Clinical Epigenetics

METHODOLOGY   **Open Access**

Gene–methylation interactions: discovering region-wise DNA methylation levels that modify SNP-associated disease risk

Julia Romanowska[1,2,3*] , Øystein A. Haaland[1], Astanand Jugessur[1,3,4], Miriam Gjerdevik[1,4], Zongli Xu[5], Jack Taylor[5], Allen J. Wilcox[5], Inge Jonassen[2], Rolv T. Lie[1,3] and Håkon K. Gjessing[1,3]

Abstract

Background: Current technology allows rapid assessment of DNA sequences and methylation levels at a single-site resolution for hundreds of thousands of sites in the human genome, in thousands of individuals simultaneously. This has led to an increase in epigenome-wide association studies (EWAS) of complex traits, particularly those that are poorly explained by previous genome-wide association studies (GWAS). However, the genome and epigenome are intertwined, e.g., DNA methylation is known to affect gene expression through, for example, genomic imprinting. There is thus a need to go beyond single-omics data analyses and develop interaction models that allow a meaningful combination of information from EWAS and GWAS.

Results: We present two new methods for genetic association analyses that treat offspring DNA methylation levels as environmental exposure. Our approach searches for statistical interactions between SNP alleles and DNA methylation (G×Me) and between parent-of-origin effects and DNA methylation (PoO×Me), using case-parent triads or dyads. We use summarized methylation levels over nearby genomic region to ease biological interpretation. The methods were tested on a dataset of parent–offspring dyads, with EWAS data on the offspring. Our results showed that methylation levels around a SNP can significantly alter the estimated relative risk. Moreover, we show how a control dataset can identify false positives.

Conclusions: The new methods, G×Me and PoO×Me, integrate DNA methylation in the assessment of genetic relative risks and thus enable a more comprehensive biological interpretation of genome-wide scans. Moreover, our strategy of condensing DNA methylation levels within regions helps overcome specific disadvantages of using sparse chip-based measurements. The methods are implemented in the freely available R package Haplin (https://cran.r-project.org/package=Haplin), enabling fast scans of multi-omics datasets.

Keywords: Integrative analysis, Statistical interaction effect, DNA methylation, Genome-wide data, Parent-of-origin, Case-parent triads, Haplin

*Correspondence: Julia.Romanowska@uib.no
[1] Department of Global Public Health and Primary Care, University of Bergen, N-5020, Bergen, Norway
[2] Computational Biology Unit, University of Bergen, N-5020, Bergen, Norway
[3] Centre for Fertility and Health, Norwegian Institute of Public Health, N-0213, Oslo, Norway
Full list of author information is available at the end of the article

Background

Genome-wide association studies (GWAS) of single-nucleotide polymorphisms (SNPs) have contributed enormously to our understanding of the genetic underpinnings of various complex diseases. However, it has become increasingly clear that the heritability of a disease cannot be fully explained by GWAS alone, prompting researchers to examine rare variants, other types of omics data, and alternative disease mechanisms in an attempt to explain the *missing heritability*.

The term "epigenome" is widely used to encapsulate all the epigenetic processes involved in regulating gene expression in the entire genome. Epigenome-wide association studies (EWASes) have become a relatively common complementary omics to GWAS as a result of major advances in high-throughput microarray-based technologies for measuring DNA methylation (DNAm) [1, 2]. Among several epigenetic modifications characterized to date, DNAm is by far the most studied epigenetic mark in humans. It is a process by which a methyl group binds to the cytosine (C) at a CpG dinucleotide, resulting in activation or repression of gene expression through mechanisms that are highly region- and context-dependent [3–5]. Furthermore, it has been shown that the state of methylation is controlled by several enzymes [6] and is influenced by environmental exposures [7].

A standard GWAS analysis for a dichotomous phenotype computes relative risks (RRs) between all SNPs and the phenotype. However, it is easy to envision that the effect of a SNP on the phenotype can be modified by DNAm levels in the nearby regions, for instance, when DNAm affects the gene expression. In statistical terms, this corresponds to finding an interaction between the SNP and nearby CpG methylation levels. For instance, the RRs may differ depending on whether DNAm levels are low, medium, or high. Here, we refer to this as a gene–methylation interaction effect (G×Me).

We are aware of only two studies in the literature that have explored G×Me effects [8, 9]. The authors of those papers analyzed several SNPs and one CpG in a candidate gene for asthma and detected a statistically significant interaction between a specific SNP–CpG pair. The relative risk of asthma associated with the SNP increased with an increasing level of methylation at the CpG site. Both studies were, however, limited in that they only investigated a few SNPs in one gene and only a single CpG, which were selected a priori because they all showed significant associations with the phenotype. Developing an efficient method that could be applied to the entire genome and epigenome would thus advance the field substantially by enabling an agnostic search for G×Me interactions.

In addition to G×Me interactions, it is important to consider parent-of-origin (PoO) effects, which may account for a fraction of the unexplained heritability of a trait. Here, we define a PoO effect as the effect of a particular allele in the child depending on whether the allele is inherited from the mother or the father; see the references and discussion in our previous work [10]. While the main genetic and gene–environment (G×E) effects can be estimated using a case–control design, assessing a PoO effect requires genetic information from at least one parent, thus a dyad or triad design [11]. An advantage of these family-based designs is that it is possible to estimate G×E effects even when only the genotypes of the case families are available [12]. The ability to estimate PoO effects opens an entirely unexplored possibility, namely to study how DNAm levels influence PoO effects. Previously, we have developed models for estimating PoO×E effects in GWAS analyses, where E denotes an external environmental exposure [10, 13, 14]. By letting DNAm levels take the role of the exposure variable, we can thus evaluate PoO×Me effects, i.e., we can study how the PoO effect changes depending on the nearby DNAm levels.

While G×Me effects are interpreted analogously to those derived from a cohort or case–control setting, the PoO×Me interactions offer an intriguing extension. Since genetic imprinting can lead to a PoO effect for a phenotype association, methylation levels at CpGs located near a SNP exhibiting a PoO effect may influence the magnitude of that effect.

In this paper, we treat the level of DNAm as environmental exposure and develop new statistical methods to estimate G×Me and PoO×Me interactions from case-parent triads and dyads in a full GWAS-EWAS setting. Our implementation of the methods is applicable to any dichotomous trait where genotypes and methylation status are available on cases (affected children), and the genotypes of at least one parent can be obtained. For each SNP, we investigate DNAm in various genomic regions and not only at a single CpG site because correlated CpGs are known to exert their effects over long stretches of DNA [15–17].

G×E analyses using case-parent triads are generally robust in the sense that they only require genes and environment to be statistically independent, *conditional on* parental genotypes [18]. This somewhat technical condition would usually be satisfied when E represents an external environmental factor. However, replacing E with Me may be problematic in a few cases, such as when the SNP and a nearby CpG represent an meQTL pair, i.e., where the SNP exerts direct influence on the CpG methylation levels, and thus violates the independence assumption. While this is less of a problem for PoO×Me interactions, we show how control triads can be used to resolve this issue for G×Me analyses.

We showcase our methods on orofacial clefts (OFC), which are relatively common congenital malformations with a high heritability and recurrence risk [19]. Several

Romanowska *et al. Clinical Epigenetics*   (2020) 12:109

Page 3 of 18

GWASes on OFC have been published, confirming previously reported genes and loci for OFC and identifying new ones for further investigation (for a review, see [20]). Besides the genetic variants, environmental factors have been shown to influence the risk of OFC [21, 22]. A recent study using Mendelian randomization showed that DNAm might mediate genetic liability to clefting [23] and suggested that the causal pathway might proceed in the following direction: environment → *DNAm* → OFC. In our proof-of-concept analyses described here, we analyzed a well-vetted set of genetic variants known to be strongly associated with OFC risk [24]. This increases the chance of identifying any weaknesses in our theory and minimizes the multiple-testing burden.

Results

Implementation of the methods

As mentioned, the main idea behind our methods is to treat the level of DNAm as environmental exposure in the G×E or PoO×E analyses to estimate the change of RR of a disease, depending on DNAm levels. The workflow of the method is presented in Figs. 1 and 2, while the details of the modeling framework, implementation, and analyses are presented in the "Methods" section and in Additional file 2. Below, we give a brief overview of the implementation.

We start by summarizing the DNAm levels from several CpG sites within a specific region located near a given SNP (see Fig. 1). To ease the biological interpretation of the results, we focus on three types of genomic regions: promoter, enhancer, and gene body (for detailed definitions, see the "Genomic regions" section). Because the implementation does not depend on region type, the methods are equally well applicable to, e.g., CpG-rich regions or single CpG sites.
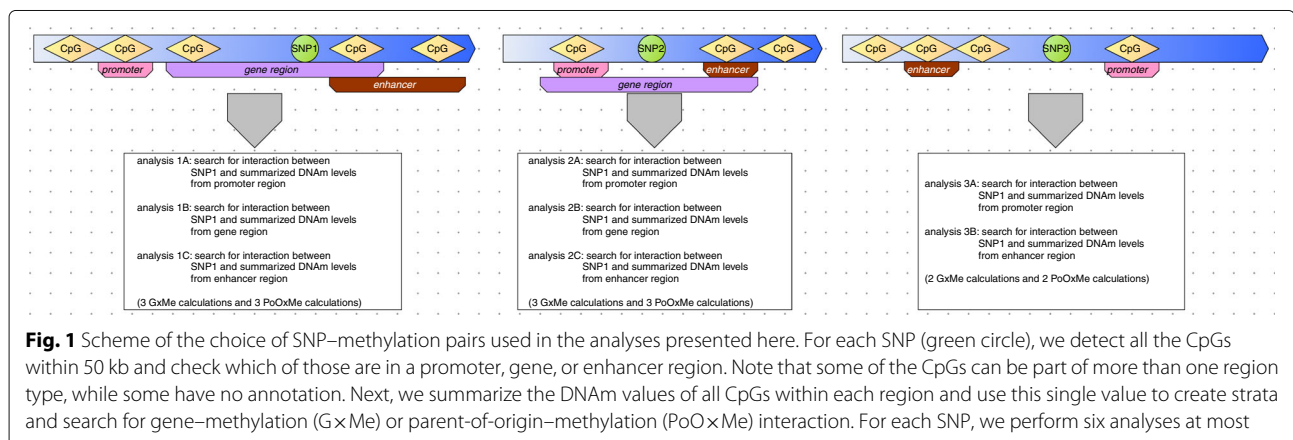
Next, to create discrete categories of the continuous DNAm levels, we average the DNAm levels of the CpGs within each region and divide the individuals into three equally sized strata, based on whether their average
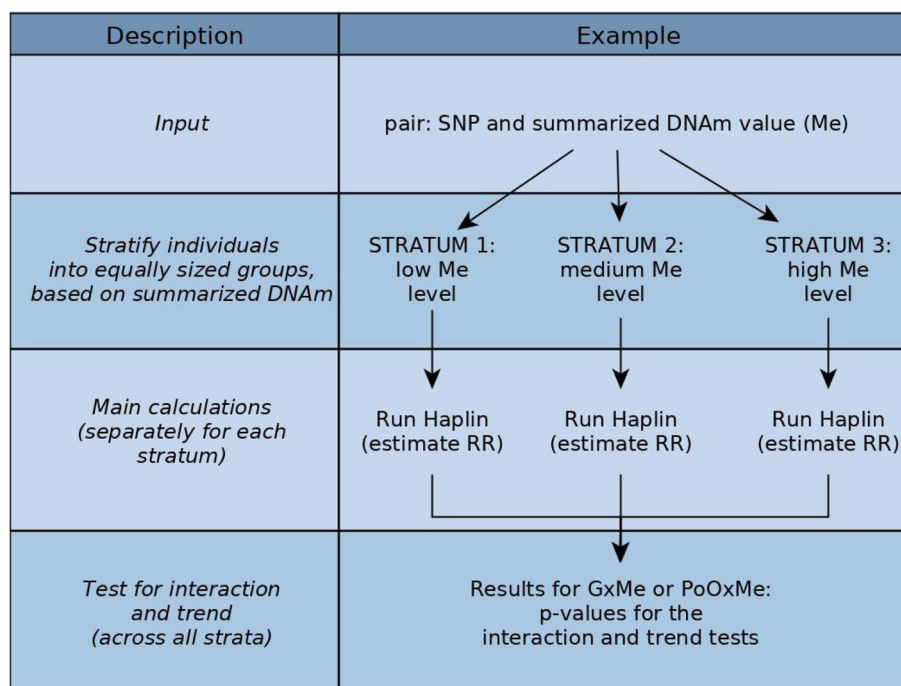
DNAm level in a specific region is low, medium, or high. In the "Discussion" section, we show that our method of summarizing the DNAm levels helps to retain important information carried by each CpG site and offsets the disadvantages of using sparse, chip-based measurements of DNAm.

