Genomics 113 (2021) 1018–1028

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Predicting DNA methylation from genetic data lacking racial diversity using shared classified random effects

J. Sunil Rao [a,*,1], Hang Zhang [a,1], Erin Kobetz [a], Melinda C. Aldrich [b], Douglas Conway [b]

[a] University of Miami, FL, United States of America
[b] Vanderbilt University Medical Center, Nashville, TN, United States of America

ARTICLE INFO

Keywords:
DNA methylation
Prediction
Mixed effects models
Racial diversity

ABSTRACT

Public genomic repositories are notoriously lacking in racially and ethnically diverse samples. This limits the reaches of exploration and has in fact been one of the driving factors for the initiation of the All of Us project. Our particular focus here is to provide a model-based framework for accurately predicting DNA methylation from genetic data using racially sparse public repository data. Epigenetic alterations are of great interest in cancer research but public repository data is limited in the information it provides. However, genetic data is more plentiful. Our phenotype of interest is cervical cancer in The Cancer Genome Atlas (TCGA) repository. Being able to generate such predictions would nicely complement other work that has generated gene-level predictions of gene expression for normal samples.

We develop a new prediction approach which uses shared random effects from a nested error mixed effects regression model. The sharing of random effects allows borrowing of strength across racial groups greatly improving predictive accuracy. Additionally, we show how to further borrow strength by combining data from different cancers in TCGA even though the focus of our predictions is DNA methylation in cervical cancer. We compare our methodology against other popular approaches including the elastic net shrinkage estimator and random forest prediction. Results are very encouraging with the shared classified random effects approach uniformly producing more accurate predictions – overall and for each racial group.

1. Background

Epigenetic markers are more and more of a focus of cancer researchers due to the fact that they form a bridge between the environment and the biology of a tumor. Hypermethylation of CpG islands are well known to result in decreased gene expression of nearby genes. In cervical cancer for instance, a number of tumor suppressor genes have been found to have undergone hypermethylation including those involved in apoptosis, cell-cycle, WNT-pathway, DNA repair, the FA-BRAC pathway, mismatch repair, metastasis and cell death and cell differentiation [5]. Additionally, African Americans have a higher incidence and worse survival and little is known about their differential methylation patterns. Since methylation is modifiable, understanding this might lead to new ways to mitigate disparities.

While the amount of epigenetic data is increasing all the time, it's still not as ubiquitously found as other genomic profiles. So it might be very useful if one could predict methylation values using the mountains of cancer-related genomic data that does exist today in the form of genotyping data. In fact, this type of thing was done by [7] who generated a gene expression predictive linear model based on SNP profiles from normal tissues using the GTEx public repository [16]. They focused on an ANOVA-type decomposition of total expression into a genetically controlled portion, a portion determined by phenotype and a portion determined by environmental and other factors. Thus their predictions were interrogating the genetically controlled portion of gene expression. Their decomposition, if applied directly to DNA methylation (DNAm), was such that DNAm $\rightarrow w_1$gDNAm $+ w_2$pDNAm $+ w_3$eDNAm, with $\sum_{j=1}^{3} w_j = 1$. The addition of covariates beyond SNPs (e.g. clinical variables) to the model would yield predictions aimed at estimating DNAm itself. However, it is very likely that these models will still be under specified (biased) and would not be capturing enough of the variation in DNAm.

So our goal is to generate highly accurate predictions of DNAm from genetic data and other covariates. We will use a new mixed model based

* Corresponding author.
E-mail address: jrao@miami.edu (J.S. Rao).
[1] Joint first authors.

https://doi.org/10.1016/j.ygeno.2020.10.036
Received 21 December 2019; Received in revised form 24 September 2020; Accepted 30 October 2020
Available online 5 November 2020
0888-7543/© 2020 Elsevier Inc. All rights reserved.

methodology to help *capture the unmeasured* sources of variation in DNAm. But our task will be doubly difficult because we also are interested in predicting race-specific DNAm and will have to address the fact that genomic data repositories notoriously lack racial and ethnic diversity. In fact, this was one of the impetuses behind NIH's new All of Us project [19] - to generate a public repository of genomic profiling data that was more representative of the general population. A clustering step of DNAm executed prior to fitting our mixed models which results in clusters with racial diversity will allow for borrowing strength across races to improve race-specific predictions.

2. Results

The TCGA Cervical and endocervical cancers (CESC) raw data contained 305 patients. The data was comprised of 482,421 methylation features (450 K platform) and 22,618 gene features with somatic copy number alterations (sCNA). The CESC methylation data and sCNA data along with clinical data which includes demographic and staging information were downloaded using R package TCGA2STAT [26].

2.1. Univariate filtering of CESC samples

Patients with race other than White or Black were excluded, which reduced sample size to 230, among which 28 (12.2%) were Black and 202 (87.8%) were White. Methylation based on raw Beta values were filtered by imposing a cutoff of 0.25 in terms of empirical standard deviation to select probes with high variability, resulting in 2472 most variable features selected. The cutoff was determined based on previous studies where [28] used a cutoff of 0.2 and [15] who used a cutoff of 0.3 in order to filter the most variable CpGs with other constraints. In this study we chose a cuttoff of 0.25. Hypothesis testing using false detection rate (FDR) control was entertained but not used because of the extreme conservativeness of the procedure with such a large number of tests to control [9]. Additionally, 70 further methylation variables were removed due to missingness resulting in 2402 methylation responses used in the analyses. The genetic data obtained from TCGA2STAT was aggregated gene level somatic copy number alterations (sCNAs) profiling from Affymetrix Genome-Wide Human SNP Array 6.0 [26], therefore genes that harboring somatic mutations were used for filtering and building statistical models as input features. The sCNAs were filtered to only including the top 0.5-percentile in terms of variance, which resulted in 114 features selected. Due to the existence of high correlation among some of the genes which would cause model singularity, one gene in any pair with a correlation greater than 0.7 were removed. The final number of sCNAs included in the analyses was 61.

2.2. Clinical variables

Clinical variables included in the analysis were *age*; *gender*; *race*; and *Stage*. The *Stage* variable had to be manually created as described in the next section. *Age* was included as continuous variable, while *gender*; *race*; *Stage* were categorical variables. Male was used as baseline level for *Gender*. Patients with race other than White and Black were excluded before filtering, as described earlier, in the analysis Black was used as baseline for *race*.

2.2.1. Defining the stage variable

Pathologic stage is an important factor in referring to cancer progression and is often grouped from I to IV with increasing severity of disease. Usually pathologic stage is determined by TNM system (tumor T, regional nodes N, and metastases M) by classifying patients with similar prognoses [6]. Table 1 below shows that all staging categories (subcategories) are absent for CESC.

The classification and definition of TNM system differs by cancer types, which in turn suggests that pathologic staging should modeled taking into account specific cancer type. AJCC provides rules for defining staging by specific cancer type using the TNM system [6] and for CESC, the information for can be found at (https://cancerstaging.org/references-tools/quickreferences/Documents/CervixMedium.pdf). According to these rules and with the available TNM variables: *pathologyTstage*, *pathologyNstage*, and *pathologyMstage*; *Stage* can be determined. We first defined each sub-category, and then combined the subcategories to form stage I to IV(1 to 4) plus an additional level, *stage.NA* since there are some patients who cannot be grouped into a stage due to limited TNM information. The distribution of the new *Stage* variable is displayed below in Table 2. Tables 3 and 4 show the original and new Stage variable by cancer type when integrating LUAD samples.

2.3. The proposed method

Suppose that we have a set of training data, y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ in the sense that their classifications are known, that is, one knows which group, i, that y_{ij} belongs to. Let y_{ij} be the measured DNAm from a training dataset of size N where $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$ with $\sum_i n_i = N$. So this represents clustered DNAm data in m distinct groups. Accompanying each y_{ij} measurement is a p-vector of individual level covariates x_{ij}.

