Application of Biostatistics and Bioinformatics Tools to Identify Putative Transcription Factor-Gene Regulatory Network of Ankylosing Spondylitis and Sarcoidosis

Dongseok Choi[1*], Srilakshmi M Sharma[2], Sirichai Pasadhika[2], Zhixin Kang[6], Christina A. Harrington[5], Justine R. Smith[2,3], Stephen R. Planck[2-4], James T. Rosenbaum[2-4]

[1]Division of Biostatistics, Department of Public Health & Preventive Medicine, [2]Casey Eye Institute, [3]Department of Cell & Developmental Biology, [4]Department of Medicine, [5]Gene Microarray Shared Resource, Oregon Health & Science University, Portland, OR 97239, USA

[6]Department of Economics, Finance, and Decision Sciences, University of North Carolina at Pembroke, NC 28372, USA

*Corresponding author:

Dongseok Choi

Oregon Health & Science University,

3181 SW Sam Jackson Park Road, CB669

OR 97239, USA

e-mail: choid@ohsu.edu

Tel.503-494-5336

Fax.503-494-4981

Short title: Predicting Transcription Factor – Gene Regulatory Network

Revised: March 3, 2008

Abstract

Transcription factors and corresponding *cis*-regulatory elements are considered key components in gene regulation. We combined biostatistics and bioinformatics tools to streamline identification of putative transcription factor-gene regulatory networks unique for two immune-mediated diseases, ankylosing spondylitis and sarcoidosis. After identifying differentially expressed genes from microarrays, we employed tightCluster to find tight clusters of potentially co-regulated genes. By subsequently applying bioinformatics tools to search for common *cis*-regulatory elements, putative transcription factor-gene regulatory networks were found. Recognition of these networks by applying this methodology could pave the way for new insights into disease pathogenesis.

Introduction:

An oligonucleotide microarray experiment can provide rich information on gene expression patterns by simultaneously screening tens of thousands of transcripts from relatively small samples of solid tissue or blood. It has been used successfully to identify genes related to diseases such as lymphoma (Hans, et al., 2004), melanoma (Dai, et al., 2007), and breast cancer (Driouch, et al., 2007). Transcription factors and corresponding *cis*-regulatory elements, also called motifs, are considered key components in regulating gene expression patterns but continue to remain elusive because the binding sites are short sequence segments, scattered widely over the genomic non-coding regions, and these motifs are difficult to locate using conventional approaches (Antes, et al., 2000; Barton, et al., 2001; Bonifer, 2000; Lee, et al., 2004; Michelson, 2002; Pennacchio and Rubin, 2001). With the complete sequence information for many eukaryotic genomes available and a publicly available online database of transcription factors, we can search for potential *cis*-regulatory elements more efficiently by combining biostatistics and bioinformatics methods for a group of co-regulated genes.

Since it is typical to detect hundreds or thousands of differentially expressed genes in a microarray experiment, it is essential to reduce the dimensions before applying any bioinformatics tools. Clustering analysis, an unsupervised learning method, has been crucial in reducing the dimensions of microarray data analyses (Choi, et al., 2006; Fachin, et al., 2007; Nikpour, et al., 2007). Traditional clustering methods such as hierarchical clustering (Kaufman and Rousseeuw, 1990) and *k*-means clustering (Hartigan and Wong, 1978) have been used for detecting groups of genes that show similar expression profiles. However, these traditional clustering methods assign every gene to a cluster and require subjective decisions with regard to the number of clusters. This number may be difficult to predict when investigating complex biological processes although one can be guided by some post-hoc criterion that are based on homogeneity within clusters. Forcing all data points into a predetermined number of clusters could increase the possibility that a cluster contains multiple subclusters of genes that are regulated by different transcription factors. This might in turn increase noise and make identification of unique transcription factors for co-regulated genes difficult. A new generation of clustering algorithms has been proposed in recent years to overcome these difficulties. These new methods do not require a pre-determined number of clusters nor do they force all genes to be grouped. Some examples are the tightClust (Tseng and Wong, 2005), split (Pena, et al., 2003), model-based clustering (mClust) (Fraley and Raphael, 2002), and dynamic tree cut (Langelder, et al., 2008). The tightClust algorithm largely consists of two parts; a tight clustering step and a sequential search for stable clusters based on bootstrapping. One unique feature of tightClust is including natural ranks of tightness of clusters by implementing a sequential search in the algorithm. The split and mClust are based on a Gaussian mixture model for clustering. The split employs the jackknife-style prediction for clustering while the mClust uses the EM algorithm. The dynamic tree cut methods are novel algorithms that can detect clusters in a dendrogram from hierarchical clustering. For more details, refer to the original publication for each algorithm.