For each DNAm stratum, we estimate the RR of the SNP relative to the trait and then test for trend and interaction between strata (see Fig. 2 and the "Implementation details" section). The interpretation of the G×Me analysis is that a statistically significant change in RR across strata would indicate an interaction. Similarly, for the PoO×Me analyses, we perform PoO analyses for each stratum and then check for a significant change across strata.

Application of the methods

To test our methods, we apply them to genotype data from mother–child dyads and genome-wide DNAm data from the children only. These data are available on controls and cases. Cases are children diagnosed with OFC and divided into following subsets: cleft lip only (CLO), cleft lip with cleft palate (CLP), cleft lip with or without cleft palate (CL/P), or cleft palate only (CPO). We focus the current analyses on the subset of SNPs that showed the strongest associations with OFC risk in a recent study that used the same genetic dataset as here [24] (Table 1). Depending on whether there are any CpGs near a SNP that could be categorized as belonging to one of promoters, enhancers, or genes (Table 2), we conduct the G×Me and PoO×Me analyses up to three times for each SNP in the case triads. We also repeat the analyses on control data to see if there are any background correlations between SNPs and DNAm. This would indicate false-positive results due to, e.g., the presence of meQTLs. In the sections below, we present the most significant results. All the *p* values are provided in Tables S1–S5 and S7–S11 in Additional File 1.

Fig. 1 Scheme of the choice of SNP–methylation pairs used in the analyses presented here. For each SNP (green circle), we detect all the CpGs within 50 kb and check which of those are in a promoter, gene, or enhancer region. Note that some of the CpGs can be part of more than one region type, while some have no annotation. Next, we summarize the DNAm values of all CpGs within each region and use this single value to create strata and search for gene–methylation (G×Me) or parent-of-origin–methylation (PoO×Me) interaction. For each SNP, we perform six analyses at most

Fig. 2 Flowchart of the method for integrating DNA methylation information into genetic association analyses. *RR* relative risk, G×Me gene–methylation interaction, PoO×Me parent-of-origin–methylation interaction

G×Me analyses

The results of the G×Me analyses are easiest to evaluate when the Wald interaction test *p* values are plotted as quantile–quantile (Q-Q) plots (Fig. 3). The Q-Q plots for CLO and CPO showed no significant interactions. The results for CL/P pointed to a possible interaction between rs12543318 at 8q21.3 and the methylation state

of the promoter-flanking region nearby (one CpG, regulatory stable ID: ENSR00001432551 for GRCh37 and ENSR00000861245 for GRCh38). The *p* value of 0.012 indicates a change in the relative risk depending on the methylation status in that region. Figure 4a also shows that the risk of CL/P changes markedly among individuals carrying at least one C-allele at rs12543318 and whose

Table 1 The SNPs selected for the current analyses, along with the names of the nearest genes (if any), and the measures of association (relative risks (RR), 95% confidence intervals (CI), and *p* values; taken from Tables 1–3 in Ref. [24])

SNP	Locus	Minor allele[a]	MAF[a]	RR	95% CI	*p* value	Cleft subtype[b]
rs12543318	*8q21.3*	c	0.31	1.51	1.31–1.75	$4.46e^{-8}$	CLO
rs987525	*8q24*	a	0.19	1.85	1.63–2.10	$1.47e^{-19}$	CLP
rs560426	*ABCA4*	g	0.44	1.24	1.10–1.41	$3.91e^{-4}$	CLP
rs3758249	*FOXE1*	t	0.36	0.78	0.69–0.88	$8.18e^{-5}$	CLP
rs642961	*IRF6*	a	0.23	1.60	1.36–1.87	$1.41e^{-8}$	CLO
rs7078160	*KIAA1598*	a	0.17	1.33	1.15–1.53	$1.04e^{-4}$	CLP
rs13041247	*MAFB*	c	0.41	0.67	0.59–0.76	$2.32e^{-9}$	CLP
rs227731	*NOG1*	g	0.47	0.74	0.64–0.85	$3.8e^{-5}$	CPO
rs742071	*PAX7*	t	0.38	1.52	1.31–1.75	$3.74e^{-8}$	CLO
rs8001641	*SPRY2*	g	0.54	0.79	0.70–0.90	$2.05e^{-4}$	CLP
rs7590268	*THADA*	g	0.25	1.27	1.12–1.46	$3.96e^{-4}$	CLP
rs1873147	*TPM1*	g	0.27	1.31	1.13–1.53	$5.82e^{-4}$	CLO

[a]The minor allele and its frequency (MAF) for the Norwegian population were taken from Table 1 in the Appendix of Ref. [24]
[b]The cleft subtype (CLO, CPO, or CLP) for which the association was the strongest

Romanowska *et al. Clinical Epigenetics*   (2020) 12:109

Page 5 of 18

Table 2 Availability of data for G×Me and PoO×Me analyses. We used genotypes and DNA methylation data and classified CpGs into one of gene, promoter, or enhancer classes. In each of those groups, we performed both G×Me and PoO×Me analyses, where at least one CpG was localized within 50 kb from the indicated SNP (gray shading; the number indicates how many CpGs were in each category)

SNP (gene/locus)	Enhancer	Promoter	Gene body
rs12543318 (*8q21.3*)		1	
rs987525 (*8q24*)	1	1	
rs560426 (*ABCA4*)	6	1	16
rs3758249 (*FOXE1*)		6	7
rs642961 (*IRF6*)		14	27
rs7078160 (*KIAA1598*)	1		1
rs13041247 (*MAFB*)		21	9
rs227731 (*NOG1*)		2	
rs742071 (*PAX7*)	1	19	54
rs8001641 (*SPRY2*)	1	1	
rs7590268 (*THADA*)	1	1	14
rs1873147 (*TPM1*)		30	29

methylation status at the CpG in question was in the middle stratum (i.e., between 0.971 and 0.975). This group had a higher risk of CL/P when the methylation level was taken into account (i.e., in the stratum "2" in Fig. 4a) versus when methylation level was not considered (no stratification, results for "all" in Fig. 4a). Thus, it is plausible that the interaction is driven by the change in RR in the middle stratum. As illustrated in Fig. 4b, the result of the stratified analysis of the control dataset did not show any trend and was not significant.

The CLP analysis resulted in one significant interaction between rs3758249 in *FOXE1* and the methylation status of the gene region (Fig. 3, a *p* value of 0.001). Intriguingly, we found the same significant interaction in the control dataset, and the pattern was similar in the analyses of cases and controls (Fig. 5). Therefore, this interaction is most likely a false positive. The results of G×Me analysis on the control dataset produced another significant interaction, between rs1873147 in *TPM1* and the methylation status of a promoter region nearby (see Fig. S9 in Additional File 1). This was not replicated in any of the case datasets, which could be in part due to sample-size issues (the control dataset is approximately 3 times larger than the case datasets).

Random-SNP analysis We checked whether the assumptions of the G×E modeling framework are met when using DNAm data as exposure by randomly picking 20 SNPs from the CL/P data (Table S6 in Additional File 1). We then applied the G×Me procedure on this dataset to see if there were any false-positive results. As

expected, these analyses did not produce any significant *p* values (Fig. 6).

PoO×Me analyses
There were a few borderline significant PoO×Me effects in the CLO, CLP, and CL/P datasets (Fig. 7). The most interesting result was the interaction between the PoO effect of the allele at rs227731, located near *NOG1*, and the methylation status of the promoter-flanking region nearby (ensembl regulatory ID ENSR00001131154 for version GRCh37 and ENSR00000559290 for version GRCh38). It is one of the most significant results among the CL/P, CLP, and CLO datasets (Fig. 7). The results displayed in Fig. 8 showed that the risk of CL/P was altered among individuals who had inherited at least one G-allele at rs227731 from the mother and in whom the methylation level in the promoter-flanking region was in the stratum "2" (i.e., average β values between 0.51 and 0.52). This group had a higher risk of CL/P compared to the results of the analysis without stratification (the "all" group). Similar patterns were observed for CLO and CLP (Figs. S11a and S12b in Additional File 1).

PoO scan We also conducted a scan over all the SNPs in the CLP dataset to search for any significant PoO effects. The top 20 hits from the scan had PoO *p* values in the range $2.1 \cdot 10^{-3}$ to 0.27 (see Table S12 in Additional File 1). We then performed PoO×Me analyses on these top 20 hits in both the CLP and the control datasets (Fig. 9). Only one SNP had a more significant *p* value than would be expected by chance. Namely, the *p* value for the interaction between the PoO effect of rs766325 and the methylation level in the promoter region nearby was above the 95% point-wise confidence interval. The RR increased with increasing methylation level of the promoter (ensembl regulatory stable ID ENSR00000923082) (see Fig. S13 in Additional File 1).

