Edge sparse PCA based on gene network for correcting cell type heterogeneity in epigenome-wide association studies

Miao Rui
Macau University of Science and Technology
Dang Qi
Macau University of Science and Technology
Huang Hai Hui
Macau University of Science and Technology
Xia Liang Yong
Shanghai Jiao Tong University
Yong Liang (yongliangresearch@gmail.com)
Macau University of Science and Technology https://orcid.org/0000-0003-2403-1971

Methodology article

Keywords: epigenome association studies (EWAS), sparse principle component analysis, SVM, Drug prediction, Small sample prediction

Posted Date: August 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52780/v1

Edge sparse PCA based on gene network for correcting cell type heterogeneity in epigenome-wide association studies

Rui Miao[1], **Qi Dang**[1], **Hai-Hui Huang**[1], **Liang-Yong Xia**[2], **Yong Liang***[1.3]

*** Correspondence:**
yongliangresearch@gmail.com

Abstract

Background

In epigenome-wide association studies (EWAS), the mixed methylation expression caused by the combination of different cell types may lead the researchers to find the false methylation site related to the phenotype of interest. In order to fix this problem, researchers have proposed some non-reference methods based on sparse principle component analysis (PCA) to correct the EWAS false discovery. However, the existing model assumes that all methylation site have the same a priori probability in each PC load, but it is known that there already has network structure in the genetic variable corresponding to the methylation site. In this paper, we show that the results of the existing EWAS correction model are still not good enough. If we can integrate the existing methylation network as prior knowledge into the sparse PCA model, we can effectively improve the correction ability of the existing model.

Result

Based on the above ideas, we propose GN-ReFAEWAS, a model which uses the prior methylation gene network structure into the PCA framework for feature extraction. This model can be used to correct the false discovery in EWAS. GN-ReFAEWAS model does not need cell counting data and can estimate cell type composition through methylation principal component data. The key of this model is to solve a sparse regularize problem of methylation network. This paper uses L_0 regularize and random sampling algorithm to solve this problem. We used one simulated data set and three real data sets for experiments and compared four existing EWAS calibration models. The experimental results show that the GN-ReFAEWAS model is superior to existing models.

Conclusion

The result proved that GN-ReFAEWAS model can provide a better estimation of cell-type composition and reduce the false positives in EWAS.

Keywords: epigenome association studies (EWAS), sparse principle component analysis, SVM, Drug prediction, Small sample prediction

Background

Introduction

The whole genome methylation association analysis can help us detect new regulatory mechanisms that are more susceptible to environmental factors [1-3]. The EWAS model selects CpG sites that are easily affected by environmental factors through screening a large number of CpG sites. Standard EWAS generally uses whole blood data, which means the value of methylation data contains information on heterogeneous cell composition [4, 5]. Existing research has shown that epigenome information is greatly affected by cell changes. It means the similarity of the interest phenotype and cell composition will lead researchers to find many false methylation sites [6-9]

Fortunately, in EWAS model, the researchers used univariate testing to detect the correlation between phenotype and methylation site. Therefore, the false discovery can be corrected by adding the cell ratio as a covariate to the EWAS model. [10-12]. Unfortunately, in many situations, researchers usually do not measure the composition ratio of cells in whole blood data. In such a case, researchers have proposed some calculation models that use reference data sets to estimate the composition of cell types [13-19]. However, the model based on the reference data set can only use a small part of different blood cells as the reference data set, and cannot use other tissues [10, 20]. More Importantly, the factors affecting methylation are very complex [21]. This may lead to wrong analysis. Therefore, due to the above limitations, the researchers have proposed many non-reference models [14, 22, 23].

In 2014, Zou et.al proposed a non-reference model base on linear mixed model (LMM) and PCA which called factored spectrally transformed linear mixed model 'EWASher' (FaST-LMM-EWASher)[7]. PCA is a natural candidate to correct the false discovery of EWAS, because the previous few main PCs are related to cell type composition. This model can screen methylation data as a covariate to correct false discovery caused by cell heterogeneity without a reference data set. However, existing research shows that the effect of this model is still not good enough. Therefore, in 2016, Rahmaniet al. proposed ReFACTor model (reference-free adjustment for cell type composition model) which based on sparse PCA model [24]. This model can better screening methylation data as a covariate to correct cell heterogeneity.

The principle of sparse PCA can effectively screen out few of differentiated CPG sites to reduce the time complexity of the model and improve the correction result. Because the researchers believe that only a small part of the methylation data is significantly different (known as differentially methylated regions, DMRs) [24, 25]. If all methylation sites are used for principal component analysis, then too much useless information may lead to wrong correction results. Therefore, in the Sparse PCA, how to find the DMRs in the methylation data accurately is the key of non- reference model to correct false discovery. The existing sparse PCA model assumes that all genes have the same a priori probability in each PC load (for example ReFACTor [24]) and use the greedy algorithm based on L_0 norm to perform sparseness [26, 27]. However, it is known that there already has network structure in the genetic variable corresponding to the methylation site [28]. Known genetic network information can be considered as specific prior knowledge. The existing models cannot integrate it as prior information into sparse PCA for covariate selection. We believe if we can integrate biological network as a priori knowledge into sparse PCA, then we can improve the selection ability of existing models, better correction for false discovery caused by cell heterogeneity. Based on this idea, we designed a reference-free model for EWAS based on gene network (GN-ReFAEWAS) which used the existing gene network as a priori information to integrate into the sparse PCA. We use the edge information of gene interaction to constrain the number of non-zero elements loaded in each PC and use an alternating iteration algorithm based on random sampling for sparseness. We conducted experiments on one simulated data set and three real data sets. Experimental results show that, compared with existing models, the GN-ReFAEWAS proposed in this paper can better control the false positive rate of EWAS.

Materials

In this chapter, we introduced a simulated data set and three real data sets used in this article.

Simulation dataset

The key of non-reference model can reduce the false of EWAS is to find alternatives to cell composition data. The first few principal components of the methylation data are considered to be related to cell type. These PCs contain methylation sites that can reflect the proportion of cell composition. We simulated a methylation dataset with 5 principal components and gene networks, only a few points in each principal component are related to cell composition (DMRs). The purpose is to evaluate whether the proposed GN-ReFAEWAS model and other existing models can correctly summarize each methylation site into the correct PC loading.

We use the formula (1) to simulate the establishment the DMRs areas of each PC loadings:

$$u_k = [[rep(0,z)]^T, x_j, x_{j+1}, \dots, x_p, [rep(0,m)]^T]^T \quad (1)$$

Here, $k \in (0,5)$ represents the k-th PC loadings, assuming PC loadings has n element. $j \in (1,n)$ represents the starting DMRs sites position of the k-th PC loadings in the entire simulation data set,for example, k = 1， then, $j = 1$. p represents the position of the last DMRs sites. $(p-j)$ represents the total number of DMRs sites contained in the k-th PC loadings(in the same simulation experiment, the DMRs areas of each PC is same). The remaining $(z+m)$ elements represents non-DMRs sites. Because the area of DMRs is not fixed. Therefore, the position of $[rep(0,m)]^T$ changes dynamically in each simulation experiment. For each DMRs sites, we assume that each cell type has a different normal distribution in each PC. For non-DMRs sites, we assume that each cell type has the same normal distribution in each PC. If all PC loadings has n elements in total, then $n = z + (p-j) + m$.

For each PCs, we use the formula (2) to generate data:

$$v_k = rnorm(s) \quad (2)$$

s represents the number of samples, $rnorm(s)$ represent that the values are generated by random sampling (normal distribution).

Next, we can generate the methylation data simulation expression matrix $X \in \mathbb{R}^{n \times s}$ as $\mathcal{G}$

$$X = d_1 u_1 v_1^T + d_2 u_2 v_2^T + \cdots + d_k u_k v_k^T + \gamma\epsilon \quad (3)$$

Here $d_k \in \mathbb{N}(0,20), \gamma = 5\ and\ \epsilon \in \mathbb{N}(0,1)$

The gene edge data in the simulation data is generated in the following way: $\mathcal{G} = \mathcal{G}_1 \cup \dots \cup \mathcal{G}_k$, where $\mathcal{G}_k$ are as: $\mathcal{G}_k = \{(z_1, e_1), (z_2, e_2), \dots, (z_i, e_i)\}$, $z_1, e_1 \in \mathbb{N}(j,p)$ and $z_1 \neq e_1$. We assume that the location of the DMRs region plays a major regulatory role in the entire network. Therefore, we will generate as much edge information as possible in the DMRs during the process of establishing the simulation data set.

Real dataset

We used three real data sets to verify the performance of the proposed model. It contains a rheumatism (RA) data set, a solid tumor data for breast and colon cancer, and a Mexican and Puerto Rican descent data set. For all real data sets, according to the recommendations of existing researchers. We discarded methylation sites with mean values above 0.8 and below 0.2 [24]. The following is a detailed introduction to the data set:

First, we used a disease data set for rheumatoid arthritis. Because the cell composition of rheumatic patients is usually very different from normal people. Therefore, there is a risk of false discovery due to unexplained cell type heterogeneity. The RA data set used in this paper contains 689 experimental and control samples. Illumina HumanMethylation450 BeadChip arrays were used to obtain methylation data. The data set is already available in Gene Expression, numbered GSE42861.

Next, we use a cancer data set based on breast and colon cancer. Because the blood of cancer patients also contains many heterosexual cell mixtures and many of its components are unknown. This makes it difficult for researchers to find reliable methylation profiles. Although thousands of publications have reported the methylation changes of hundreds of genes. However, the naive analysis performed in this situation may lead to a lack of clinically available DNA methylation biomarkers. Therefore, the method in this paper can also help EWAS analysis of cancer data sets. We have uploaded the data set to GitHub at the address: https://github.com/mr1528126360/GN-ReFAEWAS.

Finally, we conducted experiments using the differential DNA methylation data set in the Mexican and Puerto Rican descent population. This data set contains 573 individuals of Mexican and Puerto Rican descent, which contains more than 450,000 variable methylated CpG sites. Prior to this. Here, we study the environmental exposure and environmental exposure of self-identity in race, genetic ancestry, DNA methylation, and analyze the differences in the environment and genetics of different ethnic groups. We found that even when adjusting the genetic lineage, the methylation level of many methylation sites is closely related to race. The influencing factors of this data are complex. More importantly, the data set contains 94 samples that have undergone a complete blood count using automated flow cytometry (Supplementary Table S1). Under such circumstances, we can study whether the results of the proposed model and other existing models still have incorrect methylation site discovery. The data set is already available in Gene Expression, numbered GSE77716.

