10-2015

Prediction of Plant Height in *Arabidopsis thaliana* Using DNA Methylation Data

Yaodong Hu
University of Wisconsin-Madison

Gota Morota
University of Nebraska- Lincoln, morota@vt.edu

Guilherme J. M. Rosa
University of Wisconsin-Madison

Daniel Gianola
University of Wisconsin-Madison

Prediction of Plant Height in *Arabidopsis thaliana* Using DNA Methylation Data

Yaodong Hu,*,[1] Gota Morota,† Guilherme J. M. Rosa,*,‡ and Daniel Gianola*,‡,§
*Department of Animal Sciences, ‡Department of Biostatistics and Medical Informatics, and §Department of Dairy Science, University of Wisconsin, Madison, Wisconsin 53706, and †Department of Animal Science, University of Nebraska, Lincoln, Nebraska 68583

ABSTRACT Prediction of complex traits using molecular genetic information is an active area in quantitative genetics research. In the postgenomic era, many types of -omic (*e.g.*, transcriptomic, epigenomic, methylomic, and proteomic) data are becoming increasingly available. Therefore, evaluating the utility of this massive amount of information in prediction of complex traits is of interest. DNA methylation, the covalent change of a DNA molecule without affecting its underlying sequence, is one quantifiable form of epigenetic modification. We used methylation information for predicting plant height (PH) in *Arabidopsis thaliana* nonparametrically, using reproducing kernel Hilbert spaces (RKHS) regression. Also, we used different criteria for selecting smaller sets of probes, to assess how representative probes could be used in prediction instead of using all probes, which may lessen computational burden and lower experimental costs. Methylation information was used for describing epigenetic similarities between individuals through a kernel matrix, and the performance of predicting PH using this similarity matrix was reasonably good. The predictive correlation reached 0.53 and the same value was attained when only preselected probes were used for prediction. We created a kernel that mimics the genomic relationship matrix in genomic best linear unbiased prediction (G-BLUP) and estimated that, in this particular data set, epigenetic variation accounted for 65% of the phenotypic variance. Our results suggest that methylation information can be useful in whole-genome prediction of complex traits and that it may help to enhance understanding of complex traits when epigenetics is under examination.

KEYWORDS epigenetics; DNA methylation; MeDIP-Chip; phenotypic prediction; RKHS regression; GenPred; shared data resource

EPIGENETICS focuses on heritable changes of genetic materials that do not reside in the sequence of DNA, called epigenetic modifications (Riggs *et al.* 1996; Riggs and Porter 1996). Major forms of these changes are DNA methylation, histone modification, and noncoding RNAs (ncRNAs) (Rivera and Bennett 2010). DNA methylation is the most common epigenetic modification, which can have various forms depending on the targeting nucleotide of the modification (Ratel *et al.* 2006). In vertebrates and flowering plants, it is usually referred to as the covalent addition of a methyl group ($-CH_3$) to the 5-position carbon atom (^5C) of the cytosine pyrimidine ring, resulting in 5-methylcytosine (^{m5}C) (Jeltsch 2002; Meissner *et al.* 2005; Vanyushin 2006). Thus, "DNA methylation" stands for ^{m5}C throughout this article. Histone modification is the multivalent modification of histone tails of the core histones, which can be acetylation, methylation, phosphorylation, ubiquitination, and symoylation (Kouzarides 2007; Ruthenburg *et al.* 2007). Both DNA methylation and histone modification interact with the entering and binding of transcription factors (TFs) to the DNA molecule such that gene expression is altered. Usually, DNA methylation is associated with reduced gene expression (Bird 1984; Razin and Cedar 1991; Lim and Maher 2010) and histone modification can either enhance or repress expression, according to different modification targets (*e.g.*, which amino acids are at the histone tail) and modification types (*e.g.*, methylation or acetylation) (Berger 2002; Cheung and Lau 2005). Recently, ncRNAs were found to be composed of a hidden layer of internal signals that control various levels of gene expression associated with physiological and developmental processes. Their role in epigenetic regulation has been acknowledged as well (Zhou *et al.* 2010; Kaikkonen *et al.* 2011).

doi: 10.1534/genetics.115.177204
Manuscript received April 8, 2015; accepted for publication August 2, 2015; published Early Online August 6, 2015.
Available freely online through the author-supported open access option.
Data used in this article are from Johannes *et al.* (2009) and the Gene Expression Omnibus data repository under accession no. GSE37284.
[1]Corresponding author: Department of Animal Sciences, University of Wisconsin, 1675 Observatory Dr., Madison, WI 53706. E-mail: yhu32@wisc.edu

Epigenetic modifications have an important role in gene expression regulation, and malfunctioning of the regulation process can have severe consequences. In epidemiology and human genetics, many diseases and disorders, including cancer, have been confirmed to have an epigenetic basis (Jones and Baylin 2002, 2007; Jiang *et al.* 2004; Esteller 2008; Pembrey 2012; Tollefsbol 2012). For example, Prader–Willi syndrome (PWS) and Angelman syndrome (AS) are sister imprinting-related disorders involving deletion of DNA segments derived from different parents at the same genomic region (Meijers-Heijboer *et al.* 1992; Nicholls *et al.* 1998; Cassidy *et al.* 2000). Another example of epigenetics-related diseases is that of oncogenes; these exist in almost everyone's genome while only a small proportion of the population develops a cancer. Here, the promoter region of a tumor suppressor gene is usually unmethylated such that the gene is expressed normally and, therefore, it prevents the formation of a tumor. In cases where there is hypermethylation in the promoter region, the tumor suppressor is deactivated and a cancer develops (Jones and Baylin 2002; Robertson 2002; Egger *et al.* 2004).

Due to the potentially important role of epigenetics in diseases, epigenome-wide association studies (EWAS), a counterpart of genome-wide association studies (GWAS) at the epigenome level, have been conducted in recent years (MacArthur 2008; Rakyan *et al.* 2011; Bell 2013), aiming at finding associations between epigenetic polymorphisms and traits of interest, instead of using DNA polymorphisms (*e.g.*, SNPs). Although epigenetic regulation is not restricted to DNA methylation, the latter is the most commonly used biomarker in EWAS at present, because it is more stable and easier to be quantified than other epigenetic regulatory mechanisms (Flanagan 2015). In EWAS, DNA methylation across the whole genome is converted into a certain measurement reflecting the "methylation level," using methylation-sensitive enzyme digestion (Waalwijk and Flavell 1978; Kaput and Sneider 1979), methylated DNA immunoprecipitation (MeDIP) (Weber *et al.* 2005), or bisulfite sequencing (BS-Seq) that combines next-generation sequencing techniques with bisulfite conversion (Frommer *et al.* 1992), with BS-Seq being the most popular method used in methylation profiling. In BS-Seq, a DNA sample is treated with sodium bisulfite, which can convert unmethylated cytosine into uracil, whereas methylated cytosine is intact. Uracil is read as thymine in polymerase chain reaction (PCR) and sequence alignment after PCR amplification gives the counts of C (originally methylated cytosine) and T (originally unmethylated cytosine) at a single-base resolution. The ratio $C/(C+T)$ gives the absolute methylation level at that base, which is referred to as the β-value in methylation profiling literature and is usually considered as the "gold standard" in methylation quantification (Krueger *et al.* 2012). Once the methylation level is obtained, statistical methods are then applied to find associations between the "methylation profile" and the trait of interest in a selected sample.

Although some diseases associated with dysregulation of epigenetic modification at some genomic region have been found, EWAS has similar drawbacks to GWAS: it is difficult to estimate how much variation in phenotypes, especially for complex traits, is explained by epigenetic polymorphisms, even if there is evidence that they contribute to phenotypes, either biologically or statistically. Two studies attempting to solve this question have been published in recent years, with *Arabidopsis thaliana* used as experimental material (Johannes *et al.* 2009; Reinders *et al.* 2009). In Johannes *et al.* (2009), a wild-type inbred line was chosen as the paternal founder and a *ddm1* mutant was used as the maternal founder. The *ddm1* mutant was genetically identical to the wild type, except for the *DDM1* locus and few other loci. The *DDM1* locus encodes an ATPase chromatin remodeler that is involved in methylation maintenance, and the *ddm1* mutant used in their study was featured by a whole-genome-wide demethylation. The F_1 generation was obtained by crossing the wild-type (as male) and *ddm1* mutant (as female), and then it was backcrossed with the wild type (as male) to create the backcross generation (BC_1). The BC_1 individuals were selfed for several generations to construct a population of epigenetic recombinant inbred lines (epiRILs). In total, 505 epiRILs were obtained by Johannes *et al.* (2009) after four generations of selfing starting from the BC_1 generation. Since these 505 lines were (almost) isogenic at the DNA level and differed only in methylation profile, all observable phenotypic variation was then regarded as due to epigenetic and environmental factors, with the impact of genetic polymorphism at the DNA level ruled out. By examining plant height and flowering time, Johannes *et al.* (2009) found that epigenetics contributed $\sim$30% of the phenotypic variation. A similar approach was used in Reinders *et al.* (2009) with the only difference being that the genetic polymorphism in the two parental *Arabidopsis* lines was at the *Met1-3* locus, which also has an impact on the whole-genome methylation level, and that instead of a backcrossing to a parental line, selfing of F_1 was adopted. At the end of the F_8 generation, 68 epiRILs were obtained.

Both Johannes *et al.* (2009) and Reinders *et al.* (2009) found that epigenetic variation contributed to a considerable proportion of phenotypic variation, hinting that epigenetic information may help prediction of quantitative traits. When using DNA polymorphisms, whole-genome-enabled prediction models can be viewed as an extension of the single-marker regression models used in GWAS, where instead of finding genomic regions that may be associated with a complex trait, integrating all marker information for prediction and/or artificial selection is the ultimate goal. In a similar context, EWAS can also be extended for prediction, using data mining and machine learning techniques. Because methylation profiles can explain phenotypic variation and it is widely believed that DNA methylation is the most stable epigenetic modification that can be retained in either mitosis or meiosis, perhaps prediction can be enhanced by using methylation data, as foreseen by González-Recio (2012). In this study, therefore, we used DNA methylation data for building statistical models suitable for prediction purposes, with the expectation that this information could potentially supplement that from DNA polymorphisms.

