Xia et al. BMC Genetics 2018, 19(Suppl 1):78
https://doi.org/10.1186/s12863-018-0644-5

BMC Genetics

RESEARCH Open Access

Incorporating methylation genome information improves prediction accuracy for drug treatment responses

Xiaoxuan Xia[1,2†], Haoyi Weng[1,2†], Ruoting Men[1,2], Rui Sun[1,2], Benny Chung Ying Zee[1,2], Ka Chun Chong[1,2*] and Maggie Haitian Wang[1,2*]

From Genetic Analysis Workshop 20
San Diego, CA, USA. 4-8 March 2017

Abstract

Background: An accumulation of evidence has revealed the important role of epigenetic factors in explaining the etiopathogenesis of human diseases. Several empirical studies have successfully incorporated methylation data into models for disease prediction. However, it is still a challenge to integrate different types of omics data into prediction models, and the contribution of methylation information to prediction remains to be fully clarified.

Results: A stratified drug-response prediction model was built based on an artificial neural network to predict the change in the circulating triglyceride level after fenofibrate intervention. Associated single-nucleotide polymorphisms (SNPs), methylation of selected cytosine-phosphate-guanine (CpG) sites, age, sex, and smoking status, were included as predictors. The model with selected SNPs achieved a mean 5-fold cross-validation prediction error rate of 43.65%. After adding methylation information into the model, the error rate dropped to 41.92%. The combination of significant SNPs, CpG sites, age, sex, and smoking status, achieved the lowest prediction error rate of 41.54%.

Conclusions: Compared to using SNP data only, adding methylation data in prediction models slightly improved the error rate; further prediction error reduction is achieved by a combination of genome, methylation genome, and environmental factors.

Keywords: Methylation, Prediction, SNPs, Neural network, Treatment responses

Background

Increasing evidence reveals the important role of epigenetic factors in explaining the etiopathogenesis of human diseases, especially in cancer [1]. For example, Chaudhry et al. verified that *BRCA1* promoter methylation was useful in predicting the response to chemotherapy in epithelial ovarian cancer [2], and Shindo et al. found that a high methylation M-score was a significant risk factor for recurrent bladder cancer [3]. Diseases other than cancer have shown profound alterations in DNA methylation profiles [4, 5]. Consideration of the effect of epigenetic factors on disease traits has the potential to improve disease prediction, which has been adopted in several recent empirical studies [6–9]. However, it is still challenging to integrate different types of omics data into prediction models. In addition, there has been insufficient information to precisely clarify the contribution of methylation information to prediction.

In this study, a stratified drug-response prediction model is built based on an artificial neural network (ANN) to identify the contribution of methylation information to predicting the change in the circulating triglyceride (TG) level after fenofibrate intervention. Omics data, including genetic, epigenetic, and clinical factors, are used as predictors. The analysis of GAW20 real data demonstrates that the inclusion of the methylation data improves the prediction

* Correspondence: marc@cuhk.edu.hk; maggiew@cuhk.edu.hk
†Xiaoxuan Xia and Haoyi Weng contributed equally to this work.
[1]Division of Biostatistics, Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, N.T, Hong Kong, SAR, China
Full list of author information is available at the end of the article

accuracy marginally, which provides an indication for future prediction research.

Methods

GAW20 data

GAW20 real data were used in this study and were provided by the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study, which aimed to identify the genetic determinants of the responses of circulating lipid levels to fenofibrate treatment interventions. In total, 1053 individuals from families with at least 2 siblings were recruited. They all self-reported as being of white ethnicity [10]. TG levels were measured at visits 1, 2, 3, and 4, among which data from visits 1 and 2 were collected before fenofibrate intervention, whereas the other two TG measurements were made after the intervention (visits 3 and 4). At visit 1, participants were measured using a lipid profile after an overnight fast. A repeated lipid file occurred the next day during visit 2. The treatment period lasted 3 weeks, after which participants returned to the clinic for 2 consecutive days for visits 3 and 4 [10]. Meanwhile, DNA methylation levels were measured at visits 2 and 4. DNA was isolated from $CD4^+$ T cells harvested from stored buffy coats and the proportion of sample methylation was quantified at $>450{,}000$ cytosine-phosphate-guanine (CpG) sites [10].