This study demonstrated a streamlined analysis for predicting putative transcription factor-gene regulatory networks by first utilizing one of the new clustering algorithms, tightClust (Tseng and Wong, 2005), to find tight clusters of differentially expressed genes in peripheral blood samples of ankylosing spondylitis and sarcoidosis patients and control groups, and then applying a bioinformatics tool, the Promoter Analysis and Interaction Network Toolset (PAINT)

(Vadigepalli, et al., 2003) to identify putative transcription factor-gene regulatory networks based on 2000 upstream base pairs from the transcription start sites of genes in selected tight clusters.

Rheumatologists care for patients who suffer from inflammatory diseases induced by the immune system. Ankylosing spondylitis is among the most common diseases in most rheumatologic practices. Ankylosing spondylitis and its closely related conditions, including reactive arthritis, psoriatic arthritis, undifferentiated spondyloarthropathy, and the arthritis associated with inflammatory bowel disease affect greater than 1% of the population by many estimates. (Sieper, et al., 2006) Although ankylosing spondylitis classically involves the sacroiliac joints, it can also affect other organs. For example, 40% of these patients develop eye inflammation, specifically anterior uveitis, which is synonymous with iritis. (Brophy, et al., 2001) Sarcoidosis is less common in most rheumatologic practices because pulmonary disease is usually the predominant manifestation. Sarcoidosis, however, is an excellent comparator to ankylosing spondylitis because it is also immunologically mediated. In North America, ankylosing spondylitis and sarcoidosis rank first and second respectively as the two most common systemic, immune-mediated diseases associated with uveitis. (Rosenbaum, 1989) Thus far, there has been no report comparing gene expression profiles for these two diagnoses.

Microarray experiment and identifying significantly expressed genes:

Peripheral blood samples were taken from recruited patients with ankylosing spondylitis or sarcoidosis attending a tertiary clinic at the Oregon Health & Science University (OHSU) for patients with uveitis, rheumatic disease, or lung disorders. Since medications can markedly affect gene expression, all blood samples were drawn from patients who were not receiving systemic corticosteroids or any sort of systemic immunomodulatory therapy at the time of blood draw. The control group was recruited from patients attending an ophthalmology clinic for routine eye care. These controls did not have a history of autoimmune disease, were not on oral corticosteroids or any immunomodulatory therapy, and had no evidence for uveitis on eye examination. There were 11 patients with ankylosing spondylitis and 12 patients with sarcoidosis and 12 controls. The study received approval by the OHSU institutional review board.

