

RESEARCH

Open Access



# Identification of tissue-specific and common methylation quantitative trait loci in healthy individuals using MAGAR

Michael Scherer<sup>1,2,3,16</sup>, Gilles Gasparoni<sup>1</sup>, Souad Rahmouni<sup>4</sup>, Tatiana Shashkova<sup>5,6</sup>, Marion Arnoux<sup>1</sup>, Edouard Louis<sup>7</sup>, Arina Nostaeva<sup>8</sup>, Diana Avalos<sup>9,10,11</sup>, Emmanouil T. Dermitzakis<sup>9,10,11</sup>, Yurii S. Aulchenko<sup>5,8,12,13</sup>, Thomas Lengauer<sup>2</sup>, Paul A. Lyons<sup>14,15</sup>, Michel Georges<sup>4</sup> and Jörn Walter<sup>1\*</sup> 

## Abstract

**Background:** Understanding the influence of genetic variants on DNA methylation is fundamental for the interpretation of epigenomic data in the context of disease. There is a need for systematic approaches not only for determining methylation quantitative trait loci (methQTL), but also for discriminating general from cell type-specific effects.

**Results:** Here, we present a two-step computational framework *MAGAR* (<https://bioconductor.org/packages/MAGAR>), which fully supports the identification of methQTLs from matched genotyping and DNA methylation data, and additionally allows for illuminating cell type-specific methQTL effects. In a pilot analysis, we apply *MAGAR* on data in four tissues (ileum, rectum, T cells, B cells) from healthy individuals and demonstrate the discrimination of common from cell type-specific methQTLs. We experimentally validate both types of methQTLs in an independent data set comprising additional cell types and tissues. Finally, we validate selected methQTLs located in the *PON1*, *ZNF155*, and *NRG2* genes by ultra-deep local sequencing. In line with previous reports, we find cell type-specific methQTLs to be preferentially located in enhancer elements.

**Conclusions:** Our analysis demonstrates that a systematic analysis of methQTLs provides important new insights on the influences of genetic variants to cell type-specific epigenomic variation.

**Keywords:** Quantitative trait loci, DNA methylation, Tissue specificity, Computational biology

## Background

Epigenetic mechanisms, including histone modifications, small RNAs, and DNA methylation, regulate gene expression in a tissue- and cell-type-specific manner [1]. DNA methylation is a critical player in such epigenetic gene regulation that has been implicated in various biological processes including X-chromosomal inactivation [2], genomic imprinting [3], and allele-specific expression [4, 5]. DNA methylation has been shown to be highly cell

type-specific and can be used to reliably estimate the proportions of different cell types in cellular mixtures such as blood or tissue samples [6, 7]. The DNA methylation state of a defined subset of CpGs in the human genome can be measured reliably across many samples using the Illumina Infinium microarray technologies allowing to perform epigenome-wide association studies (EWAS).

DNA methylation can be affected by aging [8], sex, and a range of environmental exposures [9, 10]. Additionally, donor genotype has a strong influence on the global DNA methylation state (methylome), especially when a genetic alteration, such as a single nucleotide polymorphism (SNP), occurs at a CpG site. Since bisulfite-based

\*Correspondence: j.walter@mx.uni-saarland.de

<sup>1</sup> Department of Genetics/Epigenetics, Saarland University, Saarbrücken, Germany

Full list of author information is available at the end of the article



methods can generate unclear and uninterpretable data at annotated or predicted SNPs located at CpG dinucleotides, such positions are typically removed from the analysis of DNA methylation data [11].

However, additional genetic effects that are not located in the CpG site but in genetic variants distant to the analyzed CpG can influence its DNA methylation state. Such variants influencing DNA methylation states are referred to as methylation quantitative trait loci (methQTL). These associations can range from distances of a few bases to several megabases resulting in long-range interactions [12, 13]. The definition of proximal methQTLs varies from 500 kb to 2 mb distance between the CpG and the SNP [12–14]. MethQTLs co-localize with genetic variants associated with diseases and donor phenotypes (GWAS hits) including obstructive pulmonary disease [14], prostate cancer risk [15], osteoarthritis [16], immune-mediated disease [17], asthma [18], and smoking [19]. Furthermore, combining methQTLs with expression QTLs (eQTLs) enables the investigation of associations between DNA methylation and gene expression changes [20–22].

However, so far not much emphasis has been put into analyses to investigate if and how often methQTLs affect DNA methylation in a tissue- or cell-type-specific manner [23]. An earlier study used cultured cells including fibroblasts, T cells, and lymphoblastoid cell lines to determine largely tissue-independent methQTLs. The authors reported that the association of methQTLs with changes in gene expression was rather cell type-specific [24] in line with recently identified cell type-specific eQTLs [25]. Other studies analyzing primary human cells rather reported largely cell type-independent eQTLs [26]. One problem which may have contributed to the current mixed view on the distribution of methQTLs is that methQTLs are typically determined using statistical models and tools that have been developed for eQTL analysis (e.g., *Matrix-eQTL* [27], *fastQTL* [28], or *GEM* [29]). Without the consideration of the specific properties of DNA methylation data including the correlation of DNA methylation states of neighboring CpGs such approaches may lead to substantial biases in the calling and interpretation of methQTLs.

To alleviate this problem, we present “Methylation-Aware Genotype Association in R” (*MAGAR*)—a novel computational pipeline that performs methQTL analysis. *MAGAR* defines clusters of neighboring CpGs according to their shared behavior across samples to represent DNA methylation haplotypes and performs methQTL analysis for each of the correlation blocks independently. *MAGAR* has been implemented as an R-package and utilizes existing tools such as *fastQTL* [28], *RnBeads* [30, 31], and *PLINK* [32]. Using *MAGAR*, we investigated sorted

blood cell types (T cells, B cells) and composite bowel tissues (ileum, rectum) of healthy individuals. The identified methQTLs were analyzed for cell type-specific effects using colocalization analysis, which showed that we could discern tissue-specific from common methQTLs. Finally, we validated and reproduced our findings in additional samples and in data from two published methQTL studies.

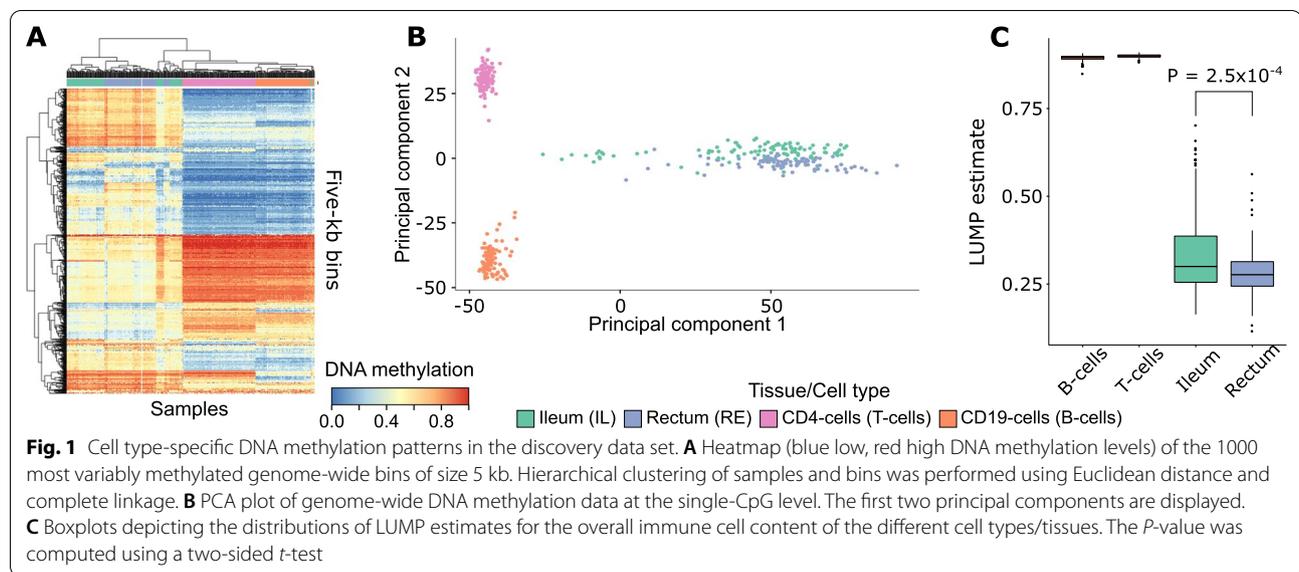
## Results

### Strong cell type-specific DNA methylation signals identified in bowel biopsies and purified blood cell types

The data set that we used for the discovery of methQTLs comprised 409 samples from ileum (IL,  $n=98$ ) and rectum (RE,  $n=95$ ) tissue biopsies and the two FACS-sorted blood cell types CD4-positive T cells ( $n=119$ ) and CD19-positive B cells ( $n=97$ ). For 29 individuals DNA methylation data were available for all four tissues/cell types within this discovery data set (Additional file 1: Figure S1). Average DNA methylation levels across all CpGs in genome-wide 5-kb bins revealed a strong cell type-specific signal that discriminates the blood cell types from the biopsies. Overall, the tissue biopsies exhibited an enhanced variation in comparison to the purified blood cell types indicating that increased cell-type heterogeneity goes along with a higher variation of DNA methylation patterns both on genome-wide bins and on the single-CpG level (Fig. 1A, B). To better understand the origins of cellular heterogeneity within the biopsy samples, we estimated the overall immune cell content of a sample using the LUMP algorithm [33] (Fig. 1C). While LUMP estimates were uniformly close to one for the two blood cell types as expected, they substantially varied across the biopsy samples. In line with previous reports [34], significantly higher immune cell content was observed in ileal compared to rectal samples.

