A case–control design for testing and estimating epigenetic effects on complex diseases

Yihan Sui, Weimiao Wu, Zhong Wang, Jianxin Wang, Zuoheng Wang and Rongling Wu

Submitted: 28th September 2012; Received (in revised form): 27th November 2012

Abstract

Epigenetic modifications may play an important role in the formation and progression of complex diseases through the regulation of gene expression. The systematic identification of epigenetic variants that contribute to human diseases can be made possible using genome-wide association studies (GWAS), although epigenetic effects are currently not included in commonly used case–control designs for GWAS. Here, we show that epigenetic modifications can be integrated into a case–control setting by dissolving the overall genetic effect into its different components, additive, dominant and epigenetic. We describe a general procedure for testing and estimating the significance of each component based on a conventional chi-squared test approach. Simulation studies were performed to investigate the power and false-positive rate of this procedure, providing recommendations for its practical use. The integration of epigenetic variants into GWAS can potentially improve our understanding of how genetic, environmental and stochastic factors interact with epialleles to construct the genetic architecture of complex diseases.

Keywords: Case-control design; epigenetic effect; quantitative genetics

INTRODUCTION

Epigenetic alterations are cellular modifications that are not dependent on DNA sequence changes but can be stimulated by environmental changes [1, 2]. The mechanisms that produce epigenetic variation typically include DNA methylation, histone modifications and non-coding RNAs, as well as many other regulatory forms. As a ubiquitous phenomenon in many taxa, epigenetic processes are thought to be a driving force for buffering against environmental

perturbations and can be causal for disease and trait formation, although they can also arise as a consequence of disease [3–9]. Current next-generation sequencing technologies have developed to a point, at which high-throughput epigenetic changes (at least for DNA methylation) can be assayed with high resolution, directly comparable to that of highly successful genome-wide single-nucleotide polymorphism chipping [10]. Such a technological advance holds a promise to identify genome-wide

Corresponding author: Rongling Wu, Center for Computational Biology, College of Biological Science and Technology, Beijing Forestry University, Beijing 100083, China. Tel.: +86-10-6233-6283; Fax: +86-10-6233-6164; E-mail: rwu@bjfu.edu.cn; Center for Statistical Genetics, The Pennsylvania State University, Hershey, PA 17033, USA. Tel.: +1-717-531-2037; Fax: +1-717-531-0480; E-mail: rwu@phs.psu.edu

Yihan Sui is a fourth-year undergraduate student in biological sciences at Beijing Forestry University. She develops statistical and bioinformatics approaches to studying the genetic control of complex traits and diseases.

Weimiao Wu is a third-year undergraduate student in biological sciences at Beijing Forestry University. Her research interest is in developing mathematical and statistical modeling of genes for dynamic traits.

Zhong Wang works as a computer scientist working at The Pennsylvania State University and Beijing Forestry University. He develops computer algorithms for statistical genetic models.

Jianxin Wang is an associate professor in information science at Beijing Forestry University. He is interested in algorithmic development for complex systems and the use of new algorithms to study genetic problems.

Zuoheng Wang is an assistant professor in biostatistics at Yale University. Her research interest is to develop statistical models for genetic mapping and genome-wide association studies.

Rongling Wu is a professor in biostatistics and bioinformatics and the director of the Center for Statistical Genetics at The Pennsylvania State University. He founded the Center for Computational Biology at Beijing Forestry University. He is interested in quantitative and statistical genetics.

epigenetic profiles, understand the functional impact of epigenetics and, ultimately, elucidate a complete picture of the genetic and non-genetic determinants of complex diseases and traits.

