University of Basel

Systems genomics analysis of complex cognitive traits

A cumulative dissertation

Submitted to the Faculty of Psychology, University of Basel,
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy

by

Virginie Freytag

Basel, Switzerland

2017

First-Supervisor : Prof. Dr. med. Andreas Papassotiropoulos

Second-Supervisor : Prof. Dr. med. Dominique J-F. de Quervain

University of Basel

Approved by the Faculty of Psychology

at the request of

Prof. Dr. med. Andreas Papassotiropoulos

Prof. Dr. med. Dominique J.-F. de Quervain

Basel, the 29.06.2017

Prof. Dr. phil. Roselind Lieb

Abstract

The study of the genetic underpinnings of human cognitive traits is deemed an important tool to increase our understanding of molecular processes related to physiological and pathological cognitive functioning. The polygenic architecture of such complex traits implies that multiple naturally occurring genetic variations, each of small effect size, are likely to influence jointly the biological processes underlying cognitive ability. Genetic association results are yet devoid of biological context, thus limiting both the identification and functional interpretation of susceptibility variants. This biological gap can be reduced by the integrative analysis of intermediate molecular traits, as mediators of genomic action. In this thesis, I present results from two such systems genomics analyses, as attempts to identify molecular patterns underlying cognitive trait variability. In the first study, we adopted a system-level approach to investigate the relationship between global age-related patterns of epigenetic variation and cortical thickness, a brain morphometric measure that is linked to cognitive functioning. The integration of both genome-wide methylomic and genetic profiles allowed the identification of a peripheral molecular signature that showed association with both cortical thickness and episodic memory performance. In the second study, we explicitly modeled the interdependencies between local genetic markers and peripherally measured epigenetic variations. We thus generated robust estimators of epigenetic regulation and showed that these estimators resulted in the identification of epigenetic underpinnings of schizophrenia, a common genetically complex disorder. These results underscore the potential of systems genomics approaches, capitalizing on the integration of high-dimensional multi-layered molecular data, for the study of brain-related complex traits.

Table of Contents

Abbreviations

C4	Complement component 4
CpG	C-phosphate-G
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
eQTL	Expression quantitative trait locus
EWAS	Epigenome-wide association study
GSEA	Gene-set enrichment analysis
GWAS	Genome-wide association study
LD	Linkage-disequilibrium
MAF	Minor Allele Frequency
mQTL	Methylation quantitative trait locus
SNP	Single nucleotide polymorphism
TROVE2	TROVE Domain Family Member 2

Acknowledgements

First of all, I would like to thank Prof. Andreas Papassotiropoulos and Prof. Dominique de Quervain for having provided me with the opportunity to embark on this scientific journey. Thank you for your guidance, support and enthusiasm that were determinant to accomplish this work.

I am also grateful to Dr. Attila Stetak for our collaboration and his help during this thesis.

It is also my pleasure to thank all my colleagues for their willingness and support through these years. Special thanks to Annette Milnik, Christian Vogler, David Coynel, Leo Gschwind, Vanja Vukojevic and Tobias Egli, for all the stimulating discussions during the course of this thesis, and their help whenever needed. I am also grateful to all members of the Divisions of Molecular and Cognitive Neuroscience for their collaborative effort in gathering all the data.

Finally, I thank my husband and son, for their understanding and permanent support over these years.

Introduction

Neuro-psychiatric disorders are among the major causes of disability worldwide (Whiteford, Ferrari, Degenhardt, Feigin, & Vos, 2015). The biological mechanisms underlying mental diseases, such as e.g. schizophrenia, remain largely unknown, thus limiting our ability to find appropriate treatments. Given the importance of inherited genetic variations in mental disease risk (Gatz et al., 2006; Gejman, Sanders, & Duan, 2010), human-centered genetic research has the potential to expand our understanding of their molecular underpinnings (Papassotiropoulos & de Quervain, 2015).

Common neuro-psychiatric disorders represent genetically complex traits with numerous genetic variations contributing to disease liability (Plomin, Haworth, & Davis, 2009). In this context, the genetic dissection of quantifiable phenotypes, genetically related to diseases, and putatively closer to the biological substrates than abstract diagnosis categories, is a proposed strategy to facilitate the identification of genetic susceptibility variants (Gottesman & Gould, 2003; Papassotiropoulos & de Quervain, 2015). Genetic factors account for a considerable part of physiological variation in cognitive traits (Kremen et al., 2007; Lee et al., 2012). From a genetic standpoint, cognitive deficits, such as memory impairments, that are manifest in many genetically complex neuro-psychiatric disorders, can be considered as the extreme ends of these heritable traits that follow normal distributions (Papassotiropoulos & de Quervain, 2015). Hence, leveraging naturally occurring genetic variations contributing to complex cognitive traits provides the means to gain insights into the molecular pathways implicated in specific physiological and pathological human cognitive processes. In turn, this might lead to the identification of new drug targets and treatment options in psychiatry (Hyman, 2013; Papassotiropoulos et al., 2013).

The genetic study of such heritable traits represents an hypothesis-generating exercise (Stranger, Stahl, & Raj, 2011) aiming at prioritizing new genes or genomic regions for further investigation that eventually may allow deciphering the molecular mechanisms contributing to trait variability. This requires both unbiased analysis of possible genetic contributions in regard of their genomic location and delivery of functionally interpretable solutions. Genome-wide association studies are such a proposed tool, in which millions of common genetic variants can be individually tested for association with a trait (Visscher, Brown, McCarthy, & Yang, 2012). The elucidation of the genetic underpinnings of brain-related phenotypes has already been started using single-marker analyses (Papassotiropoulos et al., 2011; Papassotiropoulos, Stephan, Huentelman, Hoerndli, Craig, et al., 2006). Yet, this approach does not fully account for the highly polygenic pattern and the inherent biological complexity underlying cognitive complex traits (Papassotiropoulos & de Quervain, 2011). The numerous variants of small effects, which together form the genetic substrate of many complex traits are unlikely to pass the significance threshold that results from the necessary multiple testing correction procedures. A pragmatic response to this power issue consists in increasing the sample sizes of genome-wide association studies. This initiated the development of large-scale collaborative efforts aiming at gathering multi-centric GWAS data, allowing meta and mega-analysis of various complex disorders and traits. Increasing the sample sizes successfully led to the identification of additional loci associated with common neuro-psychiatric disorders (Ripke et al., 2013, 2014) and neuro-anatomical traits (Hibar et al., 2015, 2017; Stein et al., 2012).

A majority of complex traits or diseases associated variants identified by GWAS are located in non-coding or intergenic regions of the genome (Hindorff et al., 2009) rendering their direct functional interpretation challenging (Paul, Soranzo, & Beck, 2014). We can also expect that with continuously increasing sample sizes, additional hits will be identified that will require prioritization of the genetic association signals.

In sum, the highly polygenic pattern of complex cognitive traits and the gap between genetic association signals and their biological context limit both the identification and interpretation of trait related genetic variants. The polygeneicity of complex traits in turn suggests a model in which phenotypic variability results from changes in global biological processes, arising from numerous genetic variations and environmental perturbations on the underlying molecular processes (Schadt, 2009; Weiss et al., 2012). Genetic variations are likely to contribute to phenotypic variability in complex traits through their effect on distinct aspects of gene regulation (Albert & Kruglyak, 2015; Li et al., 2016; Richards et al., 2012; Roussos et al., 2014). High-throughput -omics profiling technologies enable population-based assessment of these different layers of molecular information, such as gene expression levels or epigenetic variations. These traits represent intermediate molecular traits putatively mediating the effect of genetic variations on complex phenotypes (van der Sijde, Ng, & Fu, 2014). In turn, systems genomics approaches that rely on the integration of such intermediate traits and genotypic data have the potential to facilitate identification of molecular patterns associated with complex traits (Ritchie, Holzinger, Li, Pendergrass, & Kim, 2015).

This doctoral thesis includes two studies representing examples of such systems genomics approaches, aiming at gaining insights into the molecular processes underlying cognitive complex traits. Specifically, the two studies relied on two distinct integrative analysis of genotypic data and peripherally measured epigenetic markers assessed in healthy young adults: in a first study we adopted a systems-level approach to investigate the relationship between global age-related epigenetic patterns and cortical thickness, further amenable to genetic analysis; in a second study we explicitly integrated genotypic and epigenetic markers to allow the investigation of epigenetic underpinnings of complex cognitive traits.

This thesis includes the following two publications:

- Freytag V., Carillo-Roa T., Milnik A., Sämann PG., Vukojevic V., Coynel D., Demougin P., Egli T., Gschwind L., Jessen F., Loos E., Maier W., Riedel-Heller SG., Scherer M., Vogler C., Wagner M., Binder EB., de Quervain DJ., Papassotiropoulos A. (2017) A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and episodic memory. Nature Communications 26;8:15193. doi: 10.1038/ncomms15193.

- Freytag V, Vukojevic V, Milnik A, Vogler C, de Quervain DJ, Papassotiropoulos A. *submitted*. Genetic estimators of DNA methylation provide insights into the molecular basis of polygenic traits.

Contributions to: design of the experiment, data analysis, paper writing.

1 Complex brain-related phenotypes

The endophenotype concept was introduced in the field of psychiatry as a means to reduce the biological gap between susceptibility variants and genetically complex neuro-psychiatric diseases (Gottesman & Gould, 2003). A putative endophenotype amenable to genetic research should be quantifiable, heritable, genetically related to neuro-psychiatric diseases, and linked to clear neuro-physiological correlates to allow further substantiating a detected genetic association (Gottesman & Gould, 2003; Papassotiropoulos & de Quervain, 2015). The genetic study of such endophenotypes in healthy homogeneous populations circumvents potential confounding of genetic associations by disease-related factors.

1.1 Episodic memory

Episodic memory (EM) which refers to the capability allowing conscious retrieval of past experiences (Tulving, 2002) is a heritable complex trait amenable to genetic research (Papassotiropoulos & de Quervain, 2011). At the neural level, episodic memory depends tightly on the integrity of the medial temporal lobe comprising the hippocampus and adjacent cortices (Squire & Zola-Morgan, 1991; Tulving, 2002). Phenotypic assessment of episodic memory capacity is typically achieved by means of delayed free recall tasks, in which participants are required to retrieve visual stimuli (e.g. words, pictures) within minutes or hours following stimulus presentation. Heritability estimates for such episodic memory phenotypes suggest that naturally occurring genetic variations account for 30 to 60 % of observed phenotypic variance (Kremen et al., 2014; Panizzon et al., 2011; Volk, McDermott, Roediger III, & Todd, 2006). Impaired episodic memory is a hall-mark feature and an early manifestation of Alzheimer's disease. EM deficits have also been reported in schizophrenia patients

(Danion, Huron, Vidailhet, & Berna, 2007), thus supporting the relevance of this cognitive trait relative to neuro-psychiatric diseases. Hence, the genetic dissection of molecular pathways underlying episodic memory in healthy young adults may help to elucidate biological mechanisms implicated in neuro-psychiatric disease etiology (Heck et al., 2015).

1.2 Cortical thickness

The advances in Magnetic Resonance Imaging techniques coupled with automated algorithms enable quantitative assessment of brain sub-cortical and cortical morphometric measures allowing population based investigation of structural data (Lerch et al., 2017).

Inter-individual variability in such measures is linked to differences in cognitive functioning (Kanai & Rees, 2011), possibly through shared genetic factors (Toga & Thompson, 2005; Vuoksimaa et al., 2015; Wallace et al., 2010). Even tough the directionality of these effects remain unclear (Glahn, Thompson, & Blangero, 2007), the genetic dissection of brain neuro-anatomical phenotypes can provide an additional path to expand understanding of the molecular processes underlying cognitive functioning.

Cortical thickness is a brain structural phenotype, reflecting the amount of neurons and neuropil within the horizontal layers along the cerebral cortex (Rakic, 2009). Substantial heritability values have been reported for global cortical thickness with genetic factors estimated to account for ~70 to 80% of phenotypic variability (Panizzon et al., 2009; Winkler et al., 2010). Recent data have also described widespread decrease of cortical thickness with increasing age, observed too during early adulthood (Storsve et al., 2014; Fjell et al., 2015).

2 Genetic association analysis of complex traits

2.1 Genome-wide association studies

The genetic basis of a given trait, that is, the number, penetrance, and frequency of genetic variations affecting the phenotype, is key for the success of the implemented genetic mapping strategy. Rare disorders such as Cystic Fibrosis [*] can be caused by genetic variations within a single gene (Knowles & Drumm, 2012), with sufficiently strong effects to follow a classical Mendelian pattern of dominant or recessive inheritance. Linkage studies which rely on the co-segregation of genetic markers and a trait within families have successfully allowed chromosomal mapping and identification of highly-penetrant variants (Altshuler, Daly, & Lander, 2008).

Yet, the genetic basis of complex quantitative traits is likely to be formed by numerous genetic variations each of low effect relative to genetic variations implicated in Mendelian traits (Fisher, 1918). Given this scenario, association analysis, which allows testing the correlation between genetic markers and a phenotype in large populations of unrelated individuals, represents a powerful alternative to linkage analysis (Visscher et al., 2012).

Genome-wide association studies (GWAS) represent a population-based genetic analyses tool that can capture genetic variations underlying complex traits [**]. The most common genetic variations in the human genome are single-nucleotides polymorphisms (SNPs)(International HapMap Consortium, 2003), i.e. differences in a single base pair between chromosomes at a specific location along the DNA sequence, observed with a frequency of at least 1% in a given population. SNP alleles, which are physically close along the DNA sequence, tend to be co-inherited. This gives rise to a limited number of

[*] Estimated prevalence in European Union: 0.737/10,000 (Farrell, 2008)

[**] See (Visscher et al., 2012) for review of the 'Common-disease/Common variant' theoretical rationale that initiated the GWAS approach.

allele combinations within chromosomal stretches, termed haplotypes. This correlational pattern, referred to as linkage disequilibrium (LD), has practical implications, as in a given population, only a limited number of SNPs - 'tag SNPs' - are needed to identify the haplotypes in given genomic region. A first characterization of these patterns of variations across human populations was achieved by the International HapMap project (International HapMap Consortium, 2003). This effort pushed the development of high-throughput genotyping platforms, allowing the cost-efficient assessment of an individual's genotypes. Simultaneously it gave rise achieving an even higher resolution by employing genotype imputation at untyped marker loci, based on known LD patterns. Today, GWAS typically test genotypes at millions of individual SNPs for association with a dichotomous or continuous phenotype in large samples of unrelated individuals. Multiple-testing correction is necessary for controlling the inflation of false positives induced by the large number of tests conducted. This is typically done by Bonferroni adjustment for the total number of markers examined yielding to stringent significance thresholds.

The unbiased and hypothesis-free GWAS approach allows to pinpoint to circumscribed genetic loci associated with complex trait variability, as a first step for gaining understanding of the molecular underpinnings of those complex traits.

2.2 Complex genetic architecture

Beyond the identification of individual susceptibility loci, GWAS have provided important insights into the genetic architecture of complex polygenic traits.

Firstly, for a given trait, the variants identified by GWAS generally account for a modest fraction of the estimated trait's heritability (Price, Spencer, & Donnelly, 2015). For example, a recent meta-analysis testing the association between an intronic variant located in the *KIBRA* gene and episodic memory (Papassotiropoulos, Stephan, Huentelman, Hoerndli, Craig, et al., 2006), reported an estimated 0.5% of phenotypic variance accounted by the SNP (Milnik et al., 2012). This putative gap can be partly explained by the necessary stringent significance thresholds implied by genome-wide single-marker testing (Manolio et al., 2009). Recent tools have indeed been proposed to estimate the fraction of phenotypic variance jointly accounted for by common SNPs, irrespective of their statistical significance (Yang et al., 2010). In this seminal work, the authors could show that 45% of phenotypic variance in height, a complex trait with estimated heritability of ~ 80% (Visscher et al., 2008), could be retrieved by considering all SNPs simultaneously. Similarly, a large number of common SNPs with individual effects too small to have reached stringent significance thresholds, might collectively account for a considerable heritability fraction of cognitive traits or neuro-psychiatric diseases (Plomin, Haworth, Meaburn, Price, & Davis, 2013; Vogler et al., 2014).

Secondly, a majority of the variants identified by GWAS map to non-coding regions of the genome (Hindorff et al., 2009). Yet, the over-representation of complex trait associated variants within regulatory regions of the genome suggest that genetic variations are likely to exert their effect through gene regulation processes (Albert & Kruglyak, 2015).

3 A systems genomics perspective

3.1 Gene set enrichment analysis

The polygenic pattern of complex trait implies that phenotypic variability arises from the joint effect of multiple markers as perturbations of molecular networks. Under this rationale, gene set enrichment analyses (GSEAs) have been proposed as a powerful tool to capitalize on GWAS data (Wang, Jia, Wolfinger, Chen, & Zhao, 2011). These approaches rely on prior biological knowledge about molecular pathways. Statistical analysis consists in examining whether the aggregate of association signals at SNPs mapping to genes within a pre-specified molecular pathway, significantly deviates from random expectations. Such methods have successfully identified meaningful gene-sets associated with complex cognitive traits and related neuro-psychiatric disorders (Heck et al., 2014, 2015; Petrovska et al., 2017; Ripke et al., 2014). Hence these approaches represent an example of integrating genotypic data and pre-existing biological context information.

3.2 DNA methylation as intermediate molecular trait

Apart from the accumulation of somatic mutations, all cells of an organism carry the same DNA sequence. These cells though have diverse functions. The proper and specific functioning of a given cell requires accurate gene regulation, which is in part orchestrated by epigenetic modifications. By definition, such modifications have the potential to be maintained during somatic cell division (Berger, Kouzarides, Shiekhattar, & Shilatifard, 2009). Beyond the heterogeneity of epigenetic signatures between different cells of an individual, there is considerable inter-individual variation in these epigenetic marks. The differences can be triggered by genetic determinants, or environmental factors that possibly can lead to long lasting imprints on the genome (Fraga et al., 2005; Heijmans et al., 2008; Kaminsky et al., 2009).

Methylation of the DNA sequence is a form of epigenetic modification, which in eukaryotes, occurs only at cytosine residues, primarily in the context of CpG dinucleotides. DNA methylation is implicated in gene expression and imprinting (Deaton & Bird, 2011). More broadly, inter-individual variation in DNA methylation can be viewed as a proxy for differential gene regulation processes (Schübeler, 2015). High-throughput methylomic technology allows quantification of DNA methylation levels at up to hundred thousands of individual CpG sites (Bibikova et al., 2011) in a given tissue.

Likewise in GWAS, DNAm variation at each single CpG site can be tested for association with a given population trait. Epigenome-wide association studies investigating neuro-psychiatric disorders and related traits typically have to rely on available peripheral tissues such as whole-blood or saliva. Yet, given the tissue specificity of DNAm, inter-individual variation in whole-blood does not generally coincide with inter-individual variation in brain tissues (Hannon, Lunnon, Schalkwyk, & Mill, 2015).

Recent data from twins have reported an average heritability estimate of ~20% for whole-blood DNAm variation across all interrogated sites, with common genetic variations accounting on average for ~7% of the observed variance (van Dongen et al., 2016). Methylomic markers thus represent potentially highly-informative intermediate molecular traits, relative to the molecular effects of common genetic variants contributing to complex traits' variability (Kilpinen & Dermitzakis, 2012).

The methylome also undergoes profound changes with increasing age (Teschendorff, West, & Beck, 2013). Epigenome-wide association studies (EWAS) have repeatedly identified numerous individual markers robustly differentially methylated with age (Bell et al., 2012; Garagnani et al., 2012; Hannum et al., 2013; Zaghlool et al., 2015), allowing the derivation of epigenetic predictors for chronological age (Hannum et al., 2013; Horvath, 2013). These predictors have been shown to represent heritable traits per se and to be correlated with all-cause mortality (Marioni et al., 2015).

3.3 System-level analysis

In the scope of this thesis, a system-level approach is used to model global biological processes by considering intra and inter-individual variability within a given multi-dimensional molecular dataset. Such an endeavor broadly relies on analytical methods that allow extracting groups of related variables (e.g. genes, or CpGs) into biologically relevant units. Inter-individual variation across these modeled patterns is in turn seen as reflecting inter-individual variation across the distinct biological processes that underlie the phenotypic variability. The representation of these patterns across individuals can subsequently be tested for association with the trait under study. As intermediate molecular traits, these patterns are also per se amenable to further genetic analysis within the same population.

In Paper 1 (*A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory*) we investigated the relationship between global age-related methylomic patterns and cortical thickness by employing such a system-level modelling approach.

We applied an Independent Component Analysis method which has been shown to possibly identify relevant biological processes from -omics data (Biton et al., 2014; Rotival et al., 2011; Teschendorff, Journée, Absil, Sepulchre, & Caldas, 2007; Wexler et al., 2011). Independent Component Analysis relies on theoretical assumptions regarding the generative model of observed molecular signals: under this model, the observed molecular profiles are viewed as a mixture of independent biological processes (Liebermeister, 2002). In turn, the inferred components are simultaneously characterized by a restricted number of variables, and by their representation across the study samples. Here, we used this approach to extract age-related methylomic patterns putatively reflecting distinct biological processes. These patterns were subsequently amenable to association testing with our population study traits and genetic analysis.

3.4 Integrative mQTL analysis

Given the supposed role of variants associated with complex traits on gene regulation, expression and methylation genetic associations studies can be conducted to identify new functional SNPs related to phenotypic variation in these molecular traits (Nica & Dermitzakis, 2013). Such loci are referred to methylation quantitative trait loci (mQTLs) or expression quantitative trait loci (eQTLs), and further categorized relative to their genomic distance from the associated molecular marker, as *-cis* (typically within 1Mbp) or *-trans* (> 1Mbp). Yet, the identification of *-trans* SNPs is hampered by the multiple testing burden implied by the number of SNP-marker combinations tested, and tend to have lower effect sizes than *-cis* SNPs (Lemire et al., 2015; Mackay, Stone, & Ayroles, 2009).

Provided availability of molecular, genotypic and phenotypic data, a multi-staged strategy can be adopted to examine the relationship between phenotype associated SNPs and the molecular trait at the population level (Ritchie et al., 2015). In this case, trait associated variants identified by GWAS can for instance be examined for their association with molecular traits (eQTL or mQTL); potential molecular traits related to the SNPs can be tested back for association with the phenotype under study enabling functional annotation of the trait-related SNPs. Yet, multi-stage based analyses have to rely on stringent significance thresholds brought about by the single-SNP marker analyses, thus limiting the power for detecting markers of functional relevance (Ritchie et al., 2015).

Intermediate molecular traits can be influenced by multiple *-cis* genetic variants (Bonder et al., 2017). Recently, genetic estimators that capitalize on the joint additive effects of markers on gene expression level have been proposed for further enhancing functional annotation of susceptibility variants (Gamazon et al., 2015). These models rely on a multiple penalized regression framework (Zou & Hastie, 2005), which allows modeling the joint effect of SNPs in *-cis* on the trait and selecting a subset of predictive markers. This approach enables estimation of the genetically driven component of the observed signal, even in moderately-sized samples. In turn, each derived estimator can serve as an intermediate trait amenable to genetic association testing with a complex phenotype in an independent population. This allows investigating the relationship between genetically driven expression or a methylation trait and a population trait, without requiring individuals' molecular trait measurements. In Publication 2 (*Genetic estimators of DNA methylation provide insights into the molecular basis of polygenic traits*) we derived such robust genetic estimators of whole-blood DNAm as a tool for investigating the epigenetic underpinnings of complex cognitive traits. The association between a given trait and each estimator can also be estimated using GWAS single-markers association statistics together with a reference population correlation structure between SNPs markers (e.g. publically available HapMap panel see 2.1)(Barbeira et al., 2016; Gusev et al., 2016). Provided congruence of the GWAS studies population and LD reference panel, this extension allows investigation of the wealth of currently available GWAS summary results, even in absence of genotypic data.

Original research papers

Publication 1 A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory

nature COMMUNICATIONS

ARTICLE

Received 10 May 2016 | Accepted 8 Mar 2017 | Published 26 Apr 2017

DOI: 10.1038/ncomms15193

OPEN

A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory

Virginie Freytag[1,2], Tania Carrillo-Roa[3], Annette Milnik[1,2,4], Philipp G. Sämann[3], Vanja Vukojevic[1,2,5], David Coynel[2,6], Philippe Demougin[1,2,5], Tobias Egli[1,2], Leo Gschwind[2,6], Frank Jessen[7,8], Eva Loos[2,6], Wolfgang Maier[7,9], Steffi G. Riedel-Heller[10], Martin Scherer[11], Christian Vogler[1,2,4], Michael Wagner[7,9], Elisabeth B. Binder[3,12], Dominique J.-F. de Quervain[2,4,6,*] & Andreas Papassotiropoulos[1,2,4,5,*]

Increasing age is tightly linked to decreased thickness of the human neocortex. The biological mechanisms that mediate this effect are hitherto unknown. The DNA methylome, as part of the epigenome, contributes significantly to age-related phenotypic changes. Here, we identify an epigenetic signature that is associated with cortical thickness ($P = 3.86 \times 10^{-8}$) and memory performance in 533 healthy young adults. The epigenetic effect on cortical thickness was replicated in a sample comprising 596 participants with major depressive disorder and healthy controls. The epigenetic signature mediates partially the effect of age on cortical thickness ($P < 0.001$). A multilocus genetic score reflecting genetic variability of this signature is associated with memory performance ($P = 0.0003$) in 3,346 young and elderly healthy adults. The genomic location of the contributing methylation sites points to the involvement of specific immune system genes. The decomposition of blood methylome-wide patterns bears considerable potential for the study of brain-related traits.

[1] Division of Molecular Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland. [2] Transfaculty Research Platform Molecular and Cognitive Neurosciences, University of Basel, CH-4055 Basel, Switzerland. [3] Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, D-80804 Munich, Germany. [4] Psychiatric University Clinics, University of Basel, CH-4055 Basel, Switzerland. [5] Department Biozentrum, Life Sciences Training Facility, University of Basel, CH-4056 Basel, Switzerland. [6] Division of Cognitive Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland. [7] German Center for Neurodegenerative Diseases (DZNE), D-53175 Bonn, Germany. [8] Department of Psychiatry, University of Cologne, Medical Faculty, D-50924 Cologne, Germany. [9] Department of Psychiatry, University of Bonn, D-53105 Bonn, Germany. [10] Institute of Social Medicine, Occupational Health and Public Health, University of Leipzig, D-04103 Leipzig, Germany. [11] Center for Psychosocial Medicine, Department of Primary Medical Care, University Medical Center Hamburg-Eppendorf, D-20246 Hamburg, Germany. [12] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, Georgia 30322, USA. * These authors jointly supervised this work. Correspondence and requests for materials should be addressed to V.F. (email: virginie.freytag@unibas.ch) or to A.P. (email: andreas.papas@unibas.ch).

Human cortical thickness, a brain morphometric measure that is linked to cognitive functioning, reflects the amount of neurons and neuropil in the horizontal layers of the cortical columns that are responsible for the organization of cortical connectivity[1–3]. Recent data suggest a monotonic decrease in cortical thickness (cortical thinning) from preschool age throughout the lifespan[4], but previous studies have also described patterns of regional increase in cortical thickness during childhood[5–7].

Studies in twins and in unrelated individuals provide consistently high heritability estimates for cortical thickness (~80%), demonstrating the importance of naturally occurring genetic variation for this physiological trait[8,9]. Despite the well-known and substantial impact of age on cortical thinning, the biological mechanisms that mediate this effect are hitherto unknown. It is reasonable to assume that age-related, dynamic processes, such as epigenetic changes, represent good candidates for such mediators.

DNA methylation, the most extensively studied epigenetic modification to date, regulates important processes such as imprinting, chromosomal inactivation and gene expression[10]. Age represents one of the most potent factors known to correlate with physiological variation of global DNA methylation[11,12]. High-throughput quantification of DNA methylation at several hundreds of thousands of C-phosphate-G (CpG) sites has detected numerous CpG loci across various tissues undergoing differential methylation with age[13–16]. Interestingly, such loci have been identified within regulatory regions of genes that are known to undergo differential expression in such age-related conditions as Alzheimer's disease[13] and cancer[17]. Recently, DNA methylation markers predicting chronological age were shown to correlate with all-cause mortality[18]. DNA methylation levels can also be influenced by genetic variations[19,20] and age-related DNA methylation signatures represent heritable traits[18].

Thus, the existing data suggest that peripherally measured DNA methylation patterns might contribute to the identification of molecular underpinnings of age-related complex traits relevant to health and disease.