The assumed linear mixed model (LMM) for the training data is

$$y_i = X_i\beta + Z_i\alpha_i + \varepsilon_i, \tag{1}$$

where $y_i = (y_{ij})_{1\le j\le n_i}$, $X_i = (x_{ij}')_{1\le j\le n_i}$ is a matrix of known covariates, β is a vector of unknown regression coefficients (the fixed effects), Z_i is a known $n_i \times q$ matrix, α_i is a $q \times 1$ vector of group-specific random effects, and ε_i is an $n_i \times 1$ vector of errors. It is assumed that the α_i's and ε_i's are independent, with $\alpha_i \sim N(0, G)$ and $\varepsilon_i \sim N(0, R_i)$, where the covariance matrices G and R_i depend on a vector ψ of variance components.

Our goal is to make a classified prediction for a mixed effect associated with a set of new observations, $y_{n,j}$, $1 \le j \le n_{new}$ (the subscript n refers to "new"). Suppose that

$$y_{n,j} = x_n'\beta + z_n'\alpha_I + \varepsilon_{n,j}, \qquad 1 \le j \le n_{new}, \tag{2}$$

where x_n, z_n are known vectors, $I \in \{1, \ldots, m\}$ but one does not know which element i, $1 \le i \le m$, is equal to I. Furthermore, $\varepsilon_{n,j}$, $1 \le j \le n_{new}$ are new errors that are independent with $E(\varepsilon_{n,j}) = 0$ and $var(\varepsilon_{n,j}) = R_{new}$, and are independent with the α_i's and ε_i's. Note that the normality assumption is not always needed for the new errors. Also, the variance R_{new} of the new errors does not have to be the same as the variance of ε_{ij}, the jth component of ε_i associated with the training data. The mixed effect that we wish to predict is

$$\theta = E(y_{n,j}|\alpha_I) = x_n'\beta + z_n'\alpha_I. \tag{3}$$

Table 2
Manual classification of CESC patients into stages.

	Stage.1	Stage.2	Stage.3	Stage.4	Stage.NA	NA
CESC	44	14	15	4	153	0
NA.	0	0	0	0	0	0

Table 1
Tabulation of Pathologic stage for CESC samples.

	Stage i	Stage ia	Stage ib	Stage ii	Stage iia	Stage iib	Stage iiia	Stage iiib	Stage iv	NA
CESC	0	0	0	0	0	0	0	0	0	230

Table 3
Tabulation of pathologic stage by cancer type.

	Stage i	Stage ia	Stage ib	Stage ii	Stage iia	Stage iib	Stage iiia	Stage iiib	Stage iv	NA
CESC	0	0	0	0	0	0	0	0	0	230
LUAD	5	115	101	1	47	51	54	7	15	5
NA.	0	0	0	0	0	0	0	0	0	0

Table 4
Tabulation of revised staging variable by cancer type.

	Stage.1	Stage.2	Stage.3	Stage.4	Stage.NA	NA
CESC	44	14	15	4	153	0
LUAD	221	99	61	15	5	0
NA.	0	0	0	0	0	0

In our case, x will correspond to sCNAs, race, clinical variables and other covariates. Parameter estimation is done via ML or REML for instance although other estimators can be used - for example the best predictive estimator [13,23]. We will use the method called classified mixed model prediction (CMMP) to generate our desired predictions [14]. This method identifies a best group in the training data from which to attach that group's estimated random effect to the new observations resulting in $\alpha_{\widehat{I}}$.

Suppose that $I = i$. Then, the vectors $y_1, \ldots, y_{i-1}, (y_i', \theta)', y_{i+1}, \ldots, y_m$ are independent. Thus, we have $E(\theta|\, y_1, \ldots, y_m) = E(\theta|\, y_i)$. By the normal theory, we have

$$E(\theta|y_i) \quad = \quad x_n'\beta + z_n' G Z_i' \left(R_i + Z_i G Z_i'\right)^{-1}(y_i - X_i\beta) \tag{4}$$

In practice, however, I is unknown and treated as a parameter. In order to identify, or estimate, I, we consider the mean squared prediction error (MSPE) of θ by the BP when I is classified as i, that is $\mathrm{MSPE}_i = E\left\{\tilde{\theta}_{(i)} - \theta\right\}^2 = E\left\{\tilde{\theta}_{(i)}^2\right\} - 2E\left\{\tilde{\theta}_{(i)}\theta\right\} + E(\theta^2)$. Using the expression $\theta = \bar{y}_n - \bar{\varepsilon}_n$, where $\bar{y}_n = n_{\mathrm{new}}^{-1}\sum_{j=1}^{n_{\mathrm{new}}} y_{n,j}$ and $\bar{\varepsilon}_n$ is defined similarly, we have $E\left\{\tilde{\theta}_{(i)}\theta\right\} = E\left\{\tilde{\theta}_{(i)}\bar{y}_n\right\} - E\left\{\tilde{\theta}_{(i)}\bar{\varepsilon}_n\right\} = E\left\{\tilde{\theta}_{(i)}\bar{y}_n\right\}$. Thus, we have the expression:

$$\mathrm{MSPE}_i = E\left\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}\bar{y}_n + \theta^2\right\} \tag{5}$$

It follows that the observed MSPE corresponding to (5) is the expression inside the expectation. Therefore, a natural idea is to identify I as the index i that minimizes the observed MSPE. Because θ^2 does not depend on i, the minimizer is given by

$$\widehat{I} = \mathrm{argmin}_i\left\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}\bar{y}_n\right\} \tag{6}$$

The classified mixed-effect predictor (CMEP) of θ is then given by $\widehat{\theta} = \tilde{\theta}_{(\widehat{I})}$.

In the case say of a random intercept mixed model, if $G > R$, CMMP can generate much more accurate predictions than usual regression prediction which ignores the random effect (more precisely, it plugs in the population average value for the random effect which is zero). The CMMP prediction then becomes $x_n\widehat{\beta} + \alpha_{\widehat{I}}$. Looking more carefully at (1), we can see that the model is decomposing DNAm into a component explained by x and another component that captures the group-specific differences in DNAm. These m groups may in fact be of mixed races and v_i is capturing this. It is this feature that allows us to borrow strength across races when making race-specific DNAm predictions.

Model (1) is not however known apriori to us because the groupings are unknown. Thus we will have to first estimate these groupings.

2.3.1. Clustering CESC DNAm profiles

Clustering was performed on methylation (2402 CpGs) in the training dataset based on *pam* algorithm and optimal cluster number was determined by Gap statistic method [24] using R package cluster [17]. 5 clusters were selected and the groupid variable, indicating cluster membership, was created. The distribution of individuals by race across clusters is shown in Fig. 1. Also, other clustering methods including K-means, affinity propagation and hierarchical clustering were explored, results were similar (see supplement Figs. S1–S12).

2.3.2. Prediction results

We transformed methylation values using the M value transformation [4] and RAU value transformation [21] as below. This was done to better satisfy normality model assumptions for the mixed model framework or plausible support values for the other competing methods (described below).

$$M_i = log_2\left(\frac{Beta_i}{1 - Beta_i}\right)$$

$$RAU_i = 2(146/\pi)arcsin\left(Beta_i^{0.5}\right) - 23$$

We split the dataset into training and test partitions of 70% and 30%, respectively. Covariates included were sCNAs and clinical variables. We compared a number of different methods in terms of test set prediction accuracy based on empirical mean square prediction error (MSPE). These methods included CMMP (using group-specific random intercepts), usual linear regression model (LM) prediction and a high dimensional shrinkage estimator based on the elastic net (ENET) [29] and random forest (RF) prediction [2]. For CMMP, LM and RF, we used the selected 61 sCNAs as described earlier. For ENET, the same sCNAs were entered in the model but we went a step further. We also entertained the possibility of sCNA by race interactions. If any of these were found to be important, they would also be entered into the CMMP and LM. We also tried to look at the combination of ENET and CMMP, this was done by applying CMMP on the residuals after fitting ENET model. And then predictions combining fixed part (ENET) and the random part (CMMP) were obtained. This process was done for each of the 2402 methylation responses. For CMMP-based models, clustering of DNAm training samples was done as described above. The whole analysis was repeated for 20 training-test splits and empirical mean test MSPEs reported together with their empirical standard deviations (SD).

