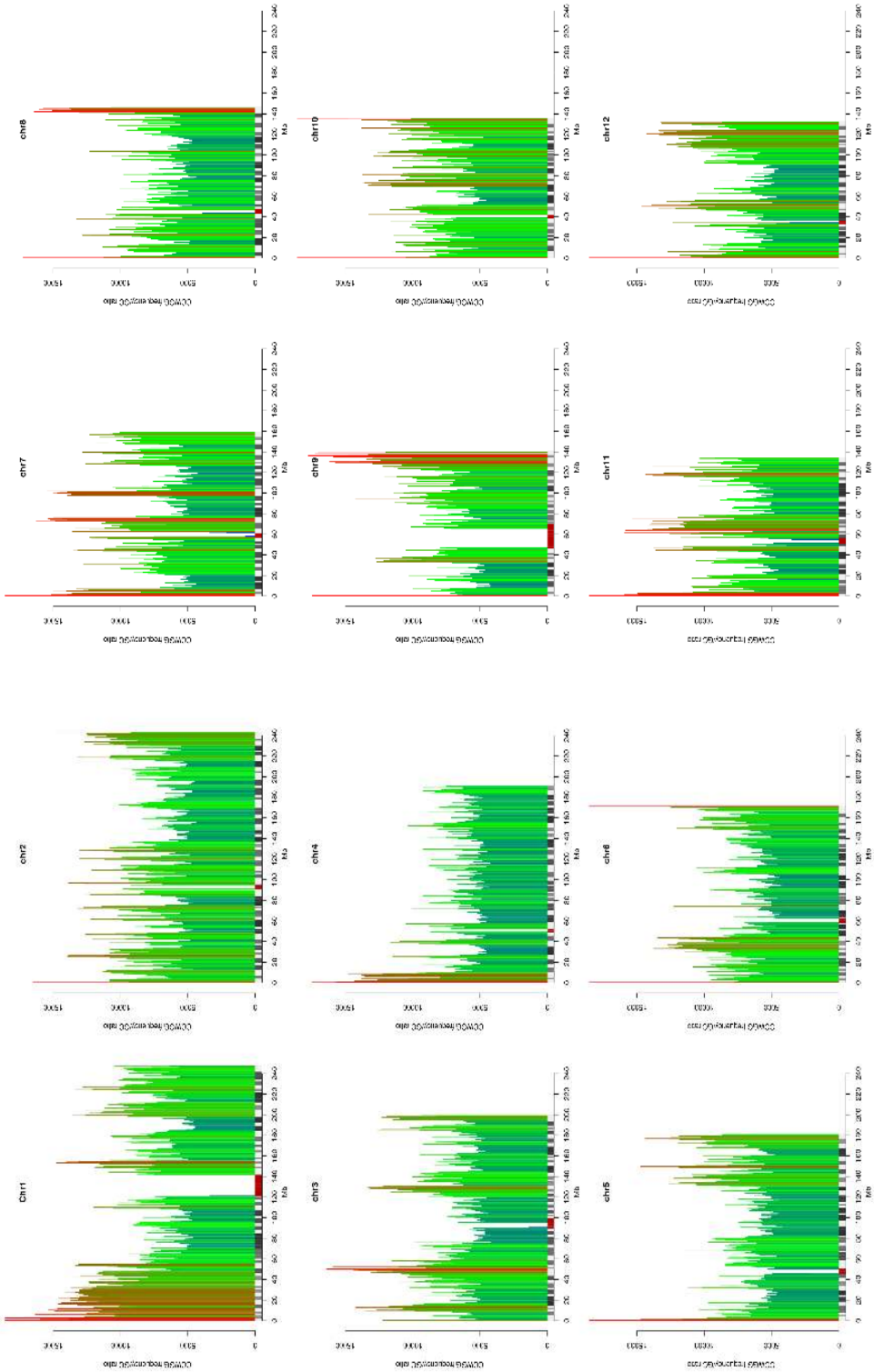
GCGC



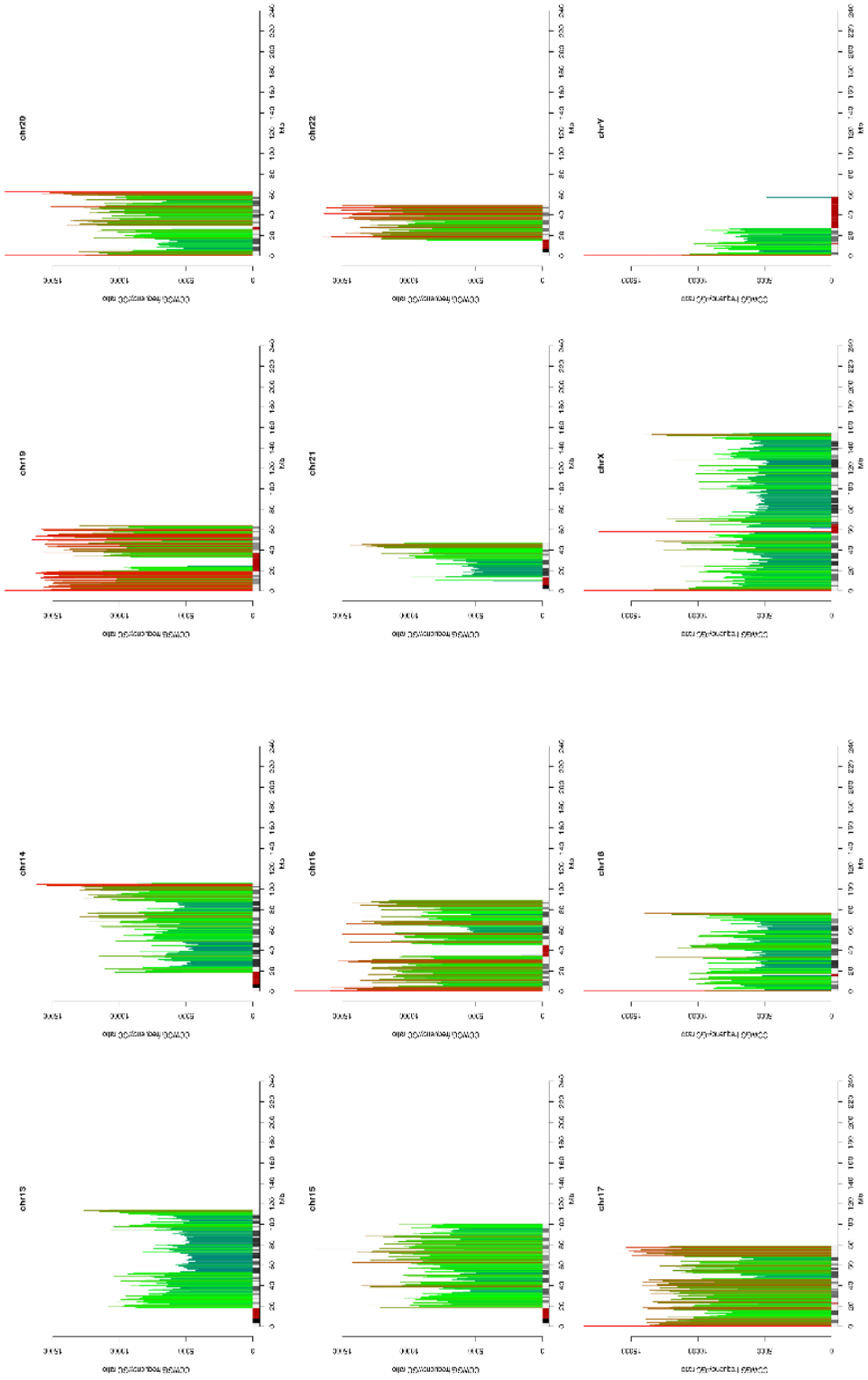
GCGC (cont)


chr20
chr22
chrY
chr19
chr21
chrX
chr14
chr16
chr18
chr13
chr15
chr17

CCWGG


chr8
chr10
chr12
chr7
chr9
chr11
chr2
chr4
chr6
Chr1
chr3
chr5
Mb
CCWGG frequency/GC ratio

CCWGG (cont)


chr20
chr22
chrY
chr19
chr21
chrX
chr14
chr16
chr18
chr13
chr15
chr17