One of the key challenges facing geneticists is how to design a powerful experiment that can capitalize on epigenetic data and detect significant epigenetic variants that are associated with complex diseases. Several authors have proposed linkage or association mapping strategies to identify the amount and pattern of quantitative variation determined by epialleles, a group of identical genes that differ in the extent of methylation [11] and retrieve epigenetic variation that cannot be estimated presently [12–15]. For example, covariances between relatives due to epigenetic transmissibility have been derived by Tal *et al.* [16]. By making crosses derived from epigenomically perturbed isogenic lines, Johannes and Colome-Tatche [17] developed an approach to estimating epigenetic variation, allowing the effects of epiallelic instability, recombination, parent-of-origin effects and transgressive segregation to be characterized. Wang *et al.* [18] integrated conventional quantitative genetic theory with DNA methylation mechanisms to provide an approach for quantifying epigenetic effects and variation in complex traits.

Given its popularity in candidate-gene association studies and genome-wide association studies (GWAS), a case-control design that compares genetic variants in a group of diseased individuals versus a group of unrelated controls has proven to be powerful for detecting and estimating the genetic variants that affect complex diseases. However, there have been no studies that can integrate epigenetic control into a case-control setting, making it impossible to elucidate a comprehensive picture of the genetic architecture of complex diseases. In this article, we show that epigenetic variants can be identified within the case-control context and provide a general procedure for testing and estimating epigenetic variation in human diseases. Epigenetic modifications can appear in multiple forms, but most studied form is DNA methylation [19]. As its most common feature, DNA methylation may occur: (i) at a cytosine-guanine dinucleotides (CpG) site [20], at which some cytosines within a CpG dinucleotide are methylated by adding a methyl group to the fifth position of the cytosine pyrimidine ring and (ii) at multiple adjacent CpG sites with varying lengths [21]. Variation in DNA methylation at a single or multiple sites is

called a methylation variable position (MVP) and a differentially methylated region (DMP), respectively. By considering MVP or DMP as the epigenetic equivalent of a SNP or multiple SNPs, Wang *et al.* [18] are able to separate and estimate the genetic and epigenetic contributions to a complex trait using a quantitative genetic and epigenetic theory. Here, we test each of these two types of contributions in a case-control design. It is our hope that the integration of epigenetic control with case-control studies provides useful guidance for designing efficient and effective GWAS to characterize a comprehensive picture of the epigenetic variation of diseases.

DESIGN

Modeling genetic and epigenetic effects

The quantitative epigenetic model is founded on the classic theory of quantitative genetics [22, 23]. To clearly describe our model, we first consider DNA methylation at a CpG site. Similar procedures can be derived for DNA methylation with multiple alleles and at multiple sites, as commonly observed in practice [19–21]. At this methylated site, two alleles A_1 and A_2 are assumed. The two alleles produce three genotypes, A_1A_1 , A_1A_2 and A_2A_2 , in a natural population. We use a to denote the additive effect of the nucleotide site due to the substitution of allele A_1 by A_2 or vice versa and d to denote the dominant effect due to the interaction between the two alleles. We express diagrammatically the values of three genotypes over an axis as follows:

Suppose there is a particular environmental signal that drives one of the alleles (say A_1) to be methylated, leading to six distinguishable genetic and epigenetic types whose genotypic values and observations are expressed as

Genotype/epigenotype	Expected value	Observation
$A_1A_1 \Rightarrow \begin{cases} A_1A_1 \text{ No methylation} \\ A_1A_e \text{ One methylation} \\ A_eA_e \text{ Two methylations} \end{cases}$	$\begin{cases} \mu_{11} = \mu + a_1 \\ \mu_{1e} = \mu + \frac{1}{2}(a_1 + a_e) + d_{1e} \\ \mu_{ee} = \mu + a_e \end{cases}$	$\begin{cases} n_{11} \\ n_{1e} \\ n_{ee} \end{cases}$
$A_1A_2 \Rightarrow \begin{cases} A_1A_2 \text{ No methylation} \\ A_2A_e \text{ One methylation} \end{cases}$	$\begin{cases} \mu_{12} = \mu - \frac{1}{2}a_e + d_{12} \\ \mu_{2e} = \mu - \frac{1}{2}a_1 + d_{2e} \end{cases}$	$\begin{cases} n_{12} \\ n_{2e} \end{cases}$
$A_2A_2 \Rightarrow A_2A_2 \text{ No methylation}$	$\mu_{22} = \mu - a_1 - a_e$	n_{22}

(2)

where the genotypic value of the trait is decomposed into different components, i.e. the overall mean (μ),

the additive effects due to the substitution of allele A_1 (a_1) and epiallele A_e by allele A_2 (a_e) and the dominance effects due to the interaction between allele A_1 and epiallele A_e (d_{1e}), between allele A_1 and allele A_2 (d_{12}) and between allele A_2 and epiallele A_e (d_{2e}).