Data quality control

In the quality control process, 39 participant outliers were removed, and only subjects without any missing data for the key variables (TG levels at visits 1 to 4, methylation value at visit 2, and genotypes) were used. A total of 523 participants were included in the analysis. For the genotype data, single-nucleotide polymorphisms (SNPs) with a minor allele frequency < 0.01 were excluded. Missing variants were imputed according to the probability distribution of the genotype in all subjects. For the methylation data, cross-reactive probes and probes containing common variants were filtered. Beta-mixture quantile normalization was used to correct for the Infinium Type I/II bias [11], and participant outliers were identified by hierarchical clustering and Eigenstrat [12].

Drug-response definition

Drug response was used as the dependent variable which could be defined as the percentage change in the TG level.

$$TG\ \ change\ \ percentage = (TG\ \ post - TG\ \ pre)/(TG\ \ pre)$$

Where *TG pre* is the average of TG levels at visits 1 and 2, and *TG post* is the average of TG levels at visits 3 and 4. It was reported that fenofibrate, which was the intervention drug for the GAW20 real data, usually reduced the plasma TG level by approximately 30 to 60% in hyperlipoproteinemia patients at a dosage of 200–400 mg daily [13]. In this regard, we defined the drug-response variable as 1 when the TG level was reduced by more than 30% after treatment, which meant the drug worked for patients. Otherwise, the drug-response variable was coded as 0, which meant that the drug did not work as expected. Consequently, as shown in Fig. 1, 301 and 222 participants were coded as 1 and 0, respectively.

Stratified variable selection and prediction modeling

The features related to drug response were selected in a stratified manner [14], first within each data type, and then aggregated in an ANN to predict the drug response [15]. ANNs are designed to perform learning tasks using a collection of computational units and a system of interlinking connections [16]. The central idea of ANN is to extract features by linearly combining the inputs and then use nonlinear functions to model the targets. Therefore, a neural

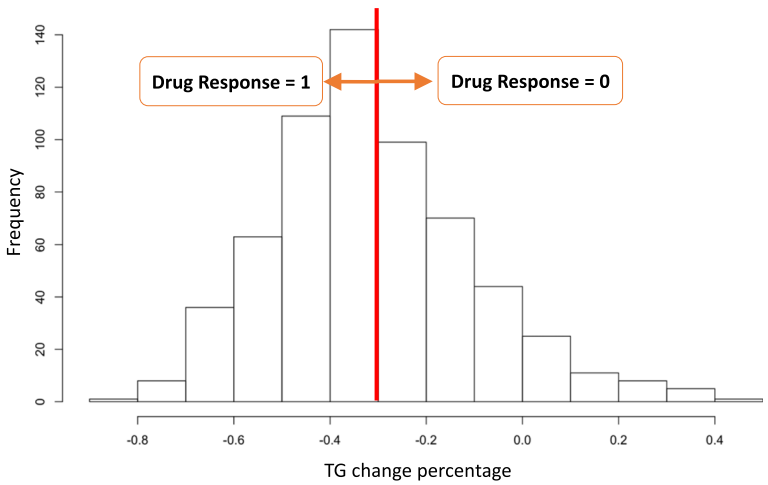

Fig. 1 Distribution of percentage change in circulating triglyceride (TG)

network can be thought of as a nonlinear generalization of linear models, which generalizations can be used for classification and regression [17]. We used the AMORE package in the R 3.3.2 GUI 1.68 Mavericks build (7288) to conduct the ANN analysis [15]. The stratification enables precise variable selection within each data type, and the ANN enables the consideration of interaction effects within and across data types [18]. Five-group cross-validation error rates and their standard deviation were calculated to evaluate prediction performance.