Materials and Methods

Data

This study used phenotypic and methylation data. The phenotypic data set is from Johannes *et al.* (2009), and it contains measurements of plant height (PH) and flowering time (FT) collected in two greenhouses for 505 *Arabidopsis* epiRILs and 2 parental lines. These data were analyzed by Johannes *et al.* (2009), using a mixed-effects model, to explore the proportion of phenotypic variance explained by different effects. Their model used greenhouses and microenvironments (*i.e.*, individual planting plots in the greenhouse) as fixed effects and the 505 epiRILs as a random factor. Greenhouse explained 39.61% and 2.45% of phenotypic variance for FT and PH, respectively, and microenvironment explained 4.12% and 0.086% of phenotypic variance for these two traits; the variance explained by random epiRIL effects accounted for ~30% for both traits. Because the microenvironment arrangement data are no longer available (F. Johannes, personal communication), we decided to perform the analysis on PH only, as FT was apparently more strongly affected by this factor. The methylation data were downloaded from the Gene Expression Omnibus data repository (accession no. GSE37284). In this data set, 123 of the 505 epiRILs and the 2 parental lines were epi-profiled, using MeDIP with a customer-designed array chip. Each line was examined at 711,320 probes (loci) located on five *Arabidopsis* chromosomes. Each probe is associated with two values: one is the rescaled $\log_2$ of the signal/background intensity ratio, which describes the enrichment of methylated cytosine proxied by that probe. This information is referred to as methyl values in subsequent discussion, and a higher methyl value indicates higher level of methylation. The other value is methylation status [methylated (M), intermediately methylated (I), or unmethylated (U)] predicted from the methyl values. Note that the methyl values are generated from enrichment intensity ratios, so these are relative, rather than absolute, values. Due to this reason, there are typically no threshold values that can be used to perform methylation status calls, and hence the status was predicted using a hidden Markov model (Colomé-Tatché *et al.* 2012), a commonly used tool in bioinformatic analysis. This predicted methylation status is referred to as methyl status hereafter. There were no missing values in the methylation data, and after removing epiRILs without phenotypic data, 114 lines remained for subsequent analysis. Therefore, each epiRIL used in this study has 1 phenotypic record on PH and paired methyl-values/methyl-status records at each of 711,320 probes (loci). For more detailed information about the methylation data, see Colomé-Tatché *et al.* (2012) and the NCBI description page. A description on data processing was given in Cortijo *et al.* (2014a).

Methods and prediction models

The methylation data described above were used by Cortijo *et al.* (2014b) to map epigenetic QTL (epiQTL) contributing to root length and FT, and three major epiQTL were found for both traits. Using analysis of variance, it was found that the broad sense (epi)heritabilities of these two traits were ~60%, and major epiQTL explained 87% and 60% of (epi)heritability in the two traits, respectively. Due to the strong contribution of methylation to variation of phenotype, we decided to explore the predictive power of this information, as suggested by González-Recio (2012). Here, we built whole (epi)genome prediction models that are analogous to whole-genome prediction models, where instead of SNP markers, methylation information was used as predictor variables. Most genome-enabled prediction studies (*e.g.*, Meuwissen *et al.* 2001; de los Campos *et al.* 2013) use a linear model with the form

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is a vector of n phenotypic records, μ is an unknown constant (intercept) common to all individuals, $\mathbf{b}$ is a vector of fixed effects with associated incidence matrix $\mathbf{X}$, $\mathbf{W}$ is an $n \times p$ matrix possessing SNP genotypic codes (*e.g.*, $W_{ij} = 0, 1,$ or 2), and $\boldsymbol{\alpha}$ is a $p \times 1$ vector of regression coefficients associated with all SNP loci. Model 1 is more statistical than biological when the $\mathbf{W}$ matrix contains epigenetic information, compared to when using genomic information such as SNP markers, with the reasons being stated next. When performing genomic prediction using SNPs, $\boldsymbol{\alpha}$ represents allelic substitution effects at these marker loci, an important concept in quantitative genetics; hence, a statistical regression coefficient can be linked to a quantitative genetic parameter. Since such a concept does not exist in quantitative epigenetic analysis, this implies that a model with this form may not lead itself to interpretability of underlying biological processes. Therefore, we adopted kernel methods for prediction purposes, from which a biological interpretation is available with less difficulty.

Kernel methods: theory: In kernel regression, phenotypes and predictor variables are often linked nonlinearly, via a kernel function. In a regression problem without nuisance variables, the relationship between an observation y_i and its corresponding covariates $\mathbf{x}_i$ is generally written as

$$y_i = g(\mathbf{x}_i) + e_i, \tag{2}$$

where y_i is the observation on the ith subject and $\mathbf{x}_i$ is, say, a $p \times 1$ vector of covariates measured on i; $g(.)$ is some function (usually unknown); and e_i is the model residual. For the purpose of describing the kernel methods, it is assumed that phenotypes (y's) and regression covariates ($\mathbf{x}$'s) are centered, so Equation 2 does not include an intercept. In standard linear regression, $g(\mathbf{x}_i)$ is $\mathbf{x}_i'\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is a vector of unknown coefficients to be inferred. The most common solution for the weights $\boldsymbol{\omega}$ is obtained by using ordinary least squares (OLS). In whole-genome prediction of complex traits, many methods use this functional form but assign some penalties (or priors) to $\boldsymbol{\omega}$ because the "curse of dimensionality" makes

OLS not applicable, and often Bayesian techniques are employed (Gianola *et al.* 2009; Gianola 2013). The linear additive model often provides a reasonable approximation to the underlying statistical architecture of a complex trait and it is easy to interpret. However, nonadditive gene action, for example epistasis, is usually not accounted for, which may lead to incorrect attributions of genetic variation .

One can define $g(\mathbf{x}_i) = E(y_i|\mathbf{x}_i)$ as the conditional expectation of y_i in Equation 2, given $\mathbf{x}_i$, which can be inferred using the Nadaraya–Watson estimator (Nadaraya 1964; Watson 1964), having the form (Silverman 1986; Gianola *et al.* 2006)

$$\hat{g}(\mathbf{x}) = \sum_{i=1}^{n} y_i \kappa_h(\mathbf{x}_i - \mathbf{x}). \tag{3}$$

In genome-enabled prediction using high-density markers, n is the number of individuals, $\mathbf{x}_i$ is the $p \times 1$ vector of SNP marker genotypes of individual i, $\mathbf{x}$ is the focal point at which the kernel function $\kappa_h(.)$ is evaluated, and h is a smoothing parameter of the kernel function. Because $\mathbf{x}_i$ possesses the marker information of individual i, $\kappa_h(\mathbf{x}_i, \mathbf{x}_j)$ measures the "genomic distance" between individuals i and j by definition. Therefore, the $n \times n$ symmetric matrix $\mathcal{K}_h = \{\kappa_h(\mathbf{x}_i, \mathbf{x}_j)\}$ measures the pairwise genomic distance of all individuals. According to Gianola and Van Kaam (2008), this kernel treatment can be written as (the "dual formulation") the linear regression model

$$\mathbf{y} = \mathcal{K}_h \boldsymbol{\alpha} + \mathbf{e}, \tag{4}$$

where $\mathbf{y}$ is an $n \times 1$ vector; $\mathcal{K}_h$ is an $n \times n$ symmetric, positive definite matrix; $\boldsymbol{\alpha}$ is an $n \times 1$ vector of regression coefficients; and $\mathbf{e}$ is the model residual with assumption $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance. Under the reproducing kernel Hilbert spaces framework (*e.g.*, Gianola and Van Kaam 2008), one assumes that $\boldsymbol{\alpha}|h \sim N(\mathbf{0}, \mathcal{K}_h^{-1}\sigma_{\mathcal{K}}^2)$ and, because $\mathcal{K}_h$ is symmetric and invertible, $\hat{\boldsymbol{\alpha}}$ is estimated as the solution to

$$\left(\mathcal{K}_h + \frac{\sigma_e^2}{\sigma_{\mathcal{K}}^2}\mathbf{I}\right)\hat{\boldsymbol{\alpha}} = \mathbf{y}. \tag{5}$$

Above, $\sigma_{\mathcal{K}}^2$ is the variance captured by the kernel. The vector $\mathcal{K}_h\hat{\boldsymbol{\alpha}}$ estimates the vector of genetic effects marked by SNPs, that is, $g(\mathbf{x})$.

Alternatively, starting from $\mathbf{y} = \mathbf{g} + \mathbf{e}$, one can minimize a loss function with form

$$\ell(\mathbf{g}|\lambda) = \|\mathbf{y} - \mathbf{g}\|^2 + \lambda\|\mathbf{g}\|_{\mathcal{H}}^2, \tag{6}$$

where λ is a regularization parameter and $\|\mathbf{g}\|_{\mathcal{H}}^2$ is the squared norm of $\mathbf{g}$ under a Hilbert space $\mathcal{H}$. According to the representer theorem of Kimeldorf and Wahba (1971), the objective function $\mathbf{g}$ is reduced to $\mathcal{K}_h\boldsymbol{\alpha}$, as in Equation 4, and Equation 6 becomes $\ell(\boldsymbol{\alpha}|\lambda) = (\mathbf{y} - \mathcal{K}_h\boldsymbol{\alpha})'(\mathbf{y} - \mathcal{K}_h\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}'\mathcal{K}_h\boldsymbol{\alpha}$. When minimizing $\ell(\boldsymbol{\alpha}|\lambda)$ by taking its first derivative with respect to $\boldsymbol{\alpha}$, Equation 5 is retrieved if $\lambda = \sigma_e^2/\sigma_{\mathcal{K}}^2$ is assumed. Because optimization of the penalty function is carried out under a Hilbert space, this approach is known as reproducing kernel Hilbert spaces (RKHS) regression, first proposed in computer sciences and machine learning (Aronszajn 1950; Kimeldorf and Wahba 1971; Wahba 1990, 1999, 2002).