Gene pathway dataset

The gene edge dataset is obtained from the following dataset: Molecular Signatures Database (MSigDB) (http://software.broadinstitute.org/gsea/msigdb). The corresponding data of methylation and genes are queried in the HumanMethylation450 file of each data set.

In the three real data sets, we collected the following network parameters: The methylation network of the rheumatism data set is the largest, which contains 14733988 edges. The cancer dataset network is relatively small and contains 27,955 edges. The Mexican and Puerto Rican descent data set methylation network contains 4008100 edges.

Result

This chapter uses a simulation data set and three real data sets for experimentation. For the simulation data set, we counted the number of erroneous methylation sites found in each PC loading. For the real data set, ReFACTor model retain up to 600 methylation sites and GN-ReFAEWAS model retain up to 600 edges. We first used Holm and FDR values to draw quantile–quantile plots. The graph can visually

show the reliability of the correlation analysis of methylation data. Next, in order to further verify the method in this article, we count the number of FDR significant sites for different models. The model with a lower number of FDR significant site obviously has a better correction effect. Because the ReFACTor model and GN-ReFAEWAS model are both sparse PCA models, both contain the parameter k (the number of methylation sites retained) and both models performed well in the experiment. To further compare the two models. This article counts the number of FDR significant methylation sites generated by the two models in the case of k = 0-600 (retain 0-600 methylation sites). Finally, this paper also uses known cell count information to count the discovery of false methylation sites in the Mexican dataset by different models.

Simulation dataset

In the actual calibration process, the no-reference model will only select a few important methylation sites (DMRs) in each PC loading as covariates. Therefore, whether the model can correctly select the methylation sites of DMRs becomes crucial. This paper establishes a simulation data set with 5 principal components. The total contains 200 methylation sites and 10 DMRs. This article uses the GN-ReFAEWAS model, ReFACTor model and PCA model for experiments. Among them, the parameters of the GN-ReFAEWAS model are set as follows: each principal component retains 5 edges, and a total of 25 edges are retained. Similarly, the 5 principal components of the ReFACTor model retain a total of 25 methylation sites. The PCA model retains the top 5 methylation sites of each principal component, for a total of 25 methylation sites. We counted the number of DMRs and non-DMRs methylation sites selected in each PC for each model.

According to the results in Table 1, The GN-ReFAEWAS model correctly extracts all 5 DMRs sites of PC1-PC5 loading. In comparison, ReFACTor model incorrectly identified 3,0,2,2 and 3 methylation sites with non-DMRs in PC1-PC5 loading. The PCA model performed the worst, in all 5 PC loadings, 0,4,2,2 and 5 methylation sites were judged wrong. This means that, in addition to the GN-ReFAEWAS model proposed in this paper, other comparison models may use the non-DMRs PC loading methylation information when correcting the dataset. This may lead the errors in the final correction results.

RA dataset

We use 5 different models to correct false discovery. As a baseline (Fig 2. F), we performed logistic regression without adjusting the cell composition data, it led to a severe expansion of the test statistics, which is consistent with the results reported in previous studies. There are 23168 FDR significant methylation sites in the RA dataset without correction (Table 2). Then, we present the GN-ReFAEWAS model to estimate the proportion of cell types obtained to adjust the data. This correction eliminates swelling by eliminating cell composition confusion. Then, we used the first PC of standard PCA, ReFACTor, FaST-LMM-EWASher and RefFreeEWAS for unsupervised cell type correction models. As shown in Figure 2, the correction effect of GN-ReFAEWAS is the best, and the PCA model is the worst. We also counted the number of FDR significant sites of different methods after correction. As shown in Table 2, the results show that the correction results of the GN-ReFAEWAS model proposed in this paper are significantly better. It uses only one PC to completely eliminate data inflation. In the comparison method, ReFACTor and LMM-EWASher models perform better. However, these two models still have 7 and 12 FDR significant methylation sites, respectively. We also show the comparison curves of the GN- ReFACTor and GN-ReFAEWAS models with retaining different methylation sites number (0-600). The results show that under the same number of methylation sites,

the GN-ReFAEWAS model has the best correction results (Fig 3,Supplementary Table S2). This indicates that the model proposed in this paper has a stronger ability to select methylation sites.

Tumor dataset

Tumor data is also one of the most common data sets for EWAS analysis. The uncorrected results show significant data swell (Fig 4. F). There are 2807 FDR significant methylation sites in the RA dataset without correction (Table 3). The GN-ReFAEWAS model produced very good correction results. As shown in Figure 4, in tumor data set, the GN-ReFAEWAS model has reached the optimal result of the experiment in this paper, which FDR significant methylation sites reached 0. For this comparison, The ReFACTor, LMM-EWASher, and RefFreeEWAS models contain 12, 35, and 22 FDR significant methylation sites, respectively. The PCA model had the worst results, with 735 FDR significant methylation sites. Figure 5 is the Number of FDR significant sites of retain different methylation sites number for tumor dataset. We found that GN-EefEWAS is still the best performing model. The model proposed in this article contains the fewest FDR significant sites in all cases(Supplementary Table S3).

Mexican and Puerto Rican descent population dataset

Finally, we selected the Mexican and Puerto Rican descent population data set for analysis. We have obtained similar results on this dataset with the first two real datasets (Fig, 6). According to Table 4, without correction, the data set contains 440 FDR significant sites. After correction, the GN-ReFAEWAS model completely eliminates FDR significant sites. The PCA model is the worst and still contains 128 FDR significant sites. In the case of retaining different methylation sites number. GN-ReFAEWAS model still shows a very obvious advantage(Fig. 7, Supplementary Table S4).

Although the GN-ReFAEWAS, ReFACTor, FaST-LMM-EWASher and RefFreeEWAS models also produced good results. However, these models still have the possibility of false positives. Therefore, we use the existing cell count data from this data set to estimate the number of false positives for each model. The final result showed that only the GN-ReFAEWAS model completely controlled the false positive rate (Table 5). The data set showed 4115 false positive methylation sites without correction. After correction, the best result is the GN-ReFAEWAS model. The model did not show any false positives. The ReFACTor and FaST-LMM-EWASher model immediately followed, with 6 and 5 false positives respectively, followed by the RefFreeEWAS model, with 273 errors, and the worst is the PCA model, with 361 errors (Table 5).

In summary, in all three real data sets, all indicators show that GN-ReFAEWAS performs best. This model controls the false positive rate of the data very well. This shows that using the known methylation network as a priori information can effectively help the model to select the DMRs methylation site, and further help researchers control the false positive rate of EWAS.

Discussion

Sparse PCA is a very effective model to correct EWAS, However, none of the existing reference-free models use any prior knowledge with network results. However, existing research has proven that the existing prior network information can greatly improve the feature selection capability of the model. Inspired by two facts, we propose a sparse PCA model based on methylated network structure for

correcting erroneous discovery caused by cell type heterogeneity in EWAS. We proposed a sparse regularizer based on network edge information, and designed a random sampling algorithm based on greedy principle to solve the sparse projection operation in the model. We used a simulated data set and three real data sets to verify the model in this paper. The results of the simulation data set show that only GN-ReFAEWAS correctly selects all methylation sites corresponding to each PC. This shows that, compared with the existing models, the prior network structure information can improve the model's feature filtering ability. The results of the real data set also show that the GN-ReFAEWAS model is superior to the other four existing models compared. The methylation information screened by this model can be effectively used as a covariate to effectively reduce the false positive rate of EWAS. In the case of using the same number of methylation sites, in general, the GN-ReFAEWAS model performs better than other models. In the third data set, GN-ReFAEWAS is the only model with no false positives.

In fact, in addition to having the effect of correcting the false positive rate of EWAS. The methylation sites in each PC screened by the GN-ReFAEWAS model may also be potential ethnic groups with the same biological function. Because the GN-ReFAEWAS model uses known gene network information as a priori knowledge. This determines that there must be a consistent relationship between the methylation variables after their screening. Therefore, the methylation sites screened by this model are highly likely to have biological connectivity. We hope this result can help other researchers.

In the future, we can further expand this model in the following aspects. First, further integrate gene connection information containing specific functions (for example, methylation sites known to be related to cell composition) as the focus of a priori network structure. Second, the model proposed in this paper only considers the edge information of the prior network. However, in some special cases, it may cause some information about methylation sites not in the network to be masked. Therefore, it may also be necessary to reasonably consider the information of methylation sites that are not in the network.

Conclusions

In summary, this paper proposes GN-ReFAEWAS, a model which uses the prior methylation gene network structure into the sparse PCA framework. Experimental results show that, compared with the existing non-reference EWAS correction model, the GN-ReFAEWAS model has a stronger ability to select DMRs methylation sites, and can provide better cell composition estimation and correction results.

Methods

In this chapter, we first introduce the network sparse regularize of the GN-ReFAEWAS model, and then propose the learning model of the GN-ReFAEWAS model and the generation model of PC and its loading. Finally, we introduced other comparison models used in this article.

Sparse network regularizer

There is a corresponding relationship between methylation sites and gene sites. Therefore, the corresponding methylation network data set can be established based on the known human gene pathway data set. In this network dataset, if two methylation sites have links, then the two sites with side information can be considered as a group. In this case, we set $\mathcal{G} = \{e_1, \dots, e_m\}$ as the set of all

edges in the network. Our goal is to select important methylation points (DMRs) from a methylation network containing m edges and use the scores of the principal components as covariates to correct false discovery caused by cell heterogeneity. Here, we presented a Methylation network Sparse penalty as formula (4):

$$\| u \|_{MN} = \underset{\forall \mathcal{G}' \subseteq \mathcal{G}, support(u) \subseteq V(\mathcal{G}')}{\text{minimize}} |\mathcal{G}'| \tag{4}$$

According to existing research, the SVD framework is usually used to solve the PCA model, then the formula (4) can be written as:

$$\underset{\|u\|_2 \leq 1, \|v\|_2 \leq 1}{\text{maximize}} u^T X v, s.t. \| u \|_{MN} \leq s \tag{5}$$

Where u and v represents methylation sites weight and sample weight ($u \subseteq \mathbb{R}^{p \times 1}$, $v \subseteq \mathbb{R}^{n \times 1}$).$X \in \mathbb{R}^{P \times N}$ represents a matrix with p methylation sites and n samples.