Figs. 2 and 3 show results of model fitting. Fig. 2 are so-called violin plots which show average test set empirical MSPE values for all 2402 methylation responses on the same plot - by race and overall. Overlaid on each violin plot are standard error (SE) bars. It is quite clear that CMMP marginally improves over other models both overall and by racial group. ENET models provide slightly lower MSPE values compared to the other methods except CMMP and that adding CMMP on top of ENET does not further improve things. Fig. 3 shows those average ENET coefficient values different than zero overall all 2402 models. The top part of the plot shows main effects and the shaded bottom part of the plot shows sCNA by race and Stage by race interactions. Among the predictive sCNAs, none seem to stand out in terms of relative size. Also, although visually looking on the same scale at the sCNA variables, Stage 4 of the stage variable seems to have been selected. This may be illustrating the association might be strongest with the most serious prognosis. The clinical variable indexing Stage 4 cancers was the most important among the clinical variables. It is also interesting to see that

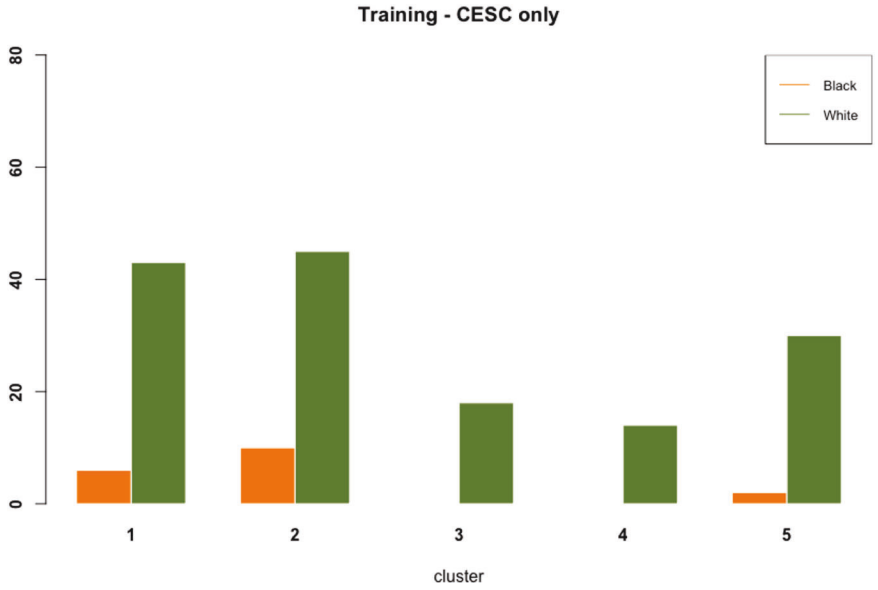

Fig. 1. Clustering of CESC samples for one training set.

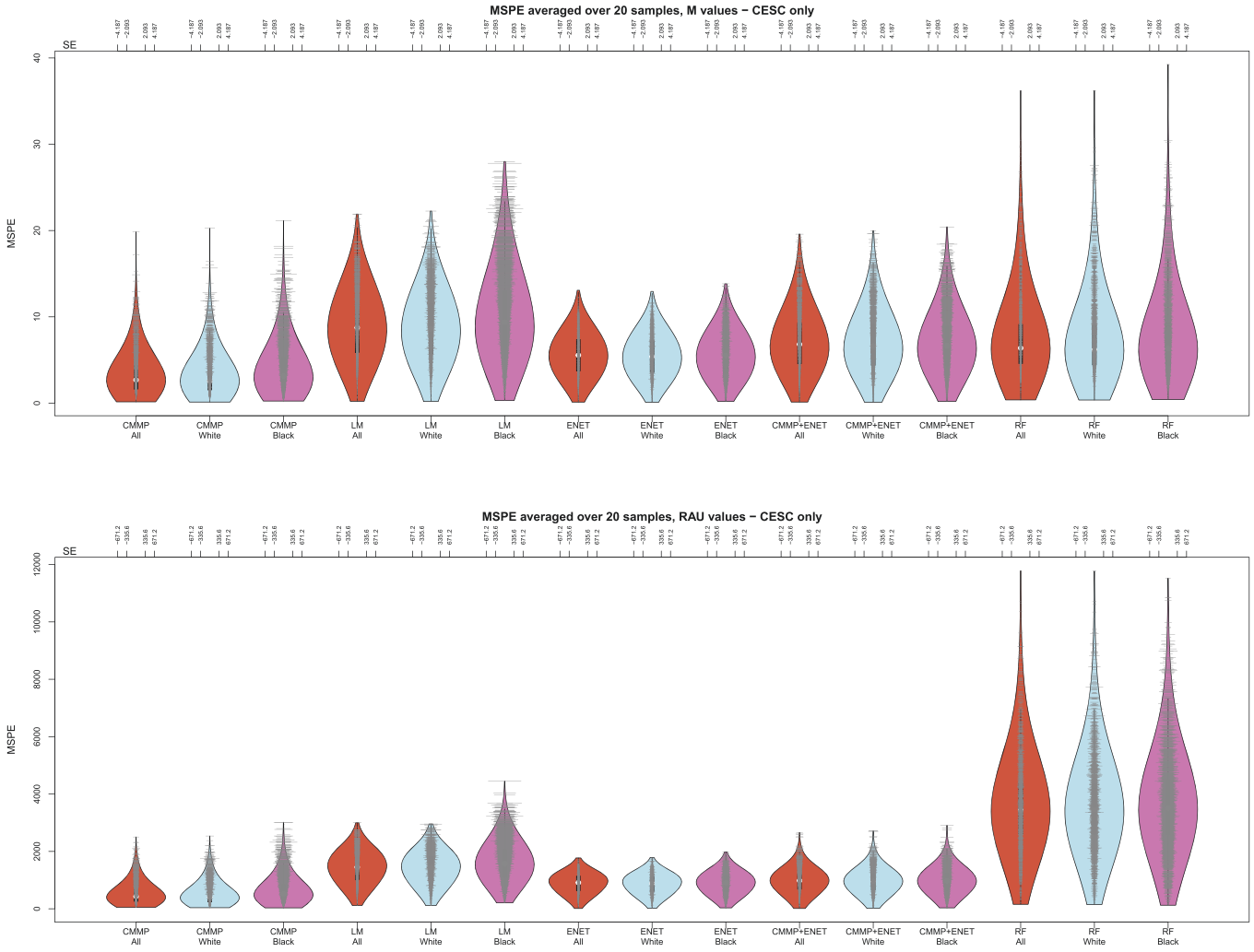

Fig. 2. Average empirical test set MSPE Violin plots with individual methylation response standard error (SE) bars overlayed. Only CESC training samples used. Top plot is the M transformation and bottom plot is the RAU transformation.