Here, we investigated the relation between peripherally measured DNA methylation and cortical thickness in healthy young adults. In a first step, we performed Independent Component Analysis (ICA)-based decomposition of whole-blood methylomic profiles to identify independent signatures of physiological variation of global DNA methylation. ICA is a decomposition method, which provides a representation of complex relationships arising from high-dimensional data, such as genome-wide expression[21,22] and brain imaging data[23]. After ICA-based decomposition, the identified methylation patterns were first tested for association with age. Age-associated methylation patterns were subsequently tested for correlation with global cortical thickness and, in case of such correlation, mediation analysis followed to assess whether these methylation patterns mediated significantly the effect of age on cortical thickness. Significant findings were subjected to further analyses, including functional annotation of CpGs contributing to the observed methylation patterns, testing for pattern association with region-specific cortical thickness and cognitive performance, and a genome-wide investigation of common genetic variations (single nucleotide polymorphisms, SNPs) that contribute to the variability of the methylomic patterns.

Results

ICA-based identification of methylomic patterns. We performed methylomic profiling (Illumina 450K Human Methylation array) of blood samples collected from $N=533$ healthy young individuals (Supplementary Table 1). After quality control, DNA methylation levels (DNAm) were quantified at 397,947 autosomal CpG sites and subsequently corrected for sex and sources of variation inferred from Surrogate Variable Analysis (see Methods).

Next, we performed ICA to achieve a low-dimensional representation of genome-wide methylation profiles. Following the ICA paradigm introduced first for gene expression data analysis[21], an individual's methylomic profile is treated as a mixture of latent variables (that is, methylomic signatures), each reflecting a combination of biological processes and exerting independent effects on DNAm. Specifically, ICA provides a representation of these signatures by decomposing the original DNAm signals into components, whose statistical inter-dependence is minimized. This property is typically achieved by favouring heavy-tailed non-gaussian distribution of the components' loadings; thus each component is characterized by a restricted set of CpGs exhibiting loadings at the extreme of the distribution. Simultaneously, each component is characterized by its representation across the study sample, giving rise to individual methylation patterns. Each of these patterns is a low-dimensional representation of a global mode of DNAm variations. Importantly, these patterns can be tested for association with traits of the study sample (Fig. 1a).

Using ICA decomposition, we obtained a total of $k=126$ independent components (see Methods). The majority of the inferred components ($n=111$) were driven by single individuals contributing to more than 10% of the pattern's variability. Given that such components represent rather singular modes of variation[24], subsequent analyses were restricted to the remaining 15 components. These components represent global modes of DNAm variation across the individuals of the study population.

Methylomic patterns related to age and cortical thickness. Participants from the methylomic profiling study underwent brain magnetic resonance imaging (MRI)(Supplementary Table 1). Global measures of cortical thickness—that is, the distance between the grey matter and white-matter boundary and the pial surface—were obtained using cortical surface-based analysis implemented in FreeSurfer (see Methods), for $N=514$ participants. Consistent with previous findings in healthy young adults[4,25], cortical thickness was negatively correlated with age ($r=-0.27$, $P=3.12\times10^{-10}$).

Two out of 15 ICA methylomic patterns (termed *ICA1* and *ICA2*) were significantly correlated with age, after Bonferroni correction for 15 comparisons (*ICA1*: $r=0.54$, $P_{\text{nominal}}=1.54\times10^{-42}$, $P_{\text{corrected}}=2.31\times10^{-41}$; *ICA2*: $r=0.29$, $P_{\text{nominal}}=4.68\times10^{-12}$, $P_{\text{corrected}}=7.02\times10^{-11}$; Fig. 1c and Supplementary Table 2). These methylomic patterns were also significantly associated with cortical thickness (*ICA2*: $r=-0.24$, $P_{\text{nominal}}=3.86\times10^{-8}$, $P_{\text{corrected}}=5.79\times10^{-7}$; *ICA1*: $r=-0.14$, $P_{\text{nominal}}=0.00162$, $P_{\text{corrected}}=0.0243$; Fig. 1b and Supplementary Table 2). No significant correlation was observed between *ICA1* and *ICA2* ($r=0.048$; nominal $P=0.27$), suggesting that the corresponding independent components capture distinct methylomic processes.

To test whether the significant correlations between *ICA2* and cortical thickness were merely attributable to the correlation between age and both types of measurements, age effects were partialled out from *ICA2* and cortical thickness (see Methods). After this adjustment, a significant correlation was exclusively detected between *ICA2* and cortical thickness ($r=-0.18$; $P=6.55\times10^{-5}$, Supplementary Table 2). The correlation remained significant ($r=-0.17$; $P=8.74\times10^{-5}$) also after correcting for individual white blood cell count (see Methods). We also examined which other available variables (that is, body mass index, smoking, alcohol consumption, frequency of cannabis use) were significantly associated with *ICA2* in addition to age. Smoking frequency was also significantly associated with *ICA2* ($r = 0.17$, $P = 0.0001$) but not with cortical thickness ($r = -0.072$, $P = 0.11$). After adjusting *ICA2* for both age and smoking frequency, its association with cortical thickness remained nearly unchanged ($r = -0.17$). No significant correlations were detected between *ICA2* and alcohol consumption ($P = 0.97$), cannabis use ($P = 0.1$) or body mass index ($P = 0.25$).

In order to capture possible non-linear age effects, we also performed an *F*-test analysis to compare the fit of a model predicting cortical thickness from a fifth degree polynomial of age (age + age^2 + … + age^5) to the fit of the same model augmented by *ICA2*. We observed a highly significant increase in adjusted R^2 with the addition of *ICA2* to the model ($F(1,507) = 15.6$, $P = 8.8 \times 10^{-5}$). Thus, the association between *ICA2* and cortical thickness is not driven by non-linear age effects.

We also used *in silico* annotation of blood cell types as described by Jaffe and Irizarry[26]. After this adjustment, *ICA2* associations with both chronological age and cortical thickness remained highly significant ($P = 2 \times 10^{-11}$ and $P = 8.3 \times 10^{-7}$,

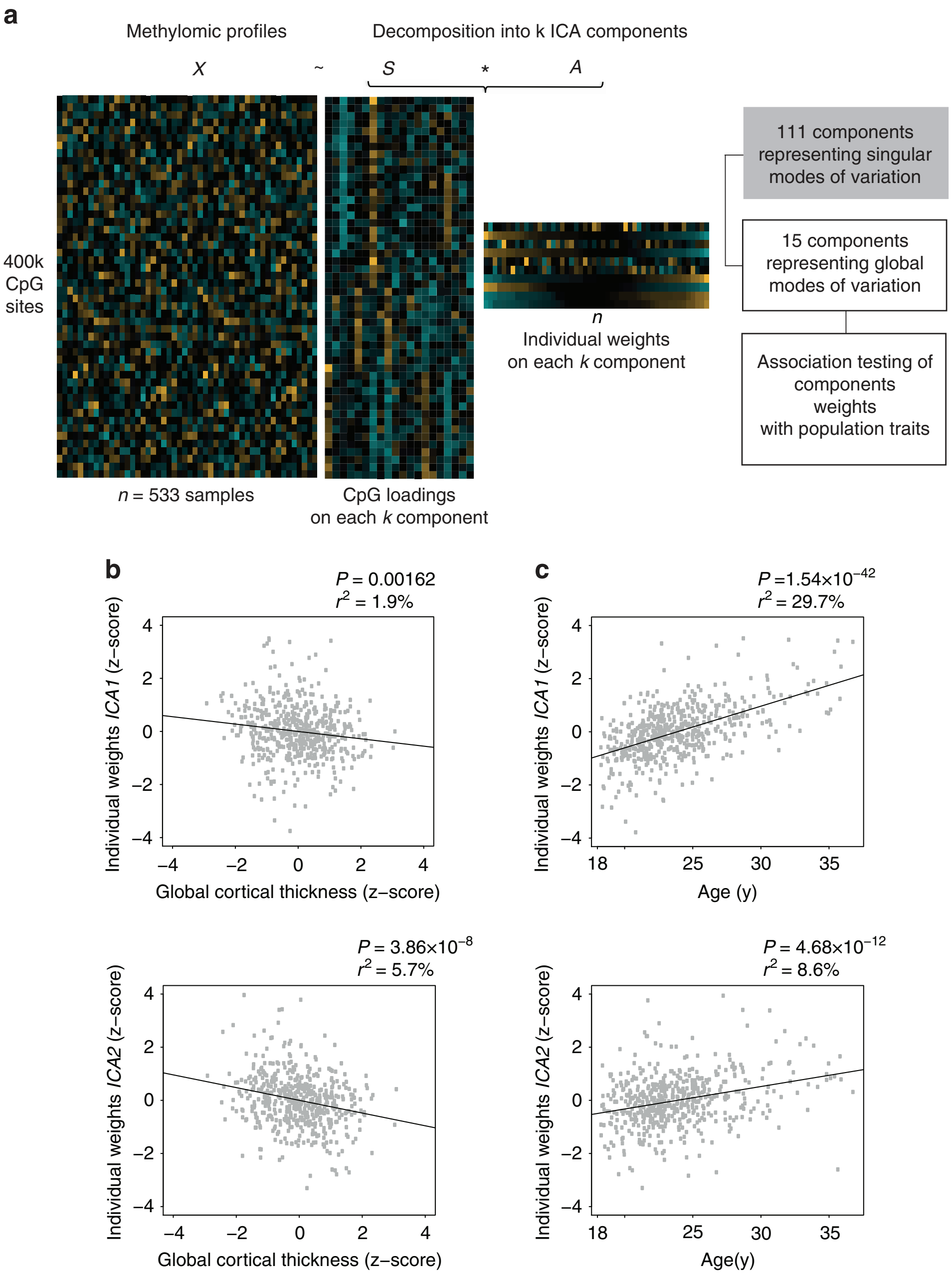


Figure 1 | ICA-based identification of DNAm patterns. (**a**) Schematic representation of the analysis workflow; ICA decomposition of genome-wide methylomic profiles (matrix *X*, $n = 533$ samples × 397,947 CpGs sites) into *k* independent components, simultaneously represented across CpGs (matrix *S* of CpGs loadings) and samples (matrix *A* of individual weights). A total of 15 components, whose corresponding weights represent global modes of DNAm across samples, were tested for association with cortical thickness and chronological age. (**b**) Two components, *ICA1* and *ICA2*, are significantly associated with cortical thickness. Horizontal axis: cortical thickness adjusted for sex, intra-cranial volume and MR-batches. Vertical axis: individual weights on *ICA* component. (**c**) *ICA1* and *ICA2* show significant association with chronological age. *P*: *P* value of association (Pearson's correlation, two-sided test); r^2: fraction of variance in component weights explained by chronological age (in %).

respectively). We also examined the association between *ICA1* and *ICA2* and chronological age in two publicly available data sets of purified blood cells ($N = 1{,}202$ monocyte samples, age range: 44–83, mean age: 60; $N = 214$ CD4+ T-cell samples, age range: 45–79, mean age: 59)[15]. In each data set, *ICA1* and *ICA2* were estimated as the linear combinations between *ICA1* and *ICA2* loadings, respectively (as inferred from the Swiss DNAm sample), and blood samples' DNAm values, adjusted for main confounders (see Supplementary Methods). In both cell-specific data sets, we observed a significant positive correlation between *ICA* patterns and chronological age (monocyte samples, $N = 1{,}202$: *ICA1*: $r = 0.67$, $P < 2.2 \times 10^{-16}$; *ICA2* $r = 0.32$, $P < 2.2 \times 10^{-16}$; CD4+ T-cell samples, $N = 214$: *ICA1*: $r = 0.70$, $P < 2.2 \times 10^{-16}$; *ICA2*: $r = 0.49$; $P = 8.6 \times 10^{-15}$), suggesting that the *ICA*–age correlations identified in whole-blood are also detectable in individual cell types. Altogether these results substantiate the lack of influence of blood cell counts on the reported associations. The correlation between cortical thickness and *ICA1*, that showed the strongest correlation with age, was not significant after adjusting for chronological age ($r = 0.01$, $P = 0.83$, Supplementary Table 2).

In addition to chronological age, we also calculated epigenetic cross-tissue- and whole-blood-based predictors in our sample as described by Horvath[27] and Hannum *et al.*[14], respectively. Both estimators yielded DNA methylation age values (that is, predictors for chronological age based on CpG methylation) that significantly correlated with actual participants' age (Horvath's predictor: $r = 0.70$, $P < 10^{-60}$; Hannum's predictor: $r = 0.71$, $P < 10^{-60}$). Neither predictor was associated with cortical thickness after adjustment for chronological age (Horvath's: $r = 0.04$, $P = 0.32$; Hannum's: $r = 0.01$, $P = 0.77$), suggesting that these predictors (like *ICA1* but, importantly, unlike *ICA2*) do not mediate the effect of age on cortical thickness.

Finally, we examined the association of *ICA2* with age and age-adjusted cortical thickness after covarying for 111 individuals who contributed more than 10% to 111 inferred components not further studied herein. Both associations remained highly significant (age: $P = 4.91 \times 10^{-12}$; age-adjusted cortical thickness: $P = 4.8 \times 10^{-5}$).

Replication study. To test the generalizability of the association between *ICA2* and cortical thickness, we studied an independent sample (termed herein the Munich sample) comprising 596 participants with major depressive disorder (MDD) and healthy controls (see Methods). The *ICA2* pattern was estimated as the linear combination between *ICA2* loadings (as inferred from the Swiss DNAm sample) and individual DNAm values of the Munich sample. In this independent sample, we observed a significant positive correlation between *ICA2* and chronological age ($N = 596$, $r = 0.48$, $P < 10^{-10}$) and a negative correlation with global cortical thickness ($N = 596$, $r = -0.31$, $P < 10^{-10}$). After adjustment for chronological age and controlling for potential confounders (diagnosis, sex, intracranial volume, MRI batch effects, time difference between MRI examination and blood drawing), the association between *ICA2* and cortical thickness remained significant ($r = -0.094$, $P = 0.011$). The same analysis in a sub-sample of $N = 163$ participants younger than 40 years (that is, within an age range similar to that of the Swiss participants) revealed an almost identical effect size ($r = -0.19$, $P = 0.009$) compared to that observed in the Swiss sample.

***ICA2* partially mediates the age–cortical thickness relation**. *ICA2* showed significant positive correlation with age and negative correlation with global cortical thickness. To investigate whether *ICA2* mediates the negative correlation between age and global cortical thickness, we conducted a mediation analysis[28]. The association between chronological age and global cortical thickness was partially (that is, $k^2 = 5.1\%$ of the maximum possible mediation effect) and significantly mediated by the methylomic pattern *ICA2* (indirect effect $= -0.051$, $P < 0.001$) (Fig. 2).

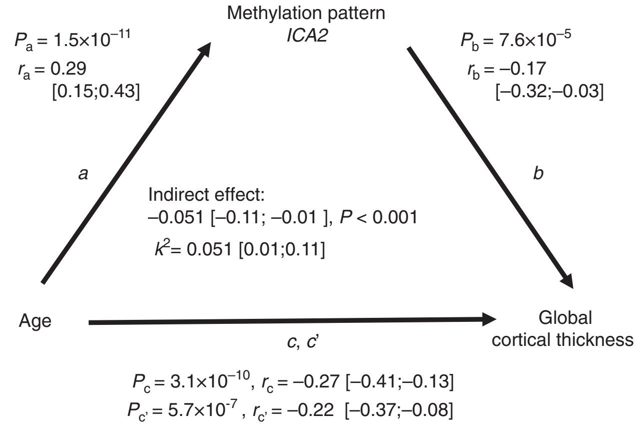

Methylation pattern
ICA2
$P_a = 1.5\times10^{-11}$
$r_a = 0.29$
[0.15;0.43]
$P_b = 7.6\times10^{-5}$
$r_b = -0.17$
[−0.32;−0.03]
a
b
Indirect effect:
−0.051 [−0.11; −0.01], $P < 0.001$
$k^2 = 0.051$ [0.01;0.11]
Age
Global
cortical thickness
c, c'
$P_c = 3.1\times10^{-10}$, $r_c = -0.27$ [−0.41;−0.13]
$P_{c'} = 5.7\times10^{-7}$, $r_{c'} = -0.22$ [−0.37;−0.08]

Figure 2 | Mediation analysis of methylomic pattern *ICA2* on the association between chronological age and global cortical thickness. Path *a* represents the effect of chronological age on *ICA2*. Path *b* represents the effect of *ICA2* on global cortical thickness after removing the effect of chronological age. Path *c* denotes the total effect of chronological age on global cortical thickness. Path *c'* represents the direct effect of chronological age on cortical thickness while controlling for the indirect effect (*a* multiplied by *b*). *r*: correlation coefficient; 99.9% confidence interval for the parameters are shown in brackets; *P*: *P* value of association. k^2: kappa-squared standardized maximum possible mediation effect.

***ICA2* is related to a specific pattern of cortical thickness**. Having detected an association between *ICA2* and global cortical thickness we next explored possible links between this methylomic pattern and regional variations in cortical thickness. Inter-individual variations in delineated brain regions often coincide with latent structural covariance patterns[29]. Exploratory factor analysis (EFA) allows depicting such distinct patterns of volumetric covariance among brain regions that can be subsequently tested for association with additional phenotypes of the population under study[30]. We therefore performed EFA, considering 68 regional brain measures of thickness (34 per hemisphere) obtained from automated parcellation of the cerebral cortex (Desikan-Killiany atlas)[31–33]. Before analysis, effects of intra-cranial volume, sex, processing batches and age, which possibly drive global correlations among brain regions, were regressed out from individual measures (see Methods). Using parallel analysis[34], we determined eight extractable factors, altogether accounting for 48.9% of variance across regional measures (Supplementary Data 1, see Methods). Factor extraction was followed by varimax orthogonal rotation. Subjects' factor scores were subsequently tested for association with the age-adjusted *ICA2* pattern. After correction for multiple testing, we identified one factor score, *F6*, that showed significant correlation with *ICA2* ($r = -0.13$, $P = 0.00314$, Bonferroni-adjusted $P = 0.025$ for eight tests conducted)(Fig. 3a and Supplementary Table 3). This factor, accounting for 4% of variance in cortical thickness measures, was characterized by a spatial pattern comprising mainly temporal areas (loadings > 0.3), with the highest loadings observed for left and right temporal poles and

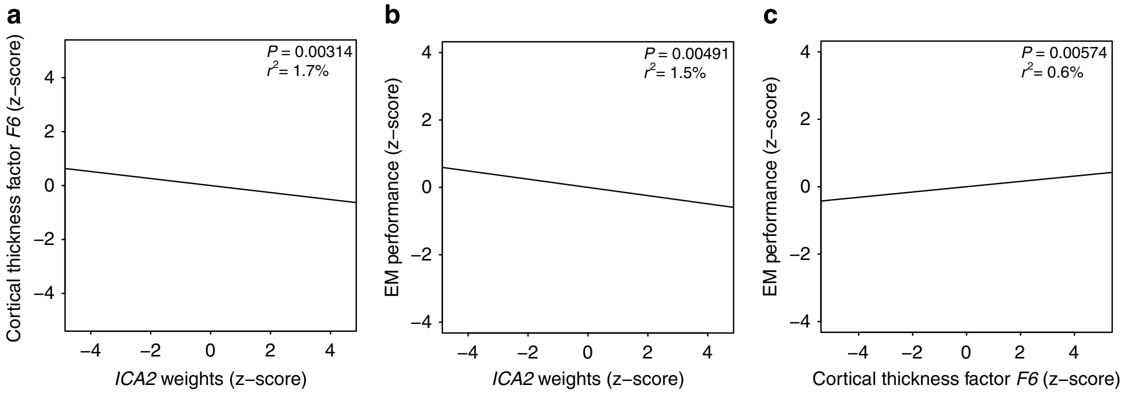


Figure 3 | Correlations between *ICA2* weights, EM performance and cortical thickness score *F6*. (**a**) Correlation between cortical thickness factor score *F6* and *ICA2* weights in the methylomic profiling sample. (**b**) Correlation between *ICA2* weights and EM performance in the methylomic profiling sample. (**c**) Correlation between cortical thickness factor score *F6* and EM performance in the combined sample ($N = 1,234$). Subjects from the methylomic profiling sample are shown in blue. *ICA2* and the EM/imaging phenotypes are adjusted for chronological age effects.

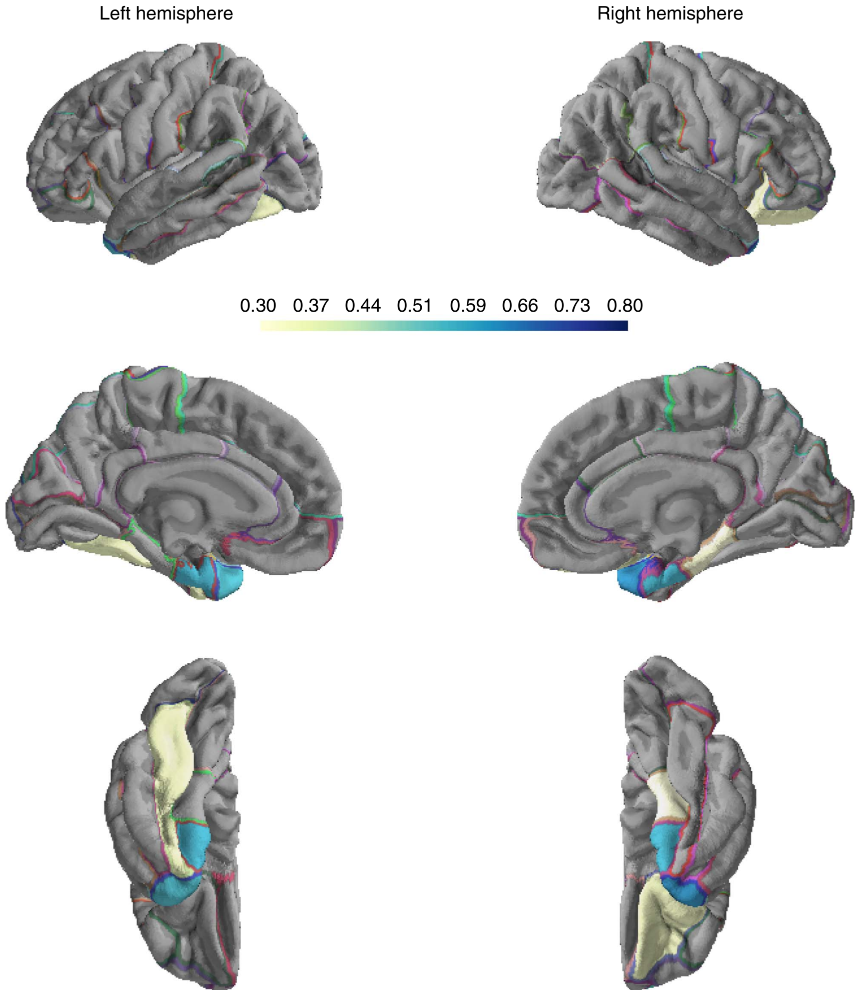


Figure 4 | Regional cortical thickness loadings on factor *F6* associated with *ICA2* methylomic profile. Absolute values for loadings are considered. Loadings < |0.3| are not shown.

entorhinal cortices (mean loadings across the four temporal regions: 0.58) (Fig. 4 and Supplementary Data 1). We also run EFA under most conservative adjustment of the 68 regional brain measures of thickness for mean global thickness, to study whether the significant regional effects observed herein are fully explained by mean global cortical thickness. We observed high mean correlation between factor loadings across the two EFA solutions ($r = 0.78$); importantly, *F6* remained stable across the two solutions with an $r = 0.89$ ($P = 6.9 \times 10^{-24}$) between loadings before/after adjustment for mean thickness. In addition, *F6* scores obtained from the mean thickness-adjusted EFA solution were still significantly associated with *ICA2* ($r = -0.09$, $P = 0.042$), suggesting that the results presented herein were not driven solely by global mean thickness.

Thus, higher values of *ICA2* are related to thinning of a circumscribed cortical pattern that harbours neuroanatomical correlates of episodic memory (EM). Therefore, we investigated the relationship between this methylomic pattern and EM. Behavioural assessment was obtained for a total of $N = 531$ subjects from the methylomic profiling study (see Methods, Supplementary Table 1). We detected a significant negative correlation between *ICA2* and EM performance ($r = -0.138$, $P = 0.00147$) (Supplementary Table 4). This association remained significant after partialling out age effects ($r = -0.122$, $P = 0.00491$) (Supplementary Table 4 and Fig. 3b) and after adjustment for white blood cell types abundance ($r = -0.105$, $P = 0.016$ after adjusting for age).

Complete MRI and EM assessments were obtained for $N = 512$ participants from the methylomic profiling sample (Supplementary Table 1). In this sample, the correlation between *F6* regional thickness score and EM performance was not significant ($r = 0.048$, $P = 0.283$)(Supplementary Table 5). In an additional independent sample of $N = 722$ healthy young adults (Basel imaging sample, Supplementary Table 1), who underwent identical MRI (in the same scanner) and cognitive assessment as the methylomic profiling sample (see Methods), the correlation between *F6* regional thickness score and EM performance was significant ($r = 0.102$, $P = 0.00617$)(Supplementary Table 5). Importantly, subjects' *F6* factor scores from this additional MRI sample were predicted based on the factor solution inferred from the methylomic profiling sample. In the combined sample ($N = 1,234$ participants), the correlation between *F6* and EM performance was significant ($r = 0.079$, $P = 0.00574$) (Fig. 3c and Supplementary Table 5).

Functional and genomic characterization of *ICA2* CpGs. Contributions to a given ICA component are commonly identified by selecting features (here: CpG sites) whose loadings, in absolute value, exceed a cut-off threshold of n_σ standard deviations from the mean of the loadings' distribution[22]. In gene expression studies, such a typical threshold ranges between 2 and 3 (ref. 35). Given the pronounced multidimensionality of the methylomic profiles, we used a stringent cut-off of $n_\sigma = 4$, which led to the selection of 970 CpGs for *ICA2* (Supplementary Data 2). The selected *ICA2* CpGs mapped to 593 genes (see Methods). Among the 970 CpGs constituting *ICA2*, one marker (cg18055007) was part of the 353 Horvath age-predicting markers[27], and four (cg20822990, cg16054275, cg16867657, cg21139312) were part of the 71 CpGs included in Hannum's DNAm age model[14]. We also examined whether, and to what extent, *ICA2* CpGs ($N = 970$) overlapped with those reported as being differentially methylated ($N = 2,037$) in smokers[36–39]. This was the case for a small fraction (3%) of the *ICA2* CpGs.

Enrichment analysis for gene ontology (GO) terms, molecular pathways and gene expression patterns, as catalogued in the Molecular Signatures Database (MsiGDB, www.broadinstitute.org/gsea/msigdb/index.jsp), was performed using the GOseq algorithm[40], which corrects for multiple CpG mapping per gene (see Methods). Using an FDR threshold < 0.05 (Benjamini–Hochberg adjustment), analysis of *ICA2* revealed significant enrichment for 76 highly overlapping gene sets, which mainly encompassed genes related to immune system function, inflammatory response and hematopoietic system (Supplementary Data 3). To explore further the nature of the immune component related to cortical thickness, we compared the DNA methylation of the 970 most prominent *ICA2* CpGs to that of blood cell subtypes and their progenitors using public data sets on 19 cell types (see Methods)[41]. We observed consistently highly significant correlations ($N = 970$ CpG sites, $P < 10^{-60}$ for all correlations) between average whole-blood DNAm values of the *ICA2* CpGs and all various cell subtypes examined (Supplementary Figs 3–5). The lowest correlation coefficients (albeit still highly significant with $P < 10^{-60}$) were observed for regulator and memory CD4+ T-cells (Supplementary Fig. 5). Generally, the correlation coefficients might suggest high concordance of the cortical thickness-related blood DNAm patterns with DNAm of B lymphocytes and of the common myeloid progenitor lineage, and relatively less concordance with DNAm of natural killer cells and T lymphocytes.

We next characterized the *ICA2*-contributing CpGs with respect to their topographical distribution across the genome (that is, island, shore, shelf, open sea regions) and to their relative location to gene transcripts. Given that a CpG site may map to multiple transcripts, each site was uniquely characterized according to following rules[42]: CpGs annotated within 1,500 bp upstream the transcription start site of at least one transcript were flagged as 'TSS'; CpGs not flagged as 'TSS' but located within a transcript (including 3′UTR, 5′UTR) were flagged as 'Genic'; all remaining CpGs were flagged as 'Intergenic'. We observed a significant shift in the distribution of CpG topographical categories as compared to the genome-wide background expectations (χ^2 test $P = 2.7 \times 10^{-53}$) with 50% of all CpGs annotated as Open Sea, while Islands CpGs were clearly under-represented (10%)(Supplementary Fig. 1A). The distribution of CpG sites across genomic context categories differed from the genome-wide background distribution (χ^2 test $P = 1.65 \times 10^{-6}$), with an increased fraction of 'Genic' CpGs and a decreased fraction of 'TSS' CpGs. We also observed a lower fraction of intergenic CpGs as compared to the background distribution (Supplementary Fig. 1B).