The generalized estimation equation (GEE) model was used to select significant SNPs and adjust for family relatedness [19]. CpG sites were selected by linear mixed model (LMM) with an empirical kinship matrix to adjust for family structure [20]. Both the mixed-effect model and GEE are theoretically suitable for the selection of the SNPs and CpG sites while controlling for family structures. The two methods differ in the way they estimate the coefficients and treat the population correlation structure. The major consideration for us was the ability of software packages to handle a binary phenotype, control family structure, and treat continuous random-effect variables. An arbitrary p value threshold of 10^{-4} was applied to filter the biomarkers for GEE and LMM so that a moderate number of predictors can be used in the prediction model. SNPs were pruned to avoid the strong influence of SNP clusters, by snpgdsLDpruning, and the linkage disequilibrium threshold was set at 0.2 [21, 22]. The empirical kinship matrix was calculated using the pruned SNPs to control for family relatedness. Other clinical variables, including sex, age, and smoking status, were also used as predictors.

Predictors were added into the prediction model step-by-step by data types. Afterward, chosen SNPs were inputted into the ANN first, followed by significant CpG sites. Finally, age, sex, and smoking status were included. This stratified method made it easy to identify the respective contribution of each category of information to prediction.

A three-layer ANN was applied with one hidden layer. The hyperbolic tangent sigmoid transfer function was used as the activation function (a) for the hidden layer, which has the following form:

$$a = transig(n) = -1 + 2/\left(1 + e^{-2n}\right)$$

A linear function was used as the activation function for the output layer (*purelin*):

$$a = purelin(n) = n$$

The learning rate and global momentum were set at 0.01 and 0.4, respectively. The preferred training method was an adaptive gradient descent with momentum. The least mean squares criterion was used to measure the proximity of the neural network prediction to its target when training the ANN.

Results

Contribution of each variable to prediction

Three types of data (SNPs, methylation, and clinical information) were included in the ANN model in a stepwise manner to compare their contributions to the prediction ability of the model. The baseline model simulates the null scenario; that is, 100 SNPs were selected from the autosomes at random and used to predict the phenotype with the ANN in 5-group cross-validation. This gave a baseline error rate of 47.15% (SD: 3.79%), representing a random-guess prediction error under the ANN. Next, including the SNP information yielded a mean test prediction error rate of 43.65% (SD: 4.79%). When methylation information was added, the prediction model achieved an error rate of 41.92% (SD: 4.64%; Wilcoxon rank sum test p value: 0.3759), which implies that the inclusion of methylation information improves the prediction model. When clinical factors (age, sex, smoking status) were also included, the error rate dropped slightly to 41.54% (SD: 5.66%, Wilcoxon rank sum test p value: 0.5) (Table 1). Figure 2 shows the changes of prediction error rate using different variable sets. Sequentially adding SNPs, CpG sites, and environmental factors gradually pushed down the prediction error rate.

Biological function of identified variables

Finally, we report the biological meaning of variables identified using all data. Many of the identified SNP and CpG markers had functions that are related to the regulation of the circulating level of TG, which is a major storage molecule for metabolic energy [23]. To list a few genes (Tables 2 and 3), *FTO* (rs10521308, p value = 9.47E-05) and *CTNNBL1* (rs2206135, p value = 7.75E-05) have both been strongly associated with obesity risk and related traits [24, 25]. The gene *DGAT1* (cg13438334, p value = 8.49E-05) plays a role in catalyzing the committed step in the biosynthesis of TGs [23], and *ALDH4A1* (cg22390041, p value = 4.97E-05) is known to catalyze ester hydrolysis, suggesting that it may lead to a change in the TG level [26].