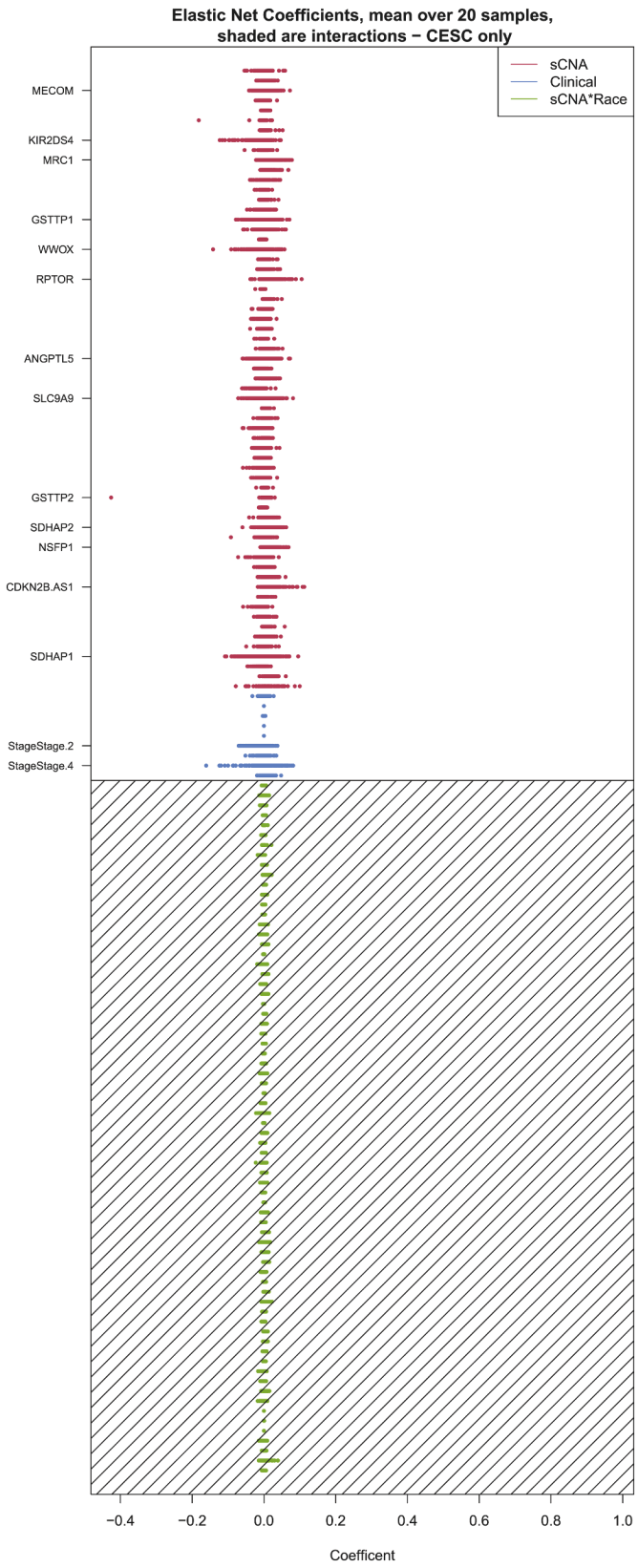

Fig. 3. Elastic net empirical average coefficient plot. Only CESC training samples used.

this plot indicates that interactions with race (right side of plot in shaded region) were not different than zero.

2.4. Combining cancer types – further opportunities to borrow strength

Model (1) can be further enhanced by including additional cancers into the analysis and then adding a fixed effect covariate for cancer type in the model. The rationale for combining cancer types is to increase data heterogeneity to a certain point while controlling for clustering. This could benefit the fit of CMMP and potentially increase the predictive ability of the analysis since we also gain additional patients this way. As mentioned in Section 2.3, as long as we can generate a clustering of DNAm where G dominates R even more than it did with CESC data alone and if we have enough diversity of cancer types within the clusters, then including additional cancers in a combined model can further improve predictions. This is also why the selection of cancers is very important, if we chose cancers that are very distinct then the clustering will separate them out and we lose the advantage of information sharing within cluster. We want to choose combinations that will create clusters that have representation of at least 2 cancer types. Also it is important to provide a point of reference as to how much the performance increases under information sharing, this is why we presented the single cancer type (CESC) model. To deal with these considerations we used a hybrid goodness of fit measure for clustering.

The first measure is the Gini statistic calculation from [11]:

$$\text{Gini}_{\max} = \underset{i \in k}{max} \sum_{j=1}^{C} \frac{n_{ij}}{n_i} \left\{ 1 - \frac{n_{ij}}{n_i} \right\},$$

where k is the number of clusters, C is the number of cancer types included, n_i is the total number of observations in cluster i, n_ij is the number of observations in cancer type j in cluster i.

The second measure is the Gap statistic from [24]:

$$\text{Gap}(k) = \mathbb{E}_n^* \{ log(W_k)) \} - log(W_k)$$

The number of clusters k is then selected via the maxSE method in [17]:

$$k_{\text{optimal}} = maxk \in K \,\text{s.t}\, \text{Gap}(k_{\max}) - \text{Gap}(k) < \text{Gap\_SE}(k_{\max})$$

where k is the number of clusters, K is the maximal number of clusters tested, $\mathbb{E}_n^*$ is the expected value taking a sample size of n, and W_k is the within cluster sum of squares pooled between all clusters, Gap_SE(k) is the standard error of the bootstrap Gap values calculated, and $k_{\max}$ is the k value that gives the global maximal Gap statistic.

Combining the two measures together allows us to identify the optimal number of clusters with sufficient cancer type diversity:

$$\text{Gap} + \text{Gini} = \text{Gini}_{\max} + \text{Gap}\left(k_{\text{optimal}}\right)$$

Fig. 4 plots this convolution measure against different combinations of cancers from TCGA (each downloaded separately and then merged with other cancers). The plot indicates that combining CESC with lung adenocarcinoma (LUAD) results in an increase in the Gap+Gini. Adding additional cancers does not improve fit much more. Thus we focused on the CESC, LUAD combination.

The TCGA Lung adenocarcinoma (LUAD) raw data contained 569 patients and removing all patients who were not Black or White further reduces the sample size to 401. The data was prepared by first combining the 450 K methylation, sCNAs, and clinical data for LUAD and CESC independently. Then the two datasets were merged by stacking and matching the corresponding variables, along with adding an indicator variable for cancer type. Once all merging was complete, there were 631 patients available of which 230 had CESC (36.5%) and 401 had LUAD (63.5%) and 78 were black or African American (12.4%) and 553 were white (87.6%). Fig. 5 summarizes the 9 clusters that were found (by the Gap statistic) and gives the racial composition of each.

