Heredity (2020) 124:658–674
https://doi.org/10.1038/s41437-020-0301-4

the geneticssociety

ARTICLE

Integration of single nucleotide variants and whole-genome DNA methylation profiles for classification of rheumatoid arthritis cases from controls

Mahmoud Amiri Roudbar[1] · Mohammad Reza Mohammadabadi[2] · Ahmad Ayatollahi Mehrgardi[2] · Rostam Abdollahi-Arpanahi[3] · Mehdi Momen[4] · Gota Morota[5] · Fernando Brito Lopes[6] · Daniel Gianola[7,8] · Guilherme J. M. Rosa[7,8]

Received: 27 September 2019 / Revised: 17 February 2020 / Accepted: 17 February 2020 / Published online: 3 March 2020

Abstract

This study evaluated the use of multiomics data for classification accuracy of rheumatoid arthritis (RA). Three approaches were used and compared in terms of prediction accuracy: (1) whole-genome prediction (WGP) using SNP marker information only, (2) whole-methylome prediction (WMP) using methylation profiles only, and (3) whole-genome/methylome prediction (WGMP) with combining both omics layers. The number of SNP and of methylation sites varied in each scenario, with either 1, 10, or 50% of these preselected based on four approaches: randomly, evenly spaced, lowest *p* value (genome-wide association or epigenome-wide association study), and estimated effect size using a Bayesian ridge regression (BRR) model. To remove effects of high levels of pairwise linkage disequilibrium (LD), SNPs were also preselected with an LD-pruning method. Five Bayesian regression models were studied for classification, including BRR, Bayes-A, Bayes-B, Bayes-C, and the Bayesian LASSO. Adjusting methylation profiles for cellular heterogeneity within whole blood samples had a detrimental effect on the classification ability of the models. Overall, WGMP using Bayes-B model has the best performance. In particular, selecting SNPs based on LD-pruning with 1% of the methylation sites selected based on BRR included in the model, and fitting the most significant SNP as a fixed effect was the best method for predicting disease risk with a classification accuracy of 0.975. Our results showed that multiomics data can be used to effectively predict the risk of RA and identify cases in early stages to prevent or alter disease progression via appropriate interventions.

Introduction

Rheumatoid arthritis (RA) is an autoimmune disease producing chronic inflammation of the joints and other areas of the body, such as blood vessels and lungs, and its prevalence ranges from 0.5 to 1% across populations (Glant et al. 2014;

Associate editor: Yuan-Ming Zhang

Supplementary information The online version of this article (https://doi.org/10.1038/s41437-020-0301-4) contains supplementary material, which is available to authorized users.

✉ Mahmoud Amiri Roudbar
mahmood.amiri225@gmail.com

[1] Department of Animal Science, Safiabad-Dezful Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education & Extension Organization (AREEO), Dezful, Iran

[2] Department of Animal Science, College of Agriculture, Shahid Bahonar University of Kerman, 76169-133 Kerman, Iran

[3] Department of Animal and Poultry Science, College of Aburaihan, University of Tehran, 465, Pakdasht, Tehran, Iran

[4] Department of Surgical Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA

[5] Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

[6] Department of Animal Sciences, Sao Paulo State University, Julio de Mesquita Filho (UNESP), Prof. Paulo Donato Castelane, Jaboticabal, SP 14884-900, Brazil

[7] Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

[8] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA

Sattar and McInnes 2005; Silman and Pearson 2002; Tanoue 1998). Leukocytes are one of the main immune cells involved in development of autoimmune reactions, probably transmitting complex signals between different tissue compartments involved in disease development (Mayadas et al. 2009).

Recent genome-wide association (GWAS) and epigenome-wide association (EWAS) studies of RA have tagged more than one hundred genetic risk loci and ten putative differentially methylated positions (DMPs) (Liu et al. 2013; Okada et al. 2014; Padyukov et al. 2011; Raychaudhuri et al. 2012; Stahl et al. 2010). In RA, genetic and epigenetic modifications can influence disease development and disease risk variation jointly (Glant et al. 2014). Other studies have shown that integration of multiple omics data can improve predictions for complex traits and diseases (e.g., Yuan et al. (2014), Vazquez et al. (2016), and Wheeler et al. (2014)).

Use of candidate loci discovered through GWAS typically does not improve prediction of complex diseases enough to be useful in personalized medicine (Li and Meyre 2014). In addition, EWAS have not produced robust associations between methylation sites and common diseases (Liu et al. 2013). On the other hand, fitting whole-genome molecular markers simultaneously into a regression model (Meuwissen et al. 2001) has been used successfully in animal and plant breeding (de los Campos et al. 2013a; Gianola and Rosa 2015) and also in human disease risk prediction (de los Campos et al. 2013b; Moser et al. 2015; Speed and Balding 2014). It is known that a much larger proportion of genetic variance can be accounted for in prediction models that use all single nucleotide polymorphisms (SNPs) simultaneously than when fitting only significant SNPs according to GWAS (de los Campos et al. 2010; Yang et al. 2010). Many studies have focused on genetic aspects of RA (Hao et al. 2014; Kapitany et al. 2005; Padyukov et al. 2011; Raychaudhuri et al. 2012; Stahl et al. 2010), but other reports have indicated the importance of environmental factors and epigenetic regulation on pathogenesis (Di Giuseppe et al. 2014; Glant et al. 2014; Goronzy et al. 2010). Most previous studies have mainly focused on pathogenesis (Choy 2012), epidemiology (Silman and Pearson 2002), and GWAS or EWAS (Liu et al. 2013; Padyukov et al. 2011; Raychaudhuri et al. 2012).

With the increasing availability of data from multiple omics layers using next-generation DNA sequencing and other high-throughput technologies, prediction of complex traits and diseases can be improved drastically. Improvements in prediction accuracy of complex traits by combining different sources of omics data (e.g., methylation patterns and gene expression patterns) in statistical models have been reported (Hu et al. 2019; Vazquez et al. 2016; Wheeler et al. 2014; Yuan et al. 2014). In a study on survival from breast cancer, a gain in prediction accuracy was achieved when layers including gene expression and DNA methylation data were added to SNP-based models (Vazquez et al. 2016). More recently, Hu et al. (2019) reported a significant improvement in prediction accuracy in rice traits by integrating transcriptome and metabolome combined with genomic data. It appears that inclusion of epigenetic data in prediction models is beneficial for improving prediction accuracy. Although the causes of RA remain unknown, genetic and epigenetic bases for the disease have been suggested (Choy 2012; Liu et al. 2013). Use of additional sources of omics data (e.g., epigenetic variants) may be helpful to increase accuracy of classification of RA inserting into cases or controls.

This study was carried out to evaluate the use of multiomics data for classifying RA cases from controls. Methylation and genotyping data collected from 689 samples (354 cases and 335 controls) were used. Classification accuracies were estimated for genomic and methylome layers separately, and then jointly. Effects of different approaches of preselection of predictor variables were assessed. We provide insights on how much multiomics data integration can improve classification accuracy of RA. Finally, we carried out a pathway analysis for the SNPs and methylation sites with the strongest effects, to assess their connection to RA.

Subjects and methods

Subjects

Data sets were obtained from the Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study involving cases of RA in Sweden. In our study, after excluding two control samples due to the lack of information for smoking status, 354 anti-CCP positive RA cases and 333 controls were used. Details of data collection are in previous studies (Liu et al. 2013; Padyukov et al. 2011). All case and control subjects were selected from the same study by matching for age, gender, smoking, and residential area at the time of diagnosis (for more details see Supplementary Table 1 in Liu et al. (2013)).

Genotyping

The EIRA sample was genotyped with the Illumina platform using Hap370CNVduo as described previously (Padyukov et al. 2011). Genotypic data were obtained with permission from EIRA investigators. A total of 301,282 autosomal SNPs were used before quality control (QC). The QC of markers was performed on all genotyped samples using the PLINK software (Purcell et al. 2007), and monomorphic SNPs, a p value from Hardy–Weinberg

equilibrium $\leq 5.0 \times 10^{-5}$, minor allele frequency ≤ 0.05, or call-rates ≥ 95% were removed. After applying such QC filters, 292,836 SNPs (p_g) remained for subsequent analysis. To quantify and control for population stratification, we used a principal components approach using the R package SNPRelate (Zheng et al. 2012) from the Bioconductor open source software (http://www.bioconductor.org/) but there was no evidence of population stratification (Supplementary Information, Fig. S1).

Methylation data

DNA extraction, preparation procedures, and methylation measurement were described in Liu et al. (2013). Briefly, after bisulfate converting of DNA, the Illumina Infinium HD Methylation Assay (Illumina) was used for measuring methylation levels in more than 485,000 sites per sample using the Infinium HumanMethylation450 BeadChips (Bibikova et al. 2011). The methylation data are available in Gene Expression Omnibus (GEO) with accession number "GSE42861". For normalization of methylation data, a quantile normalization algorithm (Fortin et al. 2014) was used to remove unwanted technical variation using Illumina's control probes. A total of 17,541 probes containing SNPs in their sequences were also removed from the final data. All methylation values with a detection p value ≥ 0.01 were set as missing. Samples and probes were checked for missing values, and 1456 probes with >5% missing values were removed from the data. All samples had <5% of missing values and were then kept for further analyses. The missing Beta-values were imputed using the R package impute with ten nearest neighbor averaging (Troyanskaya et al. 2001). After data preprocessing, 466,515 (p_m) methylation sites were available for the analyses. Methylated and unmethylated signals (M and U, respectively) were converted to Beta-values with a scale between 0 and 1 using the M/(M + U + 100) formula. All methylation array data preprocessing was conducted with the R package minfi (Aryee et al. 2014).

