Fatigue-related HIV Disease Gene-Networks identified in $CD14^+$ cells isolated from HIV infected Patients - Part II Statistical Analysis

Abstract

Purpose: In limited samples of valuable biological tissues, univariate ranking methods of microarray analyses often fail to show significant differences among expression profiles. In order to allow for hypothesis generation, novel statistical modeling systems can be of great advantage. We applied new statistical approaches to solve the issue of limited experimental data to generate new hypotheses in CD14 cells of HIV-related fatigue patients and healthy controls.

Methodology: We compared gene expression profiles of $CD14^+$ cells of low versus high fatigued, NRTI-treated HIV patients to healthy controls (n=5 each). With novel Bayesian modeling procedures, we identified 32 genes predictive of low versus high fatigue and 33 genes predictive of healthy versus HIV infection. Sparse association and liquid association networks further elucidate the possible biological pathways in which these genes are involved.

Relevance for nursing practice: Genetic networks developed in a comprehensive Bayesian framework from small sample sizes allow nursing researchers to guide future research approaches to solve such problems as HIV-related fatigue.

Implication for Practice: The findings from this pilot study may take us one step closer to the development of useful biomarker targets for fatigue status. Specific and reliable tests are needed to diagnosis, monitor and treat fatigue and mitochondrial dysfunction.

Genome-wide studies, coupled with clinical and physiological data, provide key information for the determination of biological pathways involved in many critical diseases. Without exception, the data resulting from high-throughput sequencing techniques are characterized by an extremely small number of available samples with respect to the number of candidate factors of interest. This leads to substantial difficulties in the subsequent statistical analyses. The overarching goal is to select a reduced pool of genes and traits that might play a role in the relevant disease-inducing mechanisms. Based on this initial selection, a network of interactions is identified with the goal of generating pertinent conjectures about the underlying biological processes. These conjectures are further refined based on current knowledge, then validated in additional studies and laboratory experiments.

In the statistical literature the selection of genes and traits is referred to as a variable selection problem. Univariate rankings of the strength of association between each candidate factor and disease status represent the most straightforward and wide-spread method to perform variable selection (Dudoit, Fridlyand, & Speed, 2002; Golub et al., 1999; Nguyen & Rocke, 2002; Tusher, Tibshirani, & Chu, 2001). The assessment of the statistical significance of the huge number of null hypothesis tests associated with each individual factor is a complex question that still lacks a definite answer (Benjamini & Hochberg, 1995; Efron & Tibshirani, 2002) (Storey & Tibshirani, 2003). For this reason univariate rankings often lead to the puzzling conclusion that no factors seem to be significant for the response of interest. Such a claim is not justified for two reasons: (i) the available data might be insufficient to *prove* the involvement of any single factor in the disease-generating mechanisms, but it could still show what factors are *most likely* to be involved; and (ii) combinations of two, three or more predictors also need to be taken into account. Complex hypotheses in which several factors contribute together in the

progression of a disease compound the variable selection problem because it is no longer computationally feasible to exhaustively explore all the possible candidate models. In the context of linear regression, stepwise methods (Furnival & Wilson, 1974) represent a first solution that allows the identification of the most relevant regressions without listing all candidate regressions. Unfortunately these methods can handle only small datasets due to their inability to escape local modes created by patterns of collinear factors; hence, they are unlikely to perform well in the context of genome-wide studies.

A major step forward were Markov chain Monte Carlo (MCMC) algorithms that explore the model space in a Bayesian framework, by sampling from the joint posterior distribution of the regression parameters and the candidate regressions (Chipman, George, & McCullogh, 2001; George & Mcculloch, 1993; Clyde & George, 2004). The performance of MCMC methods can be enhanced by integrating out the regression parameters, thereby creating chains that move only in the space of regressions. In this case sampling from a posterior distribution is no longer required and other stochastic search methods that aggressively move towards regions of high posterior probability in the regressions space are more desirable. Two such algorithms stand out for their performance: the shotgun stochastic search (SSS) of Hans, Dobra and West (2007) and the bounded mode stochastic search (BMSS) of Dobra (2009). Both algorithms have been proven to give excellent results when selecting the most promising factors from large datasets involving gene expression, genotype, clinical and physiological information.