Discussion
We present here two methods that incorporate DNAm data into genetic association analyses, which we call G×Me and PoO×Me. Although a simpler implementation of the method for G×Me has previously been reported [8], our setup is substantially more comprehensive in scope. We also designed a novel PoO×Me analysis that requires fewer assumptions than G×Me. We applied the two methods on genotype and DNAm data from children with OFC and their mothers, as well as from control children and their mothers. Both methods are designed and implemented to facilitate a full genome-wide screening, but for illustration, we conducted a more targeted analysis where we focused on a set of SNPs that had previously shown significant associations with OFC risk. Thus, any change in the relative risk estimate due to the DNAm
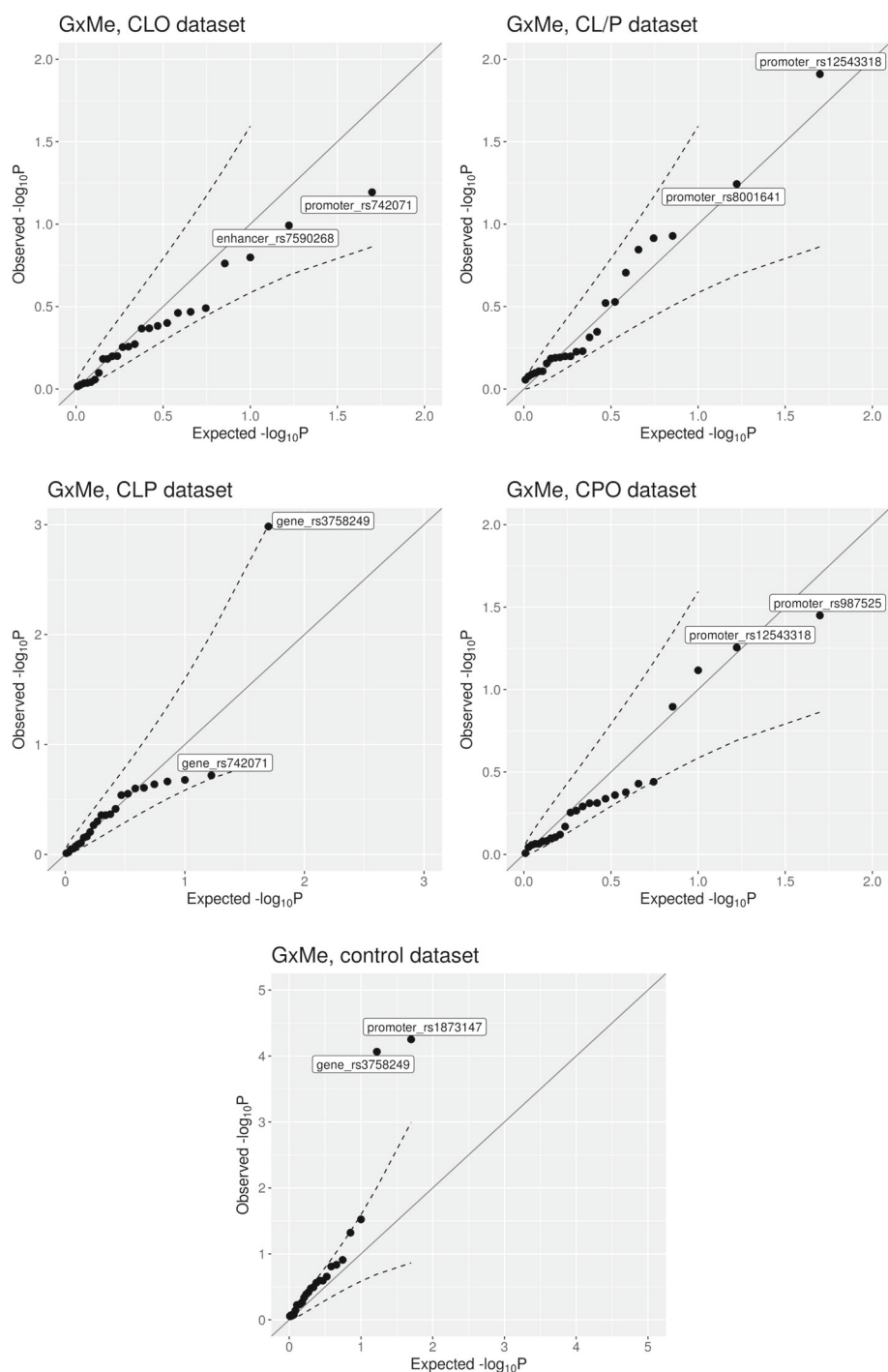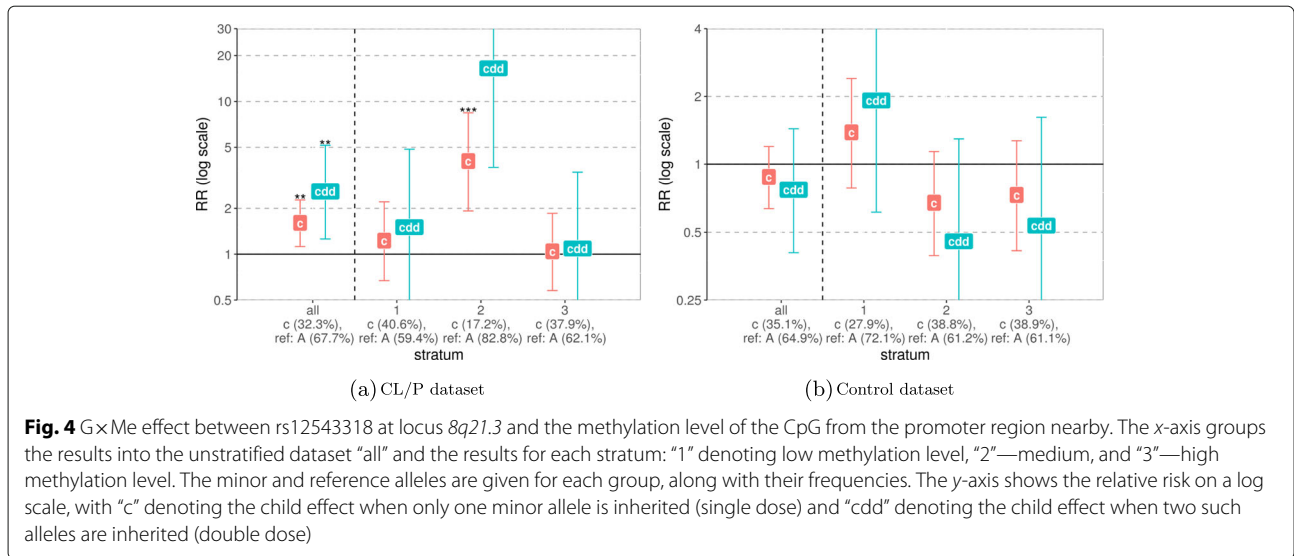
Fig. 3 Quantile–quantile plots of the interaction *p* values from the G×Me analyses. The dashed lines represent the 95% confidence interval

level should be easily detectable. Note that when judging the significance of the results after running our methods in a full genome scan mode, one can apply any standard multiple-testing method, typically based on controlling the false-discovery rate. This is because both methods return one *p* value per SNP; we do not consider SNP×SNP

interactions, which would produce a large number of correlated test results.

There are many ways of combining genetic and methylation data to predict disease prevalence or risk. For example, Shah et al. [25] tested several linear models to investigate the extent to which disease prevalence was

Fig. 4 G×Me effect between rs12543318 at locus *8q21.3* and the methylation level of the CpG from the promoter region nearby. The *x*-axis groups the results into the unstratified dataset "all" and the results for each stratum: "1" denoting low methylation level, "2"—medium, and "3"—high methylation level. The minor and reference alleles are given for each group, along with their frequencies. The *y*-axis shows the relative risk on a log scale, with "c" denoting the child effect when only one minor allele is inherited (single dose) and "cdd" denoting the child effect when two such alleles are inherited (double dose)
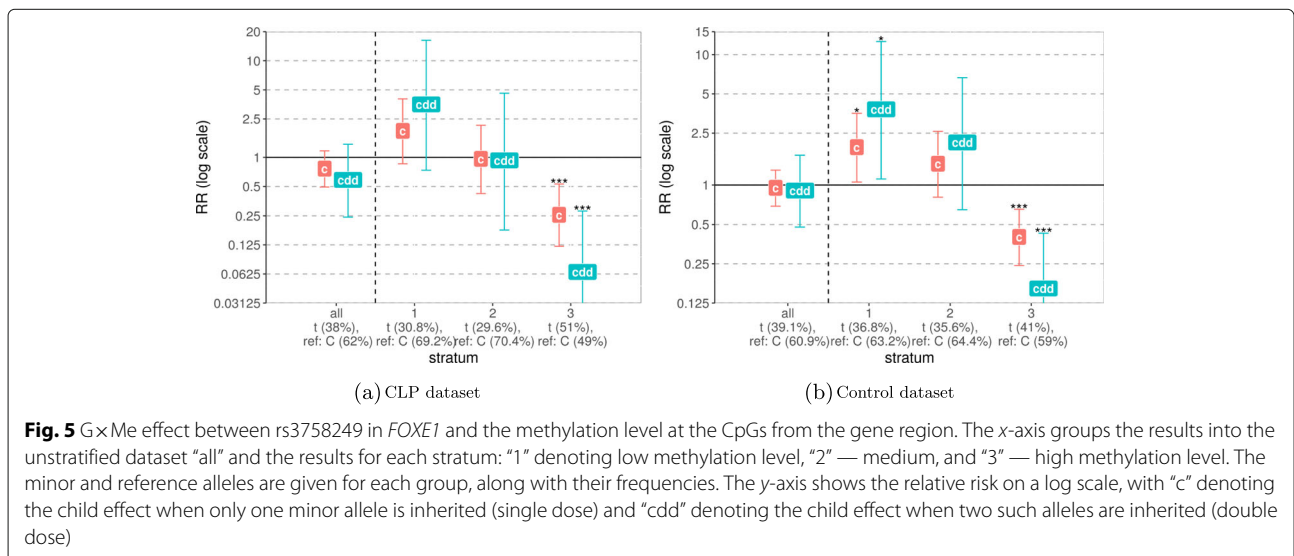
due to genetic or methylation components alone. Further, White et al. [26] performed a multi-step analysis of several omics data to identify genetic associations with neurological disorders. In another work, linear regression models were tested to identify associations between DNA methylation levels at specific CpG sites and child's birth weight [27]. These results were subsequently used to test for association between the top CpGs and genotypes. A few studies searched for an association between a specific genotype and DNAm at a chosen CpG from a neighboring region [28–30]. Another study adopted a more integrative approach by utilizing all the available genetic and methylation data to search for differentially methylated CpG sites and meQTLs associated with breast cancer [31].
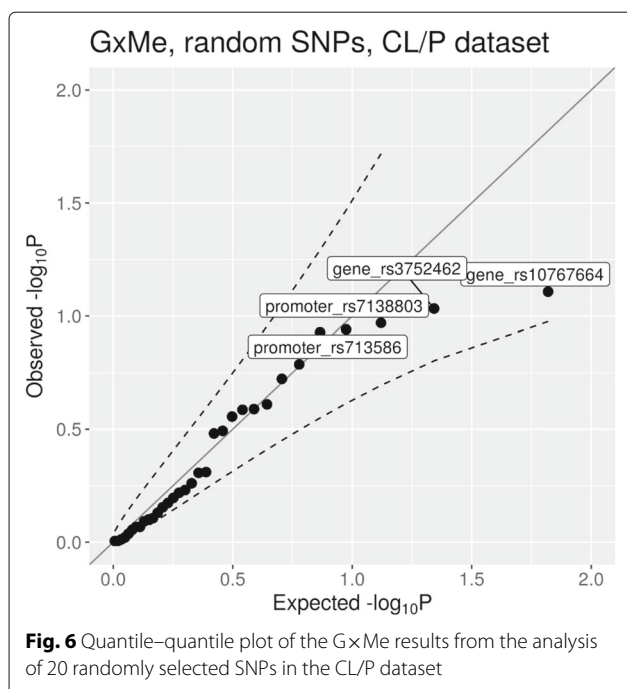
In contrast to the research mentioned above, our methods focus on dyad- or triad-based data, which are robust study designs that are particularly well-suited for early-onset disorders where biospecimens can be collected from both parents and their offspring. Moreover, we use CpG regions, and not only single CpG sites, to perform region-based CpG analyses to capture biologically relevant effects of DNAm on genotype. We implemented our new methods in the well-established R package Haplin. Our entire code is available to other researchers interested in testing and adapting the interaction models to their analyses (see the "Availability of data and materials" section below).