Prediction methods

Three modeling strategies were considered for classifying RA subjects using genomic or/and methylome data: (i) Only genomic information used for whole-genome prediction (WGP); (ii) only methylome data used for whole-methylome prediction (WMP); and (iii) genomic and methylome data jointly used for whole-genome/methylome prediction (WGMP). Several Bayesian prediction models were applied: Bayesian ridge regression (BRR) with a Gaussian prior density (de los Campos et al. 2013a); Bayes-C, a spike-slab model with a Gaussian prior density and a null-state for variable selection (Habier et al. 2011); Bayes-A with scaled-t prior density; Bayes-B, a spike-slab model with a scaled-t prior and a null-state for variable selection (Meuwissen et al. 2001); and the Bayesian LASSO (BL) which was a double-exponential (Park and Casella 2008).

Whole-genome prediction

The matrix of genomic predictor variables was $G = \{g_{ij}\}$ with $i = 1,\ldots, n$, $j = 1,\ldots, p_g$. Each element of the response vector $y = \{y_i\}$ had two possible values, i.e., presence $y_i = 1$ or absence $y_i = 0$ of RA for the ith individual. We used a probit link function $P(y_i = 1|G_i) = \Phi(\eta i)$, where Φ is a standard normal cumulative distribution function and ηi is a linear predictor given by

$$\eta_i = \mu + \sum_{1}^{p_g} g_{ij}\alpha_j.$$

Above, μ is an intercept, g_{ij} is the genotype of the ith individual at the jth marker, and α_j is the jth marker effect. The probit link assumed a latent normally distributed variable $l_i = \eta_i + \varepsilon_i$ liability (Gianola and Foulley 1983); and a measurement model $y_i = 0$ if $l_i < \gamma$, and 1 otherwise, where γ is a threshold parameter; and ε_i is an independent normal model residual with mean zero and with variance set equal to one. The density of the posterior distribution was

$$p(\theta_g|y, \omega_g) \propto p(y|\theta_g)\, p(\theta_g|\omega_g),$$

where $p(\theta_g|y, \omega_g)$ is the conditional posterior density of parameters $\theta_g = \{\mu, \sigma_e^2, \alpha\}$, including the residual variance (σ_e^2), which was assigned a scaled-inverse χ^2 prior distribution; μ was assigned a flat prior distribution, and the marker effects (α) were assigned independent and identically distributed informative priors, depending on the model; ω_g represents the genomic hyperparameters indexing the prior density of marker effects. ω_g for BRR is the variance of SNP effects (σ_α^2), for BL is the regularization parameter (λ^2) and σ_α^2, for Bayes-A is the degrees of freedom d.f.$_\alpha$ and scale parameter S_α, for Bayes-B is d.f.$_\alpha$, S_α and a mixture proportion (π), and for Bayes-C is π and σ_α^2, where π is the probability of a null effect of markers. The expression $p(y|\theta_g) = \prod_1^n \left\{ [\Phi(\eta_i)]^{y_i} [1 - \Phi(\eta_i)]^{1-y_i} \right\}$ is the conditional distribution of the phenotypes given the linear predictor, and $p(\theta_g|\omega_g) \propto p(\alpha_j|\omega_g)\, p(\sigma_e^2)$ is the joint prior distribution of model unknowns, given the hyperparameters. The prior density of marker effects, $p(\alpha_j|\omega_g)$, defines the specification of the various Bayesian methods inducing shrinkage and variable selection (Bayes-B and Bayes-C) or shrinkage only (Bayes-A, BRR, and BL with scaled-t, Gaussian, and Laplace priors, respectively). For more details, see de los Campos et al. (2013a).

Whole-methylome prediction

To regress disease status on methylation covariates, $M = \{m_{il}\}$ with $l = 1,\ldots, p_m$, we adopted the following linear predictor for WMP:

$$\tau_i = \mu + \sum_{1}^{p_m} m_{il}\beta_l,$$

where m_{il} is the methylotype of the ith individual at the lth methylome probe, and β_l is the lth probe effect. The Bayesian model used for WMP, had a similar structure to the WGP model, as follows:

$$p(\theta_m|y,\omega_m) \propto p(y|\theta_m)\,p(\theta_m|\omega_m)$$
$$\propto \prod_{1}^{n}\left\{[\Phi(\tau_i)]^{y_i}[1-\Phi(\tau_i)]^{1-y_i}\right\} p(\beta_l|\omega_m)\,p\left(\sigma_e^2\right),$$

where θ_m is the vector of unknown methylomic parameters; and ω_m contains the methylomic hyperparameters indexing the prior density of methylation effects. ω_m for BRR is the variance of methylation effects (σ_β^2); for BL it implies the regularization parameter (λ^2) and σ_β^2; for Bayes-A is the degrees of freedom (d.f.$_\beta$) and scale parameter S_β; for Bayes-B is d.f.$_\beta$, S$_\beta$, and the mixture proportion (π), and for Bayes-C it is π and σ_β^2.

Integrated genome/methylome prediction

For WGMP, G and M data were the two input layers as described earlier. The linear predictor was given by

$$\phi_i = \mu + \sum_{1}^{p_g} g_{ij}\alpha_j + \sum_{1}^{p_m} m_{il}\beta_l,$$

where notations were as in WGP and WMP. The posterior distribution of the model unknowns was

$$p(\theta_g,\theta_m|y,\omega_g,\omega_m) \propto p(y|\theta_g,\theta_m)p(\theta_g,\theta_m|\omega_g,\omega_m)$$
$$\propto \prod_{1}^{n}\left\{[\Phi(\phi_i)]^{y_i}[1-\Phi(\phi_i)]^{1-yi}\right\}p(\alpha,\beta|\omega_g,\omega_m)p\left(\sigma_e^2\right),$$

where $\alpha = \{\alpha_j\}$ and $\beta = \{\beta_j\}$ are the vectors of marker and methylation effects, respectively; $p(\theta_g, \theta_m|y, \omega_g, \omega_m)$ is the posterior density of unknown genomic and methylomic parameters (θ_g, θ_m); $p(y|\theta_g, \theta_m)$ is the conditional distribution of the phenotype given the linear predictor; and $p(\theta_g, \theta_m|\omega_g, \omega_m) = p(\theta_g|\omega_g)p(\theta_m|\omega_m)$ is the joint prior distribution of model unknowns given the sets of layer-specific regularization hyperparameters for genomic (i.e., σ_β^2, λ^2, d.f.$_\alpha$, S_α, and π) and methylomic (i.e., σ_β^2, λ^2, d.f.$_\beta$, S_β, and π) data. Note that whole-genome and whole-methylome effects were assigned independent prior distribution.

All Bayesian analyses were implemented using a Markov chain Monte Carlo (MCMC) approach, Gibbs sampling. In each cross-validation (CV) and for each model, the number of iterations of the Gibbs sampler was 200,000, with the first 100,000 samples discarded as burn in. A thinning interval of 20 was used. Thus, 5000 posterior samples were used for inferring features of the posterior distribution. We diagnosed convergence using a criterion of accuracy of estimation of a quantile using the R package coda (Plummer et al. 2006). Plots of posterior densities for variance parameters and latent effects from the best model in WGP, WMP, and WGMP are shown in Supplementary file, Fig. S2.

Treating the most significant rheumatoid arthritis risk locus as a fixed effect

In WGP, the main assumption is that quantitative traits are controlled by an infinite number of loci and each locus has an infinitely small effect (Fisher 1918). This assumption is violated if a genetic variant has a large effect. In Bayesian approaches, effects of all SNPs can be estimated jointly either without performing marker selection or by using both variable selection and shrinkage of estimates (Gianola et al. 2003; Meuwissen et al. 2001). The various methods use different prior densities to deal with different real distributions of marker effects to model different genetic scenarios. However, some prior densities may not be able to perform appropriate when phenotype is influenced by a genetic variant with a large effect.

To check for signals of major variants, we performed a logistic regression GWAS, SNP by SNP. Plink software was used to obtain p values of association tests (Purcell et al. 2007). The Manhattan plot of p values for the genomic data is shown in Fig. 1. There is a major locus on chromosome 6 with a significant effect on the disease, which is in agreement with previous studies (Plenge et al. 2007; Stahl et al. 2010; Walsh et al. 2016). Hence, the most significantly associated SNP was considered as a fixed effect (flat prior) and all other SNPs were random effects in each CV for the WGP. In all GWAS conducted within the training sets of CVs, SNP rs660895 presented the lowest p value, except in two cases in which SNP rs2395175 was the most significant, where it was the second significant marker. Nonetheless, there is a strong linkage disequilibrium (LD) between rs660895 and rs2395175, with $r^2 > 0.7$. To exclude SNPs that were in LD with the top SNP, we calculated r^2 between rs660895 and all SNPs located 1000 kb up- and downstream, and removed 14 SNPs with $r^2 > 0.2$.

Subset selection

The number of predictor variables available (466,515 methylation sites and 292,836 SNPs) was large, making

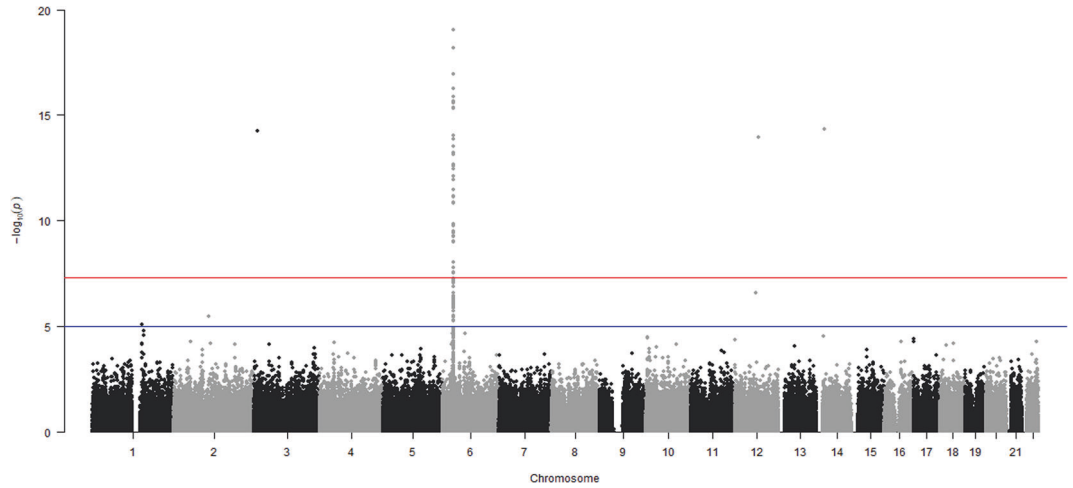

Fig. 1 Manhattan plot of *p* values for the genome-wide association study of rheumatoid arthritis. The red and blue lines represent genome-wide significance (5×10^{-8}) and suggestive (1×10^{-5}) thresholds, respectively.

the analyses computationally taxing. As an alternative, a subset selection approach was applied, where the number of predictors was reduced to 50% (146,418 SNPs and 233,258 methylation sites), 10% (with 29,284 SNPs and 46,652 methylation sites), and 1% (2928 SNPs and 4665 methylation sites) of the total number of predictors. Four different methods were used for subset selection on both SNP and methylation sites: random, evenly spaced, lowest *p* value (GWAS or EWAS), and the strongest estimated effects from a BRR model. Each subset selection method was performed in each CV analysis, so that both variable selection and parameter estimation were performed without using information from the test set samples.