The genetic and epigenetic effects can be expressed as, by solving a group of regular equations for the genotypic values [Equation (2)],

$$a_1 = \frac{1}{3}[2\mu_{11} - (\mu_{ee} + \mu_{22})] \quad (3)$$

$$a_e = \frac{1}{3}[2\mu_{ee} - (\mu_{11} + \mu_{22})] \quad (4)$$

$$d_{1e} = \mu_{1e} - \frac{1}{2}(\mu_{11} + \mu_{ee}) \quad (5)$$

$$d_{2e} = \mu_{2e} - \frac{1}{2}(\mu_{22} + \mu_{ee}) \quad (6)$$

$$d_{12} = \mu_{12} - \frac{1}{2}(\mu_{11} + \mu_{22}) \quad (7)$$

In an epigenetic study, much effort is given to test the epigenetic effect of the nucleotide site, a_e , and dominant effects due to the interactions between the alleles and epiallele, d_{1e} and d_{2e} . Next, we use a case-control design to test each of the genetic and epigenetic effects from Equations (3) to (7).

Contingency tables

From a natural human population, we draw two groups of subjects at random. One group includes n diseased individuals, whereas the second group has m controls with no disease. It is assumed that cases and controls are matched for age, sex, ethnicity, geographical location and environmental factors. All the cases and controls are genotyped for the same set of nucleotide sites, at each of which an allele is methylated to an epiallele. Based on the genotype/epigenotypes, we classify all the observations for the case and control into a (2×6) contingency table as follows:

	A_1A_1	A_1A_e	A_eA_e	A_1A_2	A_2A_e	A_2A_2	Total
Cases	n_{11}	n_{1e}	n_{ee}	n_{12}	n_{2e}	n_{22}	p_{1+}
Controls	m_{11}	m_{1e}	m_{ee}	m_{12}	m_{2e}	m_{22}	p_{2+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{+4}	p_{+5}	p_{+6}	p_{++}

(8)

Let p_{ij} denote the actual frequency of observations in the i th row ($i = 1$ for cases and 2 for controls) and j th column ($j = 1$ for A_1A_1 , ..., 6 for A_2A_2), let p_{i+} and

p_{+j} denote column and row marginals, respectively, and let p_{++} denote the grand total. If the genotypes/epigenotypes are independent of the disease, then the probability of being in p_{ij} is simply the product of the probability of being in marginal category i and j , expressed as

$$P_{ij} = \frac{p_{i+}p_{+j}}{p_{++}}$$

A Pearson statistic for testing the genotype/epigenotype-disease association can be calculated by

$$\chi_G^2 = \sum_{i=1}^2 \sum_{j=1}^6 \frac{(P_{ij} - p_{ij})^2}{P_{ij}} \quad (9)$$

which follows a chi-square distribution with $(2 - 1) \times (6 - 1) = 5$ degrees of freedom. If $\chi_G^2 > \chi_{0.05(df=5)}^2$, this suggests that the genotypes/epigenotypes are significantly associated with the disease. Otherwise, there is no significant association between the genotypes/epigenotypes and the disease.

