

SUPPLEMENTAL MATERIAL

Transite: A computational motif-based analysis platform that identifies RNA-binding proteins modulating changes in gene expression

Konstantin Krismer^{1,2,3,5,7}, Shohreh Varmeh^{2,3}, Molly A. Bird^{2,3,5}, Anna Gattinger^{3,7}, Yi Wen Kong^{2,3}, Thomas Bernwinkler^{2,3,7}, Daniel A. Anderson^{4,5}, Andreas Heinzl⁷, Brian A. Joughin^{2,3,5}, Ian G. Cannell^{2,3,8,*} and Michael B. Yaffe^{2,3,5,6,*}

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA

²Center for Precision Cancer Medicine, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02139, USA

³David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, 500 Main Street, Cambridge, MA 02139, USA

⁴Synthetic Biology Center, Massachusetts Institute of Technology, 500 Technology Square, Cambridge, MA 02139, USA

⁵Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

⁶Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA

⁷Department for Medical and Bioinformatics, University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria

⁸New York Genome Center, 101 Avenue of the Americas, New York, NY 10013, USA

*To whom correspondence should be addressed. Tel: +1 617 452 2103; Fax: +1 617 452 4978; Email: myaffe@mit.edu. Correspondence may also be addressed to Ian G. Cannell. Tel: +1 646 977 7241 ; Email: icannell@nygenome.org

1 Motif representations

Position specific scoring matrices (PSSM) are used to represent sequence motifs. Transite inherits PSSMs describing RBP binding sites from two sources (see section on motif databases in the main text). Motif databases provide PSSMs in one of three types: Position frequency matrices (PFM), position probability matrices (PPM), or position weight matrices (PWM). Internally, Transite algorithms work exclusively with PWMs in order to make subsequent calculations more efficient.

The elements of a PFM represent absolute count of each nucleotide at each position. PPMs and PWMs are derived from PFMs as follows:

$$\begin{array}{c} \text{PFM} \\ \left(\begin{array}{ccccc} & \text{A} & \text{C} & \text{G} & \text{U} \\ 1 & x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ 2 & x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n & x_{n,1} & x_{n,2} & x_{n,3} & x_{n,4} \end{array} \right) \xrightarrow{f_1} \end{array}
 \begin{array}{c} \text{PPM} \\ \left(\begin{array}{ccccc} & \text{A} & \text{C} & \text{G} & \text{U} \\ 1 & y_{1,1} & y_{1,2} & y_{1,3} & y_{1,4} \\ 2 & y_{2,1} & y_{2,2} & y_{2,3} & y_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n & y_{n,1} & y_{n,2} & y_{n,3} & y_{n,4} \end{array} \right) \xrightarrow{f_2} \end{array}
 \begin{array}{c} \text{PWM} \\ \left(\begin{array}{ccccc} & \text{A} & \text{C} & \text{G} & \text{U} \\ 1 & z_{1,1} & z_{1,2} & z_{1,3} & z_{1,4} \\ 2 & z_{2,1} & z_{2,2} & z_{2,3} & z_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n & z_{n,1} & z_{n,2} & z_{n,3} & z_{n,4} \end{array} \right) \end{array} \quad (1)$$

The conversion functions f_1 (from PFM to PPM) and f_2 (from PPM to PWM) are applied to each element of the matrix. From the PFM, containing counts of each nucleotide j at each sequence position i , a row normalization converts nucleotide counts at each position to nucleotide probabilities:

$$f_1(\mathbf{x}, i, j) = \frac{x_{i,j}}{\sum_k x_{i,k}}, \quad (2)$$

where \mathbf{x} is a PFM, and i and j are its indices. PPM elements are converted to PWM elements by taking a log ratio of the actual element-wise probabilities against the probability of the average nucleotide at the same position:

$$f_2(\mathbf{y}, i, j) = \log_2 \frac{y_{i,j}}{p_j} \quad (3)$$

where \mathbf{y} is a PPM, and p_j is the *a priori* probability of nucleotide j . In Transite, nucleotides are assumed to be equiprobable ($\Pr(\text{A}) = \Pr(\text{C}) = \Pr(\text{G}) = \Pr(\text{U}) = 0.25$).