LD-based SNP pruning

Since high levels of pairwise LD in SNP data may impair performance of genomic prediction models (Calus et al. 2016), it could be useful to generate a pruned subset of mutually uncorrelated SNPs. To evaluate effect of high LD between SNPs on our models, we produced pruned subsets of SNPs that were in approximate LD with each other via the PLINK software. SNPs were pruned based on variance inflation factor (VIF) thresholds of 2, 1.25, 1.11, and 1.01, a sliding window of 50 SNPs, and shifted forward in steps of 10% of the window size, i.e., with 5. VIFs of 2, 1.25, 1.11, and 1.01 imply multiple correlation coefficients of 0.5, 0.2, 0.1, and 0.01, respectively, for an SNP regressed on all other SNPs in each window simultaneously. These parameters allowed a pairwise comparison to remove SNPs from low (VIF = 1.01) to high LD (VIF = 2). After LD-pruning, 97,201, 57,932, 41,673, and 14,242 SNPs retained for VIF of 2, 1.25, 1.11, and 1.01, respectively, for further analysis.

Correcting methylation signatures for potential confounders

As DNA samples for methylation analysis were generally derived from a large number of individuals with distinct cellular heterogeneity, age, gender, and smoking status, we attempted to adjust for these confounders in two scenarios, i.e., correcting for cellular heterogeneity, and correcting for all available potential confounders.

Correcting methylation signatures for cellular heterogeneity

Whole blood samples are a heterogeneous mixture of cell types. A recent study showed that differential DNA methylation signatures in whole blood samples can be affected by the proportion of white blood cell types in each sample (Reinius et al. 2012). Since variation in white blood cell frequencies may affect accuracy of prediction, adjusted or nonadjusted methylation measurements for cell proportion were compared. The proportion of the major cell type in blood for each sample was estimated using an algorithm in which an external validation set consisting of signatures from purified cell samples was applied (Houseman et al. 2012). A total of six different cell types including two types of T cells (CD8T and CD4T), NK cells, B cells, monocytes, and granulocytes were used for adjustment. Cell proportions for each sample were estimated with the R package minfi. Sample-specific estimates of differential cell counts were used to adjust the methylation signatures using a linear regression model (adjusted-cell).

Correcting methylation signatures for all available potential confounders

It is shown that methylation level may also be modified by environmental factors such as cigarette smoking status or by gender and age (Breitling et al. 2011; Hannum et al. 2013). To control for such factors, methylation changes due to age, gender, smoking status, and cell type proportion were accounted by fitting a linear model for each DMP. The residuals from this model were used as new adjusted methylation signatures (adjusted-all).

Cross-validation and prediction accuracy assessment

A tenfold CV was used to assess prediction accuracy. The dataset was randomly split into ten mutually exclusive subsets. One subsample was considered as the validation data for testing the models and the other nine subsamples were considered as training data. This was repeated for all subsets until each of the ten subsamples was used exactly once as the validation data. The ten results were averaged to produce a single estimation. This whole process was repeated 20 times, which resulted in a total of 20 CV predictions. A receiver operating characteristics (ROC) was used for evaluating predicting accuracy of the models (Fawcett 2006). The CV-area under the ROC curve (CV-AUC) was used to evaluate predictive ability of each model. Therefore, 20 estimates of CV-AUC were calculated for each model. Standard deviations of the CV-AUC for each model across the 20 CVs were calculated. The average of these 20 estimates was used to compare performance of the various models.

Functional annotation for top ten SNPs and methylation sites with the strongest effect

SNPs and methylation effects were estimated from the WGP, WMP, and WGMP in the context of CV. For each SNP and methylation effect, the average across the 20 CV results was obtained. Then, the top ten SNPs and methylation sites with the largest absolute effects were selected for pathway analysis, to find connections to the disease. First, we used the R package rsnps to retrieve SNPs information by sending queries to public databases. For methylation sites, the R package minfi was used to access annotation for each position. Second, the nearest neighboring genes to SNPs and methylation sites were found employing BioMart web services through the R package biomaRt (Durinck et al. 2009). Gene lists retrieved from each prediction method were uploaded to DAVID (Huang da et al. 2009), a web-accessible program, to link them to associated diseases.

Table 1 Genomic prediction accuracy (standard errors) from the average of cross-validation AUC using SNP markers in different Bayesian methods: Bayes-A, Bayes-B, Bayes-C, Bayesian LASSO (BL), and Bayesian ridge regression (BRR).

Model	Testing sets	
	Random	Fixed + Random
Bayes-A	0.718 (0.062)[a]	0.719 (0.061)[a]
Bayes-B	0.731 (0.054)[b]	0.737 (0.057)[b]
Bayes-C	0.592 (0.064)[c]	0.704 (0.056)[d]
BL	0.592 (0.064)[c]	0.705 (0.056)[d]
BRR	0.592 (0.063)[c]	0.704 (0.056)[d]

Average AUCs in testing population with different superscript(s) are significantly different from each other (p value < 0.05).

Results

Prediction accuracy using whole-genome models

The prediction accuracy obtained by fitting either all SNPs as random (Random) or rs660895 as a fixed effect plus the other SNPs as random (Fixed + Random), using different Bayesian methods is shown in Table 1. Here we presented the average of CV-AUCs exhibited in testing sets. The model with a scaled-t prior density and variable selection (i.e., Bayes-B), had the highest accuracy of prediction in the testing set (≈0.73). Other models with priors other than Student-t density (Bayes-C, BL, and BRR) had the lowest accuracy when the major locus, rs660895, was not fitted as a fixed effect (Random model). Overall, fitting rs660895 as a fixed effect (Fixed + Random model) increased prediction accuracy, especially in Bayes-C, BL, and BRR. In contrast, Random or Fixed + Random model did not differ for either Bayes-A and Bayes-B. Therefore, the Fixed + Random fitting was used for subsequent analyses.

Genome prediction accuracy using subsets of SNP markers

Estimates of prediction accuracy from subsets of SNPs based on random, evenly spaced, GWAS p value and strongest BRR effects across the five Bayesian models are shown in Fig. 2a–d, respectively. Among these preselection SNP methods, 50% random SNP selection had the highest CV-AUC with an average of 0.742. When SNPs selection levels were reduced to 10 and 1%, the evenly spaced selection method using Bayes-B (0.716 ± 0.054) and Bayes-A (0.717 ± 0.052) models exhibited the highest accuracies. For SNPs selected based on GWAS p value (Fig. 2c), when the proportion selection decreased, accuracy decreased as well. For instance, 50% SNP selection obtained a higher accuracy than 10 and 1% SNP selection across models.

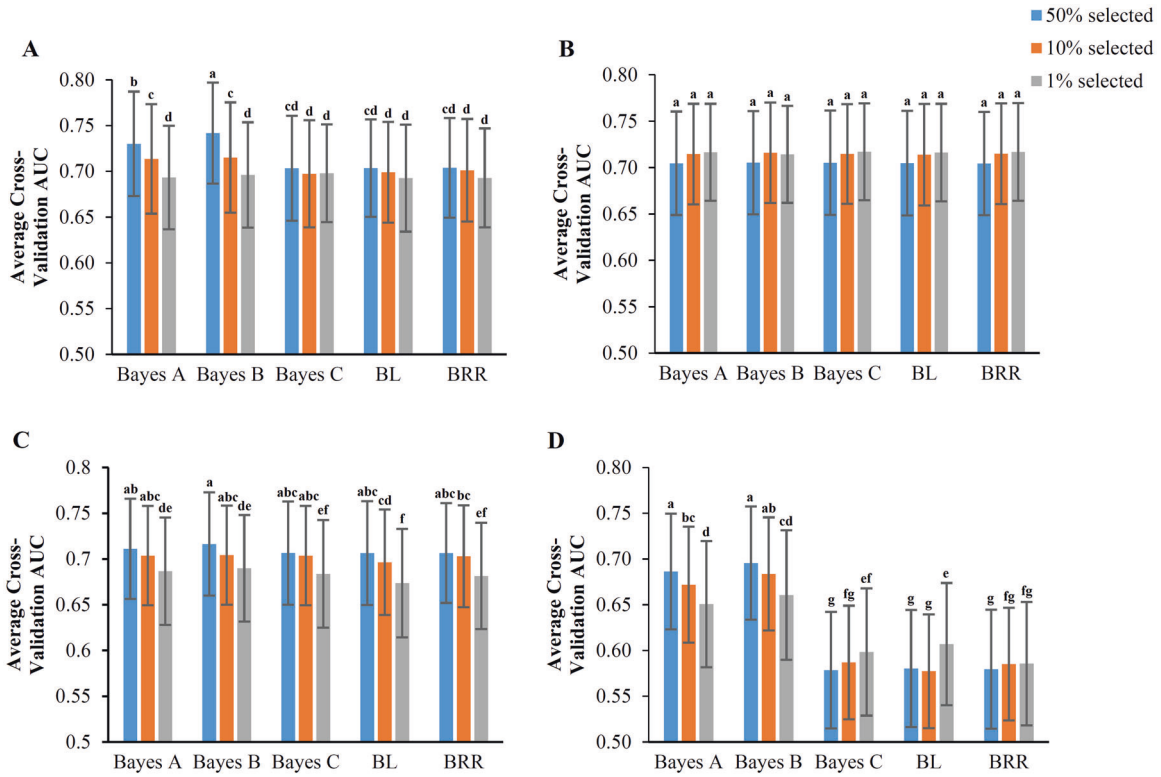

Fig. 2 Genomic prediction accuracy of three levels of selected SNPs (1, 10, and 50% of all markers) using Bayes-A, Bayes-B, Bayes-C, Bayes LASSO (BL), and Bayesian ridge regression (BRR) models and different approaches to SNP selection. Randomly (**a**), evenly spaced (**b**), GWAS *p* value (**c**), and strongest BRR effects (**d**). AUCs with different superscript(s) are significantly different from each other (p value < 0.05).