Testing and estimating genetic and epigenetic effects

Combining the principle of analyzing the contingency table [Equation (8)] with genetic and epigenetic partitioning Equations (3)–(7), we can test each of these components by reformulating new contingency tables. The contingency table for testing the additive effect of the wild-type alleles (a_1) is expressed as

	A_1A_1	$A_eA_e + A_2A_2$
Cases	$2n_{11}$	$n_{ee} + n_{22}$
Controls	$2m_{11}$	$m_{ee} + m_{22}$

(10)

The contingency table for testing the additive effect of the epiallele A_e by allele A_2 (a_e) is expressed as

	A_eA_e	$A_1A_1 + A_2A_2$
Cases	$2n_{ee}$	$n_{11} + n_{22}$
Controls	$2m_{ee}$	$m_{11} + m_{22}$

(11)

The contingency table for testing the dominance effect (d_{1e}) due to the interaction between allele A_1 and epiallele A_e is expressed as

	A_1A_e	$A_1A_1 + A_eA_e$
Cases	n_{1e}	$n_{11} + n_{ee}$
Controls	m_{1e}	$m_{11} + m_{ee}$

(12)

The contingency table for testing the dominance effect (d_{2e}) due to the interaction between allele A_2 and epiallele A_e is expressed as

	A_2A_e	$A_2A_2 + A_eA_e$
Cases	n_{2e}	$\frac{1}{2}(n_{22} + n_{ee})$
Controls	m_{2e}	$\frac{1}{2}(m_{22} + m_{ee})$

(13)

The contingency table for testing the dominance effect (d_{12}) due to the interaction between wild-type alleles A_1 and A_2 is expressed as

	A_1A_2	$A_1A_1 + A_2A_2$
Cases	n_{12}	$\frac{1}{2}(n_{11} + n_{22})$
Controls	m_{12}	$\frac{1}{2}(m_{11} + m_{22})$

(14)

The significance of each genetic and epigenetic effect above, a_1 , a_e , d_{1e} , d_{12} and d_{2e} , can be tested by calculating chi-squared test statistics based on the corresponding contingency tables, expressed as $\chi^2_{a_1}$, $\chi^2_{a_e}$, $\chi^2_{d_{1e}}$, $\chi^2_{d_{12}}$ and $\chi^2_{d_{2e}}$, respectively. Since cell counts in contingency tables [Equations (10)–(14)] are formed by a weighted combination of original observed cell counts, these test statistics may not obey a chi-square distribution with 1 degree of freedom. Liu *et al.* [24] have proved that under the null hypothesis, these test statistics are asymptotically smaller than $\chi^2_{0.05(df=1)}$. They further derived an approximation approach for determining the critical thresholds.

The effect sizes of a_1 , a_e , d_{1e} , d_{12} and d_{2e} are estimated by odds ratios (ORs). A general form of contingency tables [Equations (10)–(14)] is expressed as the joint distribution of two binary random variables X and Y , i.e.

	$Y = 1$	$Y = 2$
$X = 1$	n_1	n_2
$X = 2$	m_1	m_2

from which an OR is calculated as

$$\hat{OR} = \frac{n_1 m_2}{n_2 m_1}$$

describing the extent to which the two variables are related through a particular genetic or epigenetic effect. $OR = 1$ suggests no relationship. The confidence interval of the OR effect value is calculated as $e^{\ln \hat{OR} \pm z \cdot SE}$, where z is a standard normal deviate corresponding to the level of confidence ($z = 1.96$, 2.576 for 95 and 99% confidence, respectively), and

SE is the estimated standard deviation of the log OR calculated as $SE = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{m_1} + \frac{1}{m_2}}$.

MODEL DEMONSTRATION AND VALIDATION

The epigenetic model was demonstrated and validated through simulation studies. Suppose there is a case population in which a nucleotide site has two alleles A_1 and A_2 , with frequencies p and $1 - p$, respectively. The allele A_1 is methylated into an epiallele A_e at a rate u . Through DNA methylation, allele frequencies will change; i.e., the frequencies of non-methylated A_1 allele, epiallele A_e and allele A_2 are $(1 - u)p$, up and $1 - p$, respectively. The frequencies of six resulting genetic and epigenetic types at the nucleotide site are expressed as

