From correlation to causation: Data-driven exploration of transcriptional regulation using population genomics

Luijk, R.

2 An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs

René Luijk, J.J. Goeman, E.P. Slagboom,
B.T. Heijmans, and E.W. van Zwet

Bioinformatics, **31(3):340-345 (2015)**

Abstract

An increasing number of studies investigates the influence of local genetic variation on DNA methylation levels, so called in *cis* methylation Quantitative Trait Loci (meQTL*s*). A common multiple testing approach in genome-wide *cis*-meQTL studies limits the false discovery rate (FDR) among all CpG-SNP pairs to 0.05 and reports on CpGs from the significant CpG-SNP pairs. However, a statistical test for each CpG is not performed, potentially increasing the proportion of CpGs falsely reported on. Here, we presented an alternative approach that does properly control for multiple testing at the CpG level.

We performed *cis*-meQTL mapping for varying window sizes using publicly available SNP and 450k data, extracting the CpGs from the significant CpG-SNP pairs (FDR < 0.05). Using a new bait-and-switch simulation approach, we show that up to 50% of the CpGs found in the simulated d ata may be false positives. We present an alternative, two-step multiple testing approach using the Simes and Benjamini-Hochberg procedures that does control the FDR among the CpGs, as confirmed by the bait-and-switch simulation. This approach indicates the use of window sizes in *cis*-meQTL mapping studies that are significantly smaller than commonly adopted.

Our approach to *cis* meQTL mapping properly controls the FDR at the CpG level, is computationally fast and can also be applied to *cis* eQTL studies.

Introduction

Genome-wide association studies (GWASs) are widely used to uncover the genetic basis of complex disease. Disease-associated genetic variants identified in GWASs are commonly located in non-coding regions, leaving the molecular mechanism underlying the associations unclear [Visscher et al., 2012]. The likely mechanism involves an effect on transcriptional activity of genes nearby (*cis*) or located distantly (*trans*), for example by influencing epigenetic regulation [Mill and Heijmans, 2013]. This can be studied by investigating the relationship between genetic variation, epigenetic marks including DNA methylation and gene expression. Already, many studies have reported on associations of specific genetic variants with variation in gene expression (expression QTL or eQTLs, Small et al. [2011]; Westra et al. [2013]) and DNA methylation, in particular the methylation of cytosines in CpG dinucleotides (*e.g.*, Heijmans et al. [2007]; Shi et al. [2014]; Wagner et al. [2014]) (DNA methylation quantitative trait loci or meQTLs). Creating catalogs of meQTLs and eQTLs will be instrumental in the discovery of genetic mechanisms determining DNA methylation and gene expression, the possible interplay between the two, and eventually the etiology of common diseases. To achieve this goal, further development of sound statistical methodology will be important.

Typically in meQTL and eQTL studies, a GWAS (*i.e.*, testing hundreds of thousands to millions of single nucleotide polymorphisms, SNPs) is performed for the level of methylation of every CpG measured or the level of transcription of every gene (more generally, for every transcript or exon), respectively, leading to a vast amount of possible combinations to investigate. While we will focus on *cis* meQTL studies, we note that the same principles and problems may also apply to *cis* eQTL studies. With the recent introduction of the Illumina 450k DNA methylation array [Bibikova et al., 2011], meQTL studies have become possible investigating over 400 thousand CpGs in large numbers of subjects. To test for associations of methylation at CpGs with genetic variants *in cis*, that is locally, studies have been considering SNPs anywhere between 5 kb [Gutierrez-Arcelus et al., 2013] to 1,000 kb [Gibbs et al., 2010] from measured CpGs. Particularly large window sizes will result in hundreds of millions statistical tests and thus brings about a huge multiple testing problem. A common strategy to account for multiple testing in meQTL studies is to control the false discovery rate (FDR; Benjamini and Hochberg [1995]) of all significantly associated CpG-SNP pairs at 0.05 (*e.g.*, Grundberg et al. [2013]; Drong et al. [2013]). This means that 5% of all significantly associated CpG-SNP pairs are expected to be false positives.