### *MAGAR* facilitates the analysis of genome-wide methQTL effects

Understanding the relationship between DNA methylation and genetic variants can help to illuminate the association of genetic alterations with diseases and changes in gene expression. Thus, we are interested in defining statistically significant associations between DNA methylation and genotyping data. We call genetic variants that are associated with DNA methylation methQTLs. To alleviate the methQTL identification process, we developed the new R-based framework Methylation-Aware Genotype Association in R (*MAGAR*) that provides a comprehensive suite of tools enabling methQTL analysis leveraging the correlation of DNA methylation states of neighboring CpGs (Fig. 2A). Notably, *MAGAR* is the first package that performs data processing of raw (i.e.,

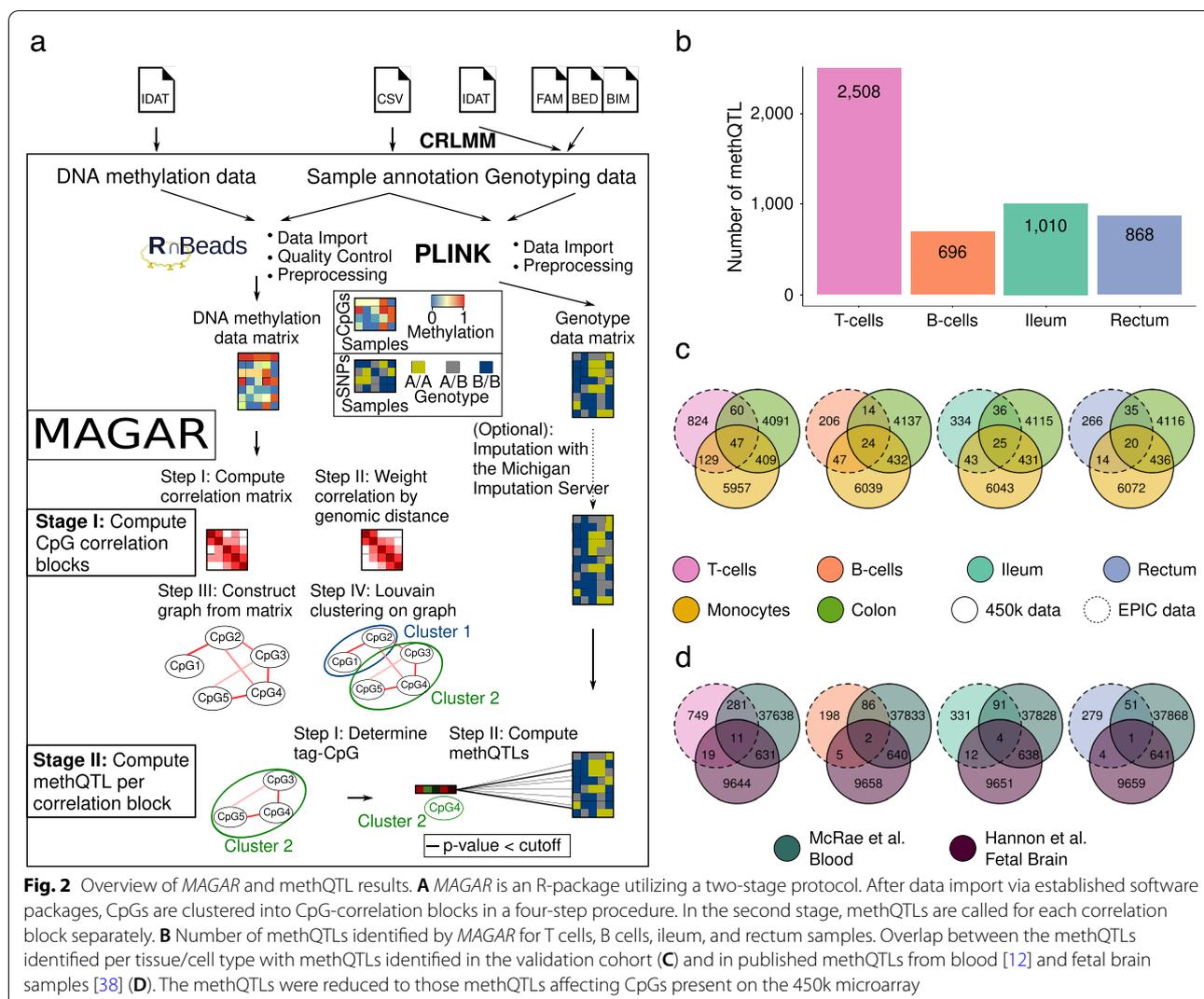


IDAT files) DNA methylation and genotyping data before returning data formatted for methQTL analysis.

In the first phase of *MAGAR*, raw data are converted and processed using the established software packages *RnBeads* [30, 31], *PLINK* [32] and *CRLMM* [35, 36]. The processing includes data filtering of CpGs and SNPs according to quality criteria (see “Methods” for details). The second phase of the package—the methQTL calling—has been implemented as a two-stage workflow as follows: initially, CpGs that exhibit high correlations of methylation states across the samples are clustered into CpG correlation blocks. *MAGAR* takes into consideration that the DNA methylation states of neighboring CpGs in the same functional or regulatory unit are usually highly correlated [37], an observation that we also made in our data (Additional file 1: Figure S2A). This assumption implies that one may not need to inspect each CpG. In fact, doing so would generate many redundant methQTLs, which becomes apparent in the association between methQTL statistics and the genomic distance (Additional file 1: Figure S2B, C). In *MAGAR*, we therefore group neighboring, highly correlated CpGs into correlation blocks. In the second stage of the process, methQTLs are determined individually for each of the CpG correlation blocks. To this end, for each correlation block, *MAGAR* determines a tag-CpG representing this block and determines statistically significant associations for each of the tag-CpGs with all SNPs within a specified genomic distance (in this instance 500 kb up- and downstream). This methQTL calling can either be performed using univariate, linear least squares or by the approach implemented in *fastQTL* [28]. The *fastQTL* software computes correlations between

DNA methylation states and SNP genotypes and uses a permutation scheme to address the multiple testing problem. Comparing *MAGAR*'s output with the output generated by *fastQTL* and *Matrix-eQTL*, we found high overall agreement of the methQTLs detected by the different tools. While *fastQTL* and *Matrix-eQTL* returned a number of methQTLs that were exclusively detected by the respective tool, the methQTLs identified by *MAGAR* were also detected by the other tools (Additional file 1: Figure S3). This indicates that *MAGAR*'s results are more reliable. We leave it to the user to decide whether to use the default linear modeling or *fastQTL* for the second stage of *MAGAR*. In general, *MAGAR* provides various options for customizing the analysis, including options for defining the CpG clustering, for defining the tag-CpG per correlation block, and for the methQTL calling approaches to be employed (linear modeling or *fastQTL*). Reasonable default values for the parameters were selected using simulation experiments (Additional file 1: Text, Figure S4). *MAGAR* returns a list of associations and corresponding statistics, which can be filtered further by the user to define methQTLs or which can be used in downstream analyses. In the analysis presented here, *MAGAR*'s output was used as input to colocalization analysis for defining tissue-specific and tissue-independent methQTLs.

Using *MAGAR*, we analyzed the ileal, rectal, T cell, and B cell methylation data (659,464 CpGs) jointly with genotype data from 5,436,098 SNPs and calculated methQTL statistics for each cell type/tissue independently. To determine significant methQTLs, we selected a Bonferroni-corrected genome-wide  $P$ -value cutoff of  $8.65 \times 10^{-11}$  (see “Methods” for details). As a result, we



found 696, 2508, 1010, and 868 methQTLs for CD19+ B cells, CD4+ T cells, ileal, and rectal biopsies, respectively (Fig. 2B, Additional file 2: Table S1). To validate the methQTLs, we used additional samples from monocytes and transverse colon from the same cohort (Additional file 1: Figure S5). Additionally, we obtained published methQTLs from two studies (blood [12] and fetal brain [38]) and compared them with the identified methQTLs. Note that the validation cohort and the published studies used DNA methylation data generated using the 450k microarray, which comprises fewer CpG sites than the EPIC array. Thus, we excluded those methQTLs from the comparison that associated with a CpG site that is exclusively present on the EPIC array. We identified some of the methQTLs found in the discovery cohort using a different, validation *P*-value cutoff (see “Methods”) in the validation cohort (Fig. 2C) and in the published data

(Fig. 2D). Notably, the overlap between the identified and the published methQTLs was significantly higher than expected by chance (Additional file 3: Table S2). As expected, the overlap of the methQTLs identified in B and T cells with the methQTLs identified using whole blood was higher than with those identified in fetal brain samples (Fig. 2D).

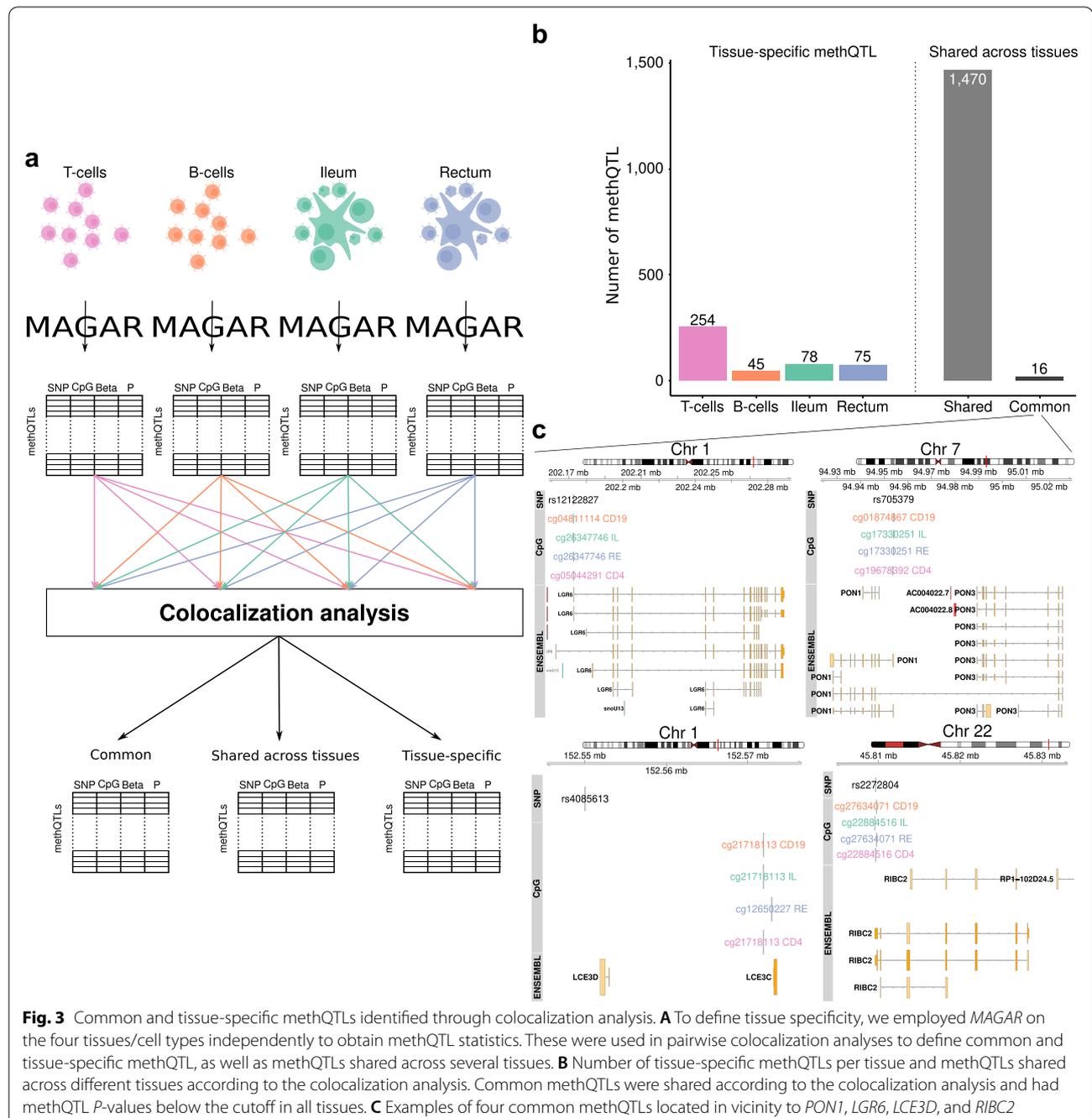
### Colocalization analysis identifies common methQTLs