Genotype/epigenotype	Frequency	Observation
$A_1A_1 \Rightarrow$	$\begin{cases} A_1A_1 \text{ No methylation} & (1 - u)^2 p^2 + D_{12} + D_{1e} \\ A_1A_e \text{ One methylation} & 2u(1 - u)p^2 - 2D_{1e} \\ A_eA_e \text{ Two methylations} & u^2 p^2 + D_{1e} + D_{2e} \end{cases}$	$\begin{cases} n_{11} \\ n_{1e} \\ n_{ee} \end{cases}$
$A_1A_2 \Rightarrow$	$\begin{cases} A_1A_2 \text{ No methylation} & 2(1 - u)p(1 - p) - 2D_{12} \\ A_2A_e \text{ One methylation} & 2up(1 - p) - 2D_{2e} \end{cases}$	$\begin{cases} n_{12} \\ n_{2e} \end{cases}$
$A_1A_2 \Rightarrow$	$A_1A_2 \text{ No methylation}$	$(1 - p)^2 + D_{12} + D_{2e}$
		n_{22}

(15)

where D_{12} , D_{1e} and D_{2e} are the coefficients of Hardy–Weinberg disequilibrium (HWD) due to a non-random association between alleles A_1 and A_2 , between allele A_1 and epiallele A_e and between allele A_2 and epiallele A_e , respectively. The genotype and epigenotype frequencies may be determined by allele and epiallele frequencies and HWD coefficients. It is possible that DNA methylation violates the previous equilibrium of the population, with the HWD quantified by D_{12} , D_{1e} and D_{2e} , whose intervals are given as

$$\begin{aligned} \max[-u^2 p^2 - up(1 - p), -(1 - u)^2 p^2 - (1 - u)p(1 - p)] &\leq D_{1e} \leq u(1 - u)p^2 \\ \max[-u^2 p^2 - u(1 - u)p^2, -(1 - p)^2 - (1 - u)p(1 - p)] &\leq D_{2e} \leq up(1 - p) \\ \max[-up(1 - p) - (1 - p)^2, -(1 - u)^2 p^2 - u(1 - u)p^2] &\leq D_{12} \leq (1 - u)p(1 - p) \end{aligned}$$

A similar situation holds for the control population.

The model was tested under four different scenarios. The first is to assume that both case and control populations have the same methylation occurrence rate and HWD coefficients. In the second scenario, the case population is assumed to have a larger occurrence rate than the control population, with the other parameters being identical. The third scenario is reciprocal in parameter design to the second one.

The last scenario assumes no HWD, although methylation occurs at the same rate for both populations. All parameter values for the four scenarios are given in Table 1, with which to calculate genotype/epigenotype frequencies and then simulate contingency table [Equation (8)]. The sample size considered is 500 cases + 500 controls, 2000 cases + 2000 controls and 5000 cases + 5000 controls.

Epigenetic effects contain three different components, the additive effect due to the substitution of epiallele A_e by allele A_2 (a_e) and the dominance effects due to the interaction between allele A_1 and epiallele A_e (d_{1e}) and between allele A_2 and epiallele A_e (d_{2e}). The power and false-positive rates of detecting each of these components from the model are

Table 1: Parameter values used to simulate the case and control populations

Scenario		p	u	D_{1e}	D_{2e}	D_{12}
1	Case	0.80	0.003	0.003	-0.003	
	Control	0.20	0.80	0.003	0.003	-0.003
2	Case	-	0.50	0.003	0.003	-0.003
	Control	0.40	0.20	0.003	0.003	-0.003
3	Case	-	0.45	0.003	0.007	-0.003
	Control	0.23	0.80	0.003	0.007	-0.003
4	Case	-	0.20	0	0	0
	Control	0.4	0.20	0	0	0

investigated. To do that, we simulated the count data in contingency table [Equation (8)] by assuming different OR for each case. The OR describes and quantifies the extent to which the two categories are related. Values of OR = 1, 1.6, 1.9 and 2.2 are regarded as no effect, small effect, moderate effect and large effect, respectively. The count data were simulated under each of these values. Note that, in order to simulate the data under these OR constraints, we relax the allele frequency of allele A_1 in the case population.