Random sampling algorithm based on greedy principle for GN-ReFAEWAS

The key to solving this problem is how to determine the value of the sample weight vector v, set $z = Xv$, then formula (6) can be expressed as:

$$\underset{\|u\|_2 \leq 1}{\text{maximize}} u^T z, s.t. \| u \|_{MN} \leq s \tag{6}$$

Traditional sparse PCA mostly adopts the principle of greedy algorithm. According to the threshold s, the first s largest methylation sites of each PC are retained. However, this algorithm has no randomness and cannot obtain the global optimal solution. This paper proposes a random sampling algorithm based on the principle of greed as a solution to the learning model of GN-ReFAEWAS. Let $z = \{z_1, \dots, z_p\}$ is the weight vector of methylation sites, $v = \{v_1, \dots, v_n\}$ is the sample weight vector of the dataset, $v = X^T u$, $u$is the weight information of the methylated sites retained after sparseness, $\mathcal{G} = \{e_1, \dots, e_m\}$ is the set of methylated network edges, N is the parameter that controls the randomness of the algorithm. The larger the value of this parameter, the higher the randomness of algorithm initialization. T is a parameter of the control algorithm to reduce the speed of randomness. S is the number of edges remaining in the methylation network after sparseness. The steps of the algorithm are as follows:

1. First, use the randomization model to initialize the sample weight vector v of the data set, and manually enter the parameters S, N and T
2. Use Xv to calculate the weight of the methylation data z, which can get the weight value of each current methylation site
3. Calculate the weight vector $\mathcal{G}$ of the edge according to the vector z and the methylated network information (for example, z_1, z_2 are the weight parameters of the two methylation sites corresponding to the edgee_1, then $e_1 = \sqrt{z_1^2 + z_2^2}$)
4. Sort the weight vector $\mathcal{G}$ of the edges from largest to smallest, reserve $(1 + N) \times S$ edges, and store the result in the temporary vector $\mathcal{G}_1$
5. Update weight $N(N \geq 0)$, $N = (N - T)$, use random sampling model to randomly extract S edges

in vector $\mathcal{G}_1$ and obtain the weight vector $\mathcal{G}'$ of the updated edge. Finally, we can get the sparse methylation site vector u， $u = \frac{\hat{u}}{\|\hat{u}\|}, where\ \hat{u} = \mathcal{P}_{\mathcal{G}}(z,s)\ and\ z = Xv$

6. Use u to update the sample weight vector v, repeat step 2 until convergence, $v = \frac{\hat{v}}{\|\hat{v}\|}, where\ \hat{v} = X^T u$
7. According to the methylation sites retained by the model, we can get the final solution (the scores of the principal components).

Among them, in step 5, $u = \frac{\hat{u}}{\|\hat{u}\|}$,We define a $s - network$ sparse projection operator： $\hat{u} = \mathcal{P}_{\mathcal{G}}(z,s)$ and for the j-th methylation site, $\left[\mathcal{P}_{\mathcal{G}}(z,s)\right]_j$:

$$\left[\mathcal{P}_{\mathcal{G}}(z,s)\right]_j = \begin{cases} z_j, & if \mathcal{G}(j) \cap supp(norm_{\mathcal{G}}(z), k) \neq \emptyset \\ 0, & otherwise \end{cases} \tag{7}$$

Where $\mathcal{G}(j)$ is an edge of the methylation network that contains methylation site i. If $\mathcal{G}(j)$is retained, then there is $\left[\mathcal{P}_{\mathcal{G}}(z,s)\right]_j = z_j$， otherwise $\left[\mathcal{P}_{\mathcal{G}}(z,s)\right]_j = 0$

Generating model of PC and PC loading

We use the following method to generate each principal component of the model (formula (8)):

$$\underset{\|u_l\|_2 \leq 1, \|v_l\|_2 \leq 1}{\text{maximize}}\ u_l^T X v_l \tag{8}$$

$$subject\ to\ \|\ u_l\ \|_{MN} \leq v_l \perp v_1, \dots, v_{l-1}$$

Where we have known that $l \geq 2$ and pairs with $\{u_i, v_i\}(i = 1, \dots, l-1)$,and v_i must be orthogonal with $v_1, \dots, v_{l-1}$. This paper adopts an alternating iteration to solve the problem of formula 8. First of all, fixed u_i and update v_l and vice versa. When u_i is fixed, formula (8) can be expressed as formula (9):

$$\underset{\|u_l\|_2 \leq 1}{\text{maximize}}\ u_l^T X v_l \tag{9}$$

$$subject\ to\ v_l \perp v_1, \dots, v_{l-1}$$

We use the Gram-Schmidt orthogonalization method in our model [29].Let $V_{l-1} = [v_1; \cdots; v_{l-1}] \in \mathbb{R}^{n \times (l-1)}$ and $V_{l-1}^{\perp} \in \mathbb{R}^{n \times (n-l+1)}$, the columns of this method are complementary in space with V_{l-1},i.e, $[V_{l-1}; V_{l-1}^{\perp}]$(Spatial orthogonality). Because $v_l \perp v_1, \dots, v_{l-1}$, v_l belong to $V_{l-1}^{\perp}$. Therefore, we can exchange the v_l as $V_{l-1}^{\perp}\beta$. We can get formula (10):

$$\underset{\|u_l\|_2 \leq 1}{\text{maximize}}\ u_l^T X V_{l-1}^{\perp} \beta_{\ l} \tag{10}$$

$$subject\ to\ \|\beta\|_2 \leq 1$$

We can get the optimal solution of the formula (11):

$$\beta = \frac{V_{l-1}^{\perp} X^T u_l}{\left\| V_{l-1}^{\perp} X^T u_l \right\|} \tag{11}$$

Where $v_l = V_{l-1}^{\perp}\beta$, then we have formula (12):

$$v_l = \frac{V_{l-1}^{\perp} {V_{l-1}^{\perp}}^T X^T u_l}{\left\| V_{l-1}^{\perp} {V_{l-1}^{\perp}}^T X^T u_l \right\|} \tag{12}$$

Where we use the following theorem to solve $V_{l-1}^{\perp} {V_{l-1}^{\perp}}^T$:

Lemma: If the space of $V_{l-1}^{\perp}$ and v_{l-1} $is\ orthogonal, then\ V_{l-1}^{\perp} {V_{l-1}^{\perp}}^T = I - V_{l-1} V_{l-1}^T$

Theorem 1. The optimal solution of problem (13) is

$$v_l = \frac{\hat{v}_l}{\|\hat{v}_l\|}, where\ \hat{v}_l = (I - V_{l-1} V_{l-1}^T) X^T u_l \tag{13}$$

Based on this Lemma, we can replace the $V_{l-1}^{\perp} {V_{l-1}^{\perp}}^T$ with $I - V_{l-1} V_{l-1}^T$.Then, we get the solution of formula 6 as $v_l = \frac{\hat{v}_l}{\|\hat{v}_l\|}$,where $\hat{v}_l = (I - V_{l-1} V_{l-1}^T) X^T u_l$

After v_l is updated, we can use the random sampling algorithm based on greedy principle proposed in this paper to update u_l. Therefore, the final alternate iteration strategy is as formula (14-15):

$$u_l \leftarrow \frac{\hat{u}}{\|\hat{u}\|}, where\ \hat{u} = \mathcal{P}_{\mathcal{G}}(z, s_l)\ and\ z = X v_l \tag{14}$$

$$v_l \leftarrow \frac{\hat{v}}{\|\hat{v}\|}, where\ \hat{v} = (I - V_{l-1} V_{l-1}^T) X^T u_l \tag{15}$$

ReFACTor

The ReFACTor model was proposed by Rahmani et al[24]. The ReFACTor model is based on the sparse PCA method and corrects EWAS false discovery by screening DMRs sites. This model is a typical sparse PCA model. The model is based on the idea of differential methylation (that is, only a small part of the methylation sites is different from other sites), combined with the sparseness idea, first select all the methylation sites according to left-singular vectors The specified k methylation sites, and finally perform PCA again to distinguish the principal components and select methylation data.

FaST-LMM-EWASher

FaST-LMM-EWASher model was proposed by Zou et al[7]. This model computes the methylome similarity between every pair of samples and then use these similarities as the covariance in the mixed model as an implicit proxy for cell-type composition.

RefFreeEWAS

RefFreeEWAS model was proposed by Houseman et al. This model closely related to surrogate variable analysis[30], The model combines the LMM method and the PCA method to obtain the final correction data through cross iteration.

Declarations

Abbreviations

EWAS: Epigenome-wide Association Studies (EWAS),PCA: Principle Component Analysis, RefFreeEWAS: Reference-free-EWAS, FaST-LMM-EWASher: factored spectrally transformed linear mixed model 'EWASher', ReFACTor: reference-free adjustment for cell type composition, GN-ReFAEWAS: reference-free model for EWAS based on gene network

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analysed during the current study are available in the following public database: Gene Expression, numbered GSE42861, GSE77716 and https://github.com/mr1528126360/GN-ReFAEWAS. The gene edge dataset is obtained from the following public dataset: Molecular Signatures Database (MSigDB) (http://software.broadinstitute.org/gsea/msigdb).

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by the Macau Science and Technology Development Funds Grands No.003/2016/AFJ and No.0055/2018/A2 from the Macau Special Administrative Region of the People's Republic of China. Macau Science and Technology Development Funds provides the computing equipment needed in this research and played no role in the design of the study, the collection, analysis, and interpretation of data and in writing the manuscript.

Authors' Contributions

RM conceived the conception. RM also designed and developed the method, acquired and analyzed the data and result. LYX, HHH and QD wrote, reviewed and revised the manuscript. YL is correspondence author. All authors have read and approved the final manuscript.

Acknowledgments

Not applicable

Author Information

1.Faculty of Information Technology, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, 999078, China

2. Biological Engineering, Shanghai Jiao Tong University, 1954 Huashan Road, Xuhui District, Shanghai, China

3. State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, 999078, China

References

[1] J. M. Flanagan, "Epigenome-wide association studies (EWAS): past, present, and future," in Cancer Epigenetics: Springer, 2015, pp. 51-63.

[2] M. Verma, "Epigenome-wide association studies (EWAS) in cancer," Current genomics, vol. 13, no. 4, pp. 308-313, 2012.

[3] K. B. Michels et al., "Recommendations for the design and analysis of epigenome-wide association studies," Nature methods, vol. 10, no. 10, p. 949, 2013.

[4] T. M. Murphy and J. Mill, "Epigenetics in health and disease: heralding the EWAS era," The Lancet, vol. 383, no. 9933, pp. 1952-1954, 2014.

[5] M. Li et al., "EWAS Atlas: a curated knowledgebase of epigenome-wide association studies," Nucleic acids research, vol. 47, no. D1, pp. D983-D988, 2019.

[6] A. E. Jaffe and R. A. Irizarry, "Accounting for cellular heterogeneity is critical in epigenome-wide association studies," Genome biology, vol. 15, no. 2, p. R31, 2014.

