Genetic Biomarkers for Bipolar Disorder

Biostatistics 5280 Final Project

Connor T. McNeill

April 26, 2023

Introduction

The dataset I will be analyzing for the project comes from a study (Catherine L Clelland 2013) from the GEO database (Tanya Barrett 2013) that looked at identifying genetic biomarkers for Bipolar Disorder. Bipolar Disorder, like many other mental illnesses, is currently diagnosed by a set of diagnostic markers contained in the DSM that must be meant when evaluated by a mental health professional. As such, there are two main issues that come up here. First, while the DSM attempts to create a uniform process for diagnosing patients, since it relies on the judgment of a medical professional, there is a risk of misdiagnosis - in fact, "69 percent of patients with bipolar disorder are misdiagnosed initially and more than one-third remained misdiagnosed for 10 years or more" (Singh and Rajput 2006). This misdiagnosis comes from a few different cases including comorbidity with other diagnoses and the fact that bipolar disorder diagnoses only happen with the occurrence of a manic (or hypomanic) episode, as that is a requirement for the diagnosis per the DSM. Many people who live with bipolar disorder are initially diagnosed with depression but are eventually diagnosed with bipolar disorder after a manic episode. If we were able to have a biological or empirical way of diagnosing bipolar disorder, there are two main advantages that would come with that. First, there would be less subjectivity with diagnoses, which would lessen the rate of misdiagnosis. Second, and more importantly, patients would be able to be treated early enough, potentially prior to having a manic episode or suicide attempt.

As such, the research question that I will be focusing on is identifying any potential genes that could serve as biomarkers for bipolar disorder and determining how well these biomarkers can predict whether someone has bipolar disorder. I will examine four statistical methods that could be utilized to answer this question and discuss which method would be best.

Methods

Pre-Processing

In order to perform any data analysis, the first required step is preprocessing. First, the raw data files were obtained from the GEO database (supplemental files for GSE46449). These files were then read in as .cel files using the `affy` `R` package. Then, a query into the GEO database was performed in order to get the phenotype and experiemental data. The dataset was then pre-processed using the Robust Multiarray Average (RMA) method with the `rma` function. First, background adjustment was performed on a raw intensity scale using the data from the data files.

Then, quantile normalization was performed using the PM intensities adjusted for background. A log transformation (base 2) was performed on the normalized data. Lastly, a median polish model was fitted $\log_2(PM_{jk}) = \mu + \alpha_k + \beta_j + e_{jk}$ where μ is the overall $\log_2$ probe set effect, α_k is the $\log_2$ probe set effect for the kth GeneChip, β_j is the $\log_2$ effect for probes $j = 1, ..., J$, and e_{jk} are the error terms. This model gives us the expression summaries that allow us to perform statistical analysis on the data. Lastly, filtering was done in two steps: first, the control probe sets (which started with "AFFX") where all removed, and second, genes with a mean expression less than 4 and IQR less than 0.5 were removed. No genes had a count of zero, which is why a value of 4 was utilized for filtering. This produced our resulting ExpressionSet object that was used to perform statistical analyses.

After preprocessing, the finalized dataset contained 6,437 genes from 88 samples. Table 1 below shows the breakdown of the samples: 49 of the samples came from patients with bipolar disorder, and 39 of the samples came from controls. While the original study further specified 6 samples coming from first-episode, never-medicated bipolar patients, these samples are just included with the general bipolar patient sample population for the sake of this analysis. Then, a random sample was taken from the samples in order to split the dataset into a training set and test set. Table 1 shows the split: it was confirmed prior to performing any methods that the ratio of bipolar patients to control subjects was approximately equal in both groups. For all of the four methods used, we will perform analysis on the transpose of the data so that the rows are the samples and the columns are the genes.

Set	**Bipolar Patients**	**Control Subjects**	**Total**
Training Set	37	29	66
Test Set	12	10	22
Total	49	39	88

Table 1: This table shows our breakdown of our samples between which patients have and did not have bipolar disorder, along with whether they are included in the training or test set.