Due to the extensive linkage disequilibrium (LD) in the human genome, individual CpGs will frequently be associated with many SNPs. Hence, a particular CpG will often occur many times in the list of significant CpG-SNP pairs. In practice, this is redundant information because LD structure renders it impossible to pinpoint the causal SNP responsible for the variation in DNA methylation using statistical means (cf. GWAS; Pearson and Manolio [2008]; Feero et al. [2010]). Hence, the results reported on and further analyses generally focus on the CpGs in the list of significant CpG-SNP associations. That is, all CpGs that significantly

associate with at least one SNP (*e.g.*, Zhang et al. [2010]; Liu et al. [2013]; van Eijk et al. [2012]). We will refer to this approach as the CpG-SNP pair-based approach.

A large proportion of CpGs among the FDR significant CpG-SNP pairs may be false positives [Bell et al., 2011; Westra et al., 2013]. To obtain a list of CpGs influenced by genetic variation *in cis* that is properly controlled for multiple testing, we propose to formally test each CpG, obtaining a single valid P-value per CpG and control the FDR among those P-values, which we will refer to as the CpG-based approach. Using a new bait-and-switch simulation scheme we compare the proportion of falsely identified CpGs using the CpG-SNP pair-based approach and our proposed CpG-based approach in simulated data.

Methods

Data

We used Illumina 450k DNA methylation data [Heyn et al., 2013] and Illumina HumanHap 550k SNP data [Niu et al., 2010] on 96 unrelated healthy Caucasian-Americans. The DNA samples were obtained from lymphoblastoid cell lines included in the Human Variation panel (sample set HD100CAU; Coriell Cell Repositories). Both data sets are publicly available from the GEO data repository (accession numbers GSE36369 and GSE24260, respectively). The SNP array data were imputed to 30,038,302 SNPs based on the 1000 Genomes CEU reference panel and using IMPUTE v2 [Howie et al., 2009]. A dosage value ranging from 0 to 2 reflected the uncertainty in the imputation for the imputed SNPs. We selected SNPs with a minor allele frequency above 5%, a minimum call rate of 95%, and an imputation quality score of at least 0.4, leaving 6,596,758 SNPs for analysis.

The quality control of the 450k array was done based on the signal intensities and detection P-values. We set any beta values [Du et al., 2010], a measure of the DNA methylation fraction, with a corresponding detection P-value lower than 0.01 to missing. Next, we removed any samples with a log2 median intensity under 10.5 in either the methylated or the unmethylated signal. In addition, we removed any probes or samples with a call rate lower than 95%. Lastly, we removed probes mapping to the sex chromosomes, mapping ambiguously to the genome [Chen et al., 2013], or with a SNP in the interrogated CpG (MAF $> 1\%$ in 1000 Genomes). These filters resulted in 423,825 probes left for analysis out of the 482,421 probes on the array targeting CpG sites. The normalization of the 450k data consisted of a correction for background signal, followed by a dye-bias correction. Both procedures were performed using the methylumi package [Davis et al., 2013]. All further analyses were done using beta values. To verify that genotype and methylation data were linked to the correct sample identified, MixupMapper was used [Westra et al., 2011]. For 77 out of 93 samples remaining after quality control, SNP and methylation data could be linked (Supplementary Tables 1 and 2).

meQTL mapping

We tested all associations between genotypes and methylation of CpGs *in cis*, that is, locally, defined by window sizes from 1 kb to 500 kb around each CpG, calculating the Spearman rank correlation between the imputed dosage values and beta values. To this end, we use the Matrix eQTL package [Shabalin, 2012]. Because the Matrix eQTL package is only able to calculate the Pearson correlation, which is less robust to outliers than the Spearman rank correlation, we pre-calculated the ranks of the observed values for all CpGs and SNPs as input for Matrix eQTL to obtain a test on the basis of the Spearman correlation. The Matrix eQTL package provides a list of all CpG-SNP pairs tested across all windows evaluated and the P-values reflecting the statistical significance of the associations. Obtaining a list of statistically significant CpG-SNP pairs was achieved by limiting the false discovery rate (FDR) among the CpG-SNP pairs to 0.05.