Equation 4 has the same form as the "animal model" (*e.g.*, Henderson 1984; Mrode 2014) widely used in animal breeding,

$$\mathbf{y} = \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{7}$$

where $\mathbf{u}$ is the vector of infinitesimal additive effects and $\mathbf{Z}$ is the associated incidence matrix. Assumptions for this model are $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, and the best linear unbiased predictor (BLUP) of $\mathbf{u}$ can be obtained by solving

$$\left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{A}^{-1}\right)\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{y}, \tag{8}$$

where σ_u^2 and σ_e^2 are the additive genetic and residual variances, respectively. Here, the additive relationship matrix $\mathbf{A}$ can be interpreted as a kernel matrix measuring the kinship between individuals based on pedigree, as discussed in de los Campos *et al.* (2009) and in Morota and Gianola (2014). Hence, the conventional animal model [pedigree-based BLUP (P-BLUP)] is a special case of RKHS regression. Similarly, the genomic BLUP (G-BLUP) proposed by VanRaden (2008) uses a genomic relationship matrix $\mathbf{G} \propto \mathbf{X}\mathbf{X}'$, with $\mathbf{X}$ being the $n \times p$ incidence matrix of marker genotypes, in lieu of the $\mathbf{A}$ matrix derived from pedigree. G-BLUP exploits "realized" relationship between individuals, using genomic information covering the entire genome. Therefore, G-BLUP is also a special case of RKHS regression. For more details on RKHS regression and its applications to animal breeding, see Gianola *et al.* (2006), Gianola and Van Kaam (2008), Gianola and de los Campos (2008), González-Recio *et al.* (2008), de los Campos *et al.* (2009, 2010), Morota *et al.* (2013), and Morota and Gianola (2014).

In general, the role of a kernel matrix in RKHS regression is to convey pairwise similarity between individuals, using a certain type of input information, with methylation profiles used here. Although the choice of the kernel function is arbitrary, as any positive-definite function can be used as a kernel function, multiple factors may affect its choice in practice. For example, the diffusion kernel adopted by Morota *et al.* (2013) has a distance function (Manhattan distance) that may not be optimal for real numbers. Hence, we chose a Gaussian kernel for the continuous methyl values. By definition, the (i,j)th element of the Gaussian kernel $\mathbf{K}$ is calculated as

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{h}\right), \tag{9}$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, in our case the $711,320 \times 1$ vectors of methyl values of epiRILs i and j, and h is the bandwidth parameter of the

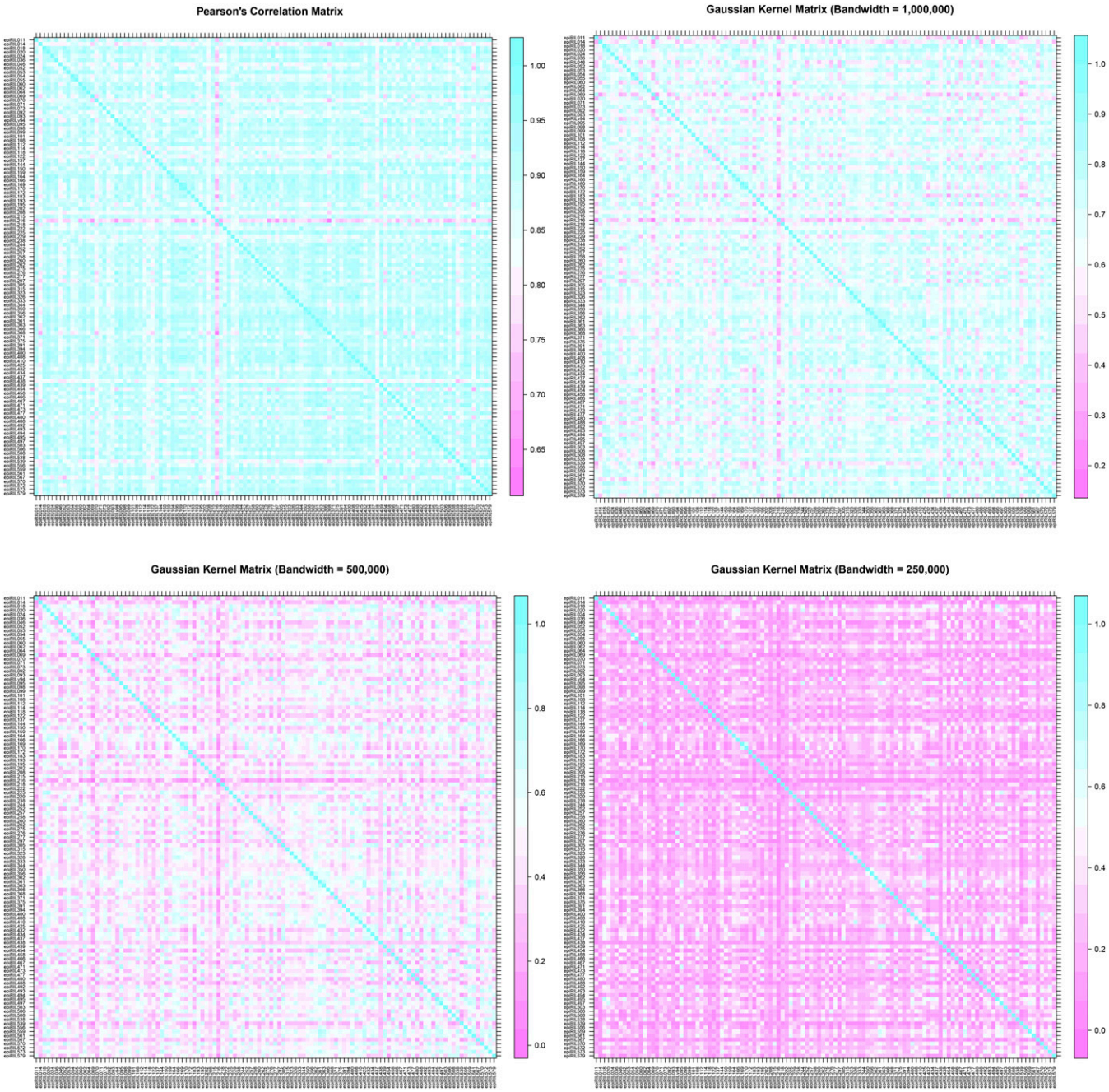

Figure 1 Visualization of several kernel matrices. The four matrices displayed are Pearson's correlation matrix (top left) and Gaussian kernels with bandwidth parameters of 1,000,000 (top right), 500,000 (bottom left), and 250,000 (bottom right).

kernel, which controls the smoothness of the fitted surface. The choice of the bandwidth parameter is important since it affects the performance of the regression. A number of algorithms have been proposed to optimize the bandwidth parameter (Jones *et al.* 1996). Here, we determined the optimal bandwidth parameter using a grid search approach under cross-validation, aiming at finding a value that maximized the predictive correlation within a model setting.

From the definition of the Gaussian kernel, all diagonal entries of the kernel matrix are 1, since the Euclidean distance between a vector and itself is always zero. Also, as the distance increases, K_{ij} approaches zero. Hence, the entries of **K** range between 0 and 1, making the kernel act as a correlation matrix. Therefore, we considered Pearson's correlation matrix **P** as a naive kernel, where $P_{ij} = \mathrm{Corr}(\mathbf{x}_i, \mathbf{x}_j)$. Advantages of using the **P** matrix are computation related: (1) it is easy to obtain, and (2) tuning a bandwidth parameter is not needed. Comparisons between prediction performances obtained using the **P** and the **K** kernels are described later. A graphical comparison between the **P** and **K** kernels is shown in Figure 1. In Figure 1, the plot at the top left corner shows the **P** matrix created from the methyl-values data. Most between-lines

correlations range from 0.7 to 0.9 and only few pairwise correlations are <0.65. The other three plots represent a **K** matrix with various h values. It can be seen that h has a big impact on the values of the **K** matrix. When h is large (1,000,000, top right corner), the majority of the entries range from 0.4 to 0.5; for intermediate h (500,000, bottom left corner), most entries are between 0.2 and 0.7; when h is small (250,000, bottom right corner), almost all entries are <0.5 except for the diagonal elements.

Given a kernel $\mathcal{K}$ and a vector of fixed effects **b** (in our case the greenhouses only), the prediction model can be written in matrix form as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathcal{K}\boldsymbol{\alpha} + \mathbf{e}, \tag{10}$$

where **y** is the vector of phenotypic records (PH here); μ is an unknown intercept common to all observations; **X** is the incidence matrix of fixed greenhouse effects; $\boldsymbol{\alpha}$ is the random vector of regressions on the kernel associated with epigenetic variation, with assumed distribution $N(\mathbf{0}, \mathcal{K}^{-1}\sigma^2_{\mathcal{K}})$, where $\sigma^2_{\mathcal{K}}$ is a variance component associated with the kernel; and **e** is the model residual with distribution $N(\mathbf{0}, \mathbf{I}\sigma^2_e)$.

Prediction using preselected probes: Our main goal is to build prediction models, using epigenetic information as a potential supplement to genomic variation (*e.g.*, SNP markers), as foreseen by González-Recio (2012). In animal and plant breeding, a training population with thousands of individuals is usually needed. However, methylation profiling experiments are extremely expensive, at least at present. Thus, cost is usually a main consideration in epigenetic studies, and data sets with hundreds of profiled individuals are commonly viewed as large-scale experiments. Due to Meissner *et al.* (2005), a molecular genetic technique called reduced representation bisulfite sequencing (RRBS) has been used to take only a small subset of all available probes as proxies to describe the methylation level of the whole genome, which may reduce experimental costs drastically and make experiments executed on a larger cohort possible. According to the mechanisms of DNA methylation known so far, cytosine in a CpG dinucleotide context (cytosine followed by guanine, where "p" indicates the phosphate bond in between) is the main target of DNA methylation in eukaryotic cells. Thus, genomic regions with high CpG content may represent the methylation profile of the entire genome and hence are chosen for BS-Seq in RRBS (note that CpG content is different from CG content; the latter evaluates cytosine and guanine frequencies separately). In the mouse, according to Meissner *et al.* (2005), these selected regions comprised only ~12 Mb of the whole genome (<0.5%), but captured most variation at the methylation level. This suggests that a subset of representative probes may perhaps provide a similar predictive performance to that from all probes. If this is the case, prediction using representative probes would be less expensive and computing burden would be lessened because generating a kernel matrix is potentially time-consuming.