Statistical Methods

The first method used was penalized logistic regression, specifically LASSO. LASSO is a penalized regression model with the goal of shrinking the number of parameters down, something that is important when our number of parameters is much greater than our sample size. This also performs model selection for us, selecting which predictors - in this case, which genes - do the best job at predicting whether a patient has bipolar disorder or not. This process is optimized by obtaining a value of λ that maximizes the AUC (area under the curve) using the independent test set. The following equation is used to estimate the model coefficients

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where the last term is our constraint term. This constraint term will shrink our model coefficients to zero for predictors that are not significant. The number of predictors selected is determined based on the optimized lambda value. In `R`, we first fitted the LASSO model using the `glmnet` function. Then, 10-fold cross-validation was used on the resulting set of lambda values with the `crossval` function in order to determine which value of lambda maximized the AUC. This value of lambda was selected and then used to (1) identify the significant predictors and (2) predict who in the test set has bipolar disorder.

The second method used was K-nearest neighbors. K-nearest neighbors (k-NN) is a supervised learning approach based of the Bayes decision rule (estimating $P(\omega = k|\mathbf{x})$ directly). More specifically, it looks at the $k = 5$ closest points to the point in order to determine whether a value is in a certain class - in this case whether a patient has bipolar disorder or not. Mathematically, we first determine the posterior probability $\hat{P}_{knn}(\omega = j|\mathbf{x}) = \frac{1}{k_N} \sum_{i=1}^{N} w_i(\mathbf{x}) z_{ij}$ and takes the value of j (either 0 or 1 in this case) that has a maximum posterior probability. To implement k-NN in `R`, we simply use the `knn` function from the `class` library, and then use the resulting object to predict who in the test set has bipolar disorder.

The third method used was a classification tree. A classification tree is a set of binary questions that are used in order to determine whether the samples is from a patient with or without bipolar disorder. Since we will be using continuous predictors, the binary question is based on whether a certain variable is greater than a cutoff value. This starting classification tree is determined using the `rpart` function from the `rpart` package. Next, the complexity parameter plot is generated - this helps us determine how we can prune the classification tree to stop splitting at a certain point. The tree is then pruned based on the optimal complexity parameter which makes the tree simpler and a better predictor. Lastly, the tree is then tested with the test set to get classification predictions on who has bipolar disorder.

The final method used was a random forest. There is a random component to this method - hence why we need to set the seed prior to running the `randomForest` function from the `randomForest` library. As the name suggests, a random forest is a bootstrap aggregation of classification trees, except that each tree is based off only a random sample of size $\sqrt{p}$ of the predictors. A split is determined for each of the predictors, and then the best splits are used to build the tree that is maximally sized. At the end, we end up with a subset of $B = 5000$ trees that are then aggregated to determine which class the observations belong to. This is represented by the out-of-bag error estimate. A variable importance plot is also generated in order to determine which genes are the most important in classification. Lastly, we will test the random forest with our test set to determine the accuracy of who in the test set has bipolar disorder.

Results

Our LASSO model was the first method utilized. Figure 1 shows how the number of coefficients included in the model shrinks as the value of lambda increases. As we see in the figure, the number of predictors included starts to decrease rapidly as lambda increases. Our optimal value of lambda is obtained by maximizing the AUC of the training set. The plot of the obtained AUC values by the lambda value that results in it is shown in Figure 2. The optimal value for lambda is 0.00424079, which produces an AUC of 0.8695. This shows that our model does extremely well at predicting whether a patient has bipolar disorder or not. The LASSO model retained 42 significant predictors. Looking at Table 2, our LASSO model correctly predicted each of the 22 samples in our test set with whether or not they have bipolar disorder.

True/Prediction	Bipolar	Control
bipolar patient	12	0
control subject	0	10

Table 2: This table shows our predicted classifications using the LASSO model.