A clear difference in prediction accuracy was observed between Bayesian models when a subset of SNPs selected based on BRR estimates was used for predictors. Models that used a scaled-t prior density (Bayes-A and Bayes-B) produced significantly higher prediction accuracies than those that assumed a normal (BRR and Bayes-C) or Laplace (BL) prior density. The higher prediction accuracies observed for Bayes-A and Bayes-B may be due to ability of these methods to estimate SNP effects better for markers linked to large-effect QTL.

Figure 3 shows the estimates of prediction accuracy for subsets of SNPs selected based on LD-pruning. In this case, Bayes-B delivered the highest prediction accuracy when a VIF of 1.11 was used to remove SNPs with LD (0.756 ± 0.057). These results suggest that reduction in levels of pairwise LD among SNPs could improve the accuracy of genomic prediction models.

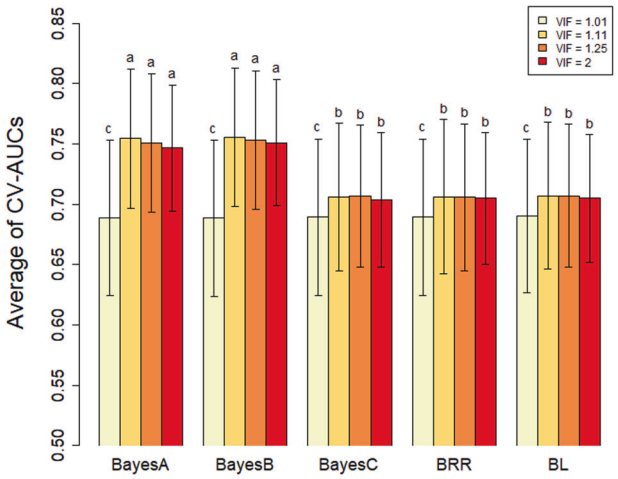

Fig. 3 Genomic prediction accuracy of selected SNPs based on LD-pruning using Bayes-A, Bayes-B, Bayes-C, Bayes LASSO (BL), and Bayesian ridge regression (BRR) models. A different range of variance inflation factor (VIF) including 2, 1.25, 1.11, and 1.01 were used to imply multiple correlation coefficients of 0.5, 0.2, 0.1, and 0.01, respectively. AUCs with different superscript(s) are significantly different from each other (p value < 0.05).

Methylome prediction accuracy using all methylation sites

Table 2 shows the methylome prediction accuracy in training and testing sets using adjusted-cell, adjusted-all, and nonadjusted methylation data with different Bayesian methods. In the adjusted-cell approach, the methylation signatures were adjusted for varying proportions of white blood cell types, whereas in the adjusted-all correction, the methylation signatures were adjusted for blood cell

Table 2 Prediction accuracy (standard errors) from the average of cross-validation AUC using all methylation data in different methods: Bayes-A, Bayes-B, Bayes-C, Bayesian LASSO (BL), and Bayesian ridge regression (BRR).

Model	Testing population		
	Nonadjusted	Adjusted-cell[1]	Adjusted-all[2]
Bayes-A	0.868 (0.044)[a]	0.821 (0.050)[b]	0.657 (0.057)[c]
Bayes-B	0.868 (0.043)[a]	0.821 (0.051)[b]	0.659 (0.057)[c]
Bayes-C	0.867 (0.044)[a]	0.820 (0.050)[b]	0.660 (0.058)[c]
BL	0.865 (0.044)[a]	0.821 (0.048)[b]	0.662 (0.058)[c]
BRR	0.867 (0.044)[a]	0.820 (0.050)[b]	0.658 (0.057)[c]

[1]Methylation signatures adjusted for cell proportions only.

[2]Methylation signatures were adjusted for all available confounders including cell proportions, age, sex, and smoking status. Average AUCs in testing populations with different superscript(s) are significantly different (p value < 0.05).

proportion and other available explanatory variables including age, sex, and smoking status. Nonadjusted methylation inputs delivered higher estimates of prediction accuracy than when adjusted-cell or adjusted-all methylation data. When nonadjusted methylation data were used, model goodness of fit was better than with adjusted methylation data. Estimated accuracy was similar across the five Bayesian models.

Prediction accuracy using subsets of methylation sites

The prediction accuracies of the various subsets of methylation sites used in five Bayesian models are shown in Fig. 4. When methylation sites were selected randomly or evenly spaced (Fig. 4a, b, respectively), there were no significant differences among the three levels of selected methylation sites ($p > 0.05$). However, when methylation sites were selected based on their estimated effects, there was a clear difference between the levels of selected. BRR subset selection was the most efficient method for selection of methylation sites, and the highest prediction accuracy was achieved when 1% of the methylation sites was selected using this method, regardless of the model used for prediction (Fig. 4d).

Prediction accuracy using integrated methylome and genomic data

Four different combinations of omics data were used and compared in terms of prediction accuracy. As Bayes-B performed better than other models on both methylation and SNP data, it was the model chosen for the integrated analysis of methylome and genomic data. First, 1% of the methylation data (using BRR as selection method) were integrated either with all of the SNP data as random effects (Random in WGP), or with all SNP effects fitted as random except for SNP rs660895 fitted as fixed (Fixed + Random method in WGP). Second, 1% of the methylation data and pruned SNPs based on LD were used jointly. Lastly, 1% of the methylation data and rs660895 as fixed were included in the model. The results of these four analyses (Table 3) showed that integrating methylation and SNP data increased prediction accuracy relative to models where SNP and methylation data were fitted separately (WGP and WMP).

Estimated SNPs effects using WGP and WGMP

SNPs effects were estimated using the average from the 20 CV results. It has been shown theoretically that SNPs effects are highly dependent on the prior distribution assumed (Gianola 2013). We verified such expectation on empirically and it was joint with estimation of allelic substitution varied over the prior adopted (Supplementary Information, Figs. S3 and S4). When Bayes-A and Bayes-B were used, the distribution of SNP effects from WGP showed a few SNPs with a major effect, whereas other SNPs had small effects (Supplementary Information, Fig. S3). The SNPs with large effects were found on chromosomes 3, 6, 12, and 14. Estimated effects of top SNPs other than those in chromosome 6 could also be due to sampling error. By fitting methylation data and SNPs information simultaneously with Bayes-B, no major SNP effect was detected (Fig. 5a, b). The correlation between SNP effect estimates using the WGMP and WGP was very low with 0.07 (Fig. 5c). This phenomenon is due to exacerbation of the $n < p$ problem when methylation sites are added to the SNP data. However, when 20 outliers based on χ^2 scores in WGP were removed, the correlation increased drastically to 0.68.

The top ten SNPs with the strongest estimated effects in WGP are shown in Table 4. The SNP with the strongest effect on RA was located on chromosome 12, at 91 Kb distance from the nearest neighboring gene, which is a long noncoding RNA (lncRNA). Six SNPs were located within or nearby four genes that were previously shown to have direct or indirect significant association with RA (Hao et al. 2014; Hirota et al. 2011; Kapitany et al. 2005; Orozco et al. 2005). More detailed functional annotations of genes related to the disease are in Supplementary Information Table S1. Butyrophilin-like 2 (BTNL2) is one of the genes listed which showed an indirect association through its strong LD with a mutation in the major histocompatibility complex, class II, DQ beta 1 (HLA-DRB1), which can increase RA risk (Orozco et al. 2005). The other three genes, including chromosome 6 open reading frame 10 (C6orf10), major histocompatibility complex, class II, DQ alpha 1(HLA-DQA1), and major histocompatibility complex, class II, DR alpha (HLA-DRA) showed a direct