Table 1 Stratified drug-response prediction model incorporating omics data

	Training error rate ± SD	Test error rate ± SD
SNP	8.59% ± 0.88%	43.65% ± 4.79%
CpG	8.88% ± 2.87%	45.00% ± 3.29%
Add useful CpG information to SNPs	0.00% ± 0.00%	41.92% ± 4.64%
Add useful CpG information to SNPs + age, sex, smoking	0.00% ± 0.00%	41.54% ± 5.66%

The error rates are average 5-fold cross-validation error rates by ANN using inputs

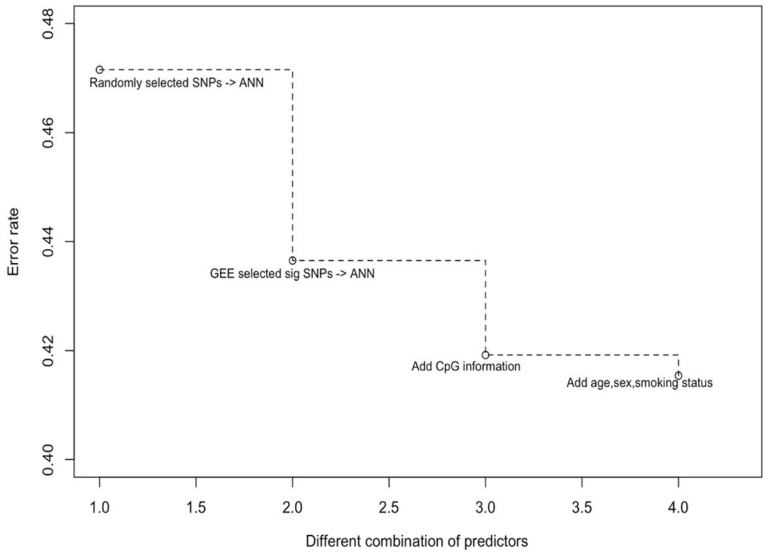

Fig. 2 Stratified drug-response prediction model: the error rate improved when adding additional variables

Discussion

Epigenetic factors are thought to be significantly associated with human diseases, making it plausible to incorporate methylation information for better disease prediction. In this study, we used an ANN to build a stratified drug-response prediction model in which SNPs, methylation, age, sex, and smoking status were considered as predictors. The GAW20 real-data analysis shows that the incorporation of methylation information could reduce the prediction error rate by approximately 4% (p value = 0.3759). The combination of significant SNPs, CpG sites, age, sex, and smoking status achieved the best prediction error rate of 41.54%.

In previous studies, Deng et al. used fusing networks to predict schizophrenia from SNPs, methylation, and functional magnetic resonance imaging data [27]. They achieved a 2.8% increase in prediction accuracy, increasing from 52.9% (using SNPs only) to 55.7% (using SNPs and methylation information). We achieved similar improvement when adding methylation information to SNP. Several reasons may account for the difference between our work and theirs. First, the cell type from which they collected methylation information for prediction is different from the GAW20 data. Methylation varies across cell types, and changes in some cell types are more environment and phenotype specific than in other cell types [4]. The GAW20 real data set methylation information was collected from $CD4^+$ T cells harvested from stored buffy coats, and the phenotype was the TG level in blood, which has a strong correlation with T-cell functions [10]. Second, family relatedness in the GAW20 real data set played a role in the lower prediction error rate. Third, 208 participants (96 cases and 112 health controls) were recruited in the study by Deng et al., whereas our study has a larger sample size of 523 participants. Finally, the method we applied uses a stratified feature selection and prediction approach. The stratification enables better power to selected variables within each stratum, compared to an all-mixture type of prediction modelling, resulting in an enhanced final prediction accuracy.