Finally, to test between-sample comparability of the identified ICA patterns, we performed ICA of the study population reported in Hannum *et al.*[14], which consists of 656 blood DNAm profiles of participants spanning a wide age-range (19–101 years, mean age: 64 years). We identified five ICA patterns that were significantly associated with age ($N = 656$, $P = 0.000043$ – $P < 10^{-60}$). We then examined the overlap between *ICA1* and *ICA2* CpGs identified in our sample and CpGs contributing to each of the Hannum age-associated IC pattern. A significant overlap with *ICA1* was observed for one pattern (termed here HICa, OR = 91, $P < 10^{-60}$). For *ICA2*, we observed a significant overlap with three age-associated Hannum patterns ($P = 1.9 \times 10^{-6}$ – $P < 10^{-60}$), with said overlap being particularly strong for one pattern (termed here HICb, OR = 49, $P < 10^{-60}$). The correlation of loadings between CpGs contributing to *ICA2* and CpGs contributing to the HICb pattern was positive and of substantial magnitude ($r = 0.87$, $P < 10^{-60}$). Thus, we observed highly significant between-sample overlap of ICA patterns despite the differences in age structure of the two populations.

***ICA2*-derived multigenic score associated with EM performance**. Given that DNA methylation patterns per se represent complex traits[18,43], we studied the genetic underpinnings of the *ICA2* pattern. As for any genetically complex trait, several genetic variants are likely to contribute jointly to inter-individual variability of DNAm variation as represented by *ICA2*. Therefore, we employed gene set enrichment analysis (GSEA)[44–46] to disentangle biologically meaningful subsets of genetic contributions to *ICA2*.

Table 1 | GSEA results for *ICA2* pattern.

Database	Gene set	No. of genes*	Nominal GSEA *P*†	FDR
Gene ontology	Leukocyte differentiation	38	7.1×10^{-5}	0.0124
Biocarta	PYK2 pathway	27	6×10^{-4}	0.0183
Gene ontology	Lymphocyte differentiation	26	8.6×10^{-5}	0.0234
Biocarta	Keratinocyte pathway	44	2×10^{-4}	0.024
Gene ontology	Haemopoiesis	71	3.22×10^{-4}	0.0407
Gene ontology	Haemopoietic or lymphoid organ development	73	2×10^{-4}	0.044

*Number of genes in gene set mapped by at least one SNP.
†Empirical enrichment *P* value at a 75th percentile cut-off.

DNA from all individuals participating in the methylomic profiling study was processed on the Affymetrix Genome-wide Human SNP Array 6.0. After standard QC, correction for minor allele frequency and deviation from Hardy–Weinberg equilibrium, a total of 733,370 autosomal SNPs were used for association analysis (see Methods).

Age-adjusted single-marker *P* values for association with *ICA2,* under an additive model, were tested for gene set enrichment using MAGENTA[44] (see Methods). Across the 1,411 tested sets we detected a significant over-representation of association signals (FDR$<$0.05) in six gene sets mainly related to immune system regulation (Table 1). Given the substantial overlap between the identified gene sets, we further combined these sets into two gene groups with minimum overlap: genes from categories GO: Lymphocyte differentiation, GO: Leukocyte differentiation and GO: Haemopoiesis were grouped into GO:0048534 (Haemopoietic or lymphoid organ development) which comprised 73 unique genes; the two remaining gene sets, that contained 12 overlapping genes, Biocarta: Pyk2 pathway and Biocarta: Keratinocyte pathway were grouped into 'Pyk2/Keratinocyte pathway', which comprised 60 unique genes. These two distinct gene groups had one gene in common.

For each of the gene groups we calculated multilocus genetic scores to capture their contributions to individual *ICA2* variability. The scores comprised 39 and 33 significant SNPs mapping to an equal number of genes from the 'GO:0048534' and 'Pyk2/Keratinocyte' groups respectively (Supplementary Data 4 and 5). Genetic scores were weighted by the direction of effect of single-marker association statistics, resulting in positive correlation of each score with the *ICA2* pattern (see Methods). As expected, both scores correlated significantly with *ICA2* variability ('GO:0048534': $r=0.53$, $P=3.26\times10^{-40}$; 'Pyk2/Keratinocyte': $r=0.45$, $P=5.9\times10^{-28}$).

The genetic score derived from the Haemopoietic and Lymphoid Organ development set (GO:0048534) was significantly correlated with EM performance ($r=-0.10$, $P=0.01$). No significant correlation was detected for the 'Pyk2/Keratinocyte'-derived genetic score ($r=0.02$, $P=0.7$).

To test the robustness of this association, we studied the correlation between the GO:0048534-derived genetic score and EM in four additional independent samples ($N=3{,}346$): three samples, including subjects from the Basel imaging sample, comprised a total of $N=2{,}603$ healthy young subjects who performed either a picture free recall or a word free recall task; an additional sample included $N=743$ elderly healthy individuals who performed a word free recall task (age range: 74–91 years, see Methods and Table 2). The genetic score correlated negatively with EM performance, resulting in a significant combined association $P=0.0003$ (Stouffer Meta-analysis, Table 2).

Table 2 | Association of Haemopoetic or Lymphoid Organ development genetic score and EM-related traits in independent samples.

Sample	*N*	Age range	EM task	r	*P*
Basel cognitive	1,445	18-35	Pictures	−0.062	0.00912
Basel imaging	534	18-35	Pictures	−0.02	0.32
Zurich	624	18-45	Words	−0.073	0.0349
AgeCode	743	74-91	Words	−0.076	0.0191
Stouffer's method meta-analysis				−0.06*	0.0003

r: Pearson's correlation coefficient. *P*: one-sided correlation test *P* value.
*sample size weighted *r*.

To test whether GSEA-derived genetic score SNPs are enriched for mQTLs of the *ICA2* CpGs, we first examined the location of these SNPs relative to the 970 CpGs constituting *ICA2*. We observed a significant over-representation (53%, $P<0.0002$) of gene score SNPs *in cis* (that is, ±1 Mbp) to *ICA2* CpGs as compared to a genome-wide random distribution (see Methods). Next, we performed mQTL analysis for each of the score SNP–*ICA2* CpG pairs. We observed significant deviation from the null uniform distribution with particular over-representation of genetic associations with effect sizes ranging from small to moderate (Supplementary Fig. 2). Thus, GSEA-derived SNPs collectively exert multiple genetic effects of small to moderate magnitude on the CpGs contributing to *ICA2*.

We also studied the association between the genetic score and cortical thickness in the methylomic sample ($N=514$). No significant correlation was observed with cortical thickness ($r=-0.06$, $P=0.08$).

Discussion

In the present study we applied ICA decomposition of whole-blood genome-wide methylomic profiles in healthy young adults (22.9 ± 3.3 years, mean ± s.d.) and detected a specific pattern of DNAm (*ICA2*) that was associated with cortical thinning and decreased EM performance. We also observed that a significant part of the well-known negative correlation between age and cortical thickness was partially mediated by *ICA2*. CpG sites that contributed to this methylation pattern mapped to genes involved in immune system regulation and inflammatory response.

Notwithstanding the robust and replicated findings presented herein, we would like to stress some limitations, which are inherent to the study design. First, the mediation analysis suggests that *ICA2* significantly, albeit partially, mediates the effect of age on cortical thickness. Given the associative nature of the data, we cannot exclude the possibility that the correlation observed between *ICA2* and cortical thickness might also be partially driven by additional non-modelled variables. Second, decomposition of genome-wide methylomic profiles comes at the cost of specificity of the inferred solution towards the genomic localization of CpG markers. The detection of CpGs contributing to the methylomic signature relies on a fixed threshold on the distribution of the components' loadings. In our case, this approach allowed relating *ICA2* broadly to genes involved in immune system function. However, the specific relationships between the identified marker sets and the phenotypes of interest can be studied only in downstream experiments focusing on single CpG sites. Third, the ICA model relies on the assumption that methylomic signals arise from a fixed set of independent sources. In the absence of *a priori* knowledge about the source signal, the number of inferred components must be determined empirically, which might impact negatively on generalizability. Integration of multiple-layers of molecular traits, such as genotypic data used in this study, is therefore important to address whether the identified patterns represent relevant features of the data set.

The cellular mechanisms underlying changes in cortical thickness are not entirely clear; however, they are most likely life phase-dependent. During development, cortical thinning might be related to events mirroring cortical maturation, such as synaptic pruning[47], whereas shrinkage of neurons, reductions of synaptic spines and lower numbers of synapses probably account for adult age-related cortical thinning[48]. In addition, myelination of lower cortical layers might cause the cortical mantle to appear thinner on MR scans[49]. This phenomenon might account for a substantial part of the observed cortical thinning during development. The correlation between cortical thickness and cognitive function also seems to be age-dependent. In adulthood and old age, cortical thinning is associated with a decline in cognitive function[3], whereas during development this relationship is dynamic with predominantly negative correlation between cognitive function and cortical thickness in early childhood to a positive correlation in late childhood and beyond[6]. Importantly, a substantial proportion of the strength of the relation between cortical thinning and cognitive decline in adults is attributable to the influence of age in each type of measure[3].

The association of *ICA2* with cortical thickness and EM performance reported herein supports observations relating the peripheral immune system to brain morphology and cognition[50,51] and is coherent with the notion that the brain and its functions is directly linked to peripheral tissues relevant to the function of the immune system[52]. The data presented herein might suggest high concordance of the cortical thickness-related blood DNAm patterns with DNAm of B lymphocytes and of the common myeloid progenitor lineage, and relatively less concordance with DNAm of natural killer cells and T lymphocytes. Nevertheless, it is important to stress that we cannot draw any mechanistic conclusions about the relationship between peripheral methylation on the one side and cortical thickness and EM performance on the other, and that no further inference can be drawn towards the contribution of a specific immune cell type to the reported associations. Indeed, the mechanisms through which the peripheral immune system exerts an influence on the central nervous system remain elusive. Direct cytokine-induced central responses or indirect cytokine-mediated changes within the central nervous system via activation of vagal-nerve afferents are being discussed among possible scenarios[53,54]. Of note, methylation sites related to *IL6R* (encoding interleukin 6 receptor), *ZC3H12D* (encoding zinc finger CCCH-type containing 12D) and *CD4* (encoding CD4 molecule) are listed among the top ten *ICA2* contributing CpGs (Supplementary Data 2) in our data. The products of these genes are centrally implicated in cytokine signalling, mRNA stability of cytokine genes and immunological response. It will be interesting to investigate whether direct measurement of the immune factors implicated herein along with traditional blood markers of the immune system will provide additional information with regard to the relation between these immune factors and cortical thickness. We speculate that this might not be the case, given the substantial volatility of such direct measurements, which mostly reflect acute state of the immune system, whereas methylation profiles reflect, at least partially, a record of past immune regulation. Nevertheless, further experimental work is warranted to test this hypothesis.

Inter-individual variability in blood cell composition is known to influence whole-blood DNAm measurements[55]. In our population of healthy young adults, no significant association between blood cell sub-types and cortical thickness or EM performance was observed (Supplementary Table 6). Moreover, the associations between *ICA2*, cortical thickness and EM were significant also after correction for blood cell composition. In addition, we observed a significant positive correlation between *ICA* patterns (*ICA1* and *ICA2*) and chronological age in the examined blood cell-specific data sets. Thus, it is unlikely that the detected associations are driven by inter-individual variability in composition of blood cell types.

In addition to studying methylation patterns, we also performed a genome-wide SNP-based analysis of *ICA2*. The reasons for this analysis were two-fold: (1) Given the fact that genetic variation is related to DNA methylation, we tested whether *ICA2*-related genetic variation can be used as a proxy for DNA methylation in larger samples, where such epigenetic measures were unavailable. (2) We hypothesized that the biological processes revealed through gene set enrichment would be similar regardless of the nature of the data input (that is, genetic versus epigenetic variation). Interestingly, the SNP-based analysis of *ICA2* revealed a robust association between variants of genes involved in the regulation of the immune system and EM in independent cohorts of young and elderly healthy adults. This suggests that the association between *ICA2*, which reflects epigenetic variation, and EM performance is, at least partially, genetically driven.

In conclusion, we adopted an ICA approach to achieve a tractable and biologically meaningful representation of genome-wide methylation profiles that are amenable to association testing. To this end we searched for methylomic profiles that arise from putatively independent biological processes, each reflected by a restricted number of CpG sites. By decomposing genome-wide DNAm profiles we identified an epigenetic mark of immune system genes linked to cortical thickness and to human memory. The well-known effect of age on cortical thinning is partially mediated by this epigenetic mark, and its genetic underpinnings also point to genes involved in immune system regulation. Thus, the decomposition of blood methylome-wide patterns bears considerable potential for the study of brain-related physiological traits. For example, peripheral markers of systemic inflammation are associated with reduced grey matter volume, both in midlife adults[50] and in the elderly[56]. Moreover, such grey matter reduction seems to mediate the negative effects of peripheral inflammation on age-related cognitive decline[50]. It will be interesting to investigate whether the peripheral DNAm profiles identified herein might be used to differentiate between physiological and pathological age-related cognitive decline and cortical thinning.

Methods

Samples. *Methylomic profiling sample.* This sample is part of an ongoing, continuously recruiting imaging genetics study of healthy young adults in the city of Basel, Switzerland. Aim of the study is to recruit large samples of healthy young adults for assessing cognitive performance measurements, personality traits, functional and anatomical MRI and genetics (based on saliva DNA) at the time-point of the main investigation. Advertising for the main investigation was done mainly in the University of Basel. Subjects were re-invited via email or at the time-point of the main investigation to an additional blood and saliva sampling. The time point of this second investigation was on average 348 days (min 1 day; max 1,384 days; median 314 days) after the main investigation. For the purpose of this study, a total of $N=568$ subjects underwent blood methylomic profiling (Data lock Apr. 2014). After pre-processing of methylomic data and genetic outliers exclusion, a total of $N=533$ subjects were included in the methylomic profiling sample (Supplementary Table 1).

Basel imaging sample. This sample is part of the same ongoing, continuously recruiting imaging genetics study as the methylomic profiling sample. A total of $N=753$ participants who were not part of the $N=533$ methylomic profiling sample underwent imaging and EM assessment. A total of $N=722$ subjects with complete imaging and EM assessment were included in the Basel imaging sample (Supplementary Table 1), among which $N=623$ subjects underwent genotyping.

Basel cognitive sample. This sample is part of an ongoing, continuously recruiting genetics study in the city of Basel, Switzerland, independent from the methylomic profiling and Basel imaging samples. A total of $N=1,622$ healthy young subjects underwent EM performance assessment and genotyping (mean age: 22.4; 66% female).

Zurich sample. This sample included a total of $N=706$ healthy young subjects recruited in Zurich, who underwent EM assessment and genotyping (mean age: 21.8; 70% female).

All participants were free of any neurological or psychiatric illness, and did not take any medication at the time of the experiment (except hormonal contraceptives). The ethics committee of the Cantons of Zurich, Basel-Stadt and Basel-Landschaft approved the experiments. All participants received general information about the study and gave their written informed consent for participation.

AgeCode sample. This sample consisted of elderly participants of the German Study on Ageing, Cognition and Dementia in primary care patients (AgeCoDe). The AgeCoDe study is an ongoing primary care-based prospective longitudinal study on early detection of mild cognitive impairment and dementia established by the German Competence Network Dementia. The sampling frame and sample selection process of the AgeCoDe study have been described in detail previously[57] (see Supplementary Methods for complete description). Sufficient DNA-samples for genome-wide genotyping were available for 782 subjects. The complete description of EM phenotypes can be found in Supplementary Methods. The AgeCoDe-study was approved by the local ethic committees of all participating centres (Ethics Committee of the Medical Association Hamburg; Ethics Committee of the University of Bonn; Medical Ethics Committee II, University of Heidelberg at the University Medical Center of Mannheim; Ethics Committee at the Medical Center of the University of Leipzig; Ethics Committee of the Medical Faculty of the Heinrich-Heine-University Düsseldorf; Ethics Committee of the TUM School of Medicine, Munich). All participants received general information about the study and gave their written informed consent for participation.

Munich sample. The Munich sample consisted of patients with first episode and recurrent unipolar depression treated as in-patients at the Max Planck Institute of Psychiatry, Munich, and healthy control subjects ($N=627$ with combined MRI and DNA availability; 423 patients, age 47.9 (s.d. 13.8) years; control subjects age 49.5 (s.d. 13.3) years), for the most part overlapping with imaging genetic and MDD association studies reported in collaboration with the ENIGMA consortium[58,59]. Other than in the flagship study[58], no bipolar patients were included for reasons of clinical homogeneity[59]. MDD diagnoses were based on clinical consensus in addition to M-CIDI or SCAN interviews, depending on the original study protocols. After pre-processing of methylomic data, and MRI-QC-based exclusions, combined data of $N=596$ subjects was available for statistical analysis. Description of methylomic profiling and structural imaging of the Munich sample are provided in Supplementary Methods. All participants gave their written informed consent after receiving general information about the study. Study protocols and the transition of anonymous data into the biobank of the Max Planck Institute of Psychiatry were approved by the ethics committee of the Ludwig Maximilian University in Munich, Germany.

Methylomic profiling. Blood samples were collected from all the subjects using BD Vaccutainer Push Button blood collection set and 10.0 ml BD Vacutainer Plus plastic whole blood tube, BD Hemogard closure with spray-coated K2EDTA (Becton, Dickinson and Company, New Jersey, USA). DNA was isolated from the remaining fraction, upon plasma removal. The isolation was performed with QIAmp Blood Maxi Kit (Qiagen AG, Hilden, Germany), using the recommended spin protocol. Subject's DNA was extracted between midday and evening (mean time = 14:30, range 13:00–20.00). Microarray-based DNA methylomic profiling from whole-blood samples was performed at ServiceXS (ServiceXS B.V., Leiden, the Netherlands). In brief, the bisulfite conversion was performed with 500 ng genomic DNA input using the EZ DNA Methylation Gold Kit (Zymo Research, Irvine, CA, USA). A bisulfite conversion quality control on the samples was performed with DNA qPCR reaction and subsequent melting curve analysis[60]. The bisulfite-converted DNA was processed and hybridized to the HumanMethylation450 BeadChip (Illumina, Inc.), according to the manufacturer's instructions. Methylation data were pre-processed using the R package RnBeads[61]. Beta values were calculated from SWAN normalized intensities[62]. Beta-values with detection P value ≥ 0.05 were considered as missing. Individual probes were excluded based on the following criteria: (1) non-CpG context probes, polymorphic probes, probes harbouring three or more SNPs in their 50mer extension (MAF ≥ 0.01), and cross-hybridizing probes, based on the annotation provided with the RnBeads package[61], (2) cross-hybridizing probes and polymorphic CpGs sites referenced in refs 63, 64, (3) detected by iterative Greedycut algorithm, (4) missing rate $\geq 5\%$ in final samples. After quality control a total of 397,947 autosomal probes remained for analysis. Samples showing divergent genetic background from the majority of Caucasian samples were excluded; these genetic outliers were identified using Bayesian Clustering Algorithm[65] on genotypic projections onto the two first principal components inferred from reference Hapmap populations (CEU, JPT, CHB). Exclusion of samples yielded a total of 533 samples entering methylomic analyses.

To rule out systematic shift in DNA methylation values induced by SWAN normalization, we compared the correlation between summary statistics of CpG sites before and after normalization. We observed high correlation for both average ($r>0.99$) and variance ($r>0.95$) of DNA methylation values across samples. We also observed high average correlation between DNAm values before and after normalization per-CpG site (average $r=0.87$), and per-sample (average $r=0.89$ after mean-centring DNAm values per CpG).

DNA methylation profiles were obtained on average 1 year after imaging acquisition. We performed a sensitivity analysis examining the association between *ICA2* and cortical thickness after regressing out the difference (Δage) between age at blood sampling and age at MRI assessment from the methylomic pattern. The association remained significant ($P=6.4\times10^{-5}$, $r=-0.18$ after adjustment for age) indicating that Δage did not affect the results of the study.

Primary phenotypes (age, cortical thickness, EM performance) were not confounded with methylomic processing covariates (plate, sentrix ID, position) (linear model minimum observed $P>0.04$).

ICA2 showed weak nominal association with Sentrix ID (Supplementary Data 6). After adjustment of *ICA2* for this technical covariate, the association with age, cortical thickness and EM performance remained highly significant (age: $P=1.6\times10^{-11}$; age-adjusted thickness: $P=1.3\times10^{-4}$, EM: $P=0.0096$).

Blood cell counting. Haematological analysis, including blood cell counts, was performed at the collection time point with Sysmex pocH-100i Automated Hematology Analyzer (Sysmex Co, Kobe, Japan).

Lymphocytes, neutrophils and overall count of basophils, monocytes and eosinophils (mixture) were available for $N=527$ participants from the methylomic profiling sample.

Structural imaging. Participants from the methylomic and Basel imaging samples underwent identical MRI assessment.

Measurements were performed on a Siemens Magnetom Verio 3T wholebody MR unit equipped with a 12-channel head coil. A high-resolution T1-weighted anatomical image was acquired using a magnetization prepared gradient echo sequence (MPRAGE) sequence with the following parameter: TE (echo time) = 3.37 ms, FOV (field of view) = 25.6 cm, acquisition matrix = 256 × 256 × 176, voxel size = 1 mm × 1 mm × 1 mm. Using a midsaggital scout image, 176 contiguous axial slices were placed along the anterior – posterior commissure (AC – PC) plane covering the entire brain with a TR = 2,000 ms (flip angle = 8°).

From the initial $N=533$ participants from the methylomic profiling sample and $N=753$ from the Basel imaging sample, a total of 50 participants were excluded due to excessive movement or scanner noise by visual inspection of T1-weighted images, or technical reasons. This yielded a total of $N=514$ from the methylomic profiling sample entering structural imaging analysis, and $N=722$ subjects from the Basel imaging sample.

T1-weighted images were processed using the publicly available FreeSurfer software (v4.5) (refs 31–33). This processing includes motion correction, removal of nonbrain tissue, automated Talairach transformation, intensity correction, volumetric segmentation, and cortical surface reconstruction and parcellation. Specifically, the three-dimensional cortical surface was reconstructed to measure volume, surface area and thickness at each surface location or vertex. After the initial surface model was constructed, a refinement procedure was applied to obtain a triangulated representation of the grey/white (GM/WM) boundary. The GM/WM boundary was then deformed outwards to obtain an explicit representation of the pial surface. Thickness measurements were obtained by calculating the distance between the GM/WM boundary and pial surfaces at each vertex across the cortical mantle[31]. Global individual measures for thickness were computed by averaging cortical vertices measurements for both hemispheres. Individual measures were adjusted for sex, intra-cranial volume and MR-technical batches (software and gradient batches) using linear regression.

ICA based identification of methylomic patterns. After probes and samples quality control, missing Beta values were imputed using the R package impute. In order to adjust the methylation signals for technical confounders and preserve effects of chronological age on methylation sites, we applied the iteratively re-weighted surrogate variable analysis algorithm implemented in the SVAR package[66], considering age at blood sampling as the outcome. Beta values were adjusted for sex and 40 inferred surrogate variables using linear regression. For each CpG, the residuals from this linear model were standardized across samples.

ICA decomposition of the standardized residuals was performed using the R package fastICA. The number of components to extract was estimated using the Random Matrix Theory algorithm[67] implemented in the R package isva[68]. Given the stochastic initialization of fastICA algorithm, we performed 30 repeats of the ICA components' estimation. All realizations of the mixing matrix (A) were clustered using hierarchical clustering, with complete linkage agglomeration, based on Pearson's correlation similarity. Final components were determined as the centrotypes of the inferred clusters.

When using such decomposition methods as ICA, multiple-correction depends on the number of identified components, which in not known *a priori*. In our case, the genome-wide methylomic data set was decomposed into 15 components that were amenable to downstream association testing. Hence, traits correlated with these 15 components were subjected to following α level adjustment: $P = 0.05/15 = 0.0033$. After having identified *ICA2* as the only pattern associated with cortical thickness, we further investigated its relationship with eight regional thickness factor scores. The α level was thus adjusted for eight tests conducted ($P = 0.05/8 = 0.00625$).

Association testing of methylomic patterns. *Swiss sample.* The association between *ICA* patterns and imaging or behavioural phenotypes was assessed using Pearson's correlation, with two-sided association test. Given the delta between age at methylomic profiling and age at main investigation, chronological age adjustment was achieved by partialling out age effects: effect of age at methylomic profiling was regressed out from methylomic patterns and age at main investigation was regressed out from the phenotypic measure, using linear regression. Adjustment of methylomic patterns for blood cell counts was performed by regressing out effects of chronological age at blood sampling and effects of each of the three white-blood cell parameters using linear regression. The obtained residuals were subsequently tested for association with the relevant imaging or EM phenotypes.

Self-reported smoking frequency was measured on a 4-point Likert scale (0 = never, 1 = occasionally, 2 = 1–5 cigarettes per day, 3 = 6–20 cigarettes per day, 4 = 20 or more cigarettes per day). Self-reported alcohol consumption and cannabis use frequencies were measured on a 3-point Likert scale (0 = never, 1 = occasionally, 2 = daily). Association testing for each indicator was performed using linear regression.

Munich sample. ICA2 patterns were calculated separately for the whole Munich sample and a subsample of <40-year-old subjects ($N = 163$). DNA methylation values were first adjusted for sex using linear regression. *ICA2* patterns were then calculated as linear combination between the scaled residuals and the inverse ICA2 loadings inferred from the Swiss sample (Supplementary Data 8). Separate Pearson's correlation analyses were performed between ICA2-scores and biographical age, and ICA2-scores and cortical thickness. In addition, partial correlation analyses were performed between ICA2-scores and cortical thickness, correcting for age at MRI, difference between age at MRI and age at blood-drawing, sex, intracranial volume and MRI batch effects. All P values reported in the replication sample are one-sided.

Exploratory factor analysis of cortical thickness measures. Average regional cortical thickness in 68 areas (34 per hemisphere) were obtained from FreeSurfer automated parcellation method based on Desikan-Atlas[31–33]. Individual measures in each sample (methylomic profiling and Basel imaging samples) were adjusted for sex, intra-cranial volume, MR-technical batches and chronological age using linear regression. EFA was performed on regional cortical thickness measures from the methylomic profiling sample ($N = 514$). Factor extraction was based on principal axis factoring method. The number of factors to extract was determined using the parallel method implemented in R package psych. The factor analysis solution was rotated using the varimax method. A variable was considered to load on a factor if its absolute loading on the factor was 0.3 or greater. Based on the factor solution inferred from the methylomic profiling sample, we extracted factor scores predictions for both the methylomic profiling sample and the independent Basel imaging sample, using regression method. For each sample, factor scores were tested for association with EM performance using Pearson's correlation, with a two-sided association test. The same association analysis was conducted combining factor scores and EM performance of the two samples (combined sample, $N = 1{,}234$).

Mediation analysis. Chronological age at main investigation, *ICA2* methylomic pattern and global cortical thickness were entered in a mediation analysis[28] using the R package MBESS. To represent the strength of the mediation we computed the indirect effect (a multiplied by b, see Fig. 1) and the k^2 square value which is interpreted as the proportion of the maximum possible indirect effect that could have occurred[69]. The 99.9% confidence intervals for these parameters were obtained on bias-corrected and accelerated bootstrapping procedure with 10,000 resamplings. Significance of the indirect effect was assessed by testing whether the confidence interval of the indirect effect excludes 0, considering interval limits from 90 to 99.9%.

Gene-set enrichment analysis of methylomic component. CpGs were mapped to transcripts based on Illumina's annotation; EntrezID gene identifiers were downloaded from the UCSC genome database. Enrichment testing was performed using the GOseq package[40] which applies stringent correction towards genes mapped by multiple CpGs across the array. Enrichment statistics were obtained using Wallenius approximation. A total of 19,518 genes mapped by the 397,947 CpGs entering the analysis were used as background. Gene sets were downloaded from the MSig DB (www.broadinstitute.org/gsea/msigdb/, curated gene lists C2 and C5).

Genetic association analyses. *Genotyping.* DNA was extracted from saliva or blood using standard protocols. All subjects were individually genotyped using the Affymetrix Human SNP Assay 6.0 according to the manufacturer's recommendation. In the methylomic profiling sample, subjects with unusual ancestry according to the majority of the sample were excluded using Bayesian clustering algorithm and Hapmap reference populations (see Methylomic profiling). Subjects were also checked for inconsistency between reported and genetically inferred sex. Individual call rate averaged to 98.3%. For the purpose of scoring analyses, subjects from the Basel imaging, Basel cognitive, Zurich and AgeCode samples were additionally excluded based on the following criteria: genome-wide call rate $<95\%$; IBD sharing defined by PI_HAT > 0.2 (one-sample of each detected pair was excluded); Bayesian Clustering[65] outlier detection on genome-wide call rate and heterozygosity rate. This yielded a total of $N = 1{,}445$ individuals entering the genetic scoring analysis for Basel cognitive sample, $N = 534$ for Basel imaging sample, $N = 624$ for Zurich sample and $N = 743$ for AgeCode sample.