Laplace smoothing: Laplace smoothing (also known as additive smoothing) is applied to avoid zeros in PFMs (and thereby zeros in PFMs and negative infinite values in PWMs). Zeros might occur if the number of sequences on which the PSSM is based, is too small to contain at least one occurrence of each nucleotide per position. In this case, pseudocounts are introduced [1]. Specifically, 0.25 was added to all cells (i.e., raw counts) of PFMs which had at least one raw count of zero.

Scoring algorithm: A given k -mer from a candidate transcript sequence may be scored by comparison to a PWM of length k . In order to calculate the score, the corresponding PWM weight for each nucleotide at each position of the candidate k -mer is summed.

Under an assumption of position-wise independence of nucleotides in their contribution to RBP-binding fitness, a random k -mer will have a score of zero, whereas positive and negative scores will correspond to higher and lower than average fitness, respectively.

2 Monte Carlo sampling

Permutation tests are a means to determine the statistical significance of a test statistic with an unknown null distribution. Since no assumptions are made about the underlying distribution of the statistic, permutation tests belong to the group of non-parametric tests. The null distribution of the statistic is obtained empirically by calculating all possible values of the statistic by rearrangement of the labels of the observations (data points). Each unique ordering of the labels is called a permutation, hence the name. Labels are categorical variables that subdivide the set of observations into groups, e.g., *treatment* and *control*. In order to build the complete empirical sampling distribution of the test statistic T based on n labeled observations, T needs to be calculated for $n!$ permutations of the observation labels. The

upper tail probability of the actual test statistic, i.e., the test statistic T calculated with the actual observations x , here denoted $T(x)$, is given as follows:

$$Pr(T(x)) = \sum_{y:T(y) \geq T(x)} Pr(y), \quad (4)$$

where y are the permuted observations.

Since the number of permutations grows factorially with the number of observations, calculating T for all permutations is infeasible even for small n . Therefore, instead of building the complete null distribution, a sample of the distribution is picked randomly to determine an estimate of the probability of $T(x)$. This process is called Monte Carlo sampling. The estimate is determined by the empirical cumulative distribution functions (lower-, upper- and two-tailed probability):

$$\hat{Pr}_L(T(x)) = \frac{\sum_{i=1}^n \mathbb{1}(T(y_i) \leq T(x)) + 1}{n + 1} \quad (5)$$

$$\hat{Pr}_U(T(x)) = \frac{\sum_{i=1}^n \mathbb{1}(T(y_i) \geq T(x)) + 1}{n + 1} \quad (6)$$

$$\hat{Pr}_T(T(x)) = \frac{\sum_{i=1}^n \mathbb{1}(|T(y_i)| \geq |T(x)|) + 1}{n + 1}, \quad (7)$$

where $\mathbb{1}$ is the indicator function and n is the sample size, i.e., the number of performed permutations.

One is added to both the numerator and the denominator to avoid p-values of zero when the actual test statistic is smaller than all of the test statistics of the permuted data [2].

A confidence interval around \hat{p} , i.e., $\hat{Pr}(T(x))$, can be calculated based on the cumulative probabilities of the binomial distribution. This interval is referred to as Clopper-Pearson interval [3]. The exact confidence limits c_l and c_u satisfy the following equations:

$$\sum_{i=n_1}^n \binom{n}{i} c_l^i (1 - c_l)^{n-i} = \alpha/2 \quad (8)$$

$$\sum_{i=0}^{n_1} \binom{n}{i} c_u^i (1 - c_u)^{n-i} = \alpha/2, \quad (9)$$

where n_1 is the number of cases where $T(y_i) \geq T(x)$ (see equation 6). If $n_1 = 0$, the lower confidence limit is 0, whereas if $n_1 = n$, the upper limit is 1.