Conclusions

Adding methylation data slightly improved the prediction accuracy for drug response using a neural network based prediction algorithm with GWAS data. The result could be constraint by the source of tissue, the outcome

Table 2 Selected SNPs that pass the threshold of 10^{-4} in the GEE model

SNP	Chromosome	Gene	Position	p Value	MAF
rs10521308	16	*FTO*	80,459,640	9.47E-05	0.05
rs2206135	20	*CTNNBL1*	35,914,069	7.75E-05	0.42
rs710711	12	*BEST3*	124,093,552	9.98E-05	0.38
rs7096710	10	*C10orf59*	63,063,177	2.92E-05	0.02
rs4851313	2	*CHST10*	100,395,434	5.47E-05	0.44

MAF minor allele frequency

Table 3 Selected CpG sites that pass the threshold of 10^{-4} in the LMM model

CpG sites	Chromosome	Gene	Position	p Value
cg13438334	8	*DGAT1*	145,550,989	8.49E-05
cg11666857	5	*SLC6A19*	1,207,464	2.44E-05
cg22390041	1	*ALDH4A1*	3,036,916	4.97E-05
cg15883716	1	*ANKRD45*	19,226,319	2.06E-06
cg01056590	1	*CABC1*	173,638,701	4.07E-06

variable and the disorder under study. Further studies in other cohorts are necessary to validate the results.

Abbreviations

ANN: Artificial neural network; CpG: Cytosine-phosphate-guanine; GAW: Genetic Analysis Workshop; GEE: Generalized estimation equation; GOLDN: Genetics of Lipid Lowering Drugs and Diet Network; LMM: Linear mixed model; SD: Standard deviation; SNPs: Single-nucleotide polymorphisms; *TG post*: The average of triglyceride levels at visits 3 and 4; *TG pre*: The average of triglyceride levels at visits 1 and 2; TG: Triglyceride

Acknowledgements

We would like to thank the GAW20 for providing XX student travel award to attend the workshop in San Diego, United States; and both reviewers for their constructive comments.

Funding

Publication of this article was supported by NIH R01 GM031575. This work is supported by the National Science Foundation of China [81473035, 31401124 to MHW], and CUHK Direct Grant [4054334].

Availability of data and materials

The data that support the findings of this study are available from the Genetic Analysis Workshop (GAW), but restrictions apply to the availability of these data, which were used under license for the current study. Qualified researchers may request these data directly from GAW.

About this supplement

This article has been published as part of *BMC Genetics* Volume 19 Supplement 1, 2018: Genetic Analysis Workshop 20: envisioning the future of statistical genetics by exploring methods for epigenetic and pharmacogenomic data. The full contents of the supplement are available online at https://bmcgenet.biomedcentral.com/articles/supplements/volume-19-supplement-1.

Authors' contributions

MHW contributed to conception and design; XX performed the analysis and XX and MHW wrote the manuscript. RS, HW, and RM contributed to interpretation of the data. BZ and KCC gave final approval of the version to be published. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Division of Biostatistics, Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, N.T, Hong Kong, SAR, China. [2]CUHK Shenzhen Research Institute, Shenzhen, China.