Validation of the link between methylation and genotype data. A per-subject crosscheck between phenotypic data, methylation data and genetic data was performed using the reported sex and sex-predictions based on the array data, as well as matching of all SNPs represented on the Illumina 450 K array to the corresponding Affymetrix SNP 6.0 genotype calls. This crosscheck allowed an unambiguous assignment of each methylation data set to the corresponding genetic and phenotypic data set.

Gene set enrichment analysis of ICA2 pattern. GSEA was performed using the MAGENTA[44] software which derives gene-centric association statistics from single-SNP association P values, while controlling for potential confounders (gene size, number of SNPs, number of independent SNPs, number of recombination hotspots, linkage disequilibrium and genetic distance). Genome-wide single-SNP association analysis was conducted on *ICA2* pattern adjusted for chronological age (at blood sampling), using an additive model. A total of 773,330 autosomal SNPs that passed individual SNP quality control in the methylomic profiling sample (exclusion criteria MAF < 0.01; HWE P value ≤ 0.0001; call-rate < 0.90) entered the analysis.

In order to capture signals from potentially regulatory variants, MAGENTA-derived gene scores were based on SNPs lying within 20 kb upstream and downstream of the extreme transcript boundaries. The GSEA algorithm includes a built-in procedure controlling for physical proximity of SNPs within a given gene set (automatic exclusion of the gene exhibiting a lower association signal in case of one SNP mapped to multiple genes within a gene set). Gene set enrichment statistic was based on the 75th percentile cut-off of the observed genome-wide gene-score distribution, which has been proposed to show optimal power for weak genetic effects as expected for complex polygenetic traits. Empirical P values were adjusted for multiple testing using FDR. Gene sets were extracted from the MSigDB v3.1 database (http://www.broadinstitute.org/gsea/msigdb), including gene sets from different online databases (KEGG, Gene Ontology GO, BioCarta and Reactome). We used a gene set size ranging between 20 and 200 genes to avoid both overly narrow and broad gene set categories, resulting in 1,411 gene sets to be analysed. Genes from the extended major histocompatibility complex region were excluded from the analysis.

Genetic scoring association analyses. The scores comprised SNPs associated with *ICA2* pattern ($P < 0.05$) mapping to an equal number of genes (that is, one most significant SNP per gene). Genetic scores were computed using the PLINK[70] score profile procedure. Scores were weighted by the direction of effect of association ($+1$ or -1) of each minor allele with *ICA2* pattern inferred from the methylomic profiling sample. Genetic score calculations were restricted to SNPs meeting the inclusion criteria: MAF ≥ 0.01; HWE $P > 0.0001$; call-rate $\geq 90\%$ within a sample. Resulting genetic profiles were adjusted by regressing out the effect of the number of missing SNPs per-subject included in the scoring procedure. Associations between the inferred scores and behavioural or imaging phenotypes were assessed using Pearson's correlation. Given the negative correlation observed between *ICA2* methylomic pattern and EM performance in the methylomic sample, genetic score correlation tests with EM performance were one-sided (lower tail).

Description of additional analyses of ICA methylomic patterns can be found in Supplementary Methods.

Data availability. The data that support the findings of this study are available from the corresponding authors on request.

References

1. Rakic, P. Evolution of the neocortex. *Nat. Rev. Neurosci.* **10,** 724–735 (2009).
2. Rash, B. G. & Grove, E. A. Area and layer patterning in the developing cerebral cortex. *Curr. Opin. Neurobiol.* **16,** 25–34 (2006).
3. Salthouse, T. A. *et al.* Breadth and age-dependency of relations between cortical thickness and cognition. *Neurobiol. Aging* **36,** 3020–3028 (2015).
4. Fjell, A. M. *et al.* Development and aging of cortical thickness correspond to genetic organization patterns. *Proc. Natl Acad. Sci. USA* **112,** 15462–15467 (2015).
5. Sowell, E. R. *et al.* Longitudinal mapping of cortical thickness and brain growth in normal children. *J. Neurosci.* **24,** 8223–8231 (2004).
6. Shaw, P. *et al.* Intellectual ability and cortical development in children and adolescents. *Nature* **440,** 676–679 (2006).
7. Raznahan, A. *et al.* Patterns of coordinated anatomical change in human cortical development: a longitudinal neuroimaging study of maturational coupling. *Neuron* **72,** 873–884 (2011).
8. Ge, T. *et al.* Massively expedited genome-wide heritability analysis (MEGHA). *Proc. Natl Acad. Sci. USA* **112,** 2479–2484 (2015).
9. Panizzon, M. S. *et al.* Distinct genetic influences on cortical surface area and cortical thickness. *Cereb. Cortex* **19,** 2728–2735 (2009).
10. Deaton, M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
11. Fraga, M. F. & Esteller, M. Epigenetics and aging: the targets and the marks. *Trends Genet.* **23,** 413–418 (2007).
12. Jones, M. J., Goodman, S. J. & Kobor, M. S. DNA methylation and healthy human aging. *Aging Cell* **14,** 924–932 (2015).
13. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13,** R97 (2012).
14. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49,** 359–367 (2013).
15. Reynolds, L. M. *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat. Commun.* **5,** 5366 (2014).
16. Bell, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* **8,** e1002629 (2012).
17. Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20,** 440–446 (2010).
18. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16,** 25 (2015).
19. Smith, A. K. *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* **15,** 145 (2014).
20. Bell, J. T. & Spector, T. D. DNA methylation studies using twins: what are they telling us? *Genome Biol.* **13,** 172 (2012).
21. Liebermeister, W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **18,** 51–60 (2002).
22. Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T. & Huang, X. A review of independent component analysis application to microarray gene expression data. *Biotechniques* **45,** 501–520 (2008).
23. Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* **23,** 137–152 (2004).
24. Rotival, M. *et al.* Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genet.* **7,** e1002367 (2011).
25. Storsve, A. B. *et al.* Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *J. Neurosci.* **34,** 8488–8498 (2014).
26. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15,** R31 (2014).
27. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14,** R115 (2013).
28. Shrout, P. E. & Bolger, N. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol. Methods* **7,** 422–445 (2002).
29. Alexander-Bloch, A., Giedd, J. N. & Bullmore, E. Imaging structural co-variance between human brain regions. *Nat. Rev. Neurosci.* **14,** 322–336 (2013).
30. Colibazzi, T. *et al.* Latent volumetric structure of the human brain: exploratory factor analysis and structural equation modeling of gray matter volumes in healthy children and adults. *Hum. Brain Mapp.* **29,** 1302–1312 (2008).
31. Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl Acad. Sci. USA* **97,** 11050–11055 (2000).
32. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* **9,** 179–194 (1999).
33. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9,** 195–207 (1999).
34. Horn, J. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30,** 179–185 (1965).
35. Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R. & Caldas, C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.* **3,** e161 (2007).
36. Bauer, M. *et al.* Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation. *Clin. Epigenetics* **7,** 83 (2016).
37. Besingi, W. & Johansson, A. Smoke-related DNA methylation changes in the etiology of human disease. *Hum. Mol. Genet.* **23,** 2290–2297 (2014).
38. Su, D. *et al.* Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. *PLoS ONE* **11,** e0166486 (2016).
39. Tsaprouni, L. *et al.* Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* **9,** 1382–1396 (2014).
40. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11,** R14 (2010).
41. Slieker, R. C. *et al.* Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* **17,** 191 (2016).
42. Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.* **15,** R37 (2014).
43. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24,** 1725–1733 (2014).
44. Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** e1001058 (2010).
45. Heck, A. *et al.* Converging genetic and functional brain imaging evidence links neuronal excitability to working memory, psychiatric disease, and brain activity. *Neuron* **81,** 1203–1213 (2014).
46. Heck, A. *et al.* Genetic analysis of association between calcium signaling and hippocampal activation, memory performance in the young and old, and risk for sporadic Alzheimer disease. *JAMA Psychiatr.* **72,** 1029–1036 (2015).
47. Huttenlocher, P. R. Synaptic density in human frontal cortex—developmental changes and effects of aging. *Brain Res.* **163,** 195–205 (1979).
48. Fjell, A. M. & Walhovd, K. B. Structural brain changes in aging: courses, causes and cognitive consequences. *Rev. Neurosci.* **21,** 187–221 (2010).
49. Deoni, S. C. L., Dean, D. C., Remer, J., Dirks, H. & O'Muircheartaigh, J. Cortical maturation and myelination in healthy toddlers and young children. *Neuroimage* **115,** 147–161 (2015).
50. Marsland, A. L. *et al.* Brain morphology links systemic inflammation to cognitive function in midlife adults. *Brain. Behav. Immun.* **48,** 195–204 (2015).
51. Weaver, J. D. *et al.* Interleukin-6 and risk of cognitive decline: MacArthur studies of successful aging. *Neurology* **59,** 371–378 (2002).
52. Louveau, A. *et al.* Structural and functional features of central nervous system lymphatic vessels. *Nature* **523,** 337–341 (2015).
53. Ek, M. *et al.* Inflammatory response: pathway across the blood-brain barrier. *Nature* **410,** 430–431 (2001).
54. Yirmiya, R. & Goshen, I. Immune modulation of learning, memory, neural plasticity and neurogenesis. *Brain. Behav. Immun.* **25,** 181–213 (2011).
55. Adalsteinsson, B. T. *et al.* Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PLoS ONE* **7,** e46705 (2012).
56. Satizabal, C. L., Zhu, Y., Mazoyer, B., Dufouil, C. & Tzourio, C. Circulating IL-6 and CRP are associated with MRI findings in the elderly: the 3C-Dijon Study. *Neurology* **6,** 720–727 (2012).
57. Luck, T. *et al.* Mild cognitive impairment in general practice: age-specific prevalence and correlate results from the German study on ageing, cognition and dementia in primary care patients (AgeCoDe). *Dement. Geriatr. Cogn. Disord.* **24,** 307–316 (2007).
58. Stein, J. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44,** 552–561 (2012).
59. Schmaal, L. *et al.* Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol. Psychiatry* **21,** 806–812 (2016).
60. Kristensen, L. S., Mikeska, T., Krypuy, M. & Dobrovic, A. Sensitive melting analysis after real time- methylation specific PCR (SMART-MSP): high-throughput and probe-free quantitative DNA methylation detection. *Nucleic Acids Res.* **36,** e42 (2008).
61. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11,** 1138–1140 (2014).
62. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* **13,** R44 (2012).

63. Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6,** 4 (2013).
64. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–209 (2013).
65. Bellenguez, C. *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28,** 134–135 (2012).
66. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).
67. Plerou, V. *et al.* Random matrix approach to cross correlations in financial data. *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.* **65,** 066126 (2002).
68. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27,** 1496–1505 (2011).
69. Preacher, K. J. & Kelley, K. Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychol. Methods* **16,** 93–115 (2011).
70. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

Acknowledgements

This work was funded by the Swiss National Science Foundation (Sinergia grant CRSI33_130080 to D.J.F.d.Q. and A.P., individual grant 320030_159740 to D.J.F.d.Q.), by the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement #602450 (IMAGEMEND), by the German Research Network on Dementia (KND) and the German Research Network on Degenerative Dementia (KNDD), German Federal Ministry of Education and Research grants 01GI0420 and 01GI0711, and by the University of Basel. We thank Elmar Merkle, Christoph Stippich and Oliver Bieri for granting access to the MRI facilities of the University Hospital Basel. Preprocessing of the brain imaging data was performed at sciCORE (http://scicore.unibas.ch/), the centre for scientific computing at the University of Basel.

The Munich sample comprised images acquired in subsamples of the Munich Antidepressant Response Signature Study and the Recurrent Unipolar Depression Case-Control study, both performed at the MPIP. We thank radiographers Rosa Schirmer, Elke Schreiter and Ines Eidner for image acquisition, data organization and preparation. The studies have been supported by a grant of the Exzellenz-Stiftung of the Max Planck Society and also been funded by the Federal Ministry of Education and Research (BMBF) in the framework of the National Genome Research Network (NGFN), FKZ 01GS0481.

Author contributions

V.F., A.P. and D.J.F.d.Q. conceived and designed the study. V.F., T.C.-R., A.M., P.G.S., V.V., D.C., P.D., T.E., L.G., E.L., C.V., D.J.F.d.Q. and A.P. analysed the data. F.J., W.M., S.G.R.-H., M.S., M.W. and E.B.B. provided samples and analysed the data. A.P., V.F. and D.J.F.d.Q. wrote the manuscript. All collaborators reviewed and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

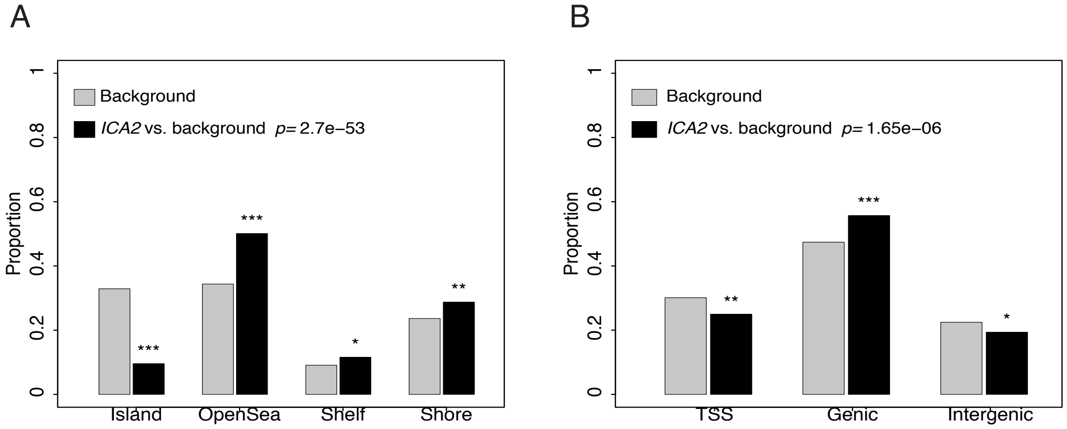
How to cite this article: Freytag, V. *et al.* A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory. *Nat. Commun.* **8,** 15193 doi: 10.1038/ncomms15193 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Figures

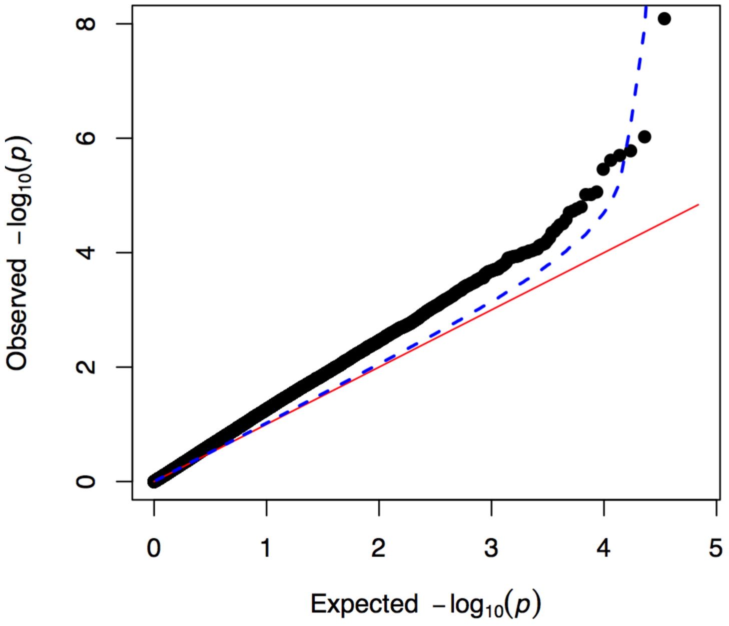
Supplementary Figure 1: Distribution of *ICA2*-contributing CpGs based on topographical distribution **(A)** and relative location to gene transcript **(B)**.

Background corresponds to the genome-wide distribution of CpGs across the arrays. *ICA2* distribution was compared to background distributions using goodness of fit χ^2 test: *p*: p-value; stars indicates the magnitude of the bins standardized residuals: *: > |2|, **: >|3|, ***: > |4|.



Supplementary Figure 2: Q-Q plot of mQTL analysis between 71 GSEA genetic score SNPs and *ICA2* CpGs.

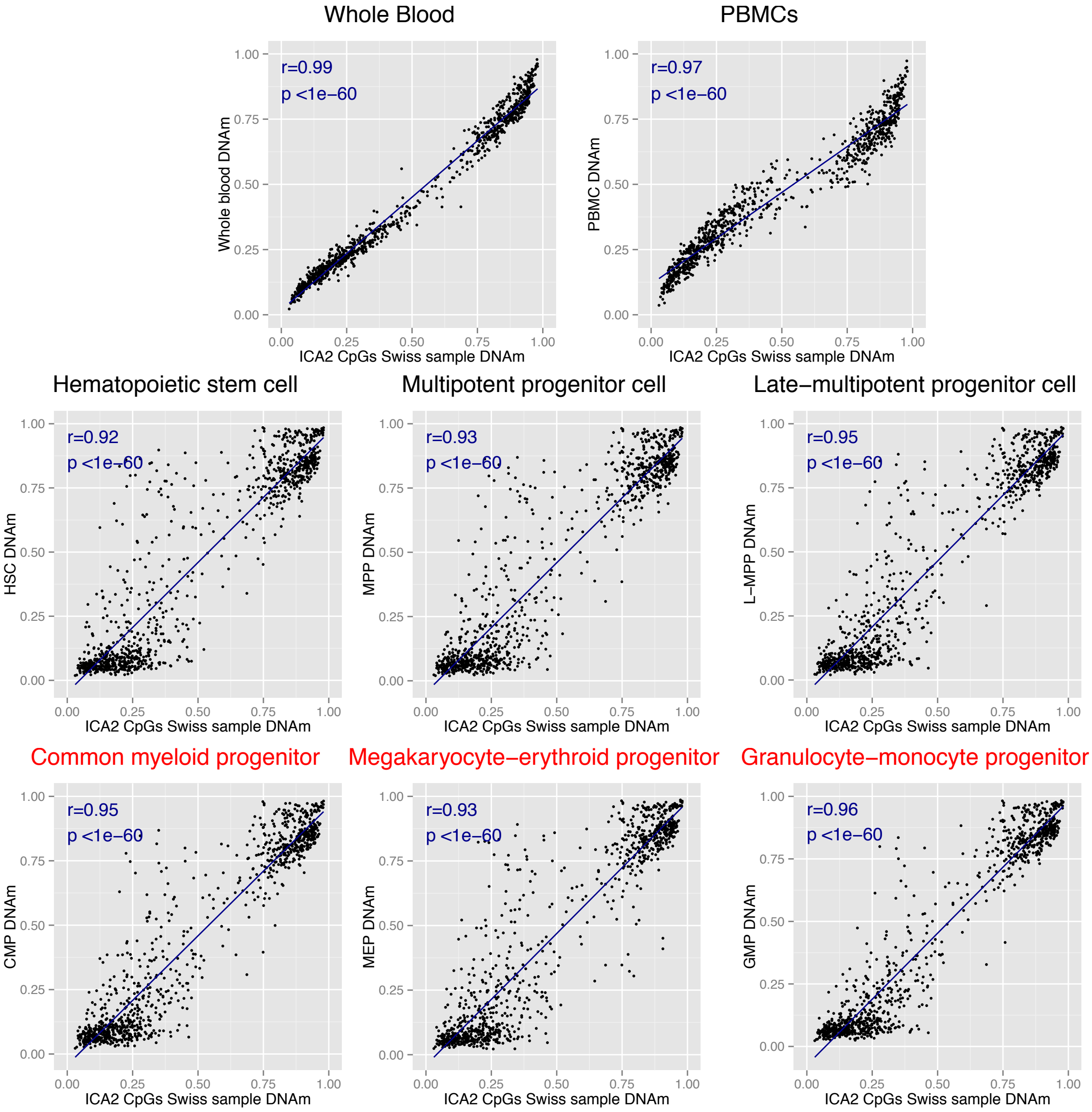
Red line shows expected uniform distribution.
Blue dashed line indicates the 95 % quantiles obtained from 1000 repeats of association testing between *ICA2* CpGs and randomly selected cis-SNPs.



Supplementary Figure 3: Average whole-blood DNAm at 970 *ICA2* CpGs versus progenitor cell specific DNAm.

Horizontal axis: average whole-blood DNAm observed in the methylomic Swiss sample (n=533).

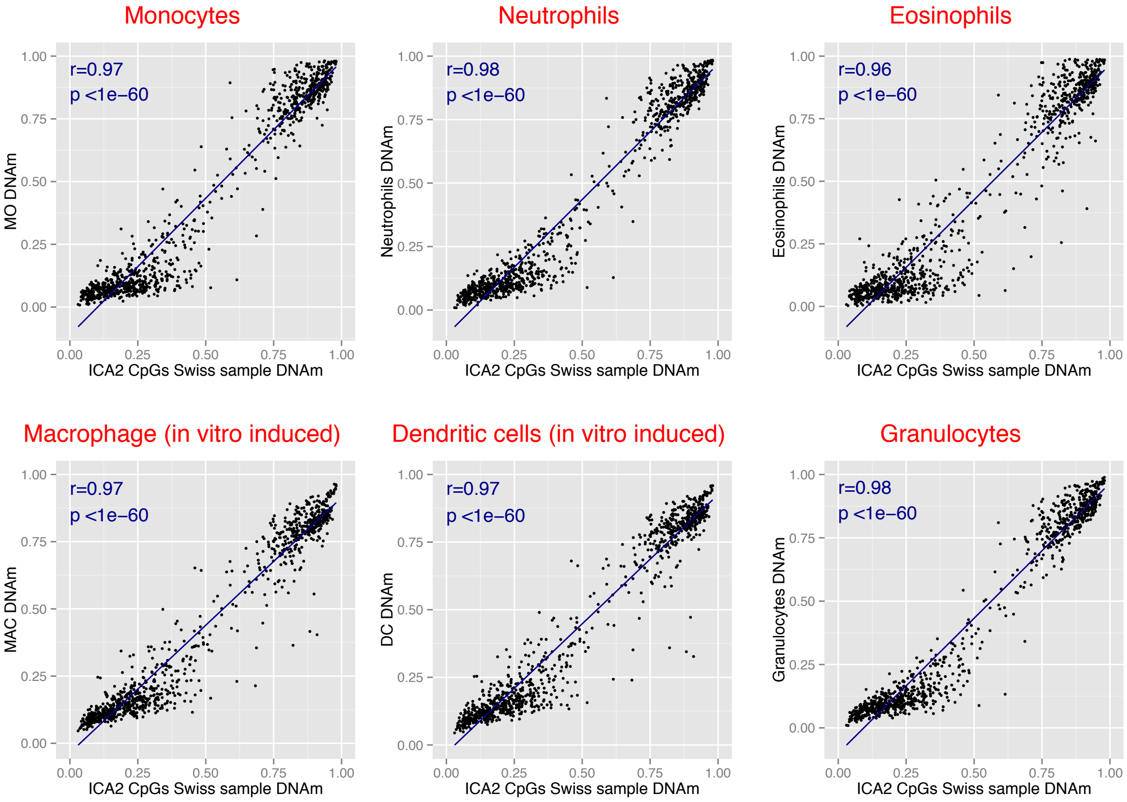
Vertical axis: average DNAm observed in progenitor cells.



Supplementary Figure 4: Average whole-blood DNAm at 970 *ICA2* CpGs versus cell subtypes specific DNAm.

Horizontal axis: average whole-blood DNAm observed in the methylomic Swiss sample (n=533).

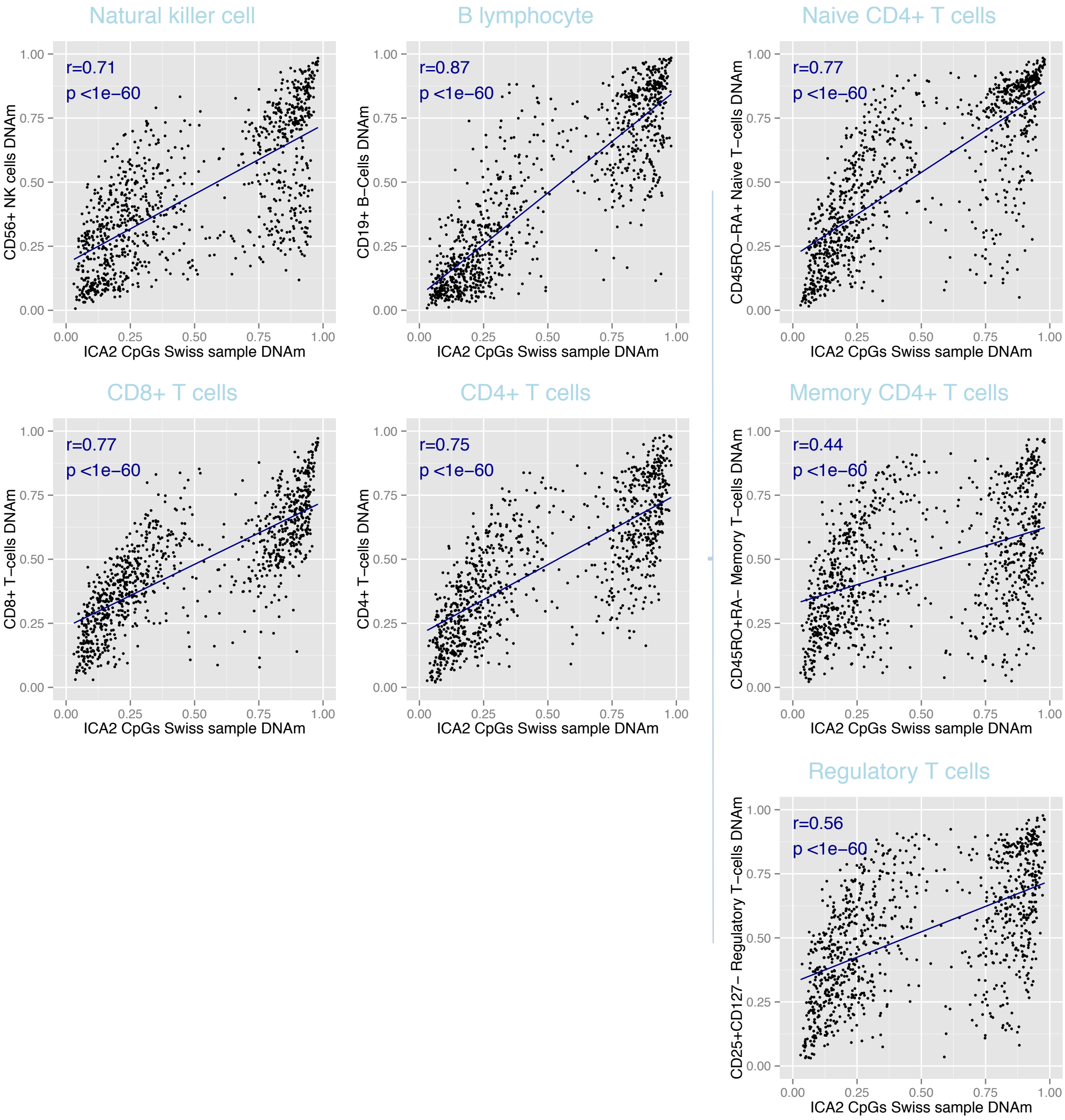
Vertical axis: average DNAm observed in specific cell subtypes.



Supplementary Figure 5: Average whole-blood DNAm at 970 *ICA2* CpGs versus lymphocytes subtypes specific DNAm.

Horizontal axis: average whole-blood DNAm observed in the methylomic Swiss sample (n=533).

Vertical axis: average DNAm observed in specific cell subtypes.



Supplementary Tables

Supplementary Table 1: Description of methylomic profiling and imaging analysis samples.

	Methylomic profiling sample	**Basel imaging sample**
N	533	722
Age at MRI/behavioral assessment in years (mean ± SD)	22.9 ± 3.3	22.9 ± 3.3
Age blood sampling in years (*)	23.9 ±3.5	/
Number of females (%)	311 (58.3%)	447 (61.9%)
Episodic Memory-*N*	531	722
Structural imaging-*N* (including Episodic Memory-*N)*	514 (512)	722 (722)

(*): blood sampling for methylomic profiling.

Supplementary Table 2: Correlations between 15 ICA methylomic patterns and global thickness.

(a) Cortical measures were adjusted for sex, intracranial volume and MR-technical batches using linear regression. (b) Chronological age effects were further partialled out from cortical measures and *ICA* component. r: Pearson's correlation coefficient; *p*: two-sided test p-value.

