

SUPPLEMENTARY INFORMATION:

Eigen- R^2 for Dissecting Variation in High-dimensional Studies, by Lin S. Chen and John D. Storey

METHODS

Suppose that the high-dimensional biological data set is organized as an $m \times n$ matrix \mathbf{Y} , where the rows of \mathbf{Y} represent different response variables and the columns represent different observations. For example, in a gene-expression study, the rows of \mathbf{Y} are the m genes and the columns of \mathbf{Y} are the n arrays. Suppose also that an additional variable $\mathbf{z} = (z_1, z_2, \dots, z_n)^T$ has been measured. For example, \mathbf{z} could be a clinical variable, genotype, or treatment. If $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in})^T$ is the data for response variable i , then we may fit a model of \mathbf{y}_i on \mathbf{z} to obtain fitted values $\hat{\mathbf{y}}_i$. The proportion of variation in \mathbf{y}_i that is explained by \mathbf{z} is then:

$$R_{\hat{\mathbf{y}}_i}^2 = \frac{\hat{\sigma}_{\hat{\mathbf{y}}_i}^2}{\hat{\sigma}_{\mathbf{y}_i}^2} = \frac{\sum_{j=1}^n (\hat{y}_{ij} - \bar{y}_i)^2}{\sum_{j=1}^n (y_{ij} - \bar{y}_i)^2},$$

where \bar{y}_i is the mean of $\hat{\mathbf{y}}_i$ and \bar{y}_i is the mean of \mathbf{y}_i . It then follows that mean- R^2 is the average of these across all response variables, $\sum_{i=1}^m R_{\hat{\mathbf{y}}_i}^2 / m$.

Instead of taking the average of the $R_{\hat{\mathbf{y}}_i}^2$, we have developed an approach to employ principal components analysis to measure an overall R^2 . In principal components analysis, a singular value decomposition (SVD) is applied to the data matrix, decomposing \mathbf{Y} into the following: $\mathbf{Y} = \mathbf{UDV}^T$, where the matrices \mathbf{U} and \mathbf{V} are column orthogonal so that $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, and \mathbf{D} is a diagonal matrix. The columns of \mathbf{V} are the right eigenvectors and the columns of \mathbf{U} are the left eigenvectors. We are particularly interested in the right eigenvectors, because these represent aggregated trends in the response variables. Specifically, the first column of \mathbf{V} is the linear combination of all response variables that explains the most variation in the data, called the first right eigenvector. The second column of \mathbf{V} is the linear combination of all response variables that explaining the most variation in the data once the first eigenvector has been removed, and so on. The proportion of total variation captured by the i th eigenvector is $\pi_i = d_i^2 / \sum_{l=1}^L d_l^2$, where d_i is the eigenvalue of the i th eigenvector, which is obtained from the i th diagonal entry of \mathbf{D} .

When \mathbf{Y} is a matrix of expression data, the right eigenvectors have been called ‘‘eigen-genes.’’ Alter, Brown, & Botstein (2000) first proposed this terminology and showed that principal components analysis is a useful tool for identifying major trends in expression data. The top few eigen-genes tend to capture biologically relevant trends present in a number of genes. In a more general setting, one could call the columns of \mathbf{V} ‘‘eigen-response-variables.’’

Let \mathbf{v}_i be the i th column of \mathbf{V} and let $\hat{\mathbf{v}}_i$ be the fitted values when modeling \mathbf{v}_i in terms of \mathbf{z} . For each of these, we can calculate an R^2 value, denoted by $R_{\hat{\mathbf{v}}_i}^2$. Since π_i of the total variation in the data is explained by \mathbf{v}_i , $R_{\hat{\mathbf{v}}_i}^2$ should be weighted by π_i . Additionally, since each pair of eigen-response-variables is uncorrelated, the variation explained by \mathbf{z} in \mathbf{v}_i is orthogonal to the variation explained by \mathbf{z} in \mathbf{v}_j where $j \neq i$. Therefore, as an overall measure of R^2 , we proposed to take the average of the $R_{\hat{\mathbf{v}}_i}^2$, weighted

by their respective π_i :

$$\text{eigen-}R^2 = \sum_{i=1}^n \pi_i R_{\hat{\mathbf{v}}_i}^2.$$

PROPOSED ALGORITHM

The following algorithm summarizes the eigen- R^2 calculation:

Step 1. Let \mathbf{Y} be an $m \times n$ matrix, where the rows of \mathbf{Y} represent different response variables and have been mean-centered, and the columns represent different observations. Use singular value decomposition to decompose the matrix of response variables as $\mathbf{Y} = \mathbf{UDV}^T$.

