Quantitative studies of DNA methylation and gene expression in neuropsychiatric traits

Kristel R. van Eijk

The studies described in this thesis were performed at the Brain Center Rudolf Magnus, Department of Psychiatry, and at the Department of Medical Genetics, University Medical Center Utrecht, The Netherlands

© Kristel R. van Eijk, Utrecht 2014

Cover design: Guus van Eijk - Grafisch ontwerper
Printed & Lay Out by: Proefschriftmaken.nl || Uitgeverij BOXPress
Published by: Uitgeverij BOXPress, 's-Hertogenbosch

Quantitative studies of DNA methylation and gene expression in neuropsychiatric traits

Kwantitatieve studies naar DNA methylatie en genexpressie in neuropsychiatrische aandoeningen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op woensdag 25 juni 2014 des middags te 12.45 uur

door

Kristel Rianne van Eijk

geboren op 8 juni 1986 te Culemborg

Promotor: Prof. dr. ir. R.A. Ophoff

Co-promotor: Dr. M.P.M. Boks

Content

Chapter 1: Introduction and Outline of this thesis 7

Chapter 2: Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects 21

Chapter 3: Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes 49

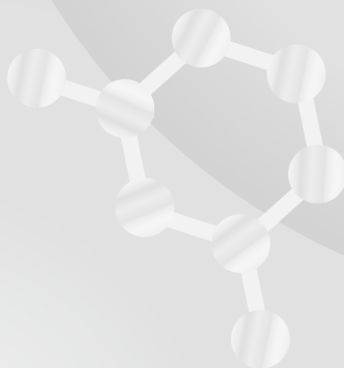
Chapter 4: Identification of schizophrenia-associated loci by combining DNA methylation and gene expression data from whole blood 65

Chapter 5: Characterization of genome-methylome interactions in 22 nuclear pedigrees 79

Chapter 6: Interplay between miRNA-137 and DNA methylation in human brain highlights neurotransmitter-related pathways 101

Chapter 7: Summary and Discussion 123

Nederlandse samenvatting 149
Dankwoord / acknowledgements 159
List of publications 167
Curriculum Vitae 179



Chapter 1

Introduction and Outline of this thesis

Our genes and environment together make us who we are. However, monozygotic twins, genetically identical and raised together, can develop different diseases, suggesting that there might be other factors that can influence behavior and health.

Epigenetics

After many years of genetic research, it has become clear that there are multiple layers of genomic information and that our genes (together with environmental influences) do not determine everything. In the past decade, the field of epigenetics has seen enormous growth and progress. Epigenetics means "on top of" or "surrounding" genetics and it defines changes in gene function that can be transmitted through cell division but are not due to alterations in the underlying DNA sequence[10], see also **box 1**. Rakyan *et al*. described the epigenome as "a complete collection of epigenetic marks, such as DNA methylation and histone modifications, and other molecules that can transmit epigenetic information, such as non-coding RNAs, that exist in a cell at any given point in time"[11]. The epigenome of a cell is highly dynamic and genetic and environmental factors can influence the epigenomic state. Epigenetics is crucial in the development of different tissues and organs. The DNA sequence

Box 1: Example of epigenetics

During the Dutch famine (known as the "Hongerwinter") of 1944/45 during World War II, food supply was reduced due to food transport blockades by the Germans. Additionally, the cold winter prohibited food delivery by boat because canals froze over. Food rations were dramatically reduced and peoples' diets lacked sufficient vitamins and proteins. Pregnant women were especially affected. Individuals conceived during this famine and born after the war were prenatally exposed to this condition, which led to enhance disease susceptibility for a subset of diseases. Exposure to famine during early gestation has been shown to be associated with for example, schizophrenia[3; 4] and glucose intolerance, the latter leading to adverse outcomes such as stress sensitivity and obesity[6].

Besides a mother's diet, also other prenatal conditions have been associated with disease susceptibility of offspring and future generations[7]. For example, maternal stress during gestation has been associated with neurodevelopmental[7; 8] and psychiatric disorders. One study found an increase in the glucocorticoid receptor promoter of these children, which can lead to psychiatric disorders later in life[12].

These examples indicate a heritable component, which can be transferred to the next generation[13; 14].

in all cells within an individual is identical, but genes are switched on or off to control various cellular mechanisms, such as differentiating cells into different tissue types[14]. Interestingly, recent discoveries have revealed that such epigenetic changes are reversible and thus add flexibility to the genome[15; 16]. This allows genomic adaptation to a changing environment, and possibly leads to phenotypic variation as well.

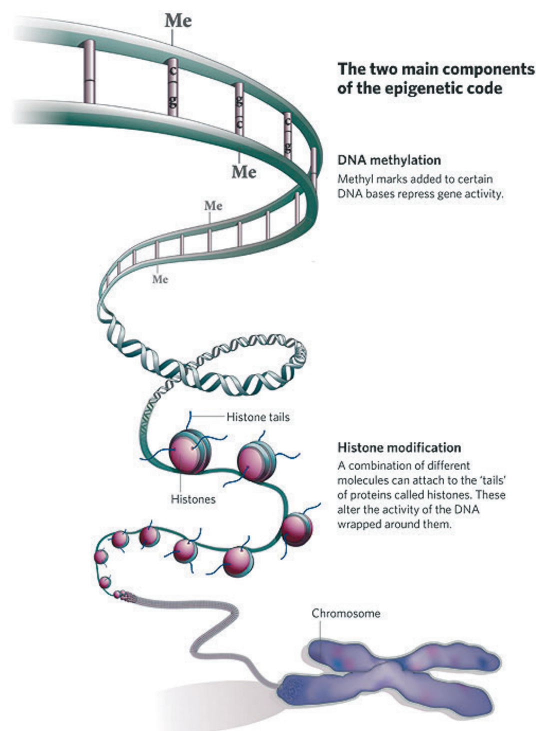


Figure 1: DNA methylation and histone modification (used with permission from Qiu 2006[9])

As described in the definition of the epigenome, there are several types of epigenetic modifications[14; 17]. The main factors are histone modification and DNA methylation (**Figure 1**). Histones are proteins that enable the DNA strand to fold into compact structures[18]. The binding of molecules to the histone tails can affect their ability to open or close the chromatin, which is necessary to facilitate or prevent gene expression, respectively. DNA methylation is the most studied and best characterized epigenetic mark, and also the easiest to study in large sample sizes. DNA methylation is the focus of this thesis and will be described below.

DNA methylation

DNA methylation involves the covalent binding of a methyl group to a Cytosine-5 at a C-phosphate-G (CpG) site in DNA by a methyltransferase. CpG sites are relatively rare in the genome but more common at promoter regions of genes, where they cluster in CpG islands (CGIs). CGIs are defined as regions of 200-500 base pairs with a rich CG content (>50-55%) and a CpG observed/expected ratio of 0.6/0.65[19; 20], although these definitions are open to interpretation[21]. Most CGIs are unmethylated but when they are methylated at promoter regions, they are in general associated with transcriptional repression of the corresponding gene[22]. More than half of the genes in the human genome contain CGIs, mostly at promoter and transcription start sites (TSS), though they also exist within gene bodies. The rest of the genome is depleted for CpGs, and methylation of CpGs in gene bodies does not necessarily lead to blocking transcription elongation[22]. Gene body methylation has even been associated

with transcriptional activation[16]. In addition, non-CGI methylation is more dynamic and tissue-specific than CGI methylation. Mainly long-term silenced genes exhibit promoter CGI methylation, such as imprinted genes or X chromosomal genes, as will be described in the next section[22]. In addition to alterations at CGIs, the regions around these islands, so-called shores, spanning up to 2 kilobases (kb), are more variable and more frequently involved in differential methylation[23; 24]. The Human Epigenome project showed that there is inter-individual variation in DNA methylation, with approximately 50% of the CpG sites having more than 50% variation across all samples. Furthermore, DNA methylation is familial, tissue-specific and can change over time, suggesting it may be partially under genetic control[25].

For example, single nucleotide polymorphisms (SNPs) can regulate CpG methylation. Association studies have been conducted in several (disease) populations, leading to identification of SNPs that cause an increase or decrease of methylation. The findings of these so-called methylation quantitative trait loci (mQTLs) support the role for an underlying genetic mechanism to DNA methylation changes. These mQTLs can also affect allele-specific methylation (ASM), meaning only one allele of the chromosome in the cell is methylated. In addition, it has been hypothesized recently that SNPs not only affect the mean methylation level but also the variability of methylation levels[26].
Other factors that have an effect on DNA methylation include age[27-30]; DNA methylation shows larger differences between twins with increasing age[13; 31], sex [28; 32], medication[33-35] and environment[34; 36-38].

Twin studies are useful for studying epigenetics since monozygotic twins have nearly identical genomes and usually have a shared environment. Despite having identical genotypes, monozygotic twin pairs can be discordant for disease, meaning that one individual of the twin pair may develop a disease while the other remains healthy. Differing phenotypes between monozygotic twin pairs suggests epigenetic mechanisms could underlie disease. A study of methylation patterns involving monozygotic twins uncovered a significant amount of methylation variation between twins, suggesting that molecular mechanisms of heritability are also due to epigenetics[39].

Recent advances in next-generation sequencing and micro-array technology (**Box 2**) provide techniques to measure DNA methylation levels genome-wide and in a large number of individuals[40; 41]. These epigenome-wide association studies (EWAS), similar in approach to genome-wide association studies (GWAS), enable identification of differentially methylated regions (DMRs) between individuals (i.e. cases and controls) or cell lines. Bioinformatics tools are necessary for handling these large amounts of data and for conducting extensive data normalization and quality control to reduce technical noise and artifacts[40].

Box 2: Detection methods:

Several methods have been developed to detect DNA methylation but the focus of this thesis involves unbiased whole genome approaches using Illumina Infinium technology.

Illumina's Infinium Methylation Assay interrogates the methylation status of thousands of CpG's using a microarray-like technology[2]. For the human genome Illumina has developed two BeadChips: the HumanMethylation27 BeadChip[2], covering 27,000 CpG sites, and the more recently released HumanMethylation450 BeadChip[3; 5] covering 450,000 CpG sites. DNA samples are bisulfite treated, turning unmethylated cytosines into uracils, while methylated cytosines remain unchanged. Subsequently, DNA is amplified on a whole-genome level, and is hybridized onto the BeadChips, which are then scanned.

The methylation status is given as a β value, which is a continuous variable between 0 (absent methylation) and 1 (fully methylated), representing the ratio of the intensity of the methylated bead type to the sum of the methylated and unmethylated bead type.

The role of DNA methylation

DNA methylation is involved in several processes, such as X chromosome inactivation, parental imprinting, (embryonic) development, cell specialization, and gene expression.

X chromosome inactivation is a process that allows for only one of the two copies of the X chromosome in females to be expressed. This mechanism prevents females from having twice as much gene expression of genes on the X chromosome than their male counterparts, who only have one copy of the X chromosome[16; 22]. Parental imprinting is the expression of genes in a parent-of-origin manner. Imprinted alleles are silenced and only the non-silenced allele from either the father or the mother is expressed[16; 22]. Furthermore, DNA methylation contributes to cell specialization, tissue differentiation and development by switching genes on or off in certain cells[14].

Gene expression is often interpreted as the intermediate between genotype and phenotype. Only a small proportion of the human genome is transcribed[42]. The role of DNA methylation in gene expression is thought to include interference with transcription factor binding and recruitment of repressors such as histone deacetylases[13; 43]. Although promoter CGI methylation is associated with reduced expression of the corresponding gene, the association between methylation and expression is far more complex[22; 44]. For example, the direction of effect between changes in methylation and in gene expression is not always clear[22]; it can

be difficult to determine if methylation alterations are the cause or the consequence of altered expression, that might in turn lead to phenotypic differences between individuals[11; 34; 40]. In addition, environmental factors (such as smoking) and genetic variation can affect both methylation status and disease, leading to a methylation-disease association due to confounding[11].

DNA methylation and implications in disease

In this thesis, we focus on complex neuropsychiatric disorders, specifically schizophrenia. Schizophrenia is a common mental disorder affecting approximately 1% of the population. It is characterized by two kinds of symptoms: "positive symptoms" consisting of hallucinations, delusions, and paranoia, and "negative symptoms" consisting of loss of interest, and lack of energy. Many different factors including genes and environment jointly contribute to an individual's susceptibility to the disorder. Genetic variation plays a role and the heritability has been estimated to be as high as 80%[45-49]. Extensive research on the genetic background of schizophrenia has proven its highly polygenic nature in which hundreds of genes are likely to play a role[50]. Recent GWAS have identified a number of susceptibility loci but the vast majority of the estimated heritability remains unexplained. The International Schizophrenia Consortium showed that the polygenic basis of schizophrenia involves thousands of common alleles with very small effect that could explain at least 30% of the heritability[50]. Recently, this percentage has been increased to 50%, when it has been estimated that over 8,000 independent common variants contribute to the genetic basis of schizophrenia[51]. Both common genetic variants with small effects as well as rare variants with large effects may contribute to the liability of developing schizophrenia[52] (**Figure 2**). To date, no common genetic variant confers in itself more than a small increase in risk to this disorder[53], implying that epigenetics could explain a part of this "missing heritability".

Epigenetic alterations provide a new and important tool to study disease etiology for psychiatric disorders in general and schizophrenia in particular. The field of cancer research has already made great progress with the study of epigenetics. In cancer etiology and progression, the role of epigenetic modifications is well established[54-56] and DNA

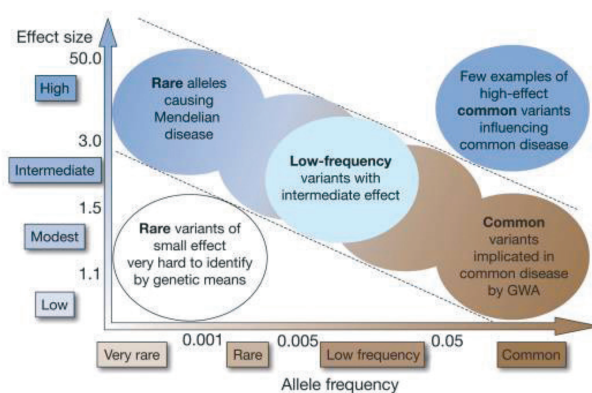


Figure 2: Common versus rare variants
(used with permission from Manolio et al. 2009[1])

methylation appears to be an important mechanism. For example, repression of tumor-suppressor genes by gain in methylation (hypermethylation) can cause unlimited tumor-growth. In addition, hypomethylation (loss of methylation) can cause increased expression of growth-promoting genes. The complexity of these processes is demonstrated by the fact that both hypomethylation and hypermethylation can lead to both gene activation and gene silencing in cancer[34]. In addition to genetic variation, epigenetics provides an extra layer of variation and might mediate the relationship between genotype and environment[34].

Although we understand some of the mechanisms behind DNA methylation, the proportion of inter-individual variation in DNA methylation that contributes to disease is largely unknown. The field of epigenomics in complex common diseases is expanding, and several observations support the role for epigenetics in disease etiology[11]. These observations include identification of epigenetic differences in monozygotic twins discordant for disease, and the increase of complex diseases in the population. The examples in **Box 1** also suggest a non-genetic form of inheritance of epigenetic factors. Combining different forms of genomic information is essential to unravel the etiology of many complex diseases.

The aim of this thesis is to explore the relationships and processes between these different types of genomic information (i.e. genotype, DNA methylation, and gene expression), and layer them together to gain more insight into the mechanisms underlying the neuropsychiatric disorder schizophrenia.

Outline of this thesis

In this thesis we aim to gain more insight into the mechanisms underlying DNA methylation and gene expression, and into how these processes may play a role in susceptibility to schizophrenia.

In **Chapter 2**, we study the association between gene expression and DNA methylation in whole blood using healthy controls. Next, we integrate genotypes to investigate genetic regulation of methylation and expression levels and determine causal relationships between these three-way associations. In **Chapter 3**, we use the results of a large meta-analysis to investigate whether the most-associated SNPs with schizophrenia regulate gene expression. Subsequently, we test the transcripts associated with these SNPs for differential expression in schizophrenia patients compared to healthy subjects. The results from this large meta-analysis are also used in **Chapter 4**, in which we examine whether biologically relevant SNPs (according to DNA methylation and gene expression associations) are enriched for schizophrenia susceptibility loci (results from the same meta-analysis as Chapter 3). In **Chapter 5**, we investigate associations between DNA methylation and genotypes using three different approaches and explore the genetic contribution to variability in DNA methylation using 22 nuclear families each containing one child diagnosed with schizophrenia. In **Chapter 6**, we explore the association of microRNA 137, which has been previously linked to schizophrenia, with DNA methylation in brain tissue of schizophrenia and bipolar patients, and healthy controls. Finally, in **Chapter 7** we present a discussion of our findings.

References

1. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.
2. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K.L. (2009). Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1, 177-200.
3. Roseboom, T.J., Painter, R.C., van Abeelen, A.F., Veenendaal, M.V., and de Rooij, S.R. (2011). Hungry in the womb: what are the consequences? Lessons from the Dutch famine. Maturitas 70, 141-145.
4. Susser, E., Neugebauer, R., Hoek, H.W., Brown, A.S., Lin, S., Labovitz, D., and Gorman, J.M. (1996). Schizophrenia after prenatal famine. Further evidence. Arch Gen Psychiatry 53, 25-31.
5. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics 98, 288-295.
6. Roseboom, T., de Rooij, S., and Painter, R. (2006). The Dutch famine and its long-term consequences for adult health. Early Hum Dev 82, 485-491.
7. Bale, T.L., Baram, T.Z., Brown, A.S., Goldstein, J.M., Insel, T.R., McCarthy, M.M., Nemeroff, C.B., Reyes, T.M., Simerly, R.B., Susser, E.S., et al. (2010). Early life programming and neurodevelopmental disorders. Biol Psychiatry 68, 314-319.
8. Khashan, A.S., Abel, K.M., McNamee, R., Pedersen, M.G., Webb, R.T., Baker, P.N., Kenny, L.C., and Mortensen, P.B. (2008). Higher risk of offspring schizophrenia following antenatal maternal exposure to severe adverse life events. Arch Gen Psychiatry 65, 146-152.
9. Qiu, J. (2006). Epigenetics: unfinished symphony. Nature 441, 143-145.
10. Egger, G., Liang, G., Aparicio, A., and Jones, P.A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. Nature 429, 457-463.
11. Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet 12, 529-541.
12. Radtke, K.M., Ruf, M., Gunter, H.M., Dohrmann, K., Schauer, M., Meyer, A., and Elbert, T. (2011). Transgenerational impact of intimate partner violence on methylation in the promoter of the glucocorticoid receptor. Transl Psychiatry 1, e21.
13. Bird, A. (2007). Perceptions of epigenetics. Nature 447, 396-398.
14. Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447, 425-432.
15. Rando, O.J., and Verstrepen, K.J. (2007). Timescales of genetic and epigenetic inheritance. Cell 128, 655-668.
16. Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9, 465-476.

17. Houston, I., Peter, C.J., Mitchell, A., Straubhaar, J., Rogaev, E., and Akbarian, S. (2013). Epigenetics in the human brain. Neuropsychopharmacology 38, 183-197.
18. Ito, T. (2007). Role of histone modification in chromatin dynamics. J Biochem 141, 609-614.
19. Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99, 3740-3745.
20. Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol 196, 261-282.
21. Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. Biostatistics 11, 499-514.
22. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13, 484-492.
23. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41, 178-186.
24. Irizarry, R.A., Wu, H., and Feinberg, A.P. (2009). A species-generalized probabilistic model-based definition of CpG islands. Mamm Genome 20, 674-680.
25. Bjornsson, H.T., Sigurdsson, M.I., Fallin, M.D., Irizarry, R.A., Aspelund, T., Cui, H., Yu, W., Rongione, M.A., Ekstrom, T.J., Harris, T.B., et al. (2008). Intra-individual change over time in DNA methylation with familial clustering. Jama 299, 2877-2883.
26. Feinberg, A.P., and Irizarry, R.A. (2010). Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A 107 Suppl 1, 1757-1764.
27. Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., et al. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet 8, e1002629.
28. Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S., and Ophoff, R.A. (2009). The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PLoS One 4, e6767.
29. Horvath, S., Zhang, Y., Langfelder, P., Kahn, R.S., Boks, M.P., van Eijk, K., van den Berg, L.H., and Ophoff, R.A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol 13, R97.
30. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol 14, R115.
31. Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A 102, 10604-10609.
32. El-Maarri, O., Becker, T., Junen, J., Manzoor, S.S., Diaz-Lacava, A., Schwaab, R., Wienker, T., and Oldenburg, J. (2007). Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. Hum Genet 122, 505-514.

33. Dong, E., Nelson, M., Grayson, D.R., Costa, E., and Guidotti, A. (2008). Clozapine and sulpiride but not haloperidol or olanzapine activate brain DNA demethylation. Proc Natl Acad Sci U S A 105, 13614-13619.
34. Feinberg, A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature 447, 433-440.
35. Melas, P.A., Rogdaki, M., Osby, U., Schalling, M., Lavebratt, C., and Ekstrom, T.J. (2012). Epigenetic aberrations in leukocytes of patients with schizophrenia: association of global DNA methylation with antipsychotic drug treatment and disease onset. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 26, 2712-2718.
36. Alasaari, J.S., Lagus, M., Ollila, H.M., Toivola, A., Kivimaki, M., Vahtera, J., Kronholm, E., Harma, M., Puttonen, S., and Paunio, T. (2012). Environmental stress affects DNA methylation of a CpG rich promoter region of serotonin transporter gene in a nurse cohort. PLoS One 7, e45813.
37. Waterland, R.A., and Jirtle, R.L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Mol Cell Biol 23, 5293-5300.
38. Weaver, I.C., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., and Meaney, M.J. (2004). Epigenetic programming by maternal behavior. Nat Neurosci 7, 847-854.
39. Kaminsky, Z.A., Tang, T., Wang, S.C., Ptak, C., Oh, G.H., Wong, A.H., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C., et al. (2009). DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet 41, 240-245.
40. Bock, C. (2012). Analysing and interpreting DNA methylation data. Nat Rev Genet 13, 705-719.
41. Laird, P.W. (2010). Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet 11, 191-203.
42. Strachan, T., and Read, A.P. (2004). Human molecular genetics 3.(London ; New York: Garland Press).
43. Sengupta, N., and Seto, E. (2004). Regulation of histone deacetylase activities. J Cell Biochem 93, 57-67.
44. Siegfried, Z., and Simon, I. (2010). DNA methylation and gene expression. Wiley Interdiscip Rev Syst Biol Med 2, 362-371.
45. Cardno, A.G., and Gottesman, II. (2000). Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. American journal of medical genetics 97, 12-17.
46. Lichtenstein, P., Yip, B.H., Bjork, C., Pawitan, Y., Cannon, T.D., Sullivan, P.F., and Hultman, C.M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 373, 234-239.
47. Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60, 1187-1192.
48. van Dongen, J., and Boomsma, D.I. (2013). The evolutionary paradox and the missing heritability of schizophrenia. Am J Med Genet B Neuropsychiatr Genet 162, 122-136.

49. Bienvenu, O.J., Davydow, D.S., and Kendler, K.S. (2011). Psychiatric 'diseases' versus behavioral disorders and degree of genetic influence. Psychol Med 41, 33-40.

50. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.

51. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45, 1150-1159.

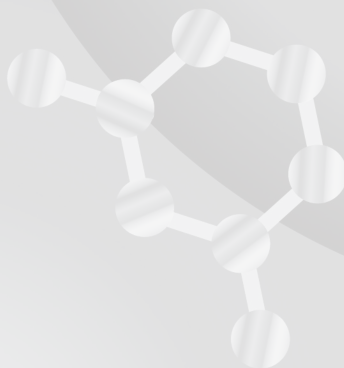
52. Maric, N.P., and Svrakic, D.M. (2012). Why schizophrenia genetics needs epigenetics: a review. Psychiatr Danub 24, 2-18.

53. Bray, N.J., Leweke, F.M., Kapur, S., and Meyer-Lindenberg, A. (2010). The neurobiology of schizophrenia: new leads and avenues for treatment. Curr Opin Neurobiol 20, 810-815.

54. Feinberg, A.P., and Tycko, B. (2004). The history of cancer epigenetics. Nat Rev Cancer 4, 143-153.

55. Kulis, M., and Esteller, M. (2010). DNA methylation and cancer. Adv Genet 70, 27-56.

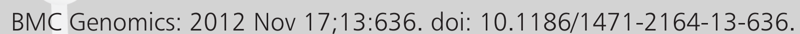
56. Lechner, M., Boshoff, C., and Beck, S. (2010). Cancer epigenome. Adv Genet 70, 247-276.



Chapter 2

Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects

Kristel R van Eijk, Simone de Jong, Marco PM Boks, Terry Langeveld, Fabrice Colas, Jan H Veldink, Carolien GF de Kovel, Esther Janson, Eric Strengman, Peter Langfelder, René S Kahn, Leonard H van den Berg, Steve Horvath, Roel A Ophoff



BMC Genomics: 2012 Nov 17;13:636. doi: 10.1186/1471-2164-13-636.

Abstract

Background

The predominant model for regulation of gene expression through DNA methylation is an inverse association in which increased methylation results in decreased gene expression levels. However, recent studies suggest that the relationship between genetic variation, DNA methylation and expression is more complex.

Results

Systems genetic approaches for examining relationships between gene expression and methylation array data were used to find both negative and positive associations between these levels. A weighted correlation network analysis reveals that i) both transcriptome and methylome are organized in modules, ii) co-expression modules are generally not preserved in the methylation data and vice-versa, and iii) highly significant correlations exist between co-expression and co-methylation modules, suggesting the existence of factors that affect expression and methylation of different modules (i.e., *trans* effects at the level of modules). We observed that methylation probes associated with expression in *cis* were more likely to be located outside CpG islands, whereas specificity for CpG island shores was present when methylation, associated with expression, was under local genetic control. A structural equation model based analysis found strong support in particular for a traditional causal model in which gene expression is regulated by genetic variation via DNA methylation instead of gene expression affecting DNA methylation levels.

Conclusions

Our results provide new insights into the complex mechanisms between genetic markers, epigenetic mechanisms and gene expression. We find strong support for the classical model of genetic variants regulating methylation, which in turn regulates gene expression. Moreover we show that, although the methylation and expression modules differ, they are highly correlated.

Keywords

DNA methylation, Gene expression, Association, Epigenetics, WGCNA

Background

Epigenetics has been described as the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states[1]. DNA methylation is one of several forms of epigenetic modifications and involves the covalent binding of a methyl group to a Cytosine-5 at a C-phosphate-G (CpG) site. These sites are relatively rare in the genome but more common at promoter regions of genes, also called CpG islands (CGIs). CpGs in these islands are less likely to be methylated than CpGs outside these islands. Recent studies have shown that specifically the CpGs in the shore of CGIs are most frequently involved in differential methylation between tissues or experimental groups[2,3]. Increased methylation of CpG islands at 5' end of a gene is associated with gene repression. Possible mechanisms for repression include interference with transcription factor binding or through the recruitment of repressors such as histone deacetylases[4].

Although one would expect DNA methylation at CGIs and expression of the nearby gene to be inversely correlated, this is not necessarily the case. Recent reports also identified positive associations between expression and methylation levels[5-7]. However, negative associations between methylation and expression were found to be enriched particularly in CGIs[6] and promoter regions[5].

Around 30% of gene expression levels in cell lines[8] and 23% of DNA methylation levels in blood are heritable[9] and genetic variation associated with expression and methylation levels has been identified in several organisms[6,10-12], tissues[13] and populations[14]. Local (*cis*) and distal (*trans*) associations of genetic variation with gene expression levels have been observed. With the arrival of high-throughput DNA methylation assays, methylation quantitative trait loci (mQTLs) can now be studied genome-wide in any tissue or cell type of interest. Similar to expression QTLs (eQTLs), more *cis* than *trans* regulation has been identified[5-7] but peak enrichment for mQTLs is located in much closer proximity to transcription start sites than that of eQTLs[6].

Attempts to identify three-way associations between genetic variants, expression and methylation on a genome-wide scale in four different brain regions did not identify co-regulation of methylation and expression by the same genetic variants[6], while a study of cerebellar samples did identify three-way associations for a number of genes[7]. In lymphoblastoid cell lines of 77 individuals of the Yoruba Hapmap population, co-regulation of expression and methylation levels by the same genetic variants was also found, suggesting a shared mechanism, whereby a genetic variant influences methylation, which in turn influences expression levels[5]. Strong evidence exists that both patterns of CpG methylation[15,16] and gene expression[13,17,18] differ between tissues.

The aims of the current study are i) to relate expression levels to methylation levels, ii) to relate co-expression modules (clusters of expression probes) to co-methylation modules, iii)

and to study the relationship between genetic markers, methylation and expression in whole blood of a relatively large (n=148) set of healthy human subjects. For the genetic analysis, we examined the associations of methylation and expression levels and identified genetic markers associated with these levels. To infer directionality in the relationships between genetic variants, methylation and expression, we calculated local edge orienting (LEO) scores based on structural equation models[19]. This method has been applied successfully before and will aid in elucidating the nature of relationship between genetic variation, methylation and expression[20-23].

Results

Associations between methylation and expression levels

A multivariate linear model analysis for regressing a gene expression level on a methylation level and age and gender resulted in the identification of 522 negative and 276 positive *cis* associations between methylation and expression levels (False Discovery Rate (FDR) 5% corrected). A negative association between methylation and transcript level means that increased methylation levels correlate with decreased expression levels, whereas a positive correlation includes levels that both increase or decrease. These associations involved 517 different *cis*-acting CpG loci (from 461 unique genes) and 495 corresponding expression probes (representing 452 unique genes). For *trans* effects, we found evidence for 844 negative and 1,806 positive associations between methylation and expression levels involving 705 different methylation probes (from 630 distinct genes), and 170 different expression probes (representing 157 unique genes). Full results are given in **Table 1** and Additional file 1: **Table S1**. Because of the stringent Bonferroni corrections for multiple testing with the number of methylation probes multiplied by the number of expression probes, the effect sizes of surviving *trans* effects were significantly larger than for *cis* effects with adjusted explained variance (R^2) ranging from 23 to 60 percent for *trans* effects and 0.8 to 50 percent

Table 1 Number of probes constituting significant methylation and expression combinations and their association with SNPs

	unique	+	-	**overlap**	**SNP *cis***
	Cis associations				
Methylation	517	224	354	61	69 probes (13.3%), 86 independent loci
Expression	495	214	336	55	62 probes (12.5%), 73 independent loci
	Trans associations				
Methylation	705	585	230	110	1 probe (<1%)
Expression	170	101	117	48	0 probes

for *cis* regulation (Additional file 2: **Figure S1a**). Another trend that we observed was that *cis* effects are enriched for negative correlations (65.4% overall) while positive correlations between DNA methylation and gene expression are more frequently observed with *trans* associations (68.2%; Fisher's Exact test for count data $p<2.2e\text{-}16$), (Additional file 2: **Figure S1b**).

This table shows the significant methylation and expression combinations, subdivided into *cis* and *trans* associations. The first column shows the counts of unique probes (for methylation and expression). The second and third columns indicate the number of probes positively (+) or negatively (−) associated. The fourth column indicates the overlapping probes: methylation or expression probes that are associated with expression or methylation levels in both directions. The last column indicates the number (and %) of unique probes associated with SNPs and the number of independent (pruned r^2 of 0.2) loci in *cis*.

DNA methylation and gene expression are regulated by genetic variants

Expression levels and methylation levels that were significantly associated with each other were separately tested for regulation by genetic variants. The methylation and expression levels were taken as phenotypes and a linear model of allele dosage, with age and gender as covariates, was tested using PLINK[24]. We focused on local (*cis*) effects only and observed that approximately 13.7% of methylation signals and 12.5% of gene expression levels are associated with single nucleotide polymorphisms (SNPs). Results are given in **Table 1**, where the number of independent loci, associated with probes, is reported. These were retrieved by pruning the SNPs with an R^2 of >0.2 to prevent reporting many SNP associations of the same signal due to linkage disequilibrium (LD). Full results are in Additional file 3: **Table S2**.

Cis-acting methylation sites under genetic control are over-represented in CpG island shores

We examined the regional distribution of methylation sites (n=517) that are associated with nearby gene expression levels and observed a significant overrepresentation of these loci outside CpG islands and shores compared to all probes present on the Illumina array (50.9% vs 26%; Fisher's Exact $p<2.2e\text{-}16$). This coincided with a significant underrepresentation of DNA methylation signal at CpG islands (13.5% vs. 42%, Fisher's Exact $p<2.2e\text{-}16$) and a modest increase at the shores flanking CpG islands (35.6% vs. 32%, Fisher's Exact $p=0.056$). The regional distribution of DNA methylation associated with gene expression is somewhat different when DNA methylation is under genetic control. In case of *cis* genetic regulation we observed a further enrichment of DNA methylation at shores of CpG islands (53.4%, Fisher's Exact $p=1.3e\text{-}4$), whereas *trans* genetic regulation shows the opposite effect and

is less frequently observed for DNA methylation at shores (24.4%, Fisher's Exact $p=3.9e-5$). The overall results are presented in **Table 2**.

Table 2: Distribution of results over CpG islands and shores. Methylation probes were classified into three categories according to UCSC browser (http://genome.ucsc.edu/); CpG islands, CGI shores (up to 2kb around an island) and outside islands or shores. Differences compared to Illumina Human Methylation27K array were tested using Fisher's Exact for count data (Bonferroni threshold p:0.05/9=0.006). A downward arrow indicates significantly lower percentage of probes while an upward arrow indicates significantly higher percentage of observations compared to the overall probe distribution on the Illumina array.

Location	Illumina Human Methylation 27		Methylation & expression *cis*		Methylation & expression *trans*		Methylation & expression & SNP *cis*	
Island	11,582	42%	**70**	**13.5%, *p*<2.2e-16↓**	269	38.2%, *p*=0.04	**11**	**15.1%, *p*=1.1e-06↓**
Island shore (2kb)	8,718	32%	184	35.6%, *p*=0.056	**172**	**24.4%, *p*=3.9e-05↓**	**39**	**53.4%, *p*=1.3e-04↑**
Outside island/shore	7,278	26%	**263**	**50.9%, *p*<2.2e-16↑**	**264**	**37.4%, *p*=2.5e-10↑**	23	31.5%, ns
Total	27,578		517		705		73	

Methylation probes were classified into three categories according to UCSC browser (http://genome.ucsc.edu/); CpG islands, CGI shores (up to 2kb around an island) and outside islands or shores. Differences compared to Illumina Human Methylation27K array were tested using Fisher's Exact for count data (Bonferroni threshold p:0.05/9=0.006). A downward arrow indicates significantly lower percentage of probes while an upward arrow indicates significantly higher percentage of observations compared to the overall probe distribution on the Illumina array.

Causal relationships between *cis*-acting methylation and expression probes

To study the causal relationship between methylation and expression levels that were significantly associated, we focused the analysis on pairs of methylation and expression levels with a common *cis*-acting SNP. We selected the top 20 methylation probes, associated with 19 expression probes that were significantly associated with 147 single common SNPs. Since alleles can be considered fixed features of a genome, we selected SNPs as causal anchors and used a model with residuals of the 20 methylation and 19 expression probes corrected for age and gender. For the causal scenario SNP → Methylation → Expression, we found 44 combinations (29.9%) with a LEO score above 0.8, involving seven unique genes (**Table 3**). Of these, 20 combinations have a strikingly high LEO score of 3 or higher; for most of these 20 combinations, the model fitting p-value of the causal model SNP → Methylation → Expression is above 0.01, indicating a good fit and lending further credence to these results

(Additional file 4: **Table S3**). For the model SNP → Expression → Methylation, we found 10 combinations (6.8%) with a LEO score above 0.8, involving again seven unique genes (**Table 3**). The model fitting p-values of these combinations are generally worse (below 0.01), indicating that the linear structural equations models do not fit the data as well and suggesting caution in interpreting the results. A full list of combinations is given in Additional file 4: **Table S3**. Some SNPs were found to be in high linkage disequilibrium (LD), especially in the Major Histocompatibility Complex (MHC) region on chromosome 6. Therefore only the top SNPs are listed in **Table 3**. We choose to investigate these two models since we were interested in the causal direction between DNA methylation and gene expression, after regulation by genetic variation, excluding models 4 and 5. Model 3, was not informative since we already selected SNPs for association with both methylation and expression.

Table 3 Top results LEO analysis, Results for top SNPs

Gene Symbol	M&E	CGI	LEO model	LEO score	P-value	Top SNP	Chr	Bp	Full name
BTN3A2	–	Outside	S→M→E	6.90	0.15	rs2093169	6	26,603,078	butyrophilin, subfamily 3, member A2
HP	–	Outside	S→M→E	4.24	0.82	rs8044555	16	70,710,256	haptoglobin
CTSW	–	Outside	S→M→E	2.73	0.13	rs11227306	11	65,335,248	cathepsin W
NAPRT1	–	Shore	S→M→E	2.69	0.11	rs4874159	8	144,742,093	nicotinate phosphoribosyltrans-ferase domain containing 1
PHACS	–	Shore	S→M→E	1.50	2.9e-03	rs4755227	11	44,078,659	1-aminocyclopropane-1-carbox-ylate synthase homolog
PNMA3	+	Shore	S→M→E	1.36	0.16	rs6627737	X	151,971,610	nicotinate phosphoribosyltrans-ferase domain containing 1
CDC16	–	Island	S→M→E	1.09	0.01	rs11147317	13	113,957,498	cell division cycle 16 homolog (S. cerevisiae)
HRASLS3	–	Shore	S→E→M	2.42	7.2e-04	rs2030731	11	63,130,224	phospholipase A2, group XVI
TACSTD2	–	Island	S→E→M	2.08	9.5e-03	rs11207272	1	58,846,018	tumor-associated calcium signal transducer 2
SRXN1	–	Shore	S→E→M	1.87	5.4e-03	rs6076864	20	569,825	sulfiredoxin 1 homolog (S. cerevisiae)
C21orf56	–	Outside	S→E→M	1.30	2.8e-03	rs8133866	21	46,423,604	chromosome 21 open reading frame 56
BTN3A2	–	Outside	S→E→M	1.14	6.4e-05	rs12199613	6	26,475,197	butyrophilin, subfamily 3, member A2
WBSCR27	–	Shore	S→E→M	0.95	8.6e-04	rs11763011	7	72,922,084	Williams Beuren syndrome chromosome region 27
GSTM3	–	Island	S→E→M	0.88	1.4e-03	rs11807	1	110,062,265	glutathione S-transferase mu 3 (brain)

This table contains top probes resulting from causality analysis (LEO scores >0.8). The top seven genes fit the causal scenario SNP → Methylation → Expression (S→M→E), while the bottom seven genes fit the reverse model in which DNA methylation is regulated by gene expression that is under genetic control (S→E→M). The Gene Symbol is given in the first column. The second column indicates whether the methylation and expression levels are associated negatively (−) or positively (+). The third column indicates whether the methylation probe is located in a CpG island (CGI), in the shore, or outside both. The columns "LEO model", "LEO score" and "P-value" indicate which causal model fits best with the corresponding LEO score and P-value. This model fitting p-value is calculated using the model chi-square statistic statistic. The chi-square statistic tests the null hypothesis that the model is correct, thus a p-value > 0.01 indicates good fit. The next column indicates the SNP most significantly associated. The last three columns contain chromosome number and base pair location (NCBI build 36) of the SNP and full name of the gene.

A locus in the *BTN3A2* gene passed the LEO threshold of 0.8 for both models SNP → Methylation → Expression (LEO score 6.2 based on causal anchor rs9467632) and SNP → Expression → Methylation (LEO score 1.14 based on causal anchor rs12199613). The two SNPs that were used as causal anchors are in moderate LD (R^2=0.092, D'=0.68 based on 1000 Genomes Pilot 1 CEU population[25]. The significant results in both directions could indicate a bi-directional causal interaction between expression and methylation. However, while the model SNP → Methylation → Expression fits the data well (model fitting p-value p=0.10), the model SNP → Expression → Methylation does fits the data poorly (model fitting p-value p=6.4e-5). Thus, while the evidence for the SNP → Methylation → Expression model for *BTN3A2* is strong, the evidence for the SNP → Expression → Methylation model is weak.

Weighted correlation network analysis of expression and methylation data

We separately constructed co-expression and co-methylation networks from the expression and methylation data, respectively (Additional file 5- Supplementary Methods), using the Weighted Correlation Network Analysis framework WGCNA[26,27]. In expression data (13,843 genes) we identified 23 co-expression modules (labeled 1–23) with sizes ranging from 32 to 1,520 genes. Additional file 6; **Table 1** provides a brief overview of the expression modules along with 10 top hub genes (genes with highest module membership) in each module. A total of 7,743 (56% of total) genes were assigned to a module while 6,091 background genes were not assigned to a module. Background genes are labeled 0 and colour-coded in grey. Gene ontology (GO) enrichment analysis revealed significant enrichment of multiple modules in various GO terms (**Table 4**), which provides evidence that these modules are biologically meaningful. A table listing module membership of all genes in expression modules is provided in Additional file 7.