The simulated data were analyzed by the epigenetic model, from which to calculate test statistics, $\chi^2_{a_e}$, $\chi^2_{d_{1e}}$, $\chi^2_{d_{2e}}$, under different effect sizes. In each case, simulation was replicated 1000 times, which were used to calculate the power for epigenetic effect detection for the data of OR > 1 and false-positive rates for the data of OR = 1. In general, even for a small sample size (500 cases + 500 controls), the model has good power (>0.80) to detect any size of epigenetic dominant effect d_{2e} under each scenario, while having reasonably low false-positive rates (<0.06) (Table 2). There is adequate power to detect the additive effect of epiallele when a moderate sample is used (2000 cases + 2000 controls). To detect a small epigenetic dominant effect d_{1e} , we need a large sample size (5000 cases + 5000 controls). It should be pointed out that the model cannot detect epigenetic dominant effect d_{1e} under

Table 2: Power analysis of epigenetic detection at different magnitudes of epigenetic effects (described by OR values) using various sample sizes under different simulation scenarios

Scenario	None (OR = 1)			Small (OR = 1.6)			Moderate (OR = 1.9)			Large (OR = 2.2)		
	a_e	d_{1e}	d_{2e}	a_e	d_{1e}	d_{2e}	a_e	d_{1e}	d_{2e}	a_e	d_{1e}	d_{2e}
500 cases + 500 controls												
1	5	2	2	27	9	84	51	6	99	64	19	100
2	6	2	4	12	12	56	21	23	90	34	53	97
3	7	0	6	30	5	78	49	13	93	72	15	95
4	4	3	3	10	1	45	13	2	69	20	6	86
2000 cases + 2000 controls												
1	4	4	1	76	24	100	99	45	100	99	67	100
2	5	4	5	51	45	97	82	81	100	96	99	100
3	4	3	2	76	21	100	99	43	100	100	57	100
4	4	3	1	27	2	97	50	2	100	71	1	100
5000 cases + 5000 controls												
1	3	4	0	98	51	100	100	81	100	100	100	100
2	3	1	3	86	80	100	99	99	100	100	100	100
3	6	4	2	100	41	100	100	77	100	100	94	100
4	3	4	0	64	2	100	82	3	100	95	1	100

Note: a_e is the additive effect due to the substitution of epiallele A_e by allele A_2 , d_{1e} and d_{2e} are the dominance effects due to the interaction between allele A_1 and epiallele A_e and between allele A_2 and epiallele A_e , respectively.

Scenario 4 in which case this effect parameter is canceled out. The false-positive rates are always low for any sample size under any simulation scenario.

DISCUSSION

Given their tremendous importance in contributing to genetic variation and adapting to environmental perturbations, epigenetic alterations have been increasingly studied, mostly focusing the molecular characterization of the underlying formation mechanisms [19–21, 25, 26]. Although this has helped greatly to improve our understanding of the origin of epigenetic alterations, a systematic detection of this phenomenon through GWAS in a commonly used case–control design is still not largely feasible [19]. Here, we have integrated quantitative genetic and epigenetic principles established by Wang *et al.* [18] into a case–control setting, allowing the significance of genetic and epigenetic effects on a disease to be tested and analyzed. In the previous studies, quantitative genetic principles were incorporated with a case–control design to analyze the associations between epistasis and disease [27, 28].

By defining various types of genetic and epigenetic effects, the approach described reformats case–control observational data into a 2×2 contingency table and then tests the associations between different genotypes or epigenotypes and disease using Pearson test statistics. An OR is calculated to estimate genetic and epigenetic effects on disease. This approach is more advantageous over a classic log-linear regression approach in many aspects. First, it is computationally simple and efficient based on statistical analyses of individual 2×2 contingency tables. As shown by simulation studies in Liu *et al.* [24], a log-linear approach incorporates all possible factors into a single model, which is easily intractable when the number of factors increases. Second, the new approach integrates quantitative genetic principles into a case–control setting, allowing statistical inference to be more biologically meaningful over that from classic approaches. Third, the approach described can be readily extended to any complicated situation, in which genetic and epigenetic effects arising from multiple methylated sites as well as the interactions between these effects and environmental factors are important contributors to disease. The approach presented is capable of testing the components of epigenetic effects including the additive effect of the methylated epiallele (a_e) and the

dominant effects between the epiallele and wild-type alleles (d_{1e} and d_{2e}). It is possible that epigenetic alterations trigger the disease singly or jointly through these different effect types [29]. For a particular case–control data set, the model allows the test and estimation of each of these effects. We performed a set of simulation studies to investigate the statistical properties of the model in analyzing and detecting epigenetic effects for a practical case–control data set. The results from the simulation demonstrate the utilization and usefulness of the model.