	Thickness *N = 514*				**Age *N = 533***	
	No age adjustment		**Age adjustment**			
ICA	**r**	***p***	**r**	***p***	**r**	***p***
ICA1	-0.14	0.00162	0.01	0.83	0.54	1.54E-42
ICA2	-0.24	3.86E-08	-0.18	6.55E-05	0.29	4.68E-12
ICA3	-0.03	0.548	-0.01	0.901	0.08	0.0578
ICA4	0.01	0.837	-0.01	0.806	-0.08	0.0691
ICA5	-0.06	0.166	-0.08	0.0721	-0.06	0.166
ICA6	0.1	0.0282	0.08	0.0545	-0.06	0.166
ICA7	0.01	0.819	0	0.937	-0.05	0.281
ICA8	0.04	0.414	0.04	0.396	-0.04	0.337
ICA9	0.09	0.0368	0.08	0.0585	-0.04	0.339
ICA10	-0.03	0.467	-0.03	0.566	0.03	0.42
ICA11	0.03	0.465	0.04	0.311	0.03	0.454
ICA12	-0.05	0.224	-0.06	0.153	-0.02	0.608
ICA13	-0.06	0.171	-0.06	0.148	-0.02	0.647
ICA14	0	0.992	0.01	0.885	0.02	0.68
ICA15	-0.04	0.369	-0.04	0.339	-0.02	0.695

Supplementary Table 3: Correlations between cortical thickness factor scores and *ICA2*.

r: Pearson's correlation coefficient. *p*: two-sided correlation test p-value.

Factor	***ICA2*** ***N = 514***	
	r	***p***
F1	-0.112	0.0108
F2	-0.09	0.0406
F3	-0.065	0.142
F4	-0.08	0.0707
F5	-0.013	0.774
F6	-0.13	0.00314
F7	0.004	0.935
F8	-0.013	0.776

Supplementary Table 4: Correlations between 15 ICA methylomic components and EM performance.

Age adjustment: chronological age effects were partialled out from each measure.
r: Pearson's correlation coefficient. *p*: two-sided correlation test p-value.

ICA	**Methylomic profiling sample** ***N= 531***			
	No Age adjustment		**Age adjustment**	
	r	***p***	***r***	***p***
ICA1	-0.043	0.328	-0.003	0.944
ICA2	-0.138	0.00147	-0.122	0.00491
ICA3	0.011	0.809	0.017	0.696
ICA4	-0.018	0.683	-0.024	0.586
ICA5	0.021	0.627	0.017	0.699
ICA6	0.06	0.165	0.057	0.19
ICA7	-0.007	0.875	-0.009	0.829
ICA8	0.09	0.0375	0.089	0.0404
ICA9	-0.002	0.972	-0.005	0.909
ICA10	-0.015	0.739	-0.012	0.785
ICA11	-0.068	0.119	-0.066	0.13
ICA12	-0.011	0.804	-0.013	0.763
ICA13	0.017	0.691	0.016	0.711
ICA14	-0.023	0.598	-0.021	0.622
ICA15	0.039	0.364	0.039	0.374

Supplementary Table 5: Correlations between cortical thickness factor scores and EM performance.

r: Pearson's r correlation coefficient; *p*: two-sided correlation test p-value.

Factor	**Combined sample (*N*=1234)**		**Basel imaging sample (*N*=722)**		**Methylomic profiling sample (*N*=512)**	
	r	***p***	**r**	***p***	**r**	***p***
F1	-0.054	0.0575	-0.086	0.0215	-0.01	0.823
F2	-0.045	0.115	-0.072	0.0548	-0.006	0.901
F3	-0.012	0.663	-0.023	0.541	0.001	0.973
F4	-0.024	0.395	0.006	0.869	-0.071	0.107
F5	-0.05	0.0812	-0.04	0.288	-0.066	0.139
F6	0.079	0.00574	0.102	0.00617	0.048	0.283
F7	-0.007	0.814	-0.018	0.634	0.008	0.85
F8	0.026	0.361	0.05	0.183	-0.009	0.839

Supplementary Table 6: Association between age, global cortical thickness and EM performance and WBC counts.

Results from linear models analysis with phenotype as dependent variable and WBC counts as explanatory variables. *t*: t-statistic value; *F*: Overall effect F-statistic value; *p*: p-value
(a): Adjustment for sex, intra-cranial volume, MR batches and chronological age.
(b): Adjustment for sex and chronological age.
(c): Basophils, Eosinophils and Monocytes.

Phenotype	**N**		***t***	***p***
Age	527	Lymphocytes	0.84	0.40
		Neutrophils	0.31	0.76
		Mixture (c)	0.57	0.57
			F(3,523) = 0.60, *p* = 0.62	
Global cortical thickness (a)	509	Lymphocytes	-0.87	0.39
		Neutrophils	-2.32	0.02
		Mixture (c)	0.31	0.75
			F(3,505) = 2.2, *p* = 0.087	
EM performance (b)	525	Lymphocytes	1.69	0.093
		Neutrophils	0.21	0.84
		Mixture (c)	1.1	0.29
			F(3,521) = 2.1, *p* = 0.10	
Cortical thickness F6	509	Lymphocytes	0.65	0.51
		Neutrophils	-0.66	0.51
		Mixture (c)	-0.28	0.78
			F(3,505) = 0.28, *p* = 0.84	

Supplementary Notes

1. Description of Munich sample

1.1 Structural imaging

MRI acquisition: High resolution T1-weighted images were acquired at the Neuroimaging Core Unit of the MPIP on a clinical 1.5 Tesla MR scanner (Signa/Signa Excite, General Electric, for sequence details see [1,2]). *MRI data processing:* Gross morphological abnormalities such as tumor or territorial infarction, ventricle asymmetries or arachnoid cysts preventing automated image processing, extensive white matter disease or motion artefacts were exclusion criteria prior to the formation of this combined sample. The surface-based segmentation stream of FreeSurfer (version 5.3, installed on 64-bit Linux workstations) was applied to all T1-weighted images, with substeps as described in the Structural Imaging section. Visual QC of cortical segmentation quality was performed on the basis of standardized protocols (http://enigma.ini.usc.edu/protocols/imaging-protocols) and led to exclusion of 12 subjects. As phenotypes of interest, left and right cortical thickness (the average of which is ref. to as cortical thickess [CT]), and intracranial volume derived indirectly from the spatial registration procedure.

1.2 Methylomic profiling

DNA was extracted from whole blood using the Gentra Puregene Blood Kit (QIAGEN). Quality and quantity of the DNA were assessed by NanoDrop 2000 Spectrophotometer (Thermo Scientific) and Quant-iT Picogreen (Invitrogen). Genomic DNA was bisulfite converted using the Zymo EZ-96 DNA Methylation Kit (Zymo Research) and genome-wide methylome levels were assessed with the Illumina Infimium HumanMethylation 450K BeadChip array. Hybridization and processing was performed according to manufacturer' s instructions. Intensity read outs, normalization and estimation and beta values were obtained using the Minfi package (version 1.21.0) in Bioconductor [3]. Beta values for the pre-selected 397,947 autosomal probes from the Swiss sample were calculated from SWAN normalized intensities. After pre-processing of methylomic data, and MRI-QC based exclusions, combined data of N=596 subjects was available for statistical analysis.

2. Description of AgeCode sample

Briefly, participants were recruited between January 2003 and November 2004 in six German study centers (Bonn, Düsseldorf, Hamburg, Leipzig, Mannheim, Munich) via general practitioners (GP) connected to the respective study sites. Inclusion criteria were age of 75 years and older, absence of dementia (according to the GP's judgment) and at least one contact with the GP within the last 12 months. Exclusion criteria were GP consultations by home visits only, residence in a nursing home, presence of a severe illness with an anticipated fatal outcome within three months, insufficient German language abilities, deafness or blindness, lack of ability to provide an informed consent and status as being only an occasional patient of the participating GP. A total of 3'327 subjects were successfully contacted and assessed with structured clinical interviews at their homes. A total of 110 individuals were excluded after the first interview due to presence of dementia or an actual age below 75 (falsely classified as 75 or older in the sample selection process). For the present analyses, data from baseline and three follow-up measurements with 18 months intervals were available. In a primary care-based sample of older individuals, conditions can be present that affect cognition and the reliability of neuropsychological tests. In order to generate a sample of healthy elderly individuals we further employed the following selection criteria at baseline: Age between 75 and 90 years, German as native language, at least school-leaving certificate, absence of severe hearing or vision impairments, absence of insufficient test motivation as judged by the interviewer, absence of disturbing factors during neuropsychological testing and absence of all of the following comorbid conditions: Parkinson's disease, epilepsy, alcohol abuse, stroke, multiple sclerosis, evidence of depression (a score of 6 or higher on the Geriatric Depression Scale [4]), traumatic brain injury with unconsciousness of more than 30 minutes, visible neurological malfunctions and dementia according to DSM-IV criteria [5]. In addition, we excluded subjects who converted to dementia up to the third follow-up or without neuropsychological test data available on baseline and all follow-up visits. After application of these selection criteria, a total of 1244 subjects remained in the sample. Sufficient DNA-samples for genome-wide genotyping were available for 782 subjects.

3. Description Episodic Memory phenotypes

3.1 Methylomic and Basel imaging samples

While undergoing fMRI acquisition, all participants completed a picture delayed free recall task. Stimuli consisted of 72 emotional and neutral pictures (24 negative, 24 positive and 24 neutral) taken from the International Affective Picture System (IAPS) [6] and from in-house standardized picture sets. Four additional pictures showing neutral objects were used to control for primacy and recency effects in memory. These pictures were not included in the analysis. Additionally, 24 scrambled pictures were included. Their background contained the color information of all pictures used in the experiment and was overlaid with a crystal and distortion filter (Adobe Photoshop CS3, Adobe Systems Inc., San Jose, CA, USA). On the foreground geometrical figures of varying shape, size and orientation were shown. Pictures were presented for 2.5 s each in a quasi-randomized order so that a maximum of four pictures of the same category (e.g. animals, humans, landscape) and valence occurred consecutively. Between the pictures a fixation-cross appeared on the screen for 500ms and the trials were separated by a variable intertrial period of 9-12 s. During this time subjects were asked to rate the presented picture for valence (negative, neutral, positive) and arousal (large, medium, small) on a three-point rating scales (Self Assessment Manikin). Scrambled pictures were rated according to their shape (vertical, symmetric or horizontal) and size (large, medium, small). Subjects were not instructed to recall the pictures later (incidental recall). The delayed free recall was performed outside of the scanner, 10 min after presentation of all photographs. To document performance for the delayed recall of positive, negative, and neutral pictures, subjects had to describe each picture by writing it down in a few words. A picture was judged as correctly recalled if the rater could identify the presented picture based on the subject's description. Two blinded investigators independently rated the descriptions for recall success (inter-rater reliability > 99%). For the pictures, which were judged differently by the two raters (i.e. a particular picture was judged as correctly recalled by one rater but not the other), a third independent and blinded rater made a final decision with regard to whether the particular picture could be considered as successfully recalled. The number of correctly recalled pictures served as a phenotype. For initial association testing with *ICA2* pattern, EM performance was adjusted for sex effects using linear regression. For additional analyses, including genetic scoring analyses, EM performance was further adjusted for chronological age effect.

3.2 Basel cognitive sample

Participants performed the same pictures free recall task as described for the methylomic profiling sample, without fMRI assessment. EM performance, used in genetic scoring analysis, was adjusted for sex and chronological age effects using linear regression.

3.3 Zurich sample

Subjects viewed six series of five semantically unrelated nouns presented at a rate of one word per second with the instruction to learn the words for immediate free recall after each series. In addition, subjects underwent an unexpected delayed free-recall test of the learned words after 5min (episodic memory). The number of correctly recalled words (hits) was the relevant output. EM performance, used in genetic scoring analysis, was adjusted for sex and chronological age effects using linear regression.

3.4 AgeCode

Delayed recall performance as quantified by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) battery [7] served as phenotype. Subjects were presented a list of 10 words three times (presentation per word: 2 seconds), each time presented in a different order. After each run, subjects freely recalled as many words as possible. The number of correctly remembered items (free recall) after a 10 min delay served as the phenotypic measure. EM performance, used in genetic scoring analysis, was adjusted for sex and chronological age effects using linear regression.

4. Analysis of *ICA* methylomic patterns

4.1 Comparison of whole-blood and cell-specific DNAm values

Average DNA methylation values were obtained from four publically available datasets from 19 cell types. Average DNAm values from hematopoietic stem cells and progenitor cells were obtained from GSE63409 [8] considering only normal bone marrow samples. Average DNAm from whole-blood, PBMCs, Natural Killer cells, B-lymphocytes, CD4 T-cells, CD8 T-cells, monocytes, neutrophils, eosinophils and granulocytes were obtained from GSE3560 [9]. Average DNAm from specific sub-types of CD4 T-cells (naive, memory and regulatory CD4 T-cells) were obtained from GSE59250 [10] considering control samples only. Average DNAm in dendritic cells and macrophage (in vitro induced) were obtained from GSE75937 [11].

4.2 ICA analysis of publically available whole-blood methylomic profiles

We analyzed whole-blood methylomic profiles from 656 samples reported in Hannum et al., 2013 [12]. In analogy to our methylomic dataset, multi-mapping or polymorphic probes were excluded from analysis. Raw intensities (methylated and unmethylated signals) were normalized using the lumi package (color-bias adjustment and quantile normalization). The BMIQ algorithm was finally applied to adjust for the difference between Type I and Type II probes used in the 450K array. Given substantial non-randomness of between-plate distribution of chronological age in this sample, we performed CoMbat adjustment for plate effect. DNA methylation values were subsequently adjusted for sex and 98 surrogate variables inferred from surrogate variable analysis (SVA). ICA decomposition on the adjusted signals yielded a total of 175 components, among which 19 were retained based on the per-subject 10% variance criterion used in our methylomic dataset. The retained ICA patterns were tested for association with age, after adjustment for estimated cell counts (CD4T, CD8T, NK, Gran, Mono, Bcell). Five patterns were significantly associated with age. In analogy to our study, CpGs contributing to these patterns were chosen so as to exhibit an absolute loading > |4| on the respective pattern.

4.3 Association of *ICA* patterns with chronological age in cell-specific methylomic profiles

We used publically available methylomic profiles from N=1202 monocytes samples (GSE56046) and N=214 CD4 T-cells samples (GSE56581) [13]. Normalized datasets deposited on GEO repository were considered for analysis. In each dataset a Surrogate Variable Analysis preserving for chronological age was performed. Individual methylomic values were adjusted for the inferred SVs using linear regression. In each dataset, *ICA1* and *ICA2* patterns were estimated as the linear combination between the inverse of genome-wide *ICA1* and *ICA2* loadings (inferred from the Swiss sample) and scaled SV-adjusted DNAm values. This score was subsequently tested for association with chronological age.

5. mQTL analysis

5.1 Testing over-representation of genetic score SNPs in -cis to ICA2 CpGs

We randomly selected an equal number of SNPs from genome-wide genotyped SNPs and assessed the occurence of SNPs found in -*cis* (± 1 Mbp) to *ICA2* CpGs. This sampling procedure was repeated 5000 times to establish the null distribution and calculate the corresponding *p*-value.

5.2 Null distribution of cis-mQTL association statistics

First we determined all SNPs located within ± 1Mbp of any of the *ICA2* CpGs ('cis-SNP pool'). Association statistics were computed between *ICA2* CpGs and *n* SNPs randomly selected from the cis-SNP pool, with *n* equal to the number of GSEA genetic score SNPs (i.e. 71 SNPs), thus providing one realization of the baseline quantile distribution. This sampling procedure was repeated 1000 times.

Supplementary References

1. Stein, J. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44,** 552–61 (2012).

2. Hibar, D. P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520,** 224–9 (2015).

3. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–1369 (2014).

4. Yesavage, J. A. *et al.* Development and validation of a geriatric depression screening scale: a preliminary report. *J. Psychiatr. Res.* **17,** 37–49 (1983).

5. American Psychiatric Association. *Diagnostical and Statistical Manual of Mental Disorders (4th ed., text rev.)*. (2000).

6. Lang, P. J., Bradley, M. & Cuthberth, B. N. International Affective Pictures System (IAPS): Affective Ratings of Pictures and Instruction Manual. *Univ. Florida, Gainesville, FL* (2008).

7. Welsh, K. A. *et al.* The consortium to establish a registry for Alzheimer's disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology* **44,** 609–14 (1994).

8. Jung, N., Dai, B., Gentles, A. J., Majeti, R. & Feinberg, A. P. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nat. Commun.* **6,** 8489 (2015).

9. Reinius, L. *et al.* Differential DNA Methylation in Purified Human Blood Cells : Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS One* **7,** e41361 doi: 10.1371/journal.pone.0041361 (2012).

10. Absher, D. M. *et al.* Genome-Wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS Genet.* **9,** e1003678 doi: 10.1371/journal.pgen.1003678 (2013).

11. Vento-Tormo, R. *et al.* IL-4 orchestrates STAT6-mediated DNA demethylation leading to dendritic cell differentiation. *Genome Biol.* **17,** 4 doi: 10.1186/s13059–015–0863–2 (2016).

12. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49,** 359–67 (2013).

13. Reynolds, L. M. *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat. Commun.* **5,** 5366 (2014).

Publication 2 Genetic estimators of DNA methylation provide insights into the molecular basis of polygenic traits

Title: Genetic estimators of DNA methylation provide insights into the molecular basis of polygenic traits

Authors: Virginie Freytag[1,2,*], Vanja Vukojevic[1,2,3], Annette Milnik[1,2,4], Christian Vogler[1,2,4], Dominique J.-F. de Quervain[2,4,5,+], Andreas Papassotiropoulos[1,2,3,4,*,+]

Affiliations:

[1]Division of Molecular Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland

[2]Transfaculty Research Platform Molecular and Cognitive Neurosciences, University of Basel, CH-4055 Basel, Switzerland

[3]Department Biozentrum, Life Sciences Training Facility, University of Basel, CH-4056 Basel, Switzerland

[4]Psychiatric University Clinics, University of Basel, CH-4055 Basel, Switzerland

[5]Division of Cognitive Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland

+ These authors jointly supervised this work.

Running title: Genetic estimators of DNA methylation

Keywords: intermediate molecular trait**,** epigenetics

ABSTRACT

The large biological distance between genetic risk loci and their mechanistic consequences in the tissue of interest limits the ability to establish functionality of susceptibility variants for genetically complex traits. Such a biological gap may be reduced through the systematic study of molecular mediators of genomic action, such as epigenetic modification. Here, we report the identification of robust genetic estimators of whole-blood CpG methylation, which can serve as intermediate molecular traits amenable to association testing with other genetically complex traits. We describe the relationship between these estimators and gene expression, demonstrate their genome-wide applicability to association testing even in the absence of individual genotypic data, and show that these estimators powerfully identify methylation-related genomic loci associated with the risk for schizophrenia, a common and genetically complex disorder. The use of genetic estimators for blood DNA methylation, which are made publically available (http://mcn.unibas.ch/files/EstiMeth_Distribution_v1.zip , password: mcnEstiMeth140510), can serve as a valuable tool for the identification of epigenetic underpinnings of complex traits.

INTRODUCTION

Improving understanding, diagnosis, and therapy of human disease has been one of the central promises of the human genome project (Editorial 2011). This promise is being increasingly fulfilled. For example, cancer research has benefited dramatically from the discoveries related to the human genome (Lander 2011), mainly because the genomic mechanisms leading to the development of many cancers are amenable to direct observation. However, the situation is slightly different in disorders for which the underlying molecular events are not easily accessible, as is the case for mental disorders (Papassotiropoulos and de Quervain 2015). Advances in the development of high-throughput genotyping and analytical software, and the launch of large collaborative efforts have led to the identification of numerous well-validated genetic risk factors for such common disorders. However, the functional relevance of most discovered loci and the molecular mechanisms behind the reported genetic association signals remain elusive (Gamazon et al. 2015).

One of the main reasons for the limited ability to establish functionality of susceptibility variants is the large biological distance between a genetic polymorphism and its related mechanistic consequences in the tissue of interest. Such biological gap may be reduced by the study of molecular mediators of genomic action, such as gene expression (Gamazon et al. 2015). For example, in such common neuropsychiatric disorders as schizophrenia, genetic susceptibility variants are significantly enriched in promoter and enhancer regions and point to a functional link between disease-associated noncoding single nucleotide polymorphisms (SNPs) and transcriptional regulation in the brain (Roussos et al. 2014). The integration of transcriptomics data in the study of the genetic factors of complex traits has significantly improved our understanding of their genetic basis (Lonsdale et al. 2013). Thus, methods that reduce the gap between genetic susceptibility and its functional consequences are expected to increase our understanding of the genetic underpinnings of genetically complex traits.

Genetic estimators for gene expression have been recently proposed in this context (Gamazon et al. 2015; Gusev et al. 2016) . These methods capitalize on the joint additive effects of *cis* markers on a given expression trait to estimate gene expression from individual genotypes. At the population level, the derived genetic estimates represent an intermediate molecular trait, amenable to association testing with the phenotype under study. This approach can be viewed as genetic correlation testing for which a significant association is interpreted as existence of shared co-localizing genetic factors between the complex phenotype and the investigated expression trait.

Here we report the generation of robust genetic estimators of epigenetic regulation as an attempt to provide insights into the molecular basis of polygenic traits by minimizing the biological gap between genetic variation and its functional impact. We focused on DNA methylation (specifically on the methylation of 5'-C-phosphate-G-3' (CpG) sites), the most extensively studied epigenetic modification to date, which directly regulates important molecular processes such as gene expression, imprinting, and chromosomal inactivation (Deaton and Bird 2011; Schübeler 2015; Lev Maor et al. 2015). High-throughput methylomic profiling studies have highlighted the strong local genetic regulation of DNA methylation (Bell et al. 2011; Gutierrez-Arcelus et al. 2013; Hannon et al. 2016b; Jaffe et al. 2016; Lemire et al. 2015), with possibly multiple co-localized markers contributing independently to variation in DNA methylation at individual CpG sites (Bonder et al. 2017).

We generated genetic estimators of DNA methylation (DNAm), that allow testing for localized shared genetic contributions between DNAm variation and complex traits. We demonstrate their applicability even to studies providing summary SNP statistics only, and show exemplarily that such estimators result in the identification of epigenetic underpinnings of a common neuropsychiatric disease.

RESULTS

Estimation of genetically driven DNA methylation

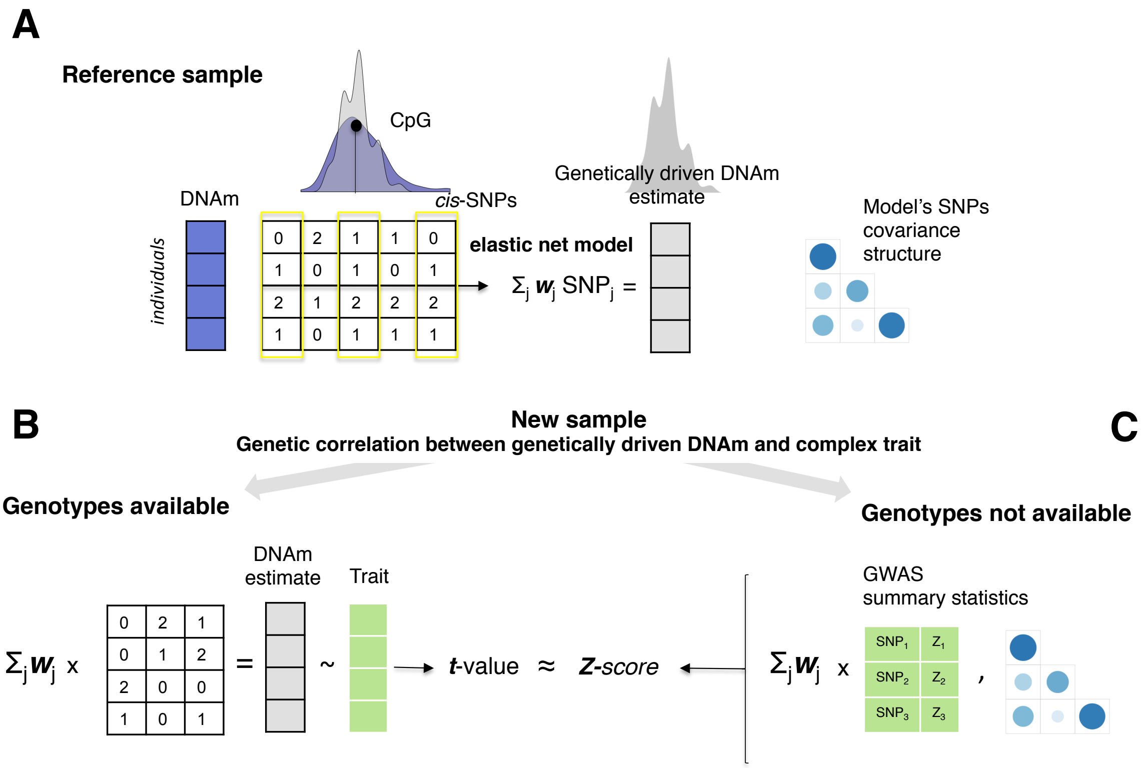
We estimated, under an additive genetic model, the genetically driven proportion of DNAm at a given CpG site, as a linear combination of SNPs in -*cis* of that site. Starting from a reference dataset of samples for which both methylation and genotypic data are measured, the weights of this linear combination can be obtained using a multiple regression approach between SNPs and corresponding DNAm. In analogy to the approach adopted previously for gene expression (Gamazon et al. 2015), we opted for a elastic net penalized multiple regression method (Zou and Hastie 2005) to infer SNP weights of the DNAm estimators (**Figure 1-A**). This method comes with the advantage of performing marker selection, thus providing sparse solutions. Subsequently, the genetically driven DNAm signal can be estimated in independent individuals as the linear combinations of the inferred weights and observed genotypes (**Figure 1-B**). In this independent sample, the derived genetic estimate of DNAm at a given CpG is amenable to genetic correlation testing with the phenotype under study (**Figure 1-B**).

Our reference dataset comprised N=533 healthy young adults (BASEL1 sample, see Methods), who underwent both whole-blood methylomic profiling and genome-wide SNP assessment. Prior to analysis, the DNAm signal was adjusted for technical and biological confounders (see Methods), and genotypes were imputed using the Michigan imputation server (https://imputationserver.sph.umich.edu/index.html, see Methods).

In the reference sample, a elastic net model was trained between common *cis*-SNPs (MAF>0.05, located within ± 1Mbp of a CpG site) and adjusted DNAm signal at each individual CpG site (see Methods). From 395,014 CpG sites investigated, a total of 236,923 non-null models (i.e., at least one site selected by penalized regression) could be fitted, with cross-validation r^2 accounting on average for 6.9 % of variance of the DNAm signal.

Figure 1: Estimation of genetically driven DNAm for genetic association testing with complex traits

(**A**) In a reference sample, a elastic net penalized multiple regression model is built between SNPs in -*cis* of a given CpG site, and DNAm signal (blue). The linear combination of the inferred weights *w* at selected genotypes (encircled in yellow) represents the genetically driven estimate of DNAm signal (grey). (**B**) The genetic model is used to estimate genetically driven DNAm in independent individuals, from observed genotypes ; this estimate can be tested for association with a sample's trait. In case genotypic data are not accessible (**C**), the association statistic can be approximated using the model's weights, the trait GWAS summary statistics (SNP to trait association) and the covariance structure of model's SNPs inferred from a reference sample (different blue dot sizes represent different covariance levels between pairs of SNPs in the reference sample). Figure 1 from Gusev et al. (Gusev et al. 2016) served as a template for this figure.



Unlike univariate testing, the elastic net approach allows for simultaneous modeling of the joint effects of multiple *cis*-markers, that are likely to impact on DNAm at a given CpG site (Bonder et al. 2017). We compared the fraction of variance of DNAm explained by the elastic net models (cross-validation r^2) with the fraction of variance explained by the single best *m*QTL identified at each CpG site. We observed substantial gain in average r^2 retrieved by the multiple regression elastic net over single-marker univariate testing (**Figure 2**). At the modeled CpGs (i.e. 236,923 non-null elastic net models), SNP-based heritability derived from recently published estimates in whole-blood samples (van Dongen et al. 2016) averaged 9%. Thus, at these CpGs, the implemented performance of our models (6.9%) was close to the maximum variance in DNAm that can theoretically be explained by common SNPs. In addition, per-CpG cross-validation r^2 showed high correlation with reported SNP-based (r=0.53) and total heritability (r=0.62) estimates across all modeled CpGs. Among CpGs for which no elastic net model could be fitted (N=158,091 CpG sites), lower SNP-based heritability was observed, with an average of 4%.