Table 4 Top GO enrichment terms for expression modules

Module	Size	Rank	p.Bcnf	Fraction	Ontology	Term name
1	1520	1	1.20E-14	0.65147	CC	membrane-bounded organelle
1	1520	2	1.90E-14	0.65007	CC	intracellular membrane-bounded organelle
2	703	1	1.30E-05	0.04885	CC	ribosome
2	703	2	1.50E-05	0.07481	BP	translation
3	658	1	0.00018	0.16382	CC	extracellular region
4	647	1	0.011	0.63711	CC	membrane-bounded organelle
6	442	1	7.60E-07	0.27229	BP	response to stress
6	442	2	2.80E-06	0.34217	BP	signal transduction
7	426	1	9.40E-07	0.18734	BP	immune system process
8	407	1	0.0042	0.66755	CC	intracellular membrane-bounded organelle
9	387	1	1.70E-05	0.82961	CC	intracellular part
10	355	1	1.80E-05	0.07855	BP	ncRNA metabolic process
10	355	2	3.60E-05	0.19335	CC	mitochondrion
12	306	1	2.00E-16	0.1134	CC	ribosome
12	306	2	4.00E-16	0.17182	CC	ribonucleoprotein complex
12	306	3	1.20E-14	0.14777	BP	translation
12	306	4	2.30E-14	0.09278	BP	viral transcription
13	260	1	0.00041	0.025	CC	hemoglobin complex
13	260	2	0.011	0.025	BP	heme biosynthetic process
14	237	1	0.014	0.61086	MF	protein binding
15	118	1	0.018	0.25688	BP	intracellular signal transduction
15	118	2	0.032	0.12844	BP	small GTPase mediated signal transduction
16	108	1	1.30E-05	0.18478	BP	translation
16	108	2	0.00017	0.1087	BP	ribosome biogenesis
17	99	1	2.30E-09	0.36957	CC	mitochondrion
18	72	1	3.30E-06	0.1791	BP	platelet activation
18	72	2	1.20E-05	0.22388	BP	blood coagulation
19	60	1	1.40E-17	0.31034	MF	structural constituent of ribosome
19	60	2	2.90E-17	0.32759	CC	ribosome
20	52	1	7.10E-23	0.32653	BP	type I interferon-mediated signaling pathway
20	52	2	7.10E-23	0.32653	BP	cellular response to type I interferon
23	32	1	0.053	0.17241	CC	external side of plasma membrane

For each module we list the top enriched GO terms. Columns list the module label, module size, rank of the enrichment p-value for that particular module, the Bonferroni-corrected enrichment p-value (the correction is performed with respect to the number of GO terms), fraction of the module genes also in the GO term, GO ontology, and GO term name. Multiple expression modules exhibit significant enrichment. Row shading separates modules for easier reading. The enrichment provides evidence that the modules are biologically meaningful.

In methylation data (13,569 genes) we identified 9 modules of sizes ranging from 37 to 1,067 genes. Additional file 6; **Table 2** provides a brief overview of the methylation

modules along with 10 top hub genes (genes with highest module membership) in each module. For reader-friendliness, methylation module labels were chosen such that modules with significant overlap with expression modules carry the same label (Methods). A total of 4,088 (30% of total) genes were assigned to a module, while 9,481 were not assigned. We observed that strong co-expression relationships tend to be more frequent than strong co-methylation. GO enrichment analysis of methylation modules revealed multiple significantly enriched categories (**Table 5**). A table listing module membership of all genes is provided in Additional file 8.

Table 5 Top GO enrichment terms for methylation modules

Module	Size	Rank	p.Bonf	Fraction	Ontology	Term name
32	1067	1	4.20E-09	0.1875	CC	extracellular region
1	1045	1	8.20E-16	0.790123	CC	intracellular
1	1045	3	1.50E-11	0.083333	CC	ribonucleoprotein complex
30	616	1	2.50E-21	0.34687	BP	anatomical structure development
30	616	2	1.60E-18	0.189509	BP	nervous system development
7	594	1	8.30E-10	0.157424	BP	immune system process
7	594	2	9.10E-10	0.177102	MF	receptor activity
3	427	1	3.20E-12	0.259947	CC	extracellular region
12	130	1	0.0032	0.075	BP	DNA recombination
2	105	1	1.70E-10	0.191919	BP	lymphocyte activation
2	105	2	2.30E-10	0.20202	BP	leukocyte activation
2	105	3	3.60E-10	0.323232	BP	immune system process

For each module we list the top enriched GO terms. Columns list the module label, module size, rank of the enrichment p-value for that particular module, the Bonferroni-corrected enrichment p-value (the correction is performed with respect to the number of GO terms), fraction of the module genes also in the GO term, GO ontology, and GO term name. Multiple methylation modules exhibit significant enrichment. Row shading separates modules for easier reading. Again, the enrichment provides evidence that the modules are biologically meaningful.

Preservation of co-expression modules in methylation data and vice versa

A natural question is whether the expression and methylation modules are related. At the most basic level one can ask whether the expression and methylation modules can be matched based on significant overlap of the genes in each module. We found that expression and methylation modules in general exhibit relatively few overlapping genes (Additional file 9), although some of the overlaps are statistically significant. The most significant overlap ($p=6e-12$) is observed between the largest co-expression module and the

largest co-methylation module. While the cross-tabulation based module overlap analysis is a simple and intuitive way of assessing module preservation, it has several limitations. In particular, it cannot be used to make strong statements about the lack of module preservation since alternative module detection methods applied to the test data may lead to different results. A rigorous module preservation analysis is based on the network module preservation statistic Zsummary (Methods) since it is independent of the vagaries of detecting modules in test data[28]. We found that the largest expression module 1 (enriched in intracellular-related terms) exhibits moderate preservation, *Zsummary*≈5. Modules 9 (enriched in intracellular-related terms), 12 (ribosome), 16 (translation), 17 (mitochondrion), and 19 (ribosome) show weak evidence of preservation, while all other expression modules show no evidence of preservation in methylation data (*Zsummary* ≤ 2, **Figure 1A**). For the methylation modules we found that modules 1 (intracellular) and 2 (lymphocyte activation) show weak to moderate evidence for preservation, while all other modules show no evidence of preservation (*Zsummary* < 2, **Figure 1B**). It is known that the Zsummary statistic tends to increase with module size, reflecting the intuition that a preservation signal observed among many genes is more significant than a similar preservation signal observed among only a few genes. To measure relative preservation irrespective of module size, the authors of[28] proposed the use of a rank-based statistic *medianRank*. Additional file 10 shows the *medianRank* statistics in this study. The modules with high Zsummary have low (i.e., near top) ranks. Hence, the two preservation statistics offer a largely consistent picture of module preservation, even though they measure very different quantities.

The weak preservation of co-expression modules in methylation data and *vice-versa* shows that in general modules (clusters) of expression probes do not correspond to modules of methylation probes. However, we found strong correlations between co-expression modules and co-methylation modules as described in the following.

Associations of expression and methylation eigengenes

Although the composition of co-expression modules is different from that of co-methylation modules, we observed strong correlations of expression and methylation module eigengenes (**Figure 1C**). A module eigengene is a mathematically optimal way of summarizing the levels of a module (Methods). For example, eigengenes of methylation modules 2 and 7 (both enriched in immune system/response terms) are strongly correlated with multiple expression eigengenes such as ME 7 (enriched in immune system process), 12 (ribosome), 15 (intracellular signal transduction), 19 (ribosome), and 22 (no significant enrichment). Methylation module eigengenes 3 (extracellular region) and 30 (anatomical structure morphogenesis, nervous system development) also relate to several expression module

eigengenes but the associations are weaker. In summary, we observed multiple strong correlations between expression and methylation module eigengenes.

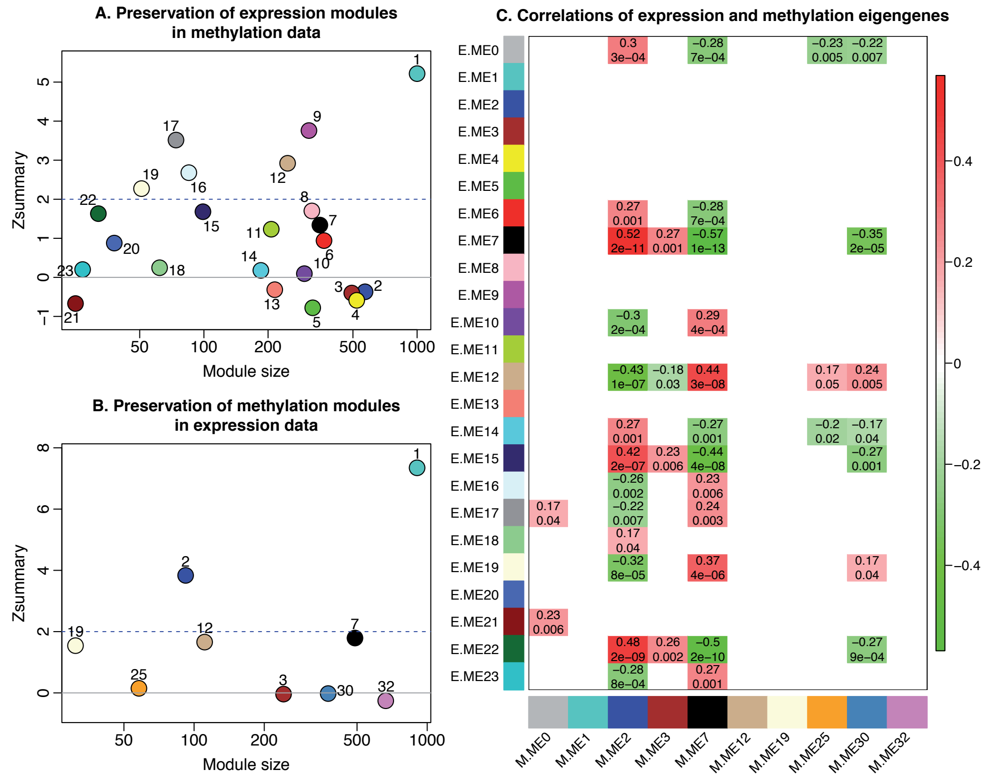


Figure 1 Preservation and association of co-expression and co-methylation modules. A. Module preservation statistic *Zsummary* that summarizes evidence of preservation of expression modules in methylation data. Each module is labelled by a numeric label and the corresponding colour. Values of *Zsummary* below 2 indicate no evidence of preservation; values between 2 and 5 indicate weak to moderate evidence for preservation. Only the largest module, labelled 1 (turquoise), exhibits *Zsummary* above 5 that can be considered moderate-strong evidence of preservation. B. Analogous plot of the *Zsummary* statistic for preservation of methylation modules in expression data. As in expression data, only the largest module (also labelled 1, turquoise) exhibits moderate-strong evidence of preservation. C. Robust correlations and the corresponding p-values of expression (y-axis) and methylation (x-axis) eigengenes. Each row corresponds to an expression eigengene (E.ME) labelled by numeric module label and colour. Each column corresponds to a methylation eigengene (M.ME) labelled by numeric module label and colour. Numbers in the table report the robust correlation and the corresponding p-value of the respective expression and methylation eigengenes. Only correlations whose p-value is below 0.05 are displayed. The table is colour-coded according to correlation such that (strong) green colour corresponds to (strong) negative correlations, and (strong) red colour corresponds to (strong) positive correlations.

Module membership of individual genes in expression and methylation modules

Weighted correlation network methods allow one to define a continuous measure of module membership for each variable in each module as the correlation of the variable profile with the module eigengene (Methods). Additional files 7 and 8 report the module membership (based on expression and methylation profiles) of all genes in all modules. Since the expression and methylation data were measured for the same set of samples, we are able to also provide the module membership of expression profiles in methylation modules and *vice-versa*. These Supplementary Files serve as a resource for relating expression and methylation probes to the modules.

Discussion

We investigated the relationship between genetic variation, DNA methylation and gene expression in a sample of 148 healthy subjects using array-based data derived from whole blood. We found both negative (levels in opposite direction) and positive (levels in same direction) associations between *cis*-acting DNA methylation probes and corresponding gene expression levels, confirming previous reports that DNA methylation and gene expression located within a *cis*-region can be both positively and negatively associated, but are predominantly negative[5-7].

In this study we applied FDR correction for multiple testing for *cis* associations between methylation and expression, but imposed a more stringent genome-wide significance threshold for *trans* effects since there is a considerable debate in the literature whether such relationships are reproducible[29,30]. This resulted in a limited number of *trans* associations that do survive this threshold but with relatively strong effect sizes. It is of note that such *trans* associations are enriched for positive correlations, whereas traditionally it is expected that methylation and expression are inversely correlated. We hypothesised that these involve genes involved in general methylation pathways, such as genes that induce the attachment of a methyl group. However, a gene ontology analysis did not show any overrepresented pathways (data not shown).

Furthermore, we observed that methylation probes with *cis*-acting effects on gene expression levels are less likely to be located in CpG islands and more likely to be present outside CGIs and shores insofar they were not regulated by genetic variation. Tissue- and cell type-specific methylation occurs much more often in gene bodies (outside island and shores) than in CpG island promoters[31], indicating that methylation at CpG sites in CpG islands is much more static, which could explain the underrepresentation of CpG sites associated with expression (and SNPs) in CpG islands. Only for those CpGs that were associated with SNPs, we did concur with previous studies showing more frequent associations with expression in island shores[2,3]. CpG sites located in shores tend to be more variable among individuals and this

might lead to an increased number of association findings. In addition, *trans* associations are less likely to be located in island shores and more likely to be positioned outside CGIs and shores. Also, *trans* associations are more likely to be positive (67%).

Identification of genetic variants (SNPs) influencing the methylation and expression levels showed that in more than 12% of methylation-expression *cis*-pairs, the methylation and/or the expression level was associated with a SNP in *cis*, suggesting genetic control of these levels.

Further analysis of genetic regulators (SNPs) of methylation and expression levels investigating the causality revealed three-way causal relationships. Previous studies have attempted to identify three-way associations in various tissues, with mixed results[6,7]. We used local structural equation models to calculate local edge orienting (LEO) scores based on using a *cis*-acting SNP as causal anchor[19,32]. We find that the traditional model of genetic variants regulating methylation, which in turn regulates gene expression to be most common in most of the three-way associations that showed significant evidence for causality (as was hypothesized in literature[5]). The set of genes for which the S→M→E model fits best does not exhibit significant enrichment for specific functions or pathways. Since the S→M→E model is expected to be ubiquitous, the lack of enrichment is not surprising. However, one of the genes that fit this model, *PNMA3*, is located on the X chromosome. Since inactivation in females may be a confounding factor when analyzing X chromosomes, we repeated the association analysis for all significant X chromosomes in males only. We observed no significant differences when using males-only, which confirms that the *PNMA3* finding is likely to be true. Strikingly, the reverse model, in which a genetic variant primarily regulates gene expression, which in turn regulates DNA methylation, was the best causal model for a number of genes (including *C21ORF56*, *HRASLS3*, *TACSTD2*, *WBSCR27*, *SRXN1*, *GSTM3*, *BTN3A2*), although the model p-values of these LEO scores were small, indicating poor fit. For example, one of these genes, *C21ORF56*, was highlighted in a previous genome-wide study where a three-way association for this gene was identified. Additional experiments indicated that genetic variation in this gene affects chromatin structure in this region[5]. The gene itself may be involved in inter-individual differences in response to DNA damaging agents[33]. These mechanisms and our data suggest that loci whereby genetic variation influences expression and in turn methylation may exist and warrants further study. The methylation and expression probes that showed a causal direction in the LEO analysis were all present within the same gene. However, we observed that of all the 798 significant *cis* associations, only 155 (19%) involved probes that represent the same gene. This may suggest that the strongest (detectable) causal correlations between DNA methylation and gene expression are likely to be local events.

The systems level analysis afforded by WGCNA reveals that both transcriptome and methylome can usefully be organized into modules. Many co-methylation and co-expression

modules are highly significantly enriched with gene ontology categories, which provides indirect evidence that these modules are biologically meaningful. Our module preservation analysis between expression and methylation data reveals that most co-expression modules are comprised of genes that do not form a module in the methylation data and vice versa. Only the largest co-expression module shows moderate to strong preservation and overlap with the largest co-methylation module. In other words, co-expression modules and co-methylation modules are largely composed of different genes. On the other hand, several pairs of expression and methylation eigengenes show highly significant positive and negative correlations. This suggests the existence of factors that affect expression and methylation of different sets of genes, i.e., *trans* effects at the module level.

A limiting factor of our study may be the fact that the Illumina 27k array covers only a selection of CpG sites and is enriched for promoter regions and CpG islands near genes. Another increasingly important issue is the potential difference between hydroxymethylation and DNA methylation that cannot be distinguished with current methylation arrays[34,35]. To date, the role of 5-hydroxymethylation is not fully understood but it is likely that 5-hydroxymethylation plays a role in demethylation[34-38]. Although there is no reason to assume a systematic influence of 5-hydroxymethylation on our results, we cannot rule this out and further refinement of methylation levels is warranted. A third possible limitation is the use of whole blood comprised of different cell types for our analysis. Yet, although whole blood does not provide the optimal resolution, these cell types can be used to study general genetic mechanisms. Given the sample size we suspect that effects of blood cell composition are limited and do not play a major role in the outcome. We measured gene expression and DNA methylation from the same blood sample so that the composition of different cell types should not substantially affect the overall outcome and conclusions. Moreover, studies have shown that a majority of the strongest eQTLs overlaps between different tissues and cell types[6,39].

Conclusions

Overall, this study contributes to our understanding about the relationship between genetic markers, methylation and expression levels in whole blood of healthy subjects. We observed *cis*-associations between methylation and expression levels to be both positive and negative, and most likely to be located outside CGIs and shores. Overrepresentation in shores, as previously found, was only present when selecting methylation/expression combinations regulated by genetic variation in *cis*. Methylation/expression combinations in *trans* are enriched for positive correlations and also located mostly outside CGI's and shores. Results from causality analyses indicate that the conventional model of genetic variants regulating methylation, which in turn regulates gene expression, is most common. This is widely

supported in literature[32]. In addition, this indicates that the causal direction analysis is a useful tool for investigating relationships between genotype, methylation and expression. Finally, we showed that methylome and transcriptome are organized into modules. Although the co-expression en co-methylation modules are generally not preserved in one another, we do find highly significant correlations between the modules. These findings suggest that there may be other (*trans*) factors affecting both methylation and expression, although in different modules. This study encompasses lookup tables for associations between methylation, gene expression, and genotype, as well as methylome and transcriptome modules, for further research.

Methods

Ethics statement

All participants gave written informed consent. This study was approved by Medical Research Ethics Committee (MREC) of the University Medical Center Utrecht, The Netherlands.

Pre-processing of genotype, methylation and expression data

Genotype, methylation and expression data were collected for different numbers of samples. For the 148 healthy subjects eventually analyzed in this paper, data was available for all three layers of genetic information after quality control, as described below. Our final data set consisted of 72 males and 76 females with a mean age of 52 (range: 19–88); all subjects were of Dutch ancestry with at least three of the four grandparents born in The Netherlands.

Genotype SNP data

Genotype data for subjects was generated on two different array platforms, 105 individuals on Illumina CytoSNP (299,173 SNPs) and 96 on Illumina 300k chips (300,299 SNPs). For each SNP platform, quality control procedures were initially performed separately using PLINK[24]. Subjects were excluded based on >5% missing genotypes and gender errors (Additional file 11). We used linkage disequilibrium (LD) based SNP pruning to select the most informative SNPs ($R^2<0.2$), only for subsequent quality control steps. This resulted in ~60k SNPs for both sets to assess heterozygosity (F<3 Standard Deviation (SD)), homozygosity (F>3SD) and relatedness by pairwise identity by descent (IBD) values (pihat >0.1). Datasets were merged with Hapmap Phase 3 individuals to check ethnicity (Additional file 12) (ethnic outliers detected by visual inspection). After these QC procedures on subjects (excluding in total 8 individuals) quality control on SNPs was performed as follows. All SNPs were filtered on missingness (>2%) and Hardy Weinberg (p>1e-6) before merging the two datasets. 84,367 SNPs were shared between the two datasets. No related samples were detected in the merged datasets (according to criteria described above). We imputed the merged

dataset with Hapmap2, release 24 using Beagle[40]. SNPs with an imputation score > 0.8 and present originally in one or both datasets were extracted and 417,708 SNPs remained for all further analyses.

DNA methylation data

Methylation data was obtained using Illumina HumanMethylation27 beadchips for two batches of 105 and 96 healthy subjects. The assay detects methylation status at CpG sites after bisulfate conversion, by means of probes designed for either methylated or unmethylated sequence. Methylation probes were classified into 3 different categories depending on the location of the probe with respect to a CpG island. Based on the UCSC Table browser (http://genome.ucsc.edu/;[41]), NCBIbuild36, categories were defined as CpG island, CpG island shore (sequences up to 2kb from an island), or outside CpG islands/shores. Ethnical outliers and samples with gender errors in genotype data were removed from the methylation data. Gender was checked by hierarchical clustering of X-chromosomal probes, excluding four individuals. Another three individuals were removed based on detection *p*-values (>0.01 for >1% of probes) and 3,027 of 27,578 probes were excluded based on detection values (p>0.01 for >1% of the samples). Both channels of the methylation array were quantile normalized independently. Beta values of a probe were calculated by dividing the methylated signal by the sum of the methylated and unmethylated signal. Next, five potential array outliers were removed in an unbiased fashion. Specifically, we used the SampleNetwork R function package[42] to calculate the Interarray based sample connectivity score Z.k. We removed samples with a Z.k value less than −3 since their connectivity is 3 standard deviations below the mean value. Batch effects of dataset, plate, array and position were removed using ComBat[43]. After these procedures, 24,561 probes remained and were mapped to the human genome using the UCSC Human BLAT Search function. In total, 25 probes did not map to the human genome, whereas 338 probes did not map uniquely (mapped more than once), and both these probes have been removed. Moreover, 904 probes that contained a SNP, based on Hapmap release 27, with a minor allele frequency (MAF) > 1% have been removed as well, leaving a total of 23,294 probes for analyses.

Gene expression data

Gene expression data was generated in two batches, one on Illumina H8 beadchip (26 healthy subjects) and one on Illumina H12 beadchip (147 healthy subjects). BeadStudio© software version 3.2.3 was used to generate background-corrected gene expression data. Data was normalized, transformed and filtered separately before merging and batch effect removal. Specifically, the datasets were separately quantile normalized and log2 transformed using the Lumi package for R[44]. Probes were filtered based on detection values generated by BeadStudio©. The detection *p*-value threshold was set at 0.01. This resulted in 17,433

expression probes overlapping between both batches. Batch effects resulting from the use of different arrays at different time points were removed using ComBat[43]. An unbiased analysis based on interarray correlations identified 16 samples from batch 2 as potential outliers, which were subsequently removed from the analysis. Of 17,433 probes, 15,983 mapped to a single genomic location, based on a previous study[45]. In addition, 465 probes contained a SNP, based on Hapmap release 27, with a MAF $>$ 1% and have been removed, leaving 15,983 probes for analyses.
DNA methylation and gene expression data have been processed using the same blood sample, excluding possible batch effects, such as the effect of different time points.

Identifying *cis* and *trans* effects between DNA methylation and gene expression

We called a methylation probe *cis* acting with respect to a given gene expression probe if there was a significant association (as defined below) within a 500kb interval between the probes. A methylation probe was called *trans* acting if it was significantly associated with the expression probe (as defined below) outside the 500kb interval.
To determine whether a significant association exists between expression and methylation levels we used a multivariate linear regression model for regressing the gene expression level (dependent variable) on the methylation level (independent variable) with age and gender as covariates. We took methylation levels as independent variable since we are interested in the epigenetic control of gene expression levels. Associations can be positive (DNA methylation levels and gene expression levels both increase or decrease) or negative (increased methylation level corresponds with a decrease in gene expression level and vice versa). The Wald test *p*-value for the association between methylation and expression was used as significance level. Correction of the significance level for multiple testing was performed separately for identifying *cis* acting methylation probes (FDR correction) and *trans* acting methylation probes (Bonferroni correction).

Identification of *cis*-and *trans*-acting SNPs

Expression levels and methylation levels that were significantly associated with each other were tested for association with SNPs to identify *cis*-and *trans*-acting genetic variations. For this analysis, the real and imputed (imputation score $>$ 0.8) genotypes were used, and a MAF threshold of 5% for these SNPs was set.
Analogous to our previous definition, a SNP significantly associated with a given gene expression or DNA methylation probe was called *cis*-acting with respect to the probe if the SNP and the probe were within 500kb of each other, and *trans*-acting if they were more than 500kb apart.
To determine whether a significant relationship exists between a SNP and a methylation or expression level we again used a multivariate linear regression model for regressing the

methylation or expression level (dependent variable) on the SNP (independent variable) with age and gender as covariates. The regressions were performed using the PLINK software[24]. Correction for multiple testing was performed separately for *cis*-acting SNPs (0.05 divided by the number of probes) and *trans*-acting SNPs (0.05 divided by the number of possible combinations ($p<0.05/(\#probes*417,708)$).

Evaluating causal relationships using local edge orienting scores of observed *cis* effects

To evaluate the fit of different causal models involving 3 variables (i.e., a *cis*-acting SNP, a *cis*-acting methylation probe, and a corresponding expression probe), we calculated the single marker local edge orienting score (LEO.NB.SingleMarker) as described elsewhere[19,32]. In short, a SNP can be used as causal anchor for evaluating the causal relationships between methylation and expression levels if the SNP is associated with at least one of them. We use the SNP as causal anchor for calculating the LEO score since genotypes are fixed at each locus as opposed to variable methylation and expression levels[19]. In this case, one can evaluate the fit of the following five models describing the causal relationships between a SNP (denoted S), a methylation probe (M) and an expression probe (E): model 1: S→M→E; model 2: S→E→M; model 3: M←S→E; model 4: S→E←M; model 5: S→M←E. For each causal model a chi-square test based model fitting p-value was calculated with the structural equation modelling (SEM) R package[46]. The relative fit of causal model 1 (SNP→Methylation→Expression) was assessed using the single anchor local edge orienting score (LEO.SingleMarker), which is the logarithm (base 10) of the ratio of the model fitting *p*-value divided by that of the next best fitting alternative model[19]. Thus a positive LEO. SingleMarker score indicates that the causal model S→M→E fits the data better than all other competing models. As significance threshold we used the LEO threshold of 0.8, as recommended in[19] based on extensive simulations as well as empirical studies. We decided to focus on local *cis* effects since there is considerable debate in the literature whether *trans* relationships are reproducible[29,30]. Since we were interested in causal direction for predetermined three-way associations, we only selected SNPs associated with both the methylation and expression levels in *cis*. To protect the causal analysis from biases due to age and gender, we utilized residuals of methylation and expression levels corrected for age and gender in the causal analysis using a linear regression by Limma in R[47].

Weighted correlation network analysis of gene expression and methylation data

A detailed description of our correlation module based analyses can be found in Additional file 5. Here we provide a terse summary. Weighted correlation network analysis implemented in the WGCNA R package[26,27] was first applied to the expression data to identify co-expression modules. Co-expresssion modules correspond to clusters of interconnected genes defined as

branches of a hierarchical cluster tree. Since modules are defined without respect to gene ontology information they are initially labelled by arbitrary integers and coded by colours. Next WGCNA was applied to the methylation data to find co-methylation modules. For easier interpretation of the relationships between expression and methylation modules, we use the same module labels for modules that show significant overlap. The matching of module labels was performed using the function matchLabels from the WGCNA R package; it is based on significance of module overlaps quantified using Fisher's exact test. Weighted networks have the advantage of preserving the continuous nature of co-expression and co-methylation information, which is particularly useful when studying module preservation. To assess the preservation of expression and methylation modules in the corresponding complementary data set, we use the network module preservation statistics described in[28] and implemented in the function modulePreservation in the WGCNA R package. Network module preservation statistics assess whether the density and connectivity patterns of modules defined in a reference data set are preserved in a test data set. Network preservation statistics do not require that modules be identified in the test data set and hence independent of the ambiguities associated with module identification in the test data set. The permutation test of the modulePreservation function leads to a composite module preservation statistic referred to as Zsummary. The Zsummary statistic of a given module summarizes the evidence that the network connections of the module are more significantly preserved than those of random set of genes of equal size. We adopted the following recommended significance thresholds for Zsummary[26-28]: Zsummary<2 implies no evidence that the module is preserved, 2<Zsummary<10 implies weak to moderate evidence, and Zsummary>10 implies strong evidence for module preservation. Thus, we report Zsummary for each expression and methylation module in the methylation and expression test data sets, respectively.
Since modules group together highly correlated variables, it is advantageous to summarize the variable profiles using a single representative. We use the module eigengene E, defined as the first principal component of the standardized matrix containing the variables in the module. The module eigengene can be intuitively understood as a weighted average of the variable profiles in the module.

Abbreviations

CpG, Cytosine-Phosphate-Guanine; CGI, CpG Island; QTL, Quantitative Trait Loci; LEO, Local Edge Orienting; FDR, False Discovery Rate; SNP, Single Nucleotide Polymorphism; LD, Linkage Disequilibrium; MHC, Major Histocompatibility Complex; WGCNA, Weighted Gene Co-expression Network Analysis; GO, Gene Ontology; MREC, Medical Research Ethics

Committee; SD, Standard Deviation; IBD, Identity By Descent; SEM, Structural Equation Modeling

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

KRE, SJ and PL wrote the paper. KRE, SJ, TL and PL analyzed the data. MPMB, SH and RAO designed the study. TL, FC and PL wrote scripts for the analyses. RAO, MPMB, SH, JHV, LHB and RSK provided data, materials, and analysis tools. EJ and ES processed the data. CGFK advised on statistics. The manuscript has been seen and approved by all listed authors.

Acknowledgements

The authors thank Bobby Koeleman for advice on statistics and Flip Mulder and Yafeng Zhang for bioinformatics support. We thank SARA Computing and Networking Services (www.sara.nl) for their support in using the Lisa Compute Cluster, which facilitates the Genetic Cluster Computer.

References

1. Bird A: Perceptions of epigenetics. *Nature* 2007, **447**(7143):396–398.
2. Irizarry RA, *et al*: The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 2009, **41**(2):178–186.
3. Irizarry RA, Wu H, Feinberg AP: A species-generalized probabilistic model-based definition of CpG islands. *Mammalian genome: official journal of the International Mammalian Genome Society* 2009, **20**(9–10):674–680.
4. Sengupta N, Seto E: Regulation of histone deacetylase activities. *J Cell Biochem* 2004, **93**(1):57–67.
5. Bell JT, *et al*: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 2011, **12**(1):R10.
6. Gibbs JR, *et al*: Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 2010, **6**(5):e1000952.
7. Zhang D, *et al*: Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet* 2010, **86**(3):411–419.
8. Monks SA, *et al*: Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* 2004, **75**(6):1094–1105.
9. Boks MP, *et al*: The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One* 2009, **4**(8):e6767.
10. Cookson W, *et al*: Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009, **10**(3):184–194.
11. Stranger BE, Dermitzakis ET: The genetics of regulatory variation in the human genome. *Hum Genomics* 2005, **2**(2):126–131.
12. Stranger BE, *et al*: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007, **315**(5813):848–853.
13. Nica AC, *et al*: The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 2011, **7**(2):e1002003.
14. Stranger BE, *et al*: Population genomics of human gene expression. *Nat Genet* 2007, **39**(10):1217–1224.
15. Ghosh S, *et al*: Tissue specific DNA methylation of CpG islands in normal human adult somatic tissues distinguishes neural from non-neural tissues. *Epigenetics: official journal of the DNA Methylation Society* 2010, **5**(6):527–538.
16. Liang P, *et al*: Genome-wide survey reveals dynamic widespread tissue-specific changes in DNA methylation during development. *BMC Genomics* 2011, **12**(1):231.
17. Huang GJ, *et al*: High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res* 2009, **19**(6):1133–1140.

18. Petretto E, *et al*: Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet* 2006, **2**(10):e172.
19. Aten JE, *et al*: Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst Biol* 2008, **2**:34.
20. Chen Y, *et al*: Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008, **452**(7186):429–435.
21. Inouye M, *et al*: An immune response network associated with blood lipid levels. *PLoS Genet* 2010, **6**((9):e1001113.
22. Plaisier CL, *et al*: A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet* 2009, **5**(9):e1000642.
23. Presson AP, *et al*: Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst Biol* 2008, **2**:95.
24. Purcell S, *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, **81**(3):559–575.
25. Johnson AD, *et al*: SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008, **24**(24):2938–2939.
26. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008, **9**:559.
27. Zhang B, Horvath S: A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
28. Langfelder P, *et al*: Is my network module preserved and reproducible? *PLoS Comput Biol* 2011, **7**(1):e1001057.
29. Peirce JL, *et al*: How replicable are mRNA expression QTL? *Mammalian genome: official journal of the International Mammalian Genome Society* 2006, **17**(6):643–656.
30. van Nas A, *et al*: Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics* 2010, **185**(3):1059–1068.
31. Maunakea AK, *et al*: Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 2010, **466**(7303):253–257.
32. Horvath S: Structural Equation Models and Directed Networks. In *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Edited by Horvath S. New York: Springer Book; 2011:279–320.
33. Fry RC, *et al*: Genomic predictors of interindividual differences in response to DNA damaging agents. *Genes Dev* 2008, **22**(19):2621–2626.
34. Jin SG, Kadam S, Pfeifer GP: Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* 2010, **38**(11):e125.
35. Dahl C, Gronbaek K, Guldberg P: Advances in DNA methylation: 5-hydroxymethylcytosine revisited. *Clinica chimica acta; international journal of clinical chemistry* 2011, **412**(11–12):831–836.

36. Guo JU, *et al*: Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. *Cell Cycle* 2011, **10**(16):2662–2668.

37. Tan L, Shi YG: Tet family proteins and 5-hydroxymethylcytosine in development and disease. *Development* 2012, **139**(11):1895–1902.

38. Branco MR, Ficz G, Reik W: Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet* 2012, **13**(1):7–13.

39. Ding J, *et al*: Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am J Hum Genet* 2010, **87**(6):779–789.

40. Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009, **84**(2):210–223.

41. Karolchik D, *et al*: The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004, **32**(Database issue):D493–D496.

42. Oldham MC, Langfelder P, Horvath S: Sample Networks for Enhancing Cluster Analysis of Genomic Data: Application to Huntington's Disease. *BMC Syst Biol* 2011, **6**:63.

43. Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, **8**(1):118–127.

44. Du P, Kibbe WA, Lin SM: lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008, **24**(13):1547–1548.

45. Fehrmann RS, *et al*: Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet* 2011, **7**(8):e1002197.

46. Fox J: Structural equation modeling with the sem package in R. *Structural Equation Modeling-a Multidisciplinary Journal* 2006, **13**(3):465–486.

47. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.

Additional data files

The following additional data are available with the online version of this paper.

Additional_file_1 as PDF

Additional file 1: Table S1 Comprises two tables that list all significant methylation and expression associations in *cis* (S1a), and *trans* (S1b).

Additional_file_2 as PDF

Additional file 2: Figure S1 Are two figures that show the coefficient and explained variance of associations between methylation and expression.

Additional_file_3 as XLS

Additional file 3: Table S2 Contains tables with all significant *cis* mQTLs (S2a) and eQTLs (S2b).

Additional_file_4 as XLS

Additional file 4: Table S3 Is a table with all LEO results. Combinations that have a LEO score above 0.8 for the model S>M>E are shown in light yellow of which LEO scores above 3 are shown in dark yellow. For the reverse model (S>E>M) combinations with a LEO score above 0.8 are shown in orange. Significant p-values (above 0.01) are coloured in green.

Additional_file_5 as PDF

Additional file 5 Contains supplementary methods, namely, a more detailed description of Weighted Correlation Network Analysis (WGCNA).

Additional_file_6 as PDF

Additional file 6 Is an overview of the modules identified in the expression (Table 1) and methylation (Table 2) data.

Additional_file_7 as CSV

Additional file 7 Includes a table of continuous module membership kME_i of all expression profiles in all expression modules. Each row in the table corresponds to one gene expression profile. Columns give the gene Entrez idenitifier, module label, and *kME* and the corresponding (uncorrected) p-values for each module. Expression modules are labelled by E.0, E.1, etc.

Additional_file_8 as CSV

Additional file 8 Includes a table of continuous module membership kME_i of all methylation profiles in all methylation modules. Each row in the table corresponds to one methylation profile. Columns give the gene Entrez idenitifier, module label, and *kME* and the corresponding (uncorrected) p-values for each module. Methylation modules are analogously labeled by M.0, M.1, etc.

Additional_file_9 as PDF

Additional file 9 Shows the overlap of expression and methylation modules. Each row corresponds to an expression module (labelled by the numeric labels, colours and total number of genes in the module, on the left), and each column corresponds to a methylation module (labelled the numeric labels, colours, and total number of genes in the module, at the bottom). Numbers in the table indicate number of genes in the overlap, and the Fisher exact test p-value for the overlap. Only overlaps whose p-value is below 0.05 are shown. The table is coloured such that significant overlaps are coloured in strong red colour. Most overlaps are quite small but some are nevertheless statistically highly significant.

Additional_file_10 as PDF

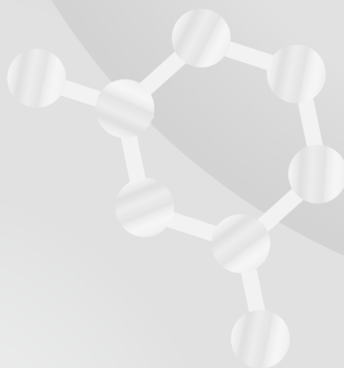
Additional file 10 Shows the *medianRank* statistics for the Module preservation with in (A) preservation of expression modules in methylation data, and in (B) preservation of methylation modules in expression data.

Additional_file_11 as XLS

Additional file 11 Is a table with the number of excluded samples per step.

Additional_file_12 as PDF

Additional file 12 Is a clusterplot of all samples together with Hapmap phase 3 populations.



Chapter 3

Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes

S de Jong, KR van Eijk, DWLH Zeegers, E Strengman, E Janson, JH Veldink, LH van den Berg, Schizophrenia PGC, W Cahn, RS Kahn, MPM Boks, RA Ophoff

European Journal of Human Genetics*:* 2012 Sep;20(9):1004-8. doi: 10.1038/ejhg.2012.38. Epub 2012 Mar 21.

Abstract

There is genetic evidence that schizophrenia is a polygenic disorder with a large number of loci of small effect on disease susceptibility. Genome-wide association studies (GWASs) of schizophrenia have had limited success, with the best finding at the MHC locus at chromosome 6p. A recent effort of the Psychiatric GWAS consortium (PGC) yielded 5 novel loci for schizophrenia. In this study we aim to highlight additional schizophrenia susceptibility loci from the PGC study by combining the top association findings from the discovery stage (9,394 schizophrenia cases and 12,462 controls) with expression QTLs (eQTLs) and differential gene expression in whole blood of schizophrenia patients and controls. We examined the 6,192 single nucleotide polymorphisms (SNPs) with significance threshold at $p<0.001$. Expression QTLs were calculated for these SNPs in a sample of healthy controls (n=437). The transcripts significantly regulated by the top SNPs from the GWAS meta-analysis were subsequently tested for differential expression in an independent set of schizophrenia cases and controls (n=202). After correction for multiple testing, the expression QTL analysis yielded 40 significant *cis*-acting effects of the SNPs. Seven of these transcripts show differential expression between cases and controls. Of these, the effect of three genes (*RNF5*, *TRIM26* and *HLA-DRB3*) coincided with the direction expected from meta-analysis findings and were all located within the MHC region. Our results identify new genes of interest and highlight again the involvement of the MHC region in schizophrenia susceptibility.

Introduction

Schizophrenia is a severe mental disorder, affecting about 1% of the population worldwide. Heritability is estimated to be around 80%, but the underlying genes are largely unknown[1]. Large-scale, genome-wide studies have identified rare genomic microdeletions as well as common variants associated with the disease[2-6].

In a recent study, Purcell and colleagues demonstrated that data from genome-wide association studies (GWAS) for schizophrenia are compatible with a very large number of loci with common alleles (N>3,000), each with a very small contribution to disease susceptibility (odds ratios <1.05)[1]. Alternative approaches may be necessary to decipher the genetic basis of schizophrenia and related disorders[7]. We therefore aim to combine different layers of genomic information to uncover genetic signal from common variants that would not be identified by current GWAS approaches.

A recent meta-analysis comparing 9,394 cases to 12,462 controls resulted in identification of numerous common variants[8] with sub threshold association with schizophrenia (6,192 SNPs with $p<0.001$). However, as Purcell et al. showed, variants associated with the disease may not reach genome-wide significance[1]. Therefore, it is likely that there are more true positives in the top 6,192 SNPs than were identified by performing a standard case-control association analysis. Subtle effects of these SNPs on gene expression could be a functional mechanism by which they confer risk for development of schizophrenia[9,10]. Recently, it has been shown that true GWAS hits are enriched for expression QTLs[10-13]. Therefore, variations influencing gene expression are more likely to be contributing to the phenotype. To this end, we generated eQTLs for the top 6,192 SNPs ($p<0.001$ in meta-analysis). Next, we tested whether the identified transcripts are differentially expressed between patients and healthy controls. These analyses have the potential to provide further support of the involvement of these SNPs in schizophrenia and may highlight additional schizophrenia candidate genes that have not been identified using genome-wide significance thresholds.

Although gene expression in whole blood is only moderately correlated with gene expression in brain tissue[1,14-16] several studies suggest that gene expression in blood could serve as a marker of brain-related disease states, including schizophrenia[15-21]. Therefore, gene expression profiling in blood may provide additional insight into the etiology of the disease. We performed the gene expression analyses using whole blood samples of relatively large sample of schizophrenia patients and controls.

Materials and methods

Expression QTL analysis in controls

We calculated the expression QTLs for the top SNPs[8] in a sample of 437 healthy controls for which genotypes (Illumina 370k array) and whole blood gene expression data (Illumina H-12 beadchip) was available as described before[10,22]. In short, this data set consists of 244 males and 193 females with a mean age of 62 years, who where recruited as controls in a study of gene expression in ALS (Amyotrophic Lateral Sclerosis). These control subjects were selected for being in good general health and unaffected for neurological and neurodegenerative diseases; no separate screen for psychiatric disorders was performed for these subjects.

Of the 6,192 SNPs, 1,336 were already available on the array. Imputation was performed by BEAGLE version 3.0.4[23] using the Hapmap phased founder set, release 2, phase 3 (The International HapMap Project, http://www.hapmap.org) A R^2 cut-off of 0.90 resulted in 4,073 successfully imputed SNPs yielding a total of 5,409 SNPs for analysis. The gene expression data of these controls were quantile normalized and log2 transformed using the PreprocessCore package in R[24]. Expression probes were then filtered for detection value smaller than 0.90 as by manufacturer protocol, leaving 12,990 high quality probes for analysis.