The microarray experiment was performed by the Affymetrix Microarray Core in the OHSU Gene Microarray Shared Resource. RNA was amplified and labeled using the GeneChip Globin Reduction Protocol rev. 1 (Affymetrix, Inc., PreAnalytiX). Then, 10 µg of each labeled target were hybridized with a Human Genome U133 Plus 2.0 array (Affymetrix, Inc.) using standard protocols as described in the GeneChip Expression Analysis manual (www.affymetrix.com/support/technical/manual/expression_manual.affx). The U133 Plus 2.0 array contains over 54,000 probe sets for 47,000 human transcripts and variants. After hybridization, cell fluorescence intensity (CEL) files were imported into the R statistical language environment (R_Development_Core_Team, 2007) for normalization. We used the GC robust multi-array analysis (GC-RMA) developed by Wu and co-workers (Wu, et al., 2004) to correct background noise in perfect match (PM) probe data. Background corrected data were further normalized with the rank invariant probes proposed by Li and Wong (Li and Wong, 2001). The gene expression levels were summarized using a linear model estimated by the median polish algorithm. (Irizarry, et al., 2003) After normalization, we applied the significance analysis of microarrays (SAM) (Tusher, et al., 2001). This method is appropriate when the numbers of samples in a data set are small. In this method, testing yields the q-values (Storey and Tibshirani, 2003) for statistical significance, which are similar to the false discovery rate (FDR) (Benjamini and Hochberg, 1995) that controls for the expected ratio of false positives among those genes that exhibit significant levels of expression rather than for the false positives among non differentially expressed genes. For the computations, we used R and its add-on packages; “affy”, “gcrma” and “samr”.

Out of approximately 47,000 transcripts, SAM identified 5,139 transcripts as being statistically significantly different between groups with $q < 0.01$ based on robustified F-statistics implemented in SAM (Tusher, et al., 2001). This implies that about 13% of genes were differentially expressed taking multiple probe sets targeting the same genes into account. There were 487 transcripts that were upregulated only in the ankylosing spondylitis group, 881 transcripts upregulated only in the sarcoidosis patients, and 11 transcripts upregulated in both diseases. Some examples of differentially expressed genes are interleukin 1 receptor type I (IL1R1), interleukin 1 receptor type II (IL1R2) and interferon (alpha, beta and omega) receptor 1 (IFNAR1) upregulated in the ankylosing spondylitis group and signal transducer and activator of transcription (STAT1), intercellular adhesion molecule 1 (ICAM1) and interferon regulatory factor 5 (IRF5) upregulated in the sarcoidosis patients.

Descriptive comparisons of clustering algorithms:

To find tight patterns in 5,139 significantly changed transcripts, we employed the tightClust (Tseng and Wong, 2005), split (Pena, et al., 2003), model-based clustering (mClust)(Fraley and Raphael, 2002), and dynamic tree cut (Langelder, et al., 2008). For all methods, we decided to use default parameters due to the wide diversity of algorithms. For the tightClust, we tried three different target numbers of clusters; 25, 50 and 100. In addition, we tried both dynamic tree methods proposed by Langelder (Langelder, et al., 2008).

Table 1 shows descriptive statistics of clustering results. All algorithms except mClust grouped about 44% ~ 76% of data in 7 ~ 98 clusters.   Note that we combined data from multiple probe sets within a transcript for some algorithms. From Table 1, tightClust with target number of clusters 50 and 100, tightClust(50) and tightClust(100), returned smaller clusters than other methods. Note that the mClust generates quite different clustering results depending on whether the original data is standardized or not. Further, it seems that the mClust procedure only obtains a few clusters. Regarding the split procedure, it turns out that some genes with opposite direction are complied into the same cluster. For dynamic tree cut, the hierarchical dendrograms are highly impacted by different link functions used in the algorithm.

Putative transcription factor-gene regulatory networks by PAINT and TRANSFAC:

The results from tightClust(50) and tightClust(100) were further analyzed by using PAINT v3.5 (Vadigepalli, et al., 2003) in conjunction with the TRANSFAC public database (http://www.gene-regulation.com/) to identify putative transcription factor-gene regulatory networks. The PAINT fetches upstream DNA sequences of genes and passes the sequences to TRANSFAC database search algorithm for known motif sequences in the TRANSFAC database to identify enriched putative common motifs across genes. Since PAINT could not identify any enriched transcription factors for the majority of small clusters from tightClust(100), the results from tightClust(50) will be reported hereafter.