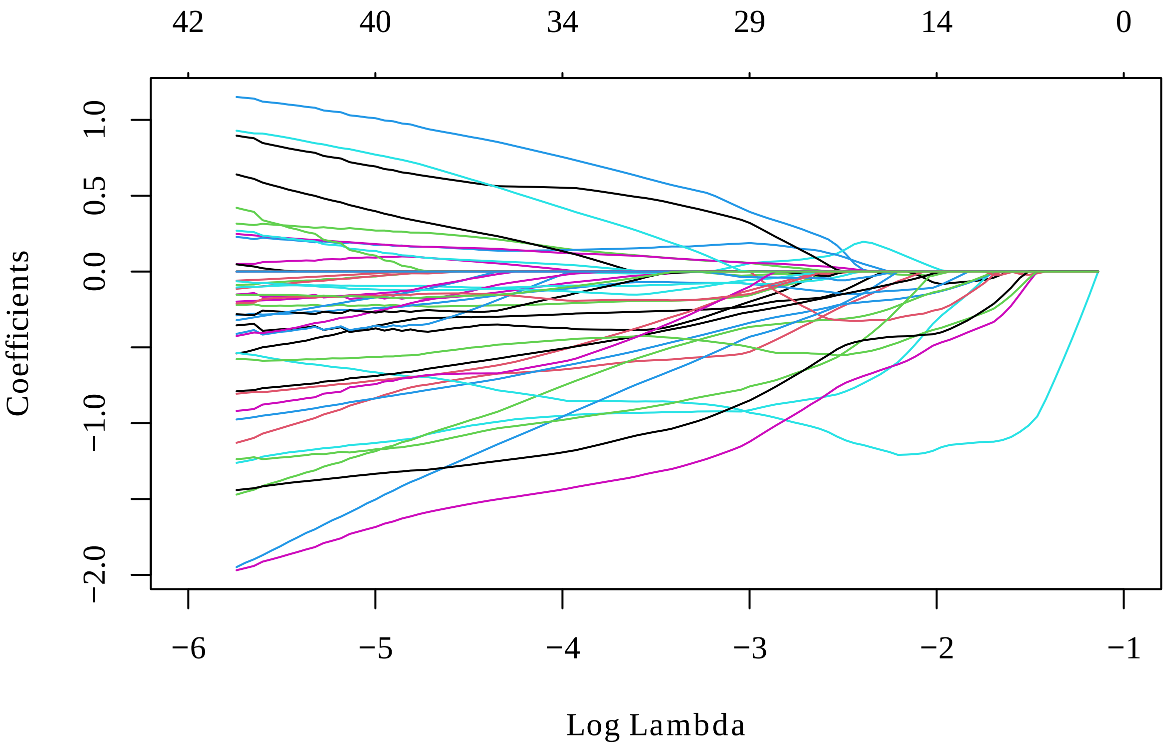

Figure 1: This figure shows the shrinkage.