Considering the size of the murine (~3000 Mb) and the *Arabidopsis* (~120 Mb) genomes, we decided to select the top (see below) 10% of the profiled probes in *Arabidopsis* such that the genomic regions in which these probes reside had ~12 Mb in total. Thus, the cost needed for the experiment would not exceed the magnitude of what was suggested by RRBS in mouse. The criterion for this selection was based on the observed/expected (O/E) CpG ratio defined as

$$\frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \times \text{Total number of nucleotides in the sequence}$$

(Gardiner-Garden and Frommer 1987), which is a statistic describing the frequency of occurrence of CpG dinucleotides. Besides CpG dinucleotides, it has been found that trinucleotides CpHpG and CpHpH (H = A, C, or T) are target sites of DNA methylation in plants as well (Henderson and Jacobsen 2007; Lister *et al.* 2008). Thus, we also calculated the O/E ratio for these two trinucleotides. According to the reference genome (TAIR7, downloaded from http://www.arabidopsis.org), the total length of the *Arabidopsis* genome is 119,186,497 bp. With 711,320 probes on the designed chip, on average there is 1 probe for every 167 bp. The average length of all probe sequences is 55.2 bp (max 75 bp, min 50 bp), which means that the DNA segment between two probes is ~112 bp long, on average. Considering that 55 bp may not be an adequate length for calculating the O/E ratio with accuracy, especially for the two trinucleotides, we decided to extend the region of examination by 120 bp to the upstream of each probe. After this extension, the estimation of O/E CpG ratio is expected to be more accurate, and the number 120 was chosen because (1) it fills the gap between two probes, so this ensures that the whole genome is under examination, and (2) the overlap between adjacent probes after extension is reduced.

To make up 10% of total probes, we chose the top 5% probes with highest O/E ratio for CpG dinucleotides and the top 2.5% probes with highest O/E ratio for each of CpHpG and CpHpH. This 2:1:1 partition comes from the fact that in *Arabidopsis*, the fractions of ^{m5}C identified in CpG, CpHpG, and CpHpH contexts are about 55%, 23%, and 22%, respectively (Lister *et al.* 2008). As a result, we selected 35,585, 17,783, and 17,783 probes based on the CpG, CpHpG, and CpHpH contents, respectively, and ended up with 65,506 probes (9.2% of all probes) in total (with some overlap between contents of different contexts). After mapping back to the genome annotation file (TAIR7, downloaded from http://www.arabidopsis.org), it was found that within these 65,506 probes, 10,044 (15.3%) were located in promoter regions of genes, and these 10,044 probes covered 40.9% of total promoters; 12,074 (18.4%) were found in coding DNA sequences (CDS); 2329 and 2005 (3.6% and 3.1%) were in 5′-UTR and 3′-UTR regions, respectively; and 2534 (3.9%) probes were in pseudogenic exons. Also, 1418 and 16,258 (2.2% and 24.8% of the 65,506 preselected probes) were found in the intron and transposon regions, respectively, with the reference information provided by Cortijo *et al.* (2014a). Finally, 18,620 (28.4%) probes did not map

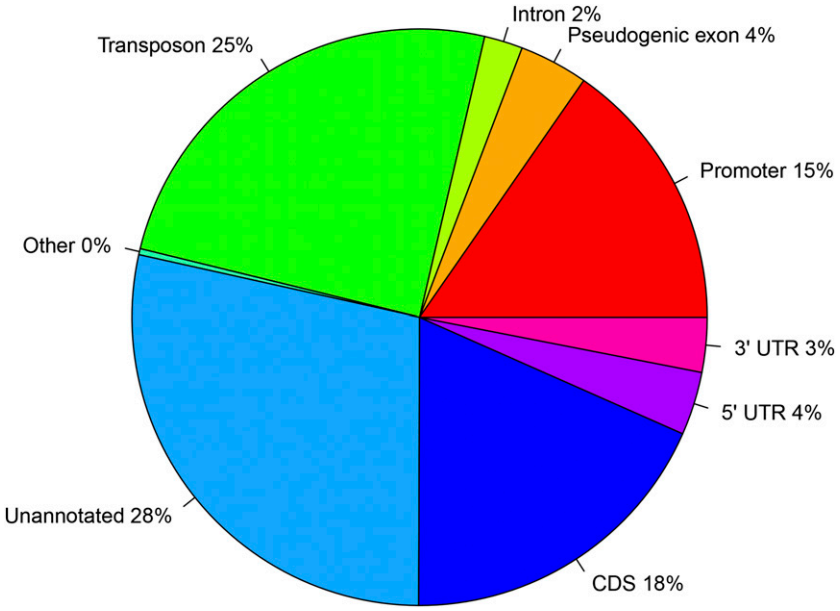

Figure 2 Distribution of the selected 65,506 probes. Most selected probes (~70%) were located in annotated regions (*e.g.*, CDS, 5′-UTR, 3′-UTR, promoter, intron, etc.).

to any annotated region according to the current annotation file. A graphical representation of the distribution of selected probes by genomic element groups is shown in Figure 2. In addition to a model using all available probes, a prediction model using these 65,506 preselected probes was built as well.

In their RRBS study, Meissner *et al.* (2005) reported that the representative subset covered $>$ 90% of gene promoter regions, while in our bioinformatic search, only 40% of the promoter regions were covered by preselected probes as described above. This is probably due to differences between species since the original RRBS method was developed in the mouse. Given the important role of gene promoter regions in epigenetic regulation of gene expression, we attempted to select a subset of probes with a different criterion such that more promoter regions could be covered. In epigenetics, CpG islands (CGIs) are CpG-rich regions that are usually unmethylated and located in the gene promoter region. In humans, at least 60~70% gene promoter regions overlap with CGIs (Illingworth and Bird 2009). CGI shores are close proximity regions (~2 kb of upstream or downstream) of CGIs (Portela and Esteller 2010). Recent studies suggested that 70% of differentially methylated regions in epigenetic reprogramming are associated with CGI shores (Doi *et al.* 2009; Ji *et al.* 2010). Therefore, probes located in CGI shores were also selected such that more gene promoter regions can be covered and these probes may constitute another (independent) subset capturing most variation of the whole-genome methylation profile. Following the definition of CGI given by Gardiner-Garden and Frommer (1987), we found 23,640 CGIs, and the probes located in the shore regions of these CGIs covered 65.6% of all promoters (Table 1). Thus, apart from different kernel matrices applied, prediction was performed using (1) all probes available in the data set, (2) preselected probes based on CpG/CpHpG/CpHpH contents (referred to as contents rule hereafter), or (3) preselected probes located in the CGI shore region (referred to as CGI rule hereafter).

Prediction using methyl-status data: When P-BLUP and G-BLUP are viewed as kernel methods, the **A**- and **G**-kernel matrices have an explicit biological meaning. For example, the kinship matrix **A** reflects the expected fraction of identical-by-descent alleles shared by a pair of relatives and the **G** matrix can be viewed as a realization of relationships given the observed molecular markers or as a "molecular similarity matrix" based on the DNA polymorphisms. Thus, variance components associated with **A** or **G** have a clear genetic basis. The correlation matrix **P** and any of the Gaussian kernels with specific bandwidth parameters used here, on the other hand, are constructed from methylation profiles and reflect only epigenetic similarity in some manner. Hence, variance components associated with these kernels do not have an easy biological interpretation except that of measuring a contribution to phenotypic variance. Further, when a Gaussian kernel **K** is used, the bandwidth parameter h has a large impact on the values in **K**, as depicted in Figure 1. As such, one may expect that various distinct $\hat{\sigma}^2_{\mathbf{K}}$ will be obtained when different values are assigned to h; hence, $\hat{\sigma}^2_{\mathbf{K}}/(\hat{\sigma}^2_{\mathbf{K}}+\hat{\sigma}^2_e)$ will vary

Table 1 Number of promoters covered by CGI (definition in Gardiner-Garden and Frommer 1987) shores

Chromosome	No. promoters	No. CGIs	No. promoters covered by CGI shores (%)
Chr1	6,354	1,716	4,111 (64.7)
Chr2	3,990	1,063	2,583 (64.7)
Chr3	4,902	1,363	3,289 (67.1)
Chr4	3,621	1,025	2,402 (66.3)
Chr5	5,677	1,624	3,709 (65.3)
Total	24,544	6,791	16,094 (65.6)

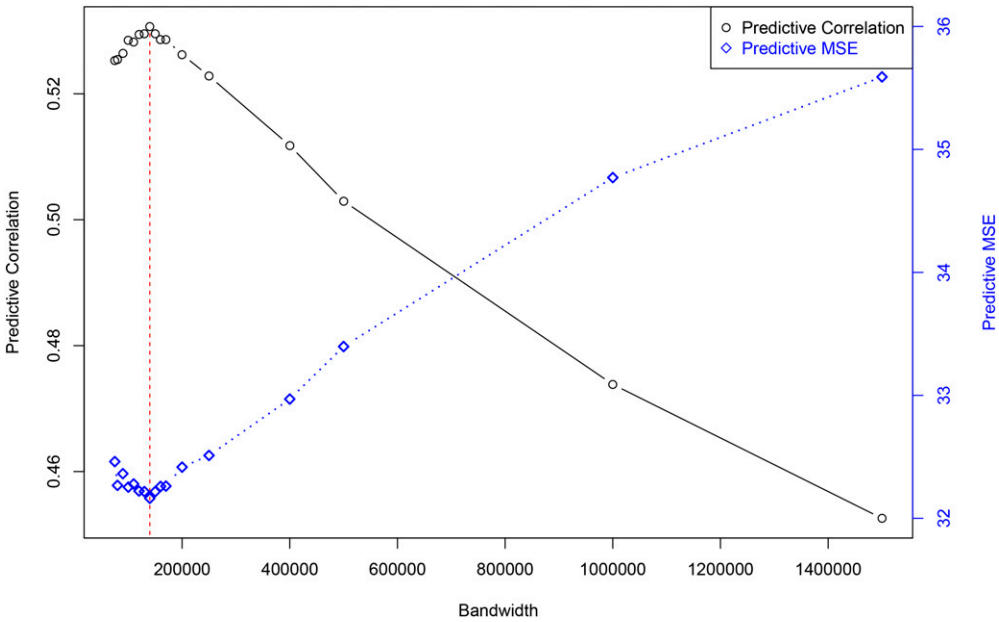

Figure 3 Predictive correlation and predictive MSE with various bandwidth parameters under LOO CV (all probes used). The kernel with the highest predictive correlation and lowest MSE is denoted by a dashed red line, where bandwidth is 140,000.

as well. To obtain a more meaningful partition of phenotypic variation explained by epigenetic polymorphisms, we built an additional kernel matrix for RKHS.