Applications in Transite: Monte Carlo sampling is used to obtain an estimate of the significance of local consistency scores (see section 6.1 for details) in SPMA spectrum classification. In this case, n , the number of local consistency scores that are sampled from the null distribution is 10,000,000. In TSMA, the null distribution of motif enrichment values (in matrix-based TSMA, $n = 2,000,000$) and k -mer enrichment values (in k -mer-based TSMA, $n = 50,000$) is obtained by randomly sampling foreground sets from the background set without replacement and recalculating motif and k -mer enrichment values, respectively.

Early stopping: In order to significantly reduce the execution time of the Monte Carlo sampling procedure without reducing the number of permutations (where they matter), the tests are implemented with an early stopping mechanism. If the observed test statistic is deemed not significant after a certain number of samples from the null distribution (in Transite, this decision is made after 5000 samples), the null distribution sampling stops.

3 Combining enrichment of motif-associated k -mers

The overall enrichment of a motif for k -mer TSMA is calculated as the geometric mean of the enrichments of associated k -mers:

$$\bar{e} = \exp\left(\frac{1}{n} \sum_{i=1}^n \log(e_i)\right), \quad (10)$$

where \mathbf{e} is the vector of enrichment values of motif-associated k -mers. The sum of logarithms is used instead of the product to avoid arithmetic underflow.

4 Methods for combining p-values

The following section describes methods to combine the significance (p-values) of enrichment values of a set of k -mers that are associated with an RNA-binding protein. These methods are used to obtain a single p-value for the overall significance of enriched or depleted RBP-associated k -mers.

In general, the methods of this section can be applied to combine the results of independent significance tests. They are commonly used in meta-analysis, where the goal is to systematically assess and integrate findings of a number of studies about a common body of research.

The problem can be specified as follows: Given a vector of n p-values p_1, \dots, p_n , find p_c , the combined p-value of the n significance tests. Most of the methods introduced here combine the p-values in order to obtain a test statistic, which follows a known probability distribution. The general procedure can be stated as:

$$T(h, C) = \sum_{i=1}^n h(p_i) * C \quad (11)$$

The function T , which returns the test statistic t , takes two arguments. h is a function defined on the interval $[0, 1]$ that transforms the individual p-values, and C is a correction term.

4.1 Fisher’s method

Fisher’s method (1932) [4], also known as the inverse chi-square method is probably the most widely used method for combining p-values. Fisher used the fact that if p_i is uniformly distributed (which p-values are under the null hypothesis), then $-2 \log p_i$ follows a chi-square distribution with two degrees of freedom. Therefore, if p-values are transformed as follows,

$$h(p) = -2 \log p, \quad (12)$$

and the correction term C is neutral, i.e., equals 1, the following statement can be made about the sampling distribution of the test statistic T_f under the null hypothesis:

$$t_f \stackrel{H_0}{\sim} \chi_{2n}^2, \quad (13)$$

where n is the number of p-values.

4.2 Stouffer’s method

Stouffer’s method [5], or the inverse normal method, uses a p-value transformation function h that leads to a test statistic that follows the standard normal distribution by transforming each p-value to its corresponding normal score. The correction term scales the sum of the normal scores by the root of the number of p-values.

$$h(p) = \Phi^{-1}(1 - p) \quad (14)$$

$$C = \frac{1}{\sqrt{n}} \quad (15)$$

$$t_s \stackrel{H_0}{\sim} N(0, 1), \quad (16)$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function.

An extension of Stouffer’s method with weighted p-values is called Lipták’s method [6].