The 12,990 expression probes were taken as quantitative traits and tested for association with the 5,409 available SNPs using a linear association of allele dosage with age and gender as covariates in PLINK[25]. To adjust for significant differences in mean age of this control sample and the schizophrenia sample described below, we included age (and gender) as covariates. Since *trans*-effects are difficult to identify in a study of this size due to limited power, we focused on *cis*-effects only, i.e. 1MB around the probe center position on either side. We used Bonferroni correction for multiple testing, setting the significance thresholds for *cis*-effects 0.05/5,409 = 9.24E-6.

Differential expression schizophrenia versus controls

We examined whether the probes associated with the top SNPs are also related to schizophrenia disease status. This set consists of 106 schizophrenia cases and 96 healthy controls including 118 male and 84 female subjects, with an average age of 39 years. Diagnoses were determined by Standardized Psychiatric interviews either The Comprehensive Assessment of Symptoms and History (CASH) or the Composite international diagnostic interview (CIDI) by trained clinicians. Schizophrenia was defined by a DSM-IV-TR diagnosis of #295.0-295.89, and #298.9. . This study was approved by Medical Research Ethics Committee (METC) of the University Medical Center Utrecht, The Netherlands. The data

was normalized (robust spline normalization, RSN), transformed (variance stabilizing transformation, VST) and filtered according to the Lumi procedure as described previously[26].

We used the Limma package[27] in R to generate a regression model with selected expression values as dependent and status as independent values. We included age and gender as covariates. We took FDR corrected *p*-value of 0.05 as significance threshold.

Results

Expression QTL analysis in controls

We identified 40 unique transcripts from MHC and non-MHC regions regulated by the top 5,409 SNPs. The distribution of expression QTL results within the *cis*-region is displayed in **Figure 1**. Since we found the signal from the MHC region in the eQTL results to be substantial, these results will be discussed separately. In total, 1,664 significant *cis*-effects on 23 unique expression probes from 578 unique SNPs in the MHC region were found. Outside the MHC, 166 *cis*-effects, were identified, representing 249 unique SNPs and 17 unique expression probes. In this instance we define unique to indicate different SNPs, without considering linkage disequilibrium (LD) between them. The eQTL effects are plotted against the rank of the SNPs in the top list in **Figure 2**. This shows the effects of the MHC region to be stronger (*p*-values ranging from 4.6E-145 to 9.1E-6) than the rest of the genome (*p*-values

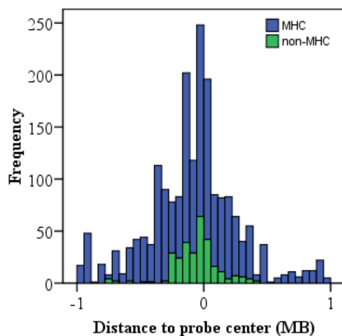


Figure 1: Distribution of eQTL effects. A region of 1MB around the center of the expression probe was taken as a threshold for *cis*-effects. The plot shows a histogram of frequency of eQTL effects (SNP-probe combinations) in this region. Effects in the MHC region are displayed in blue, effects elsewhere on the genome in green.

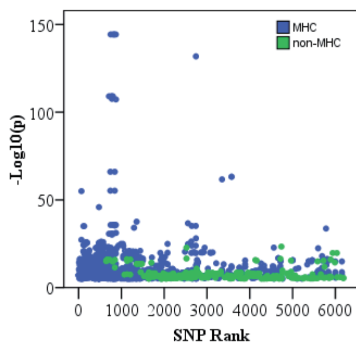


Figure 2: eQTL effects versus SNP rank. The –log10(p) value of eQTL effects (SNP-probe combinations) are plotted against the original rank of the SNP in the schizophrenia meta-analysis top 6,192. Effects in the MH C region are displayed in blue, effects elsewhere on the genome in green.

ranging from 4.3E-24 to 9.1E-6). In addition, the highest ranking genes are dominated by effects from this region. All significant eQTL effects are given in Supplementary **Table 2**.

To formally test whether the MHC is overrepresented in the eQTL results, we performed a Fishers' exact test for count data. Of the 5,409 SNPs tested, 1,071 are located within or are associated with a probe in the extended MHC region (chromosome 6, 26MB – 34MB). Even though that already represents a large part of the input, the overrepresentation in the eQTL results (unique SNPs influencing one or more expression probes) is significant when compared to SNPs located elsewhere on the genome (Fishers' $p<2.2E-16$). The fact that this region is known to have complicated and extended LD structures[28] may in part explain the fact that more SNPs are related to the same gene expression probes.

Differential expression schizophrenia versus controls

The levels of 40 unique transcripts from MHC and non-MHC regions that we identified were tested for association with schizophrenia in an independent gene expression dataset of schizophrenia cases and controls. Seven of the transcripts were differentially expressed in cases. **Table 1** presents the results together with the corresponding eQTL results. Five of the seven transcripts are from genes located in the MHC (*TRIM26*, *RNF5*, *TUBB*, *HLA-DRB3*, *HIST1H2BD*). Two differentially expressed probes are located elsewhere on the genome, *C16ORF61* on chromosome 16 and *CRELD2* on chromosome 22. All but two probes are down regulated in schizophrenia patients compared to healthy controls (*C16ORF61* and *HIST1H2BD*). Although 18 of the 106 schizophrenia patients in this dataset are antipsychotic-free (e.g. have not been treated with antipsychotic drugs during the six-month-period prior to blood sampling), the effects of medication on gene expression in blood in the overall sample might have contributed to our results.

Next, we examined whether the direction of differential expression coincided with the original meta-analysis results. When the risk allele is associated with up or down regulation of expression, this should be mirrored in the differential expression results. We therefore assume that the eQTL itself is not different between cases and controls, but the frequency of the allele associated with expression levels is. This was the case for three out of the seven genes; *TRIM26*, *RNF5* and *HLA-DRB3*, all of which are located within the MHC region. *TRIM26* is one of the genes identified as susceptibility locus in the original meta-analysis after the replication stage[8].

For our top genes (*TRIM26*, *RNF5* and *HLA-DRB3*) we examined whether the same eQTLs are also present in brain, using a publically available human brain expression dataset of 144 individuals (Gibbs et al.[40] GEO; GSE15745). Data are available for four different brain

Table 1: eQTL and differential expression results for potential schizophrenia candidate genes. Of the 40 probes identified by eQTL analysis to be regulated by the top 5,409 SNPs, seven were also found to be differentially expressed between schizophrenia cases and controls, of these the direction of *TRIM26*, *RNF5* and *HLA-DRB3* coincided with meta-analysis results. The first three columns contain Gene Symbol, Illumina Array Address ID of the expression probes and location of the genes. The next part of the table contains details of the eQTL analysis, with the top SNP, its rank in the original meta-analysis, distance to associated expression probe center and eQTL *p*-value. The number of SNPs associated with the expression level and rank range in the meta-analysis is given respectively. Finally, the direction of expression associated with the risk allele of the top SNP is given. The last part of the table contains the differential expression results, with average expression, the log fold change (LFC) between cases and controls, the *t*-statistic and the corresponding *p*-value (adjusted for testing of 40 probes).

			Expression QTL								Differential Expression			
Gene Symbol	**Array Address ID**	**Location**	SNP name	Rank top 6000	SNP-probe center (bp)	*p*-value	# SNPs	Rank range	Risk Allele	Direction of xp	AVG xp	Log FC	t	Adjusted *p*-value
TRIM26	**2370138**	**6p21.3**	**rs2844766**	**445**	**78138**	**4.5E-06**	**4**	**445 - 617**	**T**	↓	**8.86**	**-0.24**	**-6.84**	**4.0E-09**
C16ORF61	6400025	16q23.2	rs4888116	2642	51339	5.5E-06	1	2642	C	↓	9.10	0.24	5.08	1.8E-05
RNF5	**2690603**	**6p21.3**	**rs3132947**	**3215**	**28549**	**3.2E-08**	**22**	**671 - 5578**	**G**	↓	**8.40**	**-0.11**	**-4.78**	**4.8E-05**
CRELD2	6560390	22q13.33	rs138897	2183	89937	2.4E-08	26	1945 - 6118	C	↑	9.21	-0.13	-3.75	2.4E-03
TUBB	6580474	6p21.33	rs3094127	895	4464	3.4E-11	60	22 - 1120	A	↑	10.84	-0.12	-3.11	1.8E-02
HIST1H2BD	6200669	6p21.3	rs9379827	1949	5056	2.9E-18	5	1713 - 2018	A	↓	8.55	0.16	2.82	3.2E-02
HLA-DRB3	**1770504**	**6p21.3**	**rs9268858**	**748**	**25356**	**2.4E-14**	**33**	**66 - 5596**	**C**	↓	**9.77**	**-0.25**	**-2.82**	**3.2E-02**

regions: cerebellum (CRBLM), frontal cortex (FCTX), temporal cortex (TCTX) and pons. eQTL analysis was performed as described in blood, with the addition of covariates post-mortem interval, batch, and source of collection. *RNF5* is not expressed at detectable levels in any of these four regions, *HLA-DRB3* is expressed in frontal cortex and cerebellum, and *TRIM26* is expressed in all four brain regions. The top SNP regulating *HLA-DRB3* (rs9268858) was not available for the brain dataset. However, rs2395185 is in strong LD ($R^2 = 0.98$, based on HapMap3) and shows significant eQTL results for *HLA-DRB3* in frontal cortex (p=9.0E-4, beta =-0.19, t=-3.40) and cerebellum (p=3.0E-3, beta=-0.17, t=-3.08). Similarly, the top SNP regulating *TRIM26* (rs2844766) in blood was not available for the brain samples but rs1264616, with moderate LD ($R^2 = 0.45$, based on HapMap3) shows evidence of an eQTL in cerebellum (p=9.0E-3, beta=0.09, t=2.64) and pons (p-=1.0E-2, beta=0.03, t=0.91). These eQTLs show the same directionality in both blood and brain.

Discussion

Our results are noteworthy in two respects: First, we show that by combining different layers of genetic information we can successfully identify potential candidate genes for schizophrenia. By using eQTL analysis we have identified SNPs that actually have an effect on expression of transcripts that differ between patients and controls. Second, we found that the top list that we took as a starting point harbored significantly more eQTLs in the MHC region compared to elsewhere in the genome. Strong eQTLs in this region have been observed in a previous study in lymphoblastoid cell lines, derived from peripheral blood lymphocytes[29]. Specifically, *cis* eQTLS have been found for *TUBB*[30] and *HLA-DRB3*[31,32], while *RNF5* has been found to be one of the most variable genes between individuals[33]. The fact that eQTLs in the MHC region are overrepresented in our study could be due to the type of tissue studied. Blood has an important function in the (primary) immune response and eQTLs can be tissue-specific[34]. However, it is also possible that genetic control of this region is stronger than elsewhere on the genome, perhaps related to the extended LD patterns.

The MHC harbors relatively many genes and is highly polymorphic. It regulates the immune response and has been associated with a large number of immune phenotypes and diseases[35]. The association of SNPs in these MHC genes can result from mere physical closeness to the schizophrenia variant within the MHC region. However, that would not explain the difference in gene expression of MHC genes presented here, since expression transcripts are not in LD. We find that the eQTL SNPs are often related to multiple probes of different genes within the *cis*-region (up to 6 per SNP). This could point to important regulatory regions in the MHC. We therefore propose that eQTL analysis could aid in the refinement of the MHC region.

Although association of schizophrenia with this area has been consistent, the functional explanation is still ambiguous. Inflammation is suggested to be an inherent part of schizophrenia. Reports show that schizophrenia patients have either higher or lower prevalence of some autoimmune disorders than expected[36] and that the efficacy of anti-inflammatory drugs differs in patients[37]. The latter observations could support the viral hypothesis of schizophrenia[38]. However, of the three identified candidate genes in the MHC region, only one has a known immune function. This is *HLA-DRB3* (major histocompatibility complex, class II, DR beta 3). Class II molecules are expressed in antigen presenting cells. They play a central role in the immune system by presenting peptides derived from extracellular proteins. The function of *TRIM26* (tripartite motif-containing 26) in the classical class I subregion is unknown; however, it is thought to have DNA-binding activity[28]. *RNF5* (Ring finger protein 5) in the classical class III subregion is involved in cell motility. It has been shown to be a regulator of breast cancer progression through its effect on actin cytoskeletal alterations[39]. This suggests that the MHC signals associated with schizophrenia are not necessarily immune-related, but could also stem from genes in the region that are involved in non-immune related pathways.

Since schizophrenia is a brain-related disease, it was useful to replicate our findings in brain tissue. Publically available brain expression data indicated that eQTLs for *TRIM26* and *HLA-DRB3* are also present in specific brain regions. Although these results fit well with a possible involvement of these loci in schizophrenia, the lack of sufficient number of available brain tissue from patients has prevented us to examine brain-specific differential gene expression related to disease.

In summary, we have identified three genes, *TRIM26*, *RNF5* and *HLA-DRB3* that are regulated by the most significant SNPs in a recent meta-analysis and of which expression is associated with schizophrenia disease status. These effects are small but significant, indicating that the current approach allows to detect the small functional effects that may play a role in schizophrenia susceptibility[1]. The results of this study indicate that signal not only stems from the SNPs with the most significant *p*-values in the top 6,192 list, but can be linked to less significant SNPs as well. In addition to confirming one of the top findings in the meta-analysis, *TRIM26*, we also identify *RNF5* and *HLA-DRB3* as potential candidate genes for schizophrenia. The results of this study gives further insight into the relationship of SNPs with gene expression, highlight the importance of the MHC region for schizophrenia susceptibility and indicate that genetic causal variants for schizophrenia might act through regulation of expression. This approach can be fruitful in identifying phenotypic effects of SNPs highlighted by GWAS.

Acknowledgements

We thank the participants and their relatives and staff at the various recruitment centers. This work was supported by TI-pharma project T5-203 and funding from the US National Institutes of Health MH078075 and NS058980 (RAO) and the Amyotrophic Lateral Sclerosis Association (RAO and LHvdB). Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation.

Conflict of interest

The authors declare no conflict of interest.

Supplementary information

Supplementary table 1: Author list Schizophrenia Psychiatric GWAS Consortium.
Type: .pdf

Supplementary table 2: eQTL results. Results of the association of the 5,409 top SNPs with 12,990 expression probes in a control sample (n=437). The SNP name, chromosome, rank in the meta-analysis stage 1 results, basepair location is given, in addition to whether it was imputed or available in the original dataset. The next columns contain A1 (minor allele) and A2 (major allele), minor allele frequency (MAF), Hardy-Weinberg equilibrium *p*-value (HW). Results from eQTL analysis in PLINK are given in columns standard error of odds ratio (SE), regression coefficient (Beta), confidence interval (L95 and U95), t-statistic (Stat) and p-value (P_res). The Illumina array address ID, Illumina gene annotation, probe location (Center_probe) are given in the next columns. Finally, it is denoted whether a probe is in the MHC region (MHC) and found to be differentially expressed in subsequent analyses (Differential expression).
Type: .xls

Supplementary information is available at European Journal of Human Genetics' website.

References

1. Purcell SM, Wray NR, Stone JL *et al*: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460:** 748-752.
2. Shi J, Levinson DF, Duan J *et al*: Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; **460:** 753-757.
3. Stefansson H, Ophoff RA, Steinberg S *et al*: Common variants conferring risk of schizophrenia. *Nature* 2009; **460:** 744-747.
4. Stefansson H, Rujescu D, Cichon S *et al*: Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; **455:** 232-236.
5. Buizer-Voskamp JE, Muntjewerff JW, Strengman E *et al*: Genome-Wide Analysis Shows Increased Frequency of Copy Number Variation Deletions in Dutch Schizophrenia Patients. *Biol Psychiatry* 2011.
6. Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I *et al*: Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet* 2008; **83:** 504-510.
7. Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461:** 747-753.
8. Ripke S, Sanders AR, Kendler KS *et al*: Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011.
9. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M: Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009; **10:** 184-194.
10. Dubois PC, Trynka G, Franke L *et al*: Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010.
11. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010; **6:** e1000888.
12. Richards AL, Jones L, Moskvina V *et al*: Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol Psychiatry* 2011.
13. Zhong H, Beaulaurier J, Lum PY *et al*: Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* 2010; **6:** e1000932.
14. Cai C, Langfelder P, Fuller TF *et al*: Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics* 2010; **11:** 589.
15. Jasinska AJ, Service S, Choi OW *et al*: Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. *Hum Mol Genet* 2009; **18:** 4415-4427.
16. Rollins B, Martin MV, Morgan L, Vawter MP: Analysis of whole genome biomarker expression in blood and brain. *Am J Med Genet B Neuropsychiatr Genet* 2010; **153B:** 919-936.
17. Iwamoto K, Kakiuchi C, Bundo M, Ikeda K, Kato T: Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Mol Psychiatry* 2004; **9:** 406-416.

18. Tsuang MT, Nossova N, Yager T *et al*: Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: a preliminary report. *Am J Med Genet B Neuropsychiatr Genet* 2005; **133B:** 1-5.

19. Gladkevich A, Kauffman HF, Korf J: Lymphocytes as a neural probe: potential for studying psychiatric disorders. *Prog Neuropsychopharmacol Biol Psychiatry* 2004; **28:** 559-576.

20. Glatt SJ, Everall IP, Kremen WS *et al*: Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 up-regulation in schizophrenia. *Proc Natl Acad Sci U S A* 2005; **102:** 15533-15538.

21. Kuzman MR, Medved V, Terzic J, Krainc D: Genome-wide expression analysis of peripheral blood identifies candidate biomarkers for schizophrenia. *J Psychiatr Res* 2009; **43:** 1073-1077.

22. Saris CG, Horvath S, van Vught PW *et al*: Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* 2009; **10:** 405.

23. Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009; **84:** 210-223.

24. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19:** 185-193.

25. Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.

26. Du P, Kibbe WA, Lin SM: lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008; **24:** 1547-1548.

27. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3:** Article3.

28. Horton R, Wilming L, Rand V *et al*: Gene map of the extended human MHC. *Nat Rev Genet* 2004; **5:** 889-899.

29. Dixon AL, Liang L, Moffatt MF *et al*: A genome-wide association study of global gene expression. *Nat Genet* 2007; **39:** 1202-1207.

30. Stranger BE, Nica AC, Forrest MS *et al*: Population genomics of human gene expression. *Nat Genet* 2007; **39:** 1217-1224.

31. Monks SA, Leonardson A, Zhu H *et al*: Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* 2004; **75:** 1094-1105.

32. Goring HH, Curran JE, Johnson MP *et al*: Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 2007; **39:** 1208-1216.

33. Cheung VG, Jen KY, Weber T *et al*: Genetics of quantitative variation in human gene expression. *Cold Spring Harb Symp Quant Biol* 2003; **68:** 403-407.

34. Heinzen EL, Ge D, Cronin KD *et al*: Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol* 2008; **6:** e1.

35. Vandiedonck C, Knight JC: The human Major Histocompatibility Complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic* 2009; **8:** 379-394.

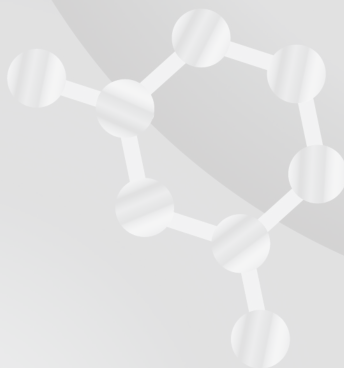
36. Eaton WW, Byrne M, Ewald H *et al*: Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *Am J Psychiatry* 2006; **163:** 521-528.

37. Laan W, Grobbee DE, Selten JP, Heijnen CJ, Kahn RS, Burger H: Adjuvant aspirin therapy reduces symptoms of schizophrenia spectrum disorders: results from a randomized, double-blind, placebo-controlled trial. *J Clin Psychiatry* 2010; **71:** 520-527.

38. Doorduin J, de Vries EF, Willemsen AT, de Groot JC, Dierckx RA, Klein HC: Neuroinflammation in schizophrenia-related psychosis: a PET study. *J Nucl Med* 2009; **50:** 1801-1807.

39. Bromberg KD, Kluger HM, Delaunay A *et al*: Increased expression of the E3 ubiquitin ligase RNF5 is associated with decreased survival in breast cancer. *Cancer Res* 2007; **67:** 8172-8179.

40. Gibbs JR, van der Brug MP, Hernandez DG *et al*: Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS genetics* 2010; **6:** e1000952.



Chapter 4

Identification of schizophrenia-associated loci by combining DNA methylation and gene expression data from whole blood

Kristel R. van Eijk, Simone de Jong, Eric Strengman, Jacobine E. Buizer-Voskamp, René S. Kahn, Marco P. Boks, Steve Horvath, Roel A. Ophoff

In press (European Journal of Human Genetics)

Abstract

Emerging evidence suggests that schizophrenia susceptibility involves variation at genetic, epigenetic, and transcriptome levels. We describe an integrated approach that leverages DNA methylation and gene expression data to prioritize genetic variation involved in disease. DNA methylation levels were obtained from whole blood of 260 schizophrenia patients and 250 unaffected controls of which a subset with gene expression data available. By assessing DNA methylation and gene expression in cases and controls we identified 432 CpG sites with differential methylation levels that are associated with differential gene expression. We hypothesized that genetic factors involved in these methylation levels may be associated with the genetic risk of schizophrenia susceptibility. To test this hypothesis, we used results from the Psychiatric Genomics Consortium (PGC) schizophrenia genome-wide association study (GWAS). We observe an enrichment of schizophrenia-associated SNPs in the mQTLs of which the associated CpG site is also correlated with differential gene expression in schizophrenia. While this enrichment was already apparent when using nominal significant thresholds, enrichment was even more pronounced when applying more stringent significance levels. One locus, previously identified as susceptibility locus in a schizophrenia GWAS, involves SNP rs11191514, which regulates DNA methylation of *CALHM1* that is also associated with differential gene expression in patients, Overall, our results suggest that epigenetic variation plays an important role in schizophrenia susceptibility and that the integration of analyses of genetic, epigenetic and gene expression profiles may be a biologically meaningful approach for identifying disease susceptibility loci, even when using whole blood data in studies of brain-related disorders.

Keywords: DNA methylation, mQTLs, gene expression, enrichment, schizophrenia

Introduction

Although a number of susceptibility loci for psychiatric disorders have been identified through genome-wide association studies, the vast majority of the estimated heritability of these traits remains unexplained. Schizophrenia (SZ), for example, is a common polygenic mental disorder affecting about 1% of the population and has an estimated heritability of 70-80%[1-3], but only a small fraction of the heritability can be attributed to known susceptibility loci[4,5]. Very large numbers of samples are needed to comprehensively identify the hundreds or possibly thousands of genetic loci involved in SZ susceptibility[6]. It has been suggested that epigenetic variation might be partly responsible for the missing heritability[7,8] and be involved in phenotypic variation[9-11]. Recent evidence shows that schizophrenia susceptibility loci are enriched for gene expression QTLs[12], and a similar enrichment exists for QTLs affecting DNA methylation and expression in bipolar disorder[13]. Given these results we hypothesized that intersecting disease-related gene expression data with disease-related methylation data might lead to identification of genetic susceptibility loci. To test this hypothesis, we combined available whole blood DNA methylation and gene expression data of schizophrenia patients and healthy controls. Using the results of the Psychiatric Genomics Consortium (PGC) schizophrenia mega GWAS[14], we examined whether methylation QTLs (mQTLs) are enriched for disease susceptibility loci. While previous studies examined gene expression and DNA methylation separately and in different ways[12,13], we consider both DNA methylation and gene expression data simultaneously in the same SZ cases and controls. Previous studies investigating enrichment of QTL signals (involving either gene expression or DNA methylation) focused on SZ-related SNPs with genome-wide significant evidence of association. We combined epigenome and transcriptome levels of information and did not restrict ourselves to only those loci with prior evidence of genome-wide association. Our results show that the combined use of gene expression and methylation data outperforms either data modality when it comes to identifying disease susceptibility loci, even in a relatively small sample size compared to GWAS.

Method

An outline of the approach is provided in **Figure 1**. Genome-wide genotype data and DNA methylation data were obtained from whole blood of 260 schizophrenia patients and 250 control subjects; for a subset of 120 cases and 120 controls, whole blood gene expression data were available as well. A detailed description of the samples, procedures and quality control steps is provided in the supplementary material. First, we examined differential methylation levels (after quantile normalization) between cases and controls. DNA methylation levels with low variability (a Beta range below 0.1) were excluded to focus on CpGs with reasonably large biological variation. We used a linear model (limma package

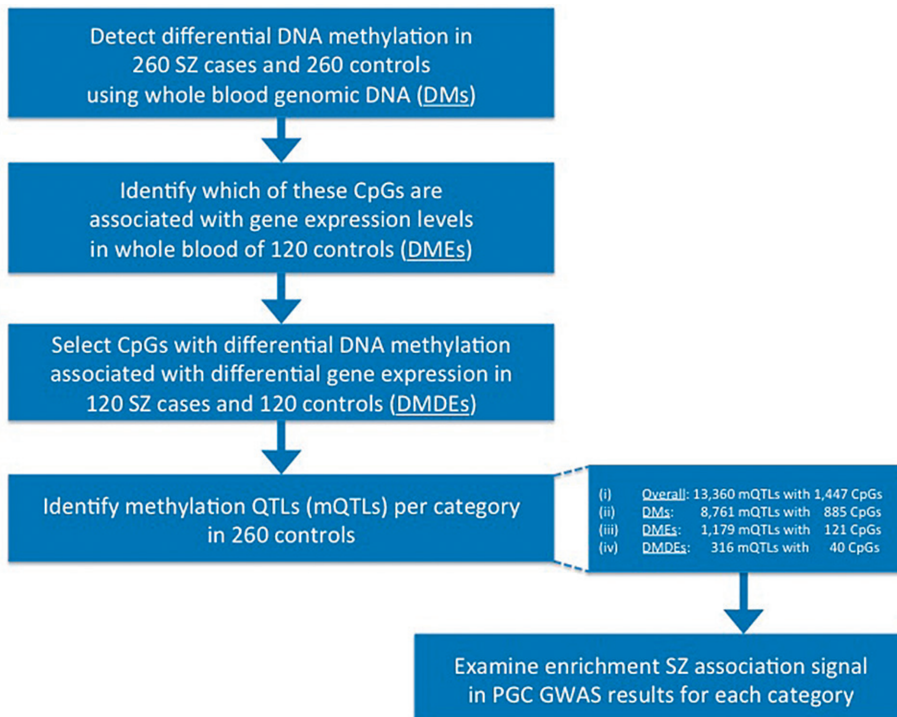


Figure 1: Outline of study with different filtering steps (see main text for details).

in R[15]) to regress methylation values on disease status, gender, and age[16]. FDR correction at the 5% level was applied to correct for multiple testing.

Next, for methylation levels that differed between cases and controls (referred to as differentially methylated loci, or abbreviated DMs), association with gene expression was investigated using a genome-wide DNA methylation-gene expression association study as described previously[17]. We focused on CpG sites with *cis* effects on expression within 500 kb interval. Locus-specific correction for multiple testing was applied for the number of expression probes within each *cis* region of that methylation probe (i.e. 0.05 divided by the number of expression probes per *cis* area). This relatively lenient significance threshold was chosen to ensure a sufficient number of probes for subsequent filtering steps and analysis. The subset of DMs associated with gene expression is referred to as DMEs, as determined in control samples. A subset of the DMEs is not only associated with gene expression levels in general but is associated with transcript levels of differentially expressed genes when comparing schizophrenia patients and controls as described previously[18]. In short, we conducted linear regression to detect differences in expression levels between cases and controls after FDR correction at the 5% level, with age and gender as covariates[18].

The DMEs associated with differential gene expression in schizophrenia are referred to as DMDEs (Differential Methylation regions with Differential gene Expression).

To identify methylation QTLs (mQTLs), we used a multivariate linear regression model for regressing all methylation values (dependent variable) on the SNPs (independent variable) with disease status, age, and gender as covariates using PLINK[19]. We defined a SNP as *cis*-acting if significantly associated ($P < 5.0e\text{-}08$) within 500kb between the SNP and the methylation probe. Next, four groups of SNPs with a minimum minor allele frequency of 0.05 were generated with increasing functional relevance with regard to DNA methylation and gene expression: (i) SNPs representing all mQTLs regardless of differential methylation between patients and controls; (ii) SNPs from the previous step that represent mQTLs with differential methylation between cases and controls (DMs); (iii) SNPs from the previous step for which the associated methylation level is associated with gene expression (DMEs); and finally (iv) SNPs selected from the previous step that are associated with differential methylation that is associated with differential gene expression between patients and controls (DMDEs). After linkage disequilibrium (LD)-pruning the SNPs with an R^2 of 0.2 using PLINK, to exclude possible signal bias and enrich for independent genetic signal, we extracted the SNPs from the Psychiatric Genomics Consortium (PGC) mega GWAS results with their association signal with schizophrenia.
Based on the significance values observed in the PGC schizophrenia GWAS, we categorized the four different SNP lists and examined the observed/expected ratios for different thresholds ranging from $P < 0.5$ to $P < 1.0e\text{-}04$.

Results

We examined whether genetic factors involved in differential methylation and gene expression in schizophrenia are enriched for schizophrenia susceptibility alleles. The results of the steps are shown in **Figure 1**. In total, 11320 CpGs were differentially methylated in SZ cases versus controls (5% FDR correction). Of these, 1095 CpGs are associated with 1226 transcripts in *cis* (after locus-specific correction). The 1226 transcripts were examined for differential expression levels between patients and controls based on data from a previous study of our group[18]. This step resulted in the identification of 391 transcripts (after 5% FDR correction); these 391 transcripts are associated with 432 CpG sites. Information about the differentially methylated probes is shown in **Supplementary Table 1**.

We subsequently calculated the mQTLs and SNPs associated with CpGs in *cis* with a *P*-value $< 5.0e\text{-}08$ and these were grouped by the biological relevance of the associated CpG site

(see methods). **Figure 1** shows the number of associated CpGs and SNPs, and the statistics of the associations are in **Supplementary Table 2**.

After calculating the observed/expected ratio of the mQTLs per PGC GWAS p-value threshold, our results show a significant pattern of enrichment of PGC schizophrenia association signal by adding DNA methylation and gene expression data (**Figure 2**). The strongest enrichment was observed for SNPs associated with DMDEs ($P = 0.0041$ with OR = 22 and 95%CI 2.60-83.7; Fisher's exact test).

We identified one DMDE locus on chromosome 10 with genome-wide significant evidence of association with schizophrenia. SNP rs11191514 at this locus represents an mQTL associated with differential DNA methylation between cases and controls that is correlated with expression of a gene that is also differentially expressed between cases and controls (**Figure 3a-c**). Additionally, this SNP is in perfect LD with rs11191580, a SNP with significant evidence of association with schizophrenia according to the PGC study (PGC $P = 2.2e\text{-}08$). SNP rs11191514 is also strongly associated with schizophrenia (PGC $P = 8.7e\text{-}08$) and likely

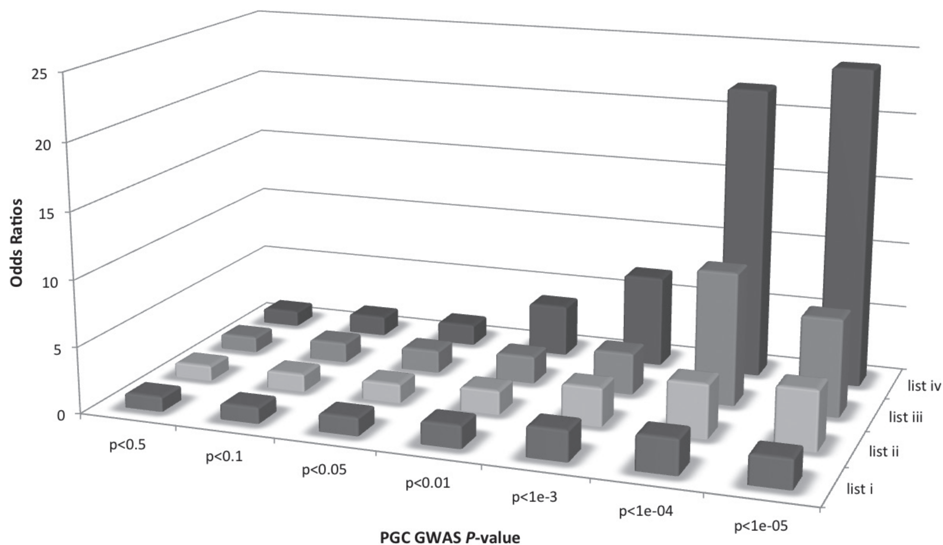


Figure 2: Enrichment of genetic association with Schizophrenia

For each of the four categories of mQTLs in the whole blood case/control data set we observed enrichment of genetic association signal based on PGC Schizophrenia genome-wide association study findings[14]. The four categories are (i) *all* mQTLs without prior selection, (ii) DMs representing mQTLs with differential methylation between cases and controls, (iii) DMEs representing mQTLs with differential DNA methylation between cases and controls associated with gene expression, and (iv) DMDEs, representing mQTLs with differential DNA methylation that is associated with differential gene expression between cases and controls. Detailed values of Odds Ratios, significance levels and confidence intervals are provided in **Supplementary Table 1**.

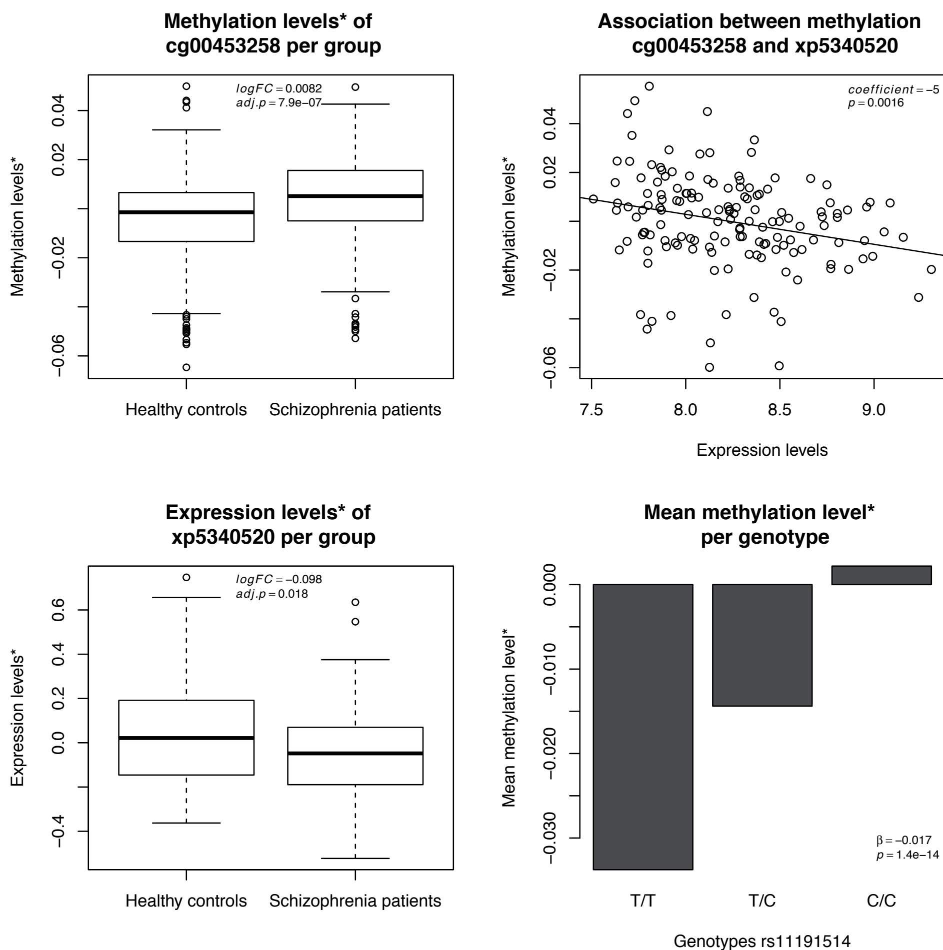


Figure 3: Correlations between rs11191514, methylation/gene expression and disease of locus with previously reported genome-wide significant evidence of association with schizophrenia.

(a) CpG site cg00453258 (representing gene *CALHM1*) is increased in schizophrenia cases versus controls; (b) this CpG site is negatively associated with gene expression, transcript probe xp5340520; (c) expression of this transcript is reduced in schizophrenia cases; (d) the mean methylation level per genotype. In the PGC schizophrenia GWAS result, the T allele is a protective allele with an OR of 0.82 (see **Supplementary Table 1**) with increased allele frequency in controls, and in our dataset associated with a decreased methylation level. This fits with the increased methylation levels in schizophrenia. *To generate these plots the residuals of the methylation values were taken after correction for age and gender in (a) and (b), and for expression in (c).

regulates DNA methylation in our data set (P = 1.4e-14, **Figure 3d**). SNP rs11191514 is located in *CNNM2*, while the methylation probe, associated with this SNP, is located at *CALHM1*, some 630KB apart (**Supplementary Figure 1**). *CALHM1*, calcium homeostasis modulator 1, regulates Ca^{2+} concentrations[20] and is highly expressed in the hippocampus[20],

which is implicated in schizophrenia[21]. Recently, a GWAS identified this whole locus as susceptibility locus for schizophrenia, including the genes *CALHM1* and *CNNM2*, $P = 3.7e\text{-}13$[22].

In addition to rs11191514, we identified four additional independent DMDE SNPs associated with schizophrenia with significance level of $p < 0.01$ (**Supplementary Table 2**), i.e. these SNPs are at best suggestive of a susceptibility locus for schizophrenia.

These additional loci highlighted in our study include *PRRT1*, *HLA-C*, and *MRPL41*. The CpG sites of these genes are all differentially methylated in patients, are associated with genotype as well as with transcripts that are differentially expressed in patients. *PRRT1*, proline-rich transmembrane protein 1, and *HLA-C* are both located within the MHC region on the short arm of chromosome 6; a region with the most significant association with schizophrenia[14,23]. Firstly, little is known about the function of *PRRT1* and this finding suggests that further study of its involvement in schizophrenia is warranted. *HLA-C* belongs to the HLA class 1 molecules that play a central role in the immune system and is located within the chromosomal region first implicated in SZ[11]. Lastly, *MRPL41* is a mitochondrial ribosomal protein and plays an important role in cell growth suppression in association with p53 and p27Kip1[24]. It is not clear how this ribosomal protein may be involved in the etiology of schizophrenia. As the MHC region is overrepresented in the PGC schizophrenia GWAS findings, we were concerned about a possible bias in our analysis. However, when we performed the same analyses without chromosome 6, we still observed an enrichment of genetic signal associated with schizophrenia ($P = 0.0037$ with OR=23 and 95%CI 2.8-89.2; Fisher's exact test). This suggests that functional enrichment analysis by using gene expression and DNA methylation from whole blood is not dependent on the MHC region.

Discussion

We investigated the enrichment of mQTLs for schizophrenia susceptibility alleles using genotype and DNA methylation data obtained from whole blood of more than 500 schizophrenia cases and controls, and of which gene expression data was available from a subset. We detected over 10,000 CpG sites that are differentially methylated in schizophrenia patients. A subset of these sites (n=1,095) was associated with gene expression with 50% of these transcripts also showing differentially gene expression in schizophrenia. Additionally, SNPs associated with differential methylation levels are enriched for schizophrenia susceptibility alleles. This enrichment was even stronger if the differential methylation level associated with these SNPs were also associated with differential expression levels. While this enrichment was already apparent when using nominal significant thresholds, enrichment was even more pronounced when applying more stringent significance levels. Previous studies have shown that top GWAS findings are enriched for mQTLs and eQTLs in bipolar disorder[13],

and for eQTLs in schizophrenia[12]. In accordance, here we show that mQTLs are enriched for schizophrenia SNPs. In addition, we used different layers of genomic information to identify mQTLs associated with CpGs with functional relevance, and find increasing enrichment. Our findings show that genetic variation affecting DNA methylation that is associated with gene expression plays an important role in schizophrenia susceptibility.

As schizophrenia is primarily a brain-related disease, a limiting factor of our study may be the use of whole blood. However, our findings show that even when using blood, we observe enrichment of schizophrenia-associated alleles, indicating that blood might be a reasonable surrogate for our approach. Replication of these analyses in brain tissue would be useful to understand the extent of enrichment of disease-specific signal when combining different genomic layers for prioritizing genomic loci. Another potential expansion of this study includes the obtaining of allele-specific methylation and gene expression information, which could provide more insight into the precise mechanism of *cis*-acting regulating SNPs and their effects on methylation and gene expression. Finally, the relationship between genetic variation, DNA methylation, gene expression and disease susceptibility is complex and warrants further study.

In summary, we identified biologically plausible schizophrenia susceptibility loci in whole blood in a relative small sample of <600 subjects. We demonstrate that enrichment of genetic data using different layers of genomic information may be an efficient approach to identify disease susceptibility loci for neuropsychiatric traits. While our results are supportive of an important role of epigenetic regulation in schizophrenia, we expect that this integrated approach based on blood DNA methylation and gene expression data from the same subjects may help prioritize SNPs from other GWAS studies as well.

Acknowledgements

We thank the patients and controls for participating in the study. This study was supported by National Institutes of Health (NIH) grants R01 DA028526, R01 NS058980, and RO1 MH 078075 (to RAO). The authors report no financial conflicts of interest that might be construed to influence the results or interpretation of our study. We thank Dr. Kim Staats for critical reading of the manuscript.

Conflict of interest

The authors have declared that no competing interests exist.