To better describe the principle of epigenetic analysis in a case–control study, we considered only a single methylated CpG site for model derivations at a time. However, given the complexity of trait inheritance, one can expect that multiple sites are methylated simultaneously, giving rise to a complex network of genetic and epigenetic interactions. There is no technical difficulty in extending the model to include two or more sites that may interact with each other, although its solution needs powerful computational facility. For two methylated sites, we need to handle a total of 25 interactions between parameter set ($a_1, a_e, d_{1e}, d_{2e}, d_{12}$) from each site. In some particular case, epigenetic contribution contains a series of pathways, therefore stimulating our model derivation to include high-order interactions [27]. Also, there are many other mechanisms, such as histone modifications and non-coding RNA modifications, that cause epigenetic perturbations [19]. In some cases, an epiallele can modify the expression of an allele, creating a spectrum of alleles that are functionally diverse [30]. All these epigenetic mechanisms and forms can be integrated into our model, with observational data rearranged in terms of 2×2 contingency tables, to better chart an entire picture of epigenetic regulation for complex diseases.

The expression of epigenetic effects is environment dependent [26, 31]. Epigenotype–environment interactions should be included in our model to better unravel the etiology of epigenetic formation and its impact on complex diseases. Recent studies show that epigenetic modifications may be reprogrammed during meiosis or transmitted from generation to generation [32, 33]. Such a dynamic feature of epigenetic modifications can be tested by developing a family-based case–control design. Currently, GWAS have been increasingly used in genetic research in light of their power to comprehend the genetic control of complex traits or diseases. Stimulated by GWAS, epigenome-wide

association studies (EWAS) have arisen as a new approach, during the past few years, to systematically study epigenetic control [34, 35], and there is now a trend to merge GWAS and EWAS [19] into a single design for dissecting the genetic regulation of complex traits or diseases. Statistical high-dimensional models for analyzing all SNPs at the same time in GWAS have now been available [36], which can be extended to capture genome-wide epigenetic control. There is no doubt that a new high-dimensional model for EWAS will enable geneticists to draw a more comprehensive picture of the epigenetic landscape for complex human diseases [19].

Key points

- Epigenetic modifications through DNA methylation, chromatin remodeling, histone modification or RNA interference have been thought to play a pivotal role in generating and maintaining new quantitative genetic variation.
- There is a pressing need on developing quantitative models that can identify and estimate epigenetic variants that contribute to complex diseases, gleaning new insight into the pattern and form of epigenetic variation.
- We integrate quantitative genetic and epigenetic principles into a case-control study framework to test the inheritance mode of genetic and epigenetic variants from GWAS. The integrative model facilitates the detection of epigenetic variants involved in human diseases.

Acknowledgements

We thank the four anonymous reviewers for their constructive comments regarding the model described and the mechanisms and forms of epigenetic modifications, which help to improve the presentation of this article.

FUNDING

This work is partially supported by NSF/IOS-0923975, NIH/UL1RR0330184, Changjiang Scholars Award and ‘Thousand-person Plan’ Award. CTSA Grant Number KL2 RR024138 from the National Center for Research Resources (NCR) and the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH) and NIH roadmap for Medical Research.