Next we applied colocalization analysis that uses summary statistics from two association studies (here methQTLs in two different tissues) to determine if an association of two traits (here CpG methylation states) to the same genetic region is significant and is likely to be caused by the same pleiotropic genetic variant. Colocalization was examined using Summary-data-based Mendelian Randomization (SMR) analysis followed by

the Heterogeneity in Dependent Instruments (HEIDI) test [39]. The SMR test indicates whether the two traits are significantly associated with the same locus and the HEIDI test interrogates whether the data are compatible with the hypothesis that both traits are affected by the same underlying functional SNP.

We only included methQTLs in the analysis that were significant at  $P$ -value lower than  $8.65 \times 10^{-11}$  in at least one tissue. The analysis is anchored at the tissue

where the methQTL exhibited a significant association and the methQTL statistics were compared with those in the other tissues. In total, 4253 colocalization tests were performed (Additional file 4: Table S3) based on the number of significant methQTLs. We defined those methQTLs as shared between two tissues/cell types that had an FDR-adjusted  $P$ -value of the SMR test lower than 0.05 and that had a HEIDI test nominal  $P$ -value larger than 0.05 (Additional file 4: Table S3).



These methQTLs are likely driven by the same genetic variant and the shared association is likely caused by a single pleiotropic variant rather than by two linked variants. Colocalization analysis was conducted for all pairs of cell types/tissues (Fig. 3A) and we define three classes of methQTLs:

1. Common methQTLs are shared across all investigated pairs of tissues/cell types according to the colocalization analysis and pass the methQTL  $P$ -value cutoff  $8.65 \times 10^{-11}$  in all tissues.
2. Shared methQTLs are shared across all the investigated pairs of tissues/cell types according to the colocalization analysis.
3. Tissue-specific methQTLs are only present in one of the tissues/cell types and not shared in any pairwise comparison according to the colocalization analysis.

We found that 16 methQTLs were shared across all of the pairwise comparisons and lay below the methQTL  $P$ -value cutoff of  $8.65 \times 10^{-11}$  in all tissues and are thus common methQTLs (Fig. 3B, Additional file 5: Table S4). The common methQTLs included well established methQTLs and eQTLs, such as the ones present in the *PON1* [40], *LGR6* [41], and *RIBC2* [42] loci (Fig. 3C). We found substantially more methQTLs shared across different tissues than tissue-specific methQTLs. Most (254) tissue-specific methQTLs were exclusively found in CD4+ T cells (Fig. 3B, Additional file 6: Table S5), and similar numbers of tissue-specific methQTLs (78, 75) were identified for ileal and rectal biopsies, respectively. Due to the definition above, common methQTLs are a subset of the shared methQTLs.

We used the validation cohort to validate the identified common and shared methQTLs further. Notably, the validation cohort samples were assayed using the 450k array and only 10 (of 16) and 689 (of 1470), respectively, of the common and shared methQTLs associated with a CpG present on the 450k array. We found that most of the common (9/10, Fisher test  $P$ -value:  $1.6 \times 10^{-4}$ ) and some of the shared QTLs (178/689, Fisher test  $P$ -value: 1) were also present in at least one of the two tissues (Additional file 1: Figure S6A,B). Additionally, four of the 10 overlapping common methQTLs (rs2272804, rs705379, rs55901738, rs10021193) were also identified in an independent study of blood samples [12] (Additional file 1: Figure S6C).

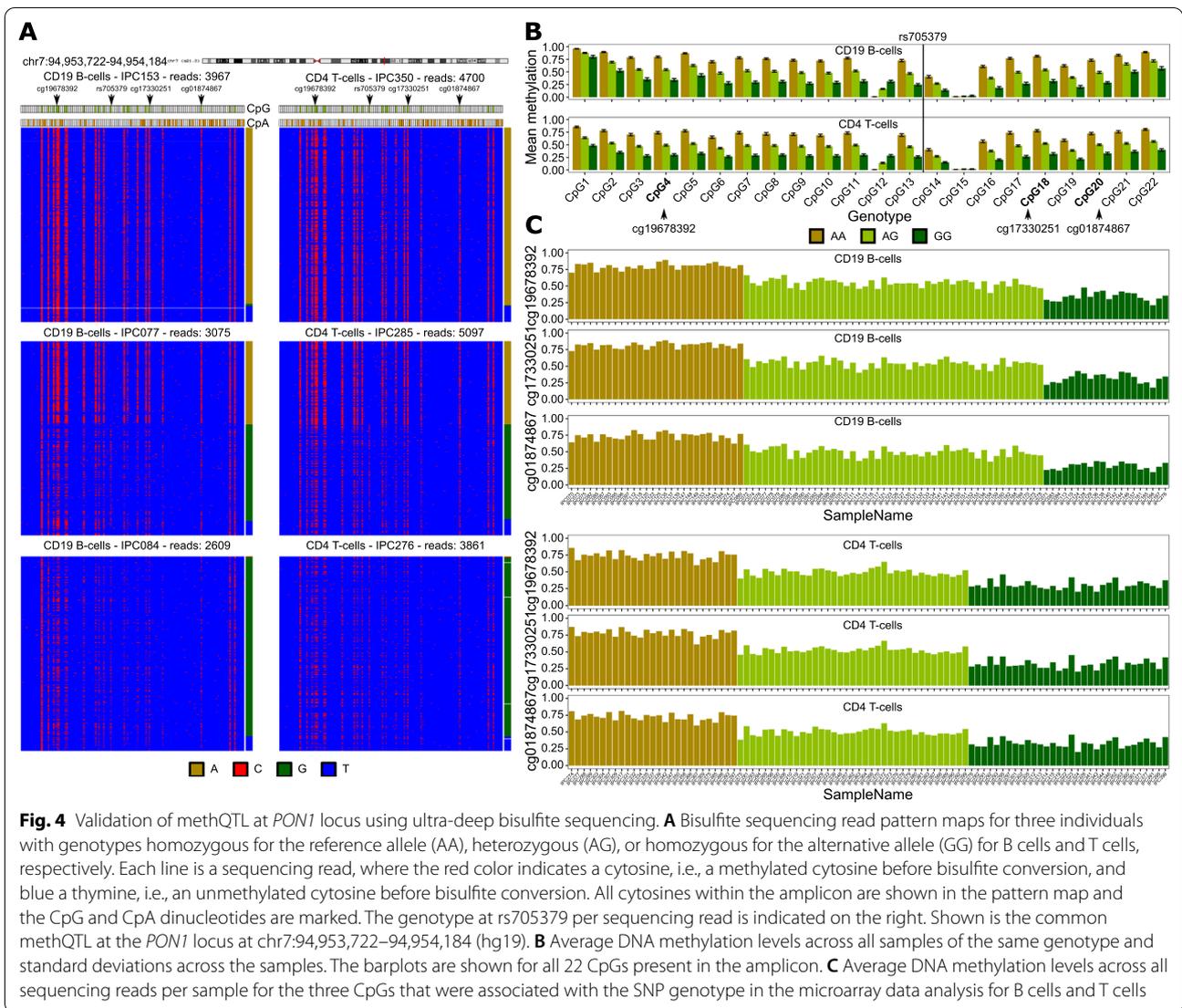
#### Common methQTL at *PON1* locus identified in independent samples using ultra-deep bisulfite sequencing

To rule out potential technology-dependent artifacts, we used local deep amplicon sequencing for the validation

of a common methQTL. We selected the methQTL at the *PON1* locus (comprising rs705379, cg19678392, cg17330251, and cg01874867), since both the SNP and the CpGs could be included into a single amplicon of size 462 base pairs. Thus, we were able to capture the genotype of the SNP and the DNA methylation state of multiple CpGs simultaneously. Notably, we associated the genotype with the CpG methylation state at the single-molecule level, since each sequencing read represents a single molecule. The results indicated a strong relationship between the genotype of rs705379 and the methylation states at all CpGs present in the amplicon, while the effect was stronger in those CpGs that were closer to the SNP (Fig. 4A). In this setting, the A genotype was associated with a high DNA methylation state of more than 50%, while the G genotype leads to a decrease of the methylation level below 25% for some CpGs (Fig. 4B). Notably, there was no one-to-one relationship between the genotype and the DNA methylation state and G genotypes co-occurred with methylated CpGs and A genotypes with unmethylated CpGs at the single-molecule level. The effect of the SNP on the DNA methylation state was consistent across all samples within a genotype, and the standard deviations across the samples within the different genotype groups were comparable (T cells: AA: 0.054, AG: 0.046, GG: 0.058; B cells: AA: 0.042, AG: 0.058, GG: 0.06, Fig. 4C). Notably, rs705379 had a high minor-allele frequency of 0.46 for the B cell and 0.47 for the T cell samples in our cohort. To further investigate whether the effects that we detected are also present for methQTLs beyond the 16 common methQTLs, we constructed two additional amplicons to capture the methQTLs shared across different cell types/tissues at the *ZNF155* (Additional file 1: Figure S7) and *NRG2* (Additional file 1: Figure S8) loci. In accordance with the results obtained in the *PON1* amplicon, we found a strong association of the genotype with DNA methylation states.