References

1. Lee SH, DeCandia TR, Ripke S *et al*: Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 2012; **44:** 247-250.
2. Sullivan PF, Daly MJ, O'Donovan M: Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* 2012; **13:** 537-551.
3. Sullivan PF, Kendler KS, Neale MC: Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry* 2003; **60:** 1187-1192.
4. Purcell SM, Wray NR, Stone JL *et al*: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460:** 748-752.
5. So HC, Gui AH, Cherny SS, Sham PC: Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet Epidemiol* 2011; **35:** 310-317.
6. Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461:** 747-753.
7. Petronis A: Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010; **465:** 721-727.
8. Labrie V, Pai S, Petronis A: Epigenetics of major psychosis: progress, problems and perspectives. *Trends Genet* 2012; **28:** 427-435.
9. Feinberg AP, Irizarry RA: Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A* 2010; **107 Suppl 1:** 1757-1764.
10. Feinberg AP: Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007; **447:** 433-440.
11. Bjornsson HT, Fallin MD, Feinberg AP: An integrated epigenetic and genetic approach to common human disease. *Trends Genet* 2004; **20:** 350-358.
12. Richards AL, Jones L, Moskvina V *et al*: Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Molecular psychiatry* 2012; **17:** 193-201.
13. Gamazon ER, Badner JA, Cheng L *et al*: Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Molecular psychiatry* 2012.
14. Ripke S, Sanders AR, Kendler KS *et al*: Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011; **43:** 969-976.
15. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3:** Article3.
16. Horvath S, Zhang Y, Langfelder P *et al*: Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol* 2012; **13:** R97.
17. van Eijk KR, de Jong S, Boks MP *et al*: Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* 2012; **13:** 636.

18. de Jong S, van Eijk KR, Zeegers DW *et al*: Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. *Eur J Hum Genet* 2012; **20:** 1004-1008.

19. Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.

20. Dreses-Werringloer U, Lambert JC, Vingtdeux V *et al*: A polymorphism in CALHM1 influences Ca2+ homeostasis, Abeta levels, and Alzheimer's disease risk. *Cell* 2008; **133:** 1149-1161.

21. Harrison PJ: The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. *Psychopharmacology* 2004; **174:** 151-162.

22. Ripke S, O'Dushlaine C, Chambert K *et al*: Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013; **45:** 1150-1159.

23. Stefansson H, Ophoff RA, Steinberg S *et al*: Common variants conferring risk of schizophrenia. *Nature* 2009; **460:** 744-747.

24. Yoo YA, Kim MJ, Park JK *et al*: Mitochondrial ribosomal protein L41 suppresses cell growth in association with p53 and p27Kip1. *Mol Cell Biol* 2005; **25:** 6603-6616.

Supplementary files are available upon request.

Supplementary Figure 1; The *CNNM2-CALHM1* locus

The most significant finding from the PGC results, SNP rs11191514, (P = 8.71e-08 with schizophrenia) is located in the *CNNM2* gene. This SNP is in perfect LD (R^2=1) with SNP rs11191580 (located in the *NT5C2* gene), which was one of the loci with genome-wide significance for involvement in schizophrenia in the PGC study (PGC P=2.2e-08). The methylation probe associated with rs11191514 is located at *CALHM1*, some 630KB apart (indicated in yellow).

Supplementary Table 1; differentially methylated probes

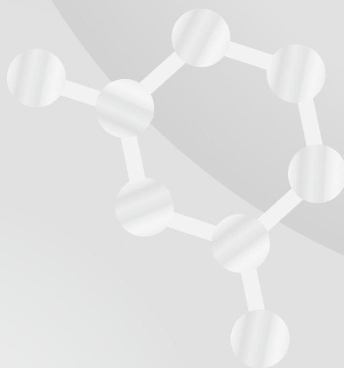
This table contains the differentially methylated probes between schizophrenia cases and controls.

Supplementary Table 2:

This file consists of the four lists with mQTLs and their associated CpG sites.

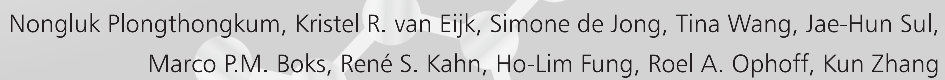
Supplementary Table 3:

The four loci described in the results are presented in this table. Results from all analyses and information about the methylation and expression probes are included.



Chapter 5

Characterization of genome-methylome interactions in 22 nuclear pedigrees



Nongluk Plongthongkum, Kristel R. van Eijk, Simone de Jong, Tina Wang, Jae-Hun Sul, Marco P.M. Boks, René S. Kahn, Ho-Lim Fung, Roel A. Ophoff, Kun Zhang

In press (PLOS ONE)

Abstract

Background: Genetic polymorphisms can shape the DNA methylome by either changing the substrate of DNA methyltransferases or alternating the affinity of *cis*-regulatory DNA binding proteins. The interactions between CpG methylation and genetic polymorphisms have been investigated by methylation quantitative trait loci (mQTL) analysis and allele-specific methylation (ASM) analysis. However, it remains unclear whether these approaches can effectively and comprehensively identify the genetic variants that contribute to inter-individual variation of DNA methylation.

Results: Here we took advantage of the family structure in 22 nuclear pedigrees, and used three independent approaches to systematically investigate the genetic influence on DNA methylation in whole blood. We applied targeted bisulfite sequencing with bisulfite padlock probes to quantify the absolute DNA methylation levels at a pre-selected set of 411,800 CpG sites in the human genome. With mid-parent offspring analysis (MPO), we identified 10,619 CpG sites that exhibit heritable methylation patterns, among which 70.1% are SNPs directly disrupting CpG dinucleotides. For the remaining 3,179 heritable CpG sites, only 45.2% were associated with *cis*-regulatory SNPs identified by mQTL analysis, and 3.1% exhibit allele-specific methylation in one of more individuals. Finally, we identified hundreds of clusters in the human genome in which the degree of variation (as opposed to the mean) of CpG methylation is associated with genetic polymorphisms, supporting a recently proposed hypothesis on the genetic influence of phenotypic plasticity.

Conclusions: This study showed that *cis*-regulatory SNPs identified by mQTL analysis account for only roughly half of the heritable CpG methylation, whereas the majority of ASM cannot be explained by consistent genetic regulatory effects among multiple individuals. Overall the extent of genome-methylome interactions is well beyond what is detectible with the commonly used mQTL and ASM analysis.

Keywords: DNA methylation, epigenetics, targeted bisulfite sequencing, bisulfite padlock probes, mid-parent offspring analysis, heritability of DNA methylation, methylation QTL, allele-specific methylation, variably methylated regions, variation-SNPs

Background

DNA methylation represents an important layer of epigenetic regulation on the activity of the human genome. Accumulating evidence suggests that the epigenome, including the methylome, varies from one individual to another[1-3]. Such variations are believed to play functional roles in the individual variations of a variety of phenotypes, including many human diseases[4; 5]. The inter-individual variation of the DNA methylomes is contributed by both genetic and environmental factors[6; 7]. With the advances in DNA methylation assays, a growing number of studies on the genetic contributions to the DNA methylomes have been reported. One type of studies relies on mQTL analysis to identify *cis*-regulatory variants associated with inter-individual variation of CpG methylation[8-11]. An alternative approach involved characterizing allele-specific methylation[2; 3; 12-15]. While an increasingly large number of SNP and CpG associations have been reported in these recent efforts, it remains unclear to what extent SNPs contribute to the inter-individual variation of DNA methylome globally, and how effective mQTL and ASM analyses are in uncovering genome-methylome interactions. In this study, we performed targeted bisulfite sequencing on human whole blood samples from 96 individuals representing 22 nuclear pedigrees, and used three independent approaches (MPO, mQTL and ASM) to characterize the associations between CpG methylation and genetic factors. In addition, we also investigated the genetic contribution to the variance of DNA methylation. We discussed the limits of mQTL and ASM analyses in characterizing the full extent of genome-methylome interactions.

Results

We characterized DNA methylation levels on genomic DNA of peripheral blood from 96 individuals in 22 nuclear pedigrees of European ancestry. Each pedigree included two unaffected parents, one proband with schizophrenia and one or two unaffected siblings. With this family structure, there are a total of 52 trios with two parents and one child. We measured CpG methylation at single base resolution by targeted bisulfite sequencing with ~330,000 bisulfite padlock probes capturing a pre-selected subset of genomic regions, including promoters, enhancers, DNase I hypersensitive sites and other regions known to be variable among different cell types[16]. Note that, like other bisulfite based methods, 5-methylcytosines and 5-hydroxymethylacytosines are indistinguishable with this assay. The quality of the data was assessed by comparing DNA methylation levels of the same CpG sites captured and measured independently on the two DNA strands, which can be treated as internal technical replicates. All samples have high correlation in methylation levels between the two strands (Pearson's correlation coefficient $R > 0.95$), indicating low technical noise. On average, we obtained methylation measurements on ~500,000 CpG sites per sample. A total of 411,800 autosomal CpG sites (416,933 CpG sites including sex chromosomes)

had valid methylation measurements in at least 80% of the samples. We filtered out the CpG sites showing low variability among the samples ("static CpG sites"), and focused on a subset of 79,604 (76,408 autosomal CpGs) variable CpG sites (with standard deviation among all samples ≥ 0.1). To verify the sample identity, genotypes called from bisulfite sequencing reads (bisREAD SNPs) were used for calculating genetic distance, and all samples were found to cluster in the correct families (data not shown). Hierarchical clustering based on the methylation levels of highly variable autosomal CpG sites (standard deviation ≥ 0.3) also showed a clustering pattern consistent with the family structure (**Figure S1** in Additional file 1).

As the sample size is too small for performing epigenome-wide association tests between CpG methylation and schizophrenia status, we examined a subset of 70 CpG sites within 100kb of six genes (ANK3, CACNA1C, FKBP5, ITIH3, ITIH4 and MIR137) identified in recent schizophrenia GWAS studies[17]. No significant association was identified after correcting for multiple tests. Therefore, in the following analyses we focused on treating CpG methylation as a quantitative trait and investigated the contributions by genetic variants.

Heritability of DNA methylation

The family-based samples in this study allowed us to examine the genetic contribution on variation of DNA methylation irrespective of the type and frequency of genetic variants (i.e. SNPs, indels, structural genomic variation). We estimated the narrow-sense heritability of DNA methylation for all variable CpG sites based on the mid-parent offspring analysis, which measures the correlation between the mean methylation levels of the parents and the methylation levels of their offspring (**Figure 1a**). To be considered as a heritable CpG site, we required $h^2 > 0.2$ in a minimum of ten trios with a FDR cutoff of 0.05 (Benjamini-Hochberg correction). We identified a total of 10,619 heritable CpGs (**Table S1** in Additional file 2), which accounts for ~13.3% of all variable CpG sites tested. This suggests that genetic factors account for over ten percent of variability in human blood DNA methylome among the samples used in this study. Further analysis revealed, however, that 70% (7,440) of heritable CpG sites overlapped with SNPs directly disrupting the CpG dinucleotides. This result indicates that the majority of heritable CpG methylation is due to genetic polymorphisms directly altering the substrates of DNA methyltransferases ("SNP-CpGs"), whereas other *cis*- or *trans*- regulatory effects account for a smaller fraction of ~30% (3,179) of heritable CpG methylation ("non-SNP CpGs") (**Figure 2a**). For these heritable non-SNP CpGs, 48.1% of the sites are in introns, 23.6% are in intergenic regions (not in 2.5kb of TSS and TES), 14.3% are within 2.5kb upstream of TSS, 11.3% are in exons, and a small fraction are in 2.5kb downstream of TES. Note that this distribution is similar to all the variable CpGs characterized in this analysis (15.6% in 2.5kb upstream of TSS, 12.6% in exons, 47.8%

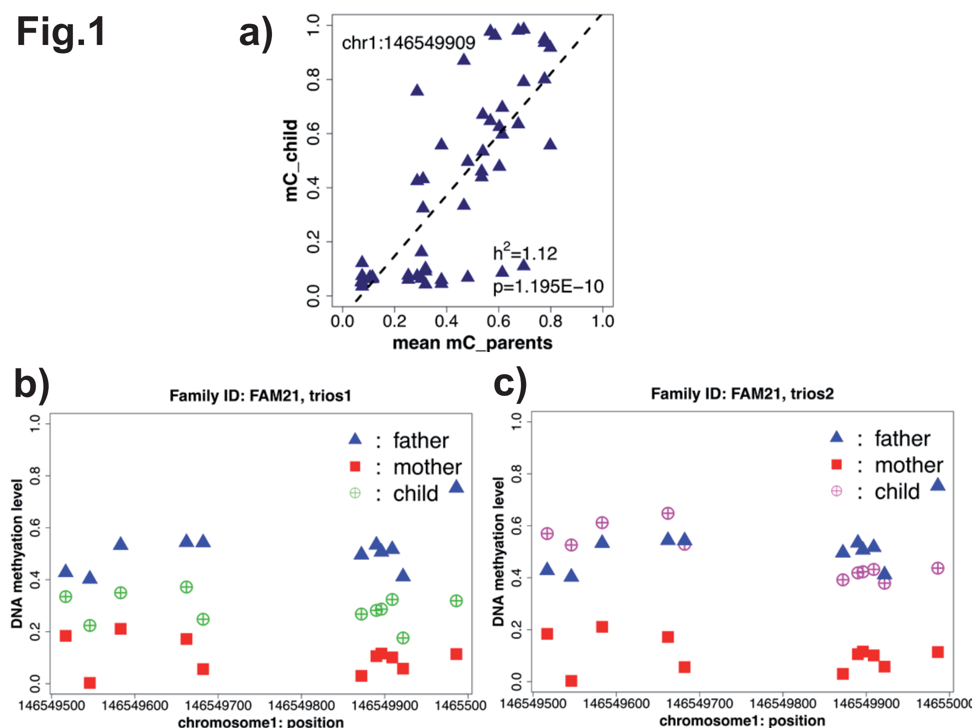


Figure 1: Identification of heritable CpG methylation by mid-parent offspring (MPO) analysis. (a) An example of mid-parent offspring regression of DNA methylation at the CpG site chr1:146549909. (b,c) DNA methylation level of heritable CpG at chr1:146549909 and the adjacent heritable CpGs on the same cluster exhibiting consistent pattern of DNA methylation between parents and their offspring on the two trios from the same family.

in introns, 2.8% in 2.5kb downstream of TES, and 21.2 in intergenic regions), and hence simply reflected the genomic targets covered by our probe set instead of indicating any enrichment to certain genomic features. The heritable non-SNP CpGs can be grouped into 92 clusters (**Table S2** in Additional file 3) each having at least five CpGs with a maximal spacing of 100,00bp within 100kb. Multiple CpG sites in a cluster exhibit highly correlated methylation among different individuals of the same family (**Figure 1b-c**), suggesting the effects of a single genetic variant or haplotype on multiple CpG sites within a cluster.

Association between genetic variation and DNA methylation

Next we treated each CpG site as a methylation quantitative trait locus (mQTL), and investigated the contribution of SNPs to the variability of DNA methylome. Note that the MPO analysis performed above covers all forms of genetic variants, whereas mQTL analysis focuses specifically on the effects of common SNPs or other genetic variants in linkage disequilibrium (LD) with the index SNPs. Here we sought to perform mQTL analysis on

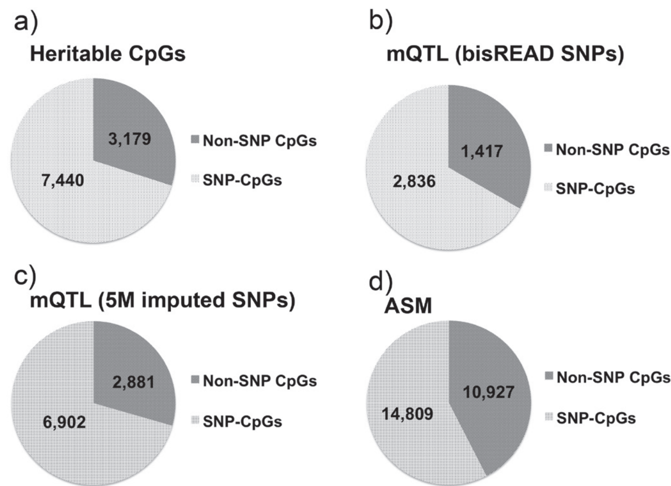


Figure 2: Fraction of non-SNP CpGs and SNP-CpG identified in MPO, mQTL, and ASM analysis. (a) Pie chart showing the number of heritable non-SNP CpGs and heritable SNP-CpGs. (b, c) Pie charts showing the fraction of mQTL associated non-SNP CpG and SNP-CpGs from mQTL analysis using bisREAD SNP data and 5M imputed SNP array data, respectively. (d) Pie chart showing the fraction of non-SNP CpG ASM and SNP-CpG ASM exist in at least one subject.

SNP genotypes available from multiple platforms, and to identify the optimal strategy through a comparison with the heritable SNPs found in the MPO analysis. From the bisulfite sequencing data, we called ~87,600 SNPs in the captured regions for each subject. After filtering out SNPs with too many missing genotypes not called in at least 75% of subjects, we obtained genotypes on 15,450 SNPs for mQTL analysis. Because these SNPs were called only in the captured regions, SNP density was low. To obtain genotypes on more SNPs for a more comprehensive mQTL mapping, we included additional genotypes on 57 of the 96 subjects generated by Affymetrix and Illumina SNP arrays. To avoid platform-specific technical differences, we performed imputation using the 1000 genomes SNP data, and obtained genotypes on approximately 5 million SNPs after quality filtering. At this density there were multiple SNPs in the same LD blocks, so we also performed association tests on a subset of 618,580 SNPs present on Illumina 1M SNP array to avoid the excessive penalty of correcting too many statistical tests. Regression analysis was performed using PLINK[18] between DNA methylation levels and genotypes of SNPs located within 1Mb upstream and downstream to the CpG sites. To account for the dependence between related individuals, QFAM analysis using PLINK were also performed. The empirical p-value (EMP) from 10,000 permutations was used to identify the significant signals along with asymptotical p-value from non-permuted data by linear regression analysis. With genotypes on the 15,450 SNPs called from bisulfite sequencing data (bisREAD SNPs), we identified 7,593 associated CpG-SNP pairs at $<5\%$ FDR (Benjamini-Hochberg, p-value $< 2.74E-05$, EMP <0.0005), consisting

of 4,253 CpG sites associated with 3,842 SNPs. After eliminating SNPs overlapping with CpGs (SNP-CpGs), we observed 2,248 association signals (representing 1,417 non-SNP CpG sites associated with 966 SNPs, **Figure 2b**). With the genotypes of 5,257,772 genome-wide SNPs, we identified a total of 644,773 CpG and SNP *cis*-associations within 1Mb distance at <5% FDR (Benjamini-Hochberg, p-value < 5.98E-06), consisting of 9,783 CpGs associated with 412,382 SNPs, with the majority of associations involving SNP-CpGs. For non-SNP CpGs, there were 144,780 associated CpG SNP pairs (2,881 CpGs associated with 63,594 SNPs) (**Figure 2c**). We next repeated the mQTL and QFAM analyses on a subset of 618,580 SNPs that were present on the Illumina 1M array. As expected, the number of associations decreased to 67,781 associations (at FDR <5%), including 2,189 non-SNP CpGs associated with 7,004 SNPs. To maximize the sensitivity, it is important to balance the penalty of including too many SNPs for the association tests with the genome coverage of the SNPs.

The majority of *cis*-regulatory SNPs locate very closely to their associated CpG sites. For the bisREAD SNPs called from the bisulfite sequencing reads, 47.6% of the SNP and CpG associations are within 2kb, including 861 CpGs associated with 507 SNPs (**Table S3**, **Figure S2a** in Additional file 1). Only 15.2% of the associations are over 100kb (**Table S3**, **Figure S2b, S2e** in Additional file 1). The enrichment of SNP and CpG associations in short distance identified here could be due to the sampling bias, because the SNPs were called only in the captured regions and tend to locate very close to CpG sites. However, with the 5,257,772 genome-wide SNPs that more uniformly capture the LD blocks in the human genome, we found that over 64% of SNP and CpG associations were within 100kb (**Table S4**, **Figure S2f** in Additional file 1), with the strongest associations mostly within 2kb (**Table S4**, **Figure S2c** in Additional file 1), including 1,640 CpGs associated with 4,706 SNPs. On the other hand, we did observe slightly more associations over larger distances (**Figure S2d, S2f** in Additional file 1). In the mQTL analysis using ~5 millions SNPs and ~600 thousands SNPs, we identified slightly different numbers of non-SNP CpGs (2,881 versus 2,189), indicating a limited gain of sensitivity in identifying mQTLs by a >8-fold increase in SNP density. However, we did observe more significant associations for many CpG-SNP pairs when using ~5 million SNPs, presumably because the chance of capturing the functional SNPs or tagging SNPs very close to the functional variants is higher with a denser set of markers. Note that bisREAD SNPs were called from our methylation sequencing data, whereas SNP genotyping experiments involved extra experimental cost. Even the number of bisREAD SNPs used in our analysis is ~340 fold less than the imputed SNPs, we managed to identify half of associations involving non-SNP CpG sites compared to 5,257,772 genome-wide SNPs. Therefore, in the situation when SNP genotyping experiments are difficult to perform due to either limited biological materials or budget, SNPs called from bisulfite sequencing data can

be used to capture a reasonable fraction of *cis*-regulatory interactions, with the caveat that long distance interactions would be under-represented.

Mapping allele-specific methylation

Taking advantage of the methylation levels and SNP genotypes called on individual bisulfite sequencing reads, we next used a third strategy to examine the influence of genetic variations on DNA methylation by allele-specific methylation (ASM) analysis. Using a recently developed computational procedure[13], we identified an average of 2,266 CpG sites per individual that exhibit a significant difference in allelic methylation (allelic methylation difference >0.2). Note that, in contrast to MPO and mQTL analysis, ASM events were called on individual subjects. Biological, genetic and technical factors can contribute to the differences of ASM events called among different samples. Consistent with previous observations by us and others[12; 13; 19], SNP-CpGs account for the majority of ASM events (69.7%-92.5%, average 86.4%). ASM events involving non-SNP CpG sites represent a minority (7.5%-30.3%, average 13.6%). Examples of ASM events are shown in **Figure S3a-3b** in Additional file 1. While we observed an average of 313 ASM events on non-SNP CpGs per subject, the majority of these ASM events were present in only a small fraction of subjects (**Table S5** in Additional file 1). After combining all overlapping ASM events, we obtained 10,927 and 14,809 ASM events on non-SNP CpGs and SNP-CpGs respectively (**Figure 2d**). For the non-SNP CpGs, 47.1% of CpGs are in introns, 26.7% are in intergenic regions, ~13.5% are in 2.5kb upstream of TSS, ~8.4% are in exons and small numbers of ASM were in 2.5 kb downstream of TES (**Figure S3c** in Additional file 1). The distribution of SNP-CpG ASM was shown in **Figure S3d** in additional file 1. Comparing to the distribution of all CpGs characterized in ASM analysis (16.0% in 2.5kb upstream of TSS, 11.9% in exons, 48.6 in introns, 2.9% in 2.5 kb downstream of TES, and 20.5% in intergenic regions), we observed a modest enrichment of ASM on non-SNP CpGs in intergenic regions.

Intersecting three different approaches in dissecting DNA methylome and genome interaction

As the three analyses were implicitly or explicitly based on different assumptions and the sensitivity/specificity was limited by various technical factors, we next compared the hits identified by the three analyses. We expect significant overlaps among the three methods unless the false positive rates are high for one or more approaches. Indeed, 49.9% of non-SNP CpGs identified in the mQTL analysis were also found heritable in the MPO analysis (**Figure 3a**), indicating that at least half of the CpG sites by mQTL mapping were the true positive signals. Even this number is lower than expected, as high as 73.3% of associations involved heritable CpGs. On the other hand, close to 54.8% of the heritable non-SNP CpGs did not overlap with mQTL hits (**Figure 3a**), indicating that the extent of genome-methylome

interactions is more than what was identified with mQTL analysis alone. This could be due to a number of reasons, including lack of statistical power due to limited sample size, and presence of longer-range *cis*-interactions over 1 megabases and/or *trans*-interactions[20], and the effects of other common or rare alleles not in LD with the SNP tested. In addition, some marginally significant sites might be included or excluded due to the choices of p-value cut-offs by the two methods. In fact, when we plotted the mQTL association signals for heritable and non-heritable CpG sites separately, the majority of CpGs most strongly associated with SNPs (low p-value) were heritable CpGs (**Figure 3b**). Non-heritable CpGs in general showed weaker association signals, especially for longer-range *cis*-interactions (**Figure 3c**). A similar pattern was observed for SNP-CpG sites (**Figure S4a-c** in Additional file 1). Therefore, the heritable CpG sites not overlapping with mQTL hits are likely to be regulated by other genetic mechanisms.

In contrast to the mQTL analysis, only very small fractions of CpG sites that exhibit ASM in at least one sample were found heritable (5.6% for non-SNP CpGs, 32.6% for SNP-CpGs), and even smaller fractions overlapped with CpG sites significant in mQTL (2.8% for non-SNP

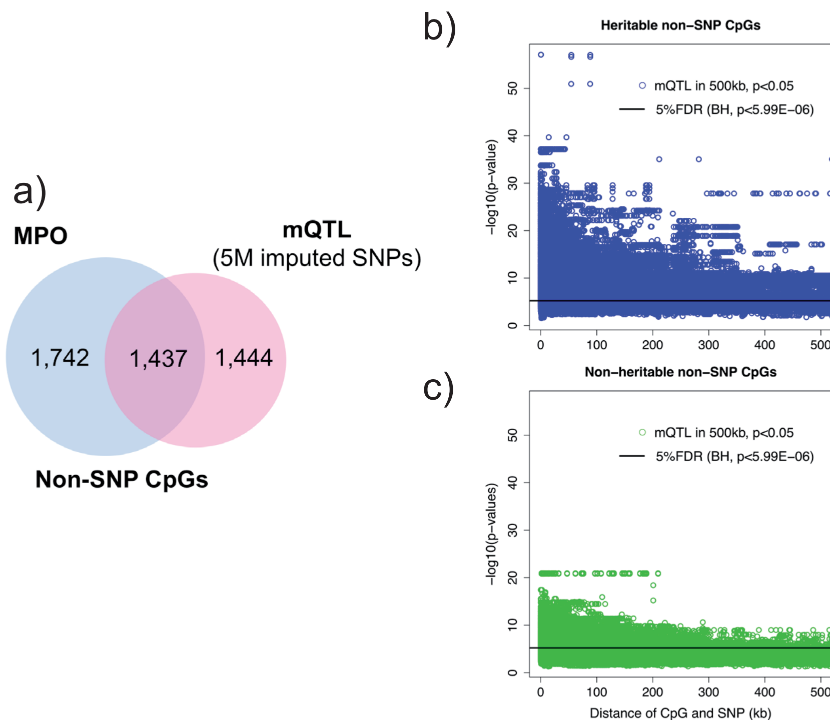


Figure 3: Mapping of CpG sites identified in MPO and mQTL analyses. (a) Venn diagrams showing overlap between non-SNP CpG sites significant in mQTL on 5,257,772 imputed SNPs and heritable CpGs. (b, c) Distribution of heritable CpGs and non-heritable CpGs and associated SNP pair distance within 500kb and their corresponding p-values from mQTL analysis on imputed SNPs.

CpGs, 21.9% for SNP-CpGs). One possibility is that our ASM calls contain too many false positives. However, when we restricted our analysis to the CpG sites that exhibit consistent ASM patterns in two or more individuals, the fractions of sites overlapping with heritable CpGs increased only moderately, far from the 49.9% overlap observed between mQTL and heritable CpGs (**Table S5** in Additional file 1). This leaves a number of other possibilities, including non-genetic parent-of-origin effects (including but not limited to imprinting), random allelic drift[21], environmental factors, potentially higher false positive rates or higher sensitivity in detecting the allelic differences with the ASM analysis. Overall, ASM appears to have very low specificity in identifying CpG sites regulated by genetic variants.

Variation-SNP affecting DNA methylation variance

Finally, we performed a search for variation-SNPs that were recently proposed to control the highly variable methylation regions recently observed in the human genome[22; 23]. Under this hypothesis, a particular allele of a SNP is associated with the gain or loss of methylation variability, as opposed to the mean methylation level, among multiple individuals. The search of such variation-SNPs involved a regression between the variance of methylation level and the genotypes. We performed a regression analysis on variance of DNA methylation at each CpG site and the genotypes of adjacent SNPs (within 1Mb). A major technical challenge is that there are only three genotypes for each SNP, and hence the sample size for each regression is limited to three, which results in a very high false positive rate. To improve the confidence for detecting true variation-SNPs, we required that a candidate SNP has consistent effects on at least five adjacent CpG sites. The false positive rate was estimated to be ~10%, by applying the same procedure on the randomly permuted methylation data. Similar to SNPs that are associated with the mean methylation level of multiple adjacent CpG sites (**Figure 4a**), we also observed SNPs associated with the variance of multiple CpGs (**Figure 4b**). A total of 1,058 variably methylated regions (VMRs) were identified (**Table S6** in Additional file 4), which were further grouped into 383 VMR clusters (**Table S7** in Additional file 5) when combining multiple VMRs that are within proximity of 100kb.

The majority of VMR clusters (316 clusters, 82.5%) locate adjacent to genes (438 genes). The largest VMR cluster involves 53 variable CpG sites in a 38kb region covering GNAS, which is a well documented imprinted gene that has a highly complex expression pattern from both strands[24; 25]. Two other large VMR clusters overlap with the HoxA gene cluster and protocadherin gamma gene cluster, both contain multiple functionally related and co-regulated genes and pseudogenes. Additional genes associated with the top ten VMR clusters are listed in **Table 1**. While the full functional consequences of such variable methylation remain largely unknown, we note that very recently four SNPs were found to be associated with rheumatoid arthritis and the variance of methylation[26]. In order to test whether the

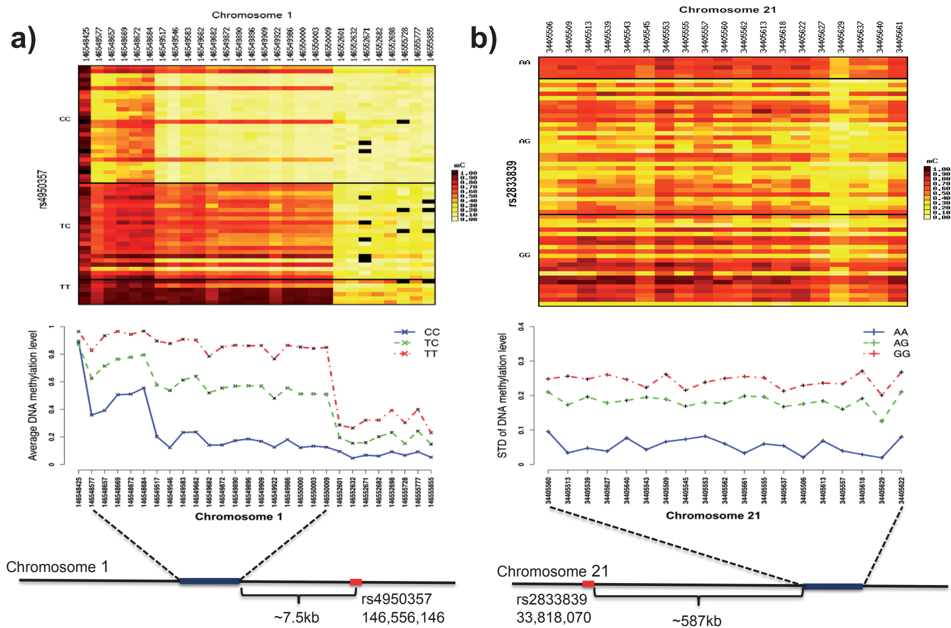


Figure 4: Genotype effects on the mean and variance of DNA methylation (a) Heatmap and line plot showing the association between rs4950357 SNP and the mean methylation of heritable CpGs cluster on chromosome 1 (chr1: 146548425-146555855). (b) The association of rs2833839 vSNP and the variance of methylation on VMR (chr21:34405506-34405661).

Table 1: **The top 10 VMR clusters and their associated genes.** The genes in blue expressed at detectible level in whole blood and were selected for association testing.

Number of variable CpGs in VMR clusters	VMR cluster coordinates	Associated genes
53	chr20:57426730-57464571	GNAS, GNAS-AS1
49	chr8:144358566-144371985	GLI4, ZNF696
47	chr7:27143370-27184750	HOXA2, HOXA3, HOXA5, HOXA6, HOXA-AS3
44	chr5:140718989-140863492	PCDHGA1,PCDHGA2,PCDHGA3,PCDHGA4,PCDHGA5, PCDHGA6,PCDHGA7,PCDHGA8,PCDHGA11,PCDHGB1, PCDHGB2,PCDHGB3,PCDHGB4,PCDHGB7,PCDHGB8P, PCDHGC3,PCDHGC4
41	chr20:32255315-32255936	ACTL10,NECAB3
35	chr5:135415001-135416725	VTRNA2-1
28	chr19:57349099-57352134	MIMT1, PEG3, ZIM2
26	chr8:145162974-145164623	KIAA1875, MAF1
26	chr11:7110142-7110456	RBMXL2
24	chr1:205818899-205819600	PM20D1

observed VMR clusters also translate into genotype-specific variation at gene expression level, we examined the top 10 VMR clusters and their respective genes in an array-based whole blood gene expression data set of independent 240 subjects[27]. Nine of the genes within the top ten VMR clusters were expressed at detectable levels (**Table 1**). Each of the probes and the variation-SNP was first tested for expression QTLs but none were detected (data not shown). Standard deviation for each probe was calculated per genotype group of corresponding variation-SNP. Even though the effects sizes are small, we observed three genes, *GNAS*, *PEG3*, and *PCDHGA5*, from different VMR clusters with genotype-specific differences on variance at gene expression level (**Figure 4b**).

Discussion

In the recent years, association mapping of molecular phenotypes, such as gene expression, DNA methylation or chromatin accessibility as quantitative traits (eQTL, mQTL, dsQTL), have revealed how genetic variants contribute to the inter-individual variability of these quantitative traits measured at the molecular level, and provided additional insights on how genetic variants modulate disease susceptibility[1; 28-32]. The recent technical advances in low-cost genome-wide DNA methylation assays (such as Illumina 450k methylation array[33], RRBS[34], and BSPP[16]), has catalyzed a new wave of epigenome-wide association studies aiming to characterizing the contribution of both genetic and environmental factors to disease susceptibility[6; 35], with encouraging progresses already in sight[26; 36-38]. However, aside from connecting genetic variants with CpG methylation, and disease phenotypes, additional questions remain to be addressed. To what extent we should expect interaction to occur between genetic variation and the variability of DNA methylation, what fraction of these interactions are we able to capture with different approaches, and what strategy should we use to efficiently capture these interactions?

In this study, we have revealed a much greater extent of genome-methylome interactions than previously recognized. By comparing the results from two orthogonal approaches (MPO, mQTL) on the DNA methylation profiles obtained from 22 nuclear pedigrees, we demonstrated that a large fraction of heritable traits on CpG methylation remain hard to detect with widely used mQTL association mapping. We observed that the vast majority of CpG sites exhibiting heritable methylation patterns is due to genetic variation at the CpG dinucleotide disrupting the methylation target. For the non-SNP CpGs, increasing the sample size will definitely improve the sensitivity in detecting weak *cis*-regulatory signals. However, we hypothesize that *trans*-regulation might account for the majority of heritable CpG sites not detectible by conventional mQTL analysis. While the anti-correlation of promoter DNA methylation and gene expression has been observed for many years, how DNA methylation

regulates gene expression has yet to be firmly established at the mechanistic level. More recent observations of positive correlation between gene-body methylation and gene expression added additional confusion on the functional roles of DNA methylation[39-42]. Stadler et al. recently demonstrated elegantly that binding of protein factors to DNA can lead to the local reduction of DNA methylation[43], providing the first direct evidence that DNA methylation in general is a passive mark for protein-DNA binding. A corollary of this observation is that, a DNA binding protein (such as a transcription factor) of which the expression is an eQTL (regulated by a genetic variant) can affect the DNA methylation level in hundreds to thousands of its binding regions genome-wide. As such, a single functional variant can regulate many mQTLs, mostly in trans, mediated by its primary effect on a single transcription factor. Connecting these mQTLs to the functional variants cannot be accomplished by simple association tests using adjacent CpG and SNPs. It requires additional information on the transcriptional factors and their direct regulating genes, which are becoming increasingly available through large-scale ChIP-Seq and DHS mapping efforts like the ENCODE project[44]. This also calls for a coherent statistical framework for association testing by incorporating the information of protein-DNA binding from genome-wide assays.

We also provided a practical assessment on the sensitivity of mQTL mapping at various SNP densities, showing that using over a million SNPs can improve the level of statistical significance with limited gain on detecting additional associated CpG sites. On the other hand, for projects based on bisulfite sequencing, the SNP genotypes called from the sequencing reads alone can be used to recover a reasonable fraction of associated CpG sites. As bisulfite sequencing is being widely adopted and algorithms for SNP calling from bisulfite data being optimized[45], this could represent an economical option for large-scale EWAS studies, with the understanding that a denser SNP map is still necessary to recover the majority of long-range regulatory effects. While we found many CpG sites that both exhibit allele-specific methylation in different individuals and show heritable methylation patterns across all the pedigrees, the majority of CpG sites identified in our ASM analysis cannot be explained by consistent effects of *cis*-regulatory variants across multiple individuals. We reason that ASM analysis is more susceptible to many non-genetic factors, including parent-of-origin effects, random allelic drift, and technical artifacts, and hence might not be that appropriate as the primary approach for identifying methylation traits regulated by genetic variants.

Finally, we provided evidence to support a recent hypothesis that genetic variants can regulate not only the mean but also the variation of molecular phenotypes, such as CpG methylation or gene expression. This is not unexpected, as gene regulatory networks are connected through positive and negative feedbacks[46; 47]. Reduction of negative feedback has been

shown to increase the variability in both prokaryotic and eukaryotic organisms[48; 49], lending mechanistic supports that genetic variants, affecting the strength of negative regulation, can result in the difference of variability for the components involved in the molecular network. Feinberg and colleagues have proposed that epigenetic variability provided a mechanism for selectable phenotypic variation[23], and provided examples of variable DNA methylation and their roles on cancer[22] and rheumatoid arthritis[26]. Although the full extent of variable DNA methylation, as well as their phenotypic consequences, remain to be further characterized with larger cohorts of genetically unrelated individuals, the observation of hundreds of VMRs in the 22 nuclear pedigrees suggests that the inherent variability of CpG methylation, and possibly other molecular phenotypes, is likely to play a broader role in human biology and diseases.

Materials and methods

Targeted bisulfite sequencing with padlock probes.

Bisulfite padlock probe design, production and sequencing were previously described[16; 41]. Briefly, genomic DNA was extracted from peripheral blood of 22 pedigrees, and approximately 1 μg of genomic DNA was bisulfite converted with EZ-96 Zymo DNA Methylation-Gold kit (Zymo Research). Approximately 250ng of bisulfite converted genomic DNAs were mixed with normalized amount of genome-wide scale padlock probes and oligo suppressors. The padlock probes annealing to targets were polymerized and ligated resulting in circularized DNA. The bisulfite sequencing libraries were generated by library-free BSPP protocol as described[16]. Two-thirds of the circularized DNA of each captured reaction were directly amplified and barcoded. The bisulfite sequencing libraries were purified with AMPure magnetic beads (Agencourt), pooled in equimolar ratios, size selected with 6% polyacrylamide gel (Invitrogen), and sequenced by Illumina HiSeq2000 sequencer.

DNA methylation data

Bisulfite sequencing data were processed as described[13; 16]. Briefly, the bisulfite reads were mapped to the in silico bisulfite-converted human genome sequences (hg19) by bisReadMapper[16]. DNA methylation level at each CpG site with minimum 10x depth coverage was calculated at the level from 0-1.

Mid-parent offspring analysis

Mid-parent offspring (MPO) analysis was performed by linear regression analysis to estimate the heritability of DNA methylation at each CpG site. DNA methylation level of the offspring in each trio was compared against the mean DNA methylation level of the parents. In total, 79,604 variable CpGs (minimum standard deviation 0.1) shared in at least 80% of

subjects were analyzed. The CpG sites with minimum non-negative slope of the fitted line representing heritability (h^2) of 0.2 and with the minimum sample size (number of trio) of 10 were defined as heritable CpGs. The Benjamini-Hochberg method was used to correct for multiple testing errors.

Methylation quantitative trait loci

Methylation quantitative trait loci (mQTL) analysis was performed by PLINK (Purcell et al, 2007) to determine the association between DNA methylation frequency of variable CpG sites as described above and SNP genotypes called from methylation data (15,450 SNPs) of 96 subjects or imputed autosomal SNP genotypes (5,257,772 SNPs) of 57 subjects. SNP genotypes with a minor allele frequency (MAF) of at least 0.05 and with a Hardy-Weinberg Equilibrium (HWE) p-value > 0.001 were included in this analysis. We used least square linear regression, and the corresponding p-values were calculated for each CpG-SNP association pair within 1Mb. FDR was calculated by Benjamini-Hochberg multiple correction method to asses the significance of the CpG-SNP association. To deal with family structure, QFAM analysis was performed. 10,000 permutations were performed and p-values were empirically calculated as the fraction of permuted data test-statistic was larger than the non-permuted data test statistic. Additional analyses were performed on subsets of imputed SNPs including 618,580 SNPs present on Illumina 1M SNP array. The SNPs that showed strong correlation with DNA methylation were extracted and annotated significant QTL as *cis* if the SNP lay within 1 Mbs of the CpG site.

Allele-specific methylation

Allele-specific methylation (ASM) analysis was performed as described (Shoemaker et al. 2010). Briefly, we generated the 2 X 2 contingency table where the two columns containing the two alleles and the two rows containing the counts of methylated and un-methylated cytosines at CpG site(s) on the read containing heterozygous SNP(s). The p-value at each CpG site was calculated by Fisher’s exact test. We identified ASM if the p-value was less than 0.001 and the methylation frequency between the two alleles was greater than 0.2.

Schizophrenia-associated CpGs

To identify CpG sites associated with schizophrenia, we performed linear regression analysis on the selected CpG sites in the regions of known SNPs associated with the disease, including ANK3, CACNA1C, FKBP5, ITIH3, ITIH-4, and MIR137. Variable CpG sites (minimum STD 0.1) within 100kb upstream of transcription start site (TSS) and 100kb downstream of transcription end site (TES) were analyzed. The difference of methylation mean between schizophrenia affected and normal controls was tested by t-test. Schizophrenia-associated CpG at specific locus were identified if regression and t-test p-value were below 0.05.