Exploring patterns of covariation in the observed data using genetic networks can further unravel complex biological processes. A network is represented as a graph whose vertices are associated with entities of interest (e.g., disease status, genes, single nucleotide polymorphisms, body measurements) and whose edges link two vertices if they are considered to be associated in

a certain way. Since the number of available samples is small, the networks must be sparse; most pairs of entities should be considered unrelated and the corresponding edges removed. Sparse networks arise quite often in biology. For example, only a small number of regulatory factors are expected to influence the evolution of any given process. There are numerous ways to build genetic networks from data. The interpretation of each network should be done based on the procedure employed to construct it. Networks build from the same data by using several statistical approaches that complement each other as they reveal various aspects of the available information. Association networks have edges given by genes whose expression levels exhibit high absolute associations as quantified by Kendall's tau, Spearman's rho or Pearson's correlation coefficient. Genes that are linked directly in association networks are likely to share the same biological functions (Butte, Tamayo, Slonim, Golub, & Kohane, 2000; Steuer, Kurths, Fiehn, & Weckwerth, 2003). Genes involved in the same biological pathways can be identified by shortest path analysis (Zhou, Kao, & Wong, 2002).

Another type of genetic networks is determined from Gaussian graphical models (Dobra et al., 2004; Schafer & Strimmer, 2005). In this case the edges in the network correspond with the non-zero elements of the precision matrix associated with a multivariate Gaussian distribution. Liquid association networks (Li, 2002) are constructed with respect to a phenotype of interest. Entities (e.g., genes and phenotypes) are connected if their relationship changes as a function of this phenotype. For example, two genes might exhibit strong co-expression in diseased samples, while in healthy samples their expression levels might be unrelated. As opposed to the association networks, the liquid association networks capture such a dynamic relationship.

In this paper we focus on the determination of genetic networks associated with fatigue and HIV status. The genes included in the networks are determined through the BMSS algorithm for logistic regressions (Dobra, 2009). Furthermore, we explore whether the edges we uncover from the data actually correspond to true biological processes.

Material and Methods

Patient Groups and Blood Draws

As part of a larger NIH intramural natural history study: "*Assessing the Relationship Between Fatigue and Mitochondrial Toxicity in Patients with HIV/AIDS* " (05-CC-0127) with a focus on HIV-related fatigue and mitochondrial toxicity, we enrolled 46 HIV positive patients on NRTI-containing and protease-inhibitor sparing ART regimens and 15 healthy controls. Results of the complete study will be reported at a later date. Patients completed self-report evaluations for fatigue, quality of life and depression. Muscle and fat biopsies were taken and peripheral blood mononuclear cells (PBMCs) were collected via aphaeresis. HIV patients were evaluated using the revised 26 item Piper Fatigue Scale with fatigue scores varying between 0-10, where (0-3) was considered no fatigue, (4-7) moderate, and (8-10) severe fatigue (Piper et al., 1998). For the purposes of this study, we combined patients listed as moderate and severe together and termed them high fatigue. In this substudy, three categories of $CD14^{+}$ cell samples were compared: cells from HIV patients with high fatigue (n=5), HIV patients with low fatigue (n=5) and healthy controls (n=5). For details on gender, age, race, CD4 count, viral load, medication regimen, drug use, and hepatitis co-infection status, please see Table 1 in Part I. There were no significant differences between healthy controls and both HIV patient groups.

$CD14^+$ Cell isolation and RNA extraction

$CD14^+$ cells were isolated from total PBMCs after aphaeresis with a negative $CD14^+$ isolation procedure, to prevent activation, using the Invitrogen Dynabead Monocyte Negative Isolation kit (Invitrogen, Carlsbad, CA) followed by $CD14^+$ positive cell sorting, according to the manufacturer's instructions. In addition, the CD16 monocyte subfraction, an activated subset of monocytes (Alexaki, Liu, & Wigdahl, 2008; Altenburg, Jin, Alkhatib, & Alkhatib; Crowe, Zhu, & Muller, 2003; Ziegler-Heitbrock, 2007) were selected against using anti-CD16 capture beads. Samples were sorted for $CD3^+/CD14^+$ monocytes using a B-D FACSAria system (Becton-Dickinson,Franklin Lakes, NJ) according to the manufacturer's protocols, see supplemental material. Cells were transferred to a QIAShredder MiniSpin Column (Qiagen, Valencia, CA) to sheer DNA, then spun and cell extracts were frozen and stored at -70°C until RNA extraction. RNA was extracted as previously described (Voss et al., 2008).