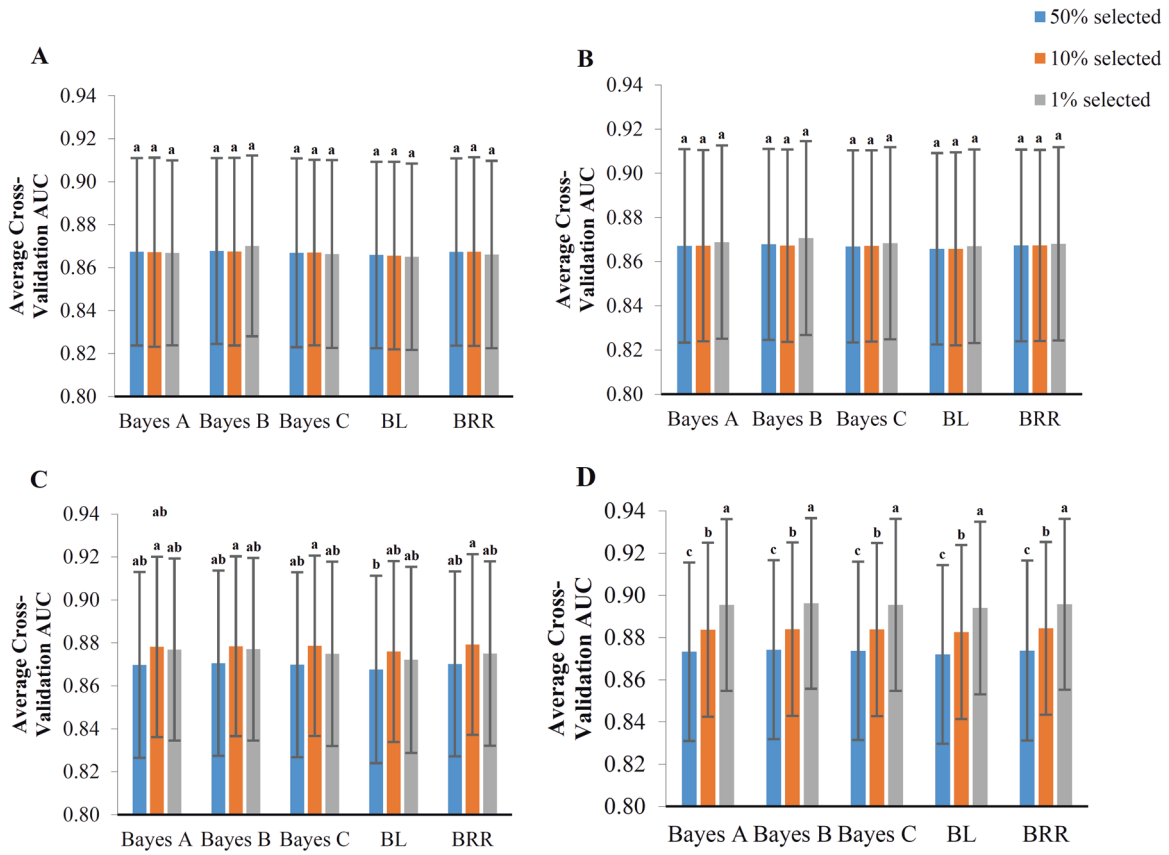

Fig. 4 Prediction accuracy of three levels of selected methylation sites (1, 10, and 50%) using Bayes-A, Bayes-B, Bayes-C, Bayesian LASSO (BL), and Bayesian ridge regression (BRR) models with different approaches to subsetting. Randomly (**a**), evenly spaced (**b**), EWA *p* value (**c**), and BRR effects (**d**). Bars with different letters indicate significant differences in average AUCs (*p* value < 0.05).

Table 3 Prediction accuracy from the average of cross-validation AUC using SNPs and methylation data simultaneously.

Integrated data	Prediction accuracy
1% of methylation selected based on BRR method	
+All SNPs as random effects except rs660895 as fixed	0.927 (0.032)[b]
+All SNPs as random effects	0.924 (0.034)[b]
+Subset of 41,673 LD-pruned SNPs + rs660895 as fixed	0.975 (0.014)[a]
+rs660895 as fixed	0.906 (0.037)[c]

Average AUCs with different superscript(s) are significantly different (*p* value < 0.05).

association with RA (Hao et al. 2014; Hirota et al. 2011; Kapitany et al. 2005). The second top SNP is located in vestigial-like 4 (VGLL4), which acts as a tumor suppressor (Jiang et al. 2015). To our knowledge, there are no studies indicating a connection between VGLL4 and RA disease risk. As this gene may play a significant role in apoptotic pathways (Jin et al. 2011), it may affect the disease risk of RA indirectly through programmed cell death (Liu and Pope 2003). The most significant SNP detected in the GWAS, rs660895, was ranked third based on SNP effects, and it is 19 Kb away from HLA-DQA1.

Five top SNPs with the strongest effect from WGMP were located nearby different classes of noncoding RNAs (ncRNA), whose functions remain to be understood (Table 5). Other top SNPs in WGMP are located close to coding genes including KHNYN, VGLL4, HLA-DRA, an uncategorized gene with ensemble code ID "ENSG00000279427" and sapiens roundabout guidance receptor 2 (ROBO2). Three of these genes, KHNYN, VGLL4, HLA-DRA appear in Table 4 for WGP. Previous studies suggest an association between a mutation in ROBO2 and risk of vesicoureteral reflux (Lu et al. 2007) and probably an effect of the expression level of the gene on prostate cancers (Choi et al. 2014); however, there are no reports on connection between ROBO2 and RA disease. In this list of top ten SNPs, only rs2395175 was connected to RA within an upstream location of HLA-DRA. The pathway analysis showed that terms associated with celiac disease, cholesterol level, and tobacco use disorder were enriched among candidate genes (see Supplementary Information, Table S2).

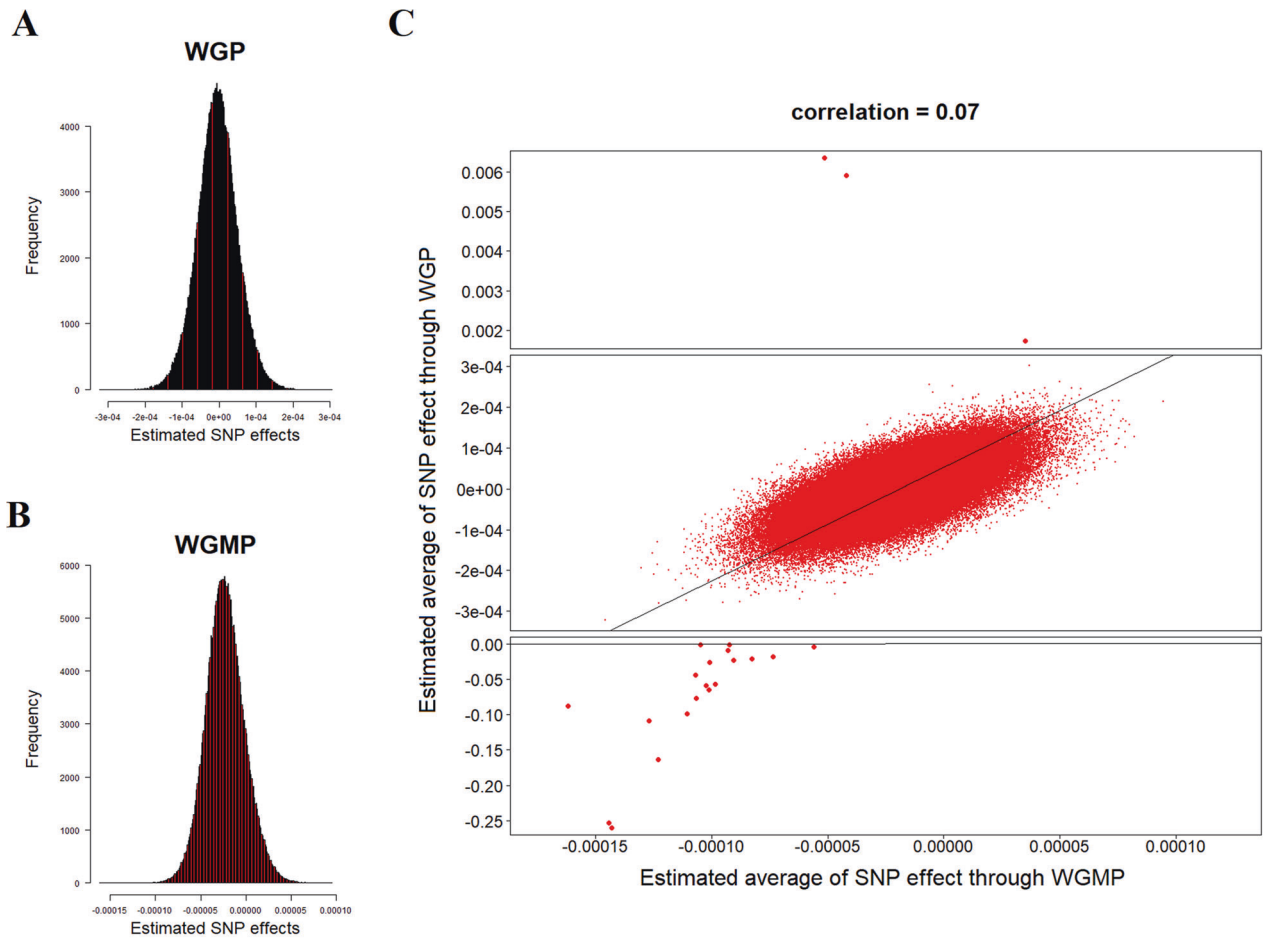

Fig. 5 Comparison of estimated SNP effects using two methods including whole-genome prediction (WGP) and whole-genome/methylome prediction (WGMP). Histogram of estimated SNP effects using WGP (**a**) and WGMP (**b**). **c** Scatter plot between estimated SNP effects using WGP and WGMP. The inset panels show SNPs with large (top panel) and small effect (below panel).

Table 4 Top ten SNPs with the strongest effects from whole-genome prediction using Bayes-B method fitting all SNPs as random effects.

SNP name	Chromosome: base pair	Alleles	Major allele	MAF[a]	Gene name[b]	Distance from gene	Effect
rs11179382	12: 72820377	C/T	T	0.269	A lncRNA with ensemble code ID ENSG00000258235	91 Kb downstream	−0.26015
rs2077507	3: 11576994	A/G	A	0.287	Vestigial-like family member 4 (VGLL4)	Body	−0.25294
rs660895	6: 32609603	A/G	A	0.198	Major histocompatibility complex, class II, DQ alpha 1 (HLA-DQA1)	19 Kb upstream	−0.16400
rs2395175	6: 32437249	A/G	G	0.098	Major histocompatibility complex, class II, DR alpha (HLA-DRA)	2.6 Kb upstream	−0.10898
rs2395163	6: 32420032	C/T	T	0.153	Butyrophilin-like 2 (BTNL2)	13 Kb downstream	−0.09879
rs1004664	14: 24424005	G/T	T	0.376	KH and NYN domain containing (KHNYN)	Body	−0.08766
rs3763309	6: 32408196	A/C	C	0.150	Butyrophilin-like 2 (BTNL2)	Body	−0.07728
rs2395157	6: 32380368	A/G	A	0.208	A ncRNA LOC101929163 (uncharacterized)	Body	−0.06514
rs3817963	6: 32400310	A/G	A	0.257	Butyrophilin-like 2 (BTNL2)	Body	−0.05923
rs6910071	6: 32315077	A/G	A	0.105	Chromosome 6 open reading frame 10 (C6orf10)	Body	−0.05694

[a]*MAF* minor allele frequency.