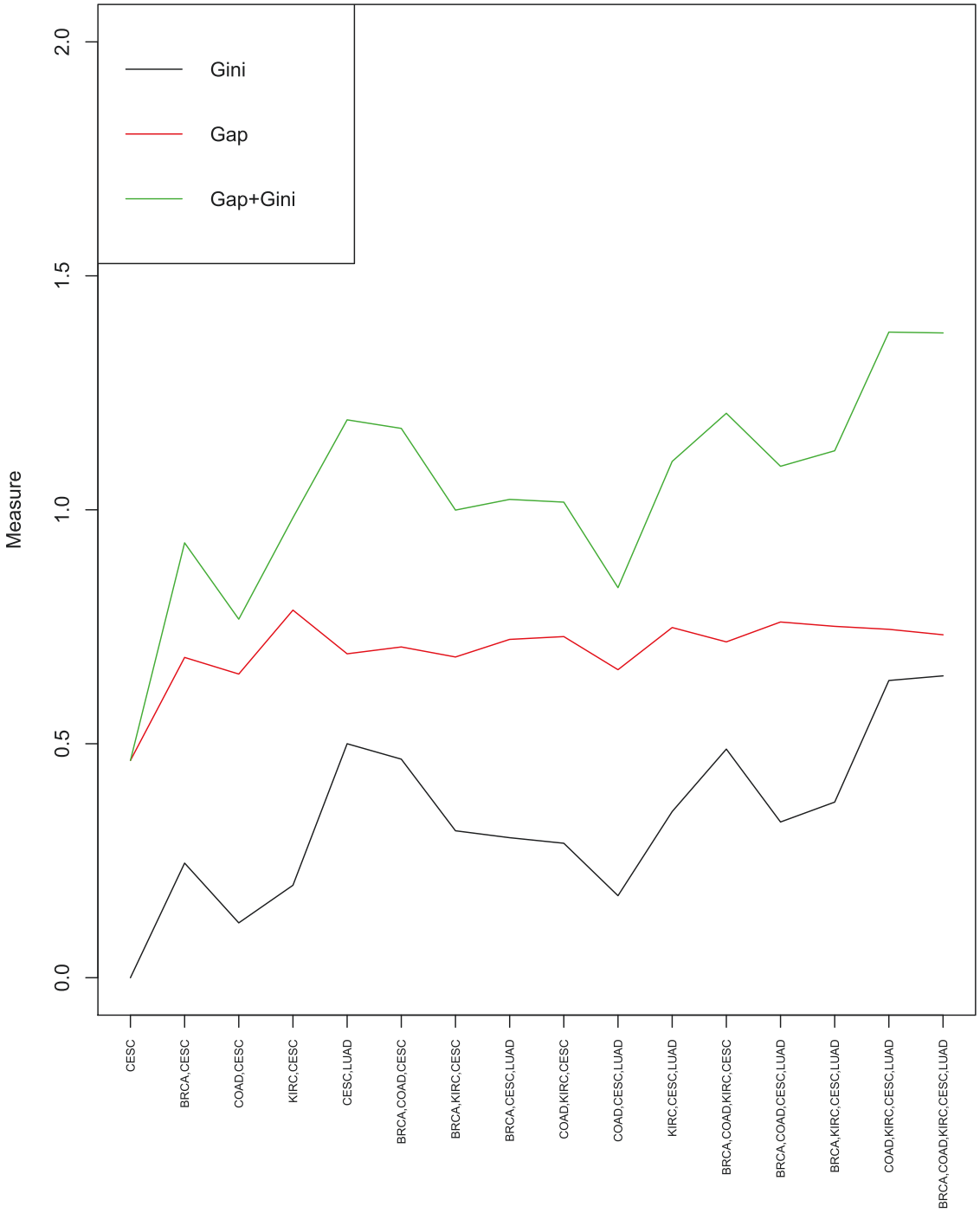

Fig. 4. Combining cancer types goodness of fit measure.

This is important since our goal is the borrowing of information between races; therefore, it is imperative that we have a mixture of representation of each race in our clusters.

Raw staging information in TCGA looked as follows:

Again we recreated the *Stage* variable for CESC as described earlier while this variable was already available for LUAD. Then for both CESC and LUAD, we combined each sub-categories to form include 1 to 4 stages and a stage.NA. The distribution of the new *Stage* variable by cancer type is displayed below.

2.4.1. Prediction performance from combined model

We used the exact same filtered sCNAs and methylation as responses from the CESC only analysis above. While this is clearly an unorthodox thing to do typically, for the purposes of this paper it makes sense. Doing so, facilitates focus on the borrowing strength of our methodology across racial groups. We will discuss this issue in more detail in the Discussion section. Once again, we split the dataset into training and test partitions of 70% and 30% respectively. Covariates included were sCNAs and clinical variables with an additional variable indicating cancer type. The analysis procedure was similar to what we described in Section 2.3.2 except that we also entertained the possibility of cancer type by stage interactions. The whole analysis was repeated for 20 training-test splits and empirical mean MSPEs *for CESC samples only* reported together with their SEs. This emphasizes the fact that CESC DNAm prediction is our

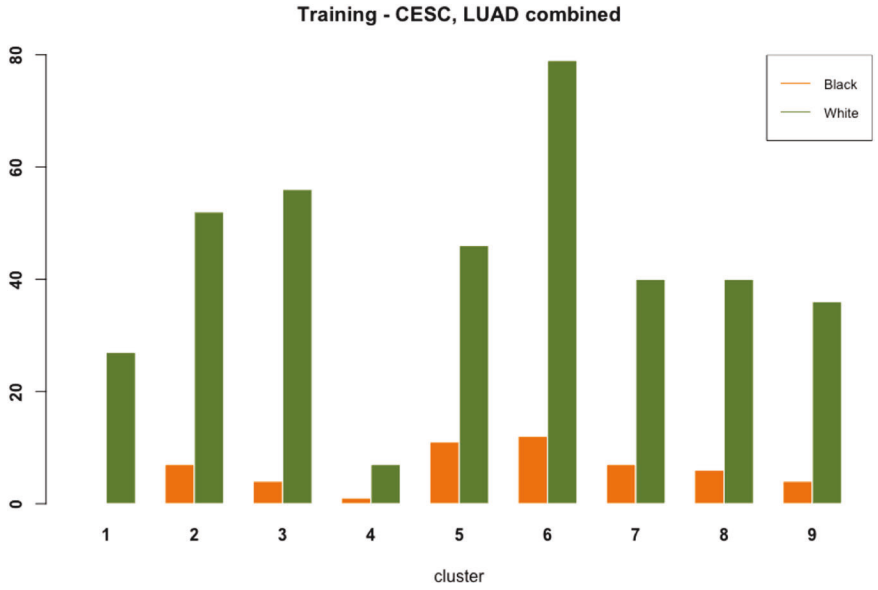

Fig. 5. Clustering for combined CESC and LUAD samples for one training set.