Obtaining an FDR controlled list of CpGs influenced by genetic variation

While the FDR among the CpG-SNP pairs is controlled at 0.05, there is no guarantee that this is also true for the set of CpGs among these pairs. No formal statistical test is performed for each CpG individually, testing the global null hypothesis H_0 of no association between the variation in methylation and genetic variation *in cis*. In order to obtain a list of CpGs that is controlled at an FDR of 0.05, we proceed as follows. First, we perform a statistical test to assess the global null hypothesis $H_{0,i}$ of no association between a CpG i and the SNPs *in cis* to obtain one valid P-value for each CpG. Next, we apply the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to these P-values to obtain an FDR controlled list of CpGs associated with genetic variation. Since commonly used software packages return P-values $p_{i,j}$ for all CpG-SNP pairs (i, j) tested, we propose to use the $p_{i,j}$ to test the global null hypothesis $H_{0,i}$. The Bonferroni correction multiplies the minimum of the observed P-values in a window by the number of such P-values

$$P_i = k_i \min(p_{1,i}, \ldots, p_{k_i,i}), \tag{2.1}$$

where k_i is the number of SNPs in that window. The Bonferroni correction is conservative in the case of dependent P-values (like in the case of LD between SNPs), since the effective number of tests done may be smaller than the number of tests corrected for by Bonferroni. Hence, we propose to use the Simes procedure [Simes, 1986] (see Supplemental Materials), a method developed specifically to test a global null hypothesis H_0. This method makes the extra assumption of positive dependence among the P-values, similar to the Benjamini-Hochberg procedure [Goeman and Solari, 2014]. The Simes procedure implicitly takes these dependencies into account, yielding a less conservative P-value than the Bonferroni correction. The Simes procedure orders the P-values belonging to CpG i in ascending order, such that $p_{(1),i} \leq \cdots \leq p_{(k_i),i}$. Next, a P-value P_i for CpG

i is calculated by multiplying each $p_{(j),i}$ by a smaller factor k_i/j and taking the minimum of these corrected P-values:

$$P_i = \min \left\{ j : \frac{k_i}{j} p_{(j),i} \right\} \tag{2.2}$$

Both the Bonferroni procedure and the Simes procedure multiply the smallest P-value $p_{(1),i}$ by k_i. However, the Simes procedure multiplies the larger $p_{i,j}$ by a smaller factor, making the Simes procedure a more liberal procedure in the case of positively correlated P-values.

Estimating the CpG level false discovery proportion in a simulated setting using the bait-and-switch simulation procedure

We have discussed two approaches to compiling a list of CpGs influenced by genetic variation: the CpG-SNP pair-based approach and our CpG-based approach. We will now discuss a novel data-based simulation scheme called the bait-and-switch simulation to provide an assessment of the performance of these approaches in terms of the proportion of CpGs falsely identified as being significantly associated with genetic variation in a realistic simulation setting. Because simulation of realistic genome-wide genotype and methylation data is hard to do from scratch, we choose to modify the current data set in such a way that we have knowledge of what null hypotheses are true, i.e. which CpGs should not associate with any genetic variation. This simulation consists of several steps and is depicted in Figure 2.2A:

1. Within-window correction: perform the Simes correction within each CpG's window separately. Take the minimum adjusted P-value as the P-value for this CpG.

2. Between-windows correction: control the FDR among the newly calculated P-values to obtain a list of FDR significant CpGs.