Summarizing the methylation patterns in regions facilitates the biological interpretation of results

We assessed the DNAm level of CpG regions instead of single CpGs and used the most straightforward method to average the β values from all the CpG sites within a region.

Fig. 5 G×Me effect between rs3758249 in *FOXE1* and the methylation level at the CpGs from the gene region. The *x*-axis groups the results into the unstratified dataset "all" and the results for each stratum: "1" denoting low methylation level, "2" — medium, and "3" — high methylation level. The minor and reference alleles are given for each group, along with their frequencies. The *y*-axis shows the relative risk on a log scale, with "c" denoting the child effect when only one minor allele is inherited (single dose) and "cdd" denoting the child effect when two such alleles are inherited (double dose)

GxMe, random SNPs, CL/P dataset

Fig. 6 Quantile–quantile plot of the G×Me results from the analysis of 20 randomly selected SNPs in the CL/P dataset

This single value from each region was then used to divide the study participants into equally sized groups according to the methylation level of each region. Although each CpG can have its unique methylation status, methylation levels of nearby CpGs have been shown to be correlated [15, 32]. Moreover, averaging the methylation level over a region has previously been successfully used to search for differentially methylated regions (DMRs) [33], when imputing missing values [34], or when defining a methylation score for a region [35]. If a given CpG displays a much wider range of methylation levels than other CpGs within the same region, this CpG will have the largest influence on the average methylation value.

However, the strategy of summarizing the methylation pattern in a region has one important caveat. For instance, let us assume that there are only two CpGs in a given region, and that, for two individuals, the β values at these CpG sites are 0.75 and 0.25, and 0.25 and 0.75, respectively. Then these two individuals would be placed in the same methylation category because the sum is 1 in both instances. We observed such a phenomenon with the most significant result of the PoO×Me analyses; namely, the interaction between rs227731 and DNAm levels of the promoter region nearby. There were only two CpGs in this region, and both had narrow and opposite β value ranges (see Fig. S3). To check whether this was problematic for our methods, we plotted the β values for each individual at each of the CpG sites separately (Fig. S14 in Additional File 1). The plots showed that there was still differential grouping of the individuals into strata, with both of the CpGs exhibiting a wide range of values

across all the strata. We then ran the PoO×Me analyses only for rs227731 and took only the β values of one of the promoter-CpGs, i.e., ch.17.52148184R. Table S13 and Fig. S15 show that the results did not change appreciably when we excluded one CpG. Notably, the influence of the methylation values on RR was the same (compared with Figs. 8, S11a, and S12b), whereas the p values were slightly higher.

Thus, averaging the methylation level of CpGs appears to be effective at summarizing the methylation level in the context of the analyses presented here. In future work, we will investigate the impact of different parameter choices and assumptions on our methods and explore how the results might be influenced by the use of a different methylation-summarizing method. Importantly, our implementation of the new methods makes them easily applicable to any other methylation-summarizing method, also concerning other genomic regions.

We chose 50 kb as the maximum distance from a given SNP to define the "nearby CpG sites" and to incorporate the promoter and gene regions in the search. Enhancers, in particular, are known to exert their effects across long distances, but the majority of enhancer–gene pairs are still located within 50 kb [36, 37]. In future developments of the methods, we will perform a more exhaustive sensitivity analysis of those parameters.

Our choice of CpGs was guided by the desire to explain the results in a biologically meaningful context. DNA methylation controls gene expression by allowing or preventing specific transcription factors to bind to promoters, enhancers, or gene bodies depending on the biological context [3, 4, 32, 38], which is why we specifically selected these three regions for our analyses here. Anastasiadi et al. [17] investigated changes in gene expression that are associated with changes in the range of DNAm within promoters, gene bodies, and gene body sub-regions. Their results indicated that, when the association is significant, one can use the mean or median value of methylation instead of the entire set of DNAm values. As a gene body may span a large region, it may potentially house so-called cryptic promoters (see, e.g., Ref.[6]), which can regulate the expression of other genes when their methylation status changes. Therefore, taking into account all CpGs within a gene region might lead to increased statistical power.

Summarizing the methylation patterns in regions may overcome the disadvantages of using data from a standard chip

Because our analyses are based on DNAm data from a microarray chip, we do not have data on all CpG sites within each of the regions considered. However, reassuringly, recent extensive genome-wide analyses of DNAm data suggest that chip-based methods may provide
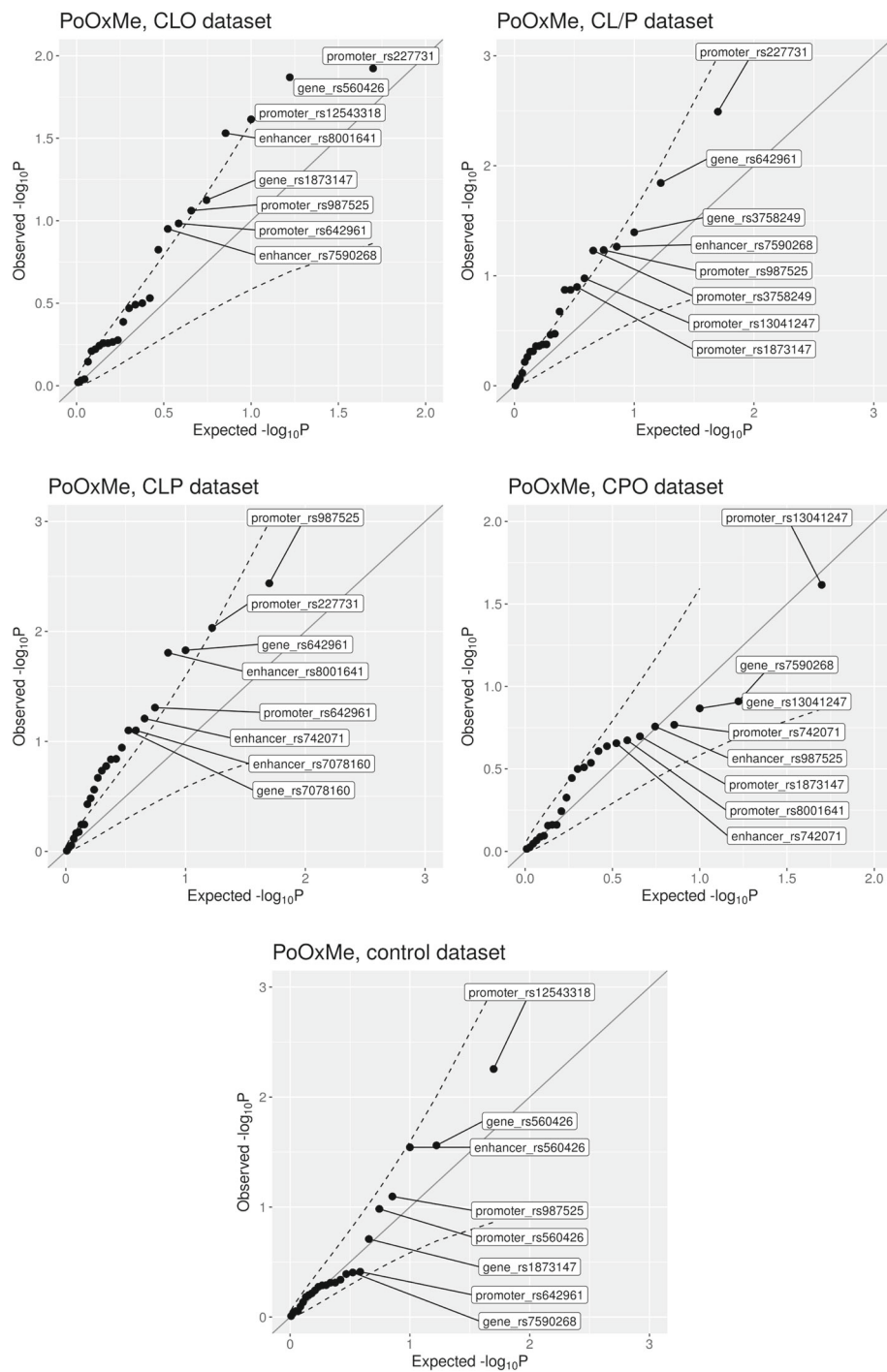
Fig. 7 Quantile–quantile plots of the interaction *p* values from the PoO×Me analyses. The dashed lines represent the 95% confidence interval

almost as much information as sequencing techniques [39]. Another potential problem when using chip-based DNAm data is the so-called gaps. As shown by Andrews et al. [40], the measured signals can sometimes be attributed not to the methylation itself, but, instead, to a mismatch occurring when a SNP is located within the DNA sequence of the probe. In those cases, plotting the β values of one CpG probe for all individuals would produce a multimodal distribution, typically with gaps in the plot. Note that while the standard pre-processing
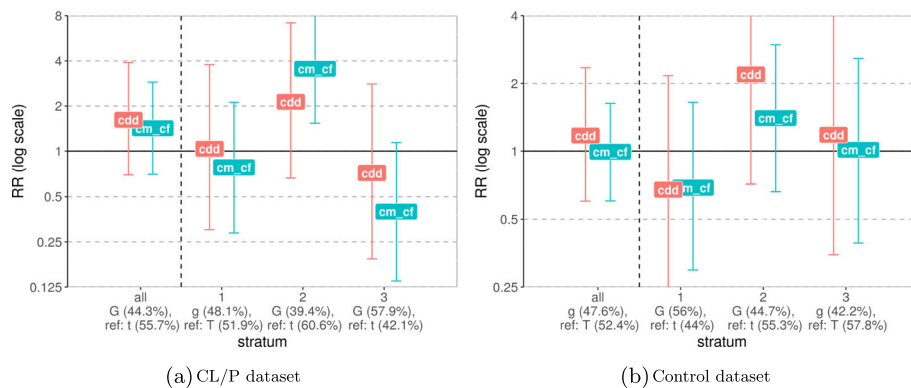
Fig. 8 PoO×Me interaction between a parent-of-origin effect of rs227731 in *NOG1* and the methylation level of the CpGs within a promoter region nearby. The *x*-axis groups the results into the unstratified dataset "all" and the results for each stratum: "1" denoting low methylation level, "2" — medium, and "3" — high methylation level. The minor and reference alleles are given for each group, along with their frequencies. The *y*-axis shows the relative risk on a log scale, with "cm_cf" denoting the parent-of-origin effect when only one minor allele is inherited (single dose) and "cdd" denoting the parent-of-origin effect when two such alleles are inherited (double dose)

procedures include removing the probes with a SNP within the sequence, each specific dataset might have its own specific SNPs. We checked for such problems in our datasets, as described in Additional File 1, Section S2.2. The `gaphunter` algorithm identified two probes as being possibly problematic. However, we did not find any correlation between the alleles and methylation levels. The impact of this problem is likely to be negligible on our methods because we took into account not only single CpG sites but summarized the methylation level within a region. Therefore, we chose to retain the data from these probes.

