Appendix to "A nonparametric approach to detect nonlinear correlation in gene expression" published in the Journal of Computational and Graphical Statistics

Y. Ann Chen[1] Jonas S. Almeida[2] Adam J. Richards[3] Peter Müller[4] Raymond J. Carroll[5] Baerbel Rohrer[6]

Appendix

A. A Permutation Test and $\widehat{D}_o$ for Local Correlation $\ell(r)$.

The permutation procedure to generate the simulated null distribution $\widehat{D}_o$ for (1) was as follows. Each independent permutation replication $z^\star$ consisted of $(N \times 1)$ vectors $x^\star$ and $y^\star$. Vectors $x^\star$ and $y^\star$ were randomly permuted from the observed x and y, respectively. The star notation distinguishes the permutation replication from the actual observations. A total of B permutation replications was independently generated: $z^\star(b)$, $b = 1, \ldots, B$. In this study, $B = 4000$ was used. The summaries $\widehat{D}^\star(r, b)$ and $\ell^\star(r, b)$ were then evaluated corresponding to each permutation replication $z^\star(b)$. The median neighbor density of the null distribution generated by B replications of permutation is used as $\widehat{D}_o(r)$ in (1). A two-sided p-value can be evaluated using $\ell^\star(r, b)$ in (2).

A one-sided p-value is approximated for each of the other two statistics, maximal local correlation (M) and mutual information (MI). Permutation replications of $S^\star$ ($S = M$

[1](corresponding author) Department of Biostatistics, Moffitt Cancer Center, Tampa, FL, USA, ann.chen@moffitt.org

[2]Department of Bioinformatics and Computational Biology, The University of Texas, M.D. Anderson Cancer Center, Houston, TX, USA

[3]Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, SC, USA

[4]Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, TX, USA

[5]Department of Statistics, Texas A & M University, College Station, TX, USA

[6](corresponding author) Departments of Ophthalmology and Neurosciences, Medical University of South Carolina, Charleston, SC, USA, rohrer@musc.edu

or MI) are evaluated corresponding to each permutation replication $z^\star(b)$. The p-value is evaluated as

$$p(S) = \#\{S^\star(b) > S\}/B, \tag{A.1}$$

where $S = M$, or MI. The statistical significance of Pearson and Spearman correlation coefficients is also estimated using permutations. Let C denote Pearson Correlation coefficient. It can take either positive or negative value, and its two-sided achieved significance level can be approximated by

$$p(C) = \#\{|C^\star(b)| > |C|\}/B. \tag{A.2}$$

B. A Permutation Test for the δ Statistics

The order of genes at each time point was randomly permuted from the observed expression data (Fig. 5). A total number of B permutation replications were generated: $z^*(b), b = 1, \ldots, B$. Distance matrices $d_{ij}^{rd\star}$ and $d_{ij}^{wt\star}$ were evaluated for each permutation replicate $z^*(b)$. As described in Appendix A, for each pair of genes (i, j), we evaluated $S^{rd\star}(b)$ and $S^{wt\star}(b)$ ($S = M, C$ or MI) between the i-th and j-th columns of the distance matrices $d_{ij}^{rd\star}$ and $d_{ij}^{wt\star}$. Each corresponding $\delta_S^\star$ ($S = M, C$ or MI) between wt and rd mice was then evaluated using (4). All computed δ_S values are exchangeable across permutations and across genes, leaving us with imputed δ_S values $\delta_S^\star(b')$, $b' = 1, \ldots, B^\star$ with $B^\star = B \cdot 16290 = 24 \cdot 16290 = 390960$ (we used $B = 24$ permutations). We controlled the false discovery rate (FDR) to adjust for multiple hypothesis testing. The two-sided achieved significant level can be approximated by

$$p(\delta_S^\star(b') = \#\{|\delta_S^\star(b')| > |\delta_S|\}/B^\star. \tag{B.1}$$

Corresponding q-values were estimated using software Q-value (Storey and Tibshirani, 2003; Storey, 2002). Statistical significance was declared when controlling FDR at 0.05 for each of the δ statistics. A FDR of 0.05 means that 5% of the features that are reported as significant are on average (repeated sampling average) false positives. The frequency of significant association change was counted for each gene. The importance of the identified

candidate genes was ranked based on the frequency of significant association changes using each δ_S. Mean ranks were listed in the presence of tied ranks.