To assess the validity of the inferred genetic estimators we examined their accuracy to predict DNAm in an independent sample comprising whole-blood methylomic profiles from *N*=319 healthy young adults (BASEL2 sample, see Methods). The correlation of model performance between training and testing samples across all modeled CpGs was high (r=0.96)(**Supplementary Figure 1**). Moreover, the average performance (i.e. proportion of variance of the DNAm signal explained by the genetic models) of the testing sample (r^2: 7.6%) was very close to the corresponding performance of the training sample (r^2: 6.9%). These findings demonstrate high stability and generalization capability of the implemented genetic models. A set of 86,710 genetic models for DNAm estimation was identified as highly robust, showing significant (FDR<0.05) and consistent correlation with DNAm across the two independent BASEL1 and BASEL2 samples (see Methods)(example shown in **Figure 3**). These genetic estimators of DNAm are termed hereafter EstiMeth models.

Figure 2: Comparison of average fraction of variance of DNAm variance explained by penalized multiple or univariate regression

Mean r^2 corresponds to cross-validation performance of the elastic net model in the BASEL1 sample, averaged across the top -*n* CpGs (yellow), and r^2 for the top identified mQTL per CpG average across the top -*n* CpGs (blue).

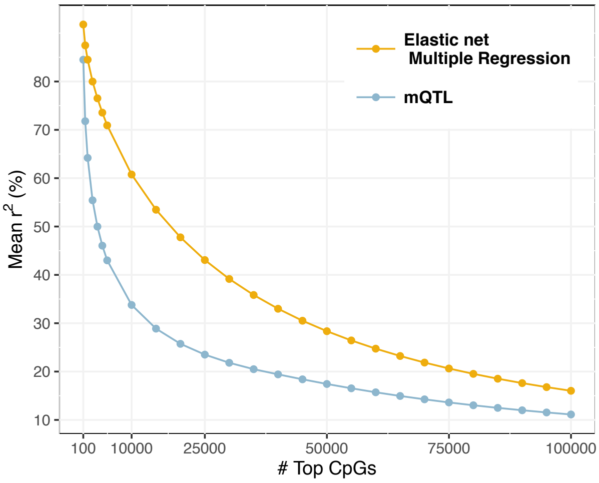
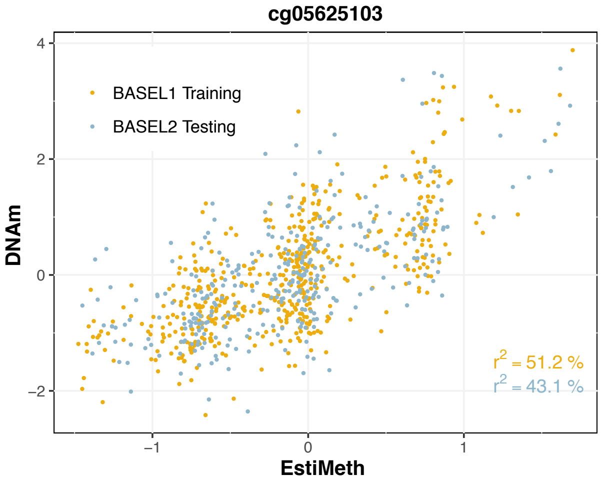


Figure 3: Example of a robust EstiMeth model

Horizontal axis represents the DNAm value estimated from the EstiMeth model at the CpG site. Vertical axis represents the observed DNAm value (adjusted for main confounders). r^2: fraction of variance of DNAm signal explained by the EstiMeth model (in %).



In real-life applications, not every SNP for a given EstiMeth might be available in the sample under study. On the other hand, EstiMeth SNPs in pair- or group-wise linkage disequilibrium might ensure robustness of the estimates also under incomplete SNP coverage. Therefore, we examined the performance of EstiMeth models after repeatedly discarding at random 10% of markers in the BASEL2 sample (see Methods). This resulted in an overall average distribution of r^2 that was very close to the original distribution (**Supplementary Figure 2**), indicating high stability of most of the models under incomplete SNP coverage. We provide EstiMeth models together with summary statistics of their performance under varying missing rates, thereby enabling the estimation of the stability of each individual model (http://mcn.unibas.ch/files/EstiMeth_Distribution_v1.zip , password: mcnEstiMeth140510).

Genetically driven DNAm is associated with gene expression of co-localizing genes

Each EstiMeth model corresponds to a CpG that is likely to be under strong genetic control. Given the role of DNAm in the regulation of gene expression (Deaton and Bird 2011) we investigated the relationship between EstiMeth CpGs and expression levels of neighboring genes. Expression levels at ~20K genes were obtained for N=408 individuals from the BASEL1 dataset (see Methods).

First, we performed genome-wide association testing between DNAm and expression levels of genes located within ±1Mbp of any CpG site (N=397,731 sites). We identified 26,925 significant associations (FDR<0.05), involving 6,160 genes and 17,867 CpGs. Among these CpGs, we observed significant over-representation of EstiMeth CpGs (78% of EstiMeth CpGs among 17,867 CpGs associated with expression; 22 % of EstiMeth CpGs across all investigated CpGs sites; Fisher's test p < 2.2e-16).

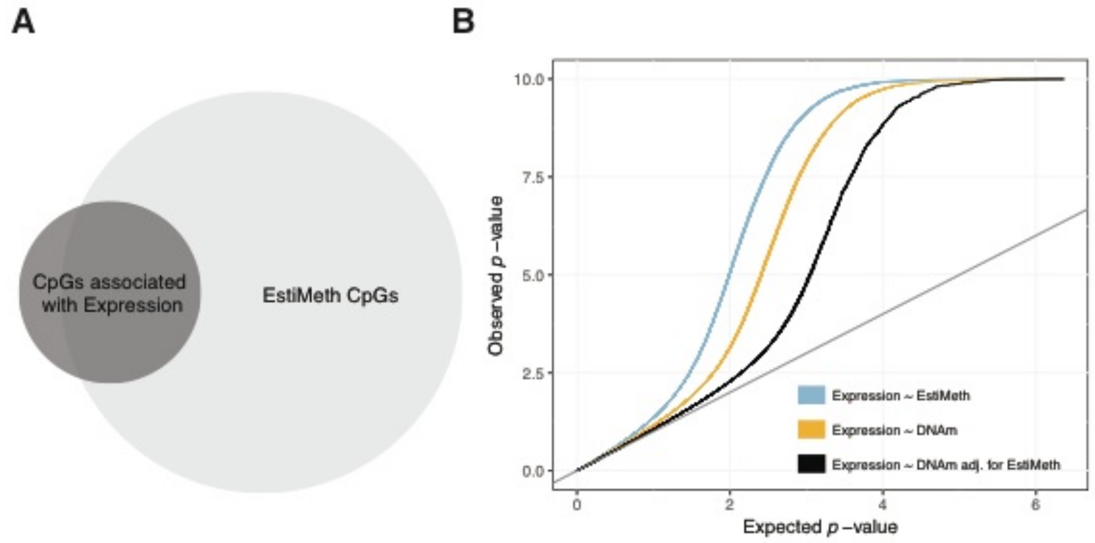
We also observed that EstiMeth CpGs, which were associated significantly with gene expression, were over-represented in shores (**Supplementary Figure 3**), which is in line with previous reports (van Eijk et al. 2012). These results indicate that genetically driven (i.e. EstiMeth) CpGs are more likely to correlate with expression of co-localizing genes. This observation might also reflect the existence of shared genetic contributions between EstiMeth CpGs and the expression of their co-localizing genes.

To test this hypothesis, we performed association testing between estimated DNAm values of each EstiMeth model and gene expression in -*cis* (±1Mbp). Given that gene expression was measured in the BASEL1 dataset, all EstiMeth models were re-implemented using the independent BASEL2 dataset as reference to prevent overfitting (see Methods). Subsequently, a total of 2 million EstiMeth-gene pairs were tested for association. We observed substantial deviation of genetic association signals from the null uniform distribution (**Figure 4-B**), with particular over-representation of large effect sizes. To further test whether EstiMeth models account for part of the shared variance observed between DNAm and expression, we also examined the DNAm-expression associations after regressing out the effect of EstiMeth estimated DNAm values. We observed a consistent and substantial decrease of detected association signals (**Figure 4-B**). Notably, within CpG-gene pairs identified as genome-wide significant (FDR<0.05), the average fraction of shared variance between DNAm and expression traits dropped from r^2= 9.6 % to 2.3% after adjustment for EstiMeth models effects. These results support the existence of shared genetic contribution between DNAm and gene expression in -*cis* , captured by the EstiMeth models.

Interestingly, the fraction of expression variance explained by the EstiMeth models was on average higher than the corresponding fraction explained by the DNAm signal alone (all EstiMeth CpG-gene pairs: r^2=0.6 % vs. r^2=0.4 %, Student t-test *p*-value <$2.2x10^{-16}$; genome-wide significant CpG-gene pairs: r^2=16.9 % vs. r^2=9.6% Student t-test p-value <$2.2x10^{-16}$) (**Figure 4-B**). This suggests that the EstiMeth models are also likely to include SNPs having DNAm-independent effects on gene expression in -*cis*. Of note, this increase in shared variance was mostly observed for CpGs located nearby the transcription start site of their associated gene (**Supplementary Figure 4**), a genomic location more likely to harbor *cis*-eQTLs (Wagner et al. 2014).

Figure 4: Relationship between DNAm, gene expression and EstiMeth

A: Overlap between CpGs associated with gene expression (FDR<0.05) and EstiMeth CpGs. **B:** QQ-plots for association between gene expression vs. EstiMeth estimates (blue), vs. DNAm (yellow) and vs. DNAm after adjustment for EstiMeth models (black). *p*-values < 1x10^{-10} not shown



Genetic correlation testing based on GWAS summary statistics

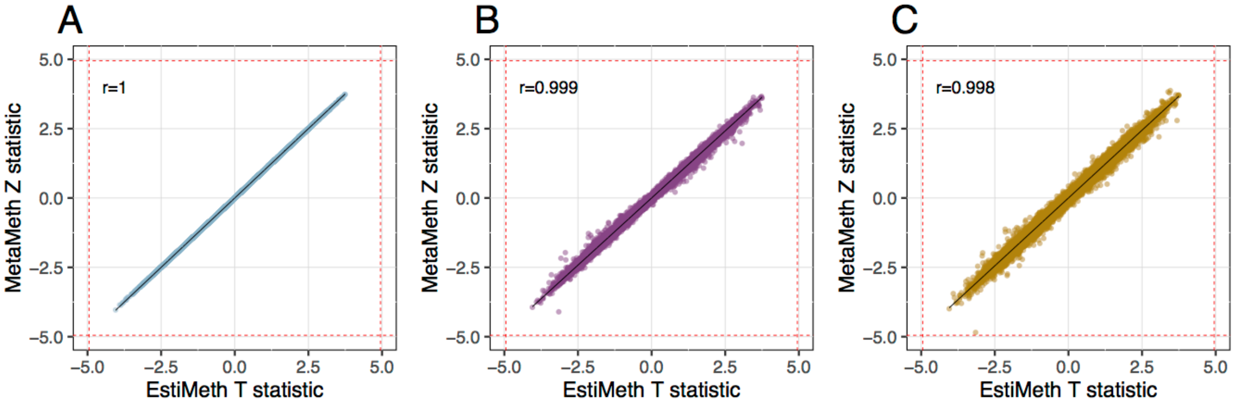
Provided availability of individual genotypic data in a given study sample, EstiMeth values can be readily obtained as the linear combination between the weights provided herein (derived from the BASEL1 sample) and the observed SNPs (**Figure 1-B**).

Yet, genotypic data from large-scale genome-wide association studies are often not directly accessible. Recently, methods have been proposed, that allow imputation of association statistics between genetic estimates of gene expression and a given trait, based solely on GWAS summary statistics (Gusev et al. 2016)(Barbeira et al. 2016). Based on this body of work, we extended the EstiMeth models to a 'MetaMeth' approach that allows genetic correlation testing from GWAS summary statistics (**Figure 1-C**)(see Methods). The *t*-value for association between the EstiMeth estimated DNAm values and the trait can be approximated using simultaneously: *(1)* EstiMeth SNPs' weights, *(2)* the standardized GWAS summary statistics (i.e. results from SNP to phenotype association), *(3)* the covariance structure of the SNPs included in the EstiMeth model (see Methods)(**Figure 1-C**). The implementation relies on the covariance structure from a reference population, provided it is genetically congruent to the population under study (**Figure 1-C**). This latter assumption represents a critical issue of the MetaMeth implementation, as slight shifts between the reference and actual population structures may potentially induce biased estimates. In order to mimic such discrepancies, we systematically assessed the validity of the MetaMeth approach on the BASEL2 sample, while using the SNP covariance structure inferred from the genetically close, yet not identical, Hapmap CEU population. Of note, all EstiMeth models were re-trained on the BASEL1 sample, restricted to SNPs present in both Hapmap and BASEL1 datasets, yielding a total of 81,807 retained models (see Methods).

Firstly, we examined the convergence of the EstiMeth (i.e., genotype-based) and MetaMeth (i.e., summary statistics-based) genetic correlation approaches by considering height as the complex trait under study. Under quasi-interchangeability of the two approaches, the MetaMeth Z-statistic should be close to the *T*-value obtained by testing directly association between height and the corresponding EstiMeth estimated DNAm values. Using the SNP covariance structure from the BASEL2 sample, i.e. the actual population structure, the correlation between genetic correlation statistics was close to 1 (**Figure 5**). We next used the SNP covariance structure derived from two independent population panels (i.e. BASEL1 sample and Hapmap CEU sample). The correlation between statistics obtained from the two approaches remained greater than 0.997 (**Figure 5**), supporting the validity of the MetaMeth approach.

Figure 5: Comparison of EstiMeth and MetaMeth association statistics with height in the BASEL2 sample

Each dot corresponds to an individual CpG included in EstiMeth models. Horizontal axis represents the *T* statistic obtained from the correlation between sex-adjusted height and genotypes based EstiMeth estimate. The vertical axis represents the MetaMeth *Z* statistic based on the SNPs covariance structure observed within the BASEL2 sample (**A**), from external independent BASEL1 sample (**B**), from external independent reference HapMap CEU sample (**C**). Red dashed lines represent critical statistics at Bonferroni adjusted significance threshold (p < 0.05/81807).



In a second stage, we estimated the Type I error rate of the MetaMeth approach. We performed a genome-wide MetaMeth scan on 1000 phenotypes generated from a normal distribution. The distribution of the minimum *p*-value obtained per run yielded a 5% quantile equal to 1.1e-06 (**Supplementary Figure 5**), which is above a Bonferroni adjusted significance threshold for a given genome-wide scan (*p* = 6.1e-7). This indicates that under realistic settings the proposed MetaMeth yields conservative association statistics.

We next compared the power of the MetaMeth and EstiMeth methods. At each CpG we repeatedly generated a trait that was associated with EstiMeth estimated DNAm values at large effect sizes (i.e. r^2=7.3% yielding an association detectable at 50% power under Bonferroni adjustment for multiple testing). Over all CpGs, the MetaMeth achieved lower average power (42.8%) as compared to the EstiMeth method (**Supplementary Figure 6**). Yet, we observed that for 17.5% of CpGs, the power of MetaMeth exceeded the power reached with the EstiMeth approach. These results indicate that in case of genuine association between EstiMeth and a trait, provided large effect sizes, the MetaMeth approach can lead to biased estimates of *T*-statistics, resulting in globally reduced power of detection.

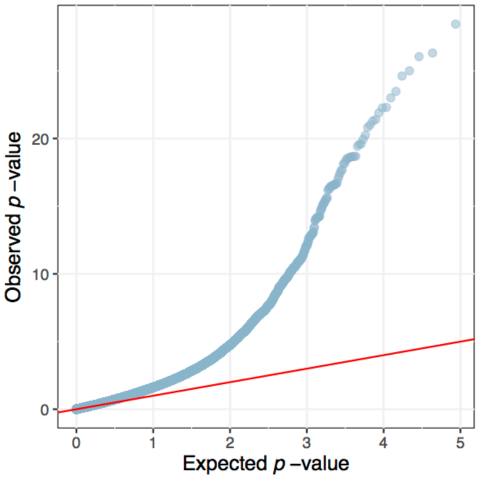
MetaMeth application to large-scale GWAS of schizophrenia

We applied MetaMeth on summary statistics obtained from the recently published schizophrenia PGC Consortium large-scale mega-analysis of GWAS-results (Ripke et al. 2014). MetaMeth statistics were derived considering the 86,710 EstiMeth models implemented on the BASEL1 SNPs panel and the corresponding BASEL1 covariance structure (see Methods).

We observed a highly significant deviation of MetaMeth statistics from the null uniform distribution (**Figure 6**). In particular, we identified a total of 469 associations withstanding genome-wide Bonferroni adjustment (unadjusted *p*-value < 5.7e-07). For the majority of these hits (n=412, 87.8 %), the corresponding EstiMeth model included at least one marker exhibiting a GWAS association *p*-value that would have reached a genome-wide GWAS Bonferroni adjustment significance threshold (*p* <5e-08)(**Supplementary Table 1**). The majority of the identified CpGs (460 out of 469) lie within ± 1Mbp of 47 regions out of the 105 reported as independent autosomal genomic loci associated with schizophrenia (average genomic loci size: 202 kbp), or are found within the extended MHC region (**Supplementary Table 1**).

Thus, these results demonstrate that MetaMeth identified, among the large number of significant susceptibility variants for such polygenic disorders as schizophrenia, the ones that impact on disease risk probably through regulation of site-specific DNAm.

Figure 6: MetaMeth analysis of large-scale GWAS for schizophrenia



DISCUSSION

We generated genetic estimators of epigenetic regulation - EstiMeth - that leverage genetic contributions to DNAm in whole-blood to identify epigenetic underpinnings of complex traits. EstiMeth models together with MetaMeth and Hapmap reference structure programs are made publicly available (http://mcn.unibas.ch/files/EstiMeth_Distribution_v1.zip , password: mcnEstiMeth140510). By capitalizing on multiple co-localized genomic loci likely to impact on the DNAm signal, we identified a set of genetic estimators accounting on average for a modest, yet highly consistent fraction of variance in DNAm across independent samples.

Inter-individual variation in DNAm correlates with variation in expression levels of co-localizing genes possibly through shared genetic factors (Bell et al. 2011; Gutierrez-Arcelus et al. 2013). Here, integration of both methylomic profiles and gene expression data revealed that EstiMeth models accounted for a substantial fraction of the shared phenotypic variance between both molecular traits. This suggests that, in line with a recent report (Shakhbazov et al. 2016), local genetic variations represent an essential factor underlying the observable inter-individual relationship between gene expression and DNAm at adjacent CpG sites. Importantly, the identified associations do not always imply direct causality between genetically driven DNAm and gene expression, as the EstiMeth models possibly include SNPs exerting independent effects on each trait (Bell et al. 2011; Gutierrez-Arcelus et al. 2013).

We also combined the genetic estimators for DNAm with recently proposed methods, that allow applicability of these estimators to SNP summary statistics solely (i.e. in the absence of individual genotypic data)(Gusev et al. 2016; Barbeira et al. 2016). This approach, applied to recent large-scale GWAS-results for schizophrenia (Ripke et al. 2014), resulted in the identification of 469 significant associations. This suggests the existence of shared genetic contributions between whole-blood DNAm and schizophrenia risk which is consistent with recent reports (Hannon et al. 2016a; Gaunt et al. 2016). Of note, it cannot be excluded that the identified associations can also be partly driven by genetic loci exerting independent effects on each trait. A majority of the identified associations implicated genome-wide significant GWAS hits, whilst encompassing slightly less than half of the 105 genomic regions associated with schizophrenia (Ripke et al. 2014). Importantly, each association suggests shared genetic contributions between schizophrenia risk and DNAm variation at a specific CpG site. We also note that MetaMeth identified significant association signals at CpGs mapping to *AS3MT* - arsenite methyltransferase (**Supplementary Table 1**). Schizophrenia genetic risk variants have been recently shown associated with expression of an *AS3MT* isoform (*AS3MT*^{d2d3}) and DNA methylation variation at this locus (Li et al. 2016). These results highlight the potential of MetaMeth to decipher, from large-scale GWAS results, trait-associated loci that are putatively mediating their effect through methylation at given CpG sites, and to prioritize specific genomic loci for downstream functional validation.

On the side of limitations, it should be stressed that the EstiMeth models were inferred and tested on moderately-sized whole-blood samples. For about one third of the investigated CpGs, characterized though by lower average SNP-based heritability, no elastic net model could be fitted from our training sample. Thus, additional local genetic contributions to DNAm might be detected with increasing sample sizes.

The fraction of variance explained by the EstiMeth models refers to DNAm signal after adjustment for main confounders. As for any -omics dataset, such confounders are usually unknown and estimated empirically in a study-specific manner. This might ultimately impact on the fraction of variance in DNAm that can be retrieved by the derived genetic estimators. For instance, in our study, we observed that these estimators showed, on average, higher performance on the testing sample as compared to the training sample. Thus, although high stability was globally observed between performance of the models across both samples, inference on multiple independent data sets, and multiple tissues, is warranted to fully appreciate their generalizability.

We also note that we generated a genetic estimator of DNAm at each single CpG site. Although this provides a straightforward way of annotating the models, it also results in certain redundancy of the estimators for highly correlated CpGs sites. This in turn leads to a number of inferred estimators, which, unlike genetic estimators for gene expression, allow only for a moderate reduction of multiple testing burden in GWAS (Gamazon et al. 2015; Gusev et al. 2016). In addition, the derived genetic estimators were built on *-cis* neighboring SNPs only. Although we globally observed high consistency of the inferred models' performance with published SNP-based heritability estimates, the performance was on average lower than the reported common SNP heritability. This gap might be explained by additional trans genetic components likely to contribute to inter-individual variability in DNAm (Lemire et al. 2015; Gaunt et al. 2016; Bonder et al. 2017). Improved accuracy might thus be achieved by extending the modeling to trans genetic components, as was recently shown for gene expression (Vervier and Michaelson 2016).

Concerning the MetaMeth extension, we could derive empirical settings that showed appropriate control of Type I error in the investigated sample, yet at the cost of decreased average power of detecting genuine associations. Robustness of the derived statistic is also tightly linked to the genetic discrepancy between the reference and study population, which might not be easily evaluated in practice. This calls for assessment of the stability of the approach on larger independent samples from varying populations.

In conclusion, we provide genetic estimators for DNAm in whole-blood, that can effectively complement genetic estimators for gene expression, to gain insight into the molecular underpinnings of complex traits.

METHODS

Study datasets

Whole-blood methylomic profiles and genotypic data were obtained from healthy young adults recruited in the course of two separate studies conducted in Basel, previously described (Milnik et al. 2016). The Basel Imaging dataset (BASEL1) included *N*=533 participants (age range: 18-37 years old; 222 males), the independent Basel Cognitive dataset (BASEL2) included a total of *N*=319 participants (age range: 18-37 years old; 97 males). The study protocols were approved by the ethics committee of the cantons of Basel-Stadt and Basel-Landschaft. All participants gave written informed consent after complete description of the study protocols. Subjects were free of any neurological or psychiatric condition and did not take medication at the time of the experiment.

Methylomic profiling

A detailed description of methylomic profiling protocols can be found in Milnik et al. (Milnik et al. 2016). Briefly, methylomic profiling was performed using the Illumina HumanMethylation450 array. Samples of non-European ancestry were identified using Hapmap references population genotypes and excluded from analysis (n=35 in BASEL1 sample, yielding *N*=533 remaining for analysis; none identified in BASEL2 sample). Beta-values were calculated from SWAN normalized intensities (Maksimovic et al. 2012). Subsequently, beta-values were *M*-transformed and adjusted for processing plate effect (z-transformation), age, sex and the main sources of technical variations inferred from principal components analysis (Milnik et al. 2016). Beta-values with detection *p*-value > 0.05 were considered as missing. Individual CpGs sites were excluded based on the following criteria: non-CpG context, non-autosomal probes, probes with a SNP mapping to the target CpG site or with three or more SNPs within the 50mer probe (maf >0.01)(based on RnBeads package annotation), multi-mapping or polymorphic CpGs (maf >0.01 in European population) reported in (Price et al. 2013; Chen et al. 2013), and probes with missing rate ≥ 5% in final samples. Prior to analysis, missing values were imputed using the R package impute (https://bioconductor.org/packages/release/bioc/html/impute.html) with *k*=10.

Genotyping

DNA was isolated from saliva sample and genotyped using the Affymetrix Genome-Wide Human SNP array 6.0 following the manufacturer's protocol. Genotype imputation was performed independently for each BASEL1 and BASEL2 sample, on the University of Michigan Imputation Server (settings for markers imputation: maf >0.01, call rate >95%). In the BASEL1 sample, approximately 5 Million imputed SNPs with minor allele frequency >0.05, Hardy-Weinberg equilibrium (HWE) *p*-value > 0.0001 and imputation score R^2>0.8 were retained for training genetic models of DNAm estimation. To allow complete evaluation of the trained models, all selected markers were considered in the BASEL2 sample.

EstiMeth models implementation

A total of 395,014 CpGs, measured in both BASEL1 and BASEL2 samples and surrounded by more than one *cis*-SNPs within ± 1Mbp were considered for analysis. At each of these individual CpG site, a elastic net (Zou and Hastie 2005) genetic additive model was fitted between all surrounding imputed *cis*-SNPs, and adjusted DNA methylation signal.

Genotypes were coded as *0:* homozygous for the major allele, *1:* heterozygous, and *2:* homozygous for the minor allele. Models were implemented using the glmnet R package with α elastic net constraint fixed to 0.5. Default standardization of genotypes (mean centering and unit variance) was applied within the training procedure which resulted in slight improvement of models performance (**Supplementary Table 1**). Beta coefficients were returned on the original genotype scale. The λ tuning parameter was determined using a 10-fold cross-validation scheme. This modeling allowed simultaneous shrinkage of individual Beta coefficients and selection of variables, thus drastically reducing the number of SNPs finally included in each model (average n=3,522 SNPs before selection, average n=26 SNPs after selection across all non-null models).

Model performance was assessed using Pearson's squared correlation r^2 between the model estimate - linear combination between elastic net inferred Beta coefficients and observed genotypes - and the actual adjusted DNAm signal; for the training data set, r^2 refers to cross-validation performance.

We derived a set of robust genotypes-based estimators for DNAm - i.e. EstiMeth models - as follows: (1) all models exhibiting a significant association between the elastic net derived cross-validation estimator and the actual values in the BASEL1 training sample (FDR<0.05 across all non-null models); (2) among those, all models exhibiting a significant association between the elastic net derived estimator and the actual values in the BASEL2 testing sample (FDR<0.05); (3) all models resulting in a positive correlation between actual and genotypes based estimated value in the testing sample. This yielded a total of 86,710 models likely reflecting a robust genetically driven DNAm signal at the corresponding CpG (minimum observed r^2 in training sample =0.94%; minimum r^2 in testing sample = 1.37%).

MetaMeth implementation

***Statistical model**:* We relied on the approach recently proposed by Barbeira et al. (Barbeira et al. 2016) for estimating genetic correlation between EstiMeth model and a trait based on GWAS summary statistics solely. Specifically, consider a given EstiMeth model comprising weights W at p SNPs. Let T_g denotes the t-value between the EstiMeth linear combination and the trait. Let $\sum_{\mathrm{p}}$ be the observed covariance matrix of the p SNPs, and Z the vector of standardized coefficients obtained from testing each SNP for association with the trait (GWAS summary statistics, *β/se(β)*).

As described by Barbeira et al. (Barbeira et al. 2016), T_g is equivalent to Z_g :

$$T_g = Z_g = \sum_{k=1}^{p} w_k \frac{\hat{\sigma}_k}{\widehat{\sigma_g}} \frac{\widehat{\beta_k}}{se(\widehat{\beta_k})} \sqrt{\frac{1-R_k^2}{1-R_g^2}} \quad \textbf{(Equation 1)}$$

with

$\widehat{\sigma_g} = \sqrt{W'\Sigma_p W}$

$\hat{\sigma}_k$ the standard deviation at SNP k

R_k^2 the proportion of phenotypic variance explained by SNP k

R_g^2 the proportion of phenotypic variance explained by EstiMeth estimator

In absence of genotypes, the R_g^2 term cannot be estimated and the covariance structure Σ_p has to be estimated from a reference population, which leads to the approximation:

$$T_g \approx MetaMeth\, Z_g = \sum_{k=1}^{p} w_k \frac{\hat{\sigma}_k}{\widehat{\sigma_{gref}}} \frac{\widehat{\beta_k}}{se(\widehat{\beta_k})} \quad \textbf{(Equation 2)}$$

with $\widehat{\sigma_{gref}} = \sqrt{W'\,\Sigma_{pref}\,W}$

This approximation has two potential caveats.