3. The data consisting of FDR significant CpGs will be called the bait set. The rest of the data, the non-significant CpGs, are called the switch set.

4. Permute the methylation values for the switch set, leaving the data in the bait set and the genotype data intact.

5. Perform the CpG-SNP pair-based approach and the CpG-based approach on the simulated data, obtaining a list of significant CpGs for each approach.

To get an estimate of the CpG level FDR, we calculate the proportion of the CpGs obtained in step 3 coming from the switch set. Although we do not know which of the CpGs in the bait set are truly associated with genetic variation, we do know that none of the CpGs in switch set have any such association. As a result, the calculated FDP is a lower bound. The CpG level FDR is the average of the different realizations of the FDP coming from many repetitions of the same simulation experiment.

Results

CpG-SNP pair-based meQTL mapping approach

We performed *cis* meQTL mapping, varying the window size from 1 kb to 500 kb. For each window size, we applied the the CpG-SNP pair-based approach, obtaining a list of statistically significant CpG-SNP pairs and a list of the CpGs among these CpG-SNP pairs, i.e. the CpGs that are associated with at least one SNP. Despite the relatively small sample size, Figure 2.1 shows that the Benjamini-Hochberg method finds an increasing number of CpG-SNP pairs with increasing window size, with a maximum of 223,428 CpG-SNP pairs at a 200 kb window and a maximum of 10,034 CpGs at the 100 kb window size. If we keep expanding the search window around each CpG the multiple testing burden becomes too great, leading to a slight decrease in the number of CpG-SNP pairs and CpGs in that list. The increase in the number of CpG-SNP pairs can be mainly attributed to linkage disequilibrium (LD). When observing a statistically significant CpG-SNP pair, LD may virtually guarantee finding more significant CpG-SNP pairs if that SNP is strongly correlated to other nearby SNPs and we expand the window around each CpG. This is illustrated by the LocusZoom plot [Pruim et al., 2010] for a CpG (cg12247378) associated with several SNPs on 22q13.1 in Figure 2.1B. Many of the SNPs associated with this CpG are in LD and will be included with an increasing window size.

Evaluating the CpG level false discovery proportion in a simulated setting using the bait-and-switch simulation

LD causes identification of the causal SNP responsible for the variation in methylation to be impossible by statistical means. Therefore, it would be more insightful to consider individual CpGs only, instead of focusing on all CpG-SNP pairs. Following the CpG-SNP pair-based approach, we report on the CpGs from the FDR significant CpG-SNP pairs found (see Figure 2.1), i.e. the CpGs that associate with at least one SNP. However, this set of CpGs has no guarantee of FDR control and likely includes many false positive CpGs.

To evaluate the CpG level FDP among the in a controlled setting, we use the bait-and-switch simulation scheme. We construct a new, simulated data set that is very similar to the original data, but allows us to compute a lower bound on the FDP among the CpGs. Performing the CpG-SNP pair-based approach for varying