[7] J. Zou, C. Lippert, D. Heckerman, M. Aryee, and J. Listgarten, "Epigenome-wide association studies without the need for cell-type composition," Nature methods, vol. 11, no. 3, pp. 309-311, 2014.

[8] H. Naeem et al., "Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array," BMC genomics, vol. 15, no. 1, p. 51, 2014.

[9] C. J. Patel, J. Bhattacharya, and A. J. Butte, "An environment-wide association study (EWAS) on type 2 diabetes mellitus," PloS one, vol. 5, no. 5, 2010.

[10] E. A. Houseman et al., "DNA methylation arrays as surrogate measures of cell mixture distribution," BMC bioinformatics, vol. 13, no. 1, p. 86, 2012.

[11] S. Graw, R. Henn, J. A. Thompson, and D. C. Koestler, "pwrEWAS: a user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS)," BMC bioinformatics, vol. 20, no. 1, p. 218, 2019.

[12] E. A. Houseman, K. T. Kelsey, J. K. Wiencke, and C. J. Marsit, "Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective," BMC bioinformatics, vol. 16, no. 1, p. 95, 2015.

[13] X. Zheng et al., "MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes," Genome biology, vol. 15, no. 7, p. 419, 2014.

[14] E. A. Houseman, J. Molitor, and C. J. Marsit, "Reference-free cell mixture adjustments in analysis of DNA methylation data," Bioinformatics, vol. 30, no. 10, pp. 1431-1439, 2014.

[15] A. M. Newman et al., "Robust enumeration of cell subsets from tissue expression profiles," Nature methods, vol. 12, no. 5, pp. 453-457, 2015.

[16] K. Yoshihara et al., "Inferring tumour purity and stromal and immune cell admixture from expression data," Nature communications, vol. 4, no. 1, pp. 1-11, 2013.

[17] [D. C. Koestler et al., "Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis," Epigenetics, vol. 8, no. 8, pp. 816-826, 2013.

[18] W. P. Accomando, J. K. Wiencke, E. A. Houseman, H. H. Nelson, and K. T. Kelsey, "Quantitative reconstruction of leukocyte subsets using DNA methylation," Genome biology, vol. 15, no. 3, p. R50, 2014.

[19] A. E. Teschendorff, C. E. Breeze, S. C. Zheng, and S. Beck, "A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies," BMC bioinformatics, vol. 18, no. 1, p. 105, 2017.

[20] L. E. Reinius et al., "Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility," PloS one, vol. 7, no. 7, 2012.

[21] [D. C. Koestler et al., "Improving cell mixture deconvolution by id entifying o ptimal dna methylation l ibraries (idol)," BMC bioinformatics, vol. 17, no. 1, p. 120, 2016.

[22] E. A. Houseman, M. L. Kile, D. C. Christiani, T. A. Ince, K. T. Kelsey, and C. J. Marsit, "Reference-free deconvolution of DNA methylation data and mediation by cell composition effects," BMC bioinformatics, vol. 17, no. 1, p. 259, 2016.

[23] K. McGregor et al., "An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies," Genome biology, vol. 17, no. 1, p. 84, 2016.

[24] E. Rahmani et al., "Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies," Nature methods, vol. 13, no. 5, p. 443, 2016.

[25] S. Li et al., "An optimized algorithm for detecting and annotating regional differential methylation," in BMC bioinformatics, 2013, vol. 14, no. 5: BioMed Central, p. S10.

[26] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," Journal of Machine Learning Research, vol. 11, no. 2, 2010.

[27] [X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," Journal of Machine Learning Research, vol. 14, no. Apr, pp. 899-925, 2013.

[28] R. Hoffmann and A. Valencia, "A gene network for navigating the literature," Nature genetics, vol. 36, no. 7, pp. 664-664, 2004.

[29] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," Neural networks, vol. 13, no. 4-5, pp. 411-430, 2000.

[30] J. Leek and J. Storey, "Bioconductor's sva package," dim (svadat), vol. 1, no. 1000, p. 20, 2007.

Figure and Figure Legend

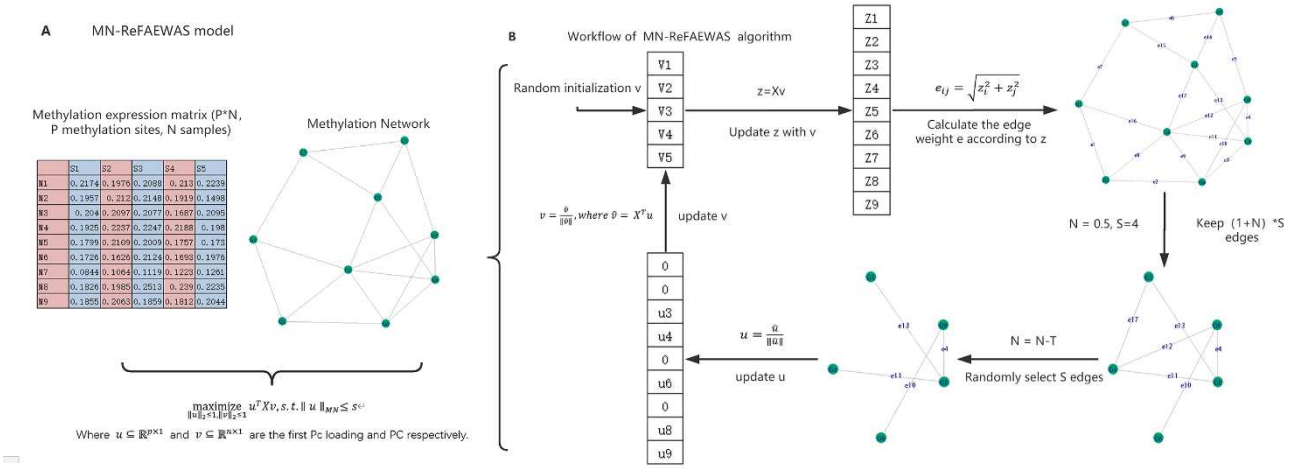

Figure 1. (A) GN-ReFAEWAS model: used to correct EWAS false discovery. (B) Workflow of GN-ReFAEWAS model.1. First initialize the parameter v randomly 2. Calculate the parameter z to get the weight of each methylation site 3. Calculate the weight of the edge according to the prior network, 4.

Perform a random sampling method based on the greedy principle. Obtain the sparse methylation site weight u, 5. Use parameter u to update parameter v, and return to step

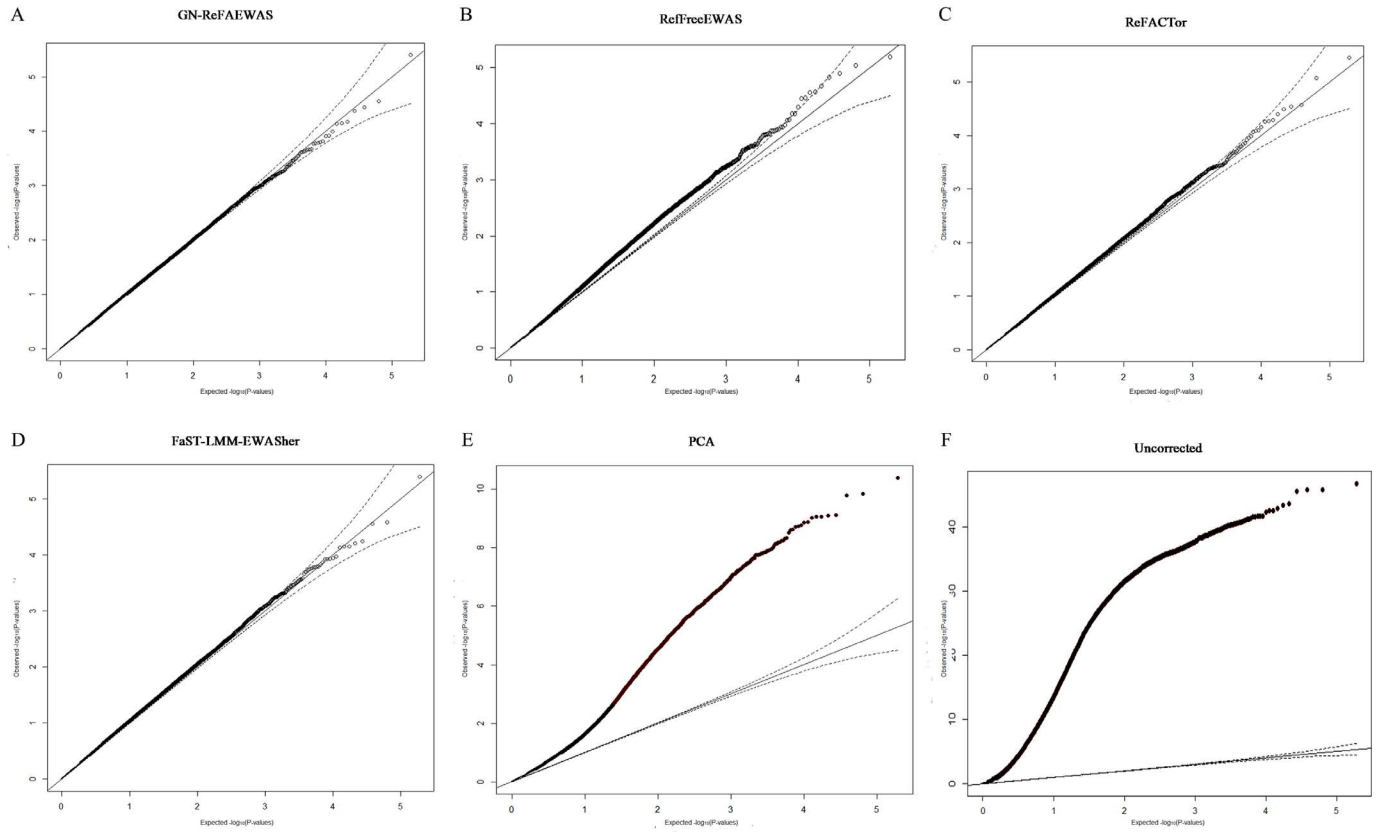

Figure. 2. The correction result of the RA data set is represented by the QQ graph of the associated test. The degree of deviation from the benchmark indicates the degree of data expansion due to cell heterogeneity. The dotted line indicates the 95% confidence interval. (A) GN-ReFAEWAS model correction results (using the first PC) (b) ReFACTor model correction results (using the first PC) (c) RefFreeEWAS model correction results (d) FaST-LMM-EWASher model correction result, (e) PCA correction result (using the first PC), and (f) no correction result.