Step 2. For each column of \mathbf{V} , denoted by \mathbf{v}_i , fit the user-specified model of \mathbf{v}_i on the independent variable(s) \mathbf{z} to obtain fitted values $\hat{\mathbf{v}}_i$, $i = 1, 2, \dots, n$. Calculate the R^2 value of this model fit as described above to obtain $R_{\hat{\mathbf{v}}_i}^2$.

Step 3. Calculate the proportion of variation explained by \mathbf{v}_i with $\pi_i = d_i^2 / \sum_{l=1}^L d_l^2$.

Step 4. Calculate the overall R^2 value as $\text{eigen-}R^2 = \sum_{i=1}^n \pi_i R_{\hat{\mathbf{v}}_i}^2$.

AN EQUIVALENCE RESULT

If the R^2 values are formed from model fits based on linear operators (which is usually the case when calculating R^2 values), then it can be shown that eigen- R^2 is equivalent to a weighted average the R^2 values calculated for each response individually. Specifically, the weights are given by the row specific variances.

Since the matrix \mathbf{Y} is row-centered, it follows that $\text{trace}(\mathbf{Y}\mathbf{Y}^T) = \sum_{j=1}^m n \hat{\sigma}_{y_j}^2$. The trace of the matrix $\mathbf{Y}^T\mathbf{Y}$ is the sum of squares of eigen-values, $\text{trace}(\mathbf{Y}^T\mathbf{Y}) = \text{trace}(\mathbf{VD}^T\mathbf{U}^T\mathbf{UDV}^T) = \sum_{l=1}^L d_l^2$. Therefore,

$$\text{trace}(\mathbf{Y}\mathbf{Y}^T) = \text{trace}(\mathbf{Y}^T\mathbf{Y}) = \sum_{l=1}^L d_l^2 = \sum_{j=1}^m n \hat{\sigma}_{y_j}^2, \quad (1)$$

Let $\hat{\mathbf{Y}}$ be the fitted matrix of \mathbf{Y} in terms of \mathbf{z} , so that each row of $\hat{\mathbf{Y}}$ is the fitted values for the corresponding row of \mathbf{Y} . Then, $\text{trace}(\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T) = \sum_{j=1}^m n \hat{\sigma}_{\hat{y}_j}^2$. Also, $R_{\hat{\mathbf{v}}_i}^2 = n \hat{\sigma}_{\hat{\mathbf{v}}_i}^2 = \hat{\mathbf{v}}_i^T \hat{\mathbf{v}}_i$. By assumption, we have that $\hat{\mathbf{v}}_i = \mathbf{H}\mathbf{v}_i$ and $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{H}$ for some matrix \mathbf{H} . (In using simple linear regression to estimate R^2 s for a variable \mathbf{z} , the fitted values $\hat{\mathbf{v}}_i = \mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{v}_i$. Let $\mathbf{H} = \mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T$. Then, for example, $\hat{\mathbf{v}}_i = \mathbf{H}\mathbf{v}_i$.) Since $\mathbf{Y} = \mathbf{UDV}^T$, we have

$$\begin{aligned} \text{trace}(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}) &= \text{trace}(\mathbf{H}^T\mathbf{VD}^T\mathbf{U}^T\mathbf{UDV}^T\mathbf{H}) \\ &= \text{trace}(\mathbf{H}^T\mathbf{VD}^T\mathbf{DV}^T\mathbf{H}), \quad \text{since } \mathbf{U}^T\mathbf{U} = \mathbf{I} \\ &= \text{trace}(\hat{\mathbf{V}}\mathbf{D}^2\hat{\mathbf{V}}^T), \end{aligned}$$

since \mathbf{H} is symmetric and \mathbf{D} is a diagonal matrix. The i th diagonal element of the diagonal matrix \mathbf{D}^2 is d_i^2 , and the i th diagonal of the

matrix $\widehat{\mathbf{V}}\widehat{\mathbf{V}}^T$ is $\widehat{\mathbf{v}}_i^T\widehat{\mathbf{v}}_i = R_{\widehat{\mathbf{v}}_i}^2$. Therefore, $\text{trace}(\widehat{\mathbf{Y}}^T\widehat{\mathbf{Y}}) = \sum_{i=1}^n d_i^2 \times R_{\widehat{\mathbf{v}}_i}^2$, and

$$\text{trace}(\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T) = \text{trace}(\widehat{\mathbf{Y}}^T\widehat{\mathbf{Y}}) = \sum_{j=1}^m n\widehat{\sigma}_{y_j}^2 = \sum_{i=1}^n d_i^2 \times R_{\widehat{\mathbf{v}}_i}^2. \quad (2)$$