Microarray protocol, data processing and gene association network determinations

A custom Affymetrix (Santa Clara, CA) microarray design, FATMITO1a520158F, containing 4712 unique probes for mitochondrial related genes, was employed in these experiments (Voss, et al., 2008). The target synthesis protocol, hybridization and data extraction were done as previously described (Voss, et al., 2008). Microarray data were submitted to the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/index.cgi) and can be found under the accession number GSE18468. We followed a microarray data analysis approach as previously described (Dobra, 2009).

Bayesian Framework for Genetic Networks

We identify candidate genes related to a binary response variable Y (e.g., HIV status or Fatigue status) by following the Bayesian inference methods described in Dobra (2009). Briefly, we assume that there are p candidate genes $X_1,\ldots X_p$. We let D be the n x (p+1) data matrix, where the rows correspond with samples, the i^{th} column corresponds with variable X_i, while the last column corresponds with variable Y. We denote by $[Y|X_A]$ the logistic regression model

$$\log it(P(Y = 1 \mid X_A = x_A)) = \beta_0 + \sum_{i \in A} \beta_i x_i,$$

where A is a subset of {1,2,…,p}. A full Bayesian specification of the logistic regression model $[Y|X_A]$ is obtained by assuming that the regression coefficients follow independent N(0,1) priors, which lead to the marginal likelihood $p(D|[Y|X_A])$ that is numerically computed through a Laplace approximation – see Appendix A in Dobra (2009). We assume that all the possible logistic regression models are a-priori equally likely; hence the posterior probability of $[Y|X_A]$ is proportional with its marginal likelihood, i.e., $P([Y \mid X_A] \mid D) \propto P(D \mid [Y \mid X_A])$.

We make use of the Bounded Mode Stochastic Search (BMSS) algorithm to identify high posterior probability logistic regressions. Our candidate genes are those genes that are involved in the top 1000 logistic regressions identified by BMSS. These regressions can be further combined in a classifier for Y. Each regression $[Y|X_A]$ receives a weight proportional with its posterior probability $p([Y|X_A]|D)$. The classifier is obtained by taking the weighted average of the top 1000 logistic regressions. This technique is called Bayesian model averaging (BMA) and is often used in the literature to build high performance predictive systems (including, for example, weather forecasting tools – see http://probcast.washington.edu/). BMA (Kass & Raftery, 1995) works especially well in the context of genome-wide studies because it avoids picking a single model from the huge pool of candidate regressions. Due to the small sample

size, there is not enough information to favor any one given model with respect to the other. Instead, model uncertainty is taken into account by considering a large number of top models. Any subsequent statistical inference (e.g., variable selection or prediction) is performed based on these top models.

BMSS can be further employed to build Bayesian dependency networks (Heckerman, Chickering, Meek, Rounthwaite, & Kadie, 2000). These are joint distributions specified by a collection of regressions of each random variable given the rest. The set of regressions associated with each variable is determined independently of the other collections by an application of BMSS. Once a dependency network has been identified, one can draw random samples from it using an ordered Gibbs sampling algorithm (Dobra, 2009) that sequentially samples from each conditional distribution as follows: (i) sample a regression for the corresponding collection of regressions, (ii) sample from the posterior distribution of parameters of this regression; and (iii) sample a data point from the fully identified regression equation.

Genetic networks can be efficiently estimated based on a large number of samples (>10,000) from a dependency network. Because the conditional distribution of each variable given the rest is identified by a weighted average of the top regressions, the resulting genetic networks are much sparser than the networks that would be constructed by estimating associations directly from the observed data. Most pair-wise dependencies are reduced to zero when the top regressions are determined by BMSS; hence, only the strongest associations are represented through edges. Furthermore, the estimation of associations from data sampled from dependency networks significantly decreases the inherent correlations between the corresponding test statistics; hence, multiple comparisons issues (Efron, 2007) are a lot less serious and the use of false discovery rate techniques are no longer required.

Results

HIV dependent genes

We first analyzed the microarray data to identify mitochondrial dysfunction genes associated with HIV. In this analysis there were 10 HIV patient samples, and 5 HIV negative controls. The 15 microarray hybridization CEL files were preprocessed (i.e., background corrected and normalized) using GC Robust Multi-array Average (see Section 10 OLS). After the removal of 112 control probes included on the Affymetrix custom chip, the resulting dataset comprised 4712 probes with 10 HIV cases and 5 controls.