[b]*ncRNA* noncoding RNA, *lncRNA* long noncoding RNA.

Table 5 Top ten SNPs with the strongest effects from whole-genome/methylome prediction using Bayes-B and fitting all SNPs as random effects.

SNP name	Chromosome: base pair	Alleles	Major allele	MAF[a]	Gene name[b]	Distance from gene	Effect
rs1004664	14: 24424005	G/T	T	0.3756	KH and NYN domain containing (KHNYN)	Body	−0.00016
rs513203	12: 63545960	A/G	A	0.4643	A ncRNA class with ensemble code ID LOC105369797	Body	−0.00015
rs2077507	3: 11576994	A/G	A	0.2875	Vestigial-like family member 4 (VGLL4)	Body	−0.00014
rs11179382	12: 72820377	C/T	T	0.2686	A lncRNA with ensemble code ID ENSG00000258235	91.53 Kb downstream	−0.00014
rs2911738	8: 71674057	A/C	A	0.3361	A lincRNA with ensemble code ID ENSG00000254277	1.24 Kb upstream	−0.00013
rs2395175	6: 32437249	A/G	G	0.0982	Major histocompatibility complex, class II, DR alpha (HLA-DRA)	2.59 Kb upstream	−0.00013
rs8049226	16: 52423230	A/G	G	0.3672	An uncategorized gene with ensemble code ID ENSG00000279427	13.54 Kb upstream	−0.00013
rs1364616	8: 71654009	G/T	G	0.3321	A lincRNA with ensemble code ID ENSG00000254277	21.29 Kb upstream	−0.00012
rs6797550	3: 76747494	A/G	G	0.3281	Sapiens roundabout guidance receptor 2 (ROBO2)	Body	−0.00012
rs2595415	11: 6897324	C/T	C	0.2821	A ncRNA LOC107984019 (uncharacterized)	Body	−0.00012

[a]*MAF* minor allele frequency.
[b]*ncRNA* noncoding *RNA*, *lncRNA* long noncoding RNA.

Estimated methylation effects based on the training population using WMP and WGMP

Methylation site effects were also estimated from the average of 20 CV results. In general, results suggested the estimated methylation effects in all different Bayesian model specifications normally distributed (Supplementary Information, Fig. S5). The distribution of estimated methylation effects using Bayes-B and 1% of selected methylation sites based on BRR are shown in Fig. 6a, b for WMP and WGMP, respectively. The correlation between methylation site effects estimated in WMP and WGMP was 0.73, which was much higher than the correlation between estimated average SNP effects from WGP and WGMP (Fig. 6c).

The top ten methylation sites with the strongest effect in 1% of selected subset based on BRR results for WMP are shown in Table 6 (Beta-value pattern of these methylation sites is also shown in Supplementary Information, Fig. S6). According to the pathway analysis, only two methylation sites were connected to RA (Supplementary Information, Table S3), including the cg24147543 site, which is located in major histocompatibility complex, class II, DR beta 1 (HLA-DRB1) and was on the top of our list. Several allelic variants in this gene have shown a significant association with RA (Weyand and Goronzy 2000). The ninth methylation site, cg18858739, which is located within the CD247 molecule was suggested to be involved in RA risk (Stahl et al. 2010). The second methylation site in the list, cg10132543, is located in the Carnitine palmitoyltransferase 1A (CPT1A) gene, which encodes an enzyme that regulates the production of reactive oxygen species (ROS) within mitochondria (Rosca et al. 2012). The ROS roles in the pathogenesis of inflammatory chronic arthropathies such as RA have been demonstrated previously (Filippin et al. 2008). Three of the listed methylation sites (cg26572452, cg04362887, and cg09717927) are located far from annotated genes. For instance, cg26572452 is located 129 kb from any gene, while sites cg04362887 and cg09717927 are 91 and 112 Kb away from the closest gene, respectively. There was no evidence for connectedness between three nearest genes, forkhead box F1 (FOXF1), Chromosome 14 Open Reading Frame 70 (C14orf7) and member RAS oncogene family (RAB28), to the top methylation sites and RA disease risk. Catenin beta 1 (CTNNB1), a mediator of the Wnt signal and also a component of E-cadherin complexes at the intercellular adhering junction, was another listed gene, which expression in synovial lining cells of the RA samples has been shown to be high (Xiao et al. 2011).

After fitting WGMP, four listed genes in top methylation sites in WMP (CPT1A, HLA-DRB1, CD247, and LINC00977) remained in the top ten list in WGMP

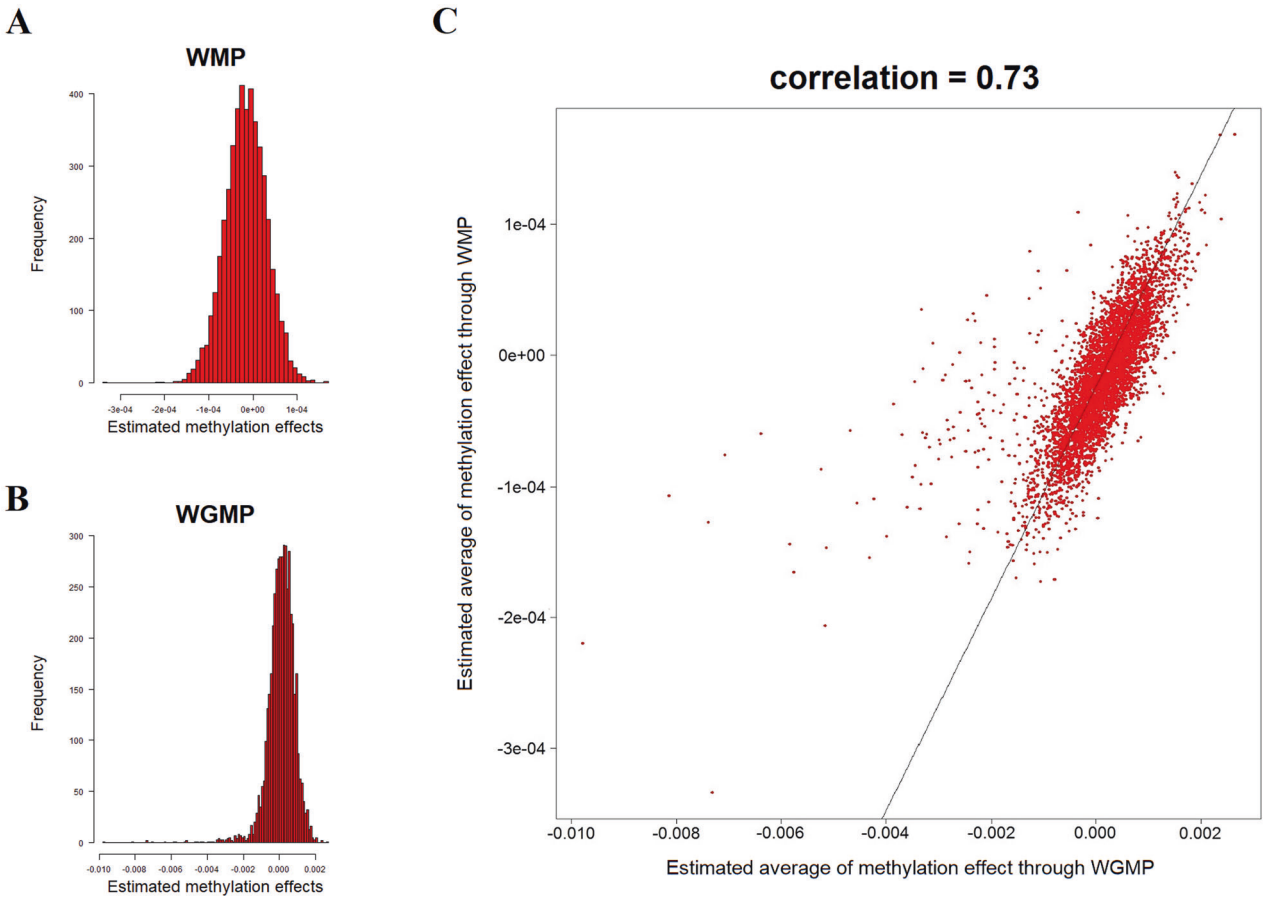

Fig. 6 Comparison of estimated methylation site effects using two methods including whole-methylome prediction (WMP) and whole-genome/methylome prediction (WGMP). Histogram of estimated methylation site effects using WMP (**a**) and WGMP (**b**). **c** Plot of estimated effects for methylation sites from WMP and WGMP.

(Table 7). Diseases that are associated with some of the listed genes are shown in Supplementary Information (Table S4). Box-plots of the Beta-values for these methylation sites for cases and controls are shown separately in the Supplementary Information (Fig. S7). To our knowledge, there is a limited number of studies showing a relationship between the five listed genes including XK related 9 (XKR9), sequence similarity 189-member B (FAM189B), MORN repeat containing 1 (MORN1), caldesmon 1 (CALD1), and peptidylprolyl isomerase like 4 (PPIL4) and RA. An association between one of the listed genes in the HLA locus, complex, class II, DR beta 1 (HLA-DRB1), and RA was reported (Okamoto et al. 2003). A previous study demonstrated that level of trimethylation of lysine 4 on histone H3 (H3K4me3) in lactamase beta 2 (LACTB2) was significantly higher in peripheral blood mononuclear cells in patients with RA (Dai et al. 2010). It is accepted that the DNA and histone lysine methylation systems are related mechanistically (Rose and Klose 2014). Consequently, methylation in this site may have some connection with H3K4me3 modification and increase RA disease risk.