#### Tissue-specific methQTLs are preferentially located in proximal enhancer elements

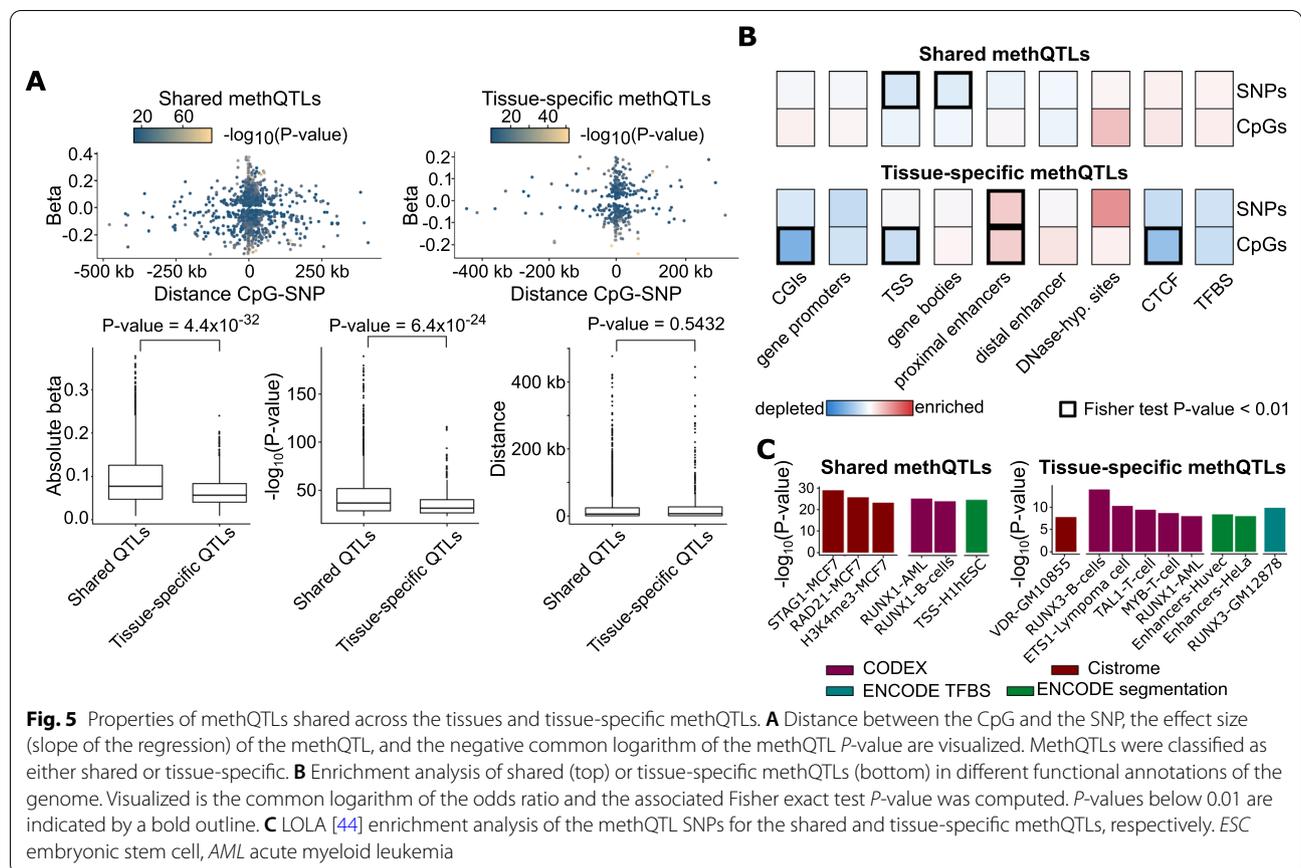
To determine characteristic properties of tissue-specific methQTLs, we compared all 452 tissue-specific methQTLs with 1470 methQTLs shared across multiple tissues (Additional file 7: Table S6). While the distance between the CpG and the SNP that significantly correlates with the DNA methylation state was not different in the two classes of methQTLs, we found both stronger effects on the DNA methylation state with respect to effect size and lower  $P$ -values for the shared methQTLs than for the tissue-specific methQTLs (Fig. 5A). To determine whether the CpGs or the SNPs of the shared and cell type-specific methQTLs are preferentially located



in particular functional regions of the genome, we performed enrichment analysis for various functional annotations such as gene promoters and proximal enhancers. We found that methQTL SNPs were depleted in regions of active transcription such as transcriptional start sites (TSS) and gene bodies for the shared methQTLs (Fig. 5B). No significant enrichment of a functional category was detected for the shared methQTLs. In contrast, the tissue-specific methQTLs were preferentially located in proximal enhancer elements according to the Ensembl Regulatory Build [43] further pointing toward the important regulatory role of enhancers in establishing cellular identity. Further indication for this hypothesis was obtained by the LOLA [44] enrichment of tissue-specific methQTLs in enhancer elements and transcription factor binding sites indicating an enhancer element in B cells

and in the B lymphocyte cell line GM12878 (Fig. 5C). Analogously, we associated the tissue-specific and shared methQTL SNPs and CpGs with overlapping gene bodies. For those overlapping genes, we performed Gene Ontology (GO) enrichment analysis [45] and detected an enrichment of the shared methQTLs towards the biological process “cell development” ( $P$ -value = 0.0069, Additional file 8: Table S7).

We aimed to validate the tissue-specific methQTLs in the validation cohort and in independent studies. While some of the ileum- and rectum-specific methQTLs identified earlier were present in the transverse colon samples, only two of them were present (at validation  $P$ -value cutoff  $9.84 \times 10^{-6}$ ) in the monocytes. Similarly, two of the T-cell-specific methQTLs were also found in transverse colon. However, more (seven for T cells, one for B cells)



were found in the CD14-positive monocytes (Additional file 1: Figure S6A). To validate whether T-cell- and B-cell-specific methQTLs actually capture effects specific to blood cell types, we compared the methQTL effect sizes in the monocytes and in transverse colon. We detected significantly higher effect sizes for the T-cell-specific methQTLs in the monocytes in comparison to transverse colon (Additional file 1: Figure S6B). Notably, not all methQTLs detected in the discovery cohort could be found in the validation cohort, since the latter has been assayed using the Infinium 450k technology. Similarly, more of the T- and B-cell-specific methQTLs were present in the methQTL study on blood samples in comparison to fetal brain samples (Additional file 1: Figure S6C).

### Discussion

Patient stratification according to mutational signatures, i.e., genotype-based markers, is already well-accepted in the clinic [46]. Recently, DNA methylation-based biomarkers are also becoming relevant in a clinical setting [47] and may contribute to clinical decision-making. The relationship between genotype and DNA methylation variation is only just beginning to be understood. As a first step towards the joint characterization of DNA

methylation patterns and genotypes, methylation quantitative trait loci (methQTL) have been identified in healthy individuals. To facilitate standardized analyses of DNA methylation and genotyping data, we developed the R-package *MAGAR* that supports processing of raw data and integrates with established bioinformatic tools. *MAGAR* is the first package providing a start-to-finish workflow for microarray-based methQTL studies and supports bisulfite sequencing data, without specifically using the information on co-methylation of neighboring CpGs present in the sequencing reads. For bisulfite sequencing data, specialized methods are available such as *IMAGE* [48]. Notably, *MAGAR* performs methQTL analysis while accounting for the correlation structure of neighboring CpGs, and is a first step toward associating genetic haplotypes with DNA methylation haplotypes. We also found that *MAGAR*'s results are concordant with results generated by alternative software tools indicating that the identified methQTLs are highly reliable. Grouping together CpGs into clusters is an approach that has also been used earlier [49, 50] in contexts different from methQTL analysis. The earlier approaches to group CpGs into correlation blocks, however, either do not take into account the genomic distance between two CpGs or

are restricted to either microarray or bisulfite sequencing data.

It remains elusive whether methQTLs are inherently cell type-specific or tissue-independent. In this study, we systematically investigated cell-type specificity of methQTLs in sorted blood cell types (CD19+ B cells, CD4+ T cells) and bowel biopsies (ileum, rectum). We found fewer tissue-specific methQTLs than methQTLs that were shared across tissues. We validated tissue specificity in additional CD14+ monocyte and transverse colon samples. Since DNA methylation is a cell type-specific epigenetic mark, it is likely that methQTLs are also cell type-specific. It remains to be shown whether these cell type-specific methQTLs preferentially co-occur with other cell type-specific epigenetic marks such as open chromatin or histone modifications. Previous methQTL studies [12, 38] identified a partially overlapping list of methQTLs, some of which were also detected in this study. Notably, the previous studies used a different strategy for identifying methQTLs (*Merlin* [51] in the blood study and *Matrix-eQTL* [27] in the fetal brain samples). While these strategies do not account for the properties of DNA methylation data, we found a substantial overlap with the methQTLs that we identified.