There were 10 clusters that included significantly upregulated transcripts in the ankylosing spondylitis group and 19 clusters that contained significantly upregulated transcripts for the sarcoidosis group. The rest consisted of significantly downregulated genes in either group. Figure 1(a) – (f) shows the heat maps of the top 3 upregulated clusters in each disease.

All 29 tight clusters with upregulated transcripts in one of the diseases were further analyzed by using PAINT (Vadigepalli, et al., 2003) in conjunction with TRANSFAC public database (http://www.gene-regulation.com/) for transcription factors within 2000 base pairs upstream from the transcription start sites after combining entries when there were multiple probe sets for a transcript and removing entries for unknown transcripts. Table **2** summarizes common transcription factors for transcripts in the top 3 tight clusters of each disease that were significantly more frequent in each cluster compared to the other transcription factors of the human genome. Given the exploratory nature and small to moderate size of clusters, we used the marginal p-value < 0.05 to define significance. Figure 2(a) – (f) shows the corresponding potential transcription factor regulatory networks with related genes. These results show that there are distinct differences in the putative transcription factor regulatory profiles of ankylosing spondylitis and sarcoidosis. Among the top tight clusters, many immune-related genes were found in the second and third clusters of ankylosing spondylitis as well as in the first cluster of sarcoidosis. In addition, we also found strong STAT1 signature in one of upregulated clusters for sarcoidosis group. The biological implication of these findings is under review elsewhere or in preparation as it is beyond the scope of this manuscript.

Discussion:

By combining tightClust with the bioinformatics tools, PAINT and TRANSFAC database, the process of making predictions for putative transcription factor regulatory mechanisms was streamlined and the potential differences in the regulatory networks were shown for two immune-mediated diseases, ankylosing spondylitis and sarcoidosis. Figure 3 illustrates the analysis flow used in this study. This is a significant step forward from a descriptive gene expression pattern analysis. This approach could reveal differences in the pathogenesis of ankylosing spondylitis and sarcoidosis and thus lead to new diagnostic, prognostic or therapeutic approaches. These predicted regulatory networks will require further verification by using other independent samples and confirmatory biological experiments.

There are other clustering algorithms that are designed to group only tight data points since this is a hot research topic in microarray studies. One example is adaptive quality-based clustering by De Smet et al. (De Smet, et al., 2002). It is a heuristic iterative two-step method which has been proposed specifically for microarray data with the similar aim of clustering only tight data points. However, we could not finish the adaptive quality-based clustering due to a processing error at the time of analysis.

Because development of the clustering algorithms that we tested was based on different methods or models, their performance assessment in terms of revealing common transcription factors is not straightforward. In addition, transcription factors are an active research topic and the TRANSFAC database is constantly growing. As our knowledge grows and our models evolve, the available algorithms will need periodic reevaluation.

Care must be taken when applying these tools. As with any analysis of these hybridization microarray assays, normalization of the values among arrays is critical. The current results were based on tight clustering with the standardization of gene expression within a transcript. The results without standardization were different and generally expression patterns within a cluster were not consistent since the tightClust used *k*-means clustering in each iteration within resample. We also observed that the average number of genes per cluster was inversely proportional to a choice of target number of clusters. We tried 25, 50 and 100 while other parameters were the same. The average numbers of genes per cluster were 131, 49 and 23 with a total of 3267, 2487 and 2248 genes clustered, respectively. It is not clear whether this property is of a concern. Having too many genes and too few genes in a cluster would both impede our ability to identify associations, such as regulatory networks, among the cluster members. In addition, since new information is frequently added to transcription factor databases, a future analysis may reveal other regulatory networks. Nonetheless, the tools described here provide a useful aid to generating plausible hypotheses to be tested in a biologic system. Although small tight clustering increases the possibility of artifactual groupings, the results make predictions of transcription factor-gene regulatory networks more efficient and are more readily analyzed for biochemical and regulatory pathways as demonstrated here for transcription factor-based networks.