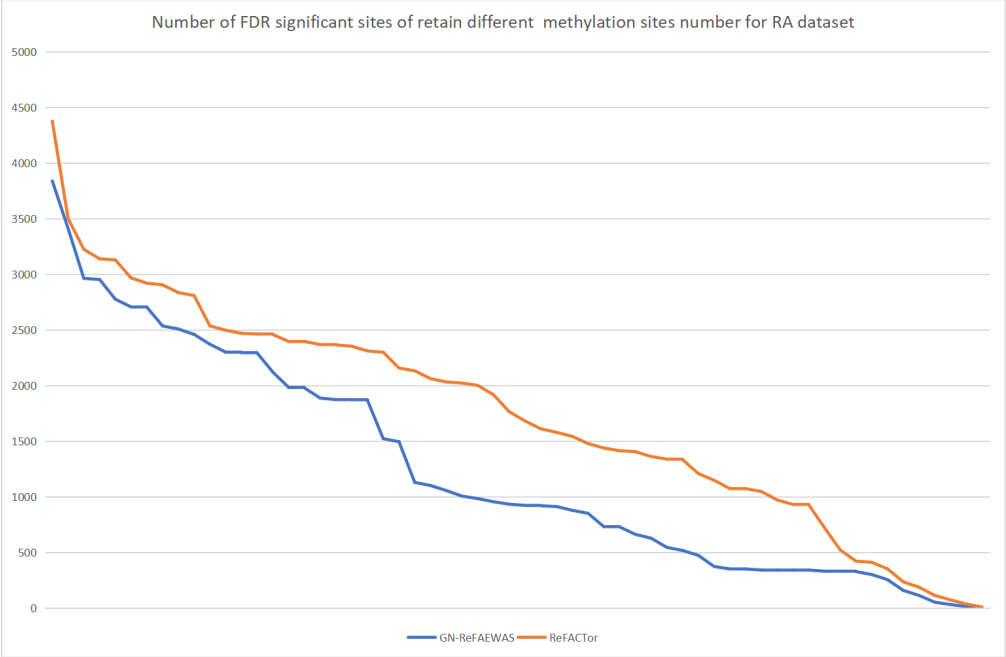

Figure. 3. Number of FDR significant sites of retain different methylation sites number for RA dataset with retaining different methylation sites number (0-600)

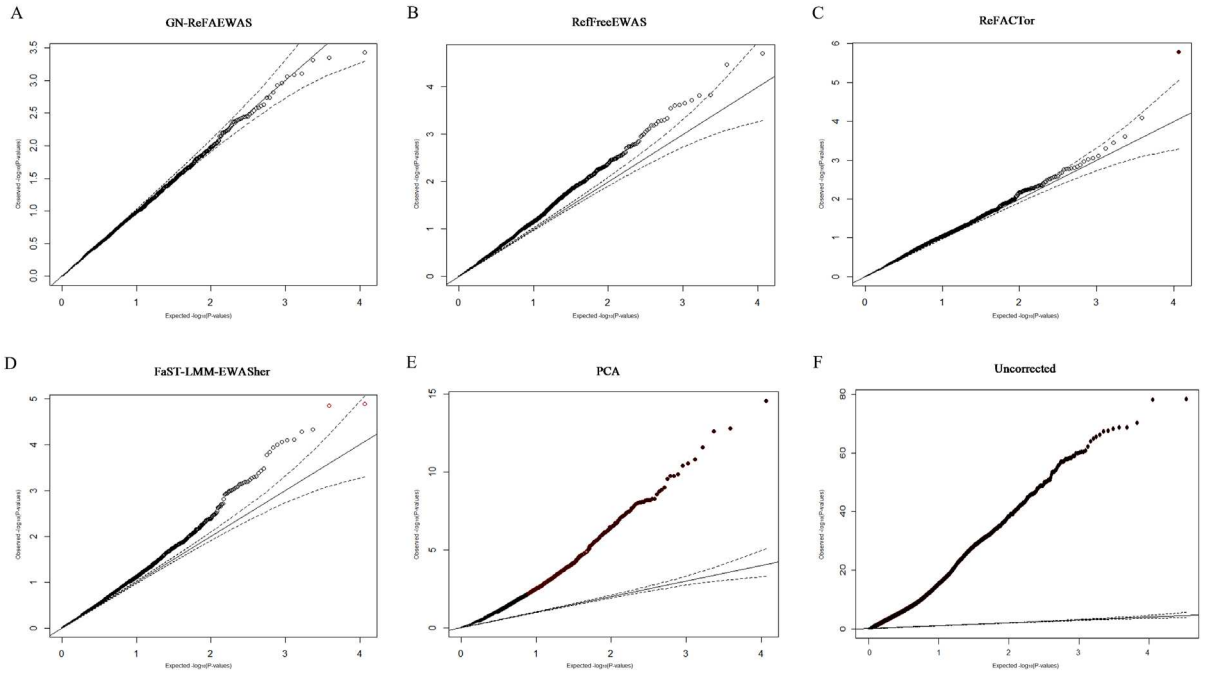

Figure. 4. The correction result of the Tumor data set is represented by the QQ graph of the associated test. The degree of deviation from the benchmark indicates the degree of data expansion due to cell heterogeneity. The dotted line indicates the 95% confidence interval. (A) GN-ReFAEWAS model correction results (using the first PC) (b) ReFACTor model correction results (using the first PC) (c) RefFreeEWAS model correction results (d) FaST-LMM-EWASher model correction result, (e) PCA correction result (using the first PC), and (f) no correction result.

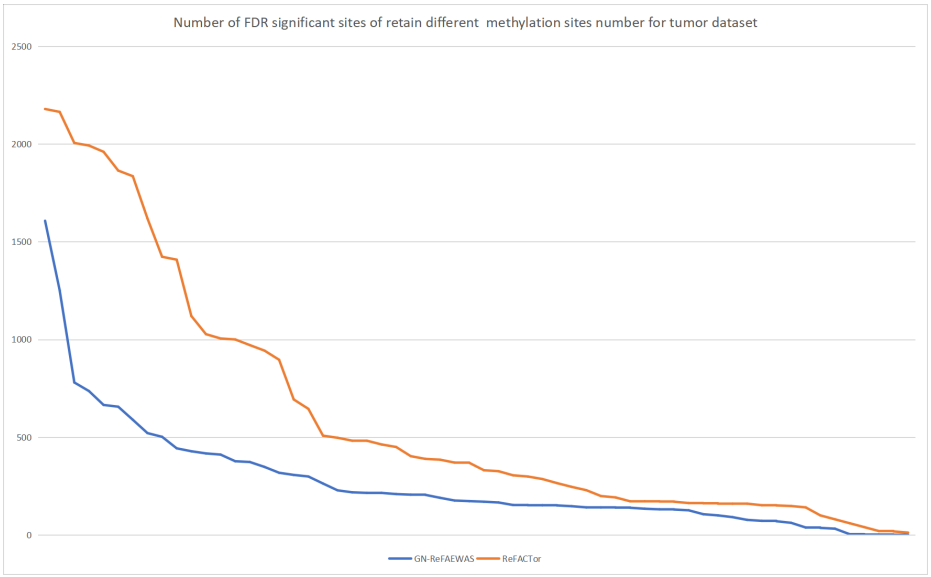

Figure. 5. Number of FDR significant sites of retain different methylation sites number for tumor dataset with retaining different methylation sites number (0-600)

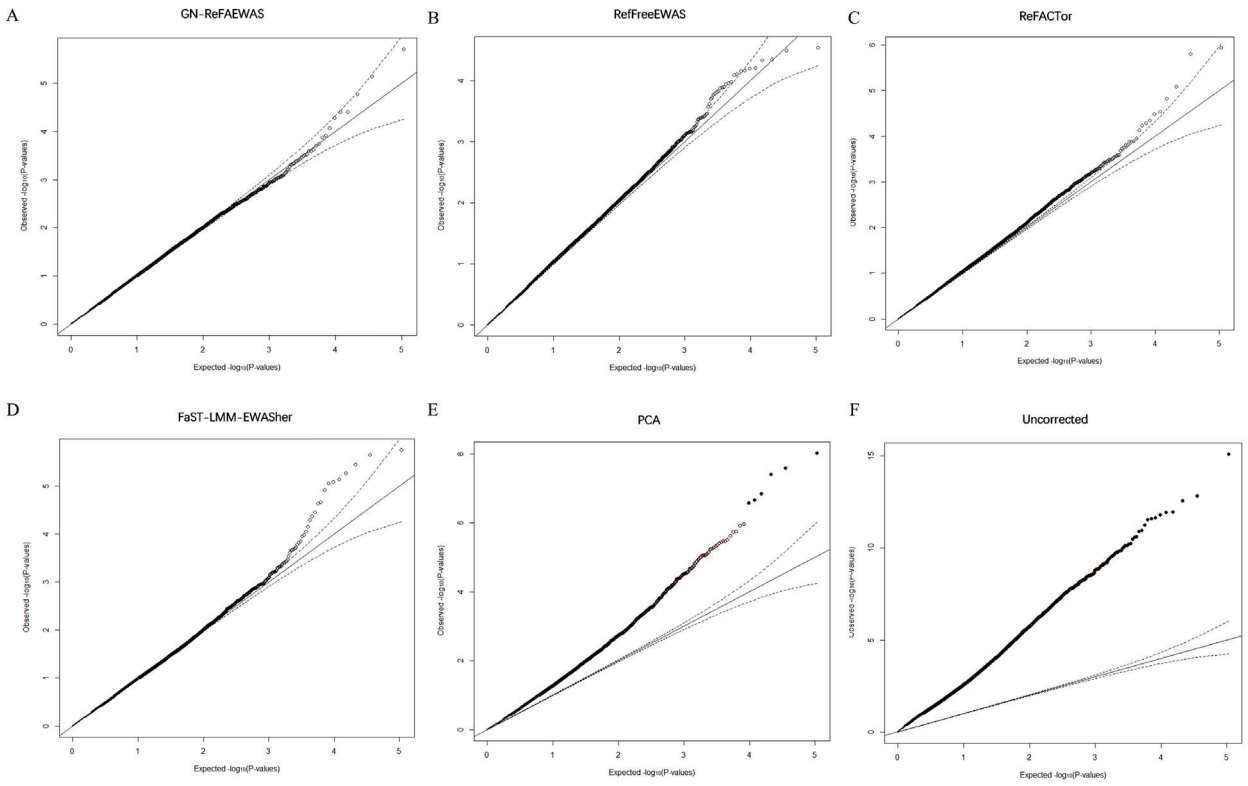

Figure. 6. The correction result of the Mexican and Puerto Rican descent population data set is represented by the QQ graph of the associated test. The degree of deviation from the benchmark indicates the degree of data expansion due to cell heterogeneity. The dotted line indicates the 95% confidence interval. (A) GN-ReFAEWAS model correction results (using the first PC) (b) ReFACTor model correction results (using the first PC) (c) RefFreeEWAS model correction results (d) FaST-

LMM-EWASher model correction result, (e) PCA correction result (using the first PC), and (f) no correction result.