We found that cell type-specific methQTLs were preferentially located in enhancer elements, which further emphasizes the importance of enhancers to establish cellular identity. However, methQTL effects were weaker in cell type-specific methQTLs compared to those shared across different cell types. It remains to be shown how methQTLs affect gene expression states in our samples. In subsequent analyses, the overlap between methQTLs and eQTLs can be explored to further understand the relationship between genome, epigenome, and transcriptome. Since the cell type-specific methQTLs were associated with the CpG methylation states to a lower extent than shared methQTLs, cell type-specific methQTLs could modulate transcript abundance in a more fine-grained manner. We would also like to point out that this observation may be due to technical rather than biological reasons. Using colocalization analysis for determining shared effects of methQTLs across tissues, a bias towards stronger effects can be introduced. Since we define tissue-specific methQTLs as those that are not shared according to the colocalization analysis, they could be weaker than the shared ones by definition.

There are some aspects of methQTLs, which remain to be investigated. It would be relevant to study cell-type specificity of methQTLs in purified cell types outside of the hematopoietic system, such as in neurons, epithelial cells, and hepatocytes. To that end, the identified common methQTLs could be further validated to determine whether they are truly tissue- and cell-type-independent.

Furthermore, *MAGAR* groups together CpGs into CpG correlation blocks, which reduces the number of redundant methQTLs detected. However, methQTLs affecting single CpGs may be missed using this method. It is well-established that genetic associations with a disease (GWAS hits) are preferentially located in non-coding regions of the genome [52]. The functional impact of such genetic variants, which can be modulated by QTLs (methQTLs, eQTLs), remains to be investigated. Additionally, DNA methylation data can be used to reliably estimate the proportions of different cell types in the samples, either using a reference data set [6] or in a reference-free way [11]; an analysis strategy known as deconvolution. Given the cell-type specificity of a subset of methQTLs identified within this study, a combination of DNA methylation-based deconvolution and identification of methQTLs could be implemented along the line of published, transcriptome-based approaches [25, 53]. By using such a method, it will be possible to investigate methQTL effects in bulk tissues without considering cell type-specific signals. Preferably, novel analysis methods, such as colocalization analysis and the integration of methQTL and DNA methylation-based deconvolution, are implemented in an easy-to-use software package such as *MAGAR*. To overcome the issue of cell-type specificity, DNA methylation can be assayed at the single-cell level and associated with genotype information from the same cell. Alternatively, more readily accessible single-cell RNA-seq data sets can be jointly analyzed with bulk methQTL studies to understand gene regulation at the single-cell level. Finally, long-read sequencing allows for simultaneously profiling the genotype and DNA methylation state of the same molecule over distances up to 10 kb, which enables associating genetic haplotypes with DNA methylation haplotypes.

## Conclusions

In summary, the relationships between genetic and epigenetic variations are currently underexplored. To facilitate the joint analysis of genotype and DNA methylation data, we present *MAGAR* as a novel software tool that accounts for the properties of DNA methylation data. In combination with colocalization analysis, we identified tissue-specific and common methQTLs with unique biological properties and genomic location. Tissue-specific and shared methQTLs identified using *MAGAR* were validated in both independent samples and were verified using an alternative local deep sequencing approach.

## Methods

### MAGAR R-package

#### MAGAR package overview

We developed “Methylation-Aware Genotype Association in R” (*MAGAR*) as a new computational framework to determine methQTLs from DNA methylation and genotyping data. *MAGAR* supports both sequencing-based assays including whole-genome (bisulfite) sequencing and microarray-based data. It is the first computational framework for performing methQTL analysis starting from raw DNA methylation and genotyping microarray data. The pipeline implemented within *MAGAR* comprises the following phases:

- i. *Data import and preprocessing* using established software packages such as *PLINK* [32], *RnBeads* [30, 31], and *CRLMM* [35, 36]. Additional modules for quality control and standard processing using these packages are available to the user. *MAGAR* supports automated imputation using the Michigan Imputation Server [54].
- ii. *MethQTL calling*, i.e., computing associations between genotype and a DNA methylation state. A two-stage approach is employed: (i) Define CpG correlation blocks as groups of CpGs that are highly correlated across the samples to mimic DNA methylation haplotypes. (ii) From each of these correlation blocks, a tag-CpG is selected as a representative of the block and associations are computed with all SNPs up to a given distance using either a linear modeling strategy or using external software tools (e.g., *fastQTL* [28]). All SNP-CpG pairs that have a *P*-value below a user-defined cutoff are returned.

#### Data import and preprocessing

**DNA methylation data** For DNA methylation data, we use the *RnBeads* software package for data handling and processing. *RnBeads* supports most DNA methylation assays yielding single-CpG methylation calls, including whole-genome/reduced-representation bisulfite sequencing (RRBS/WGBS) and the Illumina microarray series. Microarray data can be provided as raw intensity data (IDAT files) and is checked for data quality using *RnBeads*' reporting functionality. Further processing steps, such as CpG and sample filtering (e.g., SNP removal, cross-reactive site removal) and data normalization, can be performed within *RnBeads*. Although we recommend *RnBeads* for data handling, *MAGAR* supports the output of further data processing tools if they provide single-CpG methylation calls.

**Genotyping data** *MAGAR* accepts microarray and sequencing data as input. Sequencing data have to be preprocessed using genotyping pipelines [55] and converted into a format that is readable through *PLINK* (e.g., VCF files). For microarray data, *MAGAR* supports raw intensity data files as input and computes genotype calls through the *CRLMM* R-package [35, 36]. As an optional step, genotyping data can be imputed using the Michigan Imputation Server [54]. Additional data processing, such as filtering SNPs with many missing values or Hardy–Weinberg equilibrium filtering, are conducted through *PLINK*.

#### MethQTL calling

MethQTL calling within *MAGAR* follows a two-stage workflow (Fig. 2A):

- i. CpGs with highly correlated methylation states across the samples are grouped to form CpG-correlation blocks.
- ii. A tag-CpG per CpG-correlation block is associated with all SNPs in a given genomic distance to compute associations between SNP genotypes and DNA methylation states.

We elaborate on the two stages in more details below.

**Correlation block calling** To compute distinct CpG correlation blocks, i.e., groups of CpGs that exhibit high correlations of their methylation states across the samples, from a DNA methylation data matrix we developed a four-step framework:

1. To obtain a similarity matrix, compute the (Pearson) correlation coefficients between the DNA methylation states of any pair of CpGs across the samples using the *bigstatsR* R-package [56] for each chromosome separately. Similarities of two CpGs with correlation lower than 0.2 (package parameter: *cluster.cor.threshold*) are set to zero. Since matrices can grow too large to fit into main memory of standard machines, the CpGs are split per chromosome into equally sized smaller groups until a maximum number of CpGs per computation is achieved (here 40,000 CpGs, parameter: *max.cpgs*).
2. Weight the similarities according to the genomic distance between any CpG and the remaining CpGs using a Gaussian centered at the CpG of interest with standard deviation 3000 bp (parameter: *standard.deviation.gauss*). Additionally, the similarity between any pair of CpGs further apart than 500 kb is set to zero (parameter: *absolute.distance.cutoff*). Optionally, functional annotations such as the Ensembl Regula-

tory Build [43] or DNA methylation-based segmentation [57] can be used to re-define the similarities.

3. Construct the associated weighted graph from the similarity matrix, where the edge weights correspond to the similarities between the two CpGs.
4. Employ Louvain clustering [58] using the *igraph* R-package [59] on the weighted graph to obtain clusters of CpGs that are highly correlated. The obtained clusters are defined as the CpG correlation blocks.

The parameters presented here are available as package options to the user. The default parameters have been evaluated using simulations for EPIC and 450k data (Additional file 1: Text, Figure S4).

**Associating SNPs with CpG-correlation blocks** To determine whether the DNA methylation state of a CpG-correlation block is associated with the SNP, we first determine a tag-CpG per correlation block as the medoid of all CpGs in the correlation block. To compute the medoid CpG, we compute the median for each CpG in the correlation block across the samples. Then, we select the CpG that is the median of the vector of medians across the samples as the tag-CpG. Alternative tag-CpG selection methods are available through the package parameter *representative.cpg.computation*. In the next step, all SNPs closer than 500 kb to the tag-CpG are considered and a univariate, least squares regression (*lm* R function) model is created using the genotypes (encoded as 0 = homozygous reference/major allele, 1 = heterozygous, 2 = homozygous alternative/minor allele) as the features and the CpG methylation state as the response. Further covariates can be included as additional inputs into the linear model. Alternatively, *fastQTL* [28] can be used to compute associations between tag-CpGs and SNPs. The obtained *P*-values and slopes (referred to as effect sizes or beta in this work) of the linear model are used for further analysis.

#### Package options and modularity

*MAGAR* is a modular software package that allows for easy integration with additional tools. Different variants of the analysis can be specified by the package's rich option settings. For instance, CpG correlation blocks depend on various parameters including the correlation threshold between two CpGs or the standard deviation of the Gaussian distribution. We used simulation experiments to determine reasonable default parameter settings for the most widely used technologies 450k array, EPIC array, and bisulfite sequencing (Additional file 1: Text, Figure S4). However, the option setting can be tailored to the data set at hand. CpG correlation block calling can

be deactivated, resulting in the analysis scheme implemented by most published methQTL studies, i.e., associating each CpG with a SNP individually. Additionally, *MAGAR* allows for setting the parameters of the different software tools that are internally used (e.g., *RnBeads*, *PLINK*). To facilitate analyses of large-scale data sets, *MAGAR* supports multi-core processing and automatic distribution of jobs across the nodes of a high-performance computing (HPC) cluster. *MAGAR* comes with different export options, including a direct export into the format accepted by GWAS-MAP (see section "Determining tissue-specific methQTLs"). *MAGAR* is publicly available from Bioconductor (<https://bioconductor.org/packages/MAGAR>).

#### Data sets

The data sets used throughout this study have been generated in the context of the SYSCID project (<http://syscid.eu/>). The Correlated Expression Disease Association Research (CEDAR) [60] cohort data set comprises 164 individuals and we had microarray-based genotyping data available for 163 individuals as described earlier [60]. More specifically, healthy individuals were recruited at the University Hospital in Liège and bowel biopsies as well as blood draws were collected. The biopsies were obtained from rectum (RE) and ileum (IL), and blood cells were FACS-sorted into CD4-positive T cells and CD19-positive B cells. We used this data set as the discovery cohort within this study. In addition, we used a second data set from the CEDAR cohort comprising additional 197 donors (16 overlapping with the earlier ones) including transverse colon biopsies ( $n=191$ ) and CD14-positive monocytes ( $n=192$ ) as a validation cohort.