C. Mutual Information

We use similar notation as in Daub et al. (2004). With a finite set of M possible states $\{a_1, a_2, , a_{M_A}\}$ for a system A, the Shannon entropy $H(A)$ is defined as $H(A) = -\sum_{n=1}^{M_A} p(a_n) \log p(a_n)$. Similarly, the joint entropy $H(A, B)$ of two systems A and B is $H(A, B) = -\sum_{n,m} p(a_n, b_m) \log p(a_n, b_m)$. Mutual information (MI) is defined as $MI(A, B) = H(A) + H(B) - H(A, B)$. When A and B are independent, $MI(A, B) = 0$. The default setting for the R function in Daub et al. (2004) uses the widely-used binning method to bin continuous data into M discrete intervals a_n, $n = 1, \ldots, M_A$.

D. Implementation and Visualization of the Transcriptional Regulatory Network

A transcriptional regulatory network was generated using transcription factor binding sites (TFBS) of the promoter regions of the analyzed genes, independently of expression levels. The network provides a means to validate the top-ranked genes selected by the δ statistics. Each set of top ranked genes generated by each statistic was used as the "seed" to search automatically for the shortest paths between these "seed genes" using networkX libraries (https://networkx.lanl.gov/wiki). The genes are connected if they are potentially regulated by the same transcription factors. All the connections between genes and transcription factors are assumed to be weighted equally. The subnetworks generated automatically for each of the delta statistic are referred to as subnetwork-M, subnetwork-MI, subnetwork-C, respectively, based on the seeding origin.

The implementation details are as follows. Promoter sequences (1,000 bps upstream of the transcription start site, and 300 bps downstream) of the genes were retrieved from the Cold Spring Harbor Laboratory Mammalian Promoter Database (Matys et al., 2006), and scanned for transcription factor binding sites (TFBS) using the tool, MatchTM (Kel et al.,

2003), against the TRANSFAC (public version) database (Matys et al., 2006). TFBS could be obtained for 142 genes out of the 181 genes selected for analysis. The TFBS results were assembled into graph objects using a customized Python script and the NetworkX libraries. With a Python interface, PyGraphviz, networks of the genes and 62 transcription factors are visualized using the Graphviz package (http://www.graphviz.org/).

E. Computational Complexity

We summarize the computational complexity of the main steps of the method as below. The complexity of ranking variable followed by linear transformation is $O(NlogN)$. Calculation of all pair-wise neighboring distance is $O(N^2)$. Estimation of the correlation integral, which records the cumulative numbers of neighbors within a neighborhood of radius r is $O(m \cdot N^2)$. Although in our paper, for the demonstration purpose, we use the sample size N as the default choice for m, however, empirically, with $m = 10$ or 20, it already yields very detailed neighboring patterns along the radius r. When applied on a large dataset, such as a microarray dataset, we will set m as a small constant. That is, the complexity for estimating the correlation integral is $O(N^2)$. To facilitate the discussion of the computational complexity for smoothing using the automatic smoother (Vilela et al., 2007) used in our paper, let d denote the order of differences as an input of the smoother, and λ denote the smoothing parameter (larger λ gives smoother results). With the default number order of differences set as a small constant (of 5), and the default λ as a vector with 31 levels (the number of levels is also a constant). The computational complexity depends on the grid size m for the Cholesky factorization (with the complexity of $O(m^3)$). When applying on a large dataset, we will set m as a small constant as described above, and the computational complexity for the smoother is $O(C)$. Therefore, the overall complexity for our proposed method is $O(N^2)$.

References

Daub, C., Steuer, R., Selbig, J., and Kloska, S. (2004), "Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data," *BMC Bioinformatics*, 5, 118.

Kel, A., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2003), "MATCHTM: a tool for searching transcription factor binding sites in DNA sequences," *Nucl. Acids Res.*, 31, 3576–3579.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006), "TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes," *Nucl. Acids Res.*, 34, D108–110.

Storey, J. D. (2002), "A direct approach to false discovery rate," *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

Storey, J. D. and Tibshirani, R. (2003), "Statistical significance for genomewide studies," *PNAS*, 100, 9440–9445.

Vilela, M., Borges, C. C. H., Vinga, S., Vasconcelos, A. T. R., Santos, H., Voit, E. O., and Almeida, J. S. (2007), "Automated smoother for the numerical decoupling of dynamics models," *BMC Bioinformatics*, 8, 305, fully automated perfect smoother for numerical decoupling of multivariate, dynamic, S-System models.