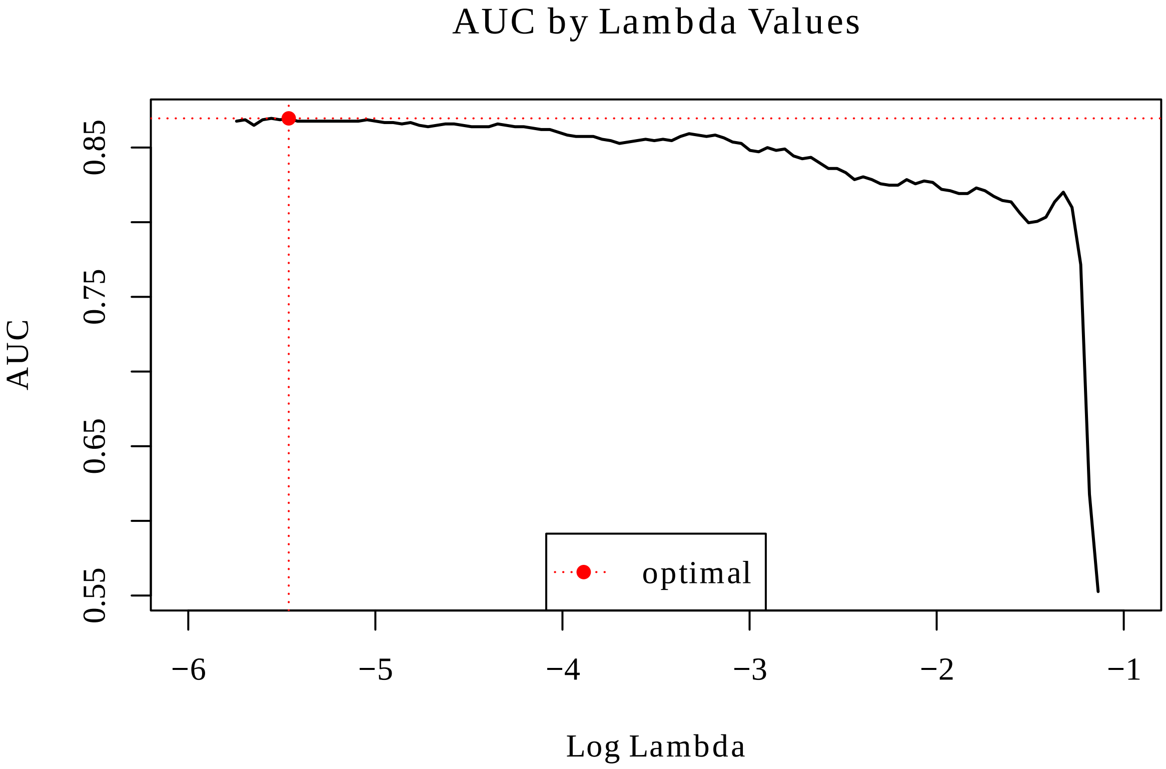

Figure 2: This plot shows the AUC for each value of lambda and shows the value of lambda that maximizes the AUC.

The next method utilized was k-Nearest Neighbors. A value of $k = 5$ was used with the `knn` function from the `class` package. Table 3 shows our classifications as predicted using `knn`. As we can see, there is a misclassification rate of 3/22, or 13.64%. It did not misclassify any control subjects, but it did misclassify 3 of the 12 patients that actually had bipolar disorder as not having the condition.

True/Prediction	Bipolar	Control
bipolar patient	9	3
control subject	0	10

Table 3: This table shows the classification of whether a patient in the test set has bipolar using k-NN.

Next we utilized the classification tree using the `rpart` function. Figure 3 shows the complexity parameter plot. We select the value of 0.3 to prune the classification tree. The pruned tree is shown in Figure 4. While the tree simplifies the classification to only one binary question, the error rate is quite high as shown in Table 4. Of the 22 samples in the test set, 7 are misclassified.

True/Predicted	Bipolar	Control
bipolar patient	10	2
control subject	5	5

Table 4: This table shows the classifications of whether a patient has bipolar disorder using our classification tree vs. the true values.