Because the methylation state of one copy at a single locus (*e.g.*, a single cytosine at a CpG dinucleotide) can be only methylated or unmethylated, the absolute methylation level β (as obtained by BS-Seq, for example) is always measured as a ratio that ranges between 0 and 1, with the numerator being the number of methylation incidences in a sample. Under some circumstances, methylation at the locus under investigation can be classified into one of the three categories: M, I, or U, according to the β-value at that locus (Meissner *et al.* 2008; Du *et al.* 2010). If two DNA segments with similar nucleotide sequence but different methylation status (*e.g.*, one is methylated and the other is not) are considered as two epialleles, this classification provides an approximation to the underlying "epigenotypes" such that M and U stand for the "epihomozygotes" for one of the two epialleles and I is the "epiheterozygote." Analogous to the SNP coding system, we can use 2, 1, and 0 to code M, I, and U and generate a kernel matrix mimicking the **G** matrix in G-BLUP (VanRaden 2008), which we call the epi-**G** matrix. This required little extra effort since methyl status was available in the methylation data set. However, this approach has some pitfalls: (1) when continuous methyl values are converted to discrete methyl status, information is lost; and (2) once a numeric coding is arrived at, many probes would be excluded from downstream analysis because their "epi-MAF" would be <0.05 (MAF, "minor allele frequency"). In the current data set, only 206,600 probes were kept for subsequent analysis after this epi-MAF filtering. Nevertheless, this epi-**G** kernel may be more biologically intuitive than a Gaussian kernel generated from methyl values since the numeric coding used to generate the epi-**G** kernel is an absolute count of a certain epiallele of an epigenotype. Thus, a prediction model can be built and implemented as in G-BLUP, and the variance component associated with epi-**G** would estimate the proportion of total variance explained by epigenetic variation with a clearer biological sense.

Data availability

Phenotypic data are from Johannes *et al.* 2009. Methylation data are downloaded froom Gene Expression Omnibus repository with accession number GSE37284.

Results

Prediction with different kernels

Considering that the data set had only 114 epiRILs, we used a leave-one-out cross-validation (LOO CV) for model evaluation throughout the study. When the correlation matrix **P** was used as a naive kernel, the predictive correlation was 0.384. When a Gaussian kernel was used, the predictive correlation varied according to the bandwidth parameter chosen. In this case, when all probes were used to create the kernel matrix, the best prediction performance was obtained when the bandwidth parameter was set to 140,000, and the predictive correlation was 0.531, with predictive mean squared error (MSE) = 32.16.

It can be seen that a reasonable predictive correlation was reached when using the Gaussian kernel, which performed much better than the correlation kernel. However, the bandwidth parameter played an important role in model performance (Figure 3). Taking the four kernels in Figure 1 as an example, **P** had entries ranging from $\sim$0.6 to 1, which means that the "dissimilarity" between each line must be distinguished within a 0.4 range. On the other hand, all three Gaussian kernels in Figure 1 ranged from $\sim$0 to 1, such that pairwise dissimilarity was better distinguished on a wider

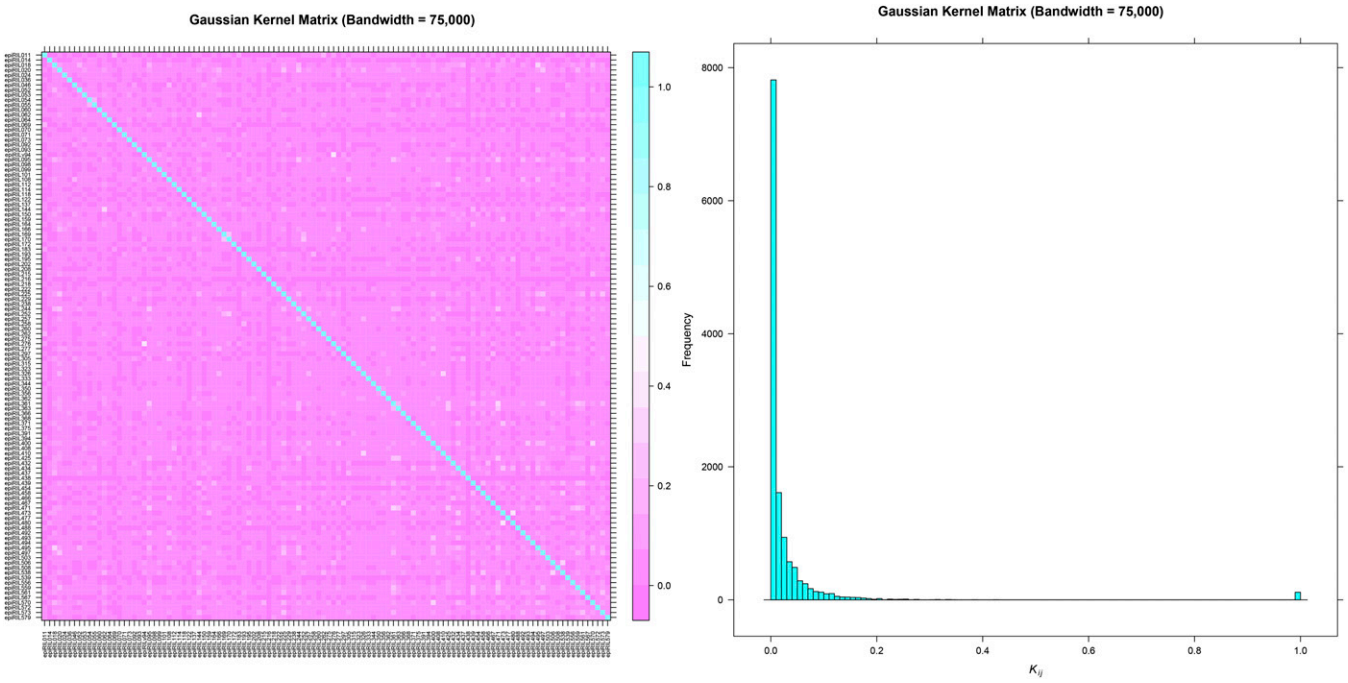

Figure 4 A Gaussian kernel with small bandwidth parameter is similar to an identity matrix. (Left) A visualization of a Gaussian kernel with bandwidth of 75,000, which makes it closer to an identity matrix than any of the other matrices in Figure 1. (Right) A histogram of the values in this kernel matrix showing that most values in this matrix are either exactly zero or very close to zero. A short bar at $K_{ij} = 1$ represents the diagonal elements of this kernel matrix.

scale. Thus, it was not surprising that the Gaussian kernel outperformed the correlation kernel. Predicting an unobserved record borrows information from observations on similar individuals. Thus, the "similarity" between lines matters. From the definition of the kernel matrix (Equation 9), the off-diagonal elements are close to zero if h is small (Figure 4). This makes the kernel matrix "confounded" with the identity matrix, which represents the variance–covariance structure of the model residuals. According to de los Campos *et al.* (2010), this type of kernel matrix captures "local" similarity, focusing mainly on the comparison of an individual with itself and a few other individuals with highest similarities. A "global" kernel with a larger bandwidth parameter, on the other hand, will also take into account comparisons between more (epi)genetically distant individuals. Therefore, the "optimal" bandwidth parameter should provide a balance between local and global comparisons between different lines, using the available data. Unless multiple kernels with different bandwidth parameters are fitted simultaneously (*e.g.*, Tusell *et al.* 2014), a kernel with an intermediate h is expected to provide the best predictive correlation (Figure 3, black solid line). A similar pattern was observed for predictive MSE (Figure 3, blue dotted line).

Prediction using preselected probes

For preselected probes, models using different kernel matrices were evaluated as well; again, the bandwidth parameter for the Gaussian kernel was determined based on a grid search via LOO CV. When using the **P** kernel, the predictive correlations for contents rule and CGI rule probes were 0.398 and 0.395, respectively, slightly higher than when all probes were used for prediction. When a Gaussian kernel was used, the highest predictive correlations for these two sets of preselected probes were 0.532 and 0.531, respectively, given an appropriate bandwidth parameter. This result was the same as when all probes were used (Table 2, Figure 5). As a comparison, we also drew 10 subsets of probes, each consisting of a random 10% of all available probes, to evaluate the usefulness of preselection of representative probes according to different criteria. Results showed that the predictive correlations using randomly selected probes were all lower than when using representative probes selected according to an explicit criterion, regardless of the kernel used in prediction.

Our results suggest that a properly selected subset of all probes is able to capture most variation at the methylome level. Therefore, prediction of a larger cohort with a limited budget is possible since only a small fraction of "loci" is needed for methylation profiling with computation time decreasing drastically. This could be very useful in livestock or crop

Table 2 Comparison between prediction results using all probes and preselected probes

Kernel		All probes	Contents rule probes	CGI rule probes
Correlation	$\text{Corr}(\mathbf{y},\hat{\mathbf{y}})$	0.384	0.398	0.395
	MSE	38.28	37.73	37.83
Gaussian	$\text{Corr}(\mathbf{y},\hat{\mathbf{y}})$	0.531	0.532	0.531
	MSE	32.16	32.08	32.13

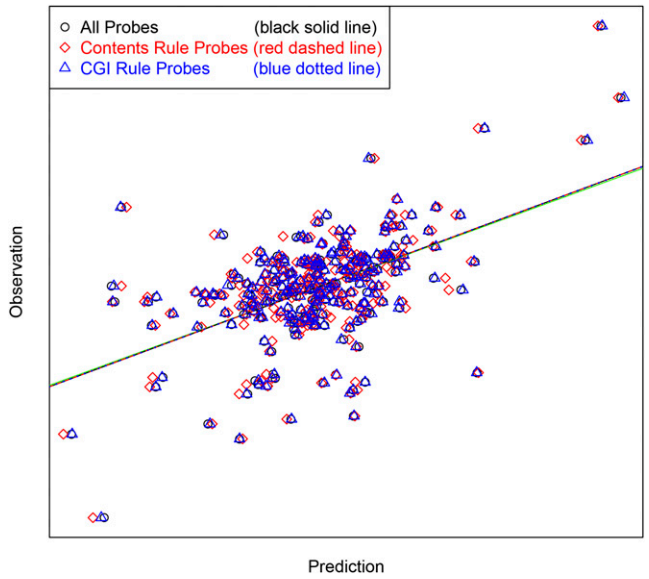

Figure 5 Graphical representation of prediction performance using different sets of probes in a Gaussian kernel. The green solid line is a 45° line passing through the origin, and the other three lines are fitted lines of regressing observation on predictions. No differences were observed for the three different sets of probes used in prediction.

production since there are usually thousands of individuals in a breeding program that need to be chipped (*i.e.*, methylation profiled), which is still very costly. Lower predictive correlations obtained using randomly selected probes indicated that the most relevant methylation variation is harbored in previously identified regions, *i.e.*, high CpG/CpGpH/CpHpH-content regions or CGI shore regions.