4.3 Mudholkar and George’s method

The logit method by Mudholkar and George [7] uses the following transformation:

$$h(p) = -\ln(p/(1 - p)) \quad (17)$$

When the sum of the transformed p-values is corrected in the following way:

$$C = \sqrt{\frac{3(5n+4)}{\pi^2 n(5n+2)}}, \quad (18)$$

the test statistic t_m is approximately t-distributed:

$$t_m \stackrel{H_0}{\sim} t_{5n+4} \quad (19)$$

4.4 Edgington's method

Edgington's method [8] is an additive procedure to combine p-values.

$$h(p) = p \quad (20)$$

The sampling distribution of the test statistic t_e under the null hypothesis is given by combinatorics:

$$Pr(t_e) = \sum_{r=0}^{\lfloor t_e \rfloor} (-1)^r \binom{n}{r} \frac{(t_e - r)^n}{n!} \quad (21)$$

4.5 Tippett's method

In Tippett's method [9] the smallest p-value is used as the test statistic t_t and the combined significance is calculated as follows:

$$Pr(t_t) = 1 - (1 - t_t)^n \quad (22)$$

5 Methods for adjusting p-values

When multiple statistical tests are performed in order to identify non-random events in a large pool of events, it is imperative to adjust either the p-values themselves or the cutoff significance level α , which is the probability of making a type I error (incorrectly rejecting the null hypothesis). Failure to do so leads to alpha error accumulation, i.e., many false positives.

Without accounting for alpha error accumulation in the k -mer enrichment step, the enrichment values of for example 204/4096 possible hexamers would be deemed significant between randomly chosen sets of sequences (assuming $\alpha = 0.05$). This is a direct consequence of the number of tests (4096 in this case) and the accepted probability of making a wrong decision per test (α).

Transite supports several methods to adjust p-values in order to avoid the multiple testing problem, all of which take a vector of p-values $\mathbf{p} \in [0, 1]^n$ and return a vector of adjusted p-values $\mathbf{q} \in [0, 1]^n$. The i th smallest or largest p-value is denoted by $p_{(i)}$, depending on whether the method belongs to the step-down (ordered from lowest to highest) or step-up (highest to lowest) group. The methods can be categorized according to the definition of type I error they control.

5.1 Familywise error rate controlling methods

The familywise error rate (FWER) is defined as

$$FWER = Pr(V > 0), \quad (23)$$

where V is the number of false positives in n tests (i.e., "the family").

Methods controlling the FWER guarantee that $FWER \leq \alpha$.

5.1.1 Holm's method

The adjusted p-values [10] obtained by Holm's method are defined as

$$q_{(i)} = \max_{j \leq i} \left(\min \left((n - j + 1)p_{(j)}, 1 \right) \right), \quad (24)$$

where $p_{(j)}$ is the j th lowest p-value and thus characterizing Holm's approach as a step-down method.

5.1.2 Hochberg's method

Hochberg's method is the step-up version of Holm's method ($p_{(i)}$ is highest p-value) and is uniformly more powerful [11].

$$q_{(i)} = \begin{cases} p_{(n)} & \text{for } i = n \\ \min(q_{(i+1)}, (n - i + 1)p_{(i)}) & \text{otherwise} \end{cases} \quad (25)$$

5.1.3 Bonferroni's method

Bonferroni corrected p-values [12] are given by

$$q_i = \min(p_i * n, 1). \quad (26)$$

It is the oldest and most conservative correction.

5.2 False discovery rate controlling methods

The false discovery rate (FDR) is defined as

$$FDR = E \left(\frac{V}{V + S} \right), \quad (27)$$

where V is the number of false positives and S the number of true positives in n tests.

Methods controlling the FDR are less conservative than the ones controlling the FWER.

5.2.1 Benjamini and Hochberg's method

Similar to Hochberg's method for controlling the familywise error rate, this method is defined as a step-up adjustment [13]:

$$q_{(i)} = \begin{cases} p_{(n)} & \text{for } i = n \\ \min(q_{(i+1)}, \frac{n}{i}p_{(i)}) & \text{otherwise} \end{cases} \quad (28)$$

Compared to the FWER controlling method, the multiplier is less conservative ($\frac{n}{i}$ to $n - i + 1$), leading to smaller adjusted p-values. This method can be used if the components (i.e., p-values) of \mathbf{p} are independent and uniformly distributed.