#### DNA methylation profiling

DNA methylation profiling of the samples in the CEDAR cohort was performed using the Illumina EPIC array. Per sample 500 ng of genomic DNA were bisulfite converted using the EZ-96 DNA methylation Gold Kit (Zymo research, Irvine, USA) according to the kit's manual, except that the final elution volume was reduced to 12  $\mu$ l. Then, four  $\mu$ l of bisulfite converted DNA was used to run on an Infinium Methylation EPIC array (Illumina, San Diego, USA) according to the manufacturer's protocol.

DNA methylation data for the validation cohort were generated earlier using the Illumina 450k microarray according to the standard protocol. Due to the small overlap of donors from the EPIC and 450k data set and due to the reduced number of CpGs available on the 450k array, we decided to analyze the data sets separately.

### Genotyping microarrays

Genotyping of the CEDAR cohort has been performed as described earlier [60]. Additional 23 donors have been genotyped using the Illumina Infinium OmniExpress-24v1.3 microarray at the GIGA-Institute Genomics core facility.

### MAGAR analysis of the CEDAR cohort

#### DNA methylation data

We used *MAGAR*, which internally uses *RnBeads*, for processing raw IDAT files obtained on the CEDAR cohort samples. A subset of samples (13 B-cell samples, 1 T-cell sample) was removed from the discovery cohort, since the samples exhibited substantially lower technical quality. CpGs were filtered for annotated SNPs in *dbSNP* [61], for sites on the sex chromosomes, and for potentially cross-reactive sites [62]. Further quality-based filtering of CpGs was conducted using *RnBeads*' GreedyCut algorithm [30]. Data were normalized using the "dasen" method from the *wateRmelon* R-package [63]. As out-

$$\frac{0.05}{(82,271 + 69,219 + 75,779 + 76,109) \times 1905} = 8.65 \times 10^{-11}.$$

come of the filtering procedure, 659,464 CpGs were retained for the analysis. The immune cell infiltration was estimated using the LUMP algorithm [33] based on 44 CpGs that are particularly hypomethylated in immune cells, 34 of which were available in the CEDAR data set. For the validation cohort (450k), we used analogous processing options, removed one sample from the 383 samples due to lower technical quality, and retained 353,388 from the 485,777 CpGs available on the microarray.

### Genotyping data

Genotyping microarray data were imported into *MAGAR* and genotypes called using the "KRLMM" algorithm implemented in the *CRLMM* R-package [35, 36] using default parameters. Genotypes were imputed using the Michigan Imputation Server [54] using *Minimac4* and the following parameters: Reference panel: "hrc-r1.1", phasing method: "shapeit", population: "eur". Imputation was performed for all 163 unique donors simultaneously and the outcome of the procedure yielded 39,127,678 SNPs. Imputed data were exported to *PLINK* [32] for further processing. We filtered for SNPs with a Hardy-Weinberg equilibrium exact test *P*-value below 0.001, a maximum number of missing values across the samples of 10%, and with minor-allele frequency below 5%. Additionally, we removed samples with more than 5% missing genotypes. Finally, 5,436,098 SNPs and all samples were retained.

### MethQTL analysis

We employed *MAGAR* on an HPC cluster to compute methQTLs for each of the tissues/cell types of the discovery data set independently (Fig. 3B). Notably, we used sex (categorical), age (continuous), BMI (continuous), smoking state (categorical), alcohol intake (categorical), ethnicity (categorical), and the first two principal components (continuous) computed on the genotype data as covariates in the analysis. *MAGAR* returns a table of methQTL summary statistics (i.e., slope of the regression, standard deviation of the estimate, *P*-value), which can be further filtered according to a *P*-value cutoff. Throughout this analysis, we termed methQTLs significant, if they have a *P*-value below a genome-wide Bonferroni-adjusted cutoff of  $8.65 \times 10^{-11}$  in the summary statistics returned by *MAGAR*. We computed the *P*-value cutoff as follows: we identified 82,271, 69,219, 75,779, and 76,109 correlation blocks for T cells, B cells, ileum, and rectum samples, respectively. On average, each CpG was tested for association with 1905 SNPs, which results in:

For each CpG that was affected by more than one methQTL, we selected the SNP with the lowest *P*-value as the lead-SNP.

### Determining tissue-specific methQTLs

To determine whether the effects observed in the four tissues independently were shared across the different samples, we employed colocalization analysis. More specifically, we used Summary-data-based Mendelian Randomization (SMR) and Heterogeneity in Dependent Instruments (HEIDI) analysis [39] implemented in the GWAS-MAP software tool. Briefly, SMR is a statistical test that indicates whether two traits (here CpG methylation states in two tissues) are significantly associated with the same genetic locus. The test is an extension of Mendelian Randomization (MR), which is used to test for a causal relationship between two traits using an instrumental variable. While classical MR requires that the two traits are measured on the same samples, these can be investigated in distinct samples or studies using SMR. The input to the SMR test are methQTL statistics (i.e., *P*-values, slopes of the regression line) obtained in two scenarios, and it returns a test statistic that indicates whether the effect observed in the two scenarios is significantly associated with the same SNP. Thus, SMR analysis determines whether the same genetic effect leads to the methQTL results that we obtained in the two tissues, but cannot discern pleiotropy (the same SNP influences two traits) from linkage (two highly correlated SNPs influence

the traits independently). Thus, for the SNPs that pass the SMR test, we employed the HEIDI test in a second step to test whether the observed effects are likely driven by pleiotropy. Briefly, the HEIDI test utilizes linkage (correlation) information of SNPs from a reference panel to determine whether the observed heterogeneity in the methQTL statistics is more likely caused by linkage than by pleiotropy. By using colocalization analysis through SMR and HEIDI, we were able to determine whether the methQTLs identified in the four tissues/cell types independently were shared or tissue-specific. We employed colocalization analysis for all pairs of tissues/cell types to determine shared methQTLs (Fig. 3B).

We selected those CpGs for colocalization analysis, which were selected as tag-CpGs in at least two tissues and that had a significant association with a lead-SNP ( $P$ -value below  $8.65 \times 10^{-11}$ ) at least in one tissue. Then, anchoring the analysis on the tissue showing the significant association, we performed the SMR test to detect if the same lead-SNP is associated with the same CpG in any of the other tissues. In case the same lead-SNP was identified in more than one tissue, the tissue/cell type with the lowest  $P$ -value was used as the starting point of the SMR analysis. In total, we performed 4253 SMR tests. The SMR  $P$ -values were adjusted for multiple testing using the Benjamini–Hochberg [64] method and we used a  $P$ -value cutoff of 0.05. In case the methQTLs measured in two tissues are significant according to the SMR test, this is an indication that the CpG methylation states are significantly correlated with the same SNP in the two tissues. Thus, we use the  $P$ -value of the SMR test as an indication of the shared effect of methQTLs in the two tissues.

For CpGs that passed the SMR test, we applied the HEIDI test to discern pleiotropy from linkage. We defined all those pairs of methQTLs with a  $P$ -value higher than 0.05 as pleiotropic interactions. The results for a different  $P$ -value cutoff (0.001) are shown in Additional file 4: Table S3. The methQTLs that had an SMR test  $P$ -value below the cutoff and had a HEIDI test  $P$ -value higher than the threshold were defined as shared across the two tissues. The methQTLs shared across all pairwise comparisons according to the colocalization analysis were termed *shared methQTLs*. Additionally, those shared methQTLs with a methQTL  $P$ -value below  $8.65 \times 10^{-11}$  in all tissues were termed *common methQTLs*.

The methQTLs that either fail the SMR test or that pass the SMR test, but also pass the HEIDI test were defined as *tissue-specific methQTLs* (Additional file 4: Table S3). Tissue-specificity was defined for each tissue individually. Finally, three classes of methQTLs were defined: tissue-specific, shared, and common methQTLs. SMR and HEIDI analysis was performed using GWAS-MAP

(<https://www.polyknomics.com/solutions/gwas-map-biomarker-and-intervention-target-discovery-platform>).

#### Characterizing tissue-specific and common methQTLs

We merged the methQTLs from the four tissues and compared the effect sizes,  $P$ -values, and the distance between the CpG and SNP of the tissue-specific with the methQTLs shared across the tissues. Additionally, we selected different functional annotations of the genome, such as Ensembl genes (version 75), associated promoter regions (defined as 1.5 kb upstream and 0.5 kb downstream of the TSS), and different functional categories according to the Ensembl regulatory build [43]. Then, we overlapped the shared/tissue-specific methQTLs with those annotations using the *GenomicRanges* [65] R-package and computed odds ratios and (one-sided) Fisher exact test  $P$ -values to investigate enrichment towards the functional annotations in comparison to all identified methQTLs. Last, we used the LOLA tool [44] to compute enrichments towards various additional functional annotations from databases such as Cistrome [66], CODEX [67], or ENCODE [68]. In contrast to the annotation enrichment analysis, we performed LOLA enrichment analysis using all CpGs/SNPs that were analyzed as the background for the enrichment.