Variation-SNP and variably methylated regions

We identified vSNPs and VMRs by performing association tests. Linear regression was performed on the variance of DNA methylation at each CpG site among individuals and the three genotype groups (AA, AB, BB) within 1Mb distance. The t-score of each CpG-SNP pair was calculated, and the false discovery rate was calculated by using different cutoff values for the test statistic values. To deal with the high rate of false positive signals, we required at least five adjacent CpG sites with maximal spacing 200 bp between CpGs showing consistent association with VMRs. We then grouped the overlapping or adjacent VMRs into clusters. We note that VMRs associated with different vSNPs could be partially overlapping, so they could be grouped into the same cluster.

Additional materials (available upon request)

Additional file 1: Contains all supplementary figures and supplementary tables 3-5 (Table S3 – S5)
Additional file 2: The table listing all heritable CpGs (Table S1)
Additional file 3: The table listing heritable non-SNP CpG clusters (Table S2)
Additional file 4: The table listing VMRs and their associated vSNPs (Table S6)
Additional file 5: The table listing VMR clusters (Table S7)

Abbreviations

MPO: mid-parent offspring analysis; mQTL: methylation quantitative trait loci; SNP: single nucleotide polymorphism; ASM: allele-specific methylation; VMRs: variably methylated regions, LD: linkage disequilibrium; TSS: transcription start site; TES: transcription end site; HWE: Hardy-Weinberg Equilibrium; BSPP: bisulfite padlock probes; RRBS: reduced representation bisulfite sequencing; FDR: false discovery rate; eQTL: expression quantitative trait loci; dsQTL: DNase I sensitivity quantitative trait loci; EWAS: epigenome-wide association study

Author' contributions

R.A.O and K.Z. designed the project. N.P., H.F. and T.W. performed experiments. N.P. K.V.E., S.D.J., J-H.S. performed data analysis, M.P.M.B. and R.S.K. were responsible for sample collection. N.P., R.A.O. and K.Z. wrote the manuscript with inputs from other co-authors.

Acknowledgements

This work is supported by GM097253 and HG003233 to K.Z., and DA028526 and MH090553 to R.A.O.

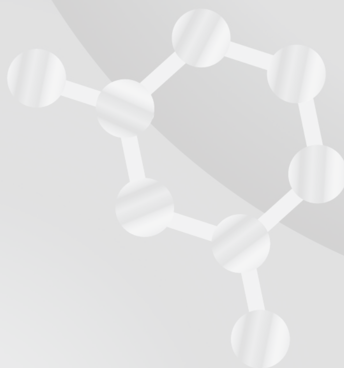
References

1. McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328, 235-239.
2. Zhang, Y., Rohde, C., Reinhardt, R., Voelcker-Rehage, C., and Jeltsch, A. (2009). Non-imprinted allele-specific DNA methylation on human autosomes. Genome Biol 10, R138.
3. Schalkwyk, L.C., Meaburn, E.L., Smith, R., Dempster, E.L., Jeffries, A.R., Davies, M.N., Plomin, R., and Mill, J. (2010). Allelic skewing of DNA methylation is widespread across the genome. American journal of human genetics 86, 196-212.
4. Tycko, B. (2010). Allele-specific DNA methylation: beyond imprinting. Human molecular genetics 19, R210-220.
5. Feinberg, A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature 447, 433-440.
6. Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet 12, 529-541.
7. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell 49, 359-367.
8. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., and Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome biology 12, R10.
9. Fraser, H.B., Lam, L.L., Neumann, S.M., and Kobor, M.S. (2012). Population-specificity of human DNA methylation. Genome biology 13, R8.
10. Gertz, J., Varley, K.E., Reddy, T.E., Bowling, K.M., Pauli, F., Parker, S.L., Kucera, K.S., Willard, H.F., and Myers, R.M. (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS genetics 7, e1002228.
11. van Eijk, K.R., de Jong, S., Boks, M.P., Langeveld, T., Colas, F., Veldink, J.H., de Kovel, C.G., Janson, E., Strengman, E., Langfelder, P., et al. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC genomics 13, 636.
12. Hellman, A., and Chess, A. (2010). Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics & chromatin 3, 11.
13. Shoemaker, R., Deng, J., Wang, W., and Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res 20, 883-889.
14. Fang, F., Hodges, E., Molaro, A., Dean, M., Hannon, G.J., and Smith, A.D. (2012). Genomic landscape of human allele-specific DNA methylation. Proceedings of the National Academy of Sciences of the United States of America 109, 7332-7337.
15. Schilling, E., El Chartouni, C., and Rehli, M. (2009). Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. Genome Res 19, 2028-2035.

16. Diep, D., Plongthongkum, N., Gore, A., Fung, H.L., Shoemaker, R., and Zhang, K. (2012). Library-free methylation sequencing with bisulfite padlock probes. Nature methods 9, 270-272.
17. Browning, R., Overmier, J.B., and Colombo, M. (2011). Delay activity in avian prefrontal cortex--sample code or reward code? Eur J Neurosci 33, 726-735.
18. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81, 559-575.
19. Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148, 816-831.
20. Greaves, I., Groszmann, M., Dennis, E.S., and Peacock, W.J. (2012). Trans-chromosomal methylation. Epigenetics : official journal of the DNA Methylation Society 7, 800-805.
21. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. Science 318, 1136-1140.
22. Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat Genet 43, 768-775.
23. Feinberg, A.P., and Irizarry, R.A. (2010). Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A 107 Suppl 1, 1757-1764.
24. Bastepe, M. (2007). The GNAS Locus: Quintessential Complex Gene Encoding Gsalpha, XLalphas, and other Imprinted Transcripts. Current genomics 8, 398-414.
25. Plagge, A., and Kelsey, G. (2006). Imprinting the Gnas locus. Cytogenetic and genome research 113, 178-187.
26. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31, 142-147.
27. Luykx, J.J., Bakker, S.C., Lentjes, E., Neeleman, M., Strengman, E., Mentink, L., Deyoung, J., de Jong, S., Sul, J.H., Eskin, E., et al. (2013). Genome-wide association study of monoamine metabolite levels in human cerebrospinal fluid. Mol Psychiatry.
28. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat Genet 39, 1217-1224.
29. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390-394.
30. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768-772.

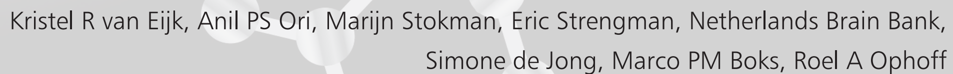
31. Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S., and Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. American journal of human genetics 86, 411-419.
32. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature 452, 423-428.
33. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics 98, 288-295.
34. Boyle, P., Clement, K., Gu, H., Smith, Z.D., Ziller, M., Fostel, J.L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., et al. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome biology 13, R92.
35. Teperino, R., Lempradl, A., and Pospisilik, J.A. (2013). Bridging epigenomics and complex disease: the basics. Cellular and molecular life sciences : CMLS.
36. Drong, A.W., Nicholson, G., Hedman, A.K., Meduri, E., Grundberg, E., Small, K.S., Shin, S.Y., Bell, J.T., Karpe, F., Soranzo, N., et al. (2013). The presence of methylation quantitative trait Loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One 8, e55923.
37. Shenker, N.S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M.A., Belvisi, M.G., Brown, R., Vineis, P., and Flanagan, J.M. (2013). Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Human molecular genetics 22, 843-851.
38. Joubert, B.R., Haberg, S.E., Nilsen, R.M., Wang, X., Vollset, S.E., Murphy, S.K., Huang, Z., Hoyo, C., Midttun, O., Cupul-Uicab, L.A., et al. (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environmental health perspectives 120, 1425-1431.
39. Ball, M.P., Li, J.B., Gao, Y., Lee, J.H., LeProust, E.M., Park, I.H., Xie, B., Daley, G.Q., and Church, G.M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nature biotechnology 27, 361-368.
40. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics 13, 484-492.
41. Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E.M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I.H., Yu, J., et al. (2009). Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 27, 353-360.
42. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.
43. Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490-495.

44. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
45. Liu, Y., Siegmund, K.D., Laird, P.W., and Berman, B.P. (2012). Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome biology 13, R61.
46. Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature 402, C47-52.
47. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science 298, 824-827.
48. Becskei, A., and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. Nature 405, 590-593.
49. Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature 463, 913-918.



Chapter 6

Interplay between miRNA-137 and DNA methylation in human brain highlights neurotransmitter-related pathways



Kristel R van Eijk, Anil PS Ori, Marijn Stokman, Eric Strengman, Netherlands Brain Bank, Simone de Jong, Marco PM Boks, Roel A Ophoff

In preparation

Abstract

Schizophrenia (SZ) is a severe brain disorder with a complex polygenetic architecture. A large GWAS meta-analysis identified microRNA-137 (miR-137) as a SZ susceptibility locus. MicroRNAs (miRNAs) and DNA methylation are both important players in the regulation of gene expression. There exists a feedback mechanism between miRNAs and DNA methylation, and the interplay between these two factors may provide more insight into gene regulation, and possibly into the etiology of schizophrenia. We examined the relationship of miR-137 expression and genome-wide DNA methylation levels in human *post-mortem* brain and found suggestive associations between miR-137 and promoter methylation of three genes, of which two are involved in neurotransmitter-related pathways. These findings might help elucidate the biological relevance of miR-137 in schizophrenia susceptibility and provide additional insight into the pathophysiology of this disorder.

Introduction

Schizophrenia (SZ) is a severe brain disorder with a complex underlying polygenetic architecture. Large-scale genome-wide efforts have identified a small number of SZ susceptibility loci with one of the strongest findings located at the noncoding gene, microRNA-137 (miR-137)[1; 2]. Further work has shown that individuals homozygous for the risk allele (TT) at this locus show reduced expression of miR-137 in human dorsolateral prefrontal cortex, a brain region implicated in the pathophysiology of SZ[3]. In addition, TCF4, a target of miR-137 was also associated with SZ and under genetic control of the SZ risk allele that is in close genomic proximity of the miR-137 gene. MiR-137 itself is thought to play a role in early neurodevelopment and neurogenesis[4; 5]. Nonetheless, the exact role of miR-137 in the etiology of the disorder remains largely unknown. Further investigation of the function of miR-137 and its regulation in the human brain might shed light on the underlying biological mechanisms of schizophrenia.

MicroRNAs (miRNAs) are endogenous small non-coding RNA molecules crucial for cell function. They are implicated in a plethora of physiological and pathophysiological processes, including brain development, function and disease[6-8]. miRNAs regulate the expression of mRNA transcripts at the posttranscriptional level through complementary binding of target sequences and induce transcript degradation or hinder protein translation. Interestingly, a subset of microRNAs has been shown to control the expression of key players of the epigenetic machinery[9].

DNA methylation, one of the most studied epigenetic mechanisms is associated with gene expression. It is a biochemical process that transfers a methyl group to the 5-position of cytosine residues in CpG dinucleotides in DNA, although non-CpG methylation can also occur in mammals[10]. Methylation is catalyzed and maintained by DNA methyltransferases (DNMTs). Promoter regions and transcription start sites of genes are often characterized by the presence of CpG islands (CGIs) harboring a high frequency of CpG dinucleotides; CpG islands are also found elsewhere in the genome including in genic regions. CpG islands in promoter regions of actively transcribed genes are mostly unmethylated[11; 12]. Silencing of the promoter can be achieved through CpG methylation, thereby regulating gene expression with possible effects on traits and phenotypes. Indeed, DNA methylation has been implicated in many biological processes and diseases, including brain development, plasticity and psychiatric disorders[13]. Importantly, a large fraction of miRNAs is postulated to be under epigenetic control. That is, about 50% of miRNA gene promoters are predicted to contain CpG islands[14].

Epigenetic regulation and miRNAs control genome-wide gene expression patterns transcriptionally and post-transcriptionally[9]. A subset of miRNAs target important genes of the epigenetic pathway, such as DNA methyltransferases, polycomb group (PcG) proteins, and histone modifiers. At the same time, the expression of miRNAs can be regulated epigenetically as shown in a number of studies[5; 9]. For example, expression of miR-137 in adult neural stem cells (aNSC) is epigenetically regulated through MeCP2, a methyl-CpG-binding protein. Furthermore, miR-137 represses the expression of Ezh2, a histone methyltransferase and PcG protein. miR-137 is thought to be an important player in neurodevelopment as it modulates proliferation, differentiation, and maturation of aNSCs both *in vitro* and *in vivo*[4; 5; 15]. This interplay between the epigenetic machinery and miRNAs forms an extra dimension in the regulation of gene expression and cellular function. Disruption in this mechanism might contribute to disease processes. More insights into the cross talk between miRNAs and epigenetics will therefore provide key information underlying and helping understand this biological phenomenon.

This study aims to investigate the interplay between miR-137, a gene recently identified to be associated with schizophrenia, and DNA methylation at gene promoter regions. Expression of microRNAs and DNA methylation of gene promoters have been shown to be dynamically regulated temporally and across regions of the mammalian nervous system, which suggests a pivotal role in brain functioning[16-21]. In addition, both have been implicated in brain dysfunction and diseases, more specifically schizophrenia[2; 22-25]. For this reason, we set out to investigate the interplay between the established schizophrenia susceptibility locus, miR-137, and genome-wide DNA methylation at gene promoter regions in human *post-mortem* brain.

Material and Methods

Subjects and tissue collection

Post-mortem brain tissue from 61 subjects (44 non-demented controls, 6 schizophrenia patients and 11 bipolar disorder patients) was obtained from the Netherlands Brain Bank (NBB). The brain tissue used in this study was collected by the NBB between 1989 and 2008. The NBB examined medical records from all subjects to confirm psychiatric diagnoses. From these individuals, a total of 235 brain samples, spanning multiple regions across the brain, were used. Brain tissue was dissected according to a standardized protocol and individual samples were subsequently flash frozen in liquid nitrogen and stored at -80°C. Samples were transported on dry ice and frozen at -80°C upon arrival. For each sample, technical variables such as pH and post mortem interval were recorded. An overview of subject and sample characteristics is presented in **Table 1.**

Table 1. Individual characteristics

	N	Gender (M/F)	Age ± SD	PMI ± SD (min)	pH ± SD
Controls	44	21/23	80 ± 9	597 ± 433	6.6 ± 0.2
SCZ	6	1/5	76 ± 10	460 ± 157	6.8 ± 0.7
BPD	11	8/3	74.8 ± 9	380 ± 82	6.4 ± 0.2

In this table, subject characteristics are presented per disease status (non-demented, schizophrenia (SCZ) and bipolar disorder (BPD)). The columns contain information on male/female ratio, age and the technical variables post-mortem interval (PMI) in minutes, and pH (± SD).

Tissue processing – DNA isolation

From all samples, approximately 25 mg of tissue was taken for extraction of DNA using Qiagen's Allprep DNA/RNA/Protein isolation kit. Brain samples were homogenized using Covaris S2 system with the following settings; duty cycle = 20%, intensity = 5, burst cycle = 100 and time = 10 seconds. From this lysate, DNA was extracted according to manufacturer's instructions. Total DNA concentrations were assessed using the Quant-iT PicoGreen RNA Assay Kit.

Tissue processing – total RNA isolation

Tissue fragments of roughly 25 mg were cut using a scalpel on a cooled glass surface. Samples were homogenized using the TissueLyser II (Qiagen) at a frequency of 23 Hz for a period of 3.5 minutes. Total RNA enriched for miRNA was extracted using Qiagen's MiRNeasy Mini Kit according to the manufacturer's protocol. 1-Bromo-3-chloropropane was used instead of chloroform as a phase separation reagent. The RNase-Free DNase Set was used to remove genomic DNA. As a quality control measure, sample RNA Integrity Number (RIN) values were determined using the 2100 Bioanalyzer (Agilent RNA 6000 Nano Kit). Total RNA concentrations were assessed using the Quant-iT RiboGreen RNA Assay Kit.

Pre-processing methylation data

Methylation data from brain tissue (n = 235) of 63 individuals was obtained with Illumina HumanMethylation27 beadchips. The assay detects methylation status at bisulfite converted CpG sites, by means of probes designed for either methylated or unmethylated sequence. Gender discrepancies were assessed by hierarchical clustering of X-chromosomal probes. Samples showing background intensity of >30% for staining, hybridization, and extension, or >35% for bisulfite conversion, as opposed to the high intensity values of the Illumina control probes were excluded from further analysis. In total, 2,581 probes with low detection

p-values (>0.01 for >1% of the samples) were excluded. There were no samples with low detection p-values (>0.01 for >1% of probes). Potential array outliers were removed in an unbiased fashion. We used the SampleNetwork R function package[26] to calculate the Inter-array-based sample connectivity score (Z.k). We removed samples with a Z.k. value below -3 since their connectivity is 3 standard deviations below the mean value[27]. Batch effects of plate, array and position were removed using ComBat[27]. Both channels of the methylation array were quantile normalized independently. Beta values of a probe (ranging from zero to one) were calculated by dividing the methylated signal by the sum of the methylated and unmethylated signal. Methylation probes were mapped to the human genome using the UCSC browser, and probes that lacked a unique mapping were removed. In addition, probes that contained a known single nucleotide polymorphism (SNP), based on Hapmap release 27 and minor allele frequency (MAF) of > 0.01, were excluded. Finally, all probes on the X and Y chromosome were also removed. After these procedures, 22,705 probes were left for further analysis. To reduce heteroscedasticity for highly methylated or unmethylated CpGs and to ensure more normally distributed methylation levels for further analysis M values were calculated[28]. Taking the log2 ratio of the intensities of the methylated probe versus the unmethylated probe resulted in M values.

Expression of microRNA-137

Complementary DNA (cDNA) was generated from total RNA enriched for miRNA using the TaqMan MicroRNA Reverse Transcription Kit and specific stem-loop primers for miR-137 and small nuclear RNA 6B (RNU6B). During the reverse transcription reaction, 15 μL of cDNA was obtained per sample using 10 ng of total RNA (2 ng/μl), 3.0 μl 5X TaqMan Small RNA Assay, 0.15 μl 100mM dNTPs, 1.0 μl MultiScribe Reverse Transcriptase (50 U/μl), 1.5 μl 10X RT buffer and 0.19 μl RNase inhibitor (20 U/μl) according to the manufacturer's protocol. Reactions were performed on a Bio-Rad C1000 Thermal Cycler for 40 cycles at 16°C for 30 minutes, 42°C for 30 minutes and 85°C for 5 minutes. RNU6B was used as an endogenous control. Showing comparatively stable expression across tissue types, RNU6B is frequently used in miRNA quantitative RT-PCR (qRT-PCR) studies to correct for sample fluctuations[29]. To evaluate background signal, one no template control (NTC) was included per miRNA assay during the RT-reaction.

The quantitative PCR (qPCR) reactions were performed in 384-well plates using the Applied Biosystems 7900 HT Real-Time PCR System, with a total volume of 10 μL per reaction. Each reaction volume contained 0.67 μL of cDNA, 0.50 μl 20X TaqMan Small RNA Assay and 5.0 μl 2X TaqMan Universal PCR Master Mix II, No UNG consistent with the manufacturer's protocol. Fifty amplification cycles, at 95°C for 15 seconds and 60°C at 60 seconds, were used to achieve a plateau phase for every miRNA assay to facilitate Ct measurements. The

qRT-PCR data was acquired using Sequence Detection Systems software version 2.4. Samples were run in triplicate per gene analyzed. The generated data was subsequently analyzed in DataAssist version 3.01. Background signal was evaluated by assessing the expression measured in the NTCs. Samples with a cycle threshold (Ct) value exceeding 40 were excluded from further analysis. Outliers among the technical replicates were removed using a refined Grubbs' outlier test. Relative miRNA expression was determined by calculating the ΔCt value using RNU6B as internal control. Samples with a ΔCt value exceeding ± 2SD per brain region were excluded from further analysis. Next, the ddCt method was used to calculate miR-137 expression relative to the average expression across all samples. Finally, natural log 2^-ddCt values were calculated and used for subsequent analyses.

Neuronal proportion correction

A Cell EpigenoType Specific (CETS) model was used to remove cell type heterogeneity bias from DNA methylation data per brain sample[30]. Briefly, neuronal proportions were estimated from neuron and glia reference profiles using 461 disease non-specific (age > 40) cell type epigenetic markers that were available on Illumina's 27K methylation array. Performance of these 461 markers was analysed using a dilution series ranging the fractions of neurons per sample. These subset of markers yielded accurate predictions of neurons and glia proportions in this dilution series. Thus, sample-specific neuronal proportions were used in subsequent analyses to account for cell-type derived variation in methylation data.

Investigating association between miR-137 and DNA methylation

Regional analysis

Linear regression models (limma package in R[31]) were used to relate 22,705 methylation values on miR-137 expression, disease status, gender, age[32], RIN values, braak stage[33], and neuronal proportion (NP). To study regional effects, analyses were carried out per brain region. FDR correction at the 5% level was applied to correct for multiple testing.

Global analysis

To increase the power of our association analysis, a linear mixed effects (LME) model (nlme package in R[34]) was conducted with methylation as outcome and miR-137 expression as predictor. Disease status, gender, age, RIN, braak stage, and NP were taken as fixed effects, and brain region and individual ID (because an individual can have several brain regions), as random effects. P-values were FDR corrected at 5% level.

Pathway analysis

Using corresponding genes annotated to the top methylation probes (p-value < 0.01), pathway analyses were conducted to investigate overrepresentation of these genes in biological processes. Several pathway tools were used including Panther[35], Webgestalt[36; 37], and David[38]. The genes were tested against the background of the Illumina 27k array, since genes can be covered by more than one probe, and Bonferroni correction was applied to correct for multiple testing.

Results

Expression of microRNA-137

Figure 1 shows the expression level of miR-137 across regions of the human brain. The highest expression was detected in the basal ganglia, cortex, and limbic system. On the contrary, cerebellum and choroid plexus exhibit the lowest expression of miR-137. The

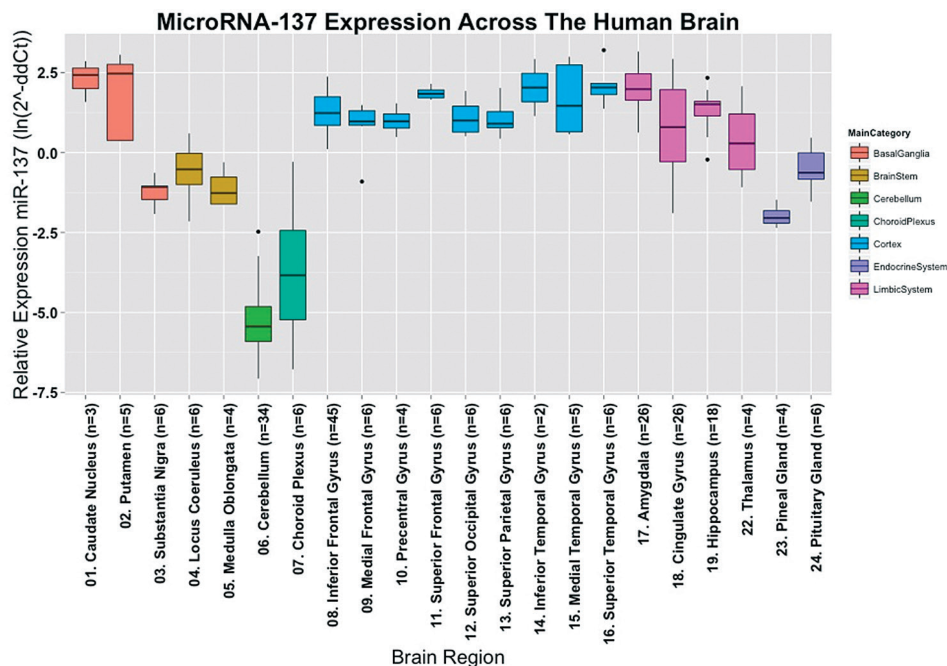


Figure 1: MicroRNA-137 expression across human brain regions

A box-and-whisker plot showing the expression of miR-137 in natural log (2^-ddCt) (see method) on the y-axis and brain regions on the x-axis. Expression levels are relative with respect to the mean expression value of all brain regions. Brain tissues are color-coded according to their brain system. The number of samples per brain region is indicated within brackets. The box represents the 25th-75th percentile and the line marks the median of the relative expression levels. The whiskers are drawn from the 10th-90th percentile.

number of samples per region differs substantially, although this does not influence the degree of variation.

Linear regression with miRNA as a predictor to determine DNA methylation shows no significant association

To account for cell-type heterogeneity in methylation data sample-specific neuronal proportions were calculated. **Figure 2** shows the NP across regions of the brain. A number of brain categories have similar neuron-to-glia ratios across the brain regions. For example, the variation of NP in the locus coeruleus and medulla oblongata, both located in the brain stem, is small. Cortical regions, on the contrary, exhibit much larger variation, both between and within regions. Proportion of neurons affected our methylation data substantially and therefore we incorporated NP as covariate in our further analyses.

We used linear models per brain region to identify associations between miR-137 and DNA methylation. Only brain regions with over 15 samples were included: hippocampus, inferior

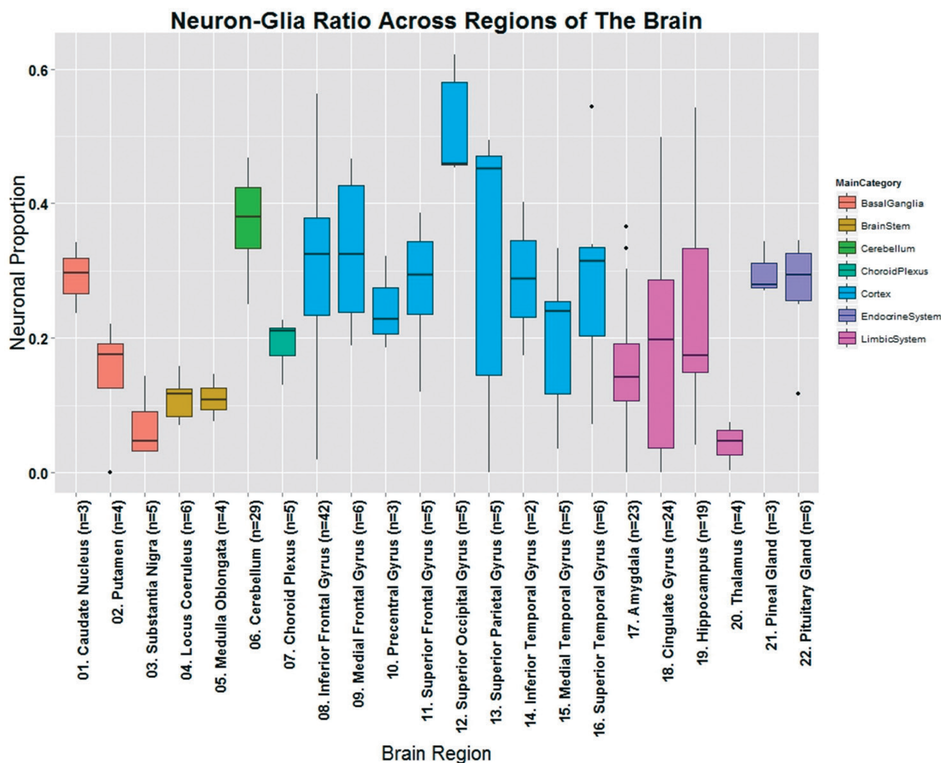


Figure 2: Neuron-to-glia ratio across human brain regions

The x-axis in this boxplot represents brain regions, color-coded according to the brain system. The y-axis shows neuronal proportions, which represents the neuron-to-glia ratios. The number of samples per brain region is indicated within brackets. The box represents the 25-75th percentile and the line marks the median of the neuron:glia ratio.

frontal cortex, cingulate cortex, and amygdala. After FDR correction we find no significant association between methylation and miR-137 in any of the four regions (results not shown).

Linear mixed effects model to study association between miR-137 and DNA methylation

To increase power for finding associations between miR-137 and DNA methylation, we utilized a linear mixed effects model, with brain region and individual ID as random effect. After 5% FDR correction for 22,705 tests, we find no significant associations between miR-137 and CpG methylation. However, as shown in the QQ plot in **Figure 3**, there are seven CpG sites that deviate from the normal distribution, of which three have an FDR q-value of 0.16. The CpG sites are located in regulatory regions of the genes *HTR2A*, *OGDHL*, and *GSPT1*; results are shown in **Table 1** and **Figure 4**. Methylation of *HTR2A*, a serotonin receptor is negatively correlated with expression levels of miR-137. Increased levels of miR-137 correspond to decreased methylation of HTR2A or vice versa with a beta coefficient of -0.09 (p-value=1.09e-05, FDR = 0.16). CpG DNA methylation levels flanking the promoter region of *OGDHL*, the Oxoglutarate Dehydrogenase-Like gene that is involved in glutamate synthesis, is positively associated with miR-137 transcription levels with a beta coefficient of 0.15 (p-value=1.9e-05, FDR=0.16). Increased expression of miR-137 correlates with increased methylation of *OGDHL*. The third CpG site that is most strongly associated with *miR-137* levels is found within the promoter region of *GSPT1,* the human homolog of a gene that is essential in the G1- to S-phase transition in yeast but with unknown function in humans (p-value=2.2e-05, FDR=0.16).

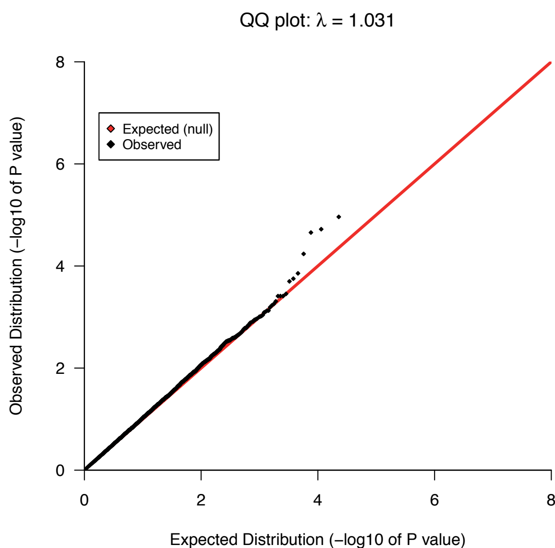


Figure 3: QQ-plot of the LME model results.

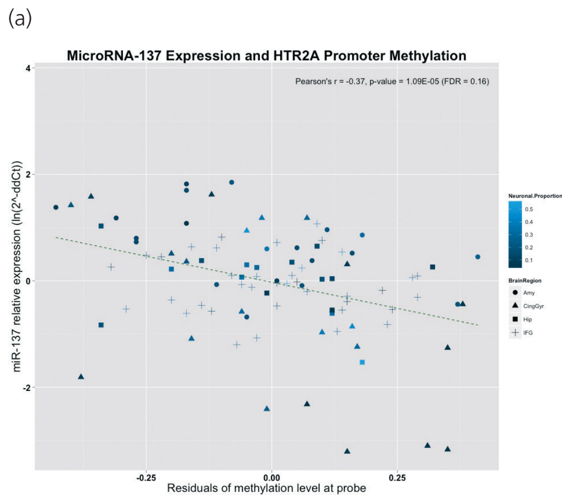
The observed versus expected p-values of LME result are shown. Methylation levels were regressed against miR-137 expression. Disease status, gender, age, RIN, braak stage, and NP were taken as fixed effects, and brain region and individual ID as random effects.

Table 1: Results from LME model analysis

CpG	Chr	Locus	Gene	Dist. to TSS	CPG ISLAND	Gene Strand	Coef	p-value	q-value
cg02250787	13	46368990	*HTR2A*	814	False	-	-0.09	1.09E-05	0.16
cg24356544	10	50640869	*OGDHL*	494	Shore	-	0.15	1.90E-05	0.16
cg21875234	16	11916906	*GSPT1*	403	True	-	-0.05	2.21E-05	0.16

The strongest associations between DNA methylation and miR-137 are shown. Columns represent name, chromosome and locus of the CpG site; Gene symbol, distance of the methylation probe to transcription start site, location of CpG site with respect to CGIs, Gene strand, beta coefficient, p-value, and FDR corrected p-value (q-value).

(a)



(b)

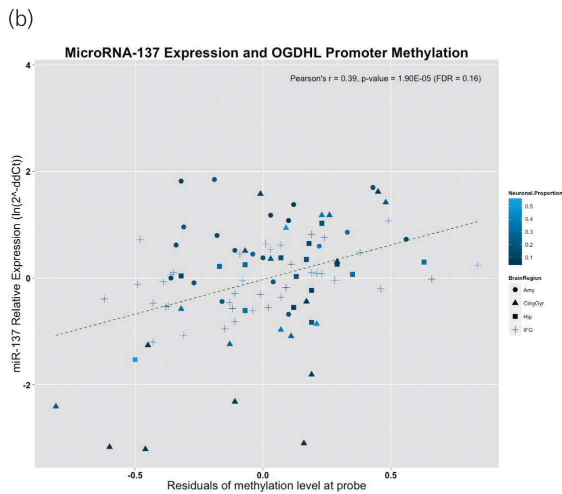


Figure 4: Correlations between miR-137 and HTR2A methylation.
Expression levels of miR-137, denoted as the natural log of 2^-ddCt, are shown on the y-axis and methylation of the CpG site on the x-axis. Data points are distinguished between brain region (shape) and neuronal proportion (color). a, shows the negative association of miR-137 with methylation of HTR2A, and b) presents the positive association between miR-137 and methylation of OGDHL.

Pathway analysis

The genes corresponding to CpG sites with DNA methylation levels associated with miRNA at (uncorrected) $p<0.01$, were not overrepresented in a pathway after adjustment for multiple testing.

Discussion

The aim of our study was to examine the relationship between expression levels of a microRNA that has been implicated in schizophrenia (miR-137) and genome-wide DNA methylation levels in adult human *post-mortem* brain. We examined expression levels of miR-137 across 24 different brain regions and observed a wide spectrum of differential expression between different brain structures. We further applied linear regression and mixed effects models to investigate the association between miR-137 and DNA methylation levels genome-wide. The most significant findings of this analysis suggest a connection between miR-137 and genes involved in serotonin and glutamate pathways.

The expression of miR-137 was examined across 24 different brain regions derived from 61 individuals. As shown in **Figure 1**, brain structures from similar regions mostly show similar expression levels of miR-137. Our results show that the largest variation of miR-137 expression exists *between* brain regions (intra-individually) and not within brain systems across different individuals, which suggests possible shared miR-137 functionality within regions of a brain system.

We observe high expression levels of miR-137 limbic system and cortical regions. This finding coincides with previous studies that found similar high expression of miR-137 in these regions of the human brain[3; 4; 39]. The highest expression levels of this microRNA in our study are observed in the basal ganglia. This in contrast with a previous study by Guella and colleagues who report relatively low expression of miR-137 in caudate nucleus and putamen of the basal ganglia[3]. In our study, expression of miR-137 in substantia nigra is slightly reduced compared to caudate nucleaus and putamen although in the same system. This could be explained by anatomical differences. Both putamen and caudate nucleus are part of the corpus striatum, while substantia nigra is part of the Mensencephalon[40]. Although part of the same dopaminergic system, these subregions are structurally and functionally different, which coincides with the different expression levels observed for miR-137. In the endocrine system and brain stem, miR-137 shows relatively low expression. This lack of miR-137 expression in brain stem has been reported recently[39], although there is not much known of expression of miR-137 in the endocrine system. We observe very low to absent expression of miR-137 in the cerebellum, similar to what has been reported before[3; 39]. We also observe very low expression levels of miR-137 in choroid plexus. Differences in expression

of miR-137 across different adult brain regions suggest miR-137 operates in a region-specific manner. Interestingly, the regions with the highest miR-137 expression levels are also the ones that show abnormalities in schizophrenia patients[41]. Especially cortical aberrations have been associated with schizophrenia, showing decreased volumes in patients. In addition, the cortical information processing is functionally abnormal in first-episode and chronic schizophrenia[41]. The basal ganglia, the region with the highest miR-137 expression in our study, is involved in learning and reward system and has been implicated in schizophrenia[42]. Furthermore, hippocampus and amygdala volumes are decreased in the onset of disease. It is possible that miR-137, which is involved in proliferation and maturation of neurons, is malfunctioning in these regions, prohibiting development of neurons. On the other hand, regions with low miR-137 expression, such as cerebellum, have been associated with schizophrenia as well, although these findings are sometimes less pronounced[43; 44]. Based on our findings, a possible role of miR-137 in cerebellum seems less likely, although we cannot exclude involvement of this microRNA in developmental stages of this brain region.

Using a linear regression model to investigate the relationship between miR-137 and DNA methylation per brain region revealed no genome-wide significant associations. QQ plots are deflated, suggesting over-correction and/or lack of power. Moreover, pathway analyses using the top 1% results (genes corresponding to methylation probes associated with miR-137 with a p-value < 0.01) yielded no further insight of enrichment of specific molecular mechanisms controlled by miR-137. The relatively low number of samples per brain region may be a critical issue if the effects sizes are small; this may have resulted in lack of power. To increase statistical power, we also applied a linear mixed effects model with brain region and individual ID as random effect. Although the QQ plot shows no signs of inflation of deflation (lambda $= 1.03$), no genome-wide significant associations were found between miR-137 transcript levels and DNA methylation after correction for multiple testing. We observed three loci with relatively strong association (q-values of 0.16) with miR-137. Our strongest observation is the correlation between miR-137 expression levels and DNA methylation at *HTR2A,* a serotonin receptor important in neuronal processes with known genetic and epigenetic regulation of its promoter[45-50]. Epigenetic down-regulation by DNA hypermethylation of this gene has previously been implicated in age of onset of schizophrenia and bipolar in brain[50] and saliva[46]. This gene has also been proposed as candidate for obsessive compulsory disorder, another psychiatric disorder[51]. Moreover, *HTR2A* is the main target for antipsychotic drugs that could cause promoter hypomethylation of this gene, enabling expression recovery[50]. Based on these previous and our current findings, it is tempting to suggest that in schizophrenia patients decreased miR-137 levels induces methylation of this serotonin transporter, which in turn might lead to gene repression thereby contributing to disease. An important question that remains is whether this correlation arises through

direct miR-137 targeting of players of the epigenetic machinery. Our findings suggest a link between a well-established schizophrenia susceptibility locus, miR-137, to the serotonin pathway, which may provide additional insight into the pathophysiology of this disorder. Further studies are needed to confirm this finding and to decipher the underlying mechanism through which miR-137 and DNA methylation of *HTR2A* are associated.

The second-best finding is the association between miR-137 expression level and DNA methylation at the gene 2-oxoglutarate dehydrogenase-like (*OGDHL*). This gene is expressed in the brain where it is involved in glutamate synthesis, an important neurotransmitter[52]. Specifically, glutamate is the major excitatory neurotransmitter in the central nervous system (CNS), and directly and indirectly involved in most aspects of normal brain functioning[53; 54]. It is thought that glutamate neurotransmission is involved in the pathogenesis of many CNS diseases[54], including neuropsychiatric disorders such as schizophrenia[55-57]. Even though the evidence of these as well as our finding is suggestive, these results may highlight the biological relevance of miR-137 in schizophrenia susceptibility. Methylation of this gene increases with higher expression levels of miR-137. The exact function of *OGDHL* and its potential role in schizophrenia is not known. Our findings need further validation before in-depth molecular studies should ensue. The third finding highlights a gene, *GSPT1*, the human homolog of a gene essential for the G1- to S-phase transition in the yeast cell cycle but with unknown function in neuronal cells[58].

Even though the findings linking miR-137 with neurotransmitter pathways in adult human brain are very appealing, we acknowledge that the modest sample size has affected the power of our study. Availability of brain tissue is limited since *post-mortem* tissue is harder to obtain in large numbers compared to whole blood samples. Moreover, the quality of *post-mortem* brain tissue can be negatively affected by a number of factors such as post-mortem delay, storage temperature, and pH (among others). Although there is not much known of these effects on DNA methylation, pre- and post-mortem circumstances may impact molecular mechanisms[59]. In our study, we examined the effects of a number of known parameters such as age, gender, Braak stage, and included these in the model when confounding. The brain samples were obtained from elderly people ranging from age 62 to 93. It is a well-established fact that there is a significant age effect on DNA methylation[19; 20; 32; 60; 61], therefore we corrected for age in our analyses. More importantly, however, is the question whether study of *post-mortem* brain tissues will be informative for a brain disorder such as schizophrenia, which is thought to be neurodevelopmental in origin. If miR-137 plays an important role in neurodevelopment and schizophrenia susceptibility, what would we expect from studying adult *post-mortem* human brain? To date, the role of miRNA-137 in brain is not fully understood. Therefore, using adult brain serves as a good starting point to

examine miR-137 expression in brain and its association with DNA methylation. The dataset used in this study consists of many brain regions, which facilitates our knowledge of miR-137 expression across different regions and structures in the brain. In *post-mortem* brain tissue of patients there is also the concern that (long-term) use of psychoactive medication may have influenced methylation levels, which is very difficult to control for[62-64]. For obtaining DNA methylation levels, we used the Illumina HumanMethylation27 Beadchip[65]. This chip interrogates 27,000 CpG sites corresponding to roughly 14,000 genes and includes mainly CpGs located in CGIs. Therefore we lack information about CpGs in a large part of the genome resulting in a biased view of the methylome. Technological advances have been made to increase the number of CpGs simultaneously tested in a single assay, either by array-based methods (e.g. the Illumina 450K arrays[66]) or by high-throughput sequencing technology[67]. Another important issue is that methylation arrays, or any method using bisulfite conversion, cannot distinguish between DNA methylation and 5' hydroxymethylation[68]. 5' hydroxymethylation is thought to be an intermediate form between methylation and de-methylation, and almost half of the methylated cytosine in the brain is in this form[5]. It is thought to play a role in epigenetic reprogramming, demethylation, and regulation of gene expression[69; 70]. We recommend that future studies should include a larger number of samples to improve power and that next-generation array- or sequencing-based tools to be used to obtain a higher resolution genomic view of the methylome profiles.
Our most significant finding implicates a known antipsychotic drug target, *HTR2A*[50]. Since gene expression and epigenetic marks are cell-type specific[71-73], we corrected the methylation data for neuronal proportion using the CETS method[30] thereby decreasing the inflated uncorrected results. Although this NP correction proved to be relevant, to date, there is no method available for implementation in microRNA data.