Discussion

This study investigated alternative modeling approaches for prediction of RA using SNP and methylation data, including the effect of correcting methylation signatures for cellular heterogeneity, the integration of genome and methylome data, and different strategies of preselection of predictors. Overall, the importance of multiomics data for prediction of disease risk was indicated.

Our results indicated that classification of RA subjects using SNPs chip data is a promising tool. Moreover, such classification can be even more efficient when methylation information is incorporated. Here, we compared predictive models using these two sources of omics information (SNP markers and methylation sites), either separately or jointly in a single model.

The accuracy of WGP depends on the number of loci and on the distribution of their effects (Li et al. 2012; Momen et al. 2018). Riedelsheimer et al. (2012) and Momen et al. (2018) reported small differences among some parametric and semiparametric WGP models using traits with drastically different genetic architectures. They found that

Table 6 Top ten methylation sites with the strongest effects from whole-methylome prediction using Bayes-B.

Methylation site[a]	Chromosome: base pair	Relation to island	Gene name	Relation to gene	Regulatory feature group	Effect
cg24147543	6: 32554481	S shelf	Major histocompatibility complex, class II, DR beta 1 (HLA-DRB1)	Body	NA	−0.00033
cg10132543	11: 68609520	N shore	Carnitine palmitoyltransferase 1A (CPT1A)	TSS200	NA	−0.00022
cg26572452*	8: 129115055	Open sea	Long intergenic nonprotein coding RNA 977 (LINC00977)	129.22 Kb upstream	Unclassified	−0.00021
cg04362887	14: 91285234	S shelf	Coiled-coil domain containing 88C (CCDC88C)	91.271 Kb downstream	NA	−0.00017
cg01695533	16: 86542905	Island	Forkhead box F1 (FOXF1)	TSS1500	Unclassified cell type specific	−0.00017
cg19741273	14: 101123263	S shore	Chromosome 14 open reading frame 70 (C14orf7)	TSS1500	NA	−0.00017
cg09717927	13: 112630399	N shore	ATP11A upstream neighbor (ATP11AUN)	112.26 Kb upstream	NA	−0.00017
cg05726118*	3: 41265374	Open sea	Catenin beta 1 (CTNNB1)	5'UTR	NA	−0.00017
cg18858739*	1: 167426278	S shore	CD247 molecule (CD247)	Body	Promoter associated cell type specific	−0.00017
cg12853199*	4: 13356876	Open sea	Member RAS oncogene family (RAB28)	13.361 Kb downstream	NA	−0.00016

[a]Methylation sites with enhancer regulatory are shown by an '*' on subscript.

Table 7 Top ten methylation sites with the strongest effects from whole-genome/methylome prediction using Bayes-B.

Methylation site[a]	Chromosome: base pair	Relation to Island	Gene name	UCSC reference gene group	Regulatory feature group	Effect
cg10132543	11: 68609520	N shore	Carnitine palmitoyltransferase 1A (CPT1A)	TSS200	NA	−0.00978
cg15821716	8: 71581642	Island	XK related 9 (XKR9); Lactamase beta 2 (LACTB2)	5′UTR and 1st Exon; TSS200	Promoter associated	−0.00814
cg00393182	1: 1.55E+08	N shore	Sequence similarity 189 member B (FAM189B)	Body	NA	−0.00739
cg24147543	6: 32554481	S shelf	Major histocompatibility complex, class II, DR beta 1 (HLA-DRB1)	Body	NA	−0.00731
cg18106803	1: 2303963	N shore	MORN repeat containing 1 (MORN1)	Body	NA	−0.00708
cg09853238*	6: 149532290	Open sea	Peptidylprolyl isomerase like 4 (PPIL4)	13.7 Kb downstream	Unclassified cell type specific	−0.00638
cg17360552	6: 32725332	Open sea	Major histocompatibility complex, class II, DQ beta 2 (HLA-DQB2)	Body	NA	−0.00585
cg18858739*	1: 167426278	S shore	CD247 molecule (CD247)	Body	Promoter associated cell type specific	−0.00576
cg15266530	7: 134516810	Open sea	Caldesmon 1 (CALD1)	5′UTR	NA	−0.00524
cg26572452*	8: 129115055	Open sea	Long intergenic nonprotein coding RNA 977 (LINC00977)	101 Kb upstream	Unclassified	−0.00516

[a]Methylation sites with enhancer regulatory are shown by an '*' on subscript.

selecting a WGP model that contemplates genetic architecture can result in a small gain in accuracy. We found that Bayes-B was the model producing the highest accuracy of prediction. Fitting SNPs with major signals as fixed effects can reduce discrepancy between models with different prior assumptions. The choice of prior distributions can have large impacts on prediction accuracy and model fitting (Table 1). Bayesian models that utilize a scaled-t prior density result in a heavy-tailed prior for the random effects and may estimate SNP effects that are linked to a large-effect QTL better. Thus, the scaled-t prior used as slab in Bayes-B perhaps better captures the underlying architecture of RA compared with other approaches. For methylation analysis, only minor differences were observed in classification accuracy among models evaluated.

There are some reports showing association between DNA methylation and diseases, including RA (Liu et al. 2013), cancer (Shenker et al. 2013; Teschendorff et al. 2009), and type 1 diabetes (Rakyan et al. 2011). A previous study with this dataset found a robust association between methylome modifications and risk of RA (Liu et al. 2013). Another EWAS on B lymphocytes with three different cohorts showed validated associations between RA risk and two CpGs located near the MHC class I-like glycoprotein, CD1C, and a cytokine that belongs to the TNF ligand family, TNFSF10 (Julià et al. 2017). It also showed that adjustment for cellular heterogeneity can reduce the confounding effect due to of cell type heterogeneity on methylation profiles in EWAS. Although this adjustment is recommended when methylation data are derived from whole blood samples (Jaffe and Irizarry 2014), our results indicated that the use of nonadjusted methylation profile can result in a higher classification accuracy. Our WMP analysis also showed detrimental effects on prediction accuracy from adjusting for cell heterogeneity and for other confounders such as gender, age, and smoking status. However, there are other possible confounder effects (e.g., lifestyle, nutrition, and environmental stress), which were not accounted for in our adjustment for methylation signatures. Previous epidemiological studies have identified smoking, gender, and age as important risk factors for RA (Di Giuseppe et al. 2014; Goronzy et al. 2010; Linos et al. 1980), and methylation of DNA may also be modified by these factors (Breitling et al. 2011; Hannum et al. 2013). It has been shown that the proportion of white blood cells in patients with RA can differ from that in healthy controls (Takeshita et al. 2019). We found that blood cell type proportions in RA cases can differ from controls (Supplementary Information, Table S5). This change may produce distinct methylation of DNA in cases and controls, leading to a more efficient separation when using unadjusted values. It is accepted that cell heterogeneity can produce false discoveries in EWAS, and many such studies have highlighted the importance of cell type correction (Liu et al. 2013). However, here we found a decrease in prediction accuracy from using the adjusted-all methylation data related to what was attained with non-adjusted methylation data.

Use of multiomics data for WGP of disease risk has increased in recent years. Choosing a tissue that can represent epigenetic modifications closely connected to the disease is important and difficult. RA is a complex disease, dependent on genetic and environmental factors (Liu et al. 2013), and our results indicated that leukocytes, one of the main classes of cells involved in the disease, can effectively represent its epigenetic regulation. Another important issue that affects prediction accuracy is the selection of the best combination of omics data. In breast cancer, for example, combining whole-genome gene expression profiles and whole-genome methylation profiles produced good predictive ability (Vazquez et al. 2016). Another comprehensive study on various types of the cancer (ovarian, renal, glioblastoma multiform, and lung squamous cell carcinoma) using omics data and clinical covariates, found that combination of molecular data with clinical variables significantly improved predictive ability for three of the diseases (Yuan et al. 2014). Our analyses also showed that consigned of two layers of omics, methylation and SNP data, can provide a higher accuracy than a single layer alone for classification of RA. Our research hints at the importance of methylation information for RA. A higher classification accuracy from using methylation data than from genotype information would suggest that epigenetic regulation is more relevant for clinical assessment of RA disease pathology.

Difficulties are present when the number of recorded individuals is smaller than the number of predictors from multilayer omics data. For fitting these models, some type of variable selection or shrinkage estimation procedure is needed (de los Campos et al. 2013a). As Bayesian methods can handle the problem of a larger number of the predictors than of samples, but choosing an appropriate method for preselection of predictors (LD-based pruning and BRR subset selection for SNPs and methylation sites, respectively) may assist in improving prediction ability. For instance, when 1% of the methylation sites was selected based on their effects estimated with BRR, the AUC had a significant increase of up to 2.8 points over the model fitted using all methylation signatures, and of 2.6 points over the model fitted using 1% randomly selected methylation sites. The improvement of predictive ability was limited probably due to the small sample size used. Although accuracy of prediction was reasonable, additional investigations using a larger sample could provide extra insights regarding the potential prediction ability of different subset selection methods of SNP markers and methylation sites. Accuracies across different layers of omics data, prediction models and

subset selection methods suggested that using Bayes-B with subset of SNPs following LD-pruning with methylation sites selected with BRR had the best classification performance. This approach can be used to effectively predict the risk of RA and identify cases in early stages to prevent or alter disease progression and potentially lead to drug-free remission.

Our pathway analysis of the top ten SNPs showed that six were connected to the disease through genes shown to have an association with RA in previous studies. One of the genes we found both with WGP and WGMP, VGLL4, has not been shown to have an association with RA in earlier studies. The effect of VGLL4, if real, could be due to its role in apoptotic pathways (Jin et al. 2011), as insufficient apoptosis of inflammatory cells was found in the RA joint (Pope 2002). HLA-DRB1 and CD247 genes have also been shown to have a significant association with RA (Stahl et al. 2010; Weyand and Goronzy 2000), and were included in our list when methylation data were used for prediction (WMP and WGMP). These two genes might be linked to RA through epigenetic regulations other than through SNP variation. The CpG site with the strongest estimated effect from WGMP analyses, cg10132543, was located within promoter sequences of CPT1A gene. This gene is involved in producing ROS, which may enhance disease risk through altering the pathogenesis of inflammatory chronic arthropathies (Filippin et al. 2008; Rosca et al. 2012). We suggest that this gene be investigated for its potential role in RA in future studies.