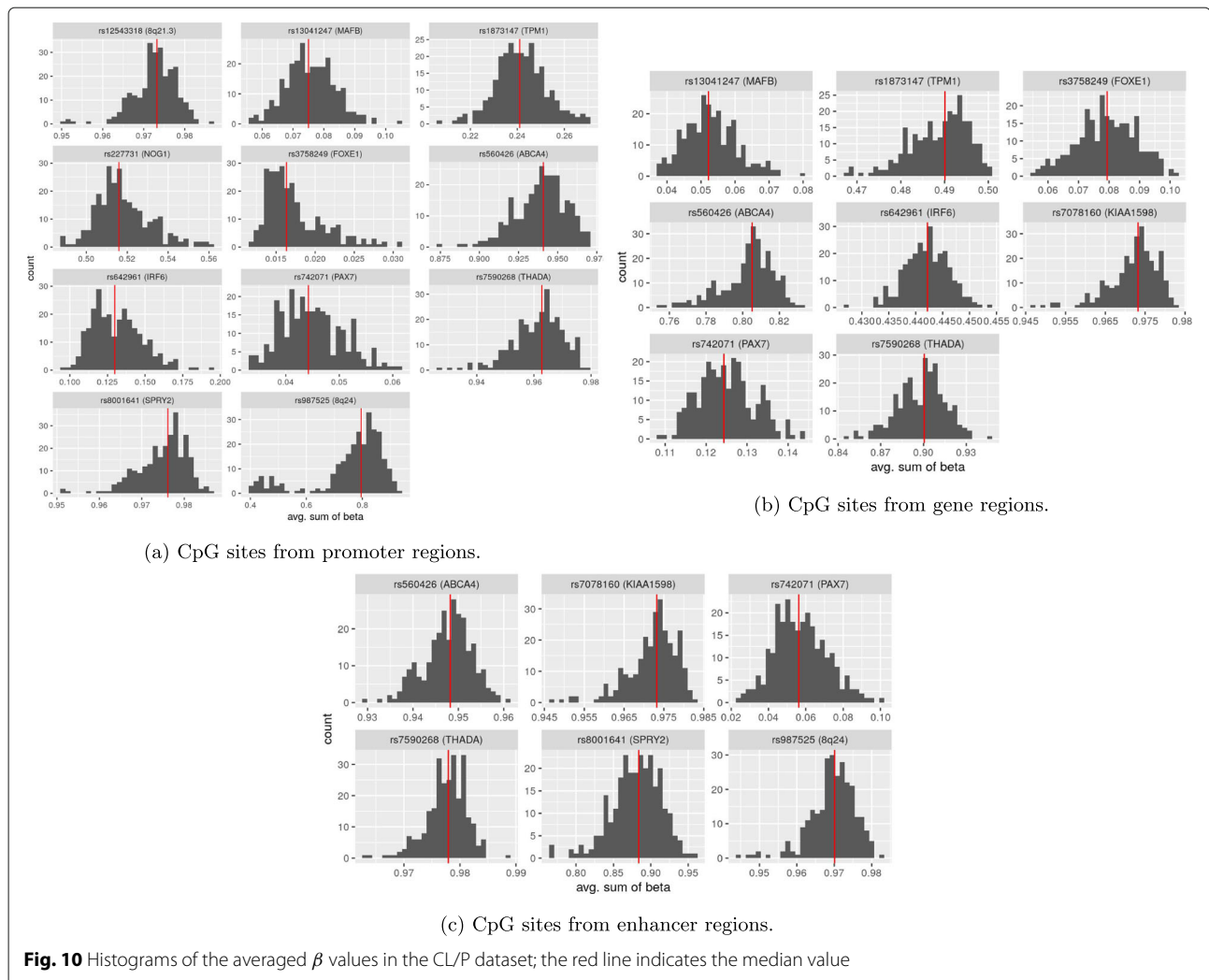**Control data are helpful in identifying false positives.**
Using log-linear models to estimate G×E and even PoO×E effects from dyad and triad studies has been thoroughly tested. Our implementation has been shown

to control for the type I error rate satisfactorily even when moderate sample sizes are used [10]. As long as the environmental exposure E is exogenous, it is reasonable to assume that the child's genes and the environment are independent of each other, conditional on parental genotypes. However, as described in the "Methods" section ("Statistical methods" section and Additional File 2), it would still be prudent to ask whether the corresponding conditions (5 and S.7) are satisfied when treating DNAm as exposure. As a simple and general test of the validity of our approach, we ran the full CpG selection strategy and G×Me analyses on 20 randomly chosen SNPs. This resulted in no false positives, indicating that the CpG selection procedure did not violate the assumption of conditional independence.

As an additional test for identifying false positives among the results for the SNPs in Table 1, we repeated
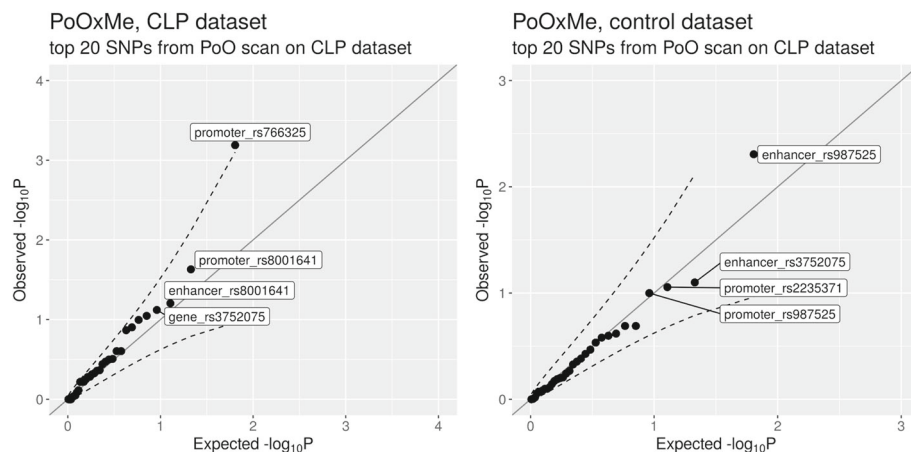
Fig. 9 Quantile–quantile plot of the *p* values from the PoO×Me analysis using the top 20 SNPs with the lowest *p* values from the GWAS scan for PoO effects of the CLP dataset. The dashed lines represent the 95% confidence interval

(a) CpG sites from promoter regions.

(b) CpG sites from gene regions.

(c) CpG sites from enhancer regions.

Fig. 10 Histograms of the averaged β values in the CL/P dataset; the red line indicates the median value

these analyses on the same SNPs in the control dataset. Intriguingly, the results of G×Me in the control dataset showed a significant interaction between rs3758249 in *FOXE1* and DNAm level in the promoter region nearby, which was also found to be significant in the analysis of the CLP dataset. We also repeated this search for interaction within the control dataset after having removed one CpG that was found in the mQTL database [41], i.e., cg13791254 (results not shown). The resulting p value of the Wald test for interaction between rs3758249 in *FOXE1* and the remaining CpGs in the promoter region was $6.4 \cdot 10^{-5}$, which is only a modest change from the original value $8.6 \cdot 10^{-5}$.

This highlights the importance of including a control dataset to sift through the initial results and to assess whether the detected G×Me interactions are genuine. The reason for observing this interaction might be that it is not specific for OFC, given that the methylation data

were generated using DNA samples from cord blood, and not from tissues that are more relevant for clefts (e.g., buccal epithelial cells or other craniofacial tissue). Nevertheless, the correlation between DNAm levels in blood and lip/palate tissues has been reported to be high [42].

The G×Me and PoO×Me methods are highly versatile
Since the main idea behind the new methods is to search for statistical interaction, the methods can be applied to a wide range of etiologic scenarios. Our implementation specifies the input as a pair consisting of a SNP and a methylation value. However, it is up to the user how these pairs are created, i.e., whether the methylation level is a value from one CpG alone or a summarized value from many contiguous CpGs, whether the CpG is located near or far away from the SNP, as long as the choice makes biological sense. Moreover, the methods are, in principle, applicable to any complex disease or binary trait.

Romanowska *et al. Clinical Epigenetics*   (2020) 12:109

Page 12 of 18

However, several points need to be addressed. These methods generally require samples from many individuals, since they are used to calculate the interaction effect, not the main effect (see, e.g., Fig. 2 in our previous work [10]). Thus, if the disease is highly polygenic, substantially more samples would be required to achieve reasonable confidence intervals for the interaction estimates. Our implementation provides more power per sample due to the calculation of the trend test and because we use only one measure of methylation level within a region instead of analyzing each CpG separately. Moreover, while the Haplin implementation is valid for dichotomous phenotypes, it should be possible to perform the search for interaction effects with software for continuous traits.

Furthermore, in our interaction analyses, we divided the dataset into three strata. This is perhaps the most reasonable minimum number of strata required, where one can apply a trend test and visually assess the changes in RR across strata. To check how sensitive the results were to the number of methylation strata, we also used two and four strata (results not shown). Generally, the four-strata analyses collapsed due to the low number of observations in each stratum and the two-strata analysis was too crude to detect trends. With larger sample sizes, a finer division than three strata may be used.

The interpretation of the significant gene–methylation interaction is not straightforward

The results of our G×Me analyses point to an interaction in the CL/P dataset between rs12543318 at 8q21.3 and the methylation state at CpG cg03309455 in the promoter-flanking region nearby (regulatory stable ID: ENSR00001432551 for GRCh37 and ENSR00000861245 for GRCh38). However, it is not easy to interpret the biological relevance of this interaction. According to the JASPAR database [43], two transcription factors (TFs) are predicted to bind to the sequence containing CpG cg03309455: ETS-related gene (ERG; uniprot ID: P11308) and Neurogenic differentiation factor 2 (NEUROD2; uniprot ID: Q15784). While ERG is a general factor expressed in 197 tissue types, according to a search in the bgee database (https://bgee.org/?page=gene&gene_id=ENSG00000157554), NEUROD2 is specifically involved in neuronal determination. rs12543318 itself is located in a non-coding region, with no genomic annotation.