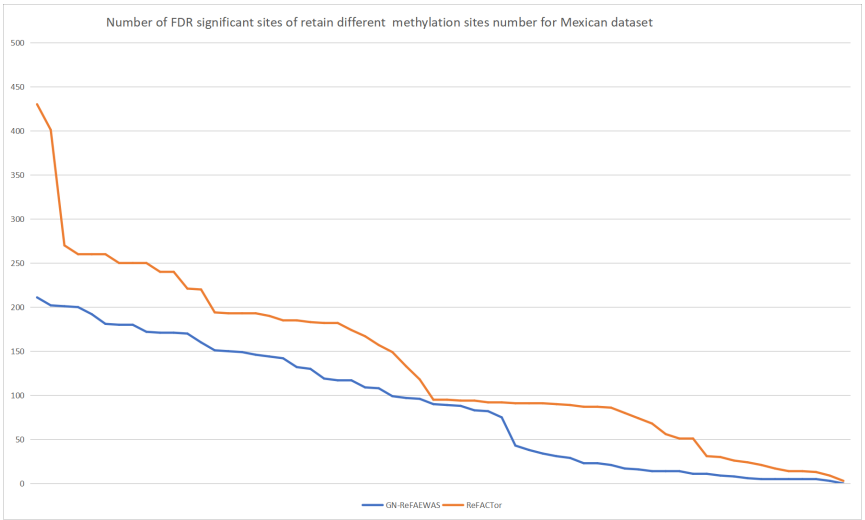

Figure. 7. Number of FDR significant sites of retain different methylation sites number for Mexican dataset with retaining different methylation sites number (0-600)

Table

Table 1 The number of DMRs and non-DMRs methlation sites selected in each PC for each model of the simulation dataset

	PC1		PC2		PC3		PC4		PC5	
	DMRS	Non-DMRS	DMRS	Non-DMRS	DMRS	Non-DMRS	DMRS	Non-DMRS	DMRS	Non-DMRS
PCA	5	0	1	4	3	2	3	2	0	5
ReFACTor	2	3	5	0	3	2	3	2	2	3
GN-ReFAEWAS	**5**	**0**	**5**	**0**	**5**	**0**	**5**	**0**	**5**	**0**

Table 2. Number of significant sites for RA dataset

	Uncorrected	PCA	ReFACTor	LMM-EWASher	RefFreeEWAS	**GN-ReFAEWAS**
Number	23168	2215	7	12	321	**0**

Table 3. Number of significant sites for Tumor dataset

	Uncorrected	PCA	ReFACTor	LMM-EWASher	RefFreeEWAS	**GN-ReFAEWAS**
Number	2807	735	12	35	22	**0**

Table 4. Number of significant sites for Mexican dataset

	Uncorrected	PCA	ReFACTor	LMM-EWASher	RefFreeEWAS	**GN-ReFAEWAS**
Number	440	128	3	10	35	**0**

Table 5. The false discovery of Mexican and puerto rican descent population dataset for different models

	Lymphocyte	Monocyte	Neutrophil granulocyte	Eosinophil granulocyte	Basophil granulocyte	Total
Uncorrected	1071	1000	1404	1211	689	4115
FaST-LMM-EWASher	1	3	0	1	0	5
ReFACTor	0	3	0	1	2	6
RefFreeEWAS	1	0	111	110	51	273
PCA	15	9	91	165	81	361
GN-ReFAEWAS	**0**	**0**	**0**	**0**	**0**	**0**

Figures

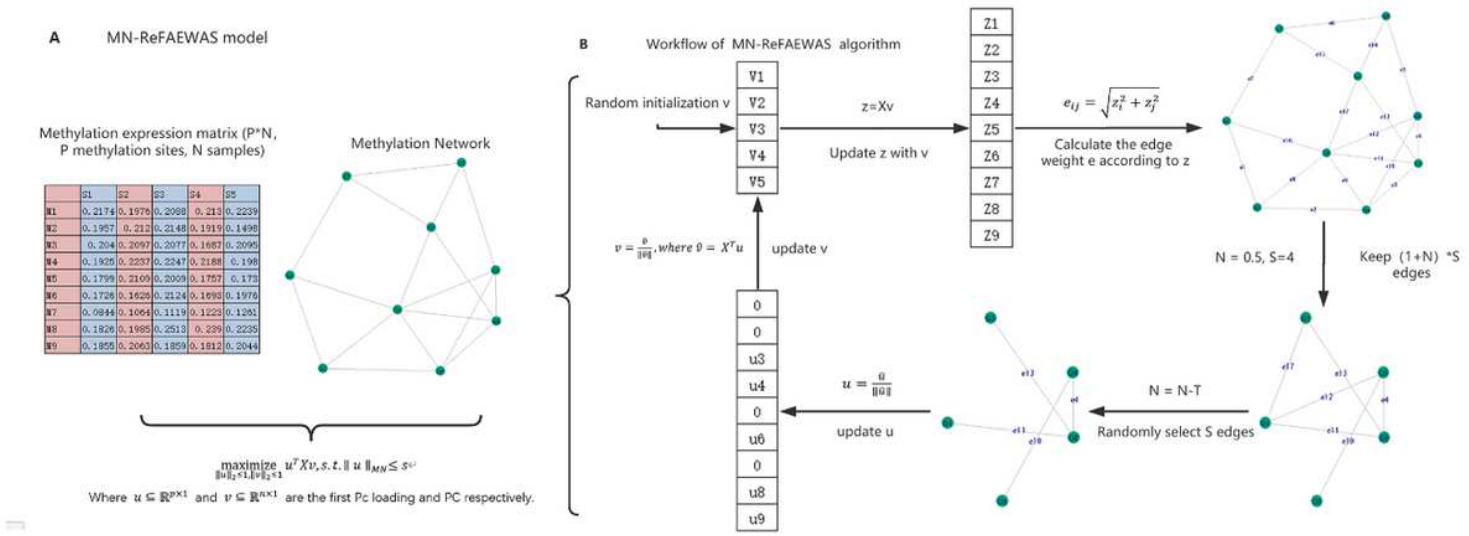

Figure 1

(A) GN-ReFAEWAS model: used to correct EWAS false discovery. (B) Workflow of GN-ReFAEWAS model.1. First initialize the parameter v randomly 2. Calculate the parameter z to get the weight of each methylation site 3. Calculate the weight of the edge according to the prior network, 4. Perform a random sampling method based on the greedy principle. Obtain the sparse methylation site weight u, 5. Use parameter u to update parameter v, and return to step

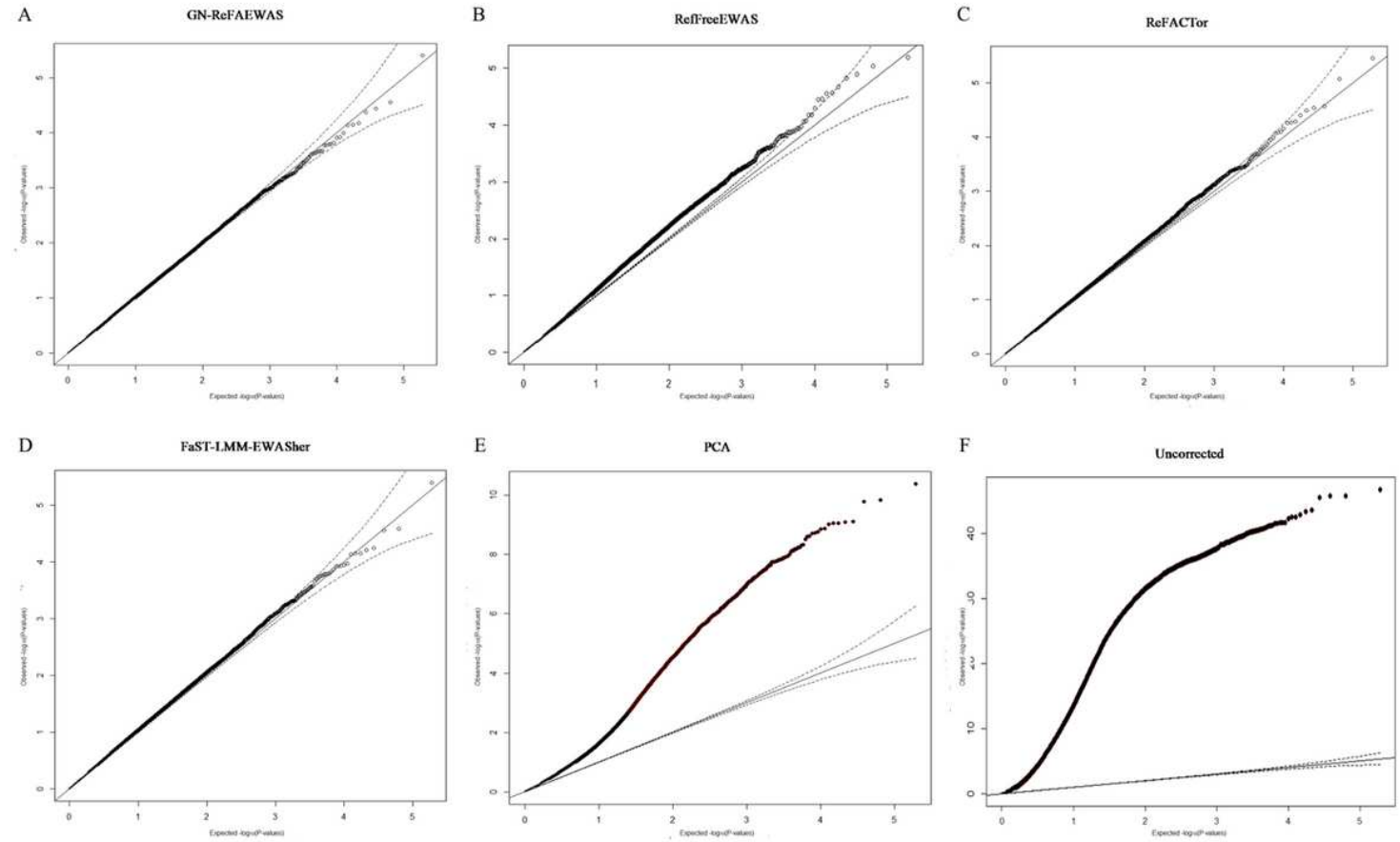

Figure 2

The correction result of the RA data set is represented by the QQ graph of the associated test. The degree of deviation from the benchmark indicates the degree of data expansion due to cell heterogeneity. The dotted line indicates the 95% confidence interval. (A) GN-ReFAEWAS model correction results (using the first PC) (b) ReFACTor model correction results (using the first PC) (c) RefFreeEWAS model correction results (d) FaST-LMM-EWASher model correction result, (e) PCA correction result (using the first PC), and (f) no correction result.