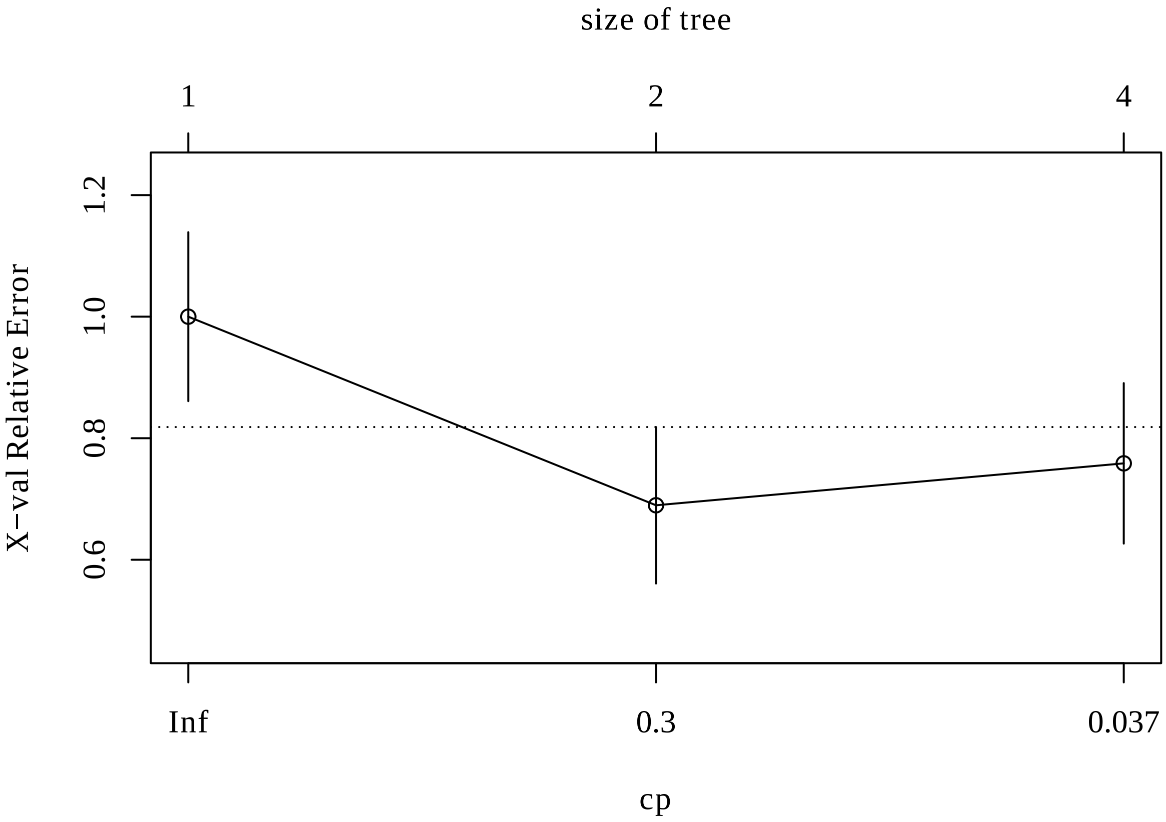

Figure 3: This figure shows the complexity parameter plot for the classification tree.

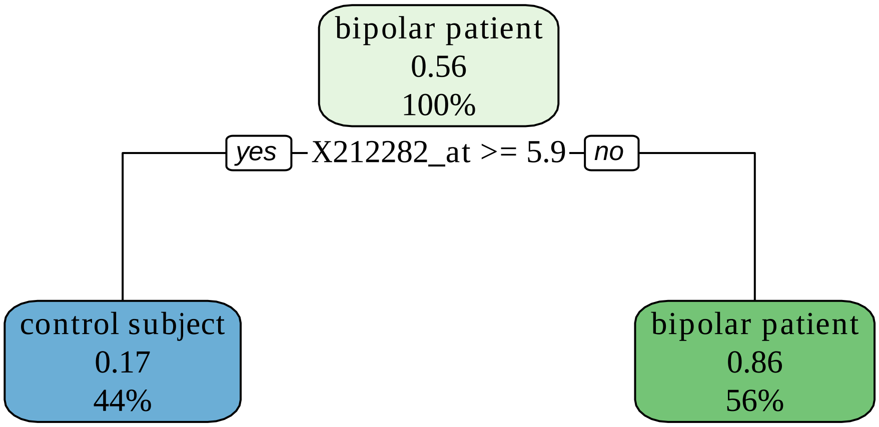

Figure 4: Above is the final, pruned classification tree.

The last method we used was random forest utilizing the `randomForest` function. 5000 trees were produced in the forest. The out-of-bag estimate of error rate was 22.73%. The important variables measured by Mean Decrease Accuracy is shown in Figure 5. The random forest performed better when performed on the test set, with an error rate of only 1/22, or 4.55%.

Test/Predicted	Bipolar	Control
bipolar patient	12	0
control subject	1	9

Table 5: The table above shows our classifications of whether or not a patient has bipolar disorder using our random forest vs. the true condition of the patient.

Conclusion

In conclusion, we examined four different supervised learning approaches to classifying whether a patient has bipolar disorder or not: LASSO (penalized regression), *k*-Nearest Neighbors, classification trees, and random forests. Overall, the LASSO model performed the best, misclassifying none of the test set samples. The random forest model performed well too, only misclassifying one of the subjects in the test set. *k*-Nearest Neighbors had a few false positives, but no false negatives, which means that the model performed decently well with classification. Lastly, the classification tree had the highest error rate, but was by-far the simplest approach as it narrowed the classification down to one particluar gene. If I were to select a method to utilize for classification, I would select the LASSO model due to its high accuracy. Additionally, it selected 42 genes that are potentially good biomarkers for identifying bipolar disorder. The results of this data analysis are promising that there could be a biological method of identifying bipolar disorder in patients, which could be a significant help in the diagnosis and treatment of the condition. However, more research needs to be done using more subjects and a higher sample size to determine whether these classification approaches could be effective in the general population.