References

1. Richard EJ. Quantitative epigenetics: DNA sequence variation need not apply. *Genes Dev* 2009;**23**:1601–5.
2. Richards EJ. Natural epigenetic variation in plant species: a view from the field. *Curr Opin Plant Biol* 2011;**14**:204–9.
3. Rutherford SL, Henikoff S. Quantitative epigenetics. *Nat Genet* 2003;**33**:6–8.
4. Richards EJ. Inherited epigenetic variation—revisiting soft inheritance. *Nat Rev Genet* 2006;**7**:395–401.
5. Richards CL, Bosdorp O, Pigliucci M. What role does heritable epigenetic variation play in phenotypic evolution? *Bioscience* 2010;**60**:232–7.
6. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007;**447**:433–40.
7. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci USA* 2010;**107**:1757–64.
8. Johannes F, Porcher E, Teixeira FK, et al. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* 2009;**5**:e1000530.
9. Eichten SR, Swanson-Wagner RA, Schnable JC, et al. Heritable epigenetic variation among maize inbreds. *PLoS Genet* 2011;**7**(11):e1002372.
10. Bell JT, Spector TD. A twin approach to unraveling epigenetics. *Trends Genet* 2011;**27**:116–25.
11. Johannes F, Colot V, Jansen RC. Epigenome dynamics: a quantitative genetics perspective. *Nat Rev Genet* 2008;**9**:883–90.
12. Maher B. Personal genomes: the case of the missing heritability. *Nature* 2008;**456**:18–21.
13. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
14. Eichler E, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010;**11**:446–50.
15. Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics* 2009;**182**:845–50.
16. Tal O, Kisdi E, Jablonka E. Epigenetic contribution to covariance between relatives. *Genetics* 2010;**184**:1037–50.
17. Johannes F, Colome-Tatche M. Quantitative epigenetics through epigenomic perturbation of isogenic lines. *Genetics* 2011;**188**:215–27.
18. Wang Z, Wang ZH, Wang JX, et al. A quantitative genetic and epigenetic model of complex traits. *BMC Bioinformatics* 2012;**13**:274.
19. Rakyen VK, Down TA, Balding DJ, et al. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;**12**:529–41.
20. Rakyen VK, Hildmann T, Novik KL, et al. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol* 2004;**2**:e405.
21. Frigola J, Song J, Stirzaker C., et al. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat Genet* 2006;**38**:540–9.
22. Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics*. London: Longman, 1996.
23. Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates, 1998.
24. Liu T, Thalamuthu A, Liu JJ, et al. Asymptotic distribution for epistatic tests in case-control studies. *Genomics* 2011;**98**:145–51.
25. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. *Science* 2001;**293**:1089–93.

26. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010;**465**: 721–7.
27. Wang Z, Liu T, Lin Z, *et al.* A general model for multilocus epistatic interactions in case-control studies. *PLoS ONE* 2010;**5**(8):e11384.
28. Zhang L, Liu R, Wang Z, *et al.* Modeling haplotype-haplotype interactions in case-control genetic association studies. *Front Genet* 2012;**3**:2.
29. Smith LM, Weigel D. On epigenetics and epistasis: hybrids and their non-additive interactions. *EMBO J* 2012;**31**: 249–50.
30. Waterland RA, Jirtle RL. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol* 2003;**23**:5293–300.
31. Furrow RE, Christiansen FB, Feldman MW. Environment-sensitive epigenetics and the heritability of complex diseases. *Genetics* 2011;**189**:1377–87.
32. Wang CG, Wang Z, Prows DR, *et al.* A computational framework for the inheritance of genomic imprinting for complex traits. *Brief Bioinformatics* 2012;**13**:34–45.
33. Wang CG, Wang Z, Luo JT, *et al.* A model for transgenerational imprinting variation in complex traits. *PLoS ONE* 2010;**5**(7):e11396.
34. Roux F, Colome-Tatche M, Edelist C, *et al.* Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* 2011;**188**:1015–7.
35. Feinberg AP, Irizarry RA, Fradin D, *et al.* Personalized epigenomic signatures that are stable over time and covary with body mass index. *Sci Transl Med* 2011;**3**(65):65er1.
36. Li JH, Das K, Fu GF, *et al.* The Bayesian lasso for genome-wide association studies. *Bioinformatics* 2011;**27**: 516–23.