A number of 33 genes are present in the top 1000 highest posterior logistic regressions corresponding with the binary outcome HIV status – see Part 1 Table 2. Bayesian model averaging involving these 1000 logistic regressions led to a classifier that correctly predicted HIV status for all 15 samples. These excellent prediction results held when performing leave-one-out cross-validation. We show the prediction results in Figure 1, where HIV status is coded as '0' for HIV negative and '1' for HIV positive.

Insert Figure 1

We constructed genetic networks that involve the 33 candidate genes and HIV status. The association network involved 286 edges associated with variables whose pairwise Kendall's tau values that were different from zero at a false discovery rate of 1%. Since highly connected genes (hubs) were more likely to play key roles in the underlying biological processes, we expressed the topology of the resulting association network by the degree of each vertex (i.e., the number of direct neighbors). We sorted the genes in decreasing order with respect to their degrees and with respect to their posterior inclusion probabilities with respect to HIV status.

The most relevant genes had larger ranks on both scales: ADCY2, CASC4, YAF2, GOT1, HDAC6. The 22 genes that were direct neighbors of HIV status in this association network together with those having strong adjacent network association between their expression levels and HIV status (having false discovery rate of 1%) and their Kendall's tau coefficients are shown in Table 2 (Part 1). Positive Kendall's tau values indicate a positive, while negative values indicate a negative, association. Positive association increases as we go from non-HIV to HIV infected expression levels. Negative associations indicate the opposite. Two of these 33 HIV associative gene probes were subsequently identified in the most recent Unigene sequence database http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene to represent the same gene, cofilin 2; 224352_s_at and 224663_s_at. Therefore there are 32 unique genes identified in this analysis of HIV association.

Figure 2 shows the liquid association network corresponding with HIV status. This graph links pairs of genes whose association as measured by Kendall's tau changes with respect to HIV status. The significance level for the corresponding permutation tests was 0.05. The genes making up the liquid association set for HIV are listed in Table 3 (Part 1). The genes that did not have any direct neighbors in this network (i.e., the singletons) are not shown. The direct neighbors of ADCY2 are GARS, TNNC1 and GOT. The hub of this network is OSBPL7 with five direct neighbors. OSBPL7 was ranked high with respect to its degree in the association network.

Insert Figure 2

Fatigue dependent genes

The 10 fatigue CEL files were preprocessed as for HIV analysis (i.e., background corrected and normalized) using GC Robust Multi-array Average (Z Wu, Irizarry, Gentleman, Murillo, & Spencer, 2004; Z. Wu & Irizarry, 2005). After the removal of 112 control probes included on the Affymetrix custom chip, the resulting dataset comprised 4712 probes with 5 low fatigued (0-3 fatigue score) samples and 5 Fatigued (4-10 fatigue score) samples.

A number of 33 genes are present in the top 1000 highest posterior logistic regressions corresponding with the binary outcome Fatigue status – see Table 3 (Part 1). We remark that the same number of genes has been present in the top logistic regression corresponding with HIV status. This is simply a coincidence and might be a consequence of the small number of samples available in our study. Bayesian model averaging involving these 1000 logistic regressions led to a classifier that correctly predicted Fatigue status for all 10 samples. These excellent prediction results held when performing leave-one-out cross-validation, as shown in Figure 3. The binary response variable associated with Fatigue status was coded as '0' for low fatigued and '1' for fatigued.

Insert Figure 3

The association network involving the 33 candidate genes and Fatigue status had 505 edges that corresponded with pairwise Kendall's tau values that were different from zero at a false discovery rate of 1%. We ranked the 33 candidate genes in decreasing order with respect to their degrees in the resulting association network. We also ranked the genes in decreasing order with respect to the posterior inclusion probabilities. The most relevant genes ranked high in both orderings. These were: two probe sets for CFL2, SULT2B1, and PAG1. One of the two probe sets for CFL2 originally was categorized as a separate gene by Unigene and thus CFL2 was