Current therapy for either ankylosing spondylitis or sarcoidosis lacks consistent efficacy. For example, while tumor necrosis factor inhibitors are proving highly effective in the treatment of ankylosing spondylitis, roughly 20 to 40% of patients demonstrate an inadequate clinical response (Davis, et al., 2007). Pharmacological techniques are rapidly emerging to allow therapeutic targeting of transcription factors and their binding sites. Such a therapeutic approach has the danger of being more toxic than the targeting of a specific cytokine, but a corollary to this is that this approach potentially will have much greater therapeutic efficacy compared to the inhibition of a specific cytokine or inflammatory mediator. The identification of clusters of upregulated transcripts should lead to novel therapeutic approaches and new insights into disease pathogenesis.

Acknowledgements:

- Supported by NIH Grants, HL077466, EY015858, EY010572; Research to Prevent Blindness Career Development Awards to JTR, SRP, and JRS, and the Stan and Madelle Rosenfeld Family Trust.

Conflict of interest notice:

- Christina A. Harrington has an equity interest in Affymetrix, Inc. This potential conflict of interest has been reviewed and managed by OHSU.

Reference:

Antes, T.J., Goodart, S.A., Huynh, C., Sullivan, M., Young, S.G. and Levy-Wilson, B. (2000) Identification and characterization of a 315-base pair enhancer, located more than 55 kilobases 5' of the apolipoprotein B gene, that confers expression in the intestine, *Journal of Biological Chemistry*, 275, 26637-26648.

Barton, L.M., Gottgens, B., Gering, M., Gilbert, J.G., Grafham, D., Rogers, J., Bentley, D., Patient, R. and Green, A.R. (2001) Regulation of the stem cell leukemia (SCL) gene: a tale of two fishes, *Proceedings of the National Academy of Sciences of the United States of America*, 98, 6747-6752.

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, 289-300.

Bonifer, C. (2000) Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required?, *Trends in Genetics*, 16, 310-315.

Brophy, S., Pavy, S., Lewis, P., Taylor, G., Bradbury, L., Robertson, D., Lovell, C. and Calin, A. (2001) Inflammatory eye, skin, and bowel disease in spondyloarthritis: genetic, phenotypic, and environmental factors, *J Rheumatol*, 28, 2667-2673.

Choi, D., Fang, Y. and Mathers, W.D. (2006) Condition-specific coregulation with cis-regulatory motifs and modules in the mouse genome, *Genomics*, 87, 500-508.

Dai, D.L., Wang, Y., Liu, M., Martinka, M. and Li, G. (2007) Bim expression is reduced in human cutaneous melanomas, *Invest Dermatol*.

Davis, J.C., van der Heijde, D.M., Braun, J., Dougados, M., Clegg, D.O., Kivitz, A.J., Fleischmann, R.M., Inman, R.D., Ni, L., Lin, S.L. and Tsuji, W. (2007) Efficacy and safety of up to 192 weeks of etanercept therapy in patients with ankylosing spondylitis, *Ann Rheum Dis*.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B. and Moreau, Y. (2002) Adaptive quality-based clustering of gene expression profiles, *Bioinformatics*, 18, 735-746.

Driouch, K., Landemaine, T., Sin, S., Wang, S. and Lidereau, R. (2007) Gene arrays for diagnosis, prognosis, and treatment of breast cancer metastatis, *Clin Exp Metastasis*.

Fachin, A.L., Mello, S.S., Sandrin-Garcia, P., Junta, C.M., Donadi, E.A., Passos, G.A. and Sakamoto-Hojo, E.T. (2007) Gene expression profiles in human lymphocytes irradiated in vitro with low doses of gamma rays, *Radiat Res*, 168, 650-665.

Fraley, C. and Raphael, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation., *Journal of the American Statistical Association* 97, 611-631.