#### Validation of methQTLs

##### Validation using independent data sets

For further validation of the methQTLs identified above, we used 191 transverse colon and 192 monocyte samples from the CEDAR cohort assayed using the Infinium 450k microarray. Genotyping and DNA methylation data were processed analogously to the discovery cohort and methQTLs were called at the  $P$ -value cutoff  $9.84 \times 10^{-6}$ . We aimed to replicate the 2508, 696, 1010, and 868 methQTLs that we identified in the four tissues/cell types and thus computed the  $P$ -value cutoff as:

$$\frac{0.05}{2508 + 696 + 1010 + 868} = 9.84 \times 10^{-6}.$$

We used sex (categorical), age (continuous), BMI (continuous), smoking (categorical), alcohol intake (categorical), ethnicity (categorical), and the first two principal components (continuous) computed on the genotype data as covariates. The resulting methQTLs were compared with the common and tissue-specific methQTLs detected in the discovery cohort, respectively. Additionally, we obtained methQTL data in tabular form from two studies identifying methQTLs in peripheral blood [12] and fetal brain samples [38], respectively. The two studies identified 52,918 (blood) and 16,811 (fetal brain) methQTLs. We only used unique SNPs with a  $P$ -value

**Table 1** Details on bisulfite amplicons screened in the study

Gene locus	chr	SNP position ID	PCR primers (5'-3') <sup>a</sup>	CpG ID	MethQTL distance (bp)	MethQTL delta <sup>b</sup> EPIC, NGS
<i>PON1</i>	7	94,953,895 rs705379	TCTTCCCTACACGACGCTCTCCGATCTgattggtggtttt gaagagtgttagttt GTGACTGGAGTTCAGACGTGTGCTCTCCGATCTccat aatcaaaactactaaatctctctaaaac	cg01874867, cg20119798	164 249	+ 14.9%, + 39.7%
<i>ZNF155</i>	19	44,488,352 rs62116613	TCTTCCCTACACGACGCTCTCCGATCTggttgataggt tagaattataggtt GTGACTGGAGTTCAGACGTGTGCTCTCCGATCTcacat acttaactcaaacacctt	cg23456212, cg20451226	182 171	+ 7.8%, + 13.3%
<i>NRG2</i>	5	139,340,779 rs6580323	TCTTCCCTACACGACGCTCTCCGATCTttatgaattt gaagaagtgttaggt GTGACTGGAGTTCAGACGTGTGCTCTCCGATCTcacata caaaactaaaactaaatcc	cg22710094	85	- 5.1%, - 20.1%

<sup>a</sup> Capital letters are NGS-compatible tags

<sup>b</sup> Absolute methylation change of homozygote minor versus major individuals

lower than  $8.65 \times 10^{-11}$  to match our criteria. To determine whether the detected overlap was larger than expected by chance, we used Fisher's exact test using all SNPs that have been used as input to the methQTL calling as the background set.

#### Validation using local deep sequencing

For validation of the common methQTLs at the *PON1* locus, as well as the shared methQTLs at the *ZNF155*, and *NRG2* loci, we performed local deep sequencing using independent samples from the CEDAR cohort. 500 ng of genomic DNA were bisulfite converted using the EZ-96 DNA methylation Gold Kit (Zymo research, Irvine, USA) according to the kit's manual. PCRs were set-up in 30 µl reactions using three µl of 10 × HotStar-Taq buffer (Qiagen, Hilden Germany), 2.4 µl of 10 mM dNTPs (Fisher Scientific, Pittsburgh, USA), 1.5 µl of 25 mM MgCl<sub>2</sub> (Qiagen), 0.3 µl each of 10 µM forward and reverse primers (Table 1), 0.5 µl of five U/µl HotStar-Taq Polymerase (Qiagen), two µl of bisulfite converted DNA, and 20 µl of aqua bidest. PCRs were performed in an ABI Veriti thermo-cycler (Life Technologies, Karlsbad, USA) using the following program: 95 °C for 15 min, 40 cycles of 95 °C for 30 s, 1.5 min of 56 °C, and one min at 72 °C, followed by five min of 72 °C and hold at 12 °C. PCR products were cleaned up using Agencourt AMPure XP Beads (Beckman Coulter, Brea, USA) and concentration was measured. All amplified products were diluted to four nM and NGS tags were finalized by a second PCR step (five cycles) with primers matching to the NGS-compatible tags and carrying a sample-specific barcode (forward 5'-3': CAAGCAGAAGACGGCATAACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT; reverse 5'-3': AATGATACGGCGACCACC GAGATCTACACXXXXXXXXTCTTCCCTACACGA

CGCTCTTCCGATC; 'X's refer to the sample barcode position) followed by a clean-up (AMPure XP). Finally, all samples (set to ten nM) were pooled, loaded on an Illumina MiSeq sequencing machine and sequenced for 2 × 300 bp paired-end reads with a MiSeq reagent kit V3 (Illumina) to ca. 10 k–20 k fold depth.

Quality control of the raw data files was performed using the *FastQC* software (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Adaptor trimming and filtering for excluding low-quality bases was conducted through *cutadapt* [69] and *Trim Galore!* ([https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)). Paired reads were joined with the *FLASH* tool [70]. Next, reads were sorted in a two-step procedure by (i) the NGS barcode adaptors to assign samples to identifiers and (ii) the initial 15 bp to assign data to the amplicons. Subsequently, the sorted data were input to *BiQAnalyzer HT* [71] using the following settings: 'analyzed methylation context' was set to "C", 'minimal sequence identity' was set to 0.9, and 'minimal conversion rate' was set to 0.95. The filtered high-quality reads were used to compute methylation levels of the respective CpGs. Finally, each read was tagged by its base call at the respective SNP position in the amplicon (*PON1@173*, *ZNF155@329*, *NRG2@255*).

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13072-021-00415-6>.

**Additional file 1:** Additional text including method descriptions and additional figures S1-S10.

**Additional file 2: Table S1.** MethQTLs identified in the four tissues/cell types using *MAGAR*.

**Additional file 3: Table S2:** Enrichment *P*-values according to Fisher's exact test for validation of the identified methQTLs in independent

samples (monocytes, transverse colon) and independent studies (blood and fetal brain).

**Additional file 4: Table S3:** Results of the colocalization analysis for different *P*-value cutoffs of the HEIDI test (0.05 and 0.001).

**Additional file 5: Table S4:** Common methQTLs across the four tissues/cell types.

**Additional file 6: Table S5:** Tissue-specific methQTLs for the four tissues/cell types.

**Additional file 7: Table S6:** MethQTLs shared across the tissues/cell types according to the colocalization analysis. The table comprises 1912 rows and we focus on the 1470 unique SNPs in the analysis.

**Additional file 8: Table S7:** GO enrichment analysis results for the shared and tissue-specific methQTLs.

### Acknowledgements

We appreciate the help of Ivan Kuznetsov with data management and analysis and would like to thank Myriam Mni and the GIGA-Institute Genomics core facility for technical assistance. We appreciate the help from the data management team at the University of Luxembourg, especially from Wei Gu.

### Authors' contributions

MS developed MAGAR, generated the figures, and performed the analysis together with GG, YSA, TS, and AN. GG generated the DNA methylation data on the samples collected by SR, EL, and MG. GG and MA performed the local deep amplicon sequencing. DA computed cis-regulatory domains on the methylation data. PAL, SR, MG, YSA, DA, ETD, TL, and JW contributed to the discussion of the results. JW, TL, and MG jointly supervised the project. MS wrote the manuscript with input from all co-authors. All authors read and approved the final manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by the EU H2020 project SYSCID (733100), an MRC Programme Grant (MR/L019027/1) to P.A.L., and by ELIXIR Luxembourg via its data hosting service. The work of A.N. was supported by the Ministry of Education and Science of the Russian Federation via the state assignment of the Novosibirsk State University (project "Graduates 2020").

### Availability of data and materials

The methylation data used within this publication is accessible at the Data Catalog hosted at the University of Luxembourg at <https://datacatalog.elixir-luxembourg.org/e/study/c7515074-b34a-11eb-9969-acde48001122>. Genotyping data have been obtained from the ArrayExpress website (<https://www.ebi.ac.uk/arrayexpress>) accession number E-MTAB-6666. MAGAR is available from GitHub at <https://github.com/MPIIComputationalEpigenetics/MAGAR/> and through Bioconductor (<https://bioconductor.org/packages/MAGAR>).

### Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

Y.S.A. is a co-founder and co-owner of PolyKnomics, a private organization active in the fields of quantitative, computational, and statistical genomics. All other authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Department of Genetics/Epigenetics, Saarland University, Saarbrücken, Germany. <sup>2</sup>Computational Biology, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. <sup>3</sup>Graduate School of Computer Science, Saarland Informatics Campus, Saarbrücken, Germany. <sup>4</sup>Unit of Animal Genomics, GIGA-Institute & Faculty of Veterinary Medicine, University of Liège,

Belgium. <sup>5</sup>Kurchatov Genomics Center of the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. <sup>6</sup>Research and Training Center on Bioinformatics, A.A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia. <sup>7</sup>Department of Gastroenterology, Liège University Hospital, CHU Liège, Liège, Belgium. <sup>8</sup>Novosibirsk State University, Novosibirsk, Russia. <sup>9</sup>Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland. <sup>10</sup>Swiss Institute of Bioinformatics (SIB), University of Geneva, Geneva, Switzerland. <sup>11</sup>Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland. <sup>12</sup>Moscow Institute of Physics and Technology (State University), Moscow, Russia. <sup>13</sup>PolyKnomics BV, 's-Hertogenbosch, The Netherlands. <sup>14</sup>Department of Medicine, University of Cambridge School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. <sup>15</sup>Cambridge Institute for Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, Cambridge CB2 0AW, UK. <sup>16</sup>Present Address: Department of Bioinformatics and Genomics, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain.