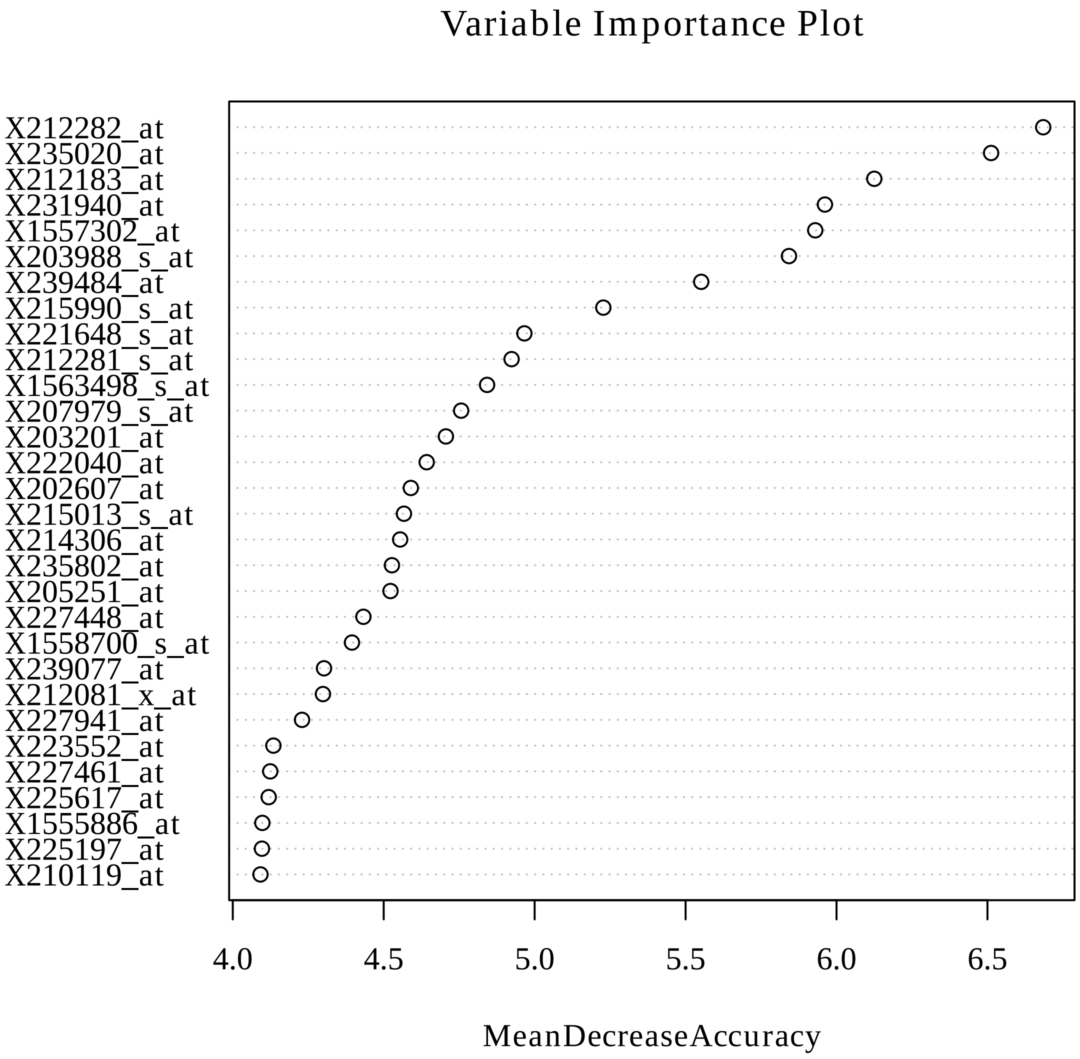

Figure 5: The above plot shows the important variables per the random forest.

References

Catherine L Clelland, L. J. P., Laura L Read (2013), “Utilization of never-medicated bipolar disorder patients towards development and validation of a peripheral biomarker profile,” *PLos One*, 8. https://doi.org/10.1371/journal.pone.0069082.

R Core Team (2022), *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Singh, T., and Rajput, M. (2006), “Misdiagnosis of bipolar disorder,” *Psychiatry*, 3, 57–63.

Tanya Barrett, P. L., Stephen E. Wilhite (2013), “NCBI GEO: Archive for functional genomics data sets–update,” *Nucleic Acids Research*, 41, D991–D995. https://doi.org/10.1093/nar/gks1193.