unknowingly placed twice on the array. Both probe sets are located on the long untranslated 3' terminal exon of CFL2, with the higher Kendall's tau probe set being close to the poly A tail thus having a higher rate of reverse transcription into target cDNA. This made our identification of 32 unique genes in the Fatigue association set. Figure 4 shows the liquid association network corresponding with fatigue status. This graph links pairs of genes whose association as measured by Kendall's tau changes with respect to Fatigue status has a significance level for the corresponding permutation tests of 0.05. ADCY2 has only one neighbor in the liquid association network, namely AGTR2. However, AGTR2 is a neighbor of HSD17B3, the hub of this network. Two independent targets were originally classified as different genes by the NCBI Unigene databases but later established as CFL2. Therefore we identified 32 unique genes out of 33 different Affymetrix probe sets. Both targets were associated with HIV status in this analysis.

Insert Figure 4

Discussion

We showed that Bayesian techniques for variable selection and network determination are a powerful tool for uncovering relevant biological pathways of interest. The edges in our genetic networks no longer represent linear associations; hence, the underlying nature of gene expression data is captured in a more efficient manner in the joint distributions specified by dependency networks. The framework from Dobra (2009) efficiently handles genome-wide studies involving any combination of continuous or discrete factors. The networks are sparse due to the regression selection process behind the BMSS algorithm. The small number of samples is properly accounted for through Bayesian model averaging.

The methodology described in Dobra (2009) expresses the conditional distribution of each variable given the rest as a finite mixture of linear regressions. Although extremely flexible, this framework does not reflect the intensity-dependent and non-linear nature of gene expression data. Furthermore, direct interactions between factors are not allowed under the current model specification. More suitable formulations should allow the inclusion of non-linear components as well as of interaction terms. Another problem relates to the existence of a joint distribution associated with a dependency network identified by BMSS. Under a general positivity condition, a dependency network uniquely determines a joint distribution up to a normalizing constant (Besag, 1974). This condition is at least loosely satisfied in our approach although it cannot be guaranteed to always hold. The work of Hobert and Casella (1998) is an excellent starting point for answering this complex question: they study the case when a dependency network determines an improper joint distribution.

Until now there have been few tools to quantify fatigue and no molecular biomarkers of HRF. Our analysis of gene expression in this study has allowed us to generate multiple hypotheses for future studies of HIV fatigue etiology. First, SULT2B1 is the gene most associated with fatigue status and is a key enzyme in androgen biosynthesis (Croxson et al., 1989; Honour, Schneider, & Miller, 1995). HIV disease is associated with low gonadal and adrenal androgen deficiencies and fatigue in men with HRF is improved with testosterone therapy (Rabkin, Wagner, McElhiney, Rabkin, & Lin, 2004). Our data are consistent with the literature linking androgen levels to HRF and newly implicates SULT2B1 and its role in androgen generation as a factor in hormonal dysregulation in HRF disease. Second, the fatigue association of the actin cytoskeleton polymerization regulators CFL2 and f-actin binding scaffolding protein PAG1 indicates that cell motility and cytoskeletal structural rigidity, factors

controlled by f-actin stability, are involved in fatigue response in $CD14^+$ cells (Van Troys et al., 2008; Wu et al., 2008; Svec, 2008; Itoh et al., 2002). Since f-actin also stabilizes signaling complexes contained in anchored lipid rafts, this can also have consequences in the regulation of cell responses to agonist stimulation. Our data suggest that changes in cellular behavior controlled by the state of the actin cytoskeletal network may play a role in HRF, possibly through unknown mechanisms in $CD14^+$ cell activity. Lastly, PROK2 is a GPCR agonist that controls torpor (state of mental or physical inactivity or insensibility), attenuates circadian rhythms and regulates normal sleep patterns, and potently promotes neutrophil chemotaxis (Li et al., 2006; Zhong, Qu, Tan, Meng, & Ferrara, 2009; Gottlieb, O'Connor, & Wilk, 2007; Jethwa et al., 2008; Monnier & Samson, 2008). This protein was negatively associated with fatigue indicating that sleep regulation is disturbed in HRF patients and consistent with the literature that sleep patterns are disturbed in HIV disease and HRF, and indicates a hormonal defect in HRF. A general hypothesis from our identification of fatigue biomarkers could be that a subset of the genes identified in this study could also be biomarkers of fatigue across many diseases with a chronic fatigue component. Future studies comparing other diseases with associated fatigue will allow us to test this hypothesis.