As seen in Fig. 4, the above interaction was found in the CL/P dataset, and there is no matching result in the control dataset. This raises the probability that the interaction is a true positive. However, it should be noted that across the individuals in our dataset, the range of β values for this CpG is narrow and higher than 0.95 (Fig. 10a). Hence, this small absolute difference in methylation value across the three methylation strata in the interaction

analysis renders a consistent biological interpretation less likely.

Parent-of-origin effect interaction with DNA methylation was significant near rs227731

There was one significant PoO×Me interaction in three of the five tested datasets, namely, the PoO effect of rs227731 interacting with the methylation values of the promoter region nearby. There were two CpGs in this region in our data: ch.17.52148184R and cg24806663. According to the newest data in the JASPAR database, there are 19 TFs that most probably bind to the cytosine that is methylated in ch.17.52148184R (Table S14 in Additional File 1). We checked the gene ontology (GO) annotations of these proteins in the QuickGO browser (https://www.ebi.ac.uk/QuickGO/). Of these 19 TFs, 11 were annotated with "multicellular organism development" (GO ID GO:0007275), while six TFs were annotated with "embryonic skeletal system morphogenesis" (GO:0048704). Of note, one of the TFs is Homeobox protein BarH-like 1 (BARX1, UniprotID: Q9HBU1), which is involved in craniofacial development and odontogenesis [44]. The JASPAR database pointed to only four TFs that most likely bind to the cytosine in cg24806663 (Table S15 in Additional File 1), three of which are involved in transcriptional repression.

At the same time, Fig. 8a shows that this interaction is due to a noticeable change in the parent-of-origin-specific relative risk among the individuals within the middle stratum of the DNAm levels. However, since we have β values that come from chip measurements, it might be that they do not adequately capture the PoO-specific methylation distribution among the cell types and DNA strands.

The new PoO×Me method is robust and can point to possible imprinting issues

The PoO×Me method presented here has a more relaxed requirement on the independence between methylation levels and genotype, as outlined below (the "Statistical methods" subsection of the "Methods" section). Furthermore, it makes full use of the dyad and triad designs. As we have shown previously [45], there are some advantages of using the dyad design instead of the full triads, as dyads sometimes provide higher statistical power relative to the number of genotyped individuals. However, since the measured β values represent an average of DNAm levels from several cell types and, importantly, an average from the two DNA strands, some issues may remain. One is whether we can correctly capture the interaction between PoO and the *averaged* methylation level; that is, how sensitive is the PoO effect in relation to methylation? A recent study detailing DNA methylome dynamics in early embryonic development [46] showed that the father's DNAm pattern may be very different

from the mother's during this early development. Only a few genes were found whose expression patterns matched the DNAm differences. It is thus likely that DNAm alone cannot induce significant changes in expression, but is rather associated with variation in gene expression [47].

Our approach to studying PoO×Me effects is closely related to the contribution of methylation to imprinting. An imprinted locus can be seen as a locus where methylation levels in the child may depend on the parent of origin of the DNA strand. This may potentially lead to an up- or downregulation of the expression of alleles on that strand in a PoO-specific manner. There are many approaches to identify genes that exhibit imprinting [48]. For example, a recent study analyzed combinations of DNAm and genotypes in the child from mother–child dyads [49]. The authors first used maternal genotypes to establish the parent-of-origin status of SNP alleles in the child's DNA and then searched for SNPs that are associated with methylation status of nearby CpG sites in a parent-of-origin-specific manner. In our notation, this corresponds to finding loci where

$$P(Me|C_{ij}) \neq P(Me|C_{ji}),$$

that is, where the distribution of methylation values at the relevant CpG depends not only on the SNP alleles themselves but also on which parent they were inherited from. Interestingly, this is closely related to our assumption (5) (see the "Statistical methods" subsection of the "Methods" section), which we use to exclude possible false positives by checking the condition in control families. While the approach of Cuellar Partida et al. [49] is not related to a specific phenotype, our analyses are focused on OFC. We thus look for parent-of-origin-specific correlations between the methylation level (*Me*) and the child's genotype (*C*) that are present among case-children but not among control-children.

Conclusions

This study implemented two new strategies to search for interactions between DNAm levels and either the genotype (G×Me) or parent-of-origin (PoO×Me) effects. In addition, we demonstrate the use of region-wise methylation levels by focusing on biologically meaningful genomic regions (promoter, gene body, and enhancer). The inclusion of these methods in our R package Haplin facilitates the ease-of-use and adaptation, as the code is open-source and free. Additionally, we performed several sensitivity analyses to test the robustness of our methods. While the triad and dyad designs allow all of the analyses presented here to be performed, we note that independent control triads or dyads are important to check and correct for false-positive results, particularly for the G×Me model.

Methods

Genotypes

Genotypes from child–mother dyads were available for cases (1311 dyads) and controls (2481 dyads). Details regarding data collection and quality checks have been provided in our previous work [24] and are summarized in Additional File 1. In the case dyads, the child was diagnosed with one of the following three subtypes of OFC: cleft lip only (CLO), cleft lip with cleft palate (CLP), or cleft palate only (CPO). In addition to these, we also analyzed another category of cleft—cleft lip with or without cleft palate (CL/P), which is a combination of CLO and CLP. For the current analyses, we use a subset of the original dataset by selecting only those families for whom DNAm data were also available for the child.

DNAm data

The Illumina HumanMethylation 450K BeadChip (Illumina, Inc., San Diego, CA) was used to assess DNAm levels at 485,577 CpG sites in the children from the dyads mentioned above. Details on how the raw data were processed are available in our recent work [50] and are summarized in Additional File 1. The β value corresponding to the methylation level at each CpG site was calculated as $\beta = I_M/(I_M + I_U + 100)$, where I_U and I_M are the intensities of the unmethylated and methylated signals, respectively, and the factor 100 is added to ensure no division by zero. After the quality control, 407,513 CpGs and 868 samples (456 controls; 105 CLO; 167 CLP and 140 CPO) remained for analysis. Matching of the children from the methylation dataset to the genotype dataset yielded 103 CLO dyads, 269 CL/P dyads, 140 CPO dyads, 166 CLP dyads, and 456 control dyads.

To inspect the entire methylation data and check for inconsistencies, we (i) performed a statistical summary of methylation levels in the regulatory regions (promoter, enhancer, gene body) and (ii) displayed the distributions of the methylation levels around the chosen SNPs. More extensive details on these analyses are provided in Additional File 1 (see Sections S2.1, S2.2, and S2.3). Overall, the distribution of the methylation levels in our data was consistent with that of other published datasets [51].

Genomic regions

CpG sites located within a maximum distance of 50 kb from each SNP were selected and categorized into specific regions (promoter, enhancer, gene body) using the newest information from the ensembl database [52]. The R package biomaRt [53] was used to extract information from the ensembl Regulatory Feature database (using GRCh37), which includes the positions of promoters and enhancers in the human genome. Based on this information, we classified a CpG into the category "promoter"

when the search returned "promoter" (Sequence Ontology [54] accession number SO:0000167) or "promoter_flanking_region" (SO:0001952). The category "enhancer" corresponded to the "enhancer" description (SO:0000165). Gene regions were derived from the ensembl genome browser (GRCh 37) via biomaRt. Because the positions of the CpGs and SNPs in our raw data were based on the human genome hg19/GRCh37 release, we did not use the newest version of the human genome assembly (hg38/GRCh38) to avoid discrepancies in positional information.

To summarize the DNAm levels per region, we averaged the β values from the CpG sites within each regulatory region. Figure 10 presents the distributions of these summarized DNAm levels (Me) per region for the largest dataset (CL/P). Almost all of the Me distributions have a normal-like shape, except for the promoter region near rs987525. The distributions within the other datasets are presented in Figs. S5–S8 in Additional File 1.

Implementation details

We used our R package Haplin [10, 55] to calculate relative risks with confidence intervals for each SNP and for each cleft category using a log-linear model (`haplinStrat` function). The significance of interactions was estimated using the Wald test (`gxe` function). Assuming a dose–response relationship across strata, we also tested for a trend to increase power. This approach has been described elsewhere [56]. The scan for PoO effects was performed using the `haplinSlide` function, with the "window size" parameter set to 1. The R code used for the current analyses is available on the Bitbucket server (see the "Availability of data and materials" section).

We use the multiplicative scale rather than the additive scale when assessing interactions. This decision is partly dictated by the relative risk being the measure of choice for case-parent triads [55, 57]. Moreover, if significant interactions are detected on a multiplicative scale, they will, typically, also be significant on an additive scale; in that respect, our results are conservative.

We used the R packages ggplot2 [58] and ggrepel [59] and the ensembl browser (http://www.ensembl.org) to create the plots. The schemes (Figs. 1 and 2) were created using the yEd software (https://www.yworks.com/products/yed).

Statistical methods

The full framework behind the new methods is presented in Additional File 2. Here, we highlight the most salient details.

G×Me interactions

Let M, F, and C denote the genotypes of the mother, father, and child, respectively, within a family triad. Here, M, F,

and C will refer to a single SNP, but could also refer to, for instance, haplotypes built from a few SNPs in close linkage disequilibrium. In particular, let $C_{ij} = A_i A_j$ denote an ordered child genotype, where allele A_i is inherited from the mother and A_j from the father. C denotes the corresponding unordered genotype. Let D denote that the child has the disease in question and $\bar{D}$ that the child is healthy. Let Me denote methylation levels in the child, which may here be low, medium, or high, as described above. We refer to the families where $Me = m$ as belonging to stratum m. The penetrance function is $P(D|C_{ij}, Me)$, which describes how the probability of disease depends on the child's genotype as well as the level of methylation. The penetrance may depend on parent of origin since $P(D|C_{ij}, Me)$ may differ from $P(D|C_{ji}, Me)$. Specifically, we assume that

$$P(D|C_{ij}, Me = m) = B^{(m)} \cdot RR_{M,i}^{(m)} \cdot RR_{F,j}^{(m)}, \qquad (1)$$

where $B^{(m)}$ is a baseline risk in stratum m, $RR_{M,i}^{(m)}$ is the relative risk associated with inheriting allele A_i from the mother in stratum m, and similarly $RR_{F,j}^{(m)}$ is the relative risk associated with inheriting allele A_j from the father in stratum m. We thus assume a multiplicative dose–response model for the allele dose. To make parameters identifiable, we assume $RR_{M,1}^{(m)} = RR_{F,1}^{(m)} = 1$ for all m, i.e., we choose A_1 as the reference allele in all strata; $RR_{M,i}^{(m)}$ and $RR_{F,i}^{(m)}$ should be estimated from the model when $i \neq 1$. More details on models that allow deviations from the multiplicative dose–response can be found elsewhere [55, 60].

To look for G×Me effects, we first assume that $RR_{M,i}^{(m)}$ and $RR_{F,i}^{(m)}$ are equal for all i and m, i.e., that the risk does not depend on the parent of origin, and let

$$RR_i^{(m)} = RR_{M,i}^{(m)} = RR_{F,i}^{(m)}.$$

A G×Me effect on the risk of disease would mean that the relative risk $RR_i^{(m)}$ changes over strata of $Me = m$ for one or more i's.

