33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

A Phenotype-Driven Dimension Reduction (PhDDR) Approach to Integrated Genomic Association Analyses

Cuilan Gao and Cheng Cheng

Abstract— **An immediate challenge in integrated genomic analysis involving several types of genomic factors all measured genome-wide is the ultra-high dimensionality. Screening all possible relationships among the genomic factors is an NP-hard problem; therefore in practice proper dimension reduction is necessary. In this paper we develop the Phenotype-Driven Dimension Reduction (PhDDR) approach to the analysis of gene co-expressions, and discuss its extensions to integration of other genetic factors. This approach is then illustrated by an application to gene co-expression analysis of treatment response of childhood leukemia.**

I. Introduction

Recent advancement of biotechnologies has enabled investigators to either directly measure, or derive, on genome-wide scale diverse types of genetic and epi-genetic factors in a large number of experimental units. The mountains of genomic data now provide opportunities to integrate different types of genomic information for a more comprehensive understanding of the underlying genomic processes affecting important health-related phenotypes. Two equally important and tightly related components in this process are genomic data integration and integrated analysis of genomic associations. Genomic data integration deals with efficient storage, cross reference, literature linkages, retrieval, and visualization of all types of genomic data. Integrated analysis of genomic association deals with effective statistical inferences to discover relationships between genomic factors and phenotypes of interest, as well as relationships among the genomic factors themselves; thereby elucidate the underlying genomic process affecting the phenotypes. Data integration facilitates integrated analyses. This paper is concerned with integrated analysis.

An integrated genomic association analysis beginning with the various genomic factors measured genome-wide encounters immediately the challenge of ultra-high dimensionality. Even with two types of genomic factors involved, say mRNA expressions for which the number of expression probes (EPs) is typically on the order of $2x10^5$, and SNPs the number of which is typically $9x10^6$, so the total number of all possible SNP-EP pairs is approximately $2x10^{12}$; clearly exhaustive screening of all possible combinations of all measured genomic factors is an NP-hard problem. Moreover, these genomic factors are usually measured on several hundred to a few thousand individuals. Often exhaustive screening of even all possible pairs is prohibitive for the computing resource in a typical academic research institution. Typically some reduction has to be performed prior to analysis, such as focusing only on *cis* SNPs for each EP.

In a cancer genomic study there is often a biological context defined by one or more phenotypes of interest; for example, to search for inter-related genomic factors jointly affecting the response to remission induction (a phenotype) in a cohort of uniformly treated patients. The specific biological context provides an opportunity to perform the effective and biologically meaningful Phenotype-Driven Dimension Reduction (PhDDR). The PhDDR approach starts with searching phenotype-specific gene co-expression sets and then extends to other types of genomic factors; whereby effectively circumvent the difficulties of ultra-high dimensionality in integrating massive numbers of different types of genomic factors into an association analysis.

We illustrate the PhDDR approach in Section II by the analysis of phenotype-specific gene co-expressions and discuss extensions to integrate other types of genomic factors. In Section III a validation inference for the discovered Phenotype-Specific Gene Co-Expression Sets (PheGCES) is developed. The use of PhDDR-PheGCES is illustrated by an application to a gene co-expression study in childhood leukemia in Section IV. Further discussion on extensions to integrate diverse types of genomic factors and some concluding remarks are made in Section V.

II. Phenotype-Specific Gene Co-expression Detection

A major component of contemporary cancer genomics is to understand how gene (mRNA) expressions and co-expressions are related with complex, clinically important phenotypes. One approach is to first construct co-expression networks or clusters [1]–[5], and then test if any of the modules are associated with the phenotypes. Exhaustive search to construct co-expression networks is NP-hard in terms of the total number of expression probes (EPs); the available methods are still computationally difficult for large number EPs ($2x10^5$); and it is often difficult and subjective to determine a proper number of clusters in a study.

In contrast, the PhDDR approach begins with testing the associations between each individual EP and the phenotype. Effective and biologically meaningful dimension reduction

Manuscript received March 26, 2011. This work was supported in part by the National Cancer Institute and National Institute of General Medical Sciences of the National Institutes of Health (U01GM 92666, U19HL065962), and by the American Lebanese Syrian Associated Charities (ALSAC).