The eigen- R^2 according to \mathbf{z} for a row-centered expression matrix is

$$\begin{aligned} \text{eigen-}R^2 &= \sum_{i=1}^n \frac{d_i^2}{\sum_{l=1}^L d_l^2} \times R_{\widehat{\mathbf{v}}_i}^2 \quad \text{by definition of eigen-}R^2 \\ &= \frac{\sum_{i=1}^n d_i^2 \times R_{\widehat{\mathbf{v}}_i}^2}{\sum_{l=1}^L d_l^2} \\ &= \frac{\sum_{i=1}^n d_i^2 \times R_{\widehat{\mathbf{v}}_i}^2}{\sum_{j=1}^m n\widehat{\sigma}_{y_j}^2} \quad \text{by Equation (1)} \\ &= \frac{\sum_{j=1}^m \widehat{\sigma}_{y_j}^2}{\sum_{j=1}^m \widehat{\sigma}_{y_j}^2} \quad \text{by Equation (2)} \quad (3) \\ &= \sum_{j=1}^m \frac{\widehat{\sigma}_{y_j}^2}{\widehat{\sigma}_{y_j}^2} \times \frac{\widehat{\sigma}_{y_j}^2}{\sum_{k=1}^m \widehat{\sigma}_{y_k}^2} \\ &= \sum_{j=1}^m w_j R_{y_j}^2, \quad \text{where } w_j = \frac{\widehat{\sigma}_{y_j}^2}{\sum_{k=1}^m \widehat{\sigma}_{y_k}^2}. \end{aligned}$$

Therefore, using models that can be fitted by applying a linear operator (e.g. ordinary least squares regression, least squares regression with spline functions), eigen- R^2 can be written as a weighted average of the feature-specific $R_{y_i}^2$ values, where the weights correspond to the response variable's contribution to the overall distribution of baseline variances. It also follows from this equivalence that if the rows of \mathbf{Y} are not only mean-centered, but also scaled to have unit standard deviation, then eigen- R^2 is equal to mean- R^2 .

COMPARING DIFFERENT R^2 MEASURES

The R package introduced here offers another weighting scheme, which involves an extra step of de-noising the right eigenvectors. We call this quantity "de-noised eigen- R^2 ." After performing the SVD of \mathbf{Y} (Step 1), we form \widehat{k} , which is the number of statistically significant eigen-vectors at a predefined threshold (Leek and Storey, 2007). (A statistically significant right eigen-vector is defined to be one that explains more variation than would be expected if there were no structure among the response variables.) We then only sum the $R_{\widehat{\mathbf{v}}_i}^2$ s over the top \widehat{k} significant eigen-vectors. The de-noised eigen- R^2 is defined as:

$$\text{de-noised eigen-}R^2 = \sum_{i=1}^{\widehat{k}} \pi_i R_{\widehat{\mathbf{v}}_i}^2,$$

where $\pi_i = d_i^2 / \sum_{l=1}^n d_l^2$. Compared with eigen- R^2 , de-noised eigen- R^2 puts zero weights on the $(n - \widehat{k})$ non-significant eigen-vectors that are estimated to be not capturing signals in \mathbf{Y} .

An alternative de-noising weighting scheme can be formed by using an empirical Bayesian approach. We call this version a

Sample size	oracle- R^2 (%)	eigen- R^2 (%)	de-noised eigen- R^2 (%)	mean- R^2 (%)	Bayesian- R^2 (%)
20	13.74	13.30	13.28	6.63	10.45
50	11.90	12.16	12.15	6.86	11.28
100	15.27	15.21	15.20	8.44	15.00

Table S1. Comparing oracle- R^2 , eigen- R^2 , de-noised eigen- R^2 , mean- R^2 and Bayesian- R^2 . All the R^2 estimates have been adjusted for sample size effect.

"Bayesian- R^2 ." Using equation (3), we replace $\widehat{\sigma}_{y_j}^2$ with

$$\begin{aligned} &\Pr(\sigma_{y_j}^2 = 0 | \mathbf{Y}, \mathbf{z}) \times 0 + \Pr(\sigma_{y_j}^2 > 0 | \mathbf{Y}, \mathbf{z}) \times \widehat{\sigma}_{y_j}^2 \\ &= \Pr(\sigma_{y_j}^2 > 0 | \mathbf{Y}, \mathbf{z}) \times \widehat{\sigma}_{y_j}^2. \end{aligned}$$

In this way, a Bayesian R^2 is calculated as:

$$\text{Bayesian-}R^2 = \frac{\sum_{j=1}^m \Pr(\sigma_{y_j}^2 > 0 | \mathbf{Y}, \mathbf{z}) \times \widehat{\sigma}_{y_j}^2}{\sum_{k=1}^m \widehat{\sigma}_{y_k}^2}$$

In order to estimate the posterior probabilities, we employ the empirical Bayes algorithm proposed in Storey *et al.* (2005).