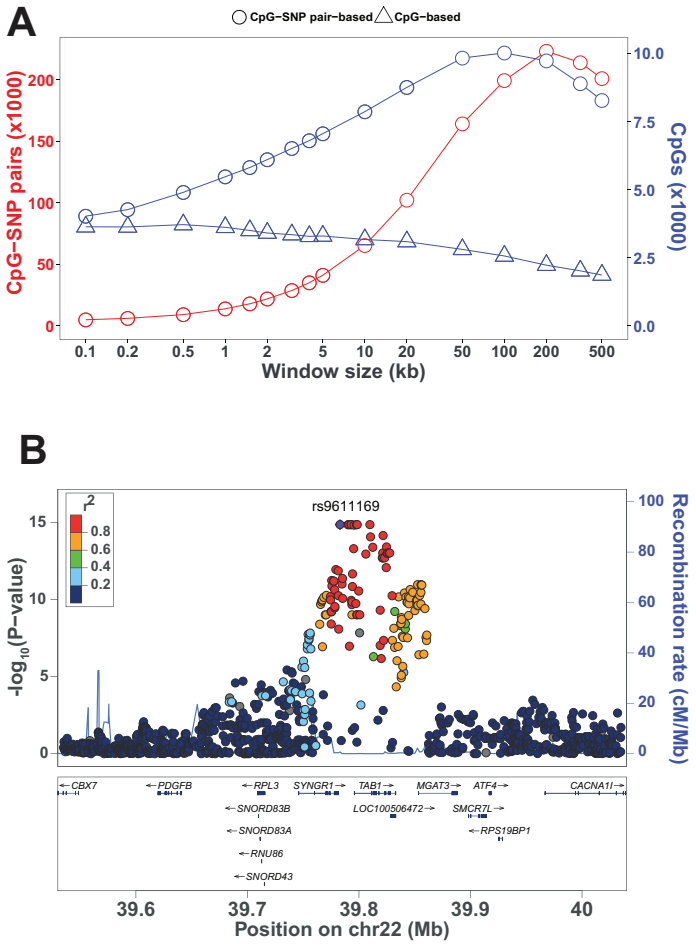

Figure 2.1: (**A**) The number of CpG-SNP pairs and the number of CpGs among them for different window sizes in the real data. The grey line shows the number of CpG-SNP pairs (FDR < 0.05). The black lines show the number of CpGs found. The two different symbols denote the CpG-SNP pair-based approach (circles) and our proposed CpG-based approach (triangles). Both the number of CpG-SNP pairs and the CpGs among them increase with window size when using the CpG-SNP pair-based approach. The CpG-based approach finds less CpGs, and reaches an optimum at at a 500 base pair window size. (**B**) CpGs associated with genetic variation are often associated with many SNPs due to LD. The LocusZoom plot shows the associations between CpG cg12247378 (22q13.1) and the SNPs in its window. The left y-axis shows the P-value corresponding to the association with the methylation levels on a -$\log_{10}$-scale, the right axis shows the recombination rate. The color coding indicates the r^2 between the SNPs, based on 1000 Genomes, build hg19. Many of the associated SNPs are in strong LD with one another.