Conclusion

This pilot study took advantage of rare CD14 cell samples collected by aphaeresis in a natural history study of mitochondrial dysfunction and HIV-related fatigue with the resources of the NIH intramural program and the most meticulous cell sorting techniques of the NCI Institute. The availability of a newly developed mitochondrial gene expression chip (huMITOchip) allowed us to investigate the gene expression patterns of CD 14 cells of fatigued and non-fatigued HIV patient compared to healthy controls. Newly developed statistical methods enabled

us to analyze these results in a way that regular microarray analyses (false-discovery rate and t-statistic based methods) would have not allowed. The development of new tools and methods are innovations that generated three major new hypotheses to further study HIV-related fatigue in much greater detail and could be tools of great value for other studies in the future.

References

Alexaki, A., Liu, Y., & Wigdahl, B. (2008). Cellular reservoirs of HIV-1 and their role in viral persistence. *Curr HIV Res, 6*(5), 388-400.

Altenburg, J. D., Jin, Q., Alkhatib, B., & Alkhatib, G. The potent anti-HIV activity of CXCL12gamma correlates with efficient CXCR4 binding and internalization. *J Virol, 84*(5), 2563-2572. doi: JVI.00342-09 [pii]10.1128/JVI.00342-09

Benjamini, Y., & 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.

Besag, J. (1974). Spatial Interaction and Statistical-Analysis of Lattice Systems. *Journal of the Royal Statistical Society Series B-Methodological, 36*(2), 192-236.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., & Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A, 97*(22), 12182-12186. doi: 10.1073/pnas.220392197 220392197 [pii]

Chipman, H., George, E. I., & McCullogh, R. E. (2001). The practical implementation of Bayesian model selection (with discussion) *In: Lahiri, P. (Ed.), Model Selection. IMS: Beachwood, OH*, 66-134.

Clyde, M., & George, E. I. (2004). Model uncertainty. *Statistical Science, 19*(1), 81-94. doi: Doi 10.1214/088342304000000035

Crowe, S., Zhu, T., & Muller, W. A. (2003). The contribution of monocyte infection and trafficking to viral persistence, and maintenance of the viral reservoir in HIV infection. *J Leukoc Biol, 74*(5), 635-641. doi: 10.1189/jlb.0503204jlb.0503204 [pii]

Croxson, T. S., Chapman, W. E., Miller, L. K., Levit, C. D., Senie, R., & Zumoff, B. (1989). Changes in the hypothalamic-pituitary-gonadal axis in human immunodeficiency virus-infected homosexual men. *J Clin Endocrinol Metab, 68*(2), 317-321.

Dobra, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics, 10*(4), 621-639. doi: DOI 10.1093/biostatistics/kxp018

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. A., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis, 90*(1), 196-212. doi: DOI 10.1016/j.jmva.2004.02.009

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*(457), 77-87.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association, 102*(477), 93-103. doi: Doi 10.1198/016214506000001211

Efron, B., & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology, 23*(1), 70-86. doi: Doi 10.1002/Gepi.01124

Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. . *Technometrics*(16), 499-511.

George, E. I., & Mcculloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association, 88*(423), 881-889.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531-537.

Gottlieb, D. J., O'Connor, G. T., & Wilk, J. B. (2007). Genome-wide association of sleep and circadian phenotypes. *BMC Med Genet, 8 Suppl 1*, S9. doi: 1471-2350-8-S1-S9 [pii]