Our results showed that one of the methylation sites with the strongest effect was close to LACTB2, which showed an increase in level of H3K4me3 in RA. This suggests a connection between methylation and H3K4me3 modification in RA to increase disease risk.

Visscher et al. (2017) showed that, to reach an adequate power from detecting association, some factors such as genotyping method, sample size, allele frequency, and effect size are important. For example, for case-control studies of disease with an allele effect size of 0.01, 0.1, and 1 phenotypic standard deviations, and a minor allele frequency of 0.01, sample size with more than 10 million, 100 thousands, and 1000 individuals, respectively, are needed. These sample sizes were calculated based on unselected population samples and for highly ascertained cases, similar to the samples of our study, power could be increased.

Data availability

The Illumina 450K array data analyzed in the current study are publicly available in GEO with accession number "GSE42861" (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861). The genotyping data are not publicly available due to EIRA policies but are available from the corresponding author upon request.

Acknowledgements We appreciate EIRA (EIRA Institute, Karolinska Institute, Stockholm, Sweden) study members for sharing genotyping data and for help. The first author also would like to thank Malachy Campbell from Virginia Polytechnic Institute and State University for his help in proofreading the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD et al. (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30:1363–1369

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM et al. (2011) High density DNA methylation array with single CpG site resolution. Genomics 98:288–295

Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet 88:450–457

Calus MP, Bouwman AC, Schrooten C, Veerkamp RF (2016) Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genet Sel Evol 48:49

Choi YJ, Yoo NJ, Lee SH (2014) Down-regulation of ROBO2 expression in prostate cancers. Pathol Oncol Res 20:517–519

Choy E (2012) Understanding the dynamics: pathways involved in the pathogenesis of rheumatoid arthritis. Rheumatology 51: v3–v11

Dai Y, Zhang L, Hu C, Zhang Y (2010) Genome-wide analysis of histone H3 lysine 4 trimethylation by ChIP-chip in peripheral blood mononuclear cells of systemic lupus erythematosus patients. Clin Exp Rheumatol 28:158

de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11:880

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013a) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345

de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013b) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9:e1003608

Di Giuseppe D, Discacciati A, Orsini N, Wolk A (2014) Cigarette smoking and risk of rheumatoid arthritis: a dose-response meta-analysis. Arthritis Res Ther 16:R61.

Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 4:1184–1191

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874

Filippin L, Vercelino R, Marroni N, Xavier R (2008) Redox signalling and the inflammatory response in rheumatoid arthritis. Clin Exp Immunol 152:415–422

Fisher R (1918) The correlation between relatives on the supposition of mendelian inheritance. Trans R Soc Edinb 52:399–433
Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ et al. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol 15:503
Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194:573–596
Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical data with a threshold model. Genetique Selection Evol 15:201
Gianola D, Perez-Enciso M, Toro MA (2003) On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163:347–365
Gianola D, Rosa GJ (2015) One hundred years of statistical developments in animal breeding. Annu Rev Anim Biosci 3:19–56
Glant TT, Mikecz K, Rauch TA (2014) Epigenetics in the pathogenesis of rheumatoid arthritis. BMC Med 12:1–5
Goronzy JJ, Shao L, Weyand CM (2010) Immune aging and rheumatoid arthritis. Rheum Dis Clin North Am 36:297–310
Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinform 12:1
Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S et al. (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell 49:359–367
Hao G-F, Li Y-S, Liu J-L, Wo M-Y (2014) Association of HLA-DQA1 (rs9272219) with susceptibility to rheumatoid arthritis in a Han Chinese population. Int J Clin Exp Pathol 7:8155–8158
Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, Doi S et al. (2011) Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. Nat Genet 43:893–896
Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH et al. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform 13:1–16
Hu X, Xie W, Wu C, Xu S (2019) A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol J 17:2011–2020
Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57
Jaffe AE, Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 15:R31
Jiang W, Yao F, He J, Lv B, Fang W, Zhu W et al. (2015) Down-regulation of VGLL4 in the progression of esophageal squamous cell carcinoma. Tumour Biol 36:1289–1297
Jin HS, Park HS, Shin JH, Kim DH, Jun SH, Lee CJ et al. (2011) A novel inhibitor of apoptosis protein (IAP)-interacting protein, vestigial-like (Vgl)-4, counteracts apoptosis-inhibitory function of IAPs by nuclear sequestration. Biochem Biophys Res Commun 412:454–459
Julià A, Absher D, López-Lasanta M, Palau N, Pluma A, Waite Jones L et al. (2017) Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. Hum Mol Genet 26:2803–2811
Kapitany A, Zilahi E, Szanto S, Szucs G, Szabo Z, Vegvari A et al. (2005) Association of rheumatoid arthritis with HLA-DR1 and HLA-DR4 in Hungary. Ann N. Y. Acad Sci 1051:263–270
Li A, Meyre D (2014) Jumping on the train of personalized medicine: a primer for non-geneticist clinicians: part 3. Clinical applications in the personalized medicine area. Curr Psychiatry Rev 10:118–132
Li W, Zhang S, Liu C-C, Zhou XJ (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics 28:2458–2466
Linos A, Worthington JW, O'Fallon WM, Kurland LT (1980) The epidemiology of rheumatoid arthritis in Rochester, Minnesota: a study of incidence, prevalence, and mortality. Am J Epidemiol 111:87–98
Liu H, Pope RM (2003) The role of apoptosis in rheumatoid arthritis. Curr Opin Pharm 3:317–322
Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A et al. (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31:142–147
Lu W, Van Eerde AM, Fan X, Quintero-Rivera F, Kulkarni S, Ferguson H et al. (2007) Disruption of ROBO2 is associated with urinary tract anomalies and confers risk of vesicoureteral reflux. Am J Hum Genet 80:616–632
Mayadas TN, Tsokos GC, Tsuboi N (2009) Mechanisms of immune complex mediated neutrophil recruitment and tissue injury. Circulation 120:2012–2024
Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
Momen M, Mehrgardi AA, Sheikhi A, Kranis A, Tusell L, Morota G et al. (2018) Predictive ability of genome-assisted statistical models under various forms of gene action. Sci Rep. 8:12309
Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLOS Genet 11:e1004969
Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K et al. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506:376–381
Okamoto K, Makino S, Yoshikawa Y, Takaki A, Nagatsuka Y, Ota M et al. (2003) Identification of IκBL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. Am J Hum Genet 72:303–312
Orozco G, Eerligh P, Sanchez E, Zhernakova S, Roep BO, Gonzalez-Gay MA et al. (2005) Analysis of a functional BTNL2 polymorphism in type 1 diabetes, rheumatoid arthritis, and systemic lupus erythematosus. Hum Immunol 66:1235–1241
Padyukov L, Seielstad M, Ong RT, Ding B, Ronnelid J, Seddighzadeh M et al. (2011) A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. Ann Rheum Dis 70:259–265
Park T, Casella G (2008) The bayesian lasso. J Am Stat Assoc 103:681–686
Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. N. Engl J Med 357:1199–1209
Plummer M, Best AN, Cowles AK, Vines AK (2006) CODA: convergence diagnosis and output analysis for MCMC. R N. 6:7–11
Pope RM (2002) Apoptosis as a therapeutic tool in rheumatoid arthritis. Nat Rev Immunol 2:527–535
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D et al. (2011) Identification of type 1 diabetes–associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet 7:e1002300
Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X et al. (2012) Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 44:291–296
Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D et al. (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE 7:e41361
Riedelsheimer C, Technow F, Melchinger AE (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452

Rosca MG, Vazquez EJ, Chen Q, Kerner J, Kern TS, Hoppel CL (2012) Oxidation of fatty acids is the source of increased mitochondrial reactive oxygen species production in kidney cortical tubules in early diabetes. Diabetes 61:2074–2083

Rose NR, Klose RJ (2014) Understanding the relationship between DNA methylation and histone lysine methylation. Biochim Biophys Acta 1839:1362–1372

Sattar N, McInnes IB (2005) Vascular comorbidity in rheumatoid arthritis: potential mechanisms and solutions. Curr Opin Rheumatol 17:286–292

Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA et al. (2013) Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet 22:843–851

Silman AJ, Pearson JE (2002) Epidemiology and genetics of rheumatoid arthritis. Arthritis Res 4:S265–S272

Speed D, Balding DJ (2014) MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 24:1550–1557

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 42:508–514

Takeshita M, Suzuki K, Kondo Y, Morita R, Okuzono Y, Koga K et al. (2019) Multi-dimensional analysis identified rheumatoid arthritis-driving pathway in human T cell. Ann Rheum Dis 78:1346–1356

Tanoue LT (1998) Pulmonary manifestations of rheumatoid arthritis. Clin Chest Med 19:667–685

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S et al. (2009) An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS ONE 4:e8274

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525

Vazquez AI, Veturi Y, Behring M, Shrestha S, Kirst M, Resende MFR et al. (2016) Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multi-omic profiles. Genetics 203:1425–1438

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA et al. (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101:5–22

Walsh AM, Whitaker JW, Huang CC, Cherkas Y, Lamberth SL, Brodmerkel C et al. (2016) Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. Genome Biol 17:79

Weyand CM, Goronzy JJ (2000) Association of MHC and rheumatoid arthritis: HLA polymorphisms in phenotypic variants of rheumatoid arthritis. Arthritis Res Ther 2:212

Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, Huang RS et al. (2014) Poly-omic prediction of complex traits: OmicKriging. Genet Epidemiol 38:402–415

Xiao CY, Pan YF, Guo XH, Wu YQ, Gu JR, Cai DZ (2011) Expression of beta-catenin in rheumatoid arthritis fibroblast-like synoviocytes. Scand J Rheumatol 40:26–33

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A et al. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotech 32:644–652

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328