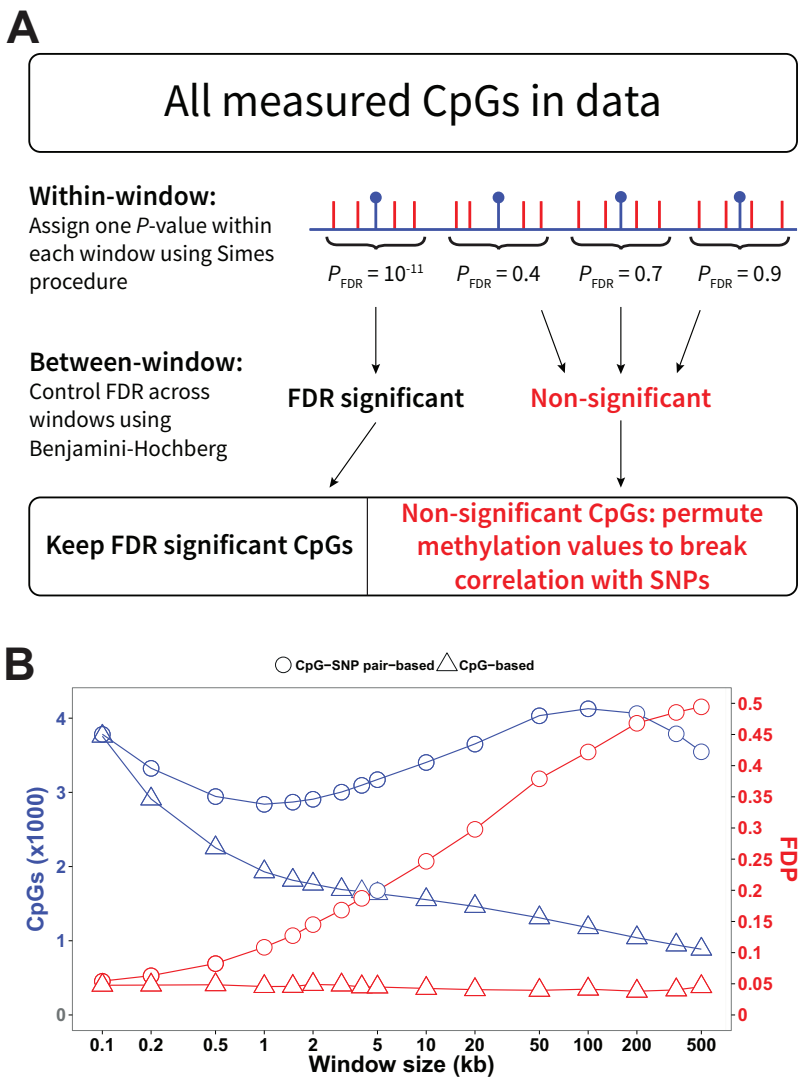

Figure 2.2: The number of CpGs found using the CpG-SNP pair-based approach, our proposed CpG-based approach and the corresponding CpG level FDP for different window sizes in the bait-and-switch simulated data. (**A**) An overview of the bait-and-switch simulation. (**B**) The grey line shows the number of CpGs. The black lines show the corresponding CpG level FDP. The two different symbols denote the CpG-SNP pair-based approach (circles) and our proposed CpG-based approach (triangles).

window sizes on the simulated data set yields a list of CpGs associated with at least one SNP and a list of all CpG-SNP pairs at an FDR of 0.05, similar to the results in Figure 2.1A. In the simulated data set we know for which CpGs we permuted the methylation values and thus are false positives (see Figure 2.2B). Strikingly, a large portion of the identified CpGs using the CpG-SNP pair-based approach seem to be false positives, especially for larger window sizes (Figure 2.2). Even when using a very small 0.5 kb window size, we find an estimated FDP of 0.1 (SE = 0.0006, based on 5 permutations), meaning at least 10% of the CpGs found among the CpG-SNP pairs in the simulated data are coming from the permuted switch set, i.e. are not truly associated with a SNP. This number greatly increases to 49.1% (FDP = 0.49, SE = 0.002, based on 5 permutations) for the 500 kb window size. While we can only claim that up to 50% of the CpGs found in the simulated data are false positives, this approach will probably yield an inflated proportion of falsely identified CpGs in the original data too. Our proposed CpG-based approach, however, controls the FDP at 0.05 (SE 0.001-0.005, based on 5 permutations) for all window sizes.

A FDR controlled list of CpGs influenced by genetic variation

To obtain a valid list of CpGs that are significantly associated with genetic variation *in cis* in the original data, we calculated one P-value per CpG, testing the global hypothesis of no association between variation in methylation any of the SNPs in its window. We calculated these P-values by means of the the Simes procedure. The Simes procedure implicitly takes into account the correlation structure among the SNPs, making it a more powerful method than, *e.g.*, the conservative Bonferroni method. After this within-window correction, we applied the Benjamini-Hochberg procedure to the resulting P-values, controlling the FDR among the CpGs to 0.05. Figure 2.1A shows that this approach identifies a maximum of 3,721 CpGs at a 500 base pair window size (black line, triangles). This suggests that strongly associated SNPs are often in close proximity to the CpG, as reported earlier [Bell et al., 2011; Gutierrez-Arcelus et al., 2013]. To show that this approach does control the FDP among the CpGs at the desired level, we again conducted the same bait-and-switch simulation experiment, applying our proposed CpG-based approach on the simulated data set. While our approach seemingly discovers fewer CpGs than the CpG-SNP pair-based approach to meQTL mapping when applied to the original data, the FDR among the CpGs identified in the simulated data is controlled at 0.05 (Figure 2.2B).