Prediction using the epi-G kernel

When using epi-**G** in prediction, the predictive correlation was 0.505 when all probes were used; predictive correlations were 0.494 and 0.507, if probes were preselected based on the contents or the CGI rules, respectively. Thus, the predictive correlation using epi-**G** was somewhat lower than that when a Gaussian kernel was used, perhaps due to the loss of information from discretization of methyl values. The estimated variance component associated with epi-**G** was 41.57 (SD = 1.69 under cross-validation) and the residual variance was estimated as 22.75 (SD = 0.45), so the proportion of total variance explained by epigenotype was 0.646, close to what was reported in Cortijo *et al.* (2014b). This proportion was 0.656 and 0.647 when only preselected probes (with two criteria, respectively) were used (Table 3). When using Gaussian kernels, on the other hand, the variance component associated with the kernel matrix represented 0.542 of the phenotypic variance (all probes used, bandwidth = 140,000), which was lower than with the epi-**G** kernel.

It is difficult to assess which kernel provides a more meaningful proportion of phenotypic variance explained by the methylation profile, since the true variance components are unknown. However, it is worth noting that in addition to a strong impact on predictive performance (Figure 3), the bandwidth parameter h had a big influence on variance component estimates as well (Figure 6). The estimated variance components associated with the Gaussian kernel were very large when h was large, and the proportion of phenotypic variation explained by the kernel matrix seemed excessive (up to 0.85). Note that the residual variance was essentially independent of the bandwidth parameter value. Therefore, caution needs to be exercised when interpreting the variance component associated with the kernel as variation explained by methylation polymorphisms. When using the epi-**G** kernel, on the other hand, the proportion of phenotypic variation explained by epigenetic polymorphisms seemed more reasonable, and predictions obtained using this kernel gave a better predictive correlation than when using the **P** kernel ($\text{Corr}(\mathbf{y},\hat{\mathbf{y}}) = 0.384$). Also, the regression of testing set observations on predicted values was 0.99, much higher than for the **P** kernel ($b_{\mathbf{y},\hat{\mathbf{y}}} = 0.90$, Figure 7).

Table 3 Estimated variance components associated with a Gaussian and an epi-G kernel

Kernel		All probes	Contents rule probes	CGI rule probes
Gaussian	$\text{Corr}(\mathbf{y},\hat{\mathbf{y}})$	0.531	0.532	0.531
	$\hat{\sigma}^2_{\mathcal{K}}$	25.59	23.75	25.61
	$\hat{\sigma}^2_e$	21.59	21.44	21.58
	$\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}}+\hat{\sigma}^2_e)$	0.542	0.525	0.543
epi-**G**	$\text{Corr}(\mathbf{y},\hat{\mathbf{y}})$	0.505	0.494	0.507
	$\hat{\sigma}^2_{\mathcal{K}}$	41.57	44.45	41.58
	$\hat{\sigma}^2_e$	22.75	23.35	22.70
	$\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}}+\hat{\sigma}^2_e)$	0.646	0.656	0.647

Discussion

Prediction using epigenomic data

It was found that methylation data produced a reasonable predictive correlation when predicting plant height in *Arabidopsis*. The kernel matrix used here reflected epigenetic similarity between epiRILs based on their methylation profiles, and such epigenomic information might complement genomic information at the DNA level. The predictive correlation and mean squared error values were similar when only preselected probes were used. Hence, use of representative probes may help to reduce the cost of methylation profiling and computing time as well, at least for prediction purposes.

Nonparametric prediction using kernel methods is relatively simpler than with Bayesian regression models based on Markov chain Monte Carlo involving an enormous number of proposal distributions. However, in most cases the variance components associated with a kernel matrix do not provide a meaningful explanation of underlying biological processes, except for P-BLUP and G-BLUP, two special cases of RKHS regression. It was observed that when a Gaussian kernel was used for prediction, the estimated variance component associated with the kernel varied much with different choices of

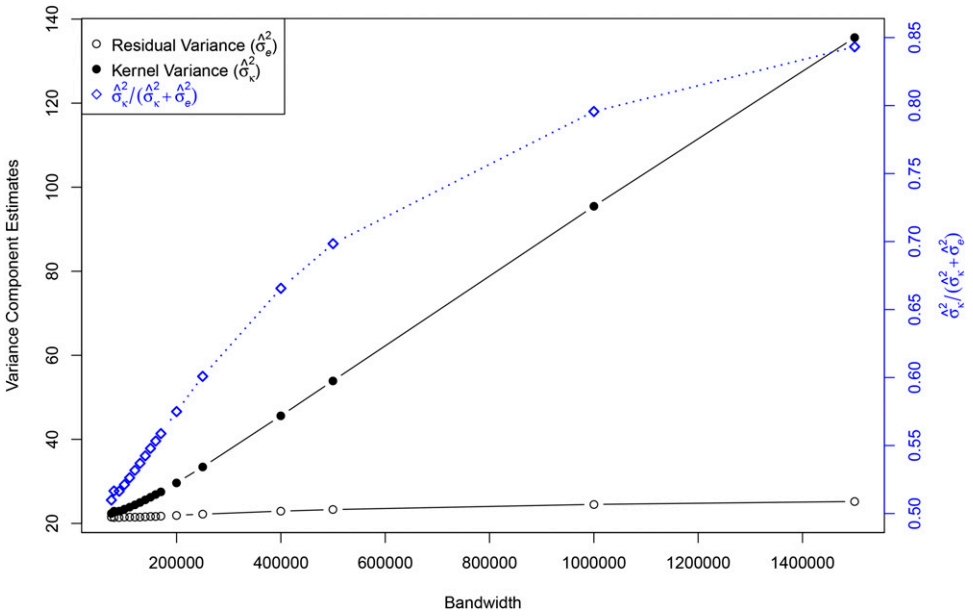

Figure 6 Estimated variance components (black lines, left *y*-axis) and variance component ratio (blue dotted line, right *y*-axis) with various bandwidth parameters of the Gaussian kernel (all probes used).

the bandwidth parameter h, probably due to the big impact of h on the values of the kernel matrix. This variation produced a wide range of $\hat{\sigma}^2_{\mathcal{K}}/(\hat{\sigma}^2_{\mathcal{K}}+\hat{\sigma}^2_e)$ ratios, making it difficult to assess phenotypic variance explained by epigenetic polymorphisms (Figure 6), but the best predictive performance was obtained when a Gaussian kernel was used. To cope with this difficulty, an epi-**G** that mimics the **G** matrix in G-BLUP was used as a kernel in RHKS regression. Since the epi-**G** kernel was generated from a discrete methyl status that was converted from continuous methyl-values data, a reduced predictive performance was observed probably due to loss of information during this data conversion process. However, the methyl-status data approximate the underlying epigenotypes of each locus. Hence the variance component associated with the epi-**G** is interpretable as in a G-BLUP model, which is clearly based on a biological concept.

In this study, we built prediction models with epigenetic information from MeDIP chip data. Alternatively, one should be able to build prediction models under the same statistical framework with BS-Seq data, which come from a combination of bisulfite conversion and next-generation sequencing (NGS) techniques with decreasing cost. Advantages of using BS-Seq data include the following: (1) unlike MeDIP chip data that rely on DNA segments, BS-Seq has single-base resolution inherited from NGS, making it more informative; and (2) instead of using the ratio between signal and background intensities to represent methylation level in a relative way, a ratio between the counts of methylated reads and total reads is used to measure the (absolute) methylation level, making it more accurate. However, when constructing the kernel matrix, any input information on methylation level, regardless of whether it is based on relative MeDIP data or on absolute BS-Seq data, can be turned into a relative measurement of epigenetic similarity, indicating that a better predictive performance may not be guaranteed when BS-Seq data are used. Moreover, the problem of information loss stemming from the discretization step when constructing the epi-**G** kernel will not be solved by the use of BS-Seq data. Therefore, even though NGS technologies are making inroads into the field of complex traits analysis, a potential next challenge is to develop a framework for BS-Seq data to take advantage of.

Despite the potential usefulness of epigenetic information in phenotype prediction suggested by our results, it should be noted that DNA methylation is reversible (*i.e.*, a methylated DNA molecule can be demethylated). Hence, methylation

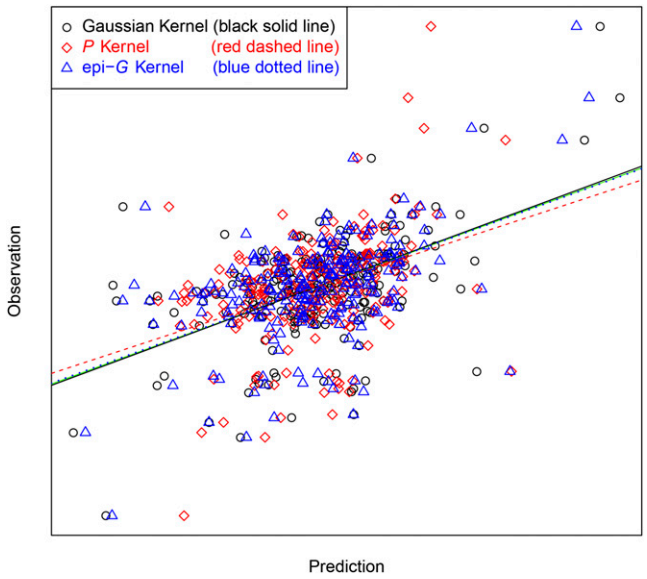

Figure 7 Graphical representation of prediction performance using different kernel matrices (all probes used). The green solid line is a 45° line passing through the origin. Regression lines using the Gaussian kernel and the epi-**G** kernel overlapped and were close to the 45° line, whereas the line from the **P** kernel had a smaller slope.

data are unstable relative to DNA polymorphisms. The reversibility of DNA methylation may produce "epimutation" events (Becker *et al.* 2011). Therefore, the entire methylome represents the dynamics of epimutations, and a particular methylation data set should be viewed as a "snapshot" of the methylome at a specific time from a specific tissue. To enhance phenotypic prediction performance further, information from multiple snapshots could be useful. Although methylation profiling is still expensive, its cost has decreased in recent years, and this trend is expected to continue.