If data have been sampled as case-parent triads and control-parent triads, we only observe the distributions $P(M, F, C, Me|D)$ and $P(M, F, C, Me|\bar{D})$, that is, the joint distributions of triad genotypes and methylation strata among the case-parent triads and control-parent triads separately. The number of triads in each category of genotypes and methylation strata is modeled as a log-linear model, and the model parameters are estimated through maximum likelihood [55, 60]. Since for the case triads,

$$
\begin{aligned}
&P(M, F, C, Me|D) \\
&= \frac{P(D|M, F, C, Me)P(Me|M, F, C)P(C|M, F)P(M, F)}{P(D)},
\end{aligned}
$$

(2)

there is a direct relationship between the parameters in the log-linear model and the penetrance model. However, the model also includes the population parental genotype distribution $P(M, F)$, the straightforward Mendelian transmission part $P(C|M, F)$, and the population disease prevalence $P(D)$. The latter may not be known, but enters the model as a scaling factor.

The basic case-parent and control-parent triad models are formulated in terms of child and both parents. However, the Haplin implementation allows mother–child and father–child dyads to be fitted within the same framework. This is achieved by using the expectation-maximization (EM) algorithm to impute the missing parent and basing inference on the likelihood that accounts for missing data. Further details on how the log-linear models is implemented can be found elsewhere [55, 56].

The factor $P(Me|M, F, C)$, i.e., the population distribution of methylation within genetic strata, is crucial in our analyses. When independent control triads are available, we make an assumption—for simplicity—of rare disease. In that case,

$$P(M, F, C, Me|\bar{D}) \approx P(Me|M, F, C)P(C|M, F)P(M, F).$$

(3)

Thus, when controls are available, there is no need to make specific assumptions about $P(Me|M, F, C)$ since it can be estimated from control triads. With only case-parent triads available, however, it is clear that assumptions about $P(Me|M, F, C)$ may interfere with inference on $P(D|M, F, C, Me)$ when observing $P(M, F, C, Me|D)$ only. In Additional File 2, we present the details on the most important situations where this becomes relevant.

PoO×Me effects
PoO effects have been described in different ways in the literature—here, we measure it as a ratio of relative risks (*RRR*). Specifically, if RR_M and RR_F are the relative risks associated with the maternally and paternally derived alleles, respectively, then the PoO effect within a stratum is $RRR = RR_M/RR_F$. See Gjerdevik et al. [10] for further details on the PoO definition.

To estimate parent-of-origin effects, we now assume that $RR_{M,i}$ is not necessarily equal to $RR_{F,i}$. In other words, the effect of allele A_i on offspring risk may depend on the parental source of the allele. Let $RRR_i = RR_{M,i}/RR_{F,i}$ be the ratio of the relative risks associated with allele A_i. Thus, $RRR_i \neq 1$ is indicative of a parent-of-origin effect. Let $RRR_i^{(m)}$ be the value of RRR_i estimated in stratum $Me = m$, and C_{ij} as defined above. As above, we assume $P(D|M, F, C_{ij}, Me) = P(D|C_{ij}, Me)$. To be able to estimate the ratio $RRR_i^{(m)}$ within each stratum, we again need to make an assumption

$$P(Me|M, F, C) = P(Me|M, F),$$

(4)

i.e., that the methylation in the child may be associated with (parental) genotypes in the population, but that the child's genotype does not directly influence methylation, conditional on parental genotypes. However, it will now suffice to assume that

$$P(Me|M, F, C_{ij}) = P(Me|M, F, C_{ji}) = P(Me|M, F, C),$$

(5)

i.e., the child's genes may influence methylation levels directly, even within parental mating types, but the mechanisms by which the child's genotype influences methylation values should not depend on whether specific alleles were inherited from the mother or from the father. Under condition (5),

$$P(M, F, C_{ij}, Me|D) = \frac{P(D|C_{ij}, Me)P(Me|M, F, C_{ij})P(M, F, C_{ij})}{P(D)}$$

$$= \frac{P(D|C_{ij}, Me)P(C|M, F, Me)P(M, F|Me)P(Me)}{P(D)}$$

$$\cdot \frac{P(C_{ij}|M, F)}{P(C|M, F)}.$$

Note that the last fraction is again determined by standard Mendelian inheritance. In a stratum-specific log-linear model, we can still obtain estimates of $RR_{M,i}^{(m)}$ and $RR_{F,i}^{(m)}$ for each stratum m. These estimates may be biased since $P(C|M, F, Me)$ may depend on both $Me = m$ and the (unordered) $C = A_iA_j$. However, the ratio $RRR_i^{(m)} = RR_{M,i}^{(m)}/RR_{F,i}^{(m)}$ will be unbiased as long as condition (5) is satisfied.

Again, if control triads are available, under the rare disease assumption, we can use estimates from the control triads to check the symmetry assumption (5), or in a combined (hybrid) analysis, unbiased estimates of $RR_{M,2}^{(m)}$ and $RR_{F,2}^{(m)}$ can be obtained.

Supplementary information

> **Additional file 1:** File containing supplementary figures and tables.
>
> **Additional file 2:** Details of the statistical methods.

Abbreviations
CL/P: Cleft lip with or without cleft palate; CLO: Cleft lip only; CLP: Cleft lip with cleft palate; CpG: Cytosine-guanine dinucleotide; CPO: Cleft palate only; DMR: Differentially methylated regions; DNAm: DNA methylation; EM: Expectation-maximization; EWAS: Epigenome-wide association study; G×Me: Gene–methylation interaction effect; GO: Gene ontology; GWAS: Genome-wide association study; meQTL: Methylation quantitative locus; OFC: Orofacial clefts; PoO: Parent-of-origin effect; PoO×Me: Parent-of-origin–methylation interaction effect; RR: Relative risk; RRR: Relative risk ratio; SNP: Single-nucleotide polymorphism; TF: Transcription factor

Romanowska *et al. Clinical Epigenetics*   (2020) 12:109

Page 16 of 18

Author details
[1]Department of Global Public Health and Primary Care, University of Bergen, N-5020, Bergen, Norway. [2]Computational Biology Unit, University of Bergen, N-5020, Bergen, Norway. [3]Centre for Fertility and Health, Norwegian Institute of Public Health, N-0213, Oslo, Norway. [4] Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, N-0473, Oslo, Norway. [5]National Institute of Environmental Health Sciences, Research Triangle Park, NC, 27709 USA.

References
1. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17(1):1–17. https://doi.org/10.1186/s13059-016-1066-1.
2. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011;3(6):771–84. https://doi.org/10.2217/epi.11.105.
3. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517(7534):321–6. https://doi.org/10.1038/nature14192.
4. Yang X, Han H, DeCarvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell. 2014;26(4):577–90. https://doi.org/10.1016/j.ccr.2014.07.028.
5. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation,. Nat Rev Genet. 2016;17(9):551–65. https://doi.org/10.1038/nrg.2016.83.
6. Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet. 2018;19(2):81–92. https://doi.org/10.1038/nrg.2017.80.
7. Liu H, Li S, Wang X, Zhu J, Wei Y, Wang Y, Wen Y, Wang L, Huang Y, Zhang B, Shang S, Zhang Y. DNA methylation dynamics: identification and functional annotation. Brief Funct Genomics. 2016;15(6):470–84. https://doi.org/10.1093/bfgp/elw029.
8. Soto-Ramírez N, Arshad SH, Holloway JW, Zhang H, Schauberger E, Ewart S, Patil V, Karmaus W. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. Clin Epigenetics. 2013;5(1):. https://doi.org/10.1186/1868-7083-5-1.
9. Mukherjee N, Lockett GA, Merid SK, Melén E, Pershagen G, Holloway JW, Arshad SH, Ewart S, Zhang H, Karmaus W. DNA methylation and genetic polymorphisms of the Leptin gene interact to influence lung function outcomes and asthma at 18 years of age. Int J Mol Epidemiol Genet. 2016;7(1):1–17.
10. Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, Jugessur A, Gjessing HK. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. Ann Hum Genet. 2018;82(2):60–73. https://doi.org/10.1111/ahg.12224.
11. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet. 1999;65(1):229–35. https://doi.org/10.1086/302466.
12. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med. 1994;13(2):153–62. https://doi.org/10.1002/sim.4780130206.
13. Wang S, Yu Z, Miller RL, Tang D, Perera FP. Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. Hum Hered. 2011;71(3):196–208. https://doi.org/10.1159/000328006.
14. Haaland ØA, Jugessur A, Gjerdevik M, Romanowska J, Shi M, Beaty TH, Marazita ML, Murray JC, Wilcox AJ, Lie RT, Gjessing HK. Genome-wide analysis of parent-of-origin interaction effects with environmental exposure (PoOxE): an application to European and Asian cleft palate trios. PLOS ONE. 2017;12(9):0184358. https://doi.org/10.1371/journal.pone.0184358.
15. Affinito O, Palumbo D, Fierro A, Cuomo M, De Riso G, Monticelli A, Miele G, Chiariotti L, Cocozza S. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics. 2020;112(1):144–50. https://doi.org/10.1016/j.ygeno.2019.05.007.
16. Lovkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016;44(11):5123–32. https://doi.org/10.1093/nar/gkw124.
17. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics Chromatin. 2018;11(1):1–17. https://doi.org/10.1186/s13072-018-0205-1.
18. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet. 2000;66(1):251–61. https://doi.org/10.1086/302707.
19. Grosen D, Bille C, Petersen I, Skytthe A, Hjelmborg JvB, Pedersen JK, Murray JC, Christensen K. Risk of oral clefts in twins. Epidemiology. 2011;22(3):313–9. https://doi.org/10.1097/EDE.0b013e3182125f9c.
20. Beaty TH, Marazita ML, Leslie EJ. Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. F1000Research. 2016;5:1–10. https://doi.org/10.12688/f1000research.9503.1.
21. Rahimov F, Jugessur A, Murray JC. Genetics of nonsyndromic orofacial clefts. Cleft Palate Craniofac J. 2012;49(1):73–91. https://doi.org/10.1597/10-178.
22. Wehby GL, Murray JC. Folic acid and orofacial clefts: a review of the evidence. Oral Dis. 2010;16(1):11–9. https://doi.org/10.1111/j.1601-0825.2009.01587.x.

23. Howe LJ, Richardson TG, Arathimos R, Alvizi L, Passos-Bueno MR, Stanier P, Nohr E, Ludwig KU, Mangold E, Knapp M, Stergiakouli E, Pourcain BS, Smith GD, Sandy J, Relton CL, Lewis SJ, Hemani G, Sharp GC. Evidence for DNA methylation mediating genetic liability to non-syndromic cleft lip/palate. Epigenomics. 2019;11(2):133–45. https://doi.org/10.2217/epi-2018-0091.

24. Moreno Uribe LM, Fomina T, Munger RG, Romitti PA, Jenkins MM, Gjessing HK, Gjerdevik M, Christensen K, Wilcox AJ, Murray JC, Lie RT, Wehby GL. A population-based study of effects of genetic loci on orofacial clefts. J Dent Res. 2017;96(11):1322–9. https://doi.org/10.1177/0022034517716914.

25. Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, Harris SE, Liewald D, Henders AK, Mendelson MM, Liu C, Joehanes R, Liang L, Levy D, Martin NG, Starr JM, Wijmenga C, Wray NR, Yang J, Montgomery GW, Franke L, Deary IJ, Visscher PM. Improving phenotypic prediction by combining genetic and epigenetic associations. Am J Hum Genet. 2015;97(1):75–85. https://doi.org/10.1016/j.ajhg.2015.05.014. http://arxiv.org/abs/NIHMS150003.

26. White CC, Yang H-S, Yu L, Chibnik LB, Dawe RJ, Yang J, Klein H-U, Felsky D, Ramos-Miguel A, Arfanakis K, Honer WG, Sperling RA, Schneider JA, Bennett DA, De Jager PL. Identification of genes associated with dissociation of cognitive performance and neuropathological burden: multistep analysis of genetic, epigenetic, and transcriptional data. PLOS Med. 2017;14(4):1002287. https://doi.org/10.1371/journal.pmed.1002287.

27. Lin X, Lim IY, Wu Y, Teh AL, Chen L, Aris IM, Soh SE, Tint MT, MacIsaac JL, Morin AM, Yap F, Tan KH, Saw SM, Kobor MS, Meaney MJ, Godfrey KM, Chong YS, Holbrook JD, Lee YS, Gluckman PD, Karnani N. Developmental pathways to adiposity begin before birth and are influenced by genotype, prenatal environment and epigenome. BMC Med. 2017;15(1):50. https://doi.org/10.1186/s12916-017-0800-1.

28. D'Addario C, Shchetynsky K, Pucci M, Cifani C, Gunnar A, Vukojević V, Padyukov L, Terenius L. Genetic variation and epigenetic modification of the prodynorphin gene in peripheral blood cells in alcoholism. Prog Neuro Psychopharmacol Biol Psychiatry. 2017;76:195–203. https://doi.org/10.1016/j.pnpbp.2017.03.012.

29. Sulkava S, Ollila HM, Alasaari J, Puttonen S, Härmä M, Viitasalo K, Lahtinen A, Lindström J, Toivola A, Sulkava R, Kivimäki M, Vahtera J, Partonen T, Silander K, Porkka-Heiskanen T, Paunio T. Common genetic variation near melatonin receptor 1A gene linked to job-related exhaustion in shift workers. Sleep. 2017;40(1):588–91. https://doi.org/10.1093/sleep/zsw011.

30. Xie B, Liu Z, Liu W, Jiang L, Zhang R, Cui D, Zhang Q, Xu S. DNA methylation and tag SNPs of the BDNF gene in conversion of amnestic mild cognitive impairment into Alzheimer's disease: a cross-sectional cohort study. J Alzheim Dis. 2017;58(1):263–74. https://doi.org/10.3233/JAD-170007.

31. Shilpi A, Bi Y, Jung S, Patra SK, Davuluri RV. Identification of genetic and epigenetic variants associated with breast cancer prognosis by integrative bioinformatics analysis,. Cancer Informat. 2017;16:1–13. https://doi.org/10.4137/CIN.S39783.

32. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–92. https://doi.org/10.1038/nrg3230.

33. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41(1):200–9. https://doi.org/10.1093/ije/dyr238.

34. Di Lena P, Sala C, Prodi A, Nardini C. Missing value estimation methods for DNA methylation data. Bioinformatics. 2019;35(19):3786–93. https://doi.org/10.1093/bioinformatics/btz134.

35. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJM, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253–7. https://doi.org/10.1038/nature09165.

36. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, Lancet D, Cohen D. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017(1):1665–80. https://doi.org/10.1093/database/bax028.

37. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012;22(3):490–503. https://doi.org/10.1038/cr.2012.15.

38. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermitzakis ET. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife. 2013;2013(2):1–18. https://doi.org/10.7554/eLife.00523.

39. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–81. https://doi.org/10.1038/nature12433.

40. Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD. "Gap hunting" to characterize clustered probe signals in Illumina methylation array data. Epigenetics Chromatin. 2016;9(1):1–21. https://doi.org/10.1186/s13072-016-0107-z.

41. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, Zheng J, Duggirala A, McArdle WL, Ho K, Ring SM, Evans DM, Davey Smith G, Relton CL. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17(1):61. https://doi.org/10.1186/s13059-016-0926-z.

42. Sharp GC, Ho K, Davies A, Stergiakouli E, Humphries K, McArdle W, Sandy J, Davey Smith G, Lewis SJ, Relton CL. Distinct DNA methylation profiles in subtypes of orofacial cleft. Clin Epigenetics. 2017;9(1):63. https://doi.org/10.1186/s13148-017-0362-2.

43. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2019;48(D1):87–92. https://doi.org/10.1093/nar/gkz1001.

44. Krivicka-Uzkurele B, Pilmane M, Akota I. Barx1, growth factors and apoptosis in facial tissue of children with clefts,. Stomatologija. 2008;10(2):62–6.

45. Gjerdevik M, Jugessur A, Haaland ØA, Romanowska J, Lie RT, Cordell HJ, Gjessing HK. Haplin power analysis: a software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls. BMC Bioinformatics. 2019;20(1):165. https://doi.org/10.1186/s12859-019-2727-3.

46. Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, Lian Y, Fan X, Hu B, Gao Y, Wang X, Wei Y, Liu P, Yan J, Ren X, Yuan P, Yuan Y, Yan Z, Wen L, Yan L, Qiao J, Tang F. Single-cell DNA methylome sequencing of human preimplantation embryos. Nat Genet. 2018;50(1):12–9. https://doi.org/10.1038/s41588-017-0007-6.

47. Chen R, Xia L, Tu K, Duan M, Kukurba K, Li-Pook-Than J, Xie D, Snyder M. Longitudinal personal DNA methylome dynamics in a human with a chronic condition. Nat Med. 2018;24(12):1930–9. https://doi.org/10.1038/s41591-018-0237-x.

48. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet. 2019;20(4):235–48. https://doi.org/10.1038/s41576-018-0092-0.

49. Cuellar Partida G, Laurin C, Ring SM, Gaunt TR, McRae A, Visscher PM, Montgomery G, Martin NG, Hemani G, Suderman M, Relton CL, Davey Smith G, Evans DM. Genome-wide survey of parent-of-origin effects on DNA methylation identifies candidate imprinted loci in humans. Hum Mol Genet. 2018;27(16):2927–39. https://doi.org/10.1093/hmg/ddy206.

50. Xu Z, Lie RT, Wilcox AJ, Saugstad OD, Taylor JA. A comparison of DNA methylation in newborn blood samples from infants with and without orofacial clefts. Clin Epigenetics. 2019;11(1):40. https://doi.org/10.1186/s13148-019-0638-9.

51. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D,

Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong Na, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer La, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra Ma, McManus MT, Sunyaev S, Thomson Ja, Tlsty TD, Tsai L-H, Wang W, Waterland Ra, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos Ja, Wang T, Kellis M, Pfenning A, ClaussnitzerYaping Liu M, Alan Harris R, David Hawkins R, Scott Hansen R, Abdennur N, Adli M, Akerman M, Barrera L, Antosiewicz-Bourget J, Ballinger T, Barnes MJ, Bates D, Bell RJA, Bennett Da, Bianco K, Bock C, Boyle P, Brinchmann J, Caballero-Campo P, Camahort R, Carrasco-Alfonso MJ, Charnecki T, Chen H, Chen Z, Cheng JB, Cho S, Chu A, Chung W-Y, Cowan C, Athena Deng Q, Deshpande V, Diegel M, Ding B, Durham T, Echipare L, Edsall L, Flowers D, Genbacev-Krtolica O, Gifford C, Gillespie S, Giste E, Glass Ia, Gnirke A, Gormley M, Gu H, Gu J, Hafler Da, Hangauer MJ, Hariharan M, Hatan M, Haugen E, He Y, Heimfeld S, Herlofsen S, Hou Z, Humbert R, Issner R, Jackson AR, Jia H, Jiang P, Johnson AK, Kadlecek T, Kamoh B, Kapidzic M, Kent J, Kim A, Kleinewietfeld M, Klugman S, Krishnan J, Kuan S, Kutyavin T, Lee A-Y, Lee K, Li J, Li N, Li Y, Ligon KL, Lin S, Lin Y, Liu J, Liu Y, Luckey CJ, Ma YP, Maire C, Marson A, Mattick JS, Mayo M, McMaster M, Metsky H, Mikkelsen T, Miller D, Miri M, Mukame E, Nagarajan RP, Neri F, Nery J, Nguyen T, O'Geen H, Paithankar S, Papayannopoulou T, Pelizzola M, Plettner P, Propson NE, Raghuraman S, Raney BJ, Raubitschek A, Reynolds AP, Richards H, Riehle K, Rinaudo P, Robinson JF, Rockweiler NB, Rosen E, Rynes E, Schein J, Sears R, Sejnowski T, Shafer A, Shen L, Shoemaker R, Sigaroudinia M, Slukvin I, Stehling-Sun S, Stewart R, Subramanian SL, Suknuntha K, Swanson S, Tian S, Tilden H, Tsai L, Urich M, Vaughn I, Vierstra J, Vong S, Wagner U, Wang H, Wang T, Wang Y, Weiss A, Whitton H, Wildberg A, Witt H, Won K-J, Xie M, Xing X, Xu I, Xua. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30. https://doi.org/10.1038/nature14248.
52. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P. Ensembl 2016. Nucleic Acids Res. 2016;44(D1): 710–6. https://doi.org/10.1093/nar/gkv1157. http://www.ensembl.org/.
53. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nat Protocol. 2009;4:1184–91. https://bioconductor.org/ packages/release/bioc/html/biomaRt.html. Accessed 18 June 2020.
54. Mungall C, Misra S, Berman B, Carlson J, Frise E, Harris N, Marshall B, Shu S, Kaminker J, Prochnik S, Smith C, Smith E, Tupy J, Wiel C, Rubin G, Lewis S. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2002;3(12):0081–1. https://doi.org/10.1186/ gb-2002-3-12-research0081.
55. Gjessing HK, Lie RT. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. Ann Hum Genet. 2006;70(3):382–96.
56. Skare Ø, Jugessur A, Lie RT, Wilcox AJ, Murray JC, Lunde A, Nguyen TT, Gjessing HK. Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts,. Ann Hum Genet. 2012;76(3):221–36. https://doi.org/10.1111/j.1469-1809.2012.00707.x.
57. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet. 2000;66(1):251–61. https://doi.org/10.1086/302707.
58. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2016. https://ggplot2-book.org/. Accessed 18 June 2020.
59. Slowikowski K. Ggrepel: automatically position non-overlapping text labels with 'ggplot2'. 2018. R package version 0.8.0. https://CRAN.R-project.org/package=ggrepel. Accessed 18 June 2020.
60. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet. 1998;62(4):969–78.
61. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. R Foundation for Statistical Computing. https://www.R-project.org/.

Publisher's Note

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