10.1186/1471-2350-8-S1-S9
Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., & Kadie, C. (2000). Dependency networks for inference, collaborative filtering and data visualization. *Journal of Machine Learning Research*(1), 1-48.
Honour, J. W., Schneider, M. A., & Miller, R. F. (1995). Low adrenal androgens in men with HIV infection and the acquired immunodeficiency syndrome. *Horm Res, 44*(1), 35-39.
Itoh, K., Sakakibara, M., Yamasaki, S., Takeuchi, A., Arase, H., Miyazaki, M. (2002). Cutting edge: negative regulation of immune synapse formation by anchoring lipid raft to cytoskeleton through Cbp-EBP50-ERM assembly. *J Immunol, 168*(2), 541-544.
Jethwa, P. H., I'Anson, H., Warner, A., Prosser, H. M., Hastings, M. H., Maywood, E. S. (2008). Loss of prokineticin receptor 2 signaling predisposes mice to torpor. *Am J Physiol Regul Integr Comp Physiol, 294*(6), R1968-1979. doi: 00778.2007 [pii]10.1152/ajpregu.00778.2007
Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795.
Li, J. D., Hu, W. P., Boehmer, L., Cheng, M. Y., Lee, A. G., Jilek, A. (2006). Attenuated circadian rhythms in mice lacking the prokineticin 2 gene. *J Neurosci, 26*(45), 11615-11623. doi: 26/45/11615 [pii] 10.1523/JNEUROSCI.3679-06.2006
Li, K. C. (2002). Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad Sci U S A, 99*(26), 16875-16880. doi: 10.1073/pnas.252466999 252466999 [pii]
Monnier, J., & Samson, M. (2008). Cytokine properties of prokineticins. *FEBS J, 275*(16), 4014-4021. doi: EJB6559 [pii]10.1111/j.1742-4658.2008.06559.x
Nguyen, D. V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics, 18*(1), 39-50.
Piper, B. F., Dibble, S. L., Dodd, M. J., Weiss, M. C., Slaughter, R. E., & Paul, S. M. (1998). The revised Piper Fatigue Scale: psychometric evaluation in women with breast cancer. *Oncol Nurs Forum, 25*(4), 677-684.
Rabkin, J. G., Wagner, G. J., McElhiney, M. C., Rabkin, R., & Lin, S. H. (2004). Testosterone versus fluoxetine for depression and fatigue in HIV/AIDS: a placebo-controlled trial. *J Clin Psychopharmacol, 24*(4), 379-385. doi: 00004714-200408000-00004 [pii]
Schafer, J., & Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics, 21*(6), 754-764. doi: bti062 [pii] 10.1093/bioinformatics/bti062
Steuer, R., Kurths, J., Fiehn, O., & Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics, 19*(8), 1019-1026.
Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A, 100*(16), 9440-9445. doi: 10.1073/pnas.1530509100 [pii]
Svec, A. (2008). Phosphoprotein associated with glycosphingolipid-enriched microdomains/Csk-binding protein: a protein that matters. *Pathol Res Pract, 204*(11), 785-792. doi: S0344-0338(08)00137-4 [pii]10.1016/j.prp.2008.06.006
Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A, 98*(9), 5116-5121. doi: 10.1073/pnas.091062498091062498 [pii]
Van Troys, M., Huyck, L., Leyman, S., Dhaese, S., Vandekerkhove, J., & Ampe, C. (2008). Ins and outs of ADF/cofilin activity and regulation. *Eur J Cell Biol, 87*(8-9), 649-667. doi: S0171-9335(08)00069-1 [pii]10.1016/j.ejcb.2008.04.001

Voss, J. G., Raju, R., Logun, C., Danner, R. L., Munson, P. J., Rangel, Z. (2008). A focused microarray to study human mitochondrial and nuclear gene expression. *Biol Res Nurs, 9*(4), 272-279. doi: 9/4/272 [pii]10.1177/1099800408315160

Wu, Y., Yoder, A., Yu, D., Wang, W., Liu, J., Barrett, T. (2008). Cofilin activation in peripheral CD4 T cells of HIV-1 infected patients: a pilot study. *Retrovirology, 5*, 95. doi: 1742-4690-5-95 [pii] 10.1186/1742-4690-5-95

Wu, Z., Irizarry, R., Gentleman, R., Murillo, F., & Spencer, F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association, 99*(468), 909-917.

Wu, Z., & Irizarry, R. A. (2005). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol, 12*(6), 882-893. doi: 10.1089/cmb.2005.12.882

Zhong, C., Qu, X., Tan, M., Meng, Y. G., & Ferrara, N. (2009). Characterization and regulation of bv8 in human blood cells. *Clin Cancer Res, 15*(8), 2675-2684. doi: 1078-0432.CCR-08-1954 [pii] 10.1158/1078-0432.CCR-08-1954

Zhou, X. H., Kao, M. C. J., & Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America, 99*(20), 12783-12788. doi: DOI 10.1073/pnas.192159399

Ziegler-Heitbrock, L. (2007). The CD14+ CD16+ blood monocytes: their role in infection and inflammation. *J Leukoc Biol, 81*(3), 584-592. doi: jlb.0806510 [pii] 10.1189/jlb.0806510