Received: 20 April 2021 Accepted: 2 August 2021

Published online: 16 September 2021

### References

- Durek P, Nordström K, Gasparoni G, Salhab A, Kressler C, de Almeida M, et al. Epigenomic profiling of human CD4+ T cells supports a linear differentiation model and highlights molecular regulators of memory development. *Immunity*. 2016;45:1148–61. <https://doi.org/10.1016/j.immuni.2016.10.022>.
- Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*. 2005;37:853–62. <https://doi.org/10.1038/ng1598>.
- Reik W, Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet*. 2001;2:21–32. <https://doi.org/10.1038/35047554>.
- Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, et al. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell*. 2012;148:816–31. <https://doi.org/10.1016/j.cell.2011.12.035>.
- Fang F, Hodges E, Molaro A, Dean M, Hannon GJ, Smith AD. Genomic landscape of human allele-specific DNA methylation. *Proc Natl Acad Sci U S A*. 2012;109:7332–7. <https://doi.org/10.1073/pnas.1201310109>.
- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform*. 2012;13:86. <https://doi.org/10.1186/1471-2105-13-86>.
- Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 2014;15:R31. <https://doi.org/10.1186/gb-2014-15-2-r31>.
- Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet*. 2018;19:371–84. <https://doi.org/10.1038/s41576-018-0004-3>.
- Almén MS, Nilsson EK, Jacobsson JA, Kalnina I, Klovin J, Fredriksson R, et al. Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity. *Gene*. 2014;548:61–7. <https://doi.org/10.1016/j.gene.2014.07.009>.
- Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenet*. 2015;7:113. <https://doi.org/10.1186/s13148-015-0148-3>.
- Scherer M, Nazarov PV, Toth R, Sahay S, Kaoma T, Maurer V, et al. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecomPipeline, MeDeCom and FactorViz. *Nat Protoc*. 2020;15:3240–63. <https://doi.org/10.1038/s41596-020-0369-6>.
- McRae AF, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, et al. Identification of 55,000 replicated DNA methylation QTL. *Sci Rep*. 2018;8:17605. <https://doi.org/10.1038/s41598-018-35871-w>.
- Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across

- the human life course. *Genome Biol.* 2016;17:61. <https://doi.org/10.1186/s13059-016-0926-z>.
14. Morrow JD, Glass K, Cho MH, Hersh CP, Pinto-Plata V, Celli B, et al. Human lung DNA methylation quantitative trait loci colocalize with chronic obstructive pulmonary disease genome-wide association loci. *Am J Respir Crit Care Med.* 2018;197:1275–84. <https://doi.org/10.1164/rccm.201707-1434OC>.
  15. Dai JY, Wang X, Wang B, Sun W, Jordahl KM, Kolb S, et al. DNA methylation and cis-regulation of gene expression by prostate cancer risk SNPs. *PLoS Genet.* 2020;16: e1008667. <https://doi.org/10.1371/journal.pgen.1008667>.
  16. Rice SJ, Cheung K, Reynard LN, Loughlin J. Discovery and analysis of methylation quantitative trait loci (mQTLs) mapping to novel osteoarthritis genetic risk signals. *Osteoarthr Cartil.* 2019;27:1545–56. <https://doi.org/10.1016/j.joca.2019.05.017>.
  17. Clark AD, Nair N, Anderson AE, Thalayasingam N, Naamane N, Skelton AJ, et al. Lymphocyte DNA methylation mediates genetic risk at shared immune-mediated disease loci. *J Allergy Clin Immunol.* 2020;145:1438–51. <https://doi.org/10.1016/j.jaci.2019.12.910>.
  18. Kim S, Forno E, Yan Q, Jiang Y, Zhang R, Boutaoui N, et al. SNPs identified by GWAS affect asthma risk through DNA methylation and expression of cis-genes in airway epithelium. *Eur Respir J.* 2020;55:1902079. <https://doi.org/10.1183/13993003.02079-2019>.
  19. Gao X, Thomsen H, Zhang Y, Breitling LP, Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. *Clin Epigenet.* 2017;9:87. <https://doi.org/10.1186/s13148-017-0387-6>.
  20. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, et al. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am J Hum Genet.* 2018;103:654–65. <https://doi.org/10.1016/j.ajhg.2018.09.007>.
  21. Zhao T, Hu Y, Zang T, Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front Genet.* 2019;10:1–8. <https://doi.org/10.3389/fgene.2019.01021>.
  22. Pierce BL, Tong L, Argos M, Demanelis K, Jasmine F, Rakibuz-Zaman M, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nat Commun.* 2018;9:804. <https://doi.org/10.1038/s41467-018-03209-9>.
  23. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol.* 2021;22:127. <https://doi.org/10.1186/s13059-021-02347-6>.
  24. Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet.* 2015;11: e1004958. <https://doi.org/10.1371/journal.pgen.1004958>.
  25. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science (80-)*. 2020;369:eaz8528. <https://doi.org/10.1126/science.aaz8528>.
  26. Peters JE, Lyons PA, Lee JC, Richard AC, Fortune MD, Newcombe PJ, et al. Insight into genotype-phenotype associations through eQTL mapping in multiple cell types in health and immune-mediated disease. *PLoS Genet.* 2016;12: e1005908. <https://doi.org/10.1371/journal.pgen.1005908>.
  27. Shabalín AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* 2012;28:1353–8. <https://doi.org/10.1093/bioinformatics/bts163>.
  28. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics.* 2016;32:1479–85. <https://doi.org/10.1093/bioinformatics/btv722>.
  29. Pan H, Holbrook JD, Karmani N, Kwok CK. Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. *BMC Bioinform.* 2016;17:299. <https://doi.org/10.1186/s12859-016-1161-z>.
  30. Assenov Y, Müller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods.* 2014;11:1138–40. <https://doi.org/10.1038/nmeth.3115>.
  31. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* 2019;20:55. <https://doi.org/10.1186/s13059-019-1664-9>.
  32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75. <https://doi.org/10.1086/519795>.
  33. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun.* 2015;6:8971. <https://doi.org/10.1038/ncomms9971>.
  34. Mowat AM, Agace WW. Regional specialization within the intestinal immune system. *Nat Rev Immunol.* 2014;14:667–85. <https://doi.org/10.1038/nri3738>.
  35. Scharpf RB, Irizarry RA, Ritchie ME, Carvalho B, Ruczinski I. Using the R package crlmm for genotyping and copy number estimation. *J Stat Softw.* 2011. <https://doi.org/10.18637/jss.v040.i12>.
  36. Ritchie ME, Carvalho BS, Hetrick KN, Tavare S, Irizarry RA. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics.* 2009;25:2621–3. <https://doi.org/10.1093/bioinformatics/btp470>.
  37. Eckhardt F, Lewin J, Cortese R, Rakyán VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet.* 2006;38:1378–85. <https://doi.org/10.1038/ng1909>.
  38. Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci.* 2015;19:48–54. <https://doi.org/10.1038/nn.4182>.
  39. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* 2016;48:481–7. <https://doi.org/10.1038/ng.3538>.
  40. Huen K, Yousefi P, Street K, Eskenazi B, Holland N. PON1 as a model for integration of genetic, epigenetic, and expression data on candidate susceptibility genes. *Environ Epigenet.* 2015;1:1–11. <https://doi.org/10.1093/eep/dvv003>.
  41. Volkov P, Olsson AH, Gillberg L, Jørgensen SW, Brøns C, Eriksson K-F, et al. A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. *PLoS ONE.* 2016;11: e0157776. <https://doi.org/10.1371/journal.pone.0157776>.
  42. Wragg D, Liu Q, Lin Z, Riggio V, Pugh CA, Beveridge AJ, et al. Using regulatory variants to detect gene–gene interactions identifies networks of genes linked to cell immortalisation. *Nat Commun.* 2020;11:343. <https://doi.org/10.1038/s41467-019-13762-6>.
  43. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. *Genome Biol.* 2015;16:56. <https://doi.org/10.1186/s13059-015-0621-5>.
  44. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics.* 2016;32:587–9. <https://doi.org/10.1093/bioinformatics/btv612>.
  45. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics.* 2007;23:257–8. <https://doi.org/10.1093/bioinformatics/btl567>.
  46. Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci Transl Med.* 2015;7:283ra53. <https://doi.org/10.1126/scitranslmed.aaa7161>.
  47. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. *Nature.* 2018;555:469–74. <https://doi.org/10.1038/nature26000>.
  48. Fan Y, Vilgalys TP, Sun S, Peng Q, Tung J, Zhou X. IMAGE: high-powered detection of genetic effects on DNA methylation using integrated methylation QTL mapping and allele-specific analysis. *Genome Biol.* 2019;20:220. <https://doi.org/10.1186/s13059-019-1813-1>.
  49. Sofer T, Schifano ED, Hoppin JA, Hou L, Baccarelli AA. A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics.* 2013;29:2884–91. <https://doi.org/10.1093/bioinformatics/btt498>.
  50. Gatev E, Gladish N, Mostafavi S, Kobor MS. CoMeBack: DNA methylation array data analysis for co-methylated regions. *Bioinformatics.* 2020;36:2675–83. <https://doi.org/10.1093/bioinformatics/btaa049>.
  51. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.* 2002;30:97–101. <https://doi.org/10.1038/ng786>.
  52. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in

- regulatory DNA. *Science* (80-). 2012;337:1190–5. <https://doi.org/10.1126/science.1222794>.
53. Aguirre-Gamboa R, de Klein N, di Tommaso J, Claringbould A, van der Wijst MG, de Vries D, et al. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinform*. 2020;21:243. <https://doi.org/10.1186/s12859-020-03576-5>.
  54. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284–7. <https://doi.org/10.1038/ng.3656>.
  55. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68–74. <https://doi.org/10.1038/nature15393>.
  56. Prive F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics*. 2018;34:2781–7. <https://doi.org/10.1093/bioinformatics/bty185>.
  57. Salhab A, Nordström K, Gasparoni G, Kattler K, Ebert P, Ramirez F, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol*. 2018;19:9–11. <https://doi.org/10.1186/s13059-018-1510-5>.
  58. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. 2008;2008:P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
  59. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJ*. 2006;1695:1–9.
  60. Momozawa Y, Dmitrieva J, Théâtre E, Deffontaine V, Rahmouni S, Charlotiaux B, et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun*. 2018;9:2427. <https://doi.org/10.1038/s41467-018-04365-8>.
  61. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308–11. <https://doi.org/10.1093/nar/29.1.308>.
  62. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17:208. <https://doi.org/10.1186/s13059-016-1066-1>.
  63. Pidsley R, Wong YCC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genom*. 2013;14:293. <https://doi.org/10.1186/1471-2164-14-293>.
  64. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57:289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
  65. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;9:e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>.
  66. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol*. 2011;12:R83. <https://doi.org/10.1186/gb-2011-12-8-r83>.
  67. Sánchez-Castillo M, Ruau D, Wilkinson AC, Ng FSL, Hannah R, Diamanti E, et al. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res*. 2015;43:D1117–23. <https://doi.org/10.1093/nar/gku895>.
  68. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74. <https://doi.org/10.1038/nature11247>.
  69. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJ*. 2011;17:10. <https://doi.org/10.14806/ej.17.1.200>.
  70. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27:2957–63. <https://doi.org/10.1093/bioinformatics/btr507>.
  71. Lutsik P, Feuerbach L, Arand J, Lengauer T, Walter J, Bock C. BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Res*. 2011;39:W551–6. <https://doi.org/10.1093/nar/gkr312>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