Hans, C.P., Weisenburger, D.D., Greiner, T.C., Gascoyne, R.D. and al, e. (2004) Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray, *Blood*, 103, 275-282.

Hartigan, J.A. and Wong, M.A. (1978) A K-means clustering algorithm, *Applied Statistics*, 28, 100-108.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, 4, 249-264.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding groups in data: An introduction to cluster analysis*. Wiley, New Jersey.

Langelder, P., Zhang, B. and Horvath, S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R., *Bioinformatics*, 24, 719-720.

Lee, D.U., Avni, O., Chen, L. and Rao, A. (2004) A distal enhancer in the interferon-gamma (IFN-gamma) locus revealed by genome sequence comparison, *Journal of Biological Chemistry*, 279, 4802-4810.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proc Natl Acad Sci U S A*, 98, 31-36.

Michelson, A.M. (2002) Deciphering genetic regulatory codes: a challenge for functional genomics, *Proc Natl Acad Sci U S A*, 99, 546-548.

Nikpour, M., Dempsey, A.A., Urowitz, M.B., Gladman, D.D. and Barnes, D.A. (2007) Association of a gene expression profile from whole blood with disease activity in systemic lupus erythematosus, *Ann Rheum Dis*.

Pena, D., Rodriguez, J. and Tiao, G. (2003) Identifying Mixtures of Regression Equations by the SAR Procedure., *Bayesian Statistics* 7, 327-347.

Pennacchio, L.A. and Rubin, E.M. (2001) Genomic strategies to identify mammalian regulatory sequences., *Nature Reviews Genetics*, 2, 100-109.

R_Development_Core_Team (2007) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, J.T. (1989) Uveitis. An internist's view, *Arch Intern Med*, 149, 1173-1176.

Sieper, J., Rudwaleit, M., Khan, M.A. and Braun, J. (2006) Concepts and epidemiology of spondyloarthritis, *Best Pract Res Clin Rheumatol*, 20, 401-417.

Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies, *Proc Natl Acad Sci U S A*, 100, 9440-9445.

Tseng, G.C. and Wong, W.H. (2005) Tight clustering: A resampling-based approach for identifying stable and tight patterns in data, *Biometrics*, 61, 10-16.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci U S A*, 98, 5116-5121.

Vadigepalli, R., Chakravarthula, P., Zak, D.E., Schwaber, J.S. and Gonye, G.E. (2003) PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification, *Omics*, 7, 235-252.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F. and Spencer, F. (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays, *Journal of the American Statistical Association*, 99, 909-917.