Firstly, as pointed by Barbeira et al, removal of the R^2 ratio can lead to remarkable underestimation of T_g for SNPs with large effect sizes. This deviation was notably observed when comparing EstiMeth and MetaMeth approaches on DNAm signal, which implicate large effect sizes. Considering the actual sample's covariance matrix, the exact statistic derived from **Equation 1** is equal to T_g (**Supplementary Figure 7-A**). Removal of the R^2 ratio in **Equation 2**, while still using the exact sample's covariance matrix, leads to deviation from the original T_g with a global decrease of derived statistics (**Supplementary Figure 7-B**). The same observation was drawn from the power study presented in **Supplementary Figure 6**.

Secondly, the divergence between the reference population and actual covariance structures can lead to biased estimates. Using Equation 2, we observed inflation of genome-wide level Type I error (**Supplementary Figure 5**). To account for this uncertainty, we penalized the denominator $\widehat{\sigma_g}$ by multiplying the diagonal of the Σ_{pref} matrix ($\Sigma_{pref} = \Sigma_{pref} + \lambda_s$ diag(Σ_{pref}) with $\lambda_s = 0.1$) (Gusev et al. 2016). We found empirically $\lambda_s = 0.1$ to achieve conservative results, at the cost of decreased power. Unless otherwise specified, all reported results were obtained using $\lambda_s = 0.1$.

Hapmap reference panel: The SNPs covariance matrices $\sum_{\mathrm{pref}}$ were inferred from the publicly available Hapmap reference genotypes panel (Phase II-III, 2010-08) of CEU unrelated individuals). Individual markers were filtered on the following criteria: genotype missing rate > 0.05, HWE p-value < 0.0001, MAF < 0.01 excluded). Selected SNPs were mapped from hg18 to hg19 annotation using the UCSC lift over tool. Finally, a total of 1,046,075 markers overlapping with the BASEL1 sample imputed markers remained for analysis.

Implementation of EstiMeth models on Hapmap SNPs: elastic net models were re-trained on the BASEL1 sample, using the restricted Hapmap SNP panel. Overall, we observed comparable performance of EstiMeth models between the full panel of 5M imputed SNPs and the restricted Hapmap SNPs panel (correlation between cross-validation r^2 across all EstiMeth CpGs > 0.99; correlation between testing r^2 > 0.96); a small proportion of models (5.6%) showed performance not reaching the minimum r^2 initially observed in the training and testing samples and were thus excluded from MetaMeth benchmarking analyses (n=81,807 models remaining).

Simulation studies: Type I error rate was assessed on 1000 repeats of genome-wide MetaMeth scan, using phenotypes randomly generated from a normal distribution. The power study was performed by generating, for each CpG, a phenotype showing an average r= 0.27 with the EstiMeth estimate. This corresponds to an effect size of 7.3%, detectable with 50% power considering the BASEL2 sample size N=319 and genome-wide significance threshold α = 0.05/81,807. This procedure was repeated 300 times per CpG.

Transcriptomic analyses

Data processing: Blood samples were collected using PAXgene Blood RNA Tubes (PreAnalytix Qiagen/BD, Switzerland). Expression profiles were obtained for *N*=408 individuals of the BASEL1 sample using with the Affymetrix GeneChip Human Transcriptome Array 2.0 (see Supplementary text), providing quantification of expression levels for ~67K transcript clusters (referred as genes). Individual expression values were adjusted for age and sex using linear regression. Expression signals were adjusted for unknown technical confounders while preserving local genuine genetic effects. This was achieved by examining the number of identified *cis*-eQTL while further adjusting expression values for increasing number of principal components (Pickrell et al. 2010; Liang et al. 2013); this procedure was repeated until no increase in the number of identified eQTLs was observed anymore, leading to 23 components retained for final adjustment of expression values. Only genes annotated to RefSeq identifiers were considered for analysis. The annotation was based on manufacturer's information (GPL17586-45144) curated using the UCSC database version oct.2015. This yielded a total of 21,186 autosomal genes entering subsequent analyses.

Association testing between DNAm and expression traits: The relationship between DNAm and expression traits was examined considering for each gene all CpGs located within ± 1Mbp from gene boundaries (N=397,731 sites out of 397,947 CpGs from BASEL1 sample). Statistical association testing was performed using Pearson's correlation test. Genome-wide significant associations were identified using Benjamini-Hochberg FDR correction. Expression signals were optimally processed for preserving genetic effects (see paragraph *Data processing*). In order to check whether this procedure possibly biased the over-representation of EstiMeth genetically driven CpGs among identified associations, the genome-wide CpG-gene association scan was re-conducted on expression data adjusted for main confounders only (batch effect adjustment using ComBat method implemented in the sva R package (Leek et al. 2012) age, sex, and the seven first principal components axes, that showed strong association with blood cell subtypes composition). This analysis led to the identification of 11,760 significant associations (FDR <0.05), implicating 8,530 CpGs, among which 74 % involved EstiMeth CpGs, convergent with results obtained from the primary analysis.

Genetic association analysis of DNA methylation and gene expression: For ensuring independence of the expression trait and the EstiMeth models, all models were re-trained on the BASEL2 sample using the same methodology as for the initial EstiMeth implementation. Out of 86,710 models, a total of 83,337 non-null models could be inferred in the BASEL2 sample and entered subsequent analyses. In turn, estimated DNAm values were obtained in the BASEL1 sample using these EstiMeth models implemented on the BASEL2 sample. These estimated values were subsequently tested for association with expression trait at their co-localizing gene (s). DNAm-expression associations were also examined under adjustment for EstiMeth models: DNAm was adjusted for EstiMeth estimated values using linear regression, and next tested for association with gene expression.

MetaMeth application to PGC data

GWAS summary statistics of PGC schizophrenia analysis (52 samples; 34,241 cases, 45,604 controls and 1,235 parent-affected offspring trios) were downloaded from https://www.med.unc.edu/pgc/files/resultfiles/scz2.snp.results.txt.gz.

MetaMeth association statistics were obtained for the more exhaustive EstiMeth models inferred from the 5M SNPs panel, considering the BASEL1 covariance structure as reference.

DATA ACCESS: The genetic estimators are made publically available (http://mcn.unibas.ch/files/EstiMeth_Distribution_v1.zip, password: mcnEstiMeth140510).

ACKNOWLEDGMENTS: This work was funded by the University of Basel and the Swiss National Science Foundation (individual grant 320030_159740 to D.Q., individual grant 320030_163434 to A.P.). Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

DISCLOSURE DECLARATION: All authors reported no potential conflicts of interests.

SUPPLEMENTARY INFORMATION FOR

Genetic estimators of DNA methylation provide insights into the molecular basis of polygenic traits

Authors: Virginie Freytag[1,2,*], Vanja Vukojevic[1,2,3], Annette Milnik[1,2,4], Christian Vogler[1,2,4], Dominique J.-F. de Quervain[2,4,5,+], Andreas Papassotiropoulos[1,2,3,4,*,+]

Affiliations:

[1]Division of Molecular Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland

[2]Transfaculty Research Platform Molecular and Cognitive Neurosciences, University of Basel, CH-4055 Basel, Switzerland

[3]Department Biozentrum, Life Sciences Training Facility, University of Basel, CH-4056 Basel, Switzerland

[4]Psychiatric University Clinics, University of Basel, CH-4055 Basel, Switzerland

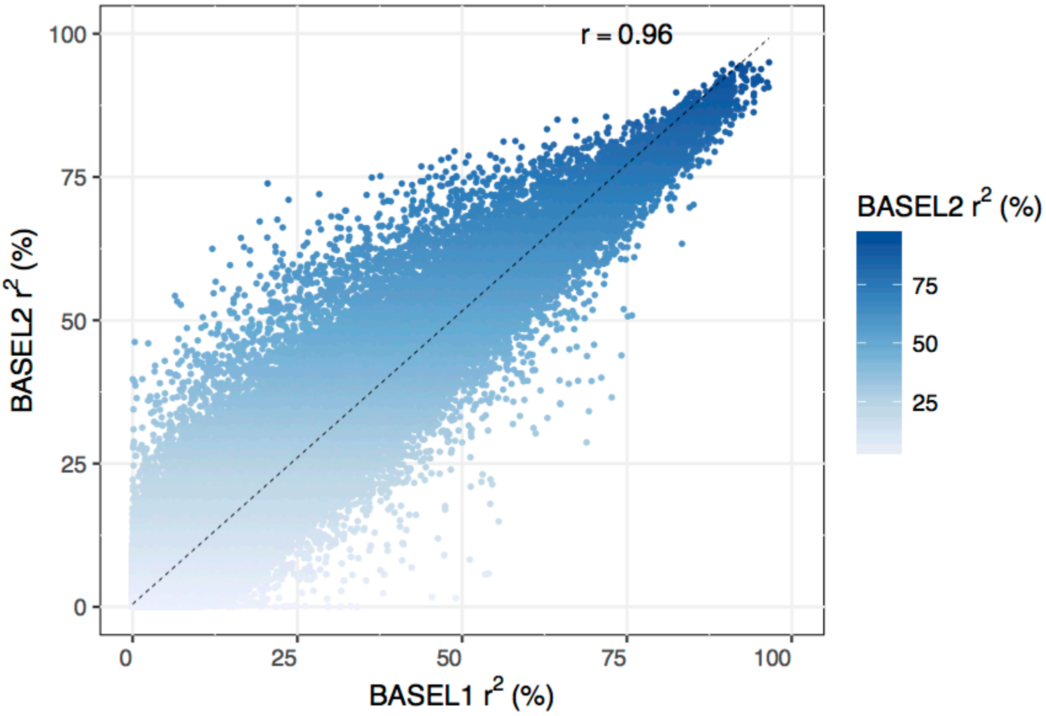
[5]Division of Cognitive Neuroscience, Department of Psychology, University of Basel, CH-4055 Basel, Switzerland

TABLE OF CONTENTS

SUPPLEMENTARY FIGURES

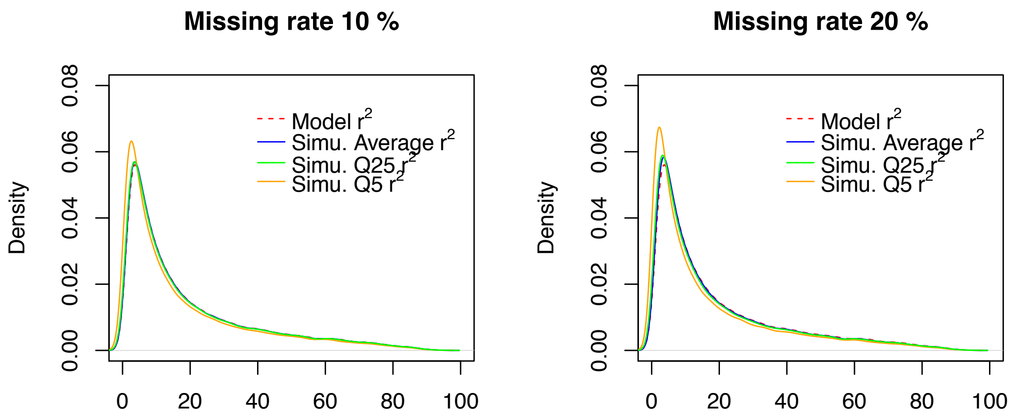
Supplementary Figure 1: Elastic net models testing vs. training performance

Horizontal axis denotes training cross-validation r^2 performance in the BASEL1 sample. Vertical axis represents performance of the models in the independent testing BASEL2 sample. Dashed line represents regression line.



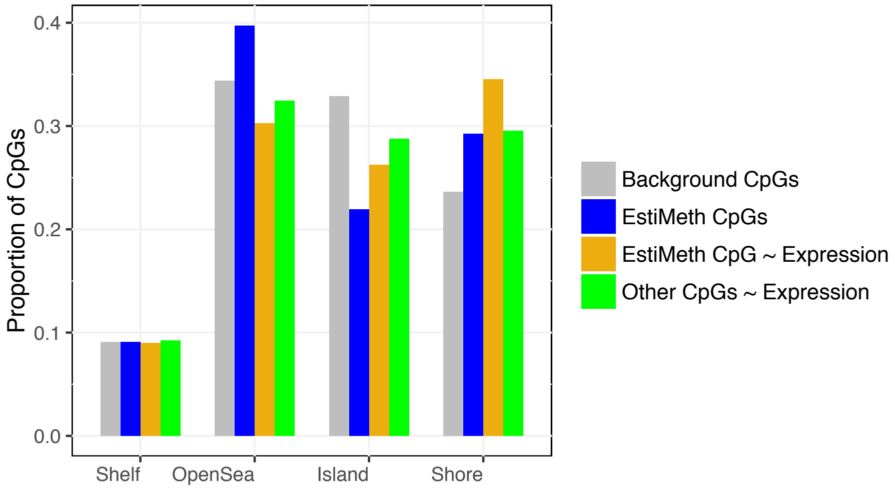
Supplementary Figure 2: Distribution of EstiMeth models' testing r^2 for varying genotyping missing rates

EstiMeth models were evaluated on the BASEL2 testing sample with 10% (left panel) or 20% (right panel) of genotypes randomly discarded. For each model, the mean, and 5th and 25th quantiles of the r^2 distribution obtained from 1000 runs was recorded. Graphs represent the density distribution of all EstiMeth models r^2 (in %) for the complete models (red dashed line) or from simulations (blue, green and orange lines).



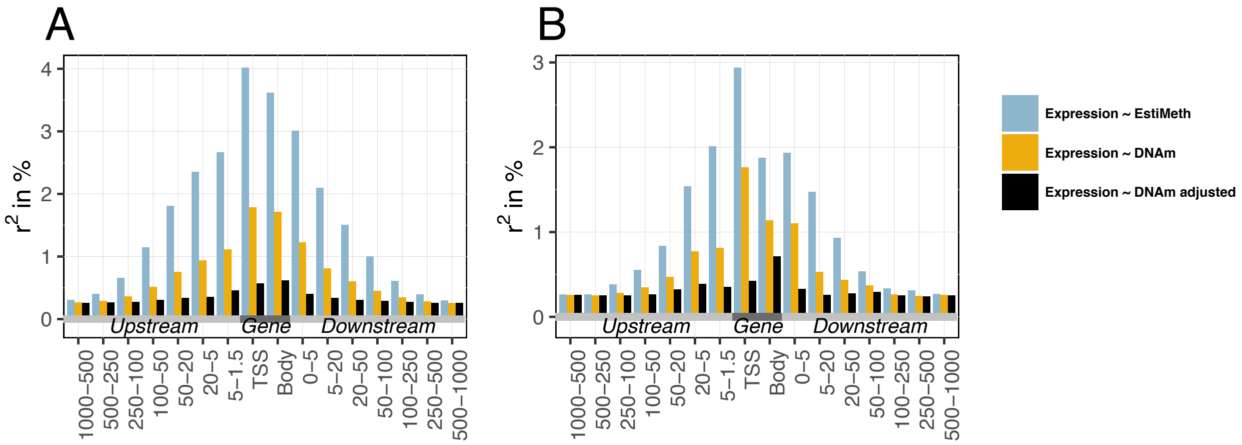
Supplementary Figure 3: Distribution of EstiMeth CpGs associated with gene expression across genomic context

Grey: Background CpGs (n = 397,947). Orange: EstiMeth CpGs associated with gene expression (n = 13,894). Blue: EstiMeth CpGs (n = 86,710). Green: CpGs associated with gene expression not included in EstiMeth CpGs (n = 3,973).



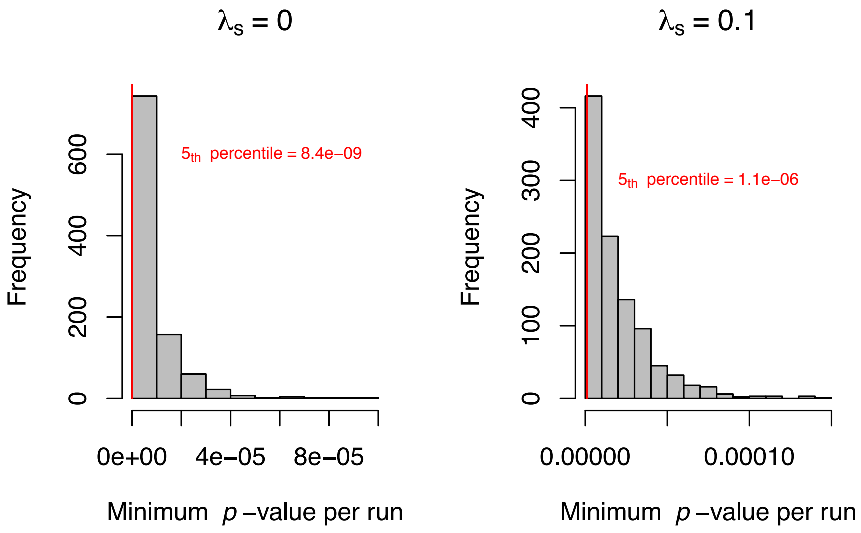
Supplementary Figure 4: Comparison of shared variance between gene expression, DNAm and EstiMeth genetic contributions across genomic locations

Horizontal axis: genomic location ; bins correspond to the distance of a given CpG relative to its associated gene in kbp. TSS is defined as 1.5kb upstream gene start. Vertical axis: fraction of shared variance (in %) between gene expression and EstiMeth (blue), DNAm (orange) or DNAm adjusted for EstiMeth effects (black). **A**: Average across ~2M EstiMeth CpG-gene association pairs . **B**: Average across EstiMeth CpG-gene association pairs identified as genome-wide significant (FDR<0.05).



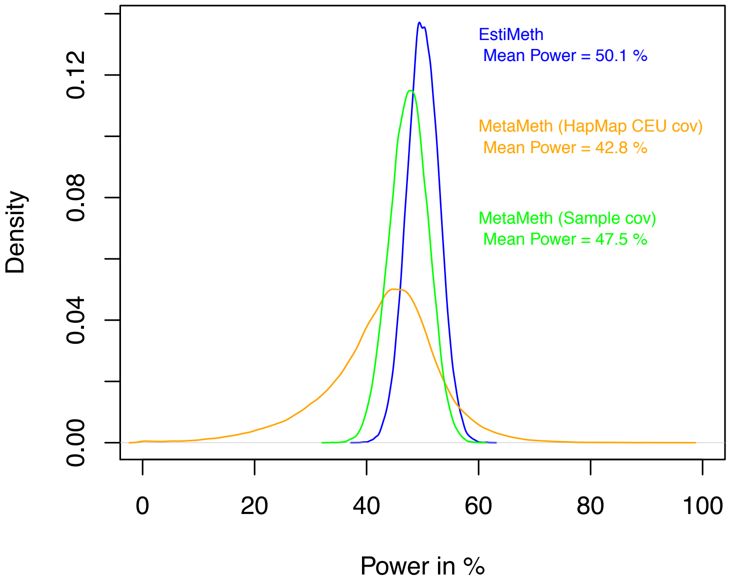
Supplementary Figure 5: Distribution of the minimum MetaMeth *p*-value per genome-wide scan under *H0*

Phenotypes were drawn from a standard normal distribution (1000 runs). For each run, a MetaMeth analysis was performed across all modeled CpGs (n = 81,807) and the random phenotype in the BASEL2 sample, using the covariance structures from HapMap CEU population. The minimum *p*-value obtained across all CpGs was retained. The left panel represents the distribution of minimum *p*-values, obtained without penalization of the *Z* statistics (Equation 2, main text); right panel represents the distribution obtained using the penalty factor retained in the MetaMeth implementation.



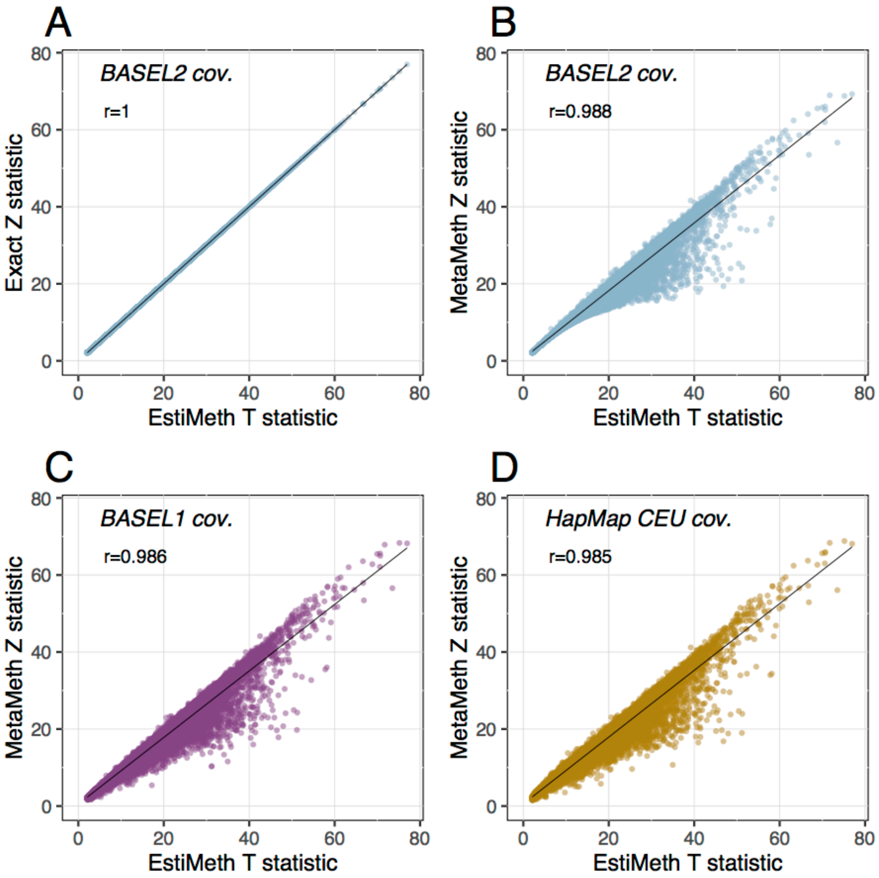
Supplementary Figure 6: Comparison of EstiMeth and MetaMeth power in the BASEL2 sample

For each EstiMeth CpG, phenotypes were generated to be associated with EstiMeth estimate with 50% power of being detected (at α = *p* <0.05/81807). The graph represents the density curves of power achieved across all CpGs, for EstiMeth (blue), MetaMeth using HapMap CEU covariance structure (orange), and MetaMeth using actual sample's covariance structure (green).



Supplementary Figure 7: Comparison of EstiMeth and MetaMeth association statistics for DNAm in the BASEL2 sample

Each dot corresponds to an individual CpG included in EstiMeth models. Horizontal axis represents the *T*-value obtained from the correlation between DNAm signal and EstiMeth estimate. In panel (**A**) the vertical axis represents *Z* statistic retrieved from Equation 1, based on the sample's SNPs covariance structure, which is equivalent to the *T*-value; in panel (**B**), the vertical axis represents the approximation MetaMeth Z statistic (Equation 2), based on the sample's covariance structure. In panels (**C**) and (**D**), the vertical axis represents the MetaMeth *Z* statistics based on SNPs covariance structure inferred from external samples BASEL1 and HapMap-CEU respectively. Black line represents regression line.



SUPPLEMENTARY TABLES

Supplementary Table 1: Comparison of elastic net performance with/without standardization of the genotypes

	Genotypes standardization	**No Genotypes standardization**
Number of Non null models	236,923	236,602
r^2 Training in % (M ± SD)	6.9 ± 14.2	6.8 ±14.1
r^2 Testing in % (M ± SD)	7.6 ± 15	7.4 ± 14.9
Selected SNPs (M ± SD)	25.6 ± 26.9	27.6 ± 30

All non-null models were considered. M: mean; SD: standard deviation.

SUPPLEMENTARY TEXT

Affymetrix HTA 2.0 array transcriptome analysis. Total RNA was further isolated with the PAXgene Blood miRNA Kit (PreAnalytix, Switzerland). Following, a second, additional purification was performed with the miRNeasy Micro Kit (Qiagen, Germany). The concentration and quality of the RNA was determined using Nanodrop 2000 (ThermoScientific, USA) and RNA Nano 6000 Kit on Bioanalyzer 2100 instrument (Agilent, USA). Next, GLOBINclear™-Human Kit (Ambion, USA) was used for a non-enzymatic depletion of the alpha and beta globin mRNA starting from 1µg of total RNA preparations derived from whole blood, following a standard procedure. The concentration and quality of the "globin-free" RNA was assessed as described above. Following, the alpha and beta globin mRNA depletion was measured by qPCR. In brief: for reverse transcription, 350ng of total RNA was denaturized for 8 min at 70°C followed by ice incubation in the presence of 25ng Anchored Oligo(dT)20 Primer (Invitrogen, USA) and 75ng Random Decamers Primers (Ambion, USA). In the RT reaction, cDNA was generated in 25µl reaction using Super RT kit (HT Biotechnology, Santa Cruz, CA USA). Upon completion of the reaction, the volume was adjusted to 200µl in Lambda DNA solution (5ng/µl final concentration; Promega, Fitchburg, WI USA). The primers were designed against splice variants that contain alpha-Globin gene: alpha-Globin Forward: 5'- GCACGCGCACAAGCT-3', and alpha-Globin Reverse: 5'-GGGTCACCAGCAGGCA-3' (Microsynth, Switzerland). The expression levels were normalized to RPLPO gene (human large ribosomal protein) using the following primers: RPLP0-Ex3-4_FW, 5'-CTCTGGAGAAACTGCTGC-3' and RPLP0-Ex3-4_RV, 5'-CTGATCTCAGTGAGGTCC-3' (Sigma Aldrich, USA). qPCR was performed using the Power SYBR Green PCR Master Mix (Life Technologies, USA) according to standard recommendations, in 12µl final volume of reaction, using 2µl of cDNA template, on RotorGene 6000A instrument (Corbett Research Pty Ltd, Sydney Australia). Cycling conditions were as follows: 95°C, 60s – 40x (95°C, 3s - 56°C, 10s – 72°C, 4s) followed by a melting curve analysis (61°C to 95°C, rising by 0.7°C / 3s) to attest amplification specificity. Threshold cycles (crossing point) were determined using Rotor-Gene software version 6.1 (Corbett Research, Australia). RPLPO was selected as reference gene for normalization after we tested several candidate-reference genes, as had been previously described [1]. Expression levels were normalized using a geometric mean level of expression [1]. Fold differences were calculated using the delta-delta Ct method [2] with the help of qBasePlus software (Biogazelle, Ghent, Belgium).

Target synthesis was performed using Ambion® WT Expression Kit (Ambion, Life Technologies, USA) starting from 250ng of high-quality "globin-free" RNA, following the standard procedure. Next, 5.16μg of target cDNA was further labeled and prepared for hybridization with the GeneChip® WT Terminal Labeling and Hybridization Kit (Affymetrix, USA). The prepared samples were loaded on Affymetrix GeneChip Human Transcriptome Array 2.0 (Cat# 902162) and hybridized for 16 hours (45°C, 60rpm) in Hybridization oven 640 (Affymetrix, USA). The arrays were washed and stained on Fluidics Stations 450 (Affymetrix) by using the Hybridization Wash and Stain Kit (Affymetrix, USA) under FS450_0001 protocol. The GeneChips were processed with an Affymetrix GeneChip Scanner 3000 7G (Affymetrix, USA). DAT images and CEL files of the microarrays were generated using Affymetrix GeneChip Command Control software (Affymetrix, USA). In order to account for technical inter-array variation we performed a full quantile-normalization; feature quantification was conducted using a median-polish on transcript-level according to the HTA 2.0. lib-set-version 0.3. (Affymetrix Power Tools version: 1.16.0). Cross-platform validation of genotyping and expression data was assessed using the MixUpMaper algorithm [3].

References

1. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3,** RESEARCH0034 (2002).

2. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29,** e45 (2001).

3. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27,** 2104–11 (2011).

REFERENCES

Barbeira A, Shah KP, Torres JM, Wheeler HE, Torstenson ES, Edwards T, Garcia T, Bell GI, Nicolae D, Cox NJ, et al. 2016. MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results. bioRxiv doi: 10.1101/045260.

Bell JT, Pai A a, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* **12**: R10.

Bonder MJ, Luijk R, Zhernakova D V, Moed M, Deelen P, Vermaat M, van Iterson M, van Dijk F, van Galen M, Bot J, et al. 2017. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* **49**: 131–138.

Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**: 203–9.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–1022.

Editorial. 2011. Best is yet to come. *Nature* **470**: 140.

Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, GTEx Consortium, Nicolae DL, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**: 1091–1098.

Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, Zheng J, Duggirala A, McArdle WL, Ho K, et al. 2016. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* **17**: 61.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright F, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**: 245–252.

Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2013**: e00523.

Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, St Clair D, Mustard C, Breen G, Therman S, et al. 2016a. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol* **17**: 176.

Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, Troakes C, Turecki G, O'Donovan MC, Schalkwyk LC, et al. 2016b. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* **19**: 48–54.

Jaffe AE, Gao Y, Deep-Soboslay A, Tao R, Hyde TM, Weinberger DR, Kleinman JE. 2016. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* **19**: 40–47.

Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* **470**: 187–197.