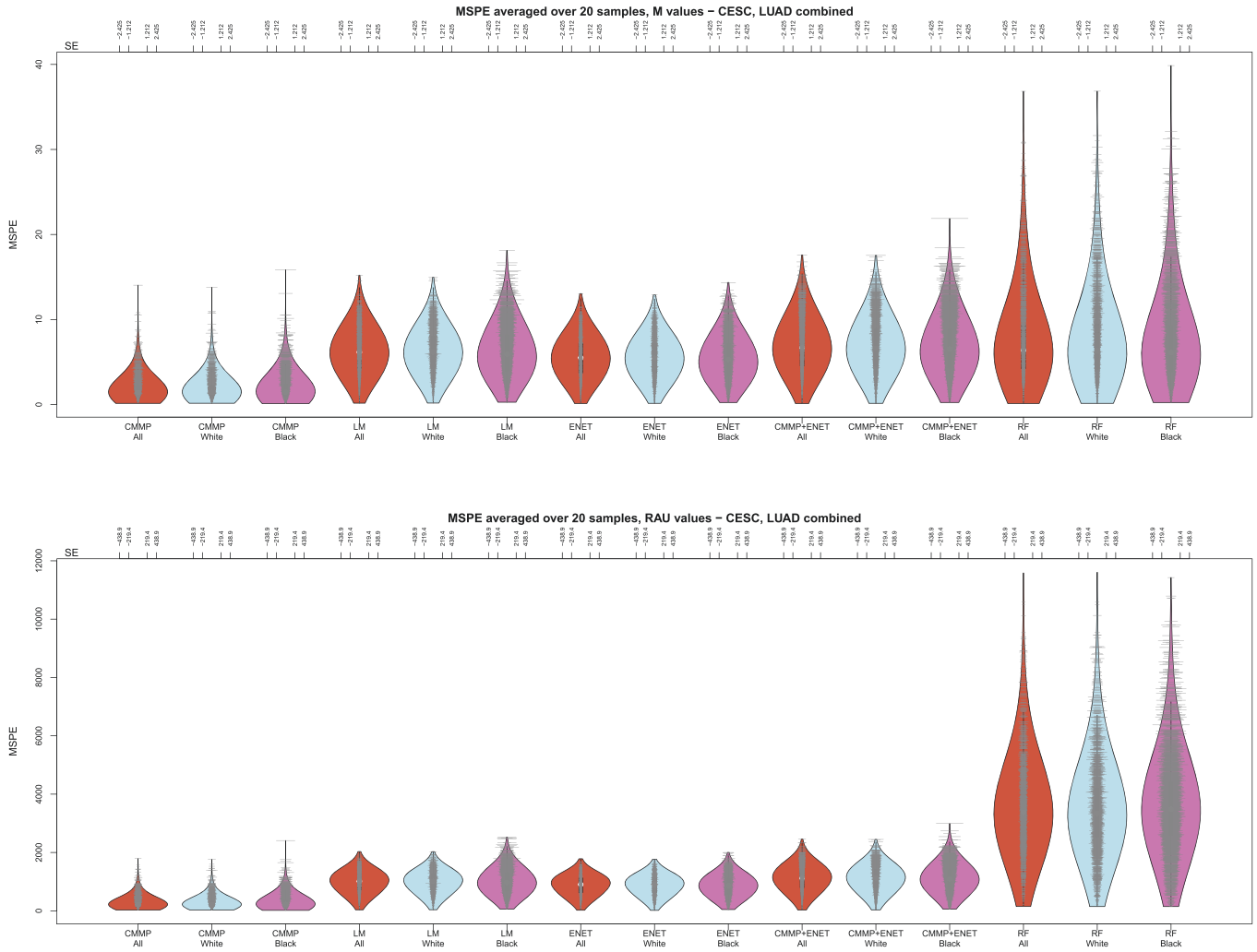

Fig. 6. Average empirical test set MSPE Violin plot with individual methylation response standard error (SE) bars overlayed for M and RAU transformations.

interest and this way, the combined model results could be directly compared against the CESC only model.

Fig. 6 shows MSPE performance across all methods by race and combined across racial groups. Plotted are violin plots for all 2402 methylation responses with SEs overlayed. What is evident is that not only are median MSPE values systematically lower (by race and overall) for CMMP, but that the distribution of values is much more skewed towards small values as compared to the other methods. What is interesting is that ENET did not seem to provide much improvement over LM with univariate sCNA filtering. Finally, we also tried ENET combined with CMMP and noted no further improvements. This is due to the fact that the shrinkage of fixed effect estimates by ENET (induced bias) makes group identification by CMMP much less effective.

2.5. An interesting robustness for using classified random effects

While much attention in the analysis of DNA methylation data has focused around what transformation would be most appropriate [4,22] or perhaps using alternate regression modeling strategies like beta regression [25,27], the clustered data mixed model that we use ends up having a remarkable robustness that we can take advantage of. As recognized by McCullogh and Neuhaus [18] some years ago, (generalized) linear mixed models can be quite robust to departures from normality in a number of ways. In particular estimates of covariate effects (both within group and between group) and prediction of random effects are quite well estimated even with large departures from normality. Only intercept estimates show significant bias. So this would mean that our classified random effects prediction approach could also be robust from such distributional departures.

So we put this fact to the test and modeled the raw beta values of methylation as a response. In supplement (Fig. S13) we have shown a same but smaller scale analysis on the 36 most variable CpGs where we generated Q-Q plots against normality for all 36 methylation responses on the raw scale, the M value scale and the RAU scale. As expected, the Beta values (raw) are quite far from normality. The transformed values improve things somewhat with the M values producing a more palatable transformation. We then fit and tested linear mixed models using our CMMP approach on the raw Beta values. In the supplement (Fig. S14) we also show boxplots of test set predictions from the CMMP, LM, ENET, ENET+CMMP and RF models, respectively (for a particular split of the data). What's very clear is that there is a maintenance of coherence for ENET models in that the predicted test set values all fall between $[0,1]$ which is the range of raw Beta values. This has to do with the amount of shrinkage to zero in the model parameter estimates which severely constrains model predictions. LM model on the other hand produces many instances where relatively large percentages of the empirical distributions fall outside of $[0,1]$ and hence demonstrate a lack of coherency. This is to be expected. On the other hand, we confirm the McCullogh and Neuhaus [18] observations and find that when extended to to CMMP, produces mostly coherent predictions on the raw scale.

Fig. 7 shows the resulting average empirical MSPE violin plot across methods using the raw Beta values. It's probably a bit unfair to plot LM as part of this plot due to the lack of coherency of this model but what is clear is we observe something similar to when working with M and RAU transformed responses. That is, CMMP produces markedly lower MSPEs as compared to other models. Since the raw Beta values are often much more interpretable biologically [4], this may suggest that one could still operate on this scale and produce very accurate and coherent test set predictions.

Ridgeline plots depicting the empirical MSPE distributions over the 100 simulation runs for each 36 methylation response were also shown in the supplement (Fig. S15). The methylation response variables are named and the actual distribution shapes can be nicely visualized. Once again, it is clearly evident that the CMMP method produces systematically lower empirical MSPE values than the other methods - again by race and overall.

We looked more closely at ENET coefficients themselves. Fig. 8 plots each variable along the y-axis and their average estimated coefficients (over the 20 training-test split runs) for all 2402 methylation responses on the x-axis. Dots are colored by covariate type. On the bottom part of the plot in the shaded area are the interaction effects that were examined. Now the important sCNAs look less impactful but the clinical variable for cancer type in the combined model is highly dominant. Again, none of the interactions with race had an impact. The lack of race-by-gene interactions was not unexpected as this has also been confirmed in other cancers like lung cancer [1,20].

A karyoplot is provided for visualizing positions of filtered genes with sCNAs on whole human chromosomes (See Fig. 9) which was done by the aid of R package karyoploteR [8]. Red colored genes are the one that showed up as important in the ENET model. It's of immediate interest to note that 4 of the 5 appear on chromosome 3. Of the red colored genes, MECOM on chromosome 3 is also known as histone-lysine *N*-methyltransferase. It is an oncogene which has a role in development, cell proliferation and differentiation. It is also one of the primary histone methyltransferases that direct cytoplasmic H3K9me1 methylation (https://www.genecards.org/cgi-bin/carddisp.pl?gene=MECOM) which in turn has been shown to significantly crosstalk with DNA methylation pathways [3].

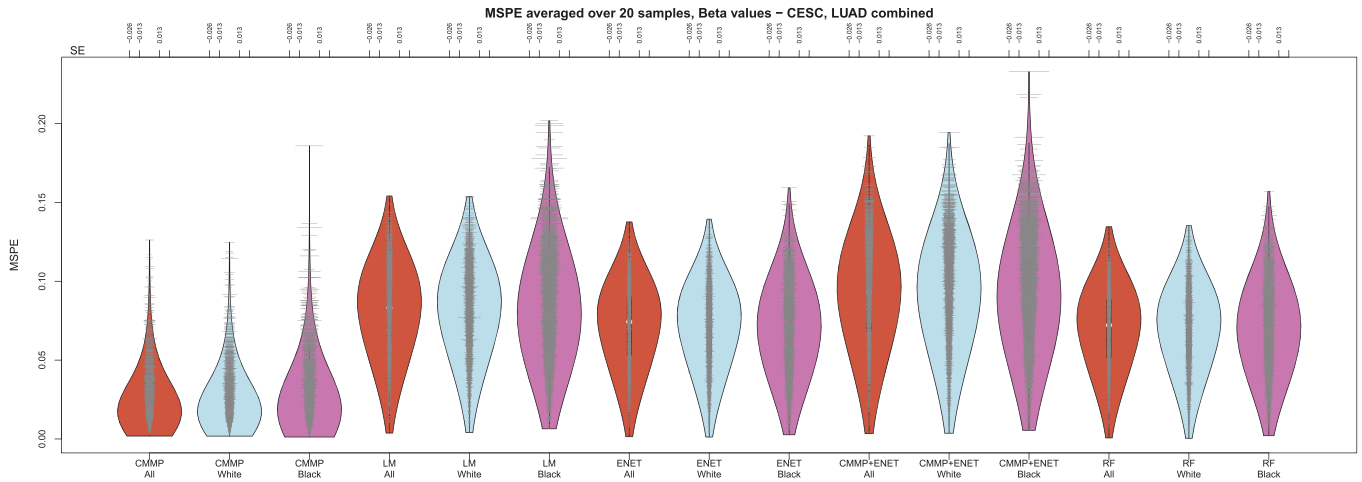

Fig. 7. Combined model average empirical test set MSPE Violin plot with individual methylation response standard error (SE) bars overlayed using raw beta values.