Figure 1: link: average, non-standardized

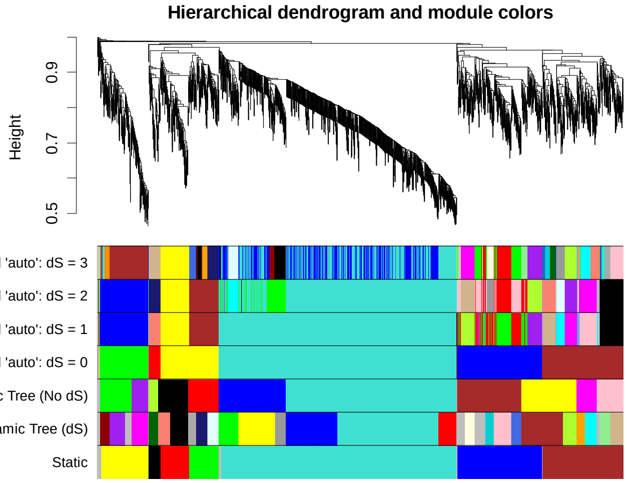


Figure 2: link: complete, non-standardized

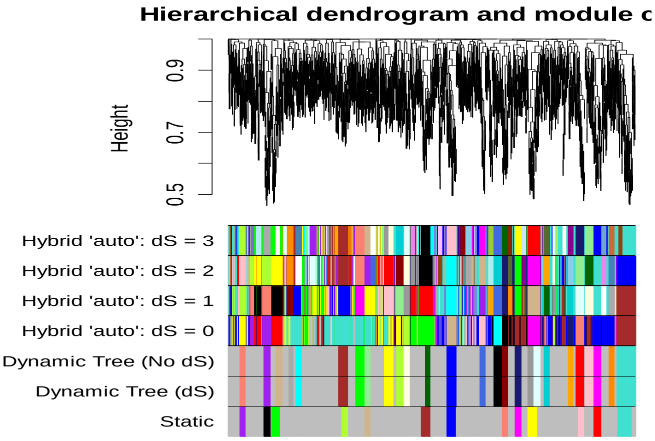
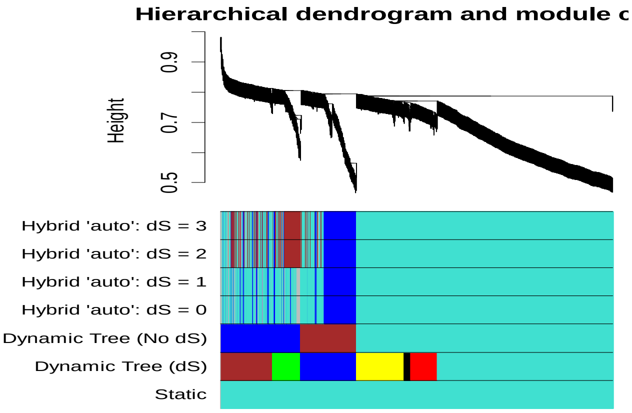