Integrating genomic and epigenomic data in prediction

Our results suggested that epigenetic information can be used alone for whole-genome prediction of plant height, as a reasonable prediction performance was obtained. Therefore, it is of interest to combine epigenetic and DNA information for the same purposes. Recently, Vázquez *et al.* (2014) showed that the inclusion of multilayer -omics data in human epidemiology can increase the predictive correlation of disease risk drastically. Likewise, Shah *et al.* (2015) presented evidence suggesting that combining genetic information with significant associations between phenotype and epiprofile may be useful for predicting body mass index in humans. These findings suggested a potential use of epigenomic data in addition to genomic data for prediction using RKHS regression since integrating multiple information sources by introducing extra kernels tends to enhance predictive performance. For example, Tusell *et al.* (2014) reported that a multikernel model performed better than a single-kernel model, as anticipated by de los Campos *et al.* (2010). Also, fitting a pedigree-based relationship matrix (the **A** matrix) and a genome-based relationship matrix (the **G** matrix) together can give a higher predictive correlation than when only one matrix was fitted (Crossa *et al.* 2010). When viewing P-BLUP and G-BLUP as special cases of RKHS regression, a prediction model with both **A** and **G** is then a model with multiple kernels. The benefit from fitting multiple kernels simultaneously can be enhanced if all kernels are mutually orthogonal (Morota *et al.* 2014), which explains the result of Crossa *et al.* (2010), since **A** and **G** may provide information from different perspectives, with **G** supplementing information that is not captured by **A**. In a recent study, it was shown that genetic and epigenetic information can be uncoupled by epimutation over an evolutionary timescale (van der Graaf *et al.* 2015), so a higher predictive correlation could be expected when information from the epigenome is included in a prediction model, as suggested by Vázquez *et al.* (2014), since this extra information is distinct from the information conveyed by DNA polymorphisms.

Biologically, the preceding phenomenon can be interpreted as follows. The DNA sequence is transcribed into RNA and subsequently translated into protein, the building blocks of final phenotypes. Therefore, information at the protein layer (proteome) is "closest" to and genomic information is most "distant" from phenotypes in this biological pathway. Hence, proteomic information might provide better predictions of phenotypes than genomic information. Similarly, the epigenomic information, which lies between that conveyed by DNA and RNA layers, might be useful, if available. However, the availability of epigenomic data does not preclude the use of genomic information. On one hand, recent studies indicated that genomic variation and epigenomic variation may interact with each other (*e.g.*, Arnold *et al.* 2013; Wachter *et al.* 2014), such that epigenetics do not have a determinant effect on the ultimate phenotype, although it is a closer layer than DNA variation. On the other hand, epigenetic status can be predicted from genomic information (*e.g.*, Benveniste *et al.* 2014; Whitaker *et al.* 2015), suggesting that the inclusion of DNA polymorphisms may enhance predictive performance. Further, DNA information is crucial for artificial selection, and the epigenetic data would be informative in such a context only if transmission between generations is verified or if it enhances DNA-based predictions. For these reasons, integrating epigenomic with genomic data may be worthwhile for prediction purposes. Unfortunately, the epiRILs population used in this study did not have any SNP data, due to genetic identity between individual lines (C. Camilleri, personal communication). Nevertheless, one can expect that data integration will always be beneficial, along the lines that using information from multiple layers is expected to give stronger predictive correlations, as indicated by González-Recio (2012) and corroborated by Vázquez *et al.* (2014).

In short, including both DNA and epigenetic information into a prediction model may be fruitful. For example, if an epi-**G** kernel were to be used along with a **G** matrix (using SNP data), the estimated variance components should help in interpreting the proportion of phenotypic variance attributed to genetic and epigenetic variation. Also, perhaps the loss of information incurred when forming the epi-**G** kernel might be compensated by **G**. Therefore, using epigenomic and genomic information together has potential and additional study is warranted.

Conclusion

We built prediction models nonparametrically, using DNA methylation data. We chose RKHS regression for prediction because, unlike with prediction using SNP data, estimated regressions using methylation data do not have an obvious interpretation that links to model parameters via some theory or biological concept. In RKHS regression, a kernel matrix describing epigenetic similarities between different epiRILs makes model interpretation less difficult. Further, the tuning procedure is easier than for a parametric model, where a Bayesian treatment and MCMC techniques are usually needed.

We used different kernels in RKHS regression, namely the naive correlation matrix **P** and a Gaussian kernel **K** with different bandwidth parameters. When the bandwidth parameter was selected appropriately, the model with a Gaussian kernel performed better than that with a **P** kernel. Since a reasonably good predictive correlation was observed, this suggested that epigenetic information may be useful in whole-genome prediction as a source of information that does not reside in a DNA sequence. Furthermore, the value of the predictive correlation was retained when using preselected

representative probes, suggesting an avenue for cost reduction in prediction studies.

The performance of RKHS regression with a Gaussian kernel was strongly affected by its associated bandwidth parameter, not only in terms of the predictive correlation and predictive mean squared error, but also with respect to the variance component associated with the kernel matrix. This is because epigenetic similarities between individuals provided by the kernel matrix are based on a relative metric, instead of an absolute one. Therefore, the proportion of variance explained by the kernel does not give a meaningful interpretation of the proportion of phenotypic variance explained by epigenetic variation. On the other hand, a kernel matrix created from coded methylation status (epi-**G**) mimicked the genomic relationship matrix **G** and gave an estimated proportion of total variance explained by epigenetic variation of ∼65%. Although a small degradation in prediction performance is incurred when this epi-**G** kernel is applied, perhaps a better understanding of the importance of epigenetic variance can be obtained.

Using epigenetic information in addition to DNA polymorphisms in prediction has been studied by other authors in human epidemiology (*e.g.*, Vázquez *et al.* 2014), and results have suggested that this extra information may lead to a pronounced impact on prediction performance. Based on their results and on the empirical observation that RKHS regression with multiple kernels performs better than a single-kernel regression (Tusell *et al.* 2014), we conclude that inclusion of epigenetic information in prediction models may be warranted and possibly useful in livestock and crop production, as suggested by González-Recio (2012).

Acknowledgments

The authors thank Frank Johannes and René Wardenaar from the Groningen Bioinformatics Centre, University of Groningen, The Netherlands; Vincent Colot from Institut de Biologie de l'Ecole Normale Supérieure, France; and Christine Camilleri from Institut Jean-Pierre Bourgin, France for their kind responses to our questions on experimental design and data collecting procedures. Editor Fred van Eeuwijk and two anonymous reviewers are also thanked for their valuable comments on our manuscript. This work was supported by the Wisconsin Agriculture Experiment Station and by a U.S. Department of Agriculture Hatch grant (142-PRJ63CV) (to D.G.).

Y.H. and G.M. conducted the study; G.J.M.R. and D.G. advised the analysis; Y.H. analyzed data and wrote the paper; and G.M., G.J.M.R., and D.G. revised the manuscript. All authors read and approved the final manuscript.

Literature Cited

Arnold, P., A. Schöler, M. Pachkov, P. J. Balwierz, H. Jørgensen *et al.*, 2013 Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. Genome Res. 23(1): 60–73.

Aronszajn, N., 1950 Theory of reproducing kernels. Trans. Am. Math. Soc. 68: 337–404.

Becker, C., J. Hagmann, J. Müller, D. Koenig, O. Stegle *et al.*, 2011 Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature 480(7376): 245–249.

Bell, C. G., 2013 Epigenome-wide association studies: potential insights into human disease, pp. 287–317 in *Epigenetic and Complex Traits*, edited by A. Naumova and C. Greenwood. Springer-Verlag, New York.

Benveniste, D., H. J. Sonntag, G. Sanguinetti, and D. Srroul, 2014 Transcription factor binding predicts histone modifications in human cell lines. Proc. Natl. Acad. Sci. USA 111(37): 13367–13372.

Berger, S. L., 2002 Histone modifications in transcriptional regulation. Curr. Opin. Genet. Dev. 12(2): 142–148.

Bird, A. P., 1984 DNA methylation *vs.* gene expression. J. Embryol. Exp. Morphol. 83(Suppl.): 31–40.

Cassidy, S. B., E. Dykens, and C. A. Williams, 2000 Prader-Willi and Angelman syndromes: sister imprinted disorders. Am. J. Med. Genet. 97(2): 136–146.

Cheung, P., and P. Lau, 2005 Epigenetic regulation by histone methylation and histone variants. Mol. Endocrinol. 19(3): 563–573.

Colomé-Tatché, M., S. Cortijo, R. Wardenaar, L. Morgado, B. Lahouze *et al.*, 2012 Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc. Natl. Acad. Sci. USA 109(40): 16240–16245.

Cortijo, S., R. Wardenaar, M. Colomé-Tatché, F. Johannes, and V. Colot, 2014a Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip, pp. 125–149 in *Plant Epigenetics and Epigenomics: Methods and Protocols* (Methods in Molecular Biology, Vol. 1112), edited by C. Spillane and P. C. McKeown. Humana Press, Clifton, NJ/Totowa, NY.