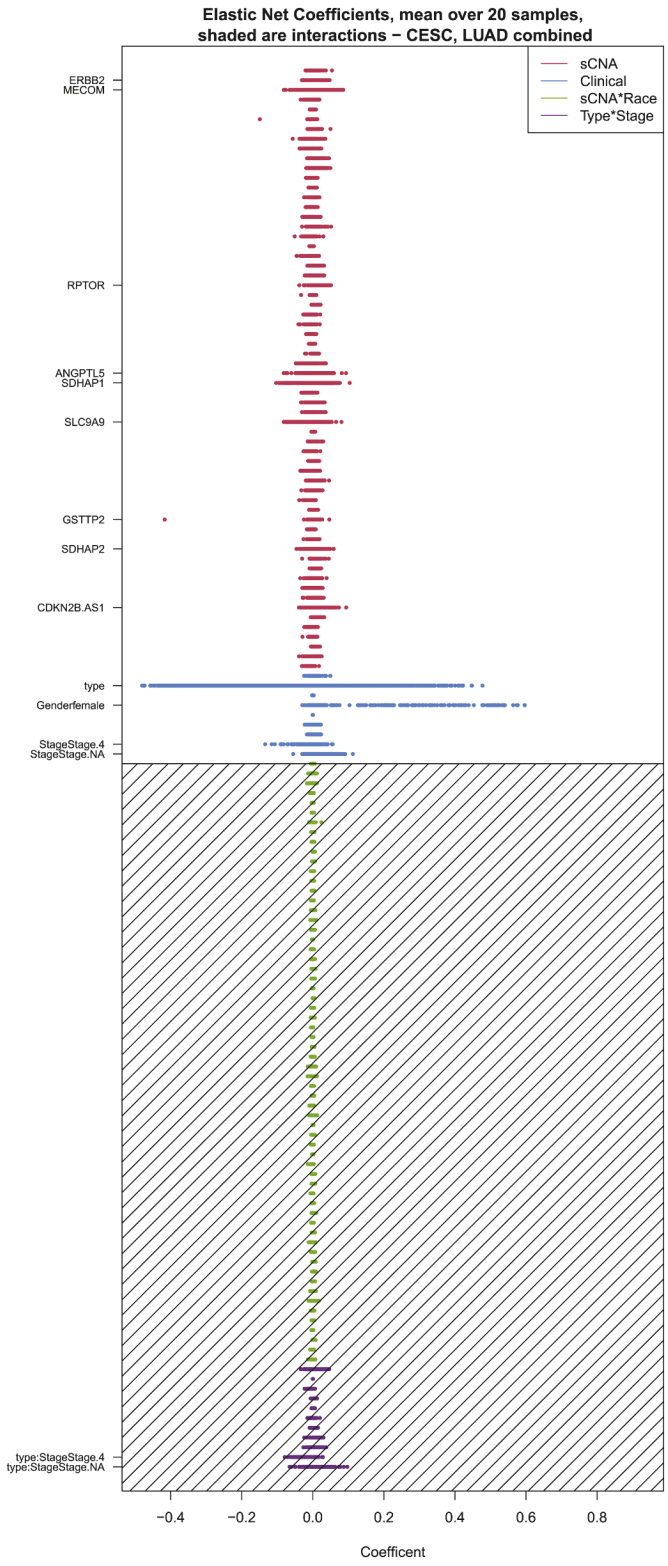

Fig. 8. Average elastic net coefficient plot with selected variable names (variance above 10% quantile) from combined analysis.

2.5.1. *Classified random effects - evidence of borrowing strength across racial groups*

To demonstrate the borrowing strength effect across races using CMMP, we plotted the classified random effects for test observations from one training-test split using the methylation response cg01315092 in Fig. 10. The y-axis plots the cluster-specific intercept (i.e. random effect) and the x-axis identifies a test patient ID. Orange circles are Black individuals and green circles are White individuals. Fully 10 out of the 10 Black individuals in the test set were assigned a non-zero classified (group-specific) random effect shared also with White individuals in the same cluster. Only 5 test set White individual was given a value of zero which indicates that because we used the no-match version of CMMP, that this test observation was not classified to any of the training set clusters and usual regression prediction was done.

3. Discussion

The most natural way to think about borrowing strength across racial groups would be through a model like the nested error regression (NER) model,

$$y_{ij} = x_{ij}^{'}\beta + w_i\alpha + v_i + e_{ij}, \tag{7}$$

where the fixed effect w_i would index race and the random effects v_i would be race-specific effects. In fact this is the strategy used in small area estimation [10] where more accurate estimates for "small areas" are achieved by borrowing strength across areas/groups via the model. Then, new observations would simply be classified by their w_n values rather than by using the CMMP strategy. This is usually termed mixed model prediction (MMP) [12]. To see why this might be suboptimal compared to what we are doing, let's take a look at the following.

A well-known formula of borrowing strength under the NER model: The best predictor (BP) of the group mean, $\theta_i = \overline{X}_i^{'}\beta + \overline{W}_i\alpha + v_i$, is

$$\overline{X}_i^{'}\beta + \overline{W}_i\alpha + \frac{n_i G}{R + n_i G}\left(\overline{y}_{i\cdot} - \overline{x}_{i\cdot}^{'}\beta - \overline{w}_{i\cdot}\alpha\right) \tag{8}$$

Thus it is seen that a BP-based method, such as the EBLUP "borrows strength" from the entire data in two ways: (i) at the population level through estimation of the population parameters, G, R; (ii) at the area level through the sample means $\overline{y}_{i\cdot}$, $\overline{x}_{i\cdot}$ and $\overline{w}_{i\cdot}$. The latter depend on the sample size, n_i. If n_i is small, the group-level strength that can be borrowed is limited. It is seen from (8) that the best one can do in estimating v_i is limited to y_i, which has sample size n_i. This is a limitation of the traditional mixed model prediction (MMP). On the other hand, CMMP estimates v_i in a different way by matching it to a cluster of the existing data, whose sample size post clustering whilst ignoring race may be much larger than n_i.

Let's return briefly to the way we kept sCNAs and methylation response filtering the same for the CESC only and the combined model. Ideally, we would repeat filtering for each model separately. This might result in different filtered sCNA and methylation responses and ultimately improve MSPE results. However, since the focus of this paper is on the new methodology of borrowing strength across racial groups, keeping the filtering the same allows us to better bring out the gains of the proposed methods.

Another clear avenue for further research would be to develop a multivariate version of our shared classified random effects prediction idea where all methylation responses would be modeled together incorporating their joint correlation structure into the model. Since it is likely that methylation values are not independent, this would accommodate that structure. We will report on this extension in future work. Finally, as with the original PrediXcan work we referenced in the Introduction for predicting gene expression values from genetic data, our approach is also gene-based. The main rationale in PrediXcan for doing so was to aid in interpretation but also reduce multiple testing issues due to working with a diminished dimensionality of the predictor space.