Figure 3: link: single, non-standardized



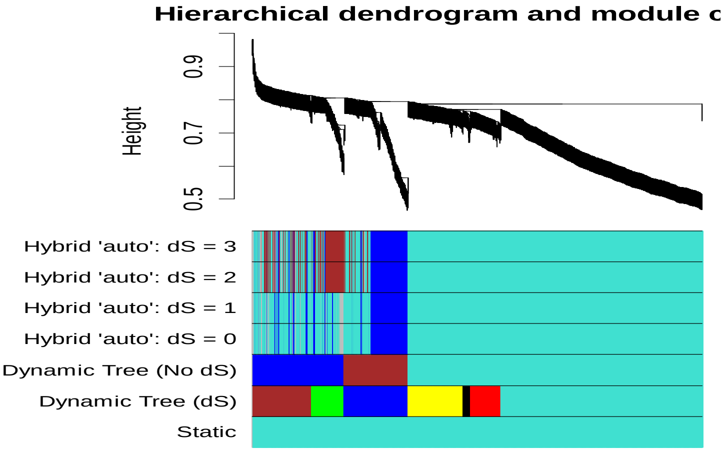


Figure 4: link: ward, non-standardized

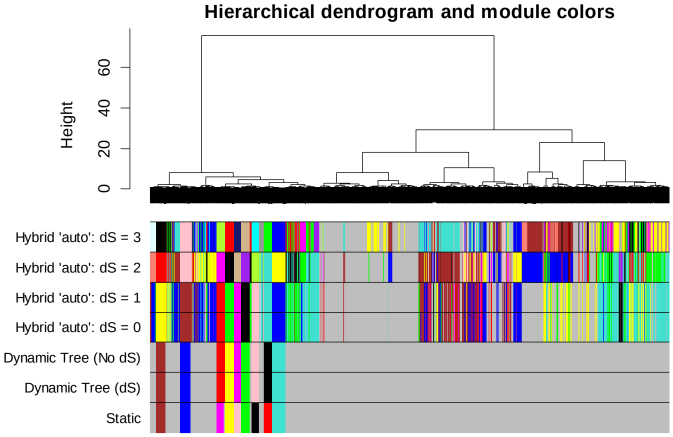


Figure 5: link: median, non-standardized

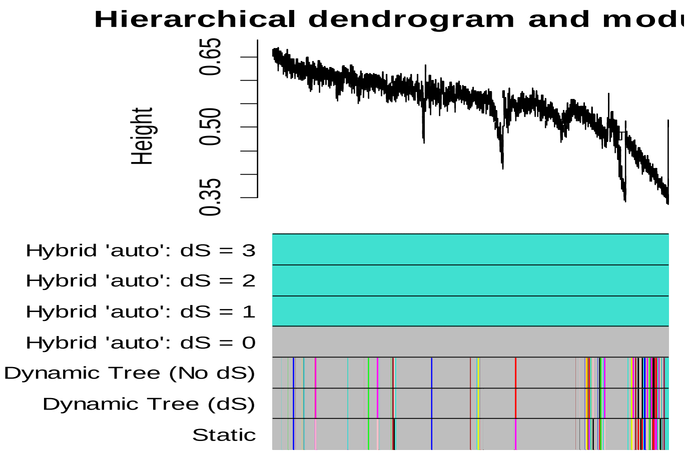


Figure 6: link: `mcquitty, non-standardized`

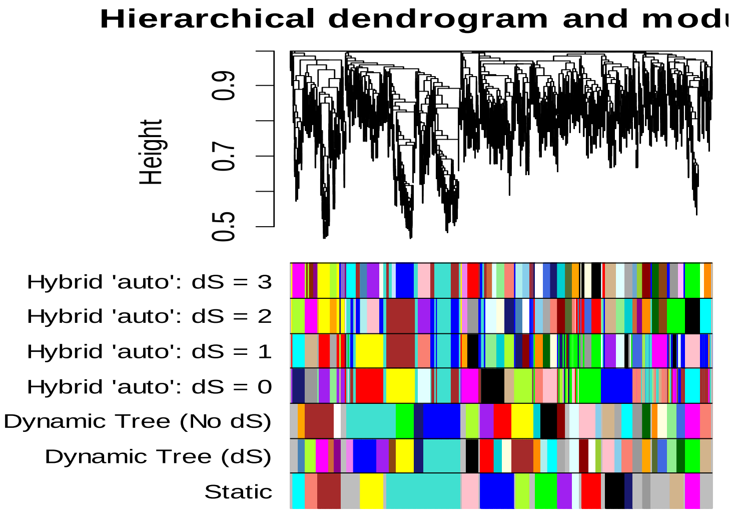
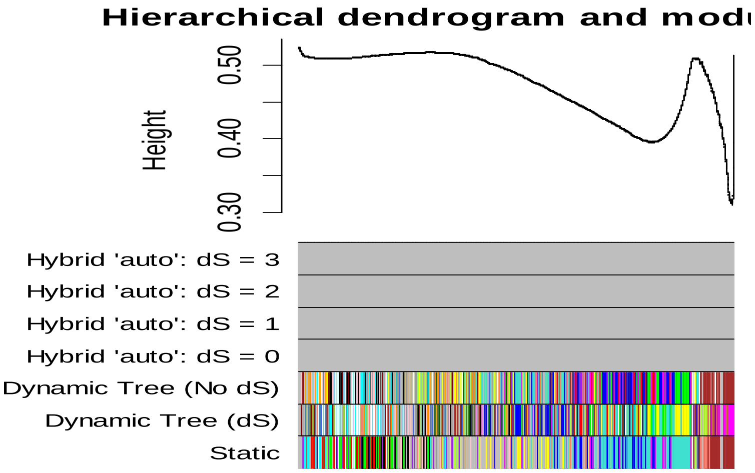


Figure 7: link: centroid, non-standardized


Hierarchical dendrogram and modu
Height
0.30
0.40
0.50
Hybrid 'auto': dS = 3
Hybrid 'auto': dS = 2
Hybrid 'auto': dS = 1
Hybrid 'auto': dS = 0
Dynamic Tree (No dS)
Dynamic Tree (dS)
Static