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**: 882–3.

Lemire M, Zaidi SHE, Ban M, Ge B, Aïssi D, Germain M, Kassam I, Wang M, Zanke BW, Gagnon F, et al. 2015. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun* **6**: 6326.

Lev Maor G, Yearim A, Ast G. 2015. The alternative role of DNA methylation in splicing regulation. *Trends Genet* **31**: 274–280.

Li M, Jaffe AE, Straub RE, Tao R, Shin JH, Wang Y, Chen Q, Li C, Jia Y, Ohi K, et al. 2016. A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nat Med* **22**: 649–656.

Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, Cookson WOC. 2013. A cross-platform analysis of 14 ,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res* **23**: 716–726.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–585.

Maksimovic J, Gordon L, Oshlack A. 2012. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**: R44.

Milnik A, Vogler C, Demougin P, Egli T, Freytag V, Hartmann F, Heck A, Peter F, Spalek K, Stetak A, et al. 2016. Common epigenetic variation in a European population of mentally healthy young adults. *J Psychiatr Res* **83**: 260–268.

Papassotiropoulos A, de Quervain DJF. 2015. Failed drug discovery in psychiatry: Time for human genome-guided solutions. *Trends Cogn Sci* **19**: 183–187.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–72.

Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, Robinson WP, Kobor MS. 2013. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**: 4.

Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, et al. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**: 421–427.

Roussos P, Mitchell AC, Voloudakis G, Fullard JF, Pothula VM, Tsang J, Stahl EA, Georgakopoulos A, Ruderfer DM, Charney A, et al. 2014. A Role for Noncoding Variation in Schizophrenia. *Cell Rep* **9**: 1417–1429.

Schübeler D. 2015. Function and information content of DNA methylation. *Nature* **517**: 321–326.

Shakhbazov K, Powell JE, Hemani G, Henders AK, Martin NG, Visscher PM, Montgomery GW, McRae AF. 2016. Shared genetic control of expression and methylation in peripheral blood. *BMC Genomics* **17**: 278.

Van Dongen J, Nivard MG, Willemsen G, Hottenga J-J, Helmer Q, Dolan C, Ehli E, Davies GE, van Iterson M, Breeze CE, et al. 2016. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun* **7**: 11115.

Van Eijk KR, De Jong S, Boks M, Langeveld T, Colas F, Veldink JH, de Kovel C, 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.

Vervier K, Michaelson JJ. 2016. SLINGER: large-scale learning for predicting gene expression. *Sci Rep* **6**: 39360.

Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol* **15**: R37.

Zou H, Hastie T. 2005. Regularization and Variable Selection via the Elastic Net. *J R Stat Soc B* **67**: 301–320.

Discussion

A large number of common genetic variations of small effect are likely to contribute to the observed phenotypic variability in complex cognitive traits. Disentangling interpretable molecular patterns from this highly polygenic architecture requires investigating the molecular systems in which common genetic variations exert their effects. The work presented in this thesis relied on two different integrative strategies to leverage epigenetic variations, as intermediate molecular traits of genomic action.

In the first study, we adopted a system-level approach that aimed at investigating the relationship of global age-related molecular epigenetic patterns and physiological variation in cortical thickness. Decomposition of whole-blood methylomic profiles in (N=533) healthy young adults allowed identification of a global age-related epigenetic signature associated with cortical thickness and episodic memory. Subsequent genetic analysis of this methylomic pattern further showed association of genetic contributions to episodic memory in an independent sample of (N=3346) healthy individuals. Finally, functional annotation and genetic study of the methylomic signature both converged in pointing to the possible involvement in immune system and function genes.

In the second study, we directly modeled the relationship between DNA methylation at a given CpG site and its local common genetic contributions. As recently proposed for studying gene expression, we applied a multiple penalized regression method to leverage the joint effect of multiple genetic variants that are likely to contribute to variability in DNA methylation signal. Thus, using the same dataset as in the first study, we could derive genetic estimators that accounted for a consistent fraction of DNAm signal across two independent samples. This approach applied to recent large-scale GWAS-results for schizophrenia, a common genetically complex disorder, led to the identification of significant associations between genetically driven whole-blood DNAm and disease risk.

The two studies presented in this thesis underscore the integration of high-throughput intermediate molecular profiles with genotypic data as an effective and essential approach to capitalize on information about molecular features for the investigation of complex cognitive traits. The analysis of -omics data in healthy young adults is indeed challenging in different aspects. Firstly, statistical limitations such as over-fitting are inherent to the multidimensional nature of these datasets. Secondly, the quantification of individual molecular signals, such as DNA methylation, is prone to technical variation (Milnik et al., 2016) and biological noise. In bulk tissue samples, such as whole-blood, inter-individual differences in cell composition notably represent a potent factor impacting on the signal (Jaffe & Irizarry, 2014). Such confounders are usually unknown and statistical methods aiming at reducing their impact have to rely on empirical evaluation, in a study-specific manner. From a system-level standpoint, the integration of genotypic data as anchor of molecular variation appears instrumental to address whether the identified molecular patterns might represent relevant features of the dataset. The results from the second study indicate that focusing on the strong local genetic components of DNAm allows the derivation of stable molecular patterns putatively associated with complex traits. This suggests that future analytical strategies that explicitly incorporate genetic information into the modeling of global molecular systems (Civelek & Lusis, 2014; Sieberts & Schadt, 2007), might further enhance the identification of relevant molecular patterns associated with phenotypic variability in complex cognitive traits. One draw-back of system-level approaches is the broadness of the identified molecular signatures. In this context focusing on the genetically driven part of these molecular networks might also help narrowing down the identified molecular patterns.

Importantly, whatever the level of complexity of the inferred models is, replication of the identified associations in independent samples must be undertaken, as it is a necessary step of genetic research (Kraft, Zeggini, & Ioannidis, 2009).

In our second study, we observed limited overlap between the derived genetic estimators and gene expression levels, consistent with the reported limited overlap between methylation QTL and expression QTL, at least in bulk tissues such as whole-blood (Gaunt et al., 2016) or brain (Gibbs et al., 2010). Recently, common genetic variants underlying complex trait variability have been shown to act on multiple components of gene regulation, detectable at the level of transcript abundance, alternative splicing, epigenetic modifications or protein levels (Li et al., 2016). Hence, these different layers of molecular data represent complementary sources of information for the functional annotation of genetic variations associated with complex traits. In this context, systematic population-based assessment of multi-layer data is key for a more comprehensive investigation of the molecular systems impacted by genetic variations.

In the present work, we relied on peripherally measured epigenetic markers measured in whole-blood samples. The relevance of peripheral profiles for the study of neuropsychiatric related traits is still an open question. The results presented herein support the potential of investigating easily accessible peripheral molecular markers for the study of brain-related traits. There is notably growing interest in the link between the immune and central nervous systems and its possible role in neuro-psychiatric disease. At the phenotypic level, immune system peripheral markers have for instance been correlated with schizophrenia status (Miller, Buckley, Seabolt, Mellor, & Kirkpatrick, 2011) and Alzheimer's disease (Swardfager et al., 2010). From a genetic standpoint, enrichment of schizophrenia associated variants has been described in immune related pathways (O'Dushlaine et al., 2015) and immune markers associated variants (Astle et al., 2016). Recently, complex genetic variations underlying risk for schizophrenia in complement 4 (*C4*) genes, located in the major histocompatibility complex, and critical for immune function, have also been linked to disease risk and the level of expression in the brain (Sekar et al., 2016). A recent study conducted by our group identified association between genetic variants in *TROVE2,* a gene implicated in autoimmunity, and aversive memory performance in healthy subjects and additionally with traumatic memory and risk for posttraumatic stress disorder in genocide survivors (Heck et al., 2017). At the moment we cannot draw a conclusion about the mechanistic link between human brain and immune system functions. However, the convergence of these findings suggests that identification of peripheral molecular patterns associated with brain-related traits might serve as a first step towards further functional investigation of this relationship.

In sum, the work presented in this thesis suggests that systems genomics analyses of peripheral molecular markers represent a valuable approach to expand understanding of the molecular underpinnings of complex brain-related traits.

References

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212.

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science*, *322*(5903), 881–888.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., … Soranzo, N. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, *167*(5), 1415–1429.e19.

Barbeira, A., Shah, K. P., Torres, J. M., Wheeler, H. E., Torstenson, E. S., Edwards, T., … Im, H. K. (2016). MetaXcan: Summary statistics based gene-level association method infers accurate PrediXcan results. *bioRxiv*, doi: https://doi.org/10.1101/045260.

Bell, J. T., Tsai, P. C., Yang, T. P., Pidsley, R., Nisbet, J., Glass, D., … Deloukas, P. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genetics*, *8*(4), e1002629.

Berger, S. L., Kouzarides, T., Shiekhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, *23*(7), 781–783.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., … Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, *98*(4), 288–295.

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., … Radvanyi, F. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Reports*, *9*(4), 1235–1245.

Bonder, M. J., Luijk, R., Zhernakova, D. V, Moed, M., Deelen, P., Vermaat, M., … Heijmans, B. T. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, *49*(1), 131–138.

Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, *15*(1), 34–48.

Danion, J. M., Huron, C., Vidailhet, P., & Berna, F. (2007). Functional mechanisms of episodic memory impairment in schizophrenia. *Canadian Journal of Psychiatry*, *52*(11), 693–701.

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, *25*(10), 1010–1022.

Farrell, P. M. (2008). The prevalence of cystic fibrosis in the European Union. *Journal of Cystic Fibrosis*, *7*(5), 450–453.

Fjell, A. M., Grydeland, H., Krogsrud, S. K., Amlien, I., Rohani, D. A., Ferschmann, L., … Walhovd, K. B. (2015). Development and aging of cortical thickness correspond to genetic organization patterns. *Proceedings of the National Academy of Sciences*, *112*(50), 15462–7.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., … Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10604–9.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R. J., … Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098.

Garagnani, P., Bacalini, M. G., Pirazzini, C., Gori, D., Giuliani, C., Mari, D., … Franceschi, C. (2012). Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell*, *11*(6), 1132–1134.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., … Pedersen, N. L. (2006). Role of Genes and Environments for Explaining Alzheimer Disease. *Archives of General Psychiatry*, *63*(2), 168-174.

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., … Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*, 61.

Gejman, P. V., Sanders, A. R., & Duan, J. (2010). The role of genetics in the etiology of schizophrenia. *Psychiatric Clinics of North America*, *33*(1), 35–66.

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S. L., … Singleton, A. B. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in Human Brain. *PLoS Genetics*, *6*(5), e1000952.

Glahn, D. C., Thompson, P. M., & Blangero, J. (2007). Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, *28*(6), 488–501.

Gottesman, I. I., & Gould, T. D. (2003). The Endophenotype Concept in Psychiatry : Etymology and Strategic Intentions. *American Journal of Psychiatry*, *160*(4), 636–645.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*, *48*(3), 245–252.

Hannon, E., Lunnon, K., Schalkwyk, L., & Mill, J. (2015). Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, *10*(11), 1024–1032.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., … Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, *49*(2), 359–67.

Heck, A., Fastenrath, M., Ackermann, S., Auschra, B., Bickel, H., Coynel, D., … Papassotiropoulos, A. (2014). Converging genetic and functional brain imaging evidence links neuronal excitability to working memory, psychiatric disease, and brain activity. *Neuron*, *81*(5), 1203–13.

Heck, A., Fastenrath, M., Coynel, D., Auschra, B., Bickel, H., Freytag, V., … Papassotiropoulos, A. (2015). Genetic analysis of association between calcium signaling and hippocampal activation, memory performance in the young and old, and risk for sporadic Alzheimer disease. *JAMA Psychiatry*, *72*(10), 1029–36.

Heck, A., Milnik, A., Vukojevic, V., Petrovska, J., Egli, T., Singer, J., … Papassotiropoulos, A. (2017). Exome sequencing of healthy phenotypic extremes links TROVE2 to emotional memory and PTSD. *Nature Human Behaviour*, *1*(March), 1–10.

Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S., … Lumey, L. H. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences*, *105*(44), 17046–17049.

Hibar, D. P., Adams, H. H., Jahanshad, N., Chauhan, G., Stein, J. L., Hofer, E., … Ikram, M. A. (2017). Novel genetic loci associated with hippocampal volume. *Nature Communications*, *8*, 13624.

Hibar, D. P., Stein, J. L., Rentería, M. E., Arias Vasquez, A., Desrivières, S., Jahanshad, N., … Medland, S. E. (2015). Common genetic variants influence human subcortical brain structures. *Nature*, *520*(7546), 224–9.

Hindorff, L. A., Sethupathy, P., Junkins, H. A, Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–7.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol*, *14* (10), R115.

Hyman, S. E. (2013). Psychiatric drug development: diagnosing a crisis. *Cerebrum*, *2013*:5.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, *426*(6968), 789–796.

Jaffe, A. E., & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, *15* (2), R31.

Kanai, R., & Rees, G. (2011) The structural basis of inter-individual differences in human behavior and cognition. *Nat Rev Neurosci,* 12(4):231-42.

Kaminsky, Z. A., Tang, T., Wang, S.-C., Ptak, C., Oh, G. H., Wong, A. H., … Petronis, A. (2009). DNA methylation profiles in monozygotic and dizygotic twins. *Nature Genetics*, *41*(2), 240–245.

Kilpinen, H., & Dermitzakis, E.T. (2012). Genetic and epigenetic contribution to complex traits. *Hum Mol Genet*, *21*(R1), R24–8.

Knowles, M. R., & Drumm, M. (2012). The influence of genetics on cystic fibrosis phenotypes. *Cold Spring Harbor Perspectives in Medicine*, *2*(12), a009548.

Kraft, P., Zeggini, E., & Ioannidis, J. P. A(2009). Replication in genome-wide association studies. *Stat Sci*, *24*(4), 561–573.

Kremen, W.S., Jacobsen, K.C., Xian, H., Eisen, S.A., Eaves, L.J., Tsuang, M.T., & Lyons, M.J. (2007). Genetics of verbal working memory processes: a twin study of middle-aged men. *Neuropsychology*, *21*(5), 569–80.

Kremen, W.S., Panizzon, M.S., Franz, C.E., Spoon, K.M., Vuoksimaa, E., Jacobson, K.C., ... Lyons, M. (2014). Genetic complexity of episodic memory: A twin approach to studies of aging. *Psychol Aging, 29*(2), 404–417.

Lee, T., Mosing, M. A., Henry, J. D., Trollor, J. N., Ames, D., Martin, N. G., … OATS Research Team (2012). Genetic influences on four measures of executive functions and their covariation with general cognitive ability: the Older Australian Twins Study. *Behav Genet*, *42*(4), 528–38.

Lemire, M., Zaidi, S. H., Ban, M., Ge, B., Aïssi, D., Germain, M., … Hudson, T. J. (2015). Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature Communications*, *6*, 6326.

Lerch, J. P., van der Kouwe, A. J. W., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K. L., … Sotiropoulos, S. N. (2017). Studying neuroanatomy using MRI. *Nature Neuroscience*, *20*(3), 314–326.

Li, Y. I., van der Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., & Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*(6285), 600–604.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, *18*(1), 51–60.

Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, *10*(8), 565–577.

Manolio, T. A, Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A, Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–53.

Marioni, R. E., Shah, S., McRae, A. F., Chen, B. H., Colicino, E., Harris, S. E., … Deary, I. J. (2015). DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, *16*, 25.

Miller, B. J., Buckley, P., Seabolt, W., Mellor, A., & Kirkpatrick, B. (2011). Meta-analysis of cytokine alterations in schizophrenia: clinical status and antipsychotic effects. *Biol Psychiatry, 70*(7), 663–671.

Milnik, A., Heck, A., Vogler, C., Heinze, H.-J., de Quervain, D. J.-F., & Papassotiropoulos, A. (2012). Association of KIBRA with episodic and working memory: a meta-analysis. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics, 159B*(8), 958–69.

Milnik, A., Vogler, C., Demougin, P., Egli, T., Freytag, V., Hartmann, F., … Vukojevic, V. (2016). Common epigenetic variation in a European population of mentally healthy young adults. *Journal of Psychiatric Research*, *83*, 260–268.

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *368*(1620), 20120362.

O'Dushlaine, C., Rossin, L., Lee, P. H., Duncan, L., Parikshak, N. N., Newhouse, S., … Breen, G. (2015). Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*, *18*(2), 199–209.

Panizzon, M. S., Fennema-Notestine, C., Eyler, L. T., Jernigan, T. L., Prom-Wormley, E., Neale, M., … Kremen, W. S. (2009). Distinct genetic influences on cortical surface area and cortical thickness. *Cerebral Cortex*, *19*(11), 2728–35.

Panizzon, M. S., Lyons, M. J., Jacobson, K. C., Franz, C.E., Grant, M.D., Eisen, S.A.,… Kremen, W.S. (2011). Genetic architecture of learning and delayed recall: a twin study of episodic memory. *Neuropsychology*, *25*(4), 488–498.

Papassotiropoulos, A., & de Quervain, D. J. F. (2011). Genetics of human episodic memory: Dealing with complexity. *Trends in Cognitive Sciences*, *15*(9), 381–387.

Papassotiropoulos, A., & de Quervain, D. J. F. (2015). Failed drug discovery in psychiatry: Time for human genome-guided solutions. *Trends in Cognitive Sciences*, *19*(4), 183–187.

Papassotiropoulos, A., Gerhards, C., Heck, A., Ackermann, S., Aerni, A., Schicktanz, N., … de Quervain, D. J.-F. (2013). Human genome-guided identification of memory-modulating drugs. *Proceedings of the National Academy of Sciences*, *110*(46), E4369–E4374.

Papassotiropoulos, A., Henke, K., Stefanova, E., Aerni, A., Müller, A., Demougin, P., … de Quervain, D. J. F. (2011). A genome-wide survey of human short-term memory. *Molecular Psychiatry*, *16*(2), 184–92.

Papassotiropoulos, A., Stephan, D. A, Huentelman, M. J., Hoerndli, F. J., Craig, D. W., Pearson, J. V, … de Quervain, D. J.-F. (2006). Common Kibra alleles are associated with human memory performance. *Science*, *314*(5798), 475–8.

Paul, D. S., Soranzo, N., & Beck, S. (2014). Functional interpretation of non-coding sequence variation: Concepts and challenges. *BioEssays*, *36*(2), 191–199.

Petrovska, J., Coynel, D., Fastenrath, M., Milnik, A., Auschra, B., Egli, T., … Heck, A. (2017). The NCAM1 gene set is linked to depressive symptoms and their brain structural correlates in healthy individuals. *Journal of Psychiatric Research*, *91*, 116–123.

Plomin, R., Haworth, C. M., & Davis, O. S. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics*, *10*(12), 872–888.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., & Davis, O. S. (2013). Common DNA markers can account for more than half of the genetic influence on cognitive abilities. *Psychological Science*, *24*(4), 562–568.

Price, A. L., Spencer, C. C., & Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1821), 20151684.

Rakic, P. (2009). Evolution of the neocortex: a perspective from developmental biology. *Nature Reviews. Neuroscience*, *10*(10), 724–35.

Richards, A.,L., Jones, L., Moskvina, V., Kirov, G., Gejman, P.V., Levinson, D.F., … O'Donovan, M. (2012). Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol Psy*, *17*(2), 193–201.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A.K., Akterin, S., … Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–59.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. A., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85–97.

Rotival, M., Zeller, T., Wild, P. S., Maouche, S., Szymczak, S., Schillert, A., … Blankenberg, S. (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genetics*, *7*(12), e1002367.

Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., … Sklar, P. (2014). A Role for Noncoding Variation in Schizophrenia. *Cell Reports*, *9*(4), 1417–1429.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, *461*(7261), 218–223.

Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, *517*(7534), 321–326.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., ... McCarroll, S.A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183.

Sieberts, S. K., & Schadt, E. E. (2007). Moving toward a system genetics view of disease. *Mammalian Genome*, *18*(6-7), 389–401.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science, 253*(5026), 1380–6.

Stein, J.,L. Medland, S.E., Vasquez, A.A., Hibar, D.P., Senstad, R.E., Winkler, A.M, … Thompson, P.M. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nature Genetics*, *44*(5), 552–61.

Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating Change. *Journal of Neuroscience*, *34*(25), 8488–8498.

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics, 187*(2), 367–383.

Swardfager, W., Lanctôt, K., Rothenburg, L., Wong, A., Cappell, J., & Herrmann, N. (2010). A meta-analysis of cytokines in Alzheimer's disease. *Biological Psychiatry*, *68*(10), 930–941.

Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R., & Caldas, C. (2007). Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Computational Biology*, *3*(8), e161

Teschendorff, A. E., West, J., & Beck, S. (2013). Age-associated epigenetic drift: Implications, and a case of epigenetic thrift? *Human Molecular Genetics*, *22*(R1), R7–R15.

Toga, A. W., & Thompson, P. M. (2005). Genetics of Brain Structure and Intelligence. *Annual Review of Neuroscience*, *28*, 1–23.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology*, *53*, 1–25.

Van der Sijde, M. R., Ng, A., & Fu, J. (2014). Systems genetics : From GWAS to disease pathways. *Biochim Biophys Acta, 1842*(10), 1903–1909.

Van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C.V., … Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, *7*, 11115.

Visscher, P.M., Hill, WG., & Wray, NR. (2008) Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics, 9*(4), 255-266.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24.

Vogler, C., Gschwind, L., Coynel, D., Freytag, V., Milnik, A., Egli, T., … Papassotiropoulos, A. (2014). Substantial SNP-based heritability estimates for working memory performance. *Translational Psychiatry*, *4*, e438.

Volk, H. E., McDermott, K. B., Roediger III, H. L., & Todd, R. D. (2006). Genetic influences on free and cued recall in long-term memory tasks. *Twin Research and Human Genetics*, *9*(5), 623–631.

Vuoksimaa, E., Panizzon, M. S., Chen, C. H., Fiecas, M., Eyler, L. T., Fennema-Notestine, C., … Kremen, W. S. (2015). The genetic association between neocortical volume and general cognitive ability is Driven by Global Surface Area Rather Than Thickness. *Cerebral Cortex*, *25*(8), 2127–2137.

Wallace, G. L., Lee, N. R., Prom-Wormley, E. C., Medland, S. E., Lenroot, R.K., Clasen, L.S., ... Giedd, J.N.(2010). A Bivariate Twin Study of Regional Brain Volumes and Verbal and Nonverbal Intellectual Skills During Childhood and Adolescence. *Behav Genet*, *40*(2), 125–134.

Wang, L., Jia, P., Wolfinger, R.D., Chen, X., & Zhao, Z. (2011). Gene set analysis of genome-wide association studies: methological issues and perspectives. *Genomics,* 98(1):1-8.

Weiss, J.N., Karma, A., MacLellan, W.R., Deng, M., Rau, C.D., Rees, C.M., … Lusis, A.J. (2012). "Good enough solutions" and the genetics of complex disease. *Circ Res*, *111*(4), 493–504.

Wexler, E. M., Rosen, E., Lu, D., Osborn, G. E., Martin, E., Raybould, H., & Geschwind, D. H. (2011). Genome-wide analysis of Wnt1-regulated transcriptional network implicates neurodegenerative pathways. *Sci Signal*, *4*(193), ra65.

Whiteford, H. A., Ferrari, A. J., Degenhardt, L., Feigin, V., & Vos, T. (2015). The global burden of mental, neurological and substance use disorders: An analysis from the global burden of disease study 2010. *PLoS One*, *10*(2), e0116820.

Winkler, A.M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P.T., … Glahn, D.C. (2010). Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetic studies. *Neuroimage*, *53*(3), 1135–1146.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–9.

Zaghlool, S. B., Al-Shafai, M., Al Muftah, W. A., Kumar, P., Falchi, M., & Suhre, K. (2015). Association of DNA methylation with age, gender, and smoking in an Arab population. *Clinical Epigenetics*, *7*, 6.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Statist. Soc. B*, *67*(2), 301–320.

Candidate's peer-reviewed publications

Coynel, D., Gschwind, L., Fastenrath, M., **Freytag, V**. Milnik, A., Spalek, K., Papassotiropoulos, A., de Quervain (2017) Picture free recall performance linked to the brain's structural connectome. *Brain and behavior* 23 May 2017.

Freytag, V.*, Probst, S.*, Hadziselimovic, N., Boglari, C., Hauser, Y., Peter, F., Gabor Fenyves B., Milnik, A., Demougin, P., Vukojevic, V, de Quervain, DJ., Papassotiropoulos, A., Stetak A. (2017) Genome-wide temporal expression profiling in *C.elegans* identifies a core gene set related to long-term memory*. J Neurosci*. Jun 7. pii: 3298-16.

*: these authors contributed equally to this work.

Freytag, V., Carrillo-Roa, T., Milnik, A., Sämann, PG., Vukojevic, V., Coynel, D., Demougin, P., Egli, T., Gschwind, L., Jessen, F., Loos, E., Maier, W., Riedel-Heller, SG., Scherer, M., Vogler, C., Wagner, M., Binder, EB., de Quervain, DJ., Papassotiropoulos, (2017) A. A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory. *Nat Commun* 26;8:15193. doi: 10.1038/ncomms15193.

Harrisberger, F., Spalek, K., Smieskova, R., Schmidt, A., Coynel, D., Milnik, A., Fastenrath, M., **Freytag, V.**, Gschwind, L., Walter, A., Vogel, T., Bendfeldt, K., de Quervain, DJ., Papassotiropoulos, A, Borgwardt S. (2014) The association of the BDNF Val66Met polymorphism and the hippocampal volumes in healthy humans: a joint meta-analysis of published and new data. *Neurosci Biobehav* 42:267-78. doi: 10.1016/j.neubiorev.2014.03.011

Heck, A., Fastenrath, M., Coynel, D., Auschra, B., Bickel, H., **Freytag V**, Gschwind, L., Hartmann, F., Jessen, F., Kaduszkiewicz, H., Maier, W., Milnik, A., Pentzek, M., Riedel-Heller, SG., Spalek, K., Vogler, C., Wagner, M., Weyerer, S., Wolfsgruber, S., de Quervain, DJ., Papassotiropoulos, A. (2015) Genetic Analysis of Association Between Calcium Signaling and Hippocampal Activation, Memory Performance in the Young and Old, and Risk for Sporadic Alzheimer Disease. *JAMA Psychiatry* 72(10):1029-36.

Heck A., Milnik, A., Vukojevic, V., Petrovska, J., Egli, T., Singer, J., Escobar, P., Sengstag, T., Coynel, D., **Freytag, V**., Fastenrath, M., Demougin, P., Loos, E., Hartmann, F., Schicktanz, N., Delarue Bizzini, B., Vogler, C., Kolassa, IT, Wilker, S., Elbert, T., Schwede, T., Beisel, C., Beerenwinkel, N., de Quervain, DJF, Papasssotiropoulos, A. Exome sequencing of healthy phenotypic extremes links TROVE2 to emotional memory and PTSD. *Nature Human Behaviour*, *1*(March), 1–10.

Luksys, G., Fastenrath, M., Coynel, D., **Freytag, V.**, Gschwind, L., Heck, A., Jessen, F., Maier, W., Milnik, A., Riedel-Heller, SG., Scherer, M., Spalek, K., Vogler, C., Wagner, M., Wolfsgruber, S., Papassotiropoulos, A., de Quervain DJ. (2015) Computational dissection of human episodic memory reveals mental process-specific genetic profiles. *Proc Natl Acad Sci* USA 112(35):E4939-48.

Milnik, A., Vogler, C., Demougin, P., Egli, T., **Freytag, V.**, Hartmann, F., Heck, A., Peter, F., Spalek, K., Stetak, A., de Quervain, DJ., Papassotiropoulos, A., Vukojevic, V. (2016) Common epigenetic variation in a European population of mentally healthy young adults. *J Psychiatr Res* 83:260-268. doi: 10.1016/j.jpsychires.2016.08.012.

Spalek, K., Coynel, D., **Freytag, V.**, Hartmann, F., Heck, A., Milnik, A., de Quervain, D., Papassotiropoulos, A. (2016) A common NTRK2 variant is associated with emotional arousal and brain white-matter integrity in healthy young subjects. *Transl Psychiatry*. 6:e758. doi: 10.1038/tp.2016.20.

Vogler, C., Gschwind, L., Coynel, D., **Freytag, V**., Milnik, A., Egli, T., Heck, A., de Quervain, D.J.-F., Papassotiropoulos, A., 2014. Substantial SNP-based heritability estimates for working memory performance. Transl. Psychiatry 4, e438. doi:10.1038/tp.2014.81

Declaration by candidate

I declare herewith that I have independently fulfilled the phD-thesis entitled "Systems genomics analysis of complex cognitive traits". The thesis consists of original research articles that have been written in collaboration with the co-authors enlisted. The articles have been published in peer-reviewed journals or submitted for publication. All references used were cited accordingly and only allowed resources were used.

Signature : ____________________

Date : ____________________