In summary, we performed the first integrated analysis of miR-137 levels and genome-wide DNA methylation levels in human brain tissues. Our findings, suggesting a link between a known schizophrenia risk factor, miR-137, and DNA methylation at *HTR2A* and *OGDHL* need further validation. Study of the exact mechanism underlying the cross talk between miR-137 and the epigenetic control of these promoter regions might provide new insights into brain function and, possibly, schizophrenia pathophysiology.

Acknowledgements

The authors thank Djie Thung Twan and Jessica van Setten for bioinformatics support. We thank SARA Computing and Networking Services (www.surfsara.nl) for their support in using the Lisa Compute Cluster, which facilitates the Genetic Cluster Computer. This study was funded by grants from the National Institutes of Health (NIH).

References

1. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45, 1150-1159.
2. Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.Y., Duan, J., Ophoff, R.A., Andreassen, O.A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43, 969-976.
3. Guella, I., Sequeira, A., Rollins, B., Morgan, L., Torri, F., van Erp, T.G., Myers, R.M., Barchas, J.D., Schatzberg, A.F., Watson, S.J., et al. (2013). Analysis of miR-137 expression and rs1625579 in dorsolateral prefrontal cortex. J Psychiatr Res 47, 1215-1221.
4. Smrt, R.D., Szulwach, K.E., Pfeiffer, R.L., Li, X., Guo, W., Pathania, M., Teng, Z.Q., Luo, Y., Peng, J., Bordey, A., et al. (2010). MicroRNA miR-137 regulates neuronal maturation by targeting ubiquitin ligase mind bomb-1. Stem Cells 28, 1060-1070.
5. Szulwach, K.E., Li, X., Smrt, R.D., Li, Y., Luo, Y., Lin, L., Santistevan, N.J., Li, W., Zhao, X., and Jin, P. (2010). Cross talk between microRNA and epigenetic regulation in adult neurogenesis. J Cell Biol 189, 127-141.
6. Barry, G. (2014). Integrating the roles of long and small non-coding RNA in brain function and disease. Molecular psychiatry.
7. Eacker, S.M., Dawson, T.M., and Dawson, V.L. (2013). The interplay of microRNA and neuronal activity in health and disease. Front Cell Neurosci 7, 136.
8. Sun, A.X., Crabtree, G.R., and Yoo, A.S. (2013). MicroRNAs: regulators of neuronal fate. Curr Opin Cell Biol 25, 215-221.
9. Sato, F., Tsuchiya, S., Meltzer, S.J., and Shimizu, K. (2011). MicroRNAs and epigenetics. Febs J 278, 1598-1609.
10. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.
11. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13, 484-492.
12. Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Gene Dev 25, 1010-1022.
13. Grayson, D.R., and Guidotti, A. (2013). The dynamics of DNA methylation in schizophrenia and related psychiatric disorders. Neuropsychopharmacology 38, 138-166.
14. Weber, B., Stresemann, C., Brueckner, B., and Lyko, F. (2007). Methylation of human microRNA genes in normal and neoplastic cells. Cell Cycle 6, 1001-1005.
15. Sun, G., Ye, P., Murai, K., Lang, M.F., Li, S., Zhang, H., Li, W., Fu, C., Yin, J., Wang, A., et al. (2011). miR-137 forms a regulatory loop with nuclear receptor TLX and LSD1 in neural stem cells. Nat Commun 2, 529.

16. Ziats, M.N., and Rennert, O.M. (2013). Identification of differentially expressed microRNAs across the developing human brain. Molecular psychiatry.
17. Numata, S., Ye, T., Hyde, T.M., Guitart-Navarro, X., Tao, R., Wininger, M., Colantuoni, C., Weinberger, D.R., Kleinman, J.E., and Lipska, B.K. (2012). DNA methylation signatures in development and aging of the human prefrontal cortex. American journal of human genetics 90, 260-272.
18. Namihira, M., and Nakashima, K. (2013). Mechanisms of astrocytogenesis in the mammalian brain. Curr Opin Neurobiol 23, 921-927.
19. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol 14, R115.
20. Hernandez, D.G., Nalls, M.A., Gibbs, J.R., Arepalli, S., van der Brug, M., Chong, S., Moore, M., Longo, D.L., Cookson, M.R., Traynor, B.J., et al. (2011). Distinct DNA methylation changes highly correlated with chronological age in the human brain. Hum Mol Genet 20, 1164-1172.
21. Beveridge, N.J., Santarelli, D.M., Wang, X., Tooney, P.A., Webster, M.J., Weickert, C.S., and Cairns, M.J. (2014). Maturation of the Human Dorsolateral Prefrontal Cortex Coincides With a Dynamic Shift in MicroRNA Expression. Schizophr Bull 40, 399-409.
22. Wockner, L.F., Noble, E.P., Lawford, B.R., Young, R.M., Morris, C.P., Whitehall, V.L., and Voisey, J. (2014). Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. Transl Psychiatry 4, e339.
23. Smalheiser, N.R., Lugli, G., Zhang, H., Rizavi, H., Cook, E.H., and Dwivedi, Y. (2014). Expression of microRNAs and Other Small RNAs in Prefrontal Cortex in Schizophrenia, Bipolar Disorder and Depressed Subjects. PLoS One 9, e86469.
24. Miller, B.H., Zeier, Z., Xi, L., Lanz, T.A., Deng, S., Strathmann, J., Willoughby, D., Kenny, P.J., Elsworth, J.D., Lawrence, M.S., et al. (2012). MicroRNA-132 dysregulation in schizophrenia has implications for both neurodevelopment and adult brain function. Proc Natl Acad Sci U S A 109, 3125-3130.
25. Kato, T., and Iwamoto, K. (2014). Comprehensive DNA methylation and hydroxymethylation analysis in the human brain and its implication in mental disorders. Neuropharmacology.
26. Oldham, M.C., P. Langfelder, S. Horvath. (2011). Sample Networks for Enhancing Cluster Analysis of Genomic Data: Application to Huntington's Disease.
27. Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D.H. (2008). Functional organization of the transcriptome in human brain. Nat Neurosci 11, 1271-1282.
28. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L., and Lin, S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11, 587.
29. Peltier, H.J., and Latham, G.J. (2008). Normalization of microRNA expression levels in quantitative RT-PCR assays: identification of suitable reference RNA targets in normal and cancerous human solid tissues. Rna 14, 844-852.

30. Guintivano, J., Aryee, M.J., and Kaminsky, Z.A. (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics 8, 290-302.

31. Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3, Article3.

32. Horvath, S., Zhang, Y., Langfelder, P., Kahn, R.S., Boks, M.P., van Eijk, K., van den Berg, L.H., and Ophoff, R.A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol 13, R97.

33. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., and Del Tredici, K. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol 112, 389-404.

34. Pinheiro JC, B.D., DebRoy S, Sarkar D. R Development Core Team. nlme: Linear and Nonlinear Mixed Effects Models. (

35. Mi, H., and Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol 563, 123-140.

36. Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 41, W77-83.

37. Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33, W741-748.

38. Huang da, W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8, R183.

39. Willemsen, M.H., Valles, A., Kirkels, L.A., Mastebroek, M., Olde Loohuis, N., Kos, A., Wissink-Lindhout, W.M., de Brouwer, A.P., Nillesen, W.M., Pfundt, R., et al. (2011). Chromosome 1p21.3 microdeletions comprising DPYD and MIR137 are associated with intellectual disability. J Med Genet 48, 810-818.

40. Lanciego, J.L., Luquin, N., and Obeso, J.A. (2012). Functional neuroanatomy of the basal ganglia. Cold Spring Harb Perspect Med 2, a009621.

41. Meyer-Lindenberg, A. (2010). From maps to mechanisms through neuroimaging of schizophrenia. Nature 468, 194-202.

42. Strauss, G.P., Waltz, J.A., and Gold, J.M. (2014). A review of reward processing and motivational impairment in schizophrenia. Schizophr Bull 40 Suppl 2, S107-116.

43. Turner, R., and Schiavetto, A. (2004). The cerebellum in schizophrenia: a case of intermittent ataxia and psychosis--clinical, cognitive, and neuroanatomical correlates. J Neuropsychiatry Clin Neurosci 16, 400-408.

44. Yeganeh-Doost, P., Gruber, O., Falkai, P., and Schmitt, A. (2011). The role of the cerebellum in schizophrenia: from cognition to molecular pathways. Clinics (Sao Paulo) 66 Suppl 1, 71-77.

45. Paquette, A.G., Lesseur, C., Armstrong, D.A., Koestler, D.C., Appleton, A.A., Lester, B.M., and Marsit, C.J. (2013). Placental HTR2A methylation is associated with infant neurobehavioral outcomes. Epigenetics 8, 796-801.

46. Ghadirivasfi, M., Nohesara, S., Ahmadkhaniha, H.R., Eskandari, M.R., Mostafavi, S., Thiagalingam, S., and Abdolmaleky, H.M. (2011). Hypomethylation of the serotonin receptor type-2A Gene (HTR2A) at T102C polymorphic site in DNA derived from the saliva of patients with schizophrenia and bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 156B, 536-545.

47. Falkenberg, V.R., Gurbaxani, B.M., Unger, E.R., and Rajeevan, M.S. (2011). Functional genomics of serotonin receptor 2A (HTR2A): interaction of polymorphism, methylation, expression and disease association. Neuromolecular Med 13, 66-76.

48. Dean, B. (2013). Epigenetic mechanisms and the serotonin 2A receptor in schizophrenia. Schizophrenia research 145, 128-129.

49. Dammann, G., Teschler, S., Haag, T., Altmuller, F., Tuczek, F., and Dammann, R.H. (2011). Increased DNA methylation of neuropsychiatric genes occurs in borderline personality disorder. Epigenetics 6, 1454-1462.

50. Abdolmaleky, H.M., Yaqubi, S., Papageorgis, P., Lambert, A.W., Ozturk, S., Sivaraman, V., and Thiagalingam, S. (2011). Epigenetic dysregulation of HTR2A in the brain of patients with schizophrenia and bipolar disorder. Schizophrenia research 129, 183-190.

51. Walitza, S., Bove, D.S., Romanos, M., Renner, T., Held, L., Simons, M., Wewetzer, C., Fleischhaker, C., Remschmidt, H., Warnke, A., et al. (2012). Pilot study on HTR2A promoter polymorphism, -1438G/A (rs6311) and a nearby copy number variation showed association with onset and severity in early onset obsessive-compulsive disorder. J Neural Transm 119, 507-515.

52. Bunik, V., Kaehne, T., Degtyarev, D., Shcherbakova, T., and Reiser, G. (2008). Novel isoenzyme of 2-oxoglutarate dehydrogenase is identified in brain, but not in heart. Febs J 275, 4990-5006.

53. Zhou, Y., and Danbolt, N.C. (2013). GABA and Glutamate Transporters in Brain. Front Endocrinol (Lausanne) 4, 165.

54. Platt, S.R. (2007). The role of glutamate in central nervous system health and disease--a review. Vet J 173, 278-286.

55. Clinton, S.M., and Meador-Woodruff, J.H. (2004). Abnormalities of the NMDA Receptor and Associated Intracellular Molecules in the Thalamus in Schizophrenia and Bipolar Disorder. Neuropsychopharmacology 29, 1353-1362.

56. Tamminga, C.A., Southcott, S., Sacco, C., Wagner, A.D., and Ghose, S. (2012). Glutamate dysfunction in hippocampus: relevance of dentate gyrus and CA3 signaling. Schizophr Bull 38, 927-935.

57. Marsman, A., van den Heuvel, M.P., Klomp, D.W., Kahn, R.S., Luijten, P.R., and Hulshoff Pol, H.E. (2013). Glutamate in schizophrenia: a focused review and meta-analysis of (1)H-MRS studies. Schizophr Bull 39, 120-129.

58. Lee, J.A., Park, J.E., Lee, D.H., Park, S.G., Myung, P.K., Park, B.C., and Cho, S. (2008). G1 to S phase transition protein 1 induces apoptosis signal-regulating kinase 1 activation by dissociating 14-3-3 from ASK1. Oncogene 27, 1297-1305.

59. Ferrer, I., Martinez, A., Boluda, S., Parchi, P., and Barrachina, M. (2008). Brain banks: benefits, limitations and cautions concerning the use of post-mortem brain tissue for molecular studies. Cell Tissue Bank 9, 181-194.

60. Florath, I., Butterbach, K., Muller, H., Bewerunge-Hudler, M., and Brenner, H. (2014). Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet 23, 1186-1201.

61. Day, K., Waite, L.L., Thalacker-Mercer, A., West, A., Bamman, M.M., Brooks, J.D., Myers, R.M., and Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol 14, R102.

62. Dong, E., Nelson, M., Grayson, D.R., Costa, E., and Guidotti, A. (2008). Clozapine and sulpiride but not haloperidol or olanzapine activate brain DNA demethylation. Proc Natl Acad Sci U S A 105, 13614-13619.

63. Feinberg, A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature 447, 433-440.

64. Melas, P.A., Rogdaki, M., Osby, U., Schalling, M., Lavebratt, C., and Ekstrom, T.J. (2012). Epigenetic aberrations in leukocytes of patients with schizophrenia: association of global DNA methylation with antipsychotic drug treatment and disease onset. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 26, 2712-2718.

65. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K.L. (2009). Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1, 177-200.

66. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics 98, 288-295.

67. Aberg, K.A., McClay, J.L., Nerella, S., Xie, L.Y., Clark, S.L., Hudson, A.D., Bukszar, J., Adkins, D., Consortium, S.S., Hultman, C.M., et al. (2012). MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case--control samples. Epigenomics 4, 605-621.

68. Jin, S.G., Kadam, S., and Pfeifer, G.P. (2010). Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Res 38, e125.

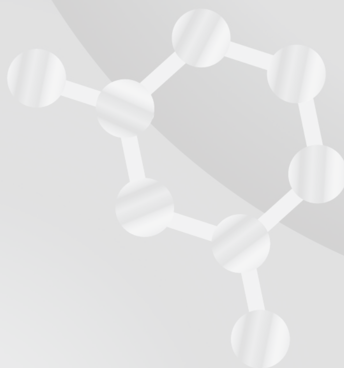
69. Branco, M.R., Ficz, G., and Reik, W. (2012). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Genet 13, 7-13.

70. Guo, J.U., Su, Y., Zhong, C., Ming, G.L., and Song, H. (2011). Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. Cell Cycle 10, 2662-2668.

71. Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet 12, 529-541.

72. Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447, 425-432.

73. Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., and Butte, A.J. (2010). Cell type-specific gene expression differences in complex tissues. Nat Methods 7, 287-289.



Summary and Discussion

The aim of this thesis was to explore the relationships and processes between different layers of genomic information (i.e. genotypes, DNA methylation, and gene expression) and to use this knowledge to gain more insight into the mechanisms underlying the neuropsychiatric disorder schizophrenia. We employed a number of approaches in a large sample of schizophrenia patients and healthy controls of Dutch descent. This sample constitutes a genetically homogenous population, which has an advantage over genetically admixed and more heterogeneous populations, which require correction for population stratification.

Our main findings are:

1. DNA methylation at many different CpG loci in the human genome is significantly correlated with gene expression, both negatively and positively (Chapter 2).
2. A subset of these methylation and expression levels is under genetic control, a small proportion of which is regulated by a shared (common) SNP (Chapter 2).
3. Using three-way associations (testing genotypes, DNA methylation and gene expression data), we identified causal relationships between methylation and expression and found the traditional model (where a SNP affects methylation that causes a change in expression) to be the most likely model in most cases (Chapter 2).
4. Methylation and expression levels can be organized into modules. Genes in the methylation and expression models generally do not overlap, however, highly significant correlations exist between co-methylation and co-expression modules (Chapter 2).
5. Expression QTL (eQTL) analysis highlights three additional schizophrenia candidate genes. These genes are regulated by SNPs associated with schizophrenia and are differentially expressed in schizophrenia patients compared to controls. (Chapter 3).
6. Genetic variations (SNPs), which regulate differential methylation that is associated with differential expression in schizophrenia, are significantly enriched for schizophrenia susceptibility loci, and provide insights into molecular mechanisms involved in disease etiology (Chapter 4).
7. Genome-methylome interactions extend well beyond what is detectible with the commonly used methylation quantitative trait loci (mQTL) and allele-specific methylation (ASM) approaches since only half of the heritable CpG methylation is regulated by *cis*-SNPs identified by mQTL analysis, and the majority of ASM cannot be explained by consistent genetic regulatory effects (Chapter 5).
8. Identification of variable methylated regions (VMRs) clusters associated with so-called "variation SNPs" supports the hypothesis that genetic variation can affect not only the mean but also the variability of methylation levels (Chapter 5).
9. MicroRNA-137 (miR-137) is highly expressed in human brain regions implicated in schizophrenia (Chapter 6).

10. Interplay between miR-137 and DNA methylation highlights neurotransmitter-related pathways, related to neuropsychiatric disorders (Chapter 6).

We will briefly summarize and discuss these findings below.

1.1 Methylation levels are associated with gene expression in *cis*, of which a subset is under genetic control by SNPs

To better understand the regulation of gene expression in relation to DNA methylation, the relationship between DNA methylation, gene expression and genotypes was examined in Chapter 2. We were the first to test thousands of transcripts with thousands of CpG sites genome-wide for association in a relatively large set of healthy controls (N=148). Many methylation-expression relations were identified revealing not only expected negative associations, but also positive associations, as has been found before[1; 2]. Although traditionally it has been thought that CpG methylation mostly inhibits or represses gene expression, for example by blocking transcription factor/RNA polymerase binding sites, these results show that these relationships are more complex. There are a number of possible explanations for this complexity.

Methylation may induce gene expression by preventing the binding of a protein to an insulator in the DNA, allowing for enhancers to bind to the promoter and initiate gene expression[3; 4]. However, the role of methylation in enhancer and insulator functions have yet to be further established[5]. Another possible explanation for positive associations between methylation and gene expression is the involvement of a third "unmeasured" factor, i.e. a gene coding for a transcription factor (TF) not present on the methylation array. Methylation of a gene could have a negative association with a transcription factor, which in turn down-regulates gene expression of a certain gene. Increased methylation causes repression of this TF (by methylating the promoter of the TF, or by blocking its binding site), prohibiting down-regulation of gene expression of that gene, possibly leading to increased expression. In addition, a recent review made the distinction between transcription start site (TSS) or promoter methylation and gene body methylation. The latter was found to be more ambiguous and context specific, and interestingly, could lead to both repression and activation[5].

Next, we found that in more than 12% of the methylation-expression *cis* pairs, methylation and/or expression was associated with a SNP in *cis*, suggesting genetic control of these levels. A subset of these SNPs regulates gene expression and the associated DNA methylation levels. These "three-way associations" are very interesting since they are likely to represent an underlying genetic mechanism for epigenetic changes. Furthermore, we introduced

directionality analysis for these three-way associations involving methylation, expression, and SNPs. We were the first to determine directionality of the association between DNA methylation and gene expression induced by genotype, using Local Edge Orienting modeling. Taking the SNP as a causal anchor, we demonstrated that in most instances, it is DNA methylation that causes a change in gene expression. This finding is in accord with the prevailing classical model of regulation of gene expression. However, we also observed a (small) number of instances in which the strongest evidence points to gene expression regulating DNA methylation levels. Further studies are needed to confirm these findings and decipher the underlying mechanisms, both for the loci supporting the classical model as well as those favoring the reversed model. Our results suggest that the causal direction analysis is a useful approach for determining directionality. Furthermore, methylation levels that are associated with SNPs and expression levels are more frequent in CpG island shores, regions spanning up to 2kb around a CpG island (CGI). Shores are more variable than CGIs or regions outside CGIs, and are found to be more frequently involved in tissue-specific differential methylation[6]. We observed that methylation changes in these shores are strongly related to transcription of associated genes, consistent with others' findings and supporting a functional role in gene expression[6]. These results indicate that our findings are biologically relevant. Finally, after organizing methylation and expression levels into modules, significant enrichment of genome ontologies was found, suggesting that these modules are biologically meaningful. Genes in the methylation and expression models generally do not overlap, however, highly significant correlations exist between co-methylation and co-expression modules, suggesting the existence of *trans* effects influencing methylation and expression in different modules.

1.2. Three candidate genes for schizophrenia identified using eQTL analysis

After investigating the association between DNA methylation and gene expression in healthy individuals, we further explored the associations with genotypes in schizophrenia in Chapter 3.

Large-scale genome-wide association studies (GWAS), conducted by the Schizophrenia Psychiatric GWAS Consortium (PGC), have led to the identification of many SNPs associated with schizophrenia[7; 8]. These analyses were conducted in thousands of samples, but the effect sizes of individual associated SNPs are very small with odds ratios (OR) below 1.3. To date, it has been estimated that >8,000 common variants contribute to schizophrenia susceptibility[7], explaining approximately 30% of the heritability[7; 9]. However, GWAS has only 22 SNPs associated with schizophrenia thus far.

We speculated that, when combining these top SNPs (associated with disease but with small effect size) with gene expression data, this extra layer of information could highlight additional candidate genes for schizophrenia that could not be detected by GWAS alone. The top SNPs from the first PGC meta-analysis[8] with a p-value < 0.001, were investigated for their association with gene expression. These transcripts were subsequently used in a schizophrenia case/control dataset to identify differentially expressed transcripts. Seven genes were differentially expressed, of which three coincided with the expected direction of the SNP found in the original meta-analysis. These three genes are all located on chromosome six, in the major histocompatibility complex (MHC) region, a region previously linked to schizophrenia in several studies[7; 9; 10]. The MHC is a difficult region to analyze and interpret due in part to its highly divers DNA sequence and linkage disequilibrium patterns[11; 12]. This diversity may hinder oligonucleotide probes from binding properly. Although, this study indicates that variants in the MHC have an effect on differential gene expression in schizophrenia cases compared to controls and may contribute to schizophrenia susceptibility, these results should be interpreted with caution due to the genomic location of these variants.

1.3. SNPs regulating differential methylation that is associated with differential gene expression are enriched for schizophrenia signal

Since GWAS hits are enriched for eQTLs[9], and because findings in Chapter 3 show that top GWAS findings represent eQTLs that are regulating differential gene expression in schizophrenia compared to controls, we hypothesized that methylation Quantitative Trait Loci (mQTLs) are enriched for SNPs previously associated with schizophrenia. Therefore we again used the results of the previously mentioned PGC meta-analysis[8]. We were the first to perform this analysis in schizophrenia patients incorporating DNA methylation and gene expression data (Chapter 4). In our study, we were interested in whether more biologically relevant mQTLs show stronger enrichment for schizophrenia loci compared to less biologically relevant mQTLs. Biologically relevant mQTLs are defined as SNPs associated with differential methylation between cases and controls and for which the CpG site is associated with differential expression. Therefore, we generated four lists of SNPs (mQTLs) with increasing biological relevance and tested to see if they were enriched in schizophrenia loci. Our analysis revealed that SNPs with high biological relevance are indeed enriched in schizophrenia susceptibility loci. The top SNP that we identified in this analysis, SNP rs11191514, is associated with methylation of *CALHM1* (calcium homeostasis modulator 1) and is located in the gene *CNNM2*. Another SNP in this gene, in complete LD with rs11191514, is one of the top ten SNPs reaching genome-wide significance in the PGC GWAS[8]. Interestingly, after we conducted this research, a paper was published where 13 more risk loci for schizophrenia were found, including a region covering the gene *CALHM1*

(for which we identified differential methylation and expression)[7]. The importance of calcium regulation in schizophrenia has been emphasized before[7; 13]. Our approach is the reverse of two recently published studies investigating whether SNPs associated with disease were enriched for mQTLs and eQTLs[14; 15]. We examined whether mQTLs are enriched for disease SNPs without restricting ourselves to those loci with prior evidence of association with schizophrenia. In summary, we demonstrated that incorporating methylation and gene expression data from whole blood of patients and controls leads to enrichment of disease-associated alleles in a much smaller sample size than is required for GWAS studies.

1.4. Characterization of genome-methylome interactions in families

For the vast majority of CpG sites, methylation is complementary at both DNA strands. However, methylation can also be allele-specific, as is the case in X-chromosome inactivation and parental imprinting. Some studies showed that methylation can also be allele-specific on autosomal chromosomes[16].

Allele specific methylation (ASM) is strongly associated with genotypes in *cis* (close to the CpG site). Studies investigating the association between genotypes and CpG methylation have been conducted using mQTL and ASM analyses. However, it is not clear to what extent these identified SNPs contribute to the variability in DNA methylation. Therefore we investigated the associations between CpGs and SNPs using three different approaches in Chapter 5. Moreover, we examined the genetic contribution to variance in DNA methylation using 22 nuclear families; each family contained one child who had been diagnosed with schizophrenia. We used targeted bisulfite sequencing with padlock probes to quantify absolute DNA methylation levels from over 400,000 CpG sites, and extracted genotypes from these reads as well. We found that there were substantial differences in the overlap of identified CpG-SNP associations between the three approaches. First of all, of the heritable CpGs established by mid-parent-offspring (MPO) analysis, 70% contains a SNP directly disrupting the CpG site. For the remaining non-SNP CpG sites, approximately half of the sites were regulated by *cis* SNPs, as identified by mQTL analysis. In addition, heritable CpGs accounted for the CpGs most strongly associated with SNPs. After examining allele-specific differences of DNA methylation within individuals, we found that SNPs at CpG sites were responsible for most ASM events. Only a small overlap between ASM CpG sites and heritable and SNP-associated CpGs was detected. We hypothesized that ASM analysis is more susceptible to non-genetic factors. Finally, we found hundreds of variable methylated regions (VMRs), covering at least five CpG sites associated with a variation SNP, supporting the hypothesis that a SNP can affect not only the mean methylation level, but also the variability of the methylation level[17].

1.5. Association of miRNA137 with methylation from brain tissue

It is increasingly clear that gene expression and methylation are in part tissue dependent[18-20]. The aforementioned studies were all conducted with DNA and RNA derived from whole blood in order to study a neuropsychiatric disease (schizophrenia). The reason for this is obvious: it is difficult to obtain brain samples from patients. Studying brain tissue from schizophrenia patients is extremely valuable. The last Chapter (6) in this thesis involves *post-mortem* brain tissue from which DNA methylation and expression levels were measured to create a very unique brain dataset. After using the results from the PGC meta-analysis in Chapters 3 and 4, we were interested in the strongest schizophrenia susceptibility locus, a transcript for the gene microRNA-137 (miR-137)[8]. MicroRNAs (miRNAs) are small non-coding RNA molecules that can regulate expression of mRNA transcripts. A subset of miRNAs controls the expression of key players of the epigenetic machinery, and in turn, epigenetic mechanisms, including DNA methylation can regulate the expression of miRNAs[21]. This interplay between miRNAs and DNA methylation may provide more insight into gene regulation, and possibly into disease processes when there is a disruption in this machinery. Therefore, we examined the relationship of miR-137 expression and genome-wide DNA methylation levels in human *post-mortem* brain. First, we studied miR-137 expression in different brain regions. We found that the regions with the highest miR-137 expression levels are also the ones that show abnormalities in schizophrenia patients[22]. These include the cortical regions, limbic system and basal ganglia. Next, we investigated interactions between miR-137 and DNA methylation levels both locally (region-specific) and globally (incorporating all regions). We found suggestive associations between miR-137 and promoter methylation of three genes. The strongest observation involves *HTR2A*, serotonin receptor type 2, of which the promoter is known to be under epigenetic control[23-28]. Moreover, this gene is important in neuronal processes, and has been recently implicated in schizophrenia[23; 27]. The second-best hit involves the gene 2-oxoglutarate dehydrogenase-like (OGDHL). This gene is expressed in the brain where it is involved in glutamate synthesis[29]. Glutamate is an important neurotransmitter and glutamate neurotransmission has been suggested to be involved in neuropsychiatric disorders including schizophrenia[30-32].

Our findings link a well-established schizophrenia susceptibility locus, miR-137, to the serotonin and glutamate pathway. Although the evidence is suggestive, it might highlight biological relevance of miR-137 in schizophrenia susceptibility and provide additional insight into the pathophysiology of this disorder.

2. Challenges

When large-scale genome-wide DNA methylation assays became available a few years ago, expectations were very high. Study of DNA methylation in cancer research resulted in remarkable discoveries[6; 33-35]. For example, hypomethylation can induce growth-promoting gene expression in tumors, and promoter hypermethylation has been associated with repression of tumor suppressor genes[33-35]. The first results of DNA methylation research in psychiatric disorders were also promising[36; 37]. However, interpretation of results from DNA methylation studies has been more challenging than expected. Some challenges and limitations should be considered in interpreting results, as we will discuss in this section. Primarily, the challenges of studying DNA methylation are technical and biological in nature.

2.1. Technical challenges:

2.1.1. The 27K Beadchip

The methylation data we used in the projects described in this thesis were obtained using the Illumina HumanMethylation27 Beadchip[38]. This beadchip (the "27K" chip) interrogates 27,000 methylation sites corresponding to roughly 14,000 protein-encoding genes. Although it covers more than half of all human genes, it is a small fraction of the CpG sites located in the genome (0.1%). In addition, the 27K chip was designed to include many known cancer candidate loci and focused on inclusion of CpGs located in CpG islands. In other words, the 27K array provides an informative but skewed view of DNA methylation in the human genome, largely ignoring regions outside CpG islands. However, since CpG islands are mostly present in promoter regions, covering an important regulatory region of many genes, these arrays are useful in analyzing methylation sites that are expected to regulate biological function. Following the 27K array, the next-generation (current) design of the Illumina methylation array contains over 450,000 loci[39], an almost 20-fold increase of genomic coverage. Still, this newer 450K DNA methylation array represents only 1.5% of CpGs in the genome. While this development shows the technical limitation of array-based screening of DNA methylation profiles, it also shows the advancements in the field of epigenetics and the rapidly developing genomic tools that are available for probing DNA methylation. In recent years, next-generation sequence-based analysis of DNA methylation has also become available[40-42]. One of these methods, bisulfite padlock sequencing, is described in Chapter 5. This method makes use of bisulfite sequencing reads with padlock probes to quantify absolute DNA methylation levels from over 400,000 CpG sites genome-wide. Sequence-based methods do not suffer from hybridization artifacts and make it possible to study allele-specific effects.

2.1.2. Batch effects

Another problem we faced when performing quality control was the enormous effects of technical variation, also known as batch effects. Most likely, batch effects in DNA methylation research is due to differences in conversion rate of unmethylated CpGs when genomic DNA is treated with bisulfate. Another technical bias may occur due to incomplete bisulphite conversion or amplification inefficiency[43]. These technical batch effects are known to exist when using array-based approaches such as Illumina Infinium chips[44]. It is critically important to handle these possible biases thoroughly since batch effects can introduce many type 1 and type 2 errors, especially when there is a relationship between batch effects and phenotype (i.e. cases/control status)[44; 45]. In our experiments, it was a real challenge to remove technical variation as much as possible, while retaining the biologically relevant variation. The largest effects were detected between the arrays (reflecting different bisulfite treatment experiments). In addition, we observed strong plate (8 arrays) and position (12 positions per array) effects on the methylation data. While avoiding batch effects is impossible, the best way to have these effects not interfere with data analysis and results is to thoroughly randomize samples (in terms of disease status and sex over the different arrays and plates. Another limitation of the array-based technology for studying DNA methylation is the presence of genetic variation at probe regions, as it affects optimal hybridization and thereby generates allele-specific methylation profiles that are not related to the phenotype of interest. CpG probes with known SNPs and those that hybridize to multiple locations in the human genome need to be removed prior to data analysis to reduce false-positive findings[44].

2.1.3. 5' hydroxymethylation

Another limitation of the Illumina arrays is that it is impossible to distinguish between DNA methylation and 5' hydroxymethylation[18; 46]. The latter is thought to be an intermediate form between methylation and de-methylation. Almost half of the methylated cytosine in the brain is in this form[47]. It is not clear what the consequences of 5' hydroxymethylation are, but it is likely to play a role in gene regulation and differentiation[48-50].

2.2. Biological challenges

2.2.1. Cell heterogeneity

Biological challenges should also be considered. An important challenge to address is cell composition heterogeneity. For most of our studies, DNA and RNA derived from whole blood was used. Whole blood mainly consists of red blood cells and plasma, and a small proportion of white blood cells and platelets. Since DNA methylation is involved in the regulation of cell differentiation, the methylation and expression marks in these cell types can be different.

The cell type composition of blood varies between individuals, and if the composition differs between cases and controls, it may lead to false positive findings. Recently, several methods have been published that take this problem into account by correcting for cell-type composition even if this composition is unknown[51-53]. It is recommended that we apply these methods and re-analyze available data sets including our own.

Since each cell type is characterized by its own epigenetic profile, we need to consider another issue when studying schizophrenia and related phenotypes. Our studies have been based (primarily) on whole blood of patients and controls. However, the target cell type of neuropsychiatric disorders is likely to be neuronal and different from cell types found in peripheral blood. This means that blood-based findings are likely to provide a limited view of epigenetic events occurring in brain-specific cells. Nevertheless, our reported findings were informative for disease (in different experiments), which suggests that some aspects of epigenetic control are preserved between different cell types and across tissues. We acknowledge, however, that it would be very valuable to replicate these findings in brain cells. The main reason to perform blood-based DNA methylation and gene expression studies of schizophrenia is based on the fact that large numbers of patients and controls can be collected. This is in sharp contrast with the available brain tissues of schizophrenia patients, which is extremely limited in sample size. Moreover, it is also very challenging to positively select one disease-specific region within the brain because of the extensive heterogeneity of cell types across different brain regions. Although we had the availability of a unique dataset consisting of many brain regions from schizophrenia and bipolar patients (n=17) and healthy subjects (n=44), the sample size resulted in our study being significantly underpowered. For the work presented in this thesis, it proved to be simply impossible to reach the same sample size for brain tissues that is widely available for whole blood. Another potential disadvantage of using brain tissue of patients is that these samples are usually obtained from elderly patients *post-mortem* while the disease is considered a neurodevelopment disorder. Moreover, the quality of *post-mortem* brain tissue can be negatively affected by a number of factors such as post-mortem delay, storage temperature, and pH. Although there is not much known about these effects on methylation, pre-mortem and post-mortem factors may intrude molecular mechanisms[54]. Also obscure is whether post-mortem adult brain tissue (obtained from mostly elderly people) is informative for neurodevelopmental disorders. These challenges have to be further explored in the future.

2.2.2. Methylation: cause or consequence?

DNA methylation is heritable and reversible, and despite many (significant) associations with disease, these associations can be either causal or consequential with respect to the phenotype[18; 33; 44]. Generally, a causal role (when methylation affects gene expression

and subsequently disorders develop) has been traditionally thought to explain these associations[33]. There are several factors leading to variation in methylation[18]. First, if this variation is inherited, it can be present in all tissues, including the germ line. There is not much known about this process[55], but several studies on this were published in mouse[56], rat[57], and human[58; 59]. One of the most striking examples involves the Agouti mice[56]. Difference in a mother's diet led to different coat color due to epigenetic modifications at a single gene. When the mother received a diet full of supplements including folic acid during pregnancy, the offspring had brown fur. However, when these supplements were restricted in the mother's diet, the offspring were born with yellow fur and were also obese and developed diabetes and tumors. This phenomenon was caused by the methylation of the agouti gene, which repressed the gene. In this case, folic acid had served as a methyl donor. Second, if variation occurred randomly, it can be present in specific tissue types or in all tissues depending on time of occurrence during the developmental stage. Third, there are multiple factors that influence DNA methylation, such as medication and environment (i.e. smoking, alcohol, diet, and stress)[60; 61]. Several studies show the effect of medication on methylation[33; 62; 63]. Considering that most schizophrenia patients use medication, as described in the next section, the confounding effects of medication make it difficult to distinguish the origin of the methylation differences between patients and healthy controls. Lastly, methylation can be under genetic control. We have examined this feature using linear models to investigate associations between genotypes (SNPs) and methylation levels.

Genotype arrays used in these studies include over 300,000 SNPs. These so-called tagging SNPs represent a number of SNPs, not necessarily present on the array, that are in linkage disequilibrium (LD) with the tag SNP. If an association is identified between a SNP and a methylation level, this SNP is not necessarily the causal variant, but may reflect signal from adjacent SNPs due to LD. Therefore it is not always clear whether the identified SNP is driving the signal. On the other hand, it demonstrates that a subset of SNPs can be used for association testing, reducing costs.

In addition to this, there is also a question about the extent to which DNA methylation really contributes to disease? Overall, variations in methylation levels between individuals are relatively small. The significant differentially methylated CpG sites we found between schizophrenia patients and controls express a maximum fold change of approximately 3%. The question remains, what does such a small change in methylation level mean biologically? To what extent does it contribute to gene expression and protein regulation? In the cell, gene expression (or "dosage") changes all the time, and may depend on a variety of factors, such as the time of the day. It is thought that a small change in DNA methylation can have major effects in expression if transcription factor binding domains or regions involved in

recruitment of repressors are implicated[64]. In this way, DNA methylation may serve as a "switch", which may act given a certain threshold level, and that threshold may vary by gene.

In this regard, methylation analysis may be very similar, but not inferior, to GWAS. GWAS, specifically meta- or mega-analyses, encompass thousands of individuals and uncover associated SNPs with small effect sizes (in complex traits). As mentioned before, ORs are often small due to very small differences in allele frequencies (~3%) between cases and controls. Therefore it is not surprising that we also find minimal effects in methylation, especially when considering our relatively small sample size compared to GWAS. Perhaps, the "common variants with small effect and rare variants with large effect" hypothesis in GWAS is also applicable to EWAS. It has also been hypothesized that the explained heritability in complex diseases is overestimated, due to epistasis (gene-gene interactions) for example[65]. However, it is not clear whether the remaining variance of disease can be attributed to epigenetics (when its not considered part of heritability) or if most of the variance is explained by SNPs already and methylation will not explain a significant proportion of this missing heritability.

In general, we are far from fully understanding the causes and consequences of the observed differences in DNA methylation[44], and the current findings do not distinguish between causal and consequential DNA methylation variation. Further longitudinal studies will be necessary to aid in our understanding of this epigenetic process. Furthermore, statistical significance does not by definition mean biological significance, stressing the importance of validation by "wet-lab" experiments.

2.2.3. The challenges of schizophrenia as a phenotype

Schizophrenia is a complex disease with large heritability (~80%), and a polygenic background. To date, common variants have been estimated to explain approximately 30-50% of the heritability[7; 9]. The effect sizes of the polymorphisms are small, with odds ratios not exceeding 1.5. Furthermore, schizophrenia is a very broad phenotype. Diagnosis is fully dependent on classification by a psychiatrist, as there are no biological markers available. Even though the Diagnostic and Statistical Manual of Mental Disorders (DSM) has been created to ensure comparability in phenotype across different psychiatrists in different countries, it still remains a subjective diagnosis.

An issue that is inherent to the study of schizophrenia and related psychiatric disorders is the confounding effect of medication use. Almost all patients are using prescription drugs or resort to self-medication that is likely to affect DNA methylation profiles in blood cells and elsewhere in the human body[62]. For this reason, results of epigenetic studies of human

neuropsychiatric traits (including those reported in this thesis) should be interpreted with some caution.

3. Future perspectives

3.1. Other techniques

Techniques and methods to investigate human genetics and diseases are growing rapidly. Genotype, methylation and expression arrays interrogate increasing numbers of SNPs or probes and high-throughput sequencing has become readily accessible. With the decreasing cost per array, sample size per study is increasing. In most of the chapters of this thesis, methylation data was collected from arrays, and in Chapter 5, bisulfite sequencing with padlock probes was used to detect methylation at CpG sites. As techniques to measure quantitative trait data like DNA methylation and gene expression develop and advance, coverage and confidence will increase. Along with array-based technologies, there are sequence-based technologies in addition to bisulfate sequencing padlock probes (BPSPP) (described in Chapter 5), such as whole-genome bisulfite sequencing (WGBS)[42] and reduced representation bisulfite sequencing (RRBS)[40; 41]. Sequencing provides higher coverage and resolution but is more expensive, although costs are likely to decrease over time[18]. Also, exome-sequencing studies involving schizophrenia patients are advancing, with a large study published recently examining rare mutations[66]. This is relevant because, due to the polygenic architecture of schizophrenia, the estimated explained variance from common alleles warrants the search for rare variants which may have much larger effect on disease susceptibility. To gain more insight into the functional relevance of (altered) DNA methylation and gene expression relations, animal models are currently being used to test the functional effects of gene mutations (i.e. using knock-out or knock-down mice) in several diseases, including schizophrenia[67]. However, schizophrenia characteristics such as delusions and hallucinations are difficult to define in a mouse, although some phenotypes such as social impairment can be observed. Also, for studying gene expression patterns and the effect of DNA methylation on expression, animal models are valuable. This approach succeeded in RETT syndrome, an autism spectrum disorder. Deletion of the gene MeCP2, which encodes for a protein that inhibits gene expression, caused RETT-like symptoms in mice. Strikingly, restoring this genes expression led to reversal of the symptoms in mice[68]. Ideally, a combination of methods, especially when validating results, is required. This will also contribute to generating new methylome maps. Furthermore, detailed knowledge of methylome structure, including correlated tissue-specific blocks will aid in improved selection of CpG sites for Epigenome-wide Association Studies (EWAS)[18]. Attempts of several consortia resulted in public resources containing epigenomic data available for research. For example, the Encyclopedia of DNA Elements (ENCODE) Consortium generated a freely available resource containing functional elements present in the human genome[69]. ENCODE also develops methods and conducts

genome-wide sequence-based studies to map functional elements, which will provide new insights into the organization and regulation of our genome. The NIH Roadmap Epigenomics Mapping Consortium launched a public resource containing human epigenome data, and aims to provide a reference epigenome map[70; 71]. This consortium also generated tools, pipelines and protocols for data generation and data analyzing, aiding in expanding our knowledge of epigenetics in normal development and disease.