Discussion

We report on a CpG-based multiple testing approach in meQTL mapping to identify individual CpGs whose methylation level is influenced by genetic variation *in cis*. Our approach is based on the application of the Simes procedure within a window around each CpG to obtain a single P-value per CpG, followed by the Benjamini-Hochberg procedure to control the FDR across CpGs. Strikingly, this approach

suggests that optimal window sizes for the identification of *cis* meQTLs are much smaller than frequently used in the literature (up to 10s of kb instead of 100s of kb). These smaller window sizes are in line with reports that SNPs strongly associated with a CpG are often in close proximity to the CpG [Bell et al., 2011; Gutierrez-Arcelus et al., 2013]. The large window sizes used in literature may stem from the CpG-SNP pair-based approach reporting on the CpGs from a list of all FDR significant CpG-SNP pairs. Using the bait-and-switch simulation we show that the latter approach yields up to 50% falsely identified CpGs in simulated data. Our proposed approach controls the CpG level FDR at the desired level and still identifies a substantial number of CpGs associated with genetic variation *in cis*.

Our method can be directly applied to the output of commonly used QTL mapping software, *e.g.*, Matrix EQTL, which returns P-values corresponding to every CpG-SNP pair tested. In addition, the current method does not require the use of permutations to control the FDR, making it a fast and easy-to-use approach. While permutations are still feasible for small 450k array data sets, this becomes burdensome for large data sets, particularly when using bisulphite-sequencing data measuring millions of CpG sites.

When calculating one P-value for the window around each CpG site it is important to account for LD between SNPs in the window. Not doing so will substantially reduce statistical power. Therefore, some methods, like the Bonferroni correction, may be too conservative. The Simes procedure implicitly takes LD into account by multiplying larger P-values with smaller factors. Although the Simes procedure seems to perform well in terms of CpGs found, it still does not fully capture the correlation structure. A possible solution would be to estimate the number of independent tests for each window, *e.g.*, using GATES [Li et al., 2011] or TATES [van der Sluis et al., 2013], accounting for the number of independent tests done. However, this may be computationally expensive. Our proposed approach is unable to distinguish between two independent SNP effects on the methylation levels of a CpG. It only allows for making claims about the global null hypothesis of no association with any genetic variant *in cis*. This approach takes into account that the causal variant cannot be identified with statistical means only. Another limitation is that there currently is no valid method to determine the optimal window size for a study prior to QTL-mapping. In general, the optimal window size will be greater for studies with higher statistical power. Our study suggests that the optimal window size will be 10-50 kb instead of the commonly used 100s kb , which will reduce statistical power by dramatically increasing the number of tests.

In this paper we introduced the bait-and-switch simulation method to estimate the true false discovery proportion among CpGs with a meQTL *in cis* in simulated data. This approach indicated up to 50% of identified CpGs in our simulated data may be false positive. While we know this is true in the simulated data, we cannot extrapolate this to the original data. It is likely that the common approach to multiple testing also brings about an increased CpG level FDR in the real data. This finding may also be an issue for *cis* expression QTL studies and possibly *trans* QTL studies. Interpretation of results based on the common approach evaluated here should be interpreted with caution.

Development of statistical methodology will aid in getting a complete catalogue of meQTLs and eQTLs that is key in understanding the mechanisms underlying the association of non-coding genetic variants with disease phenotypes.

Acknowledgements

This work was done within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

References

Bell, J. T. et al. [2011]. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines, *Genome Biol* **12**(1): R10.

Benjamini, Y. and Hochberg, Y. [1995]. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J Roy Stat Soc B Met.* pp. 289–300.

Bibikova, M. et al. [2011]. High density dna methylation array with single cpg site resolution, *Genomics* **98**(4): 288–295.

Chen, Y. et al. [2013]. Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray, *Epigenetics* **8**(2): 203.