We simulated three gene expression data sets with sample sizes $n = 20, 50$ and 100 in order to illustrate the performances of the different aggregate R^2 measures discussed here. In each data set, we simulated 500 non-expressed genes (so that the expression values represent low magnitude noise), 500 expressed genes unassociated with \mathbf{z} , and 500 expressed genes associated with \mathbf{z} . We calculated eigen- R^2 , mean- R^2 , de-noised eigen- R^2 , and Bayesian- R^2 for each data set (Table S1). We compared this to the oracle- R^2 , which we defined as the population level aggregated R^2 equal to $\sum_{j=1}^m \sigma_{y_j}^2 / \sum_{j=1}^m \sigma_{y_j}^2$. From Table S1, we can see that as sample size increases, differences among all the aggregate R^2 measures grow smaller. Eigen- R^2 and denoised eigen- R^2 estimates are similar and are close to the oracle- R^2 . The Bayesian- R^2 is comparable and relatively conservative. It should be noted that the computation for eigen- R^2 is substantially simpler than de-noised eigen- R^2 and Bayesian- R^2 .

GENETIC DISSECTION OF TRANSCRIPTIONAL VARIATION

We performed linkage analysis of the expression levels similarly to Brem *et al.*, 2002. At a P -value cutoff of 5×10^{-7} , about 9,000 out of the approximately 19 million of gene-marker combinations show significant linkage. Figure S1a plots the significantly linked gene expression trait positions against the linked marker positions on the yeast genome. We estimated the eigen- R^2 and mean- R^2 values at each locus, shown in Figure S1b. It can be seen from these plots that both quantities capture the linkage hotspots well. However, eigen- R^2 tends to capture more signal, particularly on Chromosomes II, III, XII, XIII, and XIV. The *MAT* mating locus on Chromosome III, which has been shown to play an important role in gene expression (Brem *et al.*, 2002), has the highest eigen- R^2 of 8.1%, while the mean- R^2 at that locus is 1.9% (Figure S1c).

REFERENCES

- Leek, J. T. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by "surrogate variable analysis". *PLoS Genetics*, **3**, e161.
- Storey, J. D., Akey, J. M., and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, **3**(8), e267.

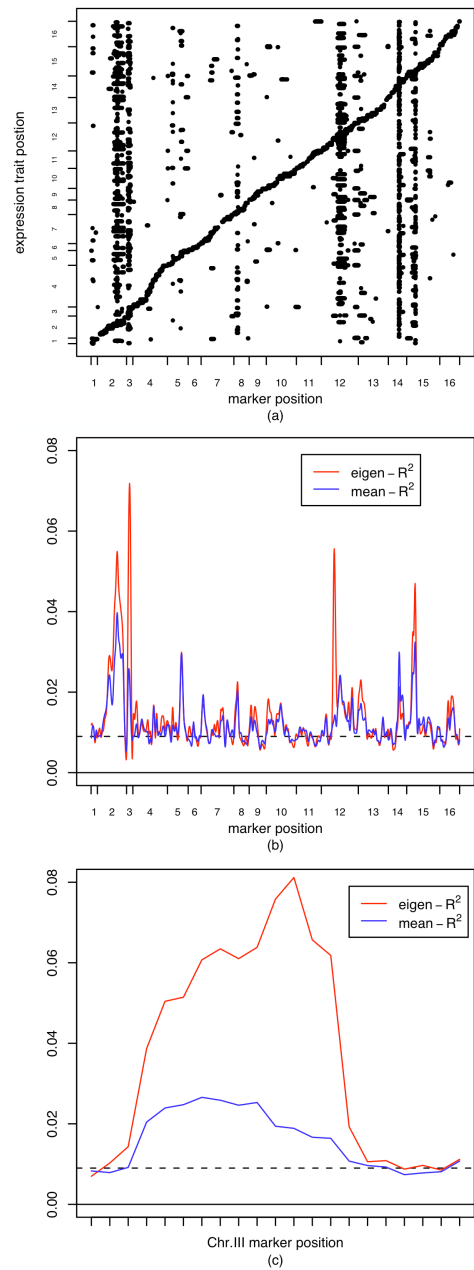


Fig. S1. (a) Tests for linkage among all gene expression trait and marker pairs. A dot indicates significant linkage. (b) Genome-wide comparison of eigen- R^2 and mean- R^2 values. The dashed horizontal line shows the expected R^2 value under no linkage signal. (c) Eigen- R^2 and mean- R^2 values on Chromosome III, which contains the important *MAT* locus.