Published: 17 September 2018

References

1. Flanagan J, Petronis A. Pharmacoepigenetics: from basic epigenetics to therapeutic applications. Drugs Pharm Sci. 2005;156:461.
2. Chaudhry P, Srinivasan R, Patel FD. Utility of gene promoter methylation in prediction of response to platinum-based chemotherapy in epithelial ovarian cancer (EOC). Cancer Investig. 2009;27(8):877–84.
3. Shindo T, Shimizu T, Nishiyama N, Niinuma T, Kitajima H, Kai M, Shinkai N, Itoh N, Tanaka T, Suzuki H, et al. Diagnosis and prediction of recurrent bladder cancer by urinary DNA methylation analysis: multicenter prospective study. Eur Urol Suppl. 2017;16(3):e206–8.
4. Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet. 2012;13(10):679–92.
5. Bullinger L, Ehrich M, Döhner K, Schlenk RF, Döhner H, Nelson MR, van den Boom D. Quantitative DNA methylation predicts survival in adult acute myeloid leukemia. Blood. 2010;115(3):636–42.
6. Cole JH, Ritchie SJ, Bastin ME, Hernández MV, Maniega SM, Royle N, Corley J, Pattie A, Harris SE, Zhang Q, et al. Brain age predicts mortality. Mol Psychiatry. 2017:1–8.
7. Howsmon DP, Kruger U, Melnyk S, James SJ, Hahn J. Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation. PLoS Comput Biol. 2017;13(3):e1005385.
8. Liu S, Chen X, Chen R, Wang J, Zhu G, Jiang J, Wang H, Duan S, Huang J. Diagnostic role of Wnt pathway gene promoter methylation in non-small cell lung cancer. Oncotarget. 2017;8(22):36354–67.
9. Peters I, Reese C, Dubrowinskaja N, Antonopoulos WI, Krause M, Dang TN, Grote A, Becker A, Hennenlotter J, Stenzl A, et al. DNA methylation signature for the assessment of metastatic risk in primary renal cell cancer. J Clin Oncol. 2017;35(6 Suppl):516.
10. Irvin MR, Zhi D, Joehanes R, Mendelson M, Aslibekyan S, Claas SA, Thibeault KS, Patel N, Day K, Jones LW, et al.: Epigenome-wide association study of fasting blood lipids in the genetics of lipid lowering drugs and diet network study. Circulation 2014; 130(7):565–572.
11. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. Brief Bioinform. 2013;15(6):929–41.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
13. Balfour JA, McTavish D, Heel RC. Fenofibrate. Drugs. 1990;40(2):260–90.
14. Wang MH, Chang B, Sun R, Hu I, Xia X, Wu WK, Chong KC, Zee BC. Stratified polygenic risk prediction model with application to CAGI bipolar disorder sequencing data. Hum Mutat. 2017;38(9):1235–9.
15. Castejón-Limas M, Ordieres Meré J, Vergara EP, Martínez-de-Pisón FJ, Pernía AV, Alba F. The AMORE package: a MORE flexible neural network package. Published April. 1014:14. Available at https://cran.r-project.org/web/packages/AMORE/index.html
16. Cheng B, Titterington DM. Neural networks: a review from a statistical perspective. Stat Sci. 1994;9(1):2–30.
17. Hastie T, Tibshirani R, Friedman J: The elements of statistical learning: data mining, inference, and prediction. Springer New York 2017. Chapter: Neural Networks.
18. Donaldson RG, Kamstra M. Neural network forecast combining with interaction effects. J Frankl Inst. 1999;336(2):227–36.
19. Chen MH, Yang Q. GWAF: an R package for genome-wide association analyses with family data. Bioinformatics. 2009;26(4):580–1.
20. Therneau TM, Therneau MT. Package "coxme". Mixed effects cox models. R package version. 2015:2. Available at https://cran.r-project.org/web/packages/coxme/coxme.pdf
21. Zheng X, Zheng MX. Package 'SNPRelate'. 2013. Available at ftp://gnu.cs.pu.edu.tw/network/CRAN/web/packages/SNPRelate/SNPRelate.pdf
22. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326–8.
23. Yen CL, Stone SJ, Koliwad S, Harris C, Farese RV. Thematic review series: glycerolipids. DGAT enzymes and triacylglycerol biosynthesis. J Lipid Res. 2008;49(11):2283–301.
24. Yoganathan P, Karunakaran S, Ho MM, Clee SM. Nutritional regulation of genome-wide association obesity genes in a tissue-dependent manner. Nutr Metab (Lond). 2012;9(1):65.
25. Liu YJ, Liu XG, Wang L, Dina C, Yan H, Liu JF, Levy S, Papasian CJ, Drees BM, Hamilton JJ, et al.: Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. Hum Mol Genet 2008;17(12):1803–1813.
26. Vasiliou V, Nebert DW. Analysis and update of the human aldehyde dehydrogenase (ALDH) gene family. Hum Genomics. 2005;2(2):138.
27. Deng S-P, Lin D, Calhoun VD, Wang Y-P. Predicting schizophrenia by fusing networks from SNPs, DNA methylation and fMRI data. Conf Proc IEEE Eng Med Biol Soc. 2016;2016:1447–50.