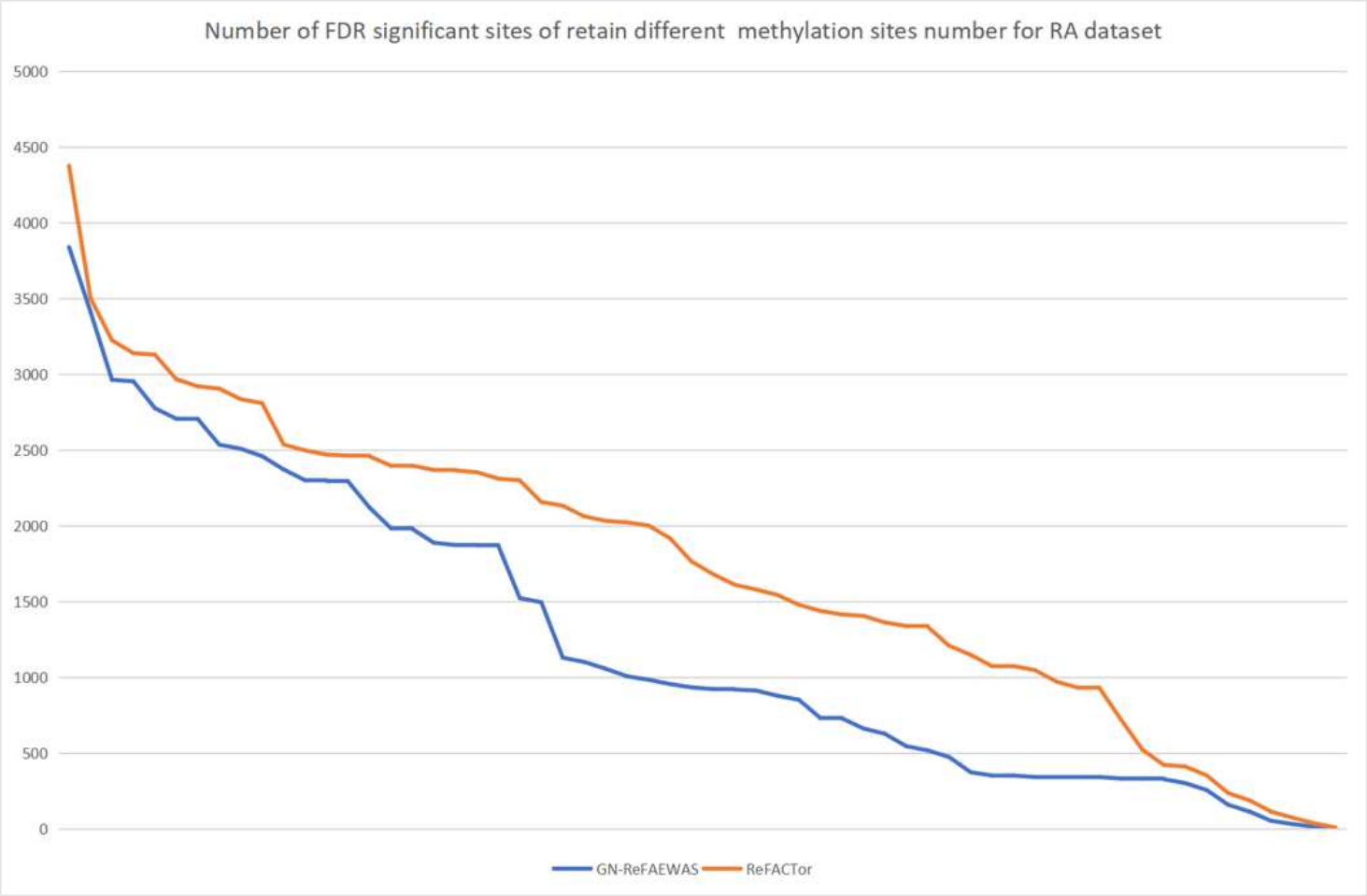

Figure 3

Number of FDR significant sites of retain different methylation sites number for RA dataset with retaining different methylation sites number (0-600)

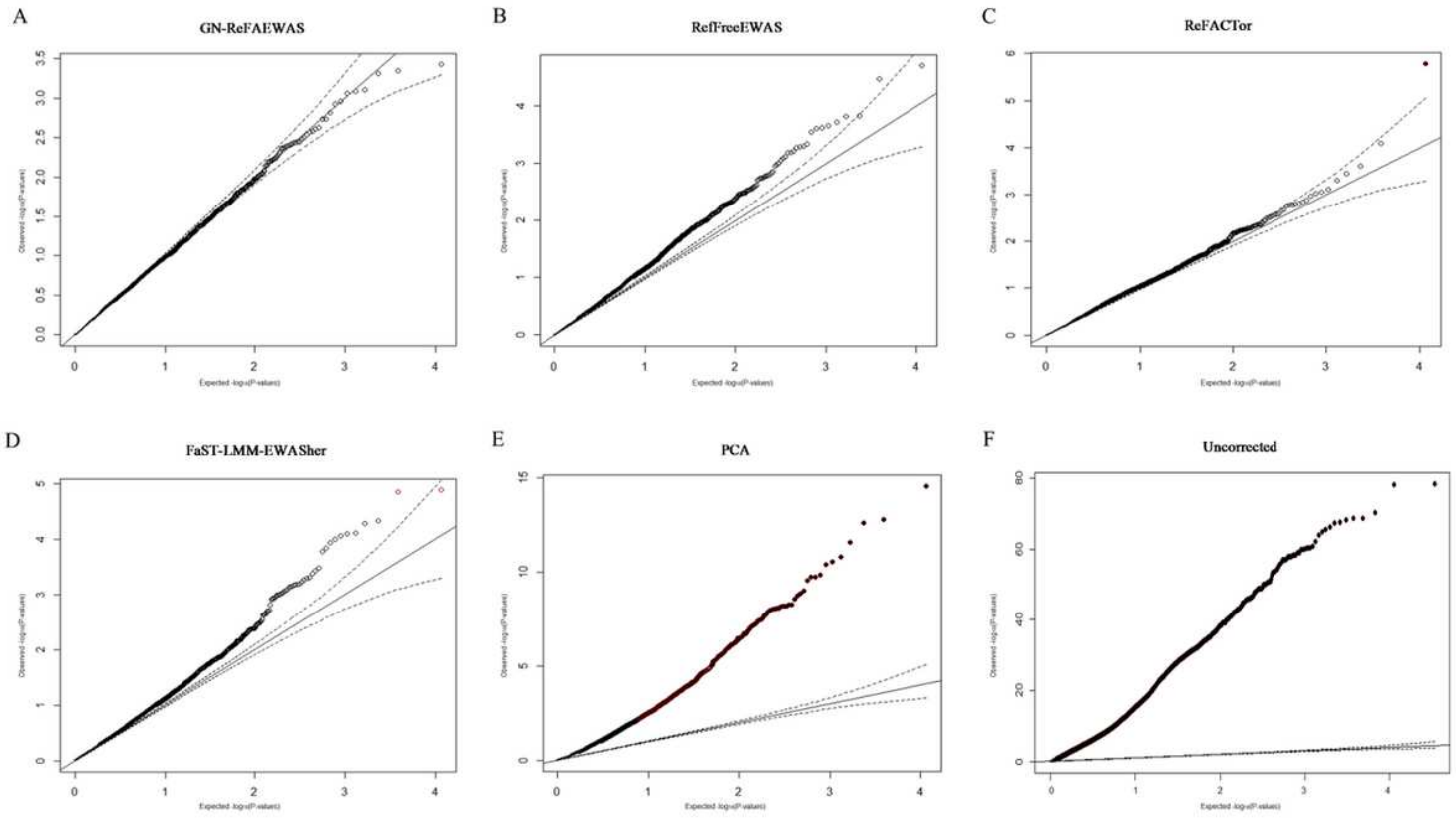

Figure 4

The correction result of the Tumor data set is represented by the QQ graph of the associated test. The degree of deviation from the benchmark indicates the degree of data expansion due to cell heterogeneity. The dotted line indicates the 95% confidence interval. (A) GN-ReFAEWAS model correction results (using the first PC) (b) ReFACTor model correction results (using the first PC) (c) RefFreeEWAS model correction results (d) FaST-LMM-EWASher model correction result, (e) PCA correction result (using the first PC), and (f) no correction result.