5.2.2 Benjamini and Yekutieli's method

If there are dependencies among the p-values or if independency cannot be guaranteed, Benjamini and Yekutieli's method [14] can be used instead:

$$q_{(i)} = \begin{cases} \gamma p_{(n)} & \text{for } i = n \\ \min(q_{(i+1)}, \gamma \frac{n}{i} p_{(i)}) & \text{otherwise} \end{cases} \quad (29)$$

where $\gamma = \sum_{i=1}^n \frac{1}{i}$.

6 Classification of spectrum plots

Two methods were developed to identify non-random spectrum plots, a local consistency score that quantifies the changes between neighboring bins, and a method based on polynomial regression models.

6.1 Local consistency score

The local consistency score quantifies the local noise of the gradient in the spectrum by calculating the deviance between the linear interpolation of the scores of two bins separated by one other, and the score of the middle bin, for each position in the spectrum. The lower the score, the more consistent the trend in the spectrum plot. Formally, the local consistency score x_c is defined as

$$x_c = \frac{1}{n} \sum_{i=1}^{n-2} \left| \frac{s_i + s_{i+2}}{2} - s_{i+1} \right|. \quad (30)$$

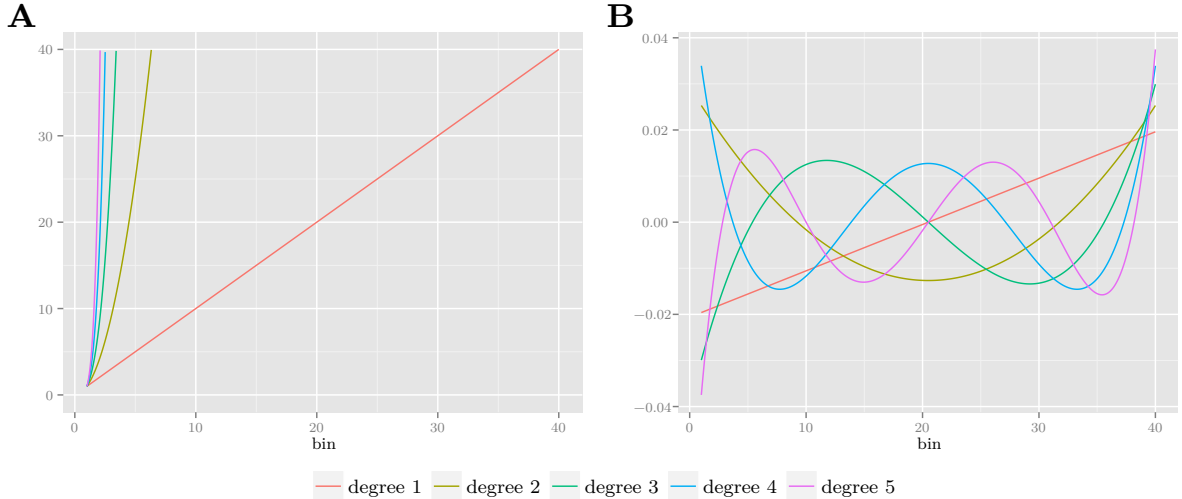


Figure 1: **Ordinary and orthogonal polynomials:** (A) The ordinary polynomials of degrees 1 to 5 are highly correlated. Moreover, polynomials of high degree can lead to floating point underflow of model coefficients. (B) Correlation between orthogonal polynomials is strongly reduced.

In order to obtain an estimate of the significance of a particular score x'_c , Monte Carlo sampling is performed by randomly permuting the coordinates of the scores vector \mathbf{s} and recomputing x_c . The probability estimate \hat{p} is given by the lower tail version of the cumulative distribution function (see equation 5), where T equals x_c in the equation above.