Davis, S. et al. [2013]. Methylumi: Handle illumina methylation data 2012, *R package* **2**(1).

Drong, A. W. et al. [2013]. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of dna methylation in adipose tissue, *PLoS One* **8**(2): e55923.

Du, P. et al. [2010]. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis, *BMC Bioinformatics* **11**(1): 587.

Feero, W. G. et al. [2010]. Genomewide association studies and assessment of the risk of disease, *N Engl J Med* **363**(2): 166–176.

Gibbs, J. R. et al. [2010]. Abundant quantitative trait loci exist for dna methylation and gene expression in human brain, *PLoS Genet* **6**(5): e1000952.

Goeman, J. J. and Solari, A. [2014]. Multiple hypothesis testing in genomics, *Stat Med* **33**(11): 1946–1978.

Grundberg, E. et al. [2013]. Global analysis of dna methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements, *Am J Hum Genet* **93**(5): 876–890.

Gutierrez-Arcelus, M. et al. [2013]. Passive and active dna methylation and the interplay with genetic variation in gene regulation, *eLife* **2**.

Heijmans, B. T. et al. [2007]. Heritable rather than age-related environmental and stochastic factors dominate variation in dna methylation of the human igf2/h19 locus, *Hum Mol Genet* **16**(5): 547–554.

Heyn, H. et al. [2013]. Dna methylation contributes to natural human variation, *Genome Res* **23**(9): 1363–1372.

Howie, B. N. et al. [2009]. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet* **5**(6): e1000529.

Li, M.-X. et al. [2011]. Gates: a rapid and powerful gene-based association test using extended simes procedure, *Am J Hum Genet* **88**(3): 283–293.

Liu, Y. et al. [2013]. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis, *Nat Biotechnol* **31**(2): 142–147.

Mill, J. and Heijmans, B. T. [2013]. From promises to practical strategies in epigenetic epidemiology, *Nat Rev Genet* **14**(8): 585–594.

Niu, N. et al. [2010]. Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines, *Genome Res* **20**(11): 1482–1492.

Pearson, T. A. and Manolio, T. A. [2008]. How to interpret a genome-wide association study, *JAMA* **299**(11): 1335–1344.

Pruim, R. J. et al. [2010]. Locuszoom: regional visualization of genome-wide association scan results, *Bioinformatics* **26**(18): 2336–2337.

Shabalin, A. A. [2012]. Matrix eqtl: ultra fast eqtl analysis via large matrix operations, *Bioinformatics* **28**(10): 1353–1358.

Shi, J. et al. [2014]. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue, *Nat Commun* **5**.

Simes, R. J. [1986]. An improved bonferroni procedure for multiple tests of significance, *Biometrika* **73**(3): 751–754.

Small, K. S. et al. [2011]. Identification of an imprinted master trans regulator at the klf14 locus related to multiple metabolic phenotypes, *Nat Genet* **43**(6): 561–564.

van der Sluis, S. et al. [2013]. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies, *PLoS Genet* **9**(1): e1003235.

van Eijk, K. R. et al. [2012]. Genetic analysis of dna methylation and gene expression levels in whole blood of healthy human subjects, *BMC Genomics* **13**(1): 636.

Visscher, P. M. et al. [2012]. Five years of gwas discovery, *Am. J. Hum. Genet.* **90**: 7–24.

Wagner, J. R. et al. [2014]. The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts, *Genome Biol* **15**(2): R37.

Westra, H. et al. [2011]. Mixupmapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects, *Bioinformatics* **27**(15): 2104–2111.

Westra, H. et al. [2013]. Systematic identification of trans eqtls as putative drivers of known disease associations, *Nat Genet* **45**(10): 1238–1243.

Zhang, D. et al. [2010]. Genetic control of individual differences in gene-specific methylation in human brain, *Am J Hum Genet* **86**(3): 411–419.