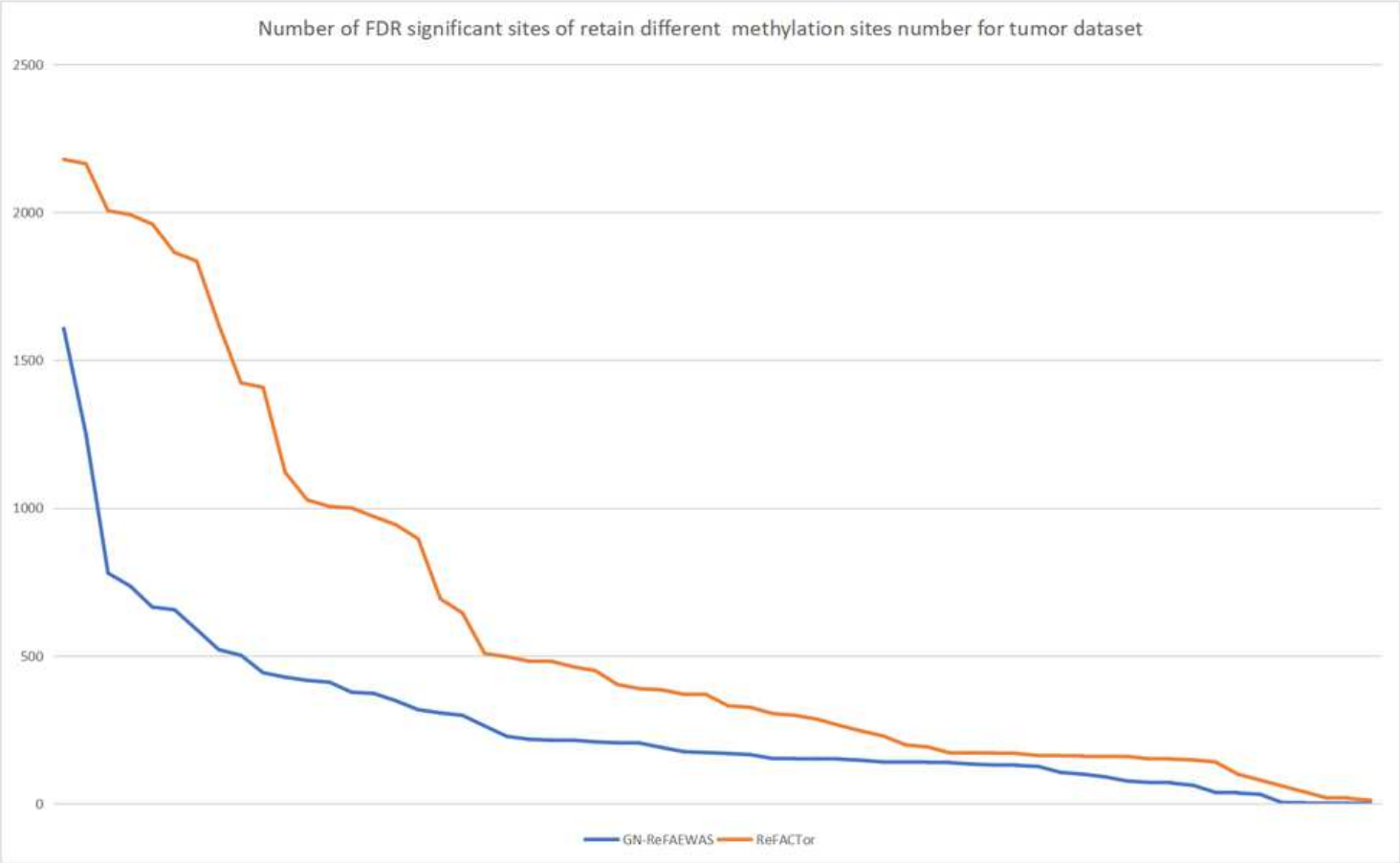

Figure 5

Number of FDR significant sites of retain different methylation sites number for tumor dataset with retaining different methylation sites number (0-600)

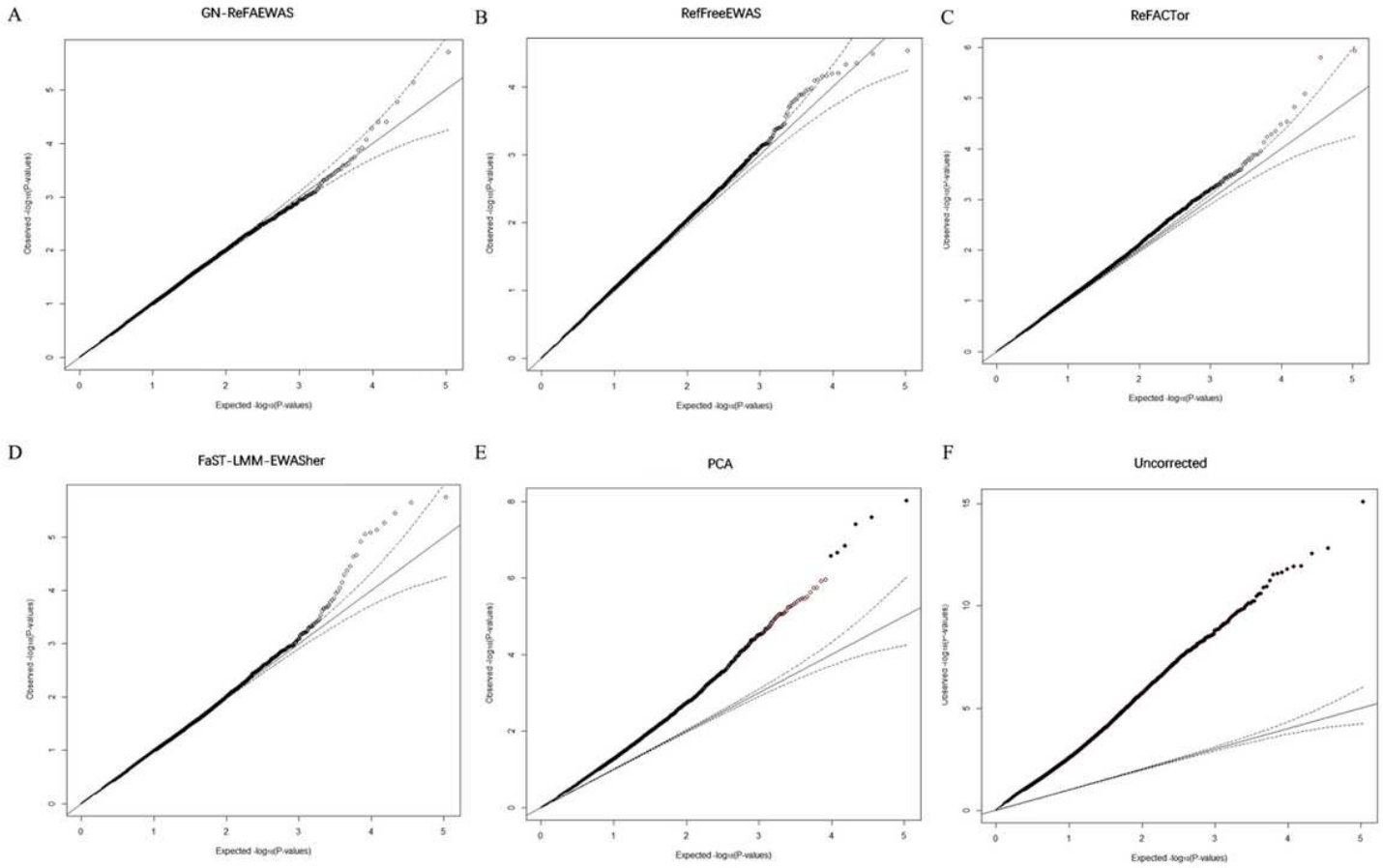

Figure 6

The correction result of the Mexican and Puerto Rican descent population data set is represented by the QQ graph of the associated test. The degree of deviation from the benchmark indicates the degree of data expansion due to cell heterogeneity. The dotted line indicates the 95% confidence interval. (A) GN-ReFAEWAS model correction results (using the first PC) (b) ReFACTor model correction results (using the first PC) (c) RefFreeEWAS model correction results (d) FaST-LMM-EWASher model correction result, (e) PCA correction result (using the first PC), and (f) no correction result.

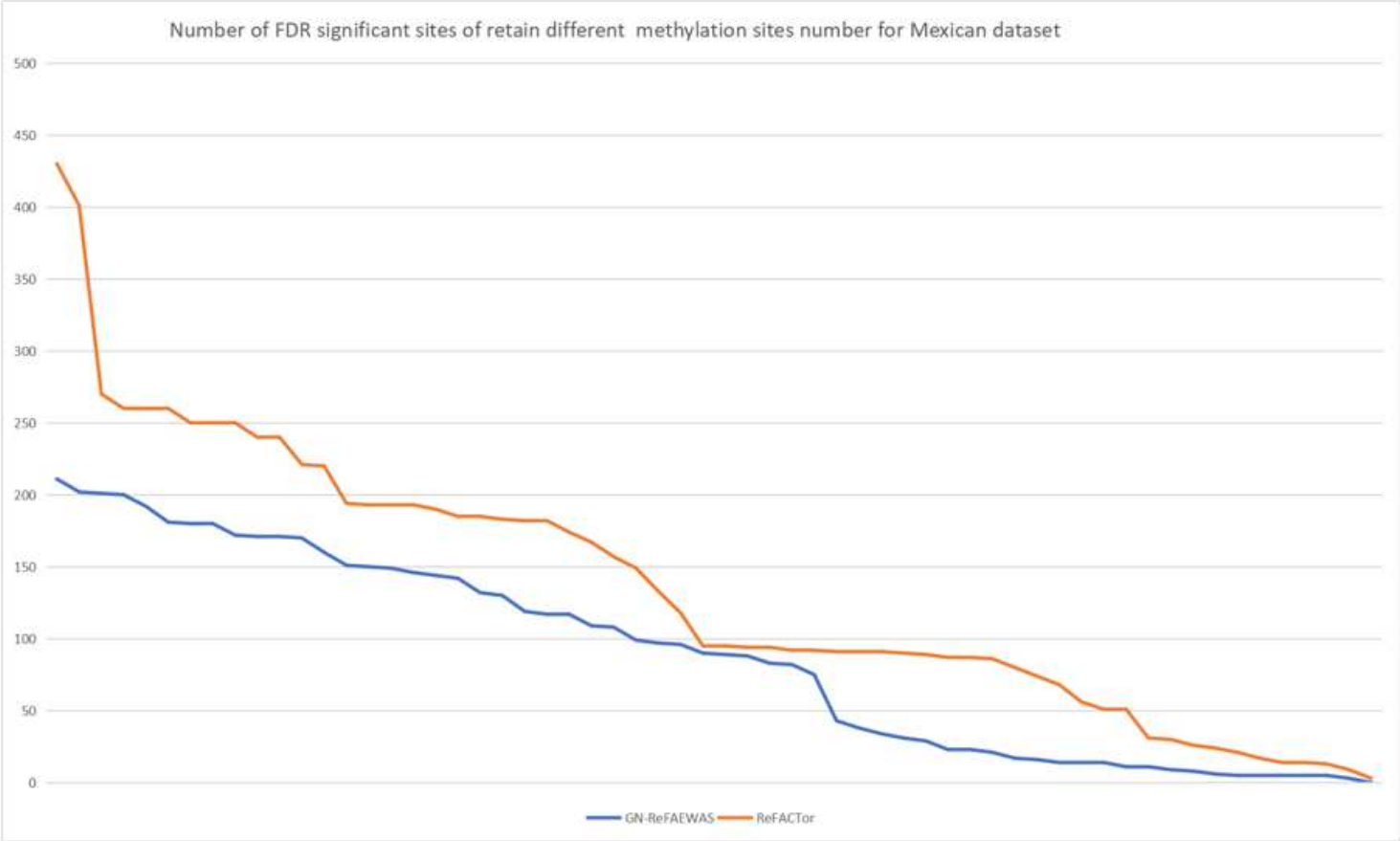

Figure 7

Number of FDR significant sites of retain different methylation sites number for Mexican dataset with retaining different methylation sites number (0-600)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS4NumberofFDRsignificantsitesMexican.xlsx
- TableS3NumberofFDRsignificantsitesTumor.xlsx
- TableS2NumberofFDRsignificantsitesRA.xlsx
- TableS1Thesamplesthathaveuncompletebloodcount.xlsx