6.2 Polynomial regression

An alternative approach to assess the consistency of a spectrum plot is via polynomial regression. In a first step, polynomial regression models of various degrees are used to fit \mathbf{s} , the vector of scores, as a function of \mathbf{b} , the vector of bin numbers. Then the model that reflects best the true nature of the data is selected by means of the F-test. Finally, the adjusted R^2 are calculated to indicate how well the model fits the data. These statistics are used as scores to rank the spectrum plots.

In general, the polynomial regression equation is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_m x_i^m + \epsilon_i, \quad (31)$$

where m is the degree of the polynomial (usually $m \leq 5$), and ϵ_i is the error term. The dependent variable \mathbf{y} is the vector of scores \mathbf{s} and \mathbf{x} to \mathbf{x}^m are the orthogonal polynomials of the vector of bin numbers \mathbf{b} .

Orthogonal polynomials are used in order to reduce the correlation between the different powers of \mathbf{b} and therefore avoid multicollinearity in the model. This is important, because correlated predictors lead to unstable coefficients, i.e., the coefficients of a polynomial regression model of degree m can be greatly different from a model of degree $m + 1$.

The orthogonal polynomials of vector \mathbf{b} are obtained by centering (subtracting the mean), QR decomposition, and subsequent normalization [15].

Given the dependent variable \mathbf{y} and the orthogonal polynomials of \mathbf{b} \mathbf{x} to \mathbf{x}^m , the model coefficients β are chosen in a way to minimize the deviance between the actual and the predicted values. Ordinary least squares is used as the estimation method for the model coefficients. After polynomial models of various degrees have been fitted to the data, the F-test is used to select the model that best fits the data. After a model has been selected, the adjusted R^2 is calculated as an additional way to evaluate the goodness of fit.

7 Transite analysis reports

Detailed information on the Transite analysis runs described in the main manuscript are available online:

- [matrix-based Transcript Set Motif Analysis report on data set GSE7880](#)

- [k-mer-based Transcript Set Motif Analysis report on data set GSE7880](#)
- [matrix-based Spectrum Motif Analysis report on data set GSE7880](#)
- [k-mer-based Spectrum Motif Analysis report on data set GSE7880](#)

References

- [1] Nishida, K., Frith, M. C., and Nakai, K. (Feb, 2009) Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.*, **37**(3), 939–944.
- [2] Phipson, B. and Smyth, G. K. (2010) Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*, **9**, Article39.
- [3] Clopper, C. J. and Pearson, E. S. (1934) The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, **26**(4), 404–413.
- [4] Fisher, R. (1932) *Statistical methods for research workers*, Edinburgh Oliver & Boyd, .
- [5] Stouffer, S., DeVinney, L., and Suchmen, E. (1949) *The American soldier: Adjustment during army life*, Princeton University Press, .
- [6] Lipták, T. (1958) On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, pp. 171–196.
- [7] Mudholkar, G. S. and George, E. O. (1979) The logit method for combining probabilities. In *Symposium on optimizing methods in statistics* pp. 345–366.
- [8] Edgington, E. S. (1972) An Additive Method for Combining Probability Values from Independent Experiments. *The Journal of Psychology*, **80**(2), 351–363.
- [9] Tippett, L. (1931) *The methods of statistics*, Williams and Norgate, .
- [10] Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.
- [11] Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**(4), 800–802.
- [12] Dunn, J. and Dunn, O. J. (1961) Multiple Comparisons Among Means. *American Statistical Association*, pp. 52–64.
- [13] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.
- [14] Benjamini, Y. and Yekutieli, D. (08, 2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**(4), 1165–1188.
- [15] Chambers, J., Hastie, T., and Pregibon, D. (1990) Statistical Models in S. In *Compstat* Springer pp. 317–321.