3.2. Extending sample size and refining patient selection

After many attempts to discover the variants contributing to schizophrenia, the majority of the estimated heritability still cannot not be explained. It is likely that larger studies with increasing sample size will uncover more common variants associated with schizophrenia[7; 72]. This will be similar for epigenetic studies. With expanding sample sizes, power to detect differential methylation in schizophrenia patients is likely to increase. The challenge is to carefully select samples. In our studies, we examined all schizophrenia cases as one group. It is possible that in EWAS (and GWAS), we need to abandon the approach of considering schizophrenia as one disorder and rather focus on symptoms (such as psychosis) as opposed to the diagnosis. This would allow for interrogating larger datasets (as other psychiatric disorders share symptoms), and improve power to find (disease-causing) associations. On the other hand, it might be worthwhile to narrow the phenotype and assign schizophrenia patients into different subgroups according to clinical features, and analyze them separately. However, sample sizes would need to increase drastically to overcome power limitations. The sample sizes achieved in GWAS studies of height and BMI, for example, will probably never be reached for psychiatric patients, even if groups are working together, because the low prevalence of psychiatric disorders is quite low (as opposed to general features present in all humans, such as height). Recruitment of schizophrenia patients is very time-consuming, as they need to be seen by clinicians, be diagnosed, and be willing to participate.

In addition, it would be very valuable to include medication-naïve patients, to exclude the medication effect; this will remain a challenge since almost all patients are on medication. Perhaps this limitation could be overcome by sampling individuals before they have the first symptoms to developing schizophrenia, and select them from a group of controls. But this requires thousands of people being collected. Besides, methylation changes contributing to the susceptibility of the disorder might not be relevant before disease onset. However, there is a pre-phase in which individuals show some first signs[73]. This might be a good moment to collect their blood. In addition, ascertainment of more families is preferable since multiplex pedigrees are enriched for causal genetic variation with higher penetrance[72], providing more information about the underlying genetics.

Foremost and key in all fields of research is collaborating with other research groups to increase sample size and knowledge. This is why large consortia were initiated, like the Psychiatric Genetics Consortium for GWAS, and ENIGMA (Enhancing Neuro Imaging Genetics through Meta-Analysis) network for imaging analyses. Epigenetics and DNA methylation consortia, such as International Human Epigenome Consortium, are essential. As the equivalent of the 1,000 genomes map, this consortium is generating 1,000 reference epigenomes, including methylomes, for many human tissues and cell types, improving our ability to conduct EWAS for many common diseases in the coming years[18]. This will shed light on the organization and regulation of the human epigenome, and the way methylation profiles differ across individuals and tissue types, and disease.

3.3. Other (epi)genetic marks

DNA methylation is not the only type of epigenetic modification and thus provides a very limited view of epigenetics. Other well-studied regulatory marks in the human genome involve histone modification, which can influence chromatin state, as well as non-coding RNAs such as microRNAs (Chapter 6). To fully understand the (epi)genetic mechanisms of disease, it is imperative to obtain these measures and combine them with other layers of (epi)genetic information and subsequently conduct case/control studies. Although DNA methylation is the most suitable for EWAS, since it is detectable at large scale, it is vital to conduct large-scale analyses for other epigenetic marks as well. Recent studies show that DNA methylation and histone modification can be associated with each other and that both histone- and DNA methyltransferases can mediate this interplay. This is extensively discussed in this review[74]. Investigating the correlations between different epigenetic variations will shed light on the epigenetic regulation of chromatin modification and subsequently gene expression. Examining chromatin structure will point to regions of interest by selecting regions that exhibit open chromatin structure and in general active transcription. In addition, extending and combining epigenetic research with gene-gene and gene-environment interactions research will facilitate our understanding of complex biological mechanisms.

3.4. Implications for disease

The time of onset for schizophrenia is typically adolescence. During this period, the brain is reorganized, and major changes in the neural systems emerge[75; 76]. The cortical regions in particular show changes. This is also observed in schizophrenia patients[22; 76]. In addition, gray matter loss has been revealed in schizophrenia patients[77]. These observations lead to the questions: are genes methylated as a consequence of these brain changes? Or were these different methylation patterns there before? And if so, what causes these patterns to change?

One hypothesis is a "two-hit" model whereby an early developmental or genetic anomaly preceding a (psychological) traumatic experience or other trigger such as stress, later in life, result in developing the disorder[78]. This fits with a model whereby methylation profiles change due to environment. Such a model is consistent with the hypothesis that psychosis does not arise in a healthy brain[73], and that genetic mutations are associated with the disorder. Many of the SNPs found so far implicate genes involved in neuronal development, and some of these variants are associated with multiple neurodevelopmental disorders like schizophrenia, autism, and bipolar disorder. Since autism is revealed in young children compared to a higher age of onset in schizophrenia, it is thought that environmental and/or epigenetic influences are contributing as well[73].

As described above, there are probably thousands of variants contributing to schizophrenia etiology. What if we eventually find all the genes and variants involved in schizophrenia? It is very likely that each individual patient has a different (epi)genetic make-up that leads to disease onset. Perhaps, a (polygenic) risk or "burden" score could be calculated that assembles information from all variants[79], representing the likelihood that a person will develop schizophrenia. This "early diagnosis", and in turn, early treatment, might help in improved outcomes.

Moreover, instead of investigating every single gene (for possible drug targets), perhaps we should move a level up and analyze the bigger picture. Combining all genes to find common pathways will lead to biological processes and molecular functions that can be studied for potential drug targets. This is very important since there is a need for more personalized medicine with fewer side effects. Although there are publications discussing the relevance of epigenetic drugs[62; 80; 81], these drugs are far from implementation in the clinic. Epigenetics will hopefully further aid in elucidating underlying biological mechanisms. Although studying DNA methylation did not reveal a major impact on a single gene, increasing sample sizes, combining data, and refining methods and techniques will make this a more promising approach for further study.

4. Conclusion

Before embarking on the projects described in this thesis, our aim was to gain more insight into molecular genetic aspects of schizophrenia by studying genetic variation, gene expression and DNA methylation jointly. With the genome-wide study in healthy controls in which we investigated the relationship of DNA methylation with gene expression in Chapter 2, we found that these relationships were far more complex than originally thought. Using Local Edge Orienting modeling we showed that in three-way associations, methylation

under genetic control is affecting gene expression, indicating the relevance of this approach in defining the causal relationships between these genomic layers. We also demonstrated the importance of using gene expression (Chapter 3) and DNA methylation (Chapter 4) data in combination with previously conducted GWAS results. This led to identification of three additional genes for schizophrenia resulting from the eQTL study, and one gene from the methylation enrichment study. Since the latter gene is a calcium ion channel, which has been associated with schizophrenia before, this might be a potential drug target and is therefore relevant for further research. Using targeted bisulfite sequencing in Chapter 5, we found that the extent of genome-methylome interactions is well beyond what is detectible with the commonly used mQTL and ASM approaches because our results showed that approximately half of the heritable CpG methylation is regulated by *cis*- SNPs as identified by mQTL analysis. However, the majority of ASM cannot be explained by consistent genetic regulatory effects and might therefore be more susceptible to non-genetic factors. Our results also support the hypothesis that genetic variation can affect the variability of the methylation level. Finally, in Chapter 6, we found that a significant schizophrenia susceptibility locus, miR-137, is highly expressed in brain regions implicated in schizophrenia, and associations of miR-137 with DNA methylation levels highlighted neurotransmitter-related pathways.

To conclude, this thesis describes systematic genome-wide approaches to relate different layers of genomic information using bioinformatics. We show that DNA methylation is a useful and relevant piece of the puzzle. In addition, we demonstrate that although DNA methylation adds to the complexity of the genomic underpinning of schizophrenia, incorporating these measures is a promising and necessary tool to unravel the complex and disabling neuropsychiatric disorder schizophrenia.

References

1. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 6, e1000952.
2. Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S., and Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. American journal of human genetics 86, 411-419.
3. Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell 137, 1194-1211.
4. Herold, M., Bartkuhn, M., and Renkawitz, R. (2012). CTCF: insights into insulator function during development. Development 139, 1045-1057.
5. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13, 484-492.
6. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41, 178-186.
7. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45, 1150-1159.
8. Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.Y., Duan, J., Ophoff, R.A., Andreassen, O.A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43, 969-976.
9. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.
10. Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietilainen, O.P., Mors, O., Mortensen, P.B., et al. (2009). Common variants conferring risk of schizophrenia. Nature 460, 744-747.
11. Traherne, J.A. (2008). Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet 35, 179-192.
12. de Bakker, P.I., and Raychaudhuri, S. (2012). Interrogating the major histocompatibility complex with high-throughput genomics. Hum Mol Genet 21, R29-36.
13. Lidow, M.S. (2003). Calcium signaling dysfunction in schizophrenia: a unifying approach. Brain Res Brain Res Rev 43, 70-84.
14. Gamazon, E.R., Badner, J.A., Cheng, L., Zhang, C., Zhang, D., Cox, N.J., Gershon, E.S., Kelsoe, J.R., Greenwood, T.A., Nievergelt, C.M., et al. (2012). Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. Molecular psychiatry.

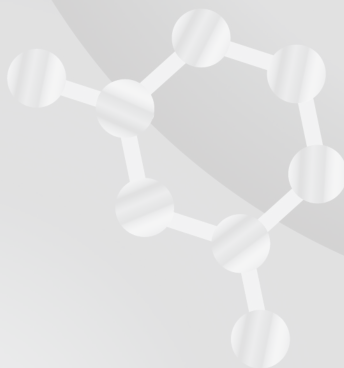
15. Richards, A.L., Jones, L., Moskvina, V., Kirov, G., Gejman, P.V., Levinson, D.F., Sanders, A.R., Purcell, S., Visscher, P.M., Craddock, N., et al. (2012). Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. Molecular psychiatry 17, 193-201.
16. Zhang, Y., Rohde, C., Reinhardt, R., Voelcker-Rehage, C., and Jeltsch, A. (2009). Non-imprinted allele-specific DNA methylation on human autosomes. Genome Biol 10, R138.
17. Feinberg, A.P., and Irizarry, R.A. (2010). Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A 107 Suppl 1, 1757-1764.
18. Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet 12, 529-541.
19. Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447, 425-432.
20. Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. Cell 155, 39-55.
21. Sato, F., Tsuchiya, S., Meltzer, S.J., and Shimizu, K. (2011). MicroRNAs and epigenetics. Febs J 278, 1598-1609.
22. Meyer-Lindenberg, A. (2010). From maps to mechanisms through neuroimaging of schizophrenia. Nature 468, 194-202.
23. Abdolmaleky, H.M., Yaqubi, S., Papageorgis, P., Lambert, A.W., Ozturk, S., Sivaraman, V., and Thiagalingam, S. (2011). Epigenetic dysregulation of HTR2A in the brain of patients with schizophrenia and bipolar disorder. Schizophrenia research 129, 183-190.
24. Dammann, G., Teschler, S., Haag, T., Altmuller, F., Tuczek, F., and Dammann, R.H. (2011). Increased DNA methylation of neuropsychiatric genes occurs in borderline personality disorder. Epigenetics 6, 1454-1462.
25. Dean, B. (2013). Epigenetic mechanisms and the serotonin 2A receptor in schizophrenia. Schizophrenia research 145, 128-129.
26. Falkenberg, V.R., Gurbaxani, B.M., Unger, E.R., and Rajeevan, M.S. (2011). Functional genomics of serotonin receptor 2A (HTR2A): interaction of polymorphism, methylation, expression and disease association. Neuromolecular Med 13, 66-76.
27. Ghadirivasfi, M., Nohesara, S., Ahmadkhaniha, H.R., Eskandari, M.R., Mostafavi, S., Thiagalingam, S., and Abdolmaleky, H.M. (2011). Hypomethylation of the serotonin receptor type-2A Gene (HTR2A) at T102C polymorphic site in DNA derived from the saliva of patients with schizophrenia and bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 156B, 536-545.
28. Paquette, A.G., Lesseur, C., Armstrong, D.A., Koestler, D.C., Appleton, A.A., Lester, B.M., and Marsit, C.J. (2013). Placental HTR2A methylation is associated with infant neurobehavioral outcomes. Epigenetics 8, 796-801.
29. Bunik, V., Kaehne, T., Degtyarev, D., Shcherbakova, T., and Reiser, G. (2008). Novel isoenzyme of 2-oxoglutarate dehydrogenase is identified in brain, but not in heart. Febs J 275, 4990-5006.

30. Clinton, S.M., and Meador-Woodruff, J.H. (2004). Abnormalities of the NMDA Receptor and Associated Intracellular Molecules in the Thalamus in Schizophrenia and Bipolar Disorder. Neuropsychopharmacology 29, 1353-1362.
31. Marsman, A., van den Heuvel, M.P., Klomp, D.W., Kahn, R.S., Luijten, P.R., and Hulshoff Pol, H.E. (2013). Glutamate in schizophrenia: a focused review and meta-analysis of (1)H-MRS studies. Schizophr Bull 39, 120-129.
32. Tamminga, C.A., Southcott, S., Sacco, C., Wagner, A.D., and Ghose, S. (2012). Glutamate dysfunction in hippocampus: relevance of dentate gyrus and CA3 signaling. Schizophr Bull 38, 927-935.
33. Feinberg, A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature 447, 433-440.
34. Feinberg, A.P., and Tycko, B. (2004). The history of cancer epigenetics. Nat Rev Cancer 4, 143-153.
35. Jones, P.A., and Baylin, S.B. (2002). The fundamental role of epigenetic events in cancer. Nat Rev Genet 3, 415-428.
36. Dempster, E.L., Pidsley, R., Schalkwyk, L.C., Owens, S., Georgiades, A., Kane, F., Kalidindi, S., Picchioni, M., Kravariti, E., Toulopoulou, T., et al. (2011). Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. Hum Mol Genet 20, 4786-4796.
37. Mill, J., Tang, T., Kaminsky, Z., Khare, T., Yazdanpanah, S., Bouchard, L., Jia, P., Assadzadeh, A., Flanagan, J., Schumacher, A., et al. (2008). Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. American journal of human genetics 82, 696-711.
38. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K.L. (2009). Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1, 177-200.
39. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics 98, 288-295.
40. Boyle, P., Clement, K., Gu, H., Smith, Z.D., Ziller, M., Fostel, J.L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., et al. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biol 13, R92.
41. Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 6, 468-481.
42. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.
43. Laird, P.W. (2010). Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet 11, 191-203.
44. Bock, C. (2012). Analysing and interpreting DNA methylation data. Nat Rev Genet 13, 705-719.

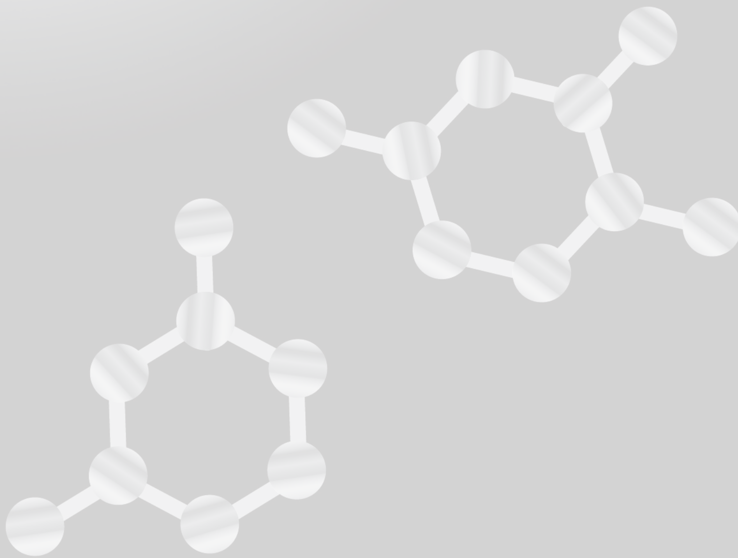
45. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11, 733-739.
46. Jin, S.G., Kadam, S., and Pfeifer, G.P. (2010). Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Res 38, e125.
47. Szulwach, K.E., Li, X., Li, Y., Song, C.X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I., et al. (2011). 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat Neurosci 14, 1607-1616.
48. Veron, N., and Peters, A.H. (2011). Epigenetics: Tet proteins in the limelight. Nature 473, 293-294.
49. Branco, M.R., Ficz, G., and Reik, W. (2012). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Genet 13, 7-13.
50. Guo, J.U., Su, Y., Zhong, C., Ming, G.L., and Song, H. (2011). Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. Cell Cycle 10, 2662-2668.
51. Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. Nat Methods.
52. Guintivano, J., Aryee, M.J., and Kaminsky, Z.A. (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics 8, 290-302.
53. Houseman, E.A., Molitor, J., and Marsit, C.J. (2014). Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data. Bioinformatics.
54. Ferrer, I., Martinez, A., Boluda, S., Parchi, P., and Barrachina, M. (2008). Brain banks: benefits, limitations and cautions concerning the use of post-mortem brain tissue for molecular studies. Cell Tissue Bank 9, 181-194.
55. Petronis, A. (2010). Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature 465, 721-727.
56. Waterland, R.A., and Jirtle, R.L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Mol Cell Biol 23, 5293-5300.
57. Weaver, I.C., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., and Meaney, M.J. (2004). Epigenetic programming by maternal behavior. Nat Neurosci 7, 847-854.
58. Lumey, L.H. (1992). Decreased birthweights in infants after maternal in utero exposure to the Dutch famine of 1944-1945. Paediatr Perinat Epidemiol 6, 240-253.
59. Kaati, G., Bygren, L.O., and Edvinsson, S. (2002). Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. Eur J Hum Genet 10, 682-688.
60. Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R., et al. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet 5, e1000602.

61. Alasaari, J.S., Lagus, M., Ollila, H.M., Toivola, A., Kivimaki, M., Vahtera, J., Kronholm, E., Harma, M., Puttonen, S., and Paunio, T. (2012). Environmental stress affects DNA methylation of a CpG rich promoter region of serotonin transporter gene in a nurse cohort. PLoS One 7, e45813.
62. Dong, E., Nelson, M., Grayson, D.R., Costa, E., and Guidotti, A. (2008). Clozapine and sulpiride but not haloperidol or olanzapine activate brain DNA demethylation. Proc Natl Acad Sci U S A 105, 13614-13619.
63. Melas, P.A., Rogdaki, M., Osby, U., Schalling, M., Lavebratt, C., and Ekstrom, T.J. (2012). Epigenetic aberrations in leukocytes of patients with schizophrenia: association of global DNA methylation with antipsychotic drug treatment and disease onset. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 26, 2712-2718.
64. Sengupta, N., and Seto, E. (2004). Regulation of histone deacetylase activities. J Cell Biochem 93, 57-67.
65. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109, 1193-1198.
66. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature.
67. Jones, C.A., Watson, D.J., and Fone, K.C. (2011). Animal models of schizophrenia. Br J Pharmacol 164, 1162-1194.
68. Guy, J., Gan, J., Selfridge, J., Cobb, S., and Bird, A. (2007). Reversal of neurological defects in a mouse model of Rett syndrome. Science 315, 1143-1147.
69. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.
70. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28, 1045-1048.
71. Chadwick, L.H. (2012). The NIH Roadmap Epigenomics Program data resource. Epigenomics 4, 317-324.
72. Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet 13, 537-551.
73. Insel, T.R. (2010). Rethinking schizophrenia. Nature 468, 187-193.
74. Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 10, 295-304.
75. Konrad, K., Firk, C., and Uhlhaas, P.J. (2013). Brain development during adolescence: neuroscientific insights into this developmental period. Dtsch Arztebl Int 110, 425-431.
76. Paus, T., Keshavan, M., and Giedd, J.N. (2008). Why do many psychiatric disorders emerge during adolescence? Nat Rev Neurosci 9, 947-957.

77. Thompson, P.M., Vidal, C., Giedd, J.N., Gochman, P., Blumenthal, J., Nicolson, R., Toga, A.W., and Rapoport, J.L. (2001). Mapping adolescent brain change reveals dynamic wave of accelerated gray matter loss in very early-onset schizophrenia. Proc Natl Acad Sci U S A 98, 11650-11655.
78. Maynard, T.M., Sikich, L., Lieberman, J.A., and LaMantia, A.S. (2001). Neural development, cell-cell signaling, and the "two-hit" hypothesis of schizophrenia. Schizophr Bull 27, 457-476.
79. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17, 1520-1528.
80. Hasan, A., Mitchell, A., Schneider, A., Halene, T., and Akbarian, S. (2013). Epigenetic dysregulation in schizophrenia: molecular and clinical aspects of histone deacetylase inhibitors. Eur Arch Psychiatry Clin Neurosci 263, 273-284.
81. Szyf, M. (2009). Epigenetics, DNA methylation, and chromatin modifying drugs. Annu Rev Pharmacol Toxicol 49, 243-263.



Nederlandse samenvatting



Na jaren van genetisch onderzoek is het duidelijk geworden dat er naast genen en omgevingsfactoren nog andere factoren een rol moeten spelen bij de ontwikkeling van eigenschappen en ziekten, namelijk epigenetica (uit het Grieks (epi -bovenop-) genetica). Epigenetische veranderingen hebben geen effect op de specifieke volgorde van het DNA (desoxyribonucleïnezuur) -ze veranderen dus niet de onderliggende genetische code- maar ze hebben wel invloed op de structuur van het DNA en daarbij effect op genfunctie, wat weer kan resulteren in verschillen in eigenschappen van de cel en het organisme. Bijvoorbeeld: de volgorde van DNA in alle cellen van het lichaam is hetzelfde, maar het aan- of uitzetten van verschillende genen maakt dat cellen differentiëren naar verschillende soorten cellen. Spiercellen hebben bijvoorbeeld dezelfde DNA volgorde als hartcellen, maar in spiercellen staan andere genen 'aan' dan in hartcellen. Dit proces wordt gereguleerd door epigenetica.

Epigenetica is een verzamelnaam voor verschillende processen en 'markers', en in dit proefschrift zal het voornamelijk gaan over DNA methylatie. DNA methylatie is een chemische modificatie van het DNA. Zogenaamde cytosines (1 van de 4 basen of bouwstenen van het DNA; cytosine, guanine, adenine en thymine) kunnen worden 'gemethyleerd' door middel van de toevoeging van een methylgroep. Dit proces vindt vooral plaats wanneer een cytosine naast een 'guanine' ligt. Vaak zien we dat deze CpG plaatsen (ook wel loci genoemd) clusteren in grote groepen, de zogenaamde CpG eilanden, die vaak in een promotor gebied liggen. Promotors liggen meestal aan het begin van een gen en reguleren expressie van dat gen. Genexpressie is de mate waarin een gen vertaald wordt naar RNA (ribonucleïnezuur) en daarna naar eiwit. Het algemene idee is dat DNA methylatie van promotors de expressie van genen kan remmen, waardoor er minder eiwit wordt gemaakt. Echter, langzaam wordt duidelijk dat dit proces complexer is dan eerder gedacht.

DNA methylatie kan variëren tussen personen en veranderen met de tijd. Er is gesuggereerd dat methylatie deels gereguleerd wordt door de onderliggende genetische code, bijvoorbeeld door varianten in het DNA genaamd SNPs ('Single Nucleotide Polymorphisms', een verandering van een enkel basenpaar). Als SNPs geassocieerd zijn met methylatie op CpG loci worden ze 'methylation quantitative trait loci' (mQTLs) genoemd. Mensen met een bepaalde variant, of allel, in het DNA hebben dan een verhoogde of verlaagde methylatie van een gen). Recente technieken maken het mogelijk methylatie op grote schaal te meten waardoor verschillen in methylatielevels geïdentificeerd kunnen worden, bijvoorbeeld tussen patiënten en gezonde mensen.

In dit proefschrift ligt de focus op studie van DNA methylatie betrokken bij neuropsychiatrische stoornissen, met name schizofrenie. Schizofrenie is een psychiatrische aandoening, die bij ongeveer 1% van de bevolking voorkomt. Het wordt gekenmerkt door een complex van

symptomen, zoals hallucinaties, wanen, onsamenhangende spraak, ernstig chaotisch gedrag en negatieve symptomen (zoals lusteloosheid, sociale terugtrekking en minder plezier vinden in activiteiten). Voor het stellen van de diagnose is het van belang dat deze symptomen langer dan een half jaar aanwezig zijn en de patiënt beperken in het dagelijks leven. Het is een van de meest invaliderende psychiatrische stoornissen. De precieze ontstaanswijze van schizofrenie is (nog) niet geheel duidelijk en waarschijnlijk een samenspel van erfelijke- en omgevingsfactoren. De erfelijke component van de variatie van de aandoening in de populatie wordt geschat op ongeveer 80%, maar ook omgevingsfactoren zoals stress, trauma en cannabis gebruik spelen waarschijnlijk een belangrijke rol.

Er is uitgebreid genetisch onderzoek gedaan naar de onderliggende factoren van schizofrenie. Hierbij wordt gezocht naar verschillen in de DNA volgorde tussen patienten en gezonde mensen. Hoogstwaarschijnlijk spelen honderden of zelfs duizenden variaties in het DNA (in en om genen) een rol, maar elk met slechts een gering effect op het ontstaan van de aandoening. Tot nu toe zijn er enkele tientallen regio's binnen het DNA ontdekt die samen ongeveer 5% van het genetische aandeel van schizofrenie verklaren.
Het bestuderen van andere aspecten van het genoom, zoals DNA methylatie, kan bijdragen in het verklaren van de nog onbekende, genetische- en omgevingsrisicofactoren.
Ondanks dat we het mechanisme achter DNA methylatie deels begrijpen is er nog weinig bekend over de gevolgen van verschillen in methylatie tussen personen en of en hoe verschillen in methylatie bijdragen aan het ontstaan van een ziekte. Een ontregeling van methylatie, en daarmee ontregeling van genexpressie, zou kunnen leiden tot grotere vatbaarheid voor bepaalde aandoeningen.

Met dit proefschrift wilden we dan ook de relatie tussen DNA (genotype), DNA methylatie en genexpressie onderzoeken, om meer inzicht te krijgen in het ontstaan van schizofrenie.

In **hoofdstuk 2** van dit boekje beschrijven we een onderzoek met als doel de invloed van de mate van methylatie op genexpressie te bekijken. Zoals hierboven beschreven, wordt over het algemeen aangenomen dat methylatie van (de promotor) van een gen leidt tot verminderde expressie van dit gen. Echter, er zijn onderzoeken die een omgekeerd effect beschrijven (methylatie leidt tot meer expressie van het betreffende gen). De tegenstrijdigheid van deze bevindingen heeft tot gevolg dat men denkt dat het mechanisme complexer is dan eerder gedacht. We hebben de relatie tussen methylatielevels en expressielevels van genen in kaart gebracht in een groep van bijna 150 gezonde mensen. Ook is de relatie van methylatie- en expressielevels met genotype onderzocht. Uit ons onderzoek blijkt nogmaals de complexiteit van regulatie van genen. In het hoofdstuk beschrijven we dat een deel van de relaties tussen methylatie en genexpressie afhankelijk zijn van bepaalde SNPs. Door deze

driehoeksverbanden (SNP, methylatie en genexpressie) nader te bestuderen, observeerden wij dat voor de meesten genen SNPs geassocieerd zijn met een verandering in DNA methylatie, wat vervolgens een effect heeft op genexpressie. Voor een aantal genen is het echter anders; genexpressie lijkt daar een effect te hebben op methylatie, maar verder onderzoek is nodig om deze bevindingen te bevestigen. Vervolgens constateerden wij dat de CpG loci -waarvan methylatielevels zijn gemeten- die geassocieerd zijn met zowel genexpressie als SNPs, vooral net buiten CpG eilanden liggen, in de flankerende gebieden van deze eilanden. Uit onderzoek was al gebleken dat methylatielevels in deze gebieden meer variëren dan methylatielevels in de eilanden zelf en ons onderzoek heeft laten zien dat deze methylatielevels vaker met expressie en genotype geassocieerd zijn. Tenslotte hebben we in dit onderzoek methylatie- en expressielevels van genen gegroepeerd in modules, waarbij genen met gelijke patronen in dezelfde module terecht kwamen. Deze modules bleken verrijkt te zijn met verschillende biologische termen, hetgeen de biologische relevantie van de modules aanduidt. Hoewel de methylatie- en expressiemodules over het algemeen slechts weinig overlappende genen hadden, was er toch een aantal modules die sterk gecorreleerd was met elkaar. Dit zou kunnen duiden op *trans* effecten op afstand.

Grootschalige genoom-wijde associatie studies (GWAS) uitgevoerd door het Psychiatrisch GWAS consortium (PGC) in duizenden patiënten en gezonde mensen hebben geleid tot de identificatie van tientallen SNPs geassocieerd met schizofrenie. Echter, deze SNPs dragen maar een fractie bij van de totale onderliggende genetische achtergrond van deze aandoening. In **hoofdstuk 3** hebben we de top 6000 SNPs, waarvan de meeste matig geassocieerd zijn met schizofrenie, gecombineerd met genexpressie data voor het bijbehorende gen. Hierdoor hoopten we meer signaal te hebben om additionele genen te kunnen identificeren. We onderzochten of de circa 6000 SNPs die het sterkst geassocieerd zijn met schizofrenie ook daadwerkelijk effect hebben op genexpressie. Vervolgens hebben we gekeken of er verschil is in genexpressie van deze genen tussen patiënten en gezonde mensen. Zeven genen werden ontdekt waarvan er drie hetzelfde signaal (effect van de SNP gaat dezelfde kant op) lieten zien als in de eerder genoemde PGC analyse. Deze drie genen liggen in een gebied op chromosoom zes dat al vaker in verband is gebracht met schizofrenie. Door het combineren van SNP en genexpressie data hebben we dus extra genen gevonden die niet ontdekt waren in de grote GWAS omdat het signaal te zwak was.

De resultaten van de PGC studie hebben we ook gebruikt in **hoofdstuk 4**, waar we deze informatie gecombineerd hebben met methylatie data. Het doel van deze studie was te onderzoeken of SNPs die methylatielevels reguleren vaker geassocieerd zijn met schizofrenie. Specifieker nog, waren we geïnteresseerd of SNPs, die methylatielevels reguleren die verschillen tussen patiënten en controles, én die geassocieerd zijn met genexpressie, meer

verrijkt zijn met schizofrenie-gerelateerde SNPs dan de SNPs die niet-specifieke methylatielevels reguleren. Eerst hebben we de methylatielevels geïdentificeerd die verschillend waren tussen patiënten en controles. Daarna onderzochten we welke methylatielevels geassocieerd waren met expressielevels, en of deze expressielevels ook verschilden tussen patiënten en controles. Vervolgens testten we door welke SNPs deze methylatielevels gereguleerd werden. Deze SNPs werden uiteindelijk opgezocht in de PGC studie om te bepalen of deze SNPs, die methylatie reguleren, sterker met schizofrenie geassocieerd zijn dan SNPs die geen methylatielevels beïnvloeden. Uit ons onderzoek is gebleken dat deze met methylatie geassocieerde SNPs inderdaad sterker gelinkt zijn aan schizofrenie. De meest significante SNP is gerelateerd aan methylatie van het gen *CALHM1*, een calcium homeostase modulator. Calcium metabolisme is al vaker in verband gebracht met schizofrenie en inmiddels is een studie gepubliceerd waarin de regio waarin dit gen ligt, geassocieerd wordt met schizofrenie. Onze studie onderstreept het belang van het combineren van methylatie en genexpressie data met SNPs, wat kan leiden tot verrijking van ziekte-gerelateerde varianten.

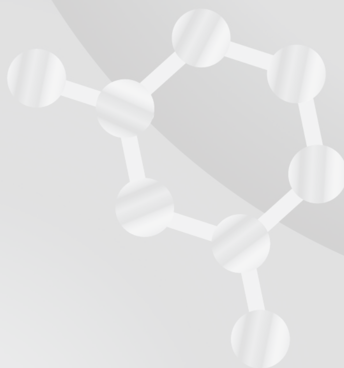
In **hoofdstuk 5** hebben we allel-specifieke methylatie (ASM) bestudeerd. Voor de meeste CpG loci is methylatie op beide DNA strengen hetzelfde. Echter, in sommige gevallen kan de methylatiestatus verschillen per DNA streng, zoals bij X-chromosoom inactivatie en bij genomische imprinting elders in het genoom. Allel-specifieke methylatie is sterk geassocieerd met genotype vlakbij de CpG locus. Er zijn inmiddels verschillende studies uitgevoerd om de relatie tussen genotype en CpG methylatie te onderzoeken maar het is niet duidelijk hoeveel de met deze methoden geïdentificeerde SNPs bijdragen aan de variabiliteit in DNA methylatie. In deze studie hebben wij op drie verschillende manieren de relatie tussen CpG methylatie en SNPs onderzocht, waarbij we ook de genetische bijdrage aan de variatie in methylatie bepaalden. Dit omvatte de in de vorige hoofdstukken uitgevoerde mQTL analyse, ASM analyse en 'mid-parent-offspring' (MPO) analyse, waarbij de methylatiepatronen van de kinderen vergeleken wordt met de ouders. Voor deze studie gebruikten we de gegevens van 22 families, waarvan binnen iedere familie één kind gediagnosticeerd is met schizofrenie. In plaats van de in eerdere hoofdstukken gebruikte SNP en methylatie chips (die 27.000 CpG loci in het genoom meten), is deze data verkregen door meer dan 400.000 CpG loci te sequencen (het bepalen van de nucleotide volgorde van het DNA). Ook de genotypes konden zo vastgesteld worden in één handeling, wat kosten en tijd bespaarde. Na het uitvoeren van deze drie analyses bleek dat de gevonden SNP-CpG associaties niet helemaal overlapten tussen de drie analyses. Soms werden bepaalde SNP-CpG combinaties slechts door één methode gedetecteerd. Ook ontdekten we dat de meeste allel-specifieke methylatie veroorzaakt werd door SNPs op CpG loci. Een voorbeeld hiervan is wanneer de ene DNA streng een cytosine bevat die gemethyleerd kan worden, terwijl de andere DNA streng op die plaats een andere nucleotide bevat die niet gemethyleerd kan worden. Er ontstaat zo

een verschil in methylatie tussen beide strengen. Tenslotte vonden we honderden variabele gemethyleerde gebieden die tenminste vijf CpG loci bevatten die geassocieerd zijn met een zogenaamde 'variatie SNP'. Dit betekent dat een bepaald allel van een SNP geassocieerd is met een stijging of daling van de methylatie variatie ten opzichte van het gemiddelde methylatienniveau, gemeten in meerdere personen. Dit ondersteunt de hypothese dat een SNP niet alleen een effect heeft op de gemiddelde methylatie waarde maar ook op de variabiliteit van de methylatie waarde.

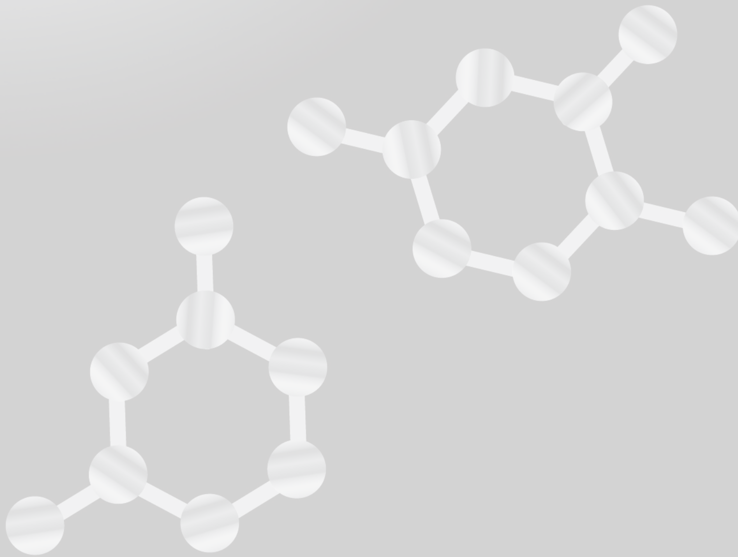
Zoals eerder genoemd zijn DNA methylatie en ook genexpressie gedeeltelijk weefselspecifiek. Omdat schizofrenie een brein aandoening is, hebben we in **hoofdstuk 6** gebruik gemaakt van beschikbare hersenweefsels in tegenstelling tot de in de hiervoor genoemde studies die uitgevoerd zijn met data uit bloed. De meest significante schizofrenie SNP uit de eerder genoemde PGC studie betreft een transcript voor het gen microRNA-137 (miR-137). MicroRNA's (micro ribonucleine zuur - miRNA's) zijn kleine niet-coderende RNA moleculen die de expressie van RNA kunnen reguleren. Naast expressie kunnen miRNA's methylatie reguleren maar DNA methylatie kan ook de expressie van miRNA's reguleren. Deze interactie tussen miRNA's en DNA methylatie kan meer inzicht verschaffen in de moleculaire regulatie en mogelijk in de ontwikkeling van zieke als deze interactie ontregeld is. Daarom hebben wij onderzocht of het eerder genoemde miRNA (miR-137) geassocieerd is met methylatielevels. Allereerst bestudeerden we de expressie van miR-137 in verschillende brein regio's. We vonden een aantal regio's waarin miR-137 in grote mate tot expressie komt, namelijk: corticale regio's, het limbisch systeem en de basale ganglia. Eerder onderzoek heeft aangetoond dat deze hersengebieden afwijkingen vertonen in schizofrenie patiënten. Naast de bevindingen van de PGC (een verschil in genotype tussen patiënten en controles) vinden wij dus ook een hoge expressie van dit gen in de gebieden die abnormaliteiten vertonen in patiënten. Vervolgens onderzochten we de interacties tussen de expressie van miR-137 en DNA methylatie van genen uit het hele genoom. Dit resulteerde in drie suggestieve associaties. De sterkste bevinding betreft het gen *HTR2A*, een serotonine receptor (type-2), waarvan bekend is dat de promotor epigenetisch gereguleerd wordt. Dit gen speelt een belangrijke rol in neuronale processen en is recentelijk zelfs in verband gebracht met schizofrenie. De tweede bevinding betreft het gen 2-oxoglutarate dehydrogenase-like (OGDHL). Dit gen komt tot expressie in de hersenen waar het een rol speelt in glutamaat synthese. Glutamaat is een belangrijke neurotransmitter en in eerdere studies is gesuggereerd dat dit proces betrokken is bij neuropsychiatrische aandoeningen zoals schizofrenie. Onze bevindingen linken een eerder gevonden schizofrenie locus, miR-137, aan mogelijk voor schizofrenie belangrijke serotonine en glutamaat signaaltransductie routes.

Tot slot bevat **hoofdstuk 7** een samenvatting met een algemene discussie. Hierin worden ook de technische en biologische beperkingen besproken waar we tijdens de beschreven onderzoeken op stuitten. Verder worden toekomstplannen aangedragen en eindigt dit hoofdstuk met een korte conclusie.

Samenvattend beschrijft dit proefschrift genoom-wijde methodes om verschillende types genomische informatie te combineren en analyseren. Door het integreren van methylatie data met genexpressie en genotype data hebben we loci geïdentificeerd die mogelijk bijdragen aan de ontwikkeling van schizofrenie en daarom interessant zijn voor vervolgonderzoek. Hiermee laten we zien dat DNA methylatie een relevant en veelbelovend hulpmiddel is in het ontrafelen van deze neuropsychiatrische aandoening.



Dankwoord / Acknowledgements



Hetgeen als eerste -en vaak als enige- gelezen wordt, is tevens ook het belangrijkste gedeelte van dit proefschrift. Zonder de hulp van veel mensen was dit proefschrift namelijk niet tot stand gekomen.

Allereerst wil ik een grote groep mensen bedanken zonder wie onderzoek doen onmogelijk is: de patiënten en gezonde personen die bereid zijn geweest hun genetisch materiaal af te staan ten behoeve van wetenschappelijk onderzoek.

Daarna mijn promotor **Roel Ophoff** en co-promotor **Marco Boks** die mij deze kans gegeven hebben en vertrouwen in mij hadden ondanks dat ik rechtstreeks van het HBO kwam. Beste **Roel**, heel veel dank gaat uit naar jou. Hoewel je na mijn eerste jaar geëmigreerd bent naar Los Angeles, ben je erg betrokken gebleven. Van dagelijks even jouw kantoor binnen lopen naar wekelijks geplande Skype calls was wel even wennen maar gelukkig waren we daarin allebei flexibel ondanks het tijdsverschil. Bedankt voor je begrip, geduld en positiviteit. Ik heb veel van je geleerd! Ook de mogelijkheid om deel te nemen aan congressen in het buitenland alsmede een project te doen bij jouw nieuwe 'thuisbasis' in LA, heb ik zeer op prijs gesteld. Je persoonlijke interesse en gesprekken over niet-werk gerelateerde zaken (zoals geloof en wetenschap) heb ik altijd gewaardeerd. Nogmaals bedankt voor alles!
Beste **Marco**, als ik jou in één woord zou moeten omschrijven is het "super-enthousiasmerende-altijd-positief-blijvende-overal-tijd-voor-makende-in-verschillende-gedaantes-verschijnende-statistiek-goeroe". Naast je eeuwig enthousiasme en optimisme maak jij van problemen (de zoveelste foutmelding in een script, artikelen die afgewezen worden) een uitdaging. Na overleg met jou zag ik alles vaak weer helemaal zitten. Jouw optimisme werkt aanstekelijk. Wel moest ik af en toe wennen aan je gedaanteverwisselingen (baard, kaal, geverfd haar), maar gelukkig bereidde je me later voor door me vast te waarschuwen in een mailtje voordat we overleg hadden. Leuk dat we nu bij dezelfde divisie werken en we vast nog wel zullen samenwerken/overleggen in de komende tijd! En om even in jouw stijl - groeten in talen uit Verweggistan- af te sluiten: Whakawhetai!

De leden van de leescommissie, **Paul de Bakker**, **Steve Horvath**, **Hilleke Hulshoff Pol**, **Nine Knoers** en **Jan Veldink**, wil ik graag bedanken voor het kritisch beoordelen van mijn manuscript.