Cortijo, S., R. Wardenaar, M. Colomé-Tatché, A. Gilly, M. Etcheverry *et al.*, 2014b Mapping the epigenetic basis of complex traits. Science 343(6175): 1145–1148.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueno *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2): 713–724.

de los Campos, G., D. Gianola, and G. J. Rosa, 2009 Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J. Anim. Sci. 87(6): 1883–1887.

de los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel, and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92(4): 295–308.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327–345.

Doi, A., I. H. Park, B. Wen, P. Murakami, M. J. Aryee *et al.*, 2009 Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. 41(12): 1350–1353.

Du, P., X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe *et al.*, 2010 Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587.

Egger, G., G. Liang, A. Aparicio, and P. A. Jones, 2004 Epigenetics in human disease and prospects for epigenetic therapy. Nature 429(6990): 457–463.

Esteller, M., 2008 Epigenetics in cancer. N. Engl. J. Med. 358(11): 1148–1159.
Flanagan, J. M., 2015 Epigenome-wide association studies (EWAS): past, present, and future, pp. 51–63 in *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis* (Methods in Molecular Biology, Vol. 1238), edited by M. Verma. Humana Press, Clifton, NJ/Totowa, NY.
Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt *et al.*, 1992 A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. USA 89(5): 1827–1831.
Gardiner-Garden, M., and M. Frommer, 1987 CpG islands in vertebrate genomes. J. Mol. Biol. 196(2): 261–282.
Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194(3): 573–596.
Gianola, D., and G. de los Campos, 2008 Inferring genetic values for quantitative traits non-parametrically. Genet. Res. 90(6): 525–540.
Gianola, D., and J. B. C. H. M. van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289–2303.
Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.
Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics 183: 347–363.
González-Recio, O., 2012 Epigenetics: a new challenge in the post-genomic era of livestock. Front. Genet. 2: 106.
González-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. Rosa *et al.*, 2008 Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178: 2305–2313.
Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, ON, Canada.
Henderson, I. R., and S. E. Jacobsen, 2007 Epigenetic inheritance in plants. Nature 447(7143): 418–424.
Illingworth, R. S., and A. P. Bird, 2009 CpG islands – "A rough guide". FEBS Lett. 583(11): 1713–1720.
Jeltsch, A., 2002 Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. ChemBioChem 3(4): 274–293.
Ji, H., L. I. Ehrlich, J. Seita, P. Murakami, A. Doi *et al.*, 2010 Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature 467(7313): 338–342.
Jiang, Y. H., J. Bressler, and A. L. Beaudet, 2004 Epigenetics and human disease. Annu. Rev. Genomics Hum. Genet. 5: 479–510.
Johannes, F., E. Porcher, F. K. Teixeira, V. Saliba-Colombani, M. Simon *et al.*, 2009 Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet. 5(6): e1000530.
Jones, M. C., J. S. Marron, and S. J. Sheather, 1996 A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. 91: 401–407.
Jones, P. A., and S. B. Baylin, 2002 The fundamental role of epigenetic events in cancer. Nat. Rev. Genet. 3(6): 415–428.
Jones, P. A., and S. B. Baylin, 2007 The epigenomics of cancer. Cell 128(4): 683–692.
Kaikkonen, M. U., M. T. Lam, and C. K. Glass, 2011 Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc. Res. 90(3): 430–440.
Kaput, J., and T. W. Sneider, 1979 Methylation of somatic *vs.* germ cell DNAs analyzed by restriction endonuclease digestions. Nucleic Acids Res. 7(8): 2303–2322.
Kimeldorf, G. S., and G. Wahba, 1971 Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 33: 82–95.
Kouzarides, T., 2007 Chromatin modifications and their function. Cell 128(4): 693–705.
Krueger, F., B. Kreck, A. Franke, and S. R. Andrews, 2012 DNA methylome analysis using short bisulfite sequencing data. Nat. Methods 9(2): 145–151.
Lim, D. H., and E. Maher, 2010 DNA methylation: a form of epigenetic control of gene expression. Obstet. Gynecol. 12(1): 37–42.
Lister, R., R. C. O'Malley, J. Tonti-Filippini *et al.*, 2008 Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133(3): 523–536.
MacArthur, D., 2008 Why do genome-wide scans fail? Genetic Future. Available at: http://www.wired.com/2008/09/why-do-genome-wide-scans-fail/.
Meijers-Heijboer, E. J., L. A. Sandkuijl, H. G. Brunner, H. J. Smeets, A. J. Hoogeboom *et al.*, 1992 Linkage analysis with chromosome 15q11–13 markers shows genomic imprinting in familial Angelman syndrome. J. Med. Genet. 29(12): 853–857.
Meissner, A., A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander *et al.*, 2005 Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 33(18): 5868–5877.
Meissner, A., T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna *et al.*, 2008 Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454(7205): 766–770.
Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
Morota, G., and D. Gianola, 2014 Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. 5: 363.
Morota, G., M. Koyama, G. J. Rosa, K. A. Weigel, and D. Gianola, 2013 Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. Genet. Sel. Evol. 45: 17.
Morota, G., P. Boddhireddy, N. Vukasinovic, D. Gianola, and S. Denise, 2014 Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. Front. Genet. 5: 56.
Mrode, R., 2014 *Linear Models for the Prediction of Animal Breeding Values*, Ed. 3. CAB International. Wallingford, UK.
Nadaraya, E. A., 1964 On estimating regression. Theory Probab. Appl. 9: 141–142.
Nicholls, R. D., S. Saitoh, and B. Horsthemke, 1998 Imprinting in Prader-Willi and Angelman syndromes. Trends Genet. 14(5): 194–200.
Pembrey, M., 2012 *An Introduction to the Genetics and Epigenetics of Human Disease*. Progress Educational Trust, London.
Portela, A., and M. Esteller, 2010 Epigenetic modifications and human disease. Nat. Biotechnol. 28(10): 1057–1068.
Rakyan, V. K., T. A. Down, D. J. Balding, and S. Beck, 2011 Epigenome-wide association studies for common human diseases. Nat. Rev. Genet. 12(8): 529–541.
Ratel, D., J. L. Ravanat, F. Berger, and D. Wion, 2006 N6-methyladenine: the other methylated base of DNA. BioEssays 28(3): 309–315.
Razin, A., and H. Cedar, 1991 DNA methylation and gene expression. Microbiol. Rev. 55(3): 451–458.
Reinders, J., B. B. Wulff, M. Mirouze, A. Mari-Ordonez, M. Dapp *et al.*, 2009 Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. Genes Dev. 23(8): 939–950.
Riggs, A. D., and T. N. Porter, 1996 Overview of epigenetic mechanisms, pp. 29–45 in *Epigenetic Mechanisms of Gene Regulation*, edited by V. E. A. Russo, R. A. Martienssen, and A. D. Riggs. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Riggs, A. D., R. A. Martienssen, and V. E. A. Russo, 1996 Introduction, pp. 1–4 in *Epigenetic Mechanisms of Gene Regulation*, edited by V. E. A. Russo, R. A. Martienssen, and A. D. Riggs. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Rivera, R. M., and L. B. Bennett, 2010 Epigenetics in humans: an overview. Curr. Opin. Endocrinol. Diabetes Obes. 17(6): 493–499.
Robertson, K. D., 2002 DNA methylation and chromatin – unraveling the tangled web. Oncogene 21(35): 5361–5379.
Ruthenburg, A. J., H. Li, D. J. Patel, and C. D. Allis, 2007 Multivalent engagement of chromatin modifications by linked binding modules. Nat. Rev. Mol. Cell Biol. 8(12): 983–994.
Shah, S., M. J. Bonder, R. E. Marioni, Z. Zhu, A. F. McRae *et al.*, 2015 Improving phenotypic prediction by combining genetic and epigenetic associations. Am. J. Hum. Genet. 97(3): 1–11.
Silverman, B., 1986 *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.
Tollefsbol, T. (Editor), 2012 *Epigenetics in Human Disease*. Academic Press, Waltham, MA.
Tusell, L., P. Perez-Rodriguez, S. Forni, and D. Gianola, 2014 Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. J. Anim. Breed. Genet. 131(2): 105–115.
van der Graaf, A., R. Wardenaar, D. A. Neumann, A. Taudt, R. G. Shaw *et al.*, 2015 Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc. Natl. Acad. Sci. USA 112 (21): 6676–6681.
VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11): 4414–4423.
Vanyushin, B. F., 2006 DNA methylation in plants. Curr. Top. Microbiol. Immunol. 301: 67–122.
Vázquez, A. I., H. W. Wiener, S. Shrestha, H. Tiwari, and G. de los Campos, 2014 Integration of multi-layer omic data for prediction of disease risk in humans, in *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*. Vancouver, BC, Canada. Available at: https://asas.org/docs/default-source/wcgalp-proceedings-oral/213_paper_10325_manuscript_1311_0.pdf?sfvrsn=2.
Waalwijk, C., and R. Flavell, 1978 DNA methylation at a CCGG sequence in the large intron of the rabbit β-globin gene: tissue-specific variations. Nucleic Acids Res. 5(12): 4631–4642.
Wachter, E., T. Quante, C. Merusi, A. Arczewska, F. Stewart *et al.*, 2014 Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. eLife 3: e03397.
Wahba, G., 1990 *Spline Models for Observational Data* (CBMS-NSF Regional Conference Series in Applied Mathematics). Society for Industrial and Applied Mathematics, Philadelphia.
Wahba, G., 1999 Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, pp. 68–88 in *Advances in Kernel Methods: Support Vector Learning*, edited by B. Schölkopf, C. J. C. Burges, and A. J. Smola. MIT Press, Cambridge, MA.
Wahba, G., 2002 Soft and hard classification by reproducing kernel Hilbert space methods. Proc. Natl. Acad. Sci. USA 99: 16524–16530.
Watson, G. S., 1964 Smooth regression analysis. Sankhyā: Ind. J. Stat. Ser. A 26: 359–372.
Weber, M., J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase *et al.*, 2005 Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. 37(8): 853–862.
Whitaker, J. W., Z. Chen, and W. Wang, 2015 Predicting the human epigenome from DNA motifs. Nat. Methods 12(3): 265–272.
Zhou, H., H. Hu, and M. Lai, 2010 Non-coding RNAs and their epigenetic regulatory mechanisms. Biol. Cell 102(12): 645–655.

Communicating editor: F. van Eeuwijk