4. Conclusions

In summary, we have developed a novel mixed effects model based approach using shared random effects to generate accurate predictions

Fig. 9. Karyoplot of filtered genes with CNAs on human chromosomes. The red colored genes are the ones which showed up as important in the elastic net model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

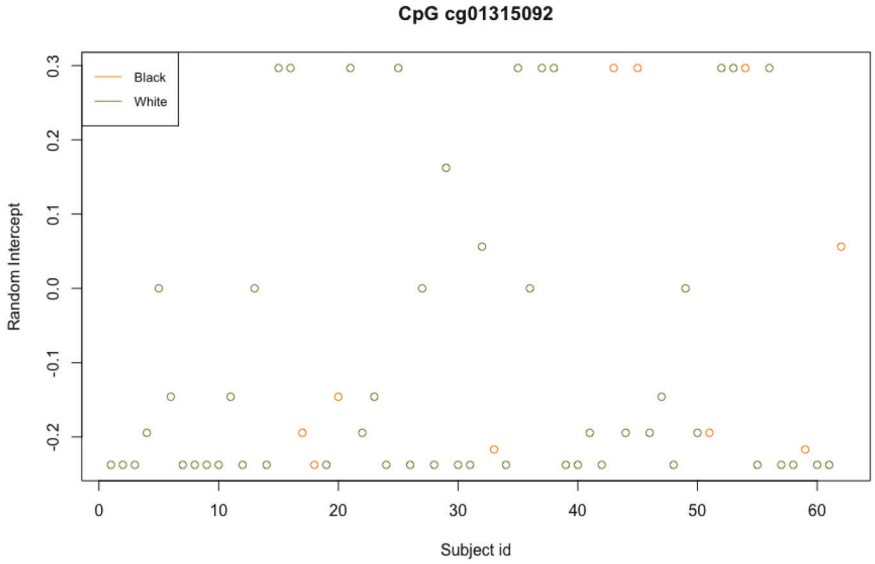

Fig. 10. Classified random effects for test data points for a particular split of full dataset.

of DNAm from gene level data profiles. These predictions demonstrate uniformly higher accuracy both in total and at the race-specific level when compared to other popular machine learning based approaches including the elastic net shrinkage estimator and random forest prediction.

Software

R scripts for implementing all methods as well as a description of the analytical pipeline can be found at URL https://github.com/hhhxz305/DNAmPredict. Each R script was named by the corresponding analysis procedure, followed by a number indicating the step. For example,

Import_setupdata.0.Rindicates this is step 0 and this code demonstrates how to import the data and set up the data ready to use for the next step.

Ethics approval and consent to participate

We used public human genomic repository data only.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available from the TCGA repository, https://portal.gdc.cancer.gov

Funding

JSR, EK, MCA and DC were all partially supported by NIH grant U54

MD010722. JSR and EK were also partially supported by NIH grant UL1-TR000460.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.10.036.

References

[1] Jill Suzanne Barnholtz-Sloan, Paola Raska, Timothy R. Rebbeck, Robert C. Millikan, Replication of gwas "hits" by race for breast and prostate cancers in European Americans and African Americans, Front. Genet. 2 (2011) 37.

[2] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[3] Du Jiamu, Lianna M. Johnson, Steven E. Jacobsen, J. Patel Dinshaw, DNA methylation pathways and their crosstalk with histone methylation, Nat. Rev. Mol. Cell Biol. 16 (2015) 519–532.

[4] Du Pan, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A. Kibbe, Lifang Hou, Simon M. Lin, Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis, BMC Bioinform. 11 (1) (2010) 587.

[5] Alfonso Dueñas-González, Marcela Lizano, Myrna Candelaria, Lucely Cetina, Claudia Arce, Eduardo Cervera, Epigenetics of cervical cancer. an overview and therapeutic perspectives, Mol. Cancer 4 (1) (2005) 38.

[6] Stephen B. Edge, David R. Byrd, AJCC Cancer Staging Manual, Springer, 2010.

[7] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, Dan L. Nicolae, Nancy J. Cox, et al., A gene-based association method for mapping traits using reference transcriptome data, Nat. Genet. 47 (9) (2015) 1091.

[8] Bernat Gel, Eduard Serra, Karyoploter: an r/bioconductor package to plot customizable genomes displaying arbitrary data, Bioinformatics 33 (19) (2017) 3088–3090, https://doi.org/10.1093/bioinformatics/btx346.

[9] C. Genovese, L. Wasserman, Operating characteristics and extensions of the false discovery rate procedure, J. R. Stat. Soc. B 64 (2002) 499–517.

[10] Ghosh Malay, J.N.K. Rao, et al., Small area estimation: an appraisal, Stat. Sci. 9 (1) (1994) 55–76.

[11] Corrado Gini, Measurement of inequality of incomes, Econ. J. 31 (121) (1921) 124–126.

[12] Jiming Jiang, Partha Lahiri, Mixed model prediction and small area estimation, Test 15 (1) (2006) 1.

[13] Jiming Jiang, Thuan Nguyen, J. Sunil Rao, Best predictive small area estimation, J. Am. Stat. Assoc. 106 (494) (2011) 732–745.

[14] J. Jiming Jiang, Sunil Rao, Jie Fan, Thuan Nguyen, Classified mixed model prediction, J. Am. Stat. Assoc. (2018) 1–11.

[15] Ruiwei Jiang, Meaghan J. Jones, Edith Chen, Sarah M. Neumann, Hunter B. Fraser, Gregory E. Miller, Michael S. Kobor, Discordance of dna methylation variance between two accessible human tissues, Sci. Rep. 5 (2015) 8257.

[16] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al., The genotype-tissue expression (gtex) project, Nat. Genet. 45 (6) (2013) 580.

[17] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, Kurt Hornik, Cluster: Cluster Analysis Basics and Extensions, 2019. R package version 2.0.8 — For new features, see the 'Changelog' file (in the package source).

[18] Charles E. McCulloch, John M. Neuhaus, Misspecifying the shape of a random effects distribution: why getting it wrong may not matter, Statistical Science (2011) 388–402.

[19] NIH, National Institutes of Health, et al., All of us Research Program, At: allofus.nih.gov. Accessed, 22, 2017.

[20] Ann G. Schwartz, Michele L. Cote, Angela S. Wenzlaff, Susan Land, Christopher I. Amos, Racial differences in the association between snps on 15q25. 1, smoking behavior, and risk of non-small cell lung cancer, J. Thorac. Oncol. 4 (10) (2009) 1195–1201.

[21] Robert L. Sherbecoe, Gerald A. Studebaker, Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units, Int. J. Audiol. 43 (8) (2004) 442–448.

[22] Kimberly D. Siegmund, Statistical approaches for the analysis of dna methylation microarray data, Hum. Genet. 129 (6) (2011) 585–595.

[23] Sanjoy K. Sinha, J.N.K. Rao, Robust small area estimation, Can. J. Stat. 37 (3) (2009) 381–399.

[24] Robert Tibshirani, Guenther Walther, Trevor Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc. B 63 (2) (2002) 411–423.

[25] Timothy J. Triche, Peter W. Laird, Kimberly D. Siegmund, Beta regression improves the detection of differential dna methylation for epigenetic epidemiology, BioRxiv (2016) 054643.

[26] Ying-Wooi Wan, Genevera I Allen, and Zhandong Liu. Tcga2stat: simple tcga data access for integrated statistical analysis in r, Bioinformatics 32 (6) (2015) 952–954.

[27] Leonie Weinhold, Simone Wahl, Sonali Pechlivanis, Per Hoffmann, Matthias Schmid, A statistical model for the analysis of beta values in dna methylation studies, BMC Bioinform. 17 (1) (2016) 480.

[28] Wanxue Xu, Mengyao Xu, Longlong Wang, Wei Zhou, Rong Xiang, Yi Shi, Yunshan Zhang, Yongjun Piao, Integrative analysis of dna methylation and gene expression identified cervical cancer-specific diagnostic biomarkers, Signal Transduct. Target. Ther. 4 (1) (2019) 1–11.

[29] Hui Zou, Trevor Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2) (2005) 301–320.