Dan m'n (oud) flexkamergenoten: **Albertien**, **Androniki**, **Anne-Marije**, **Annet**, **Balder**, **Clara**, **Eric**, **Esther**, **Femke**, **Flip**, **Karen**, **Jacobine**, **Jelena**, **Jessica**, **Jytte**, **Laurent**, **Linda**, **Maarten**, **Ruben**, **Sander**, **Sara**, **Sasha**, **Simone**, **Vinicius**.
Bedankt voor alle gezelligheid, versieringen, etentjes, brainstorms over van alles en nog wat, (jellybrain)experimenten, creativiteit, cute-overload mailtjes, slechte en grappige youtube

filmpjes, roasts, top-zoveel-lijsten van hilarische, schattige en/of bizarre gebeurtenissen, momenten, dieren en mensen! En laten we the Hoff niet vergeten! Zonder jullie had ik heel wat muzikaal talent gemist!

Jacobine, wat leuk dat je het laatste jaar nog m'n "pleeg-oma-aio" bent geweest. Bedankt voor het beoordelen van mijn artikelen en hoofdstukken en voor de alles-wat-je-moet-weten-om-te-promoveren-tips! De gezellige koffie/thee/water-drinkmomenten hoeven we gelukkig niet te missen aangezien we op een steenworp afstand van elkaar werken. En laat ik de handige huis-tuin-en keuken adviezen niet vergeten (en dan vooral in de categorie baby's en kleine kinderen). Dank daarvoor! :-) **Flip**, buurman, bedankt voor al je hulp al dan niet computer gerelateerd. SupertHoff dat je 99% van mijn computerproblemen kon oplossen door alleen even naar m'n scherm te kijken! Hopelijk mag ik straks nog gebruik maken van die special skills ;-). Wel jammer dat ik niet met je kon twisten over de smaak van koffie. Ik heb echt geprobeerd om het te leren drinken in de afgelopen 4 jaar maar ik heb het toch maar opgegeven. En natuurlijk hulde aan je photoshoppers! (Wordt het niet eens tijd voor een "best-of" boek?) **Jessica**, ik ben blij dat we tegelijk gaan promoveren (die planning maakten we 4 jaar geleden al en ik heb mijn verlofregelingen zo aangepast dat het ook precies goed uitkwam ;-)). Bedankt voor je fancy scriptjes en het lezen van mijn teksten. Ik heb veel van je geleerd (zowel wat betreft inhoudelijke dingen als daarbuiten, je hebt altijd over een antwoord op)! Onze uitjes (van Kerkrade tot Montreal) waren ook heel gezellig! Je bent een fijne kamergenoot :-). **Simone**, bedankt voor de samenwerking waardoor we 2 artikelen samen hebben! Het heeft mij erg geholpen dat ik voor mijn eerste schrijfervaring jou als voorbeeld had. Bedankt ook dat ik je altijd kon en kan mailen als ik vragen heb. Ik ben alleen vergeten te vragen hoe je alles toch zo goed kan combineren (hard werken, feesten, shoppen en dan ook nog tijd voor series, boeken en de kapper). Wel jammer dat je niet bij het tHoffste promotiefilmpje ooit gemaakt kon zijn, namelijk die van jezelf. Gelukkig hebben we de beelden nog :P. **Jytte**, wat ben ik blij dat jij het leuk vindt om in teksten te kliederen :-). Jij wist mijn soms verwarde zinnen weer in mooi Engels (of Nederlands) te verwoorden en betere uitleg uit me te trekken. Hopelijk is mijn proefschrift nu een beetje begrijpelijk geworden! Heel erg bedankt! **Ruben**, bedankt voor alle gezelligheid! En bedankt dat je mijn boekje niet hebt stukgemaakt! **Eric,** leuk dat jij al in LA zat toen ik daarheen mocht en we samen de Mac ontdekt hebben. En laten we onze roadtrips naar Hollywood en de Walmart niet vergeten (bijna even gedenkwaardig)! **Clara**, jij hebt me heel veel jaar geleden enthousiast gemaakt voor genetica toen ik stage bij je liep en dankzij jou ben ik ook hier terecht gekomen. Dankjewel voor alles wat je me hebt geleerd (inclusief cute-overload)! **Jelena**, wat leuk dat we weer collega's worden! En bedankt voor de tip ;-). Gezellig, nu kunnen we vaker carpoolen naar dat gehucht tussen Everdingen en Leerdam :-). **Esther,** bedankt voor alle gezelligheid en het organiseren van van alles en nog wat. Wel jammer

dat de cocktailcommissie is vertrokken.. **Sara,** thank you so much for critically reading my introduction and discussion (amongst others), and checking English. It was nice to have some company during the late nights in the lab (although for me it was only temporary ;-)). **Androniki**, thanks for answering my statistical questions! I cannot promise I won't bother you again with that.. **Laurent** thanks for the advice to counting chairs in an airplane to the nearest exit (I won't mention the exact theory here).

De rest van de (oud) collega's van de afdeling Medische Genetica wil ik ook bedanken voor de discussies, samenwerkingen en gezelligheid op de afdeling: **Annet**, **Behrooz**, **Bobby**, **Carla**, **Carolien**, **Diane**, **Dick**, **Edwin**, **Gijs** (x2), **Glen**, **Glenn**, **Harry**, **Ies**, **Irena**, **Ivo**, **Jackie**, **Jacqueline**, **Kirsten**, **Magdalena**, **Marijn**, **Mark**, **Marlous**, **Martin** (x2), **Mirjam**, **Monique**, **Nayia**, **Nine**, **Paul**, **Stef**, **Terry**, **Wigard**. In het bijzonder bedank ik graag **Bobby** voor het onder je hoede nemen van mij sinds Roel naar LA is gegaan en voor het nodige papierwerk dat zo her en der ondertekend moest worden; **Carolien**, bedankt dat ik altijd bij je terecht kon voor (met name statistiek) vragen!; en **Jacqueline**, dank je wel voor al je regelwerk en de gezelligheid!

De studenten **Dave**, **Terry** en **Djie** wil ik ook bedanken voor hun bijdrage aan de projecten. **Dave** (oud studiegenootje), alhoewel je eigenlijk Simone's student was hebben we toch veel samengewerkt, wat heeft geresulteerd in een mooi artikel! **Terry**, jij hebt veel werk verricht voor de grootste studie in dit boekje. Bedankt daarvoor! En **Djie**, is er ook iets wat jij niet kan? Ik heb veel van jou geleerd (misschien wel meer dan andersom ;-)) en gebruik je scripts nog steeds :-), thanks!

Van de afdeling Psychiatrie wil ik de volgende mensen bedanken voor samenwerkingen en werkoverleggen: **Afke**, **Annabel**, **Anneloes**, **Chris**, **Elemi**, **Eske**, **Jeanette**, **Jurjen**, **Maartje**, **Marijke**, **Metten**, **Mirjam** en **Steven**.

During my visit to UCLA I have met some great people: **Peter**, **Steve**, and **Yafeng**, thanks for all the bioinformatic support and collaborations (resulting in papers)! **Eric**, **Katherine**, **Marisa** and **Yoon**, thanks for the fun!
Anil, ik wil je enorm bedanken voor je inzet en grote bijdrage aan het laatste hoofdstuk in mijn boekje. Ik kreeg mailtjes van jou op tijden die gezien het tijdsverschil niet echt normaal waren, naarmate de deadline van mijn boekje naderde. Je zocht dingen tot de bodem uit en ik heb veel van je geleerd over hersenen (wat in mijn nieuwe baan weer van pas komt ;-))!

Ook een belangrijke groep mensen die misschien vaak ondergewaardeerd wordt zijn de mensen van de IT: **Eric**, **Dennis**, **Marc** en **Wim**. Bedankt dat ik altijd even binnen kon lopen

voor alles wat met computers te maken heeft! Verder wil ik de medewerkers van **Surfsara** bedanken, en in het bijzonder **Willem**; bedankt dat je altijd de tijd nam voor uitleg om mijn jobs efficiënter te maken!

Alle co-auteurs ben ik erg dankbaar voor hun bijdrage -in welke vorm dan ook- aan artikelen. Naast de mensen die ik al eerder genoemd heb zijn dat: **Fabrice**, **Ho-Lim**, **Jae-Hun**, **Kun**, **Leonard**, **Nongluk**, **René**, **Tina** en **Wiepke**. Thank you all for your collaborations!

Tot slot wil ik van deze gelegenheid gebruik maken om mijn lieve familie en vrienden te bedanken. Allereerst (ome) **Guus**; bedankt voor het ontwerpen van de kaft van mijn boekje! Het is supergaaf geworden! De rest van de familie **van Eijk** en **ten Berge**, ook bedankt voor de interesse en de gezelligheid in het oosten van het land! M'n **volleybalmaatjes**: ik ga jullie niet allemaal bij naam noemen maar bedankt allemaal voor de afleiding! Hierbij beloof ik ook meteen dat ik de komende tijd weer trouw zal komen (na wat maanden van sporadische aanwezigheid). De **buurtjes** van de Prins Bernhardstraat; mijn collega's begrijpen niet dat je kan wonen in een dorp waar niet eens OV is maar dat het zo'n fijne buurt is komt ook door jullie! De **mommies**: bedankt voor alle steun, afleiding en gezelligheid (via verschillende kanalen)!

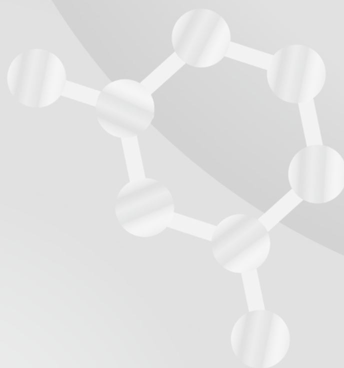
Verder wil ik al mijn vrienden en vriendinnen bedanken, die ik tijdens school en studie heb leren kennen en waarvan ik sommigen zelfs al mijn hele leven ken! In het bijzonder noem ik mijn lieve vriendinnen **Anemee**, **Ellen** en **Jorieke**. Bedankt dat jullie er altijd voor me zijn en we uren kunnen keuvelen over van alles en nog wat. Ondanks dat we allemaal totaal ander werk doen zijn jullie altijd geïnteresseerd in mijn onderzoek. **Ellen**, bedankt dat je m'n paranimf wilt zijn! **Jo** (en Irma namens Ingrid), alvast bedankt voor de foto's tijdens mijn verdediging (photoshop m'n angstzweet maar weg ;-))!

Mijn lieve **schoonouders**, **schoonzusjes** en **zwagers** wil ik bedanken voor alles! Ik ben heel blij met jullie. **Marja**, bedankt voor het vele oppassen op Ninthe. Sorry voor de keren dat Niels haar onverwachts bij jou bracht als ik ineens weer naar Utrecht ging om mijn boekje af te schrijven.

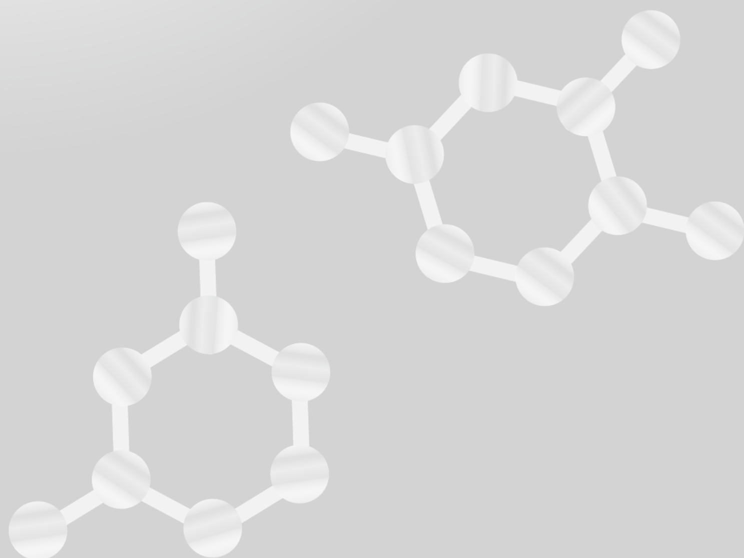
Mijn lieve zusje, **Francis**; ik had vroeger nooit gedacht dat ik dit nog eens zou zeggen maar je bent mijn grote steun en toeverlaat! Bedankt voor het op sleeptouw nemen van Ninthe, voor je hulp in huis, dat ik altijd (boodschappen)lijstjes mag doorgeven, en voor de gezellige dingen (vakanties, musicals, etentjes etc.). Kortom: bedankt dat je altijd voor me (ons) klaar staat! Ook letterlijk tijdens mijn verdediging als paranimf.

Lieve **pap** en **mam**, bedankt dat jullie er altijd voor me zijn! Bedankt voor jullie steun en vertrouwen! Jullie hebben altijd achter me gestaan in de keuzes die ik maakte. Fijn dat ik altijd bij jullie terecht kan! **Pap**, je pensioen was echt goed getimed! Bedankt voor alle honderden uren en rondjes die je hebt gewandeld met Ninthe zodat ik nog weer even wat kon doen op mijn 'vrije' dag (vooral in de laatste fase van dit traject). Lief dat ik Ninthe altijd bij je mocht brengen of dat je naar mij toe kwam. **Mam**, ik vind het heel lief dat je je enige vrije dag toegewijd hebt aan je kleindochter. Dat waardeer ik echt! Ik hoop dat ik voor Ninthe net zo'n goede moeder word als jij bent.

Lieve **Niels**, ondanks dat je zelf altijd superdruk bent met je eigen bedrijf (samen met je vader), kwam daar (vooral aan het eind) nog een hoop extra op je af zoals de zorg voor Ninthe en huishoudelijke dingen. Bedankt dat je me altijd liet werken als dat nodig was en je mij ontlastte (ja ik moet wat makkelijker gaan worden en denken). Maar ik vind het vooral schattig dat je, ondanks dat je totaal geen idee hebt waar ik mee bezig was, je toch steeds zei dat ik het wel kon en het ging halen. Bedankt voor je vertrouwen in mij en dat ik altijd mezelf kan zijn bij je! Sinds we onze lieve, mooie dochter mochten krijgen is het alleen maar leuker en fijner geworden! Dank dat je zo'n lieve papa bent voor haar.
Liefste **Ninthe**, mijn altijd vrolijke snoetje, mama houdt van jou!



Publication list



Peer reviewed Pubmed publications

1. Elbers CC, **van Eijk KR**, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC. - Using genome-wide pathway analysis to unravel the etiology of complex diseases - Genetic Epidemiology: 2009 Jul;33(5):419-31. doi: 10.1002/gepi.20395.

2. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, **van Eijk K**, Hermjakob H. The IntAct molecular interaction database in 2010 - Nucleic Acids Research: 2010 Jan;38(Database issue):D525-31. doi: 10.1093/nar/gkp878. Epub 2009 Oct 22.

3. de Jong S, **van Eijk KR**, Zeegers DW, Strengman E, Janson E, Veldink JH, van den Berg LH, Cahn W, Kahn RS, Boks MP, Ophoff RA; PGC Schizophrenia (GWAS) Consortium - Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes - European Journal of Human Genetics*:* 2012 Sep;20(9):1004-8. doi: 10.1038/ejhg.2012.38. Epub 2012 Mar 21.

4. Stein JL, Medland SE, Vasquez AA, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Bartecek R, Bergmann Ø, Bernard M, Brown AA, Cannon DM, Chakravarty MM, Christoforou A, Domin M, Grimm O, Hollinshead M, Holmes AJ, Homuth G, Hottenga JJ, Langan C, Lopez LM, Hansell NK, Hwang KS, Kim S, Laje G, Lee PH, Liu X, Loth E, Lourdusamy A, Mattingsdal M, Mohnke S, Maniega SM, Nho K, Nugent AC, O'Brien C, Papmeyer M, Pütz B, Ramasamy A, Rasmussen J, Rijpkema M, Risacher SL, Roddey JC, Rose EJ, Ryten M, Shen L, Sprooten E, Strengman E, Teumer A, Trabzuni D, Turner J, **van Eijk K,** van Erp TG, van Tol MJ, Wittfeld K, Wolf C, Woudstra S, Aleman A, Alhusaini S, Almasy L, Binder EB, Brohawn DG, Cantor RM, Carless MA, Corvin A, Czisch M, Curran JE, Davies G, de Almeida MA, Delanty N, Depondt C, Duggirala R, Dyer TD, Erk S, Fagerness J, Fox PT, Freimer NB, Gill M, Göring HH, Hagler DJ, Hoehn D, Holsboer F, Hoogman M, Hosten N, Jahanshad N, Johnson MP, Kasperaviciute D, Kent JW Jr, Kochunov P, Lancaster JL, Lawrie SM, Liewald DC, Mandl R, Matarin M, Mattheisen M, Meisenzahl E, Melle I, Moses EK, Mühleisen TW, Nauck M, Nöthen MM, Olvera RL, Pandolfo M, Pike GB, Puls R, Reinvang I, Rentería ME, Rietschel M, Roffman JL, Royle NA, Rujescu D, Savitz J, Schnack HG, Schnell K, Seiferth N, Smith C, Steen VM, Valdés Hernández MC, Van den Heuvel M, van der Wee NJ, Van Haren NE, Veltman JA, Völzke H, Walker R, Westlye LT, Whelan CD, Agartz I, Boomsma DI, Cavalleri GL, Dale AM, Djurovic S, Drevets WC, Hagoort P, Hall J, Heinz A, Jack CR Jr, Foroud TM, Le Hellard S,

Macciardi F, Montgomery GW, Poline JB, Porteous DJ, Sisodiya SM, Starr JM, Sussmann J, Toga AW, Veltman DJ, Walter H, Weiner MW; Alzheimer's Disease Neuroimaging Initiative; EPIGEN Consortium; IMAGEN Consortium; Saguenay Youth Study Group, Bis JC, Ikram MA, Smith AV, Gudnason V, Tzourio C, Vernooij MW, Launer LJ, DeCarli C, Seshadri S; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium, Andreassen OA, Apostolova LG, Bastin ME, Blangero J, Brunner HG, Buckner RL, Cichon S, Coppola G, de Zubicaray GI, Deary IJ, Donohoe G, de Geus EJ, Espeseth T, Fernández G, Glahn DC, Grabe HJ, Hardy J, Hulshoff Pol HE, Jenkinson M, Kahn RS, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meyer-Lindenberg A, Morris DW, Müller-Myhsok B, Nichols TE, Ophoff RA, Paus T, Pausova Z, Penninx BW, Potkin SG, Sämann PG, Saykin AJ, Schumann G, Smoller JW, Wardlaw JM, Weale ME, Martin NG, Franke B, Wright MJ, Thompson PM; Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium - Identification of common variants associated with human hippocampal and intracranial volumes - Nature Genetics: 2012 Apr 15;44(5):552-61. doi: 10.1038/ng.2250.

5. Bis JC, DeCarli C, Smith AV, van der Lijn F, Crivello F, Fornage M, Debette S, Shulman JM, Schmidt H, Srikanth V, Schuur M, Yu L, Choi SH, Sigurdsson S, Verhaaren BF, DeStefano AL, Lambert JC, Jack CR Jr, Struchalin M, Stankovich J, Ibrahim-Verbaas CA, Fleischman D, Zijdenbos A, den Heijer T, Mazoyer B, Coker LH, Enzinger C, Danoy P, Amin N, Arfanakis K, van Buchem MA, de Bruijn RF, Beiser A, Dufouil C, Huang J, Cavalieri M, Thomson R, Niessen WJ, Chibnik LB, Gislason GK, Hofman A, Pikula A, Amouyel P, Freeman KB, Phan TG, Oostra BA, Stein JL, Medland SE, Vasquez AA, Hibar DP, Wright MJ, Franke B, Martin NG, Thompson PM; Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium, Nalls MA, Uitterlinden AG, Au R, Elbaz A, Beare RJ, van Swieten JC, Lopez OL, Harris TB, Chouraki V, Breteler MM, De Jager PL, Becker JT, Vernooij MW, Knopman D, Fazekas F, Wolf PA, van der Lugt A, Gudnason V, Longstreth WT Jr, Brown MA, Bennett DA, van Duijn CM, Mosley TH, Schmidt R, Tzourio C, Launer LJ, Ikram MA, Seshadri S; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium.
   COLLABORATORS:
   Stein JL, Medland SE, Arias Vasquez A, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Bartecek R, Bergmann Ø, Bernard M, Brown AA, Cannon DM, Chakravarty M, Christoforou A, Domin M, Grimm O, Hollinshead M, Holmes AJ, Homuth G, Hottenga JJ, Langan C, Lopez LM, Hansell NK, Hwang KS, Kim S, Laje G, Lee PH, Liu X, Loth E, Lourdusamy A, Maniega SM, Mattingsdal M, Mohnke S, Nho K, Nugent AC, O'Brien C, Papmeyer M, Pütz B, Ramasamy A, Rasmussen J, Rijpkema M, Risacher SL, Roddey JC, Rose EJ, Ryten M, Shen L, Sprooten E, Strengman E, Teumer A, Trabzuni D, Turner J, **van Eijk K**, van Erp TG, van Tol MJ, Wittfeld K, Wolf C, Woudstra S, Aleman A, Alhusaini S,

Almasy L, Binder EB, Brohawn DG, Cantor RM, Carless MA, Corvin A, Czisch M, Curran JE, Davies G, de Almeida MA, Delanty N, Depondt C, Duggirala R, Dyer TD, Erk S, Fagerness J, Fox PT, Freimer NB, Gill M, Göring HH, Hagler DJ, Hoehn D, Holsboer F, Hoogman M, Hosten N, Jahanshad N, Johnson MP, Kasperaviciute D, Kent JW Jr, Kochunov P, Lancaster JL, Lawrie SM, Liewald DC, Mandl R, Matarin M, Mattheisen M, Meisenzahl E, Melle I, Moses EK, Mühleisen TW, Nauck M, Nöthen MM, Olvera RL, Pandolfo M, Pike GB, Puls R, Reinvang I, Rentería ME, Rietschel M, Roffman JL, Royle NA, Rujescu D, Savitz J, Schnack HG, Schnell K, Seiferth N, Smith C, Steen VM, Valdés Hernández MC, Van den Heuvel M, van der Wee NJ, Van Haren NE, Veltman JA, Völzke H, Walker R, Westlye LT, Whelan CD, Agartz I, Boomsma DI, Cavalleri GL, Dale AM, Djurovic S, Drevets WC, Hagoort P, Hall J, Heinz A, Jack CR Jr, Foroud TM, Le Hellard S, Macciardi F, Montgomery GW, Poline JB, Porteous DJ, Sisodiya SM, Starr JM, Sussmann J, Toga AW, Veltman DJ, Walter H, Weiner MW, Andreassen OA, Apostolova LG, Bastin ME, Blangero J, Brunner HG, Buckner RL, Cichon S, Coppola G, de Zubicaray GI, Deary IJ, Donohoe G, de Geus EJ, Espeseth T, Fernández G, Glahn DC, Grabe HJ, Hardy J, Hulshoff Pol HE, Jenkinson M, Kahn RS, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meyer-Lindenberg A, Morris DW, Müller-Myhsok B, Nichols TE, Ophoff RA, Paus T, Pausova Z, Penninx BW, Potkin SG, Sämann PG, Saykin AJ, Schumann G, Smoller JW, Wardlaw JM, Weale ME, Martin NG, Franke B, Wright MJ, Thompson PM. - Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium - Common variants at 12q14 and 12q24 are associated with hippocampal volume - Nature Genetics: 2012 Apr 15;44(5):545-51. doi: 10.1038/ng.2237.

6. Luykx JJ, Vinkers CH, Bakker SC, Visser WF, van Boxmeer L, Strengman E, **van Eijk KR**, Lens JA, Borgdorff P, Keijzers P, Kappen TH, van Dongen EP, Bruins P, Verhoeven NM, de Koning TJ, Kahn RS, Ophoff RA - A common variant in ERBB4 regulates GABA concentrations in human cerebrospinal fluid – Neuropsychopharmacology: 2012 Aug;37(9):2088-92. doi: 10.1038/npp.2012.57. Epub 2012 May 2.

7. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, **van Eijk K**, van den Berg LH, Ophoff RA - Aging effects on DNA methylation modules in human brain and blood tissue - Genome Biology: 2012 Oct 3;13(10):R97. [Epub ahead of print]

8. **van Eijk KR**, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, de Kovel CG, Janson E, Strengman E, Langfelder P, Kahn RS, van den Berg LH, Horvath S, Ophoff RA - Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy

human subjects - BMC Genomics: 2012 Nov 17;13:636. doi: 10.1186/1471-2164-13-636.

9. Luykx JJ, Bakker SC, Lentjes E, Neeleman M, Strengman E, Mentink L, Deyoung J, de Jong S, Sul JH, Eskin E, **van Eijk K**, van Setten J, Buizer-Voskamp JE, Cantor RM, Lu A, van Amerongen M, van Dongen EP, Keijzers P, Kappen T, Borgdorff P, Bruins P, Derks EM, Kahn RS, Ophoff RA - Genome-wide association study of monoamine metabolite levels in human cerebrospinal fluid - Molecular Psychiatry: 2014 Feb;19(2):228-34. doi: 10.1038/mp.2012.183. Epub 2013 Jan 15.

10. Buizer-Voskamp JE, Blauw HM, Boks MP, **van Eijk KR**, Veldink JH, Hennekam EA, Vorstman JA, Mulder F, Tiemeier H, Uitterlinden AG, Kiemeney LA, van den Berg LH, Kahn RS, Sabatti C, Ophoff RA - Increased paternal age and the influence on burden of genomic copy number variation in the general population - Human Genetics: 2013 Apr;132(4):443-50. doi: 10.1007/s00439-012-1261-4. Epub 2013 Jan 13.

11. Vinkers CH, Van Gastel WA, Schubart CD, **Van Eijk KR**, Luykx JJ, Van Winkel R, Joëls M, Ophoff RA, Boks MP; Genetic Risk and OUtcome of Psychosis (GROUP) Investigators, Bruggeman R, Cahn W, de Haan L, Kahn RS, Meijer CJ, Myin-Germeys I, van Os J, Wiersma D -The effect of childhood maltreatment and cannabis use on adult psychotic symptoms is modified by the COMT Val^{158}Met polymorphism - Schizophrenia Research: 2013 Oct;150(1):303-11. doi: 10.1016/j.schres.2013.07.020. Epub 2013 Aug 15.

12. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, Schumann G, Franke B, Wright MJ, Martin NG, Agartz I, Alda M, Alhusaini S, Almasy L, Almeida J, Alpert K, Andreasen NC, Andreassen OA, Apostolova LG, Appel K, Armstrong NJ, Aribisala B, Bastin ME, Bauer M, Bearden CE, Bergmann O, Binder EB, Blangero J, Bockholt HJ, Bøen E, Bois C, Boomsma DI, Booth T, Bowman IJ, Bralten J, Brouwer RM, Brunner HG, Brohawn DG, Buckner RL, Buitelaar J, Bulayeva K, Bustillo JR, Calhoun VD, Cannon DM, Cantor RM, Carless MA, Caseras X, Cavalleri GL, Chakravarty MM, Chang KD, Ching CR, Christoforou A, Cichon S, Clark VP, Conrod P, Coppola G, Crespo-Facorro B, Curran JE, Czisch M, Deary IJ, de Geus EJ, den Braber A, Delvecchio G, Depondt C, de Haan L, de Zubicaray GI, Dima D, Dimitrova R, Djurovic S, Dong H, Donohoe G, Duggirala R, Dyer TD, Ehrlich S, Ekman CJ, Elvsåshagen T, Emsell L, Erk S, Espeseth T, Fagerness J, Fears S, Fedko I, Fernández G, Fisher SE, Foroud T, Fox PT, Francks C, Frangou S, Frey EM, Frodl T, Frouin V, Garavan H, Giddaluru S, Glahn DC, Godlewska B, Goldstein RZ, Gollub RL, Grabe HJ, Grimm O, Gruber O, Guadalupe T, Gur RE, Gur RC, Göring HH, Hagenaars S, Hajek T, Hall GB, Hall J, Hardy J, Hartman CA,

Hass J, Hatton SN, Haukvik UK, Hegenscheid K, Heinz A, Hickie IB, Ho BC, Hoehn D, Hoekstra PJ, Hollinshead M, Holmes AJ, Homuth G, Hoogman M, Hong LE, Hosten N, Hottenga JJ, Hulshoff Pol HE, Hwang KS, Jack CR Jr, Jenkinson M, Johnston C, Jönsson EG, Kahn RS, Kasperaviciute D, Kelly S, Kim S, Kochunov P, Koenders L, Krämer B, Kwok JB, Lagopoulos J, Laje G, Landen M, Landman BA, Lauriello J, Lawrie SM, Lee PH, Le Hellard S, Lemaître H, Leonardo CD, Li CS, Liberg B, Liewald DC, Liu X, Lopez LM, Loth E, Lourdusamy A, Luciano M, Macciardi F, Machielsen MW, Macqueen GM, Malt UF, Mandl R, Manoach DS, Martinot JL, Matarin M, Mather KA, Mattheisen M, Mattingsdal M, Meyer-Lindenberg A, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meisenzahl E, Melle I, Milaneschi Y, Mohnke S, Montgomery GW, Morris DW, Moses EK, Mueller BA, Muñoz Maniega S, Mühleisen TW, Müller-Myhsok B, Mwangi B, Nauck M, Nho K, Nichols TE, Nilsson LG, Nugent AC, Nyberg L, Olvera RL, Oosterlaan J, Ophoff RA, Pandolfo M, Papalampropoulou-Tsiridou M, Papmeyer M, Paus T, Pausova Z, Pearlson GD, Penninx BW, Peterson CP, Pfennig A, Phillips M, Pike GB, Poline JB, Potkin SG, Pütz B, Ramasamy A, Rasmussen J, Rietschel M, Rijpkema M, Risacher SL, Roffman JL, Roiz-Santiañez R, Romanczuk-Seiferth N, Rose EJ, Royle NA, Rujescu D, Ryten M, Sachdev PS, Salami A, Satterthwaite TD, Savitz J, Saykin AJ, Scanlon C, Schmaal L, Schnack HG, Schork AJ, Schulz SC, Schür R, Seidman L, Shen L, Shoemaker JM, Simmons A, Sisodiya SM, Smith C, Smoller JW, Soares JC, Sponheim SR, Sprooten E, Starr JM, Steen VM, Strakowski S, Strike L, Sussmann J, Sämann PG, Teumer A, Toga AW, Tordesillas-Gutierrez D, Trabzuni D, Trost S, Turner J, Van den Heuvel M, van der Wee NJ, **van Eijk K**, van Erp TG, van Haren NE, van 't Ent D, van Tol MJ, Valdés Hernández MC, Veltman DJ, Versace A, Völzke H, Walker R, Walter H, Wang L, Wardlaw JM, Weale ME, Weiner MW, Wen W, Westlye LT, Whalley HC, Whelan CD, White T, Winkler AM, Wittfeld K, Woldehawariat G, Wolf C, Zilles D, Zwiers MP, Thalamuthu A, Schofield PR, Freimer NB, Lawrence NS, Drevets W; the Alzheimer's Disease Neuroimaging Initiative, EPIGEN Consortium, IMAGEN Consortium, Saguenay Youth Study (SYS) Group – The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data – Brain Imaging and Behavior: 2014 Jan 8. [Epub ahead of print]

In press

Kristel R. van Eijk, Simone de Jong, Eric Strengman, Jacobine E. Buizer-Voskamp, René S. Kahn, Marco P. Boks, Steve Horvath, Roel A. Ophoff - Identification of Schizophrenia-Associated Loci by Combining DNA Methylation and Gene Expression Data from Whole Blood

Nongluk Plongthongkum, **Kristel R. van Eijk**, Simone de Jong, Tina Wang, Jae-Hun Sul, Marco P.M. Boks, René S. Kahn, Ho-Lim Fung, Roel A. Ophoff, Kun Zhang - Characterization of genome-methylome interactions in 22 nuclear pedigrees

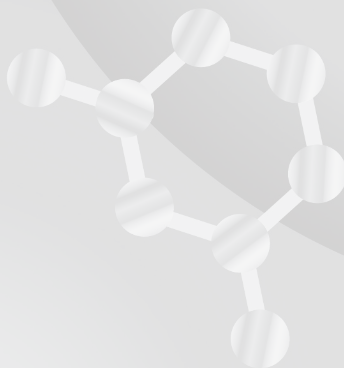
Submitted for publication

Derrek P. Hibar*, Jason L. Stein*, Miguel E. Renteria*, Alejandro Arias-Vasquez*, Sylvane Desrivieres*, Neda Jahanshad, Roberto Toro, Katharina Wittfeld, Micael Andersson, Benjamin S. Aribisala, Nicola J. Armstrong, Manon Bernard, Marc M. Bohlken, Janita Bralten, Andrew A. Brown, M. Mallar Chakravarty, Qiang Chen, Christopher R.K. Ching, Gabriel Cuellar-Partida, Anouk den Braber, Sudheer Giddaluru, Aaron Goldman, Oliver Grimm, Tulio Guadalupe, Johanna Hass, Girma Hawariat, Avram Holmes, Martine Hoogman, Deborah Janowitz, Tianye Jia, Sungeun Kim, Marieke Klein, Bernd Kraemer, Phil H. Lee, Michelle Luciano, Christine Macare, Karen A. Mather, Manuel Mattheisen, Yuri Milaneschi, Kwangsik Nho, Martina Papmeyer, Adaikalavan Ramasamy, Shannon L. Risacher, Roberto Roiz-Santiañez, Emma J. Rose, Alireza Salami, Philipp G. Sämann, Lianne Schmaal, Andrew J. Schork, Jean Shin, Lachlan T. Strike, Alexander Teumer, Marjolein M.J. van Donkelaar, **Kristel R. van Eijk**, Lars T. Westlye, Christopher D. Whelan, Anderson M. Winkler, Marcel P. Zwiers, Saud Alhusaini, Lavinia Athanasiu, Stefan Ehrlich, Marina M.H. Hakobjan, Cecilie B. Hartberg, Unn Haukvik, Angelien J.G.M. Heister, David Höhn, Dalia Kasperaviciute, David CM. Liewald, Lorna M. Lopez, Remco R.R. Makkinje, Mar Matarin, Marlies A.M. Naber, Margaret Needham, Allison C. Nugent, Benno Pütz, Natalie Royle, Li Shen, Emma Sprooten, Daniah Trabzuni, Saskia S.L. van der Marel, Kimm J.E. van Hulzen, Esther Walton, Christiane Wolf, Laura Almasy, David Ames, Sampath Arepalli, Amelia A. Assareh, Mark Bastin, Henry Brodaty, Kazima B. Bulayeva, Melanie A. Carless, Sven Cichon, Aiden Corvin, Joanne E. Curran, Michael Czisch, Greig I. de Zubicaray, Allissa Dillman, Ravi Duggirala, Thomas D. Dyer, Susanne Erk, Iryna O. Fedko, Luigi Ferrucci, Tatiana M. Foroud, Peter T. Fox, Masaki Fukunaga, Raphael Gibbs, Harald H.H. Göring, Robert C. Green, Sebastian Guelfi, Narelle K. Hansell, Catharina A. Hartman, Katrin Hegenscheid, Andreas Heinz, Dena G. Hernandez, Dirk Heslenfeld, Pieter J. Hoekstra, Florian Holsboer, Georg Homuth, Jouke-Jan Hottenga, Masashi Ikeda, Clifford R. Jack Jr., Mark Jenkinson, Robert Johnson ,, Ryota Kanai, Maria Keil, Jack W. Kent Jr, Peter Kochunov, John B. Kwok, Stephen M. Lawrie, Xinmin Liu, Dan

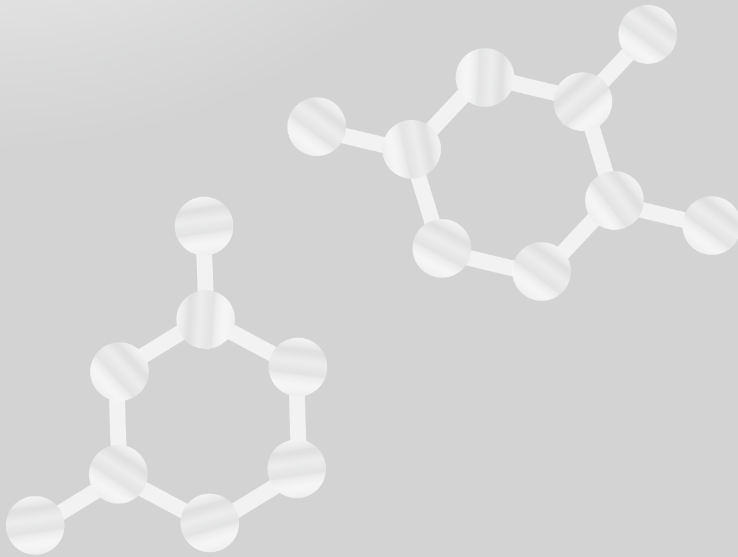
L. Longo, Katie L. McMahon, Eva Meisenzahl, Ingrid Melle, Sebastian Mohnke, Grant W. Montgomery, Jeanette C. Mostert, Thomas W. Mühleisen, Michael A. Nalls, Thomas E. Nichols, Lars G. Nilsson, Markus M. Nöthen, Kazutaka Ohi, Rene L. Olvera, Rocio Perez-Iglesias, G. Bruce. Pike, Steven G. Potkin, Ivar Reinvang, Simone Reppermund, Marcella Rietschel, Nina Romanczuk-Seiferth, Glenn D. Rosen, Dan Rujescu, Knut Schnell, Peter R. Schofield, Colin Smith, Vidar M. Steen, Jessika E. Sussmann, Anbupalam Thalamuthu, Arthur W. Toga, Bryan Traynor, Juan Troncoso, Jessica A. Turner, Dennis van 't Ent, Marcel van der Brug, Nic J.A. van der Wee, Marie-Jose van Tol, Dick J. Veltman, Thomas H. Wassink, Eric Westman, Ronald H. Zielke, Alan Zonderman, David G. Ashbrook, Reinmar Hager, Lu Lu, Francis J. McMahon, Derek W. Morris, Robert W. Williams, Han G. Brunner, Randy L. Buckner, Jan K. Buitelaar, Wiepke Cahn, Vince D. Calhoun, Gianpiero L. Cavalleri, Benedicto Crespo-Facorro, Anders M. Dale, Gareth E. Davies, Norman Delanty, Chantal Depondt, Srdjan Djurovic, Wayne C. Drevets, Thomas Espeseth, Randy L. Gollub, Beng-Choon Ho, Wolfgang Hoffmann, Norbert Hosten, René S. Kahn, Stephanie Le Hellard, Andreas Meyer-Lindenberg, Bertram Müller-Myhsok, Matthias Nauck, Lars Nyberg, Roel A. Ophoff, Massimo Pandolfo, Brenda W.J.H. Penninx, Joshua Roffman, Sanjay M. Sisodiya, Jordan W. Smoller, Hans van Bokhoven, Neeltje E.M. van Haren, Henry Völzke, Henrik Walter, Michael W. Weiner, Wei Wen, Tonya White, Ingrid Agartz, Ole A. Andreassen, John Blangero, Dorret I. Boomsma, Rachel M. Brouwer, Dara M. Cannon, Mark R. Cookson, Eco J.C. de Geus, Ian J. Deary, Gary Donohoe, Guillén Fernández, Simon E. Fisher, Clyde Francks, David C. Glahn, Hans J. Grabe, Oliver Gruber, John Hardy, Ryota Hashimoto, Hilleke E. Hulshoff Pol, Erik Jönsson, Iwona Kloszewska, Simon Lovestone, Venkata S. Mattay, Colm McDonald, Andrew M. McIntosh, Tomas Paus, Zdenka Pausova, Mina Ryten, Perminder S. Sachdev, Andrew J. Saykin, Andy Simmons, Andrew Singleton, Hilkka Soininen, Joanna A. Wardlaw, Michael E. Weale, Daniel R. Weinberger, Hieab H.H.. Adams, Lenore J.. Launer, Reinhold Schmidt, Ganesh Chauhan, Claudia Satizabal, James Becker, Lisa Yanek, Sven J.. van der Lee, Maritza Ebling, Bruce Fischl, W.T. Longstreth, Douglas Greve, Helena Schmidt, Paul Nyquist, Louis N. Vinke, Cornelia M.. van Duijn, Xue Luting, Bernard Mazoyer, Joshua C. Bis, Vilmundur Gudnason, Sudha Seshadri, M. Arfan Ikram, the ADNI, the CHARGE Consortium, EPIGEN, IMAGEN, SYS, Nicholas G. Martin*, Margaret J. Wright*, Gunter Schumann*, Barbara Franke*, Paul M. Thompson*+, Sarah E. Medland*+ - Common genetic variants influence human subcortical brain structures

In preparation

Kristel R van Eijk, Anil PS Ori, Marijn Stokman, Eric Strengman, Netherlands Brain Bank, Simone de Jong, Marco PM Boks, Roel A Ophoff - Interplay between miRNA-137 and DNA methylation in human brain highlights neurotransmitter-related pathways



Curriculum Vitae



Kristel van Eijk was born on June 8th, 1986 in Culemborg where she lived with her parents and sister. She graduated secondary school at O.R.S. Lek en Linge in Culemborg, in 2004. After backpacking in Australia and New Zealand, she started studying bioinformatics at the Hogeschool van Arnhem en Nijmegen. She did internships at the Department of Medical Genetics at the University Medical Center Utrecht (UMCU) under the supervision of Clara Elbers and Lude Franke, and at the European Bioinformatics Institute in Hinxton, Cambridge, UK, supervised by Henning Hermjakob. Kristel graduated with her bachelor's degree in 2009 and began her PhD work in September of that year under the supervision of Prof. dr. ir. Roel Ophoff and Dr. Marco Boks in the Department of Medical Genetics, and at the Brain Center Rudolf Magnus at the department of Psychiatry, at the UMCU, which resulted in this thesis. Her project involved integration of DNA methylation, gene expression, and genotype data in order to gain more insight into the neuropsychiatric disorder schizophrenia. At present, she is working at the Human Neurogenetics Unit at the Brain Center Rudolf Magnus at the UMCU. Kristel lives in Everdingen with her husband Niels and their daughter Ninthe.