C. Gao is with the Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105 USA.

C. Cheng is with the Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105 USA (Phone: 901-595-2935, Fax: 901-595-8843, email: cheng.cheng@stjude.org).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

is done by considering the top EPs significantly associated with the phenotype by proper statistical significance criteria. Then the Phenotype-Specific Gene Co-Expression Sets (PheGCES) are constructed using the top phenotype-associated EPs as leads. A PheGCES is a co-expressed gene cluster associated with the phenotype of interest; the association can be further validated externally with independent validation data (see Sections III and IV) or internally with an internal validation inference [6].

To describe the PheGCES detection algorithm it is necessary to introduce some notation first. Let n be the number of experimental units (sample size), and m be the number of EPs. Let X_i be the vector of expression data of the ith EP ($i=1,2,\dots,m$); each X_i is a n-vector, the jth component of X_i is the expression level of the ith EP in experimental unit j. Let $Y_1, Y_2, \dots, Y_n$ be the observed phenotype values; Y_j is the phenotype value of experimental unit j.

Algorithm 1: Detection of PheGCES

1. For each EP, test its association with the phenotype; obtain a phenotype-association P value for each EP.
2. Sort the EPs according to the above-obtained P values in ascending order.
3. Use the P values to select top phenotype-associated EPs by some statistical significance criteria. Let R be the number of selected EPs, and X_k^* ($k=1,\dots,R$) be the expression data vectors of these EPs.
4. Detect co-expression clusters. Specify a positive correlation threshold $\rho>0$.

```
Set $nc=0$;
REPEAT for $k=1,\dots,R$
IF (the EP corresponding to $X_k^*$ is not in any
    PheGCES yet) THEN
    Put the EP into the current PheGCES;
    $nc=nc+1$;
    REPEAT for $i=1,\dots,m$
        IF ($i$th EP is not in any PheGCES yet) THEN
          Compute the correlation coefficient
          $r(X_i, X_k^*)$ of the expression vectors
          $X_k^*$ and $X_i$;
          IF $r(X_i, X_k^*)\geq \rho$ THEN
             Put $i$th EP into the current PheGCES;
        END.IF
    END.REPEAT
END.IF
END.REPEAT
```

Depending on the data type of the phenotype, established methods are available to test the expression-phenotype associations [7]–[8]. Likewise a number of established methods are available for significance inference in Step 3 [9]–[12]. The number of selected top EPs, R, is typically on the order of 10^2-10^3. Hence the PhDDR starting with EPs can reduce the dimension of the EP space from the order of 10^5-10^6 to typically 10^2. In Step 4 nc records the current PheGCES number; at the end it is the number of PheGCES's detected. Either Pearson's or Spearman's (rank) correlation can be used for the correlation coefficient $r(X_i, X_k^*)$. It can be seen from the algorithm that each PheGCES has a "core" EP which is one of the top phenotype-associated EPs determined in Step 3, and each EP in the PheGCES is positively correlated with the core EP at the level at least ρ.

Integration of other types of genomic factors can be done straightforwardly: The inner loop "REPEAT for $i=1,\dots,m$" searching in the EPs can as well be performed among other types of genomic factors, such as microRNA expressions, methylation levels, and SNPs (coded as AA=0, AB=1, BB=2). The inclusion condition "$r(X_i, X_k^*)\geq \rho$" may need to be modified for this extension. For example, to incorporate the biological intuition that DNA methylation may down regulate gene expression, use "$-r(X_i, X_k^*)\geq \rho$" for each methylation EP data vector X_i. On the other hand if there is no reason to apply any restriction, then replace $r(X_i, X_k^*)$ by its absolute value in the condition. The search can be repeated for each type of genomic factors interrogated in the study. In the end each "PheGCES" so constructed is a set of related genetic and epi-genetic factors associated with the phenotype of interest through the core EP, and defines the vertex set of a local genomic association network for the locus (gene) represented by the core EP. Formally the network is an edge-weighted labeled graph with vertices labeled by the genomic factors and edges weighted by pair-wise correlations. Specific biological interpretation of such a local genomic network is application dependent. Generally speaking however, such a phenotype-associated local genomic network shows how a gene's effect on the phenotype is brought by the joint actions of the related genomic factors, as part of the underlying molecular-cellular process affecting the phenotype.

Of practical importance is the selection of a correlation threshold ρ for the inclusion condition. A preliminary simulation study showed that the findings can remain stable for the moderate threshold values 0.2 to 0.4; the number and contents of the detected co-expression sets are more sensitive to stringent threshold values 0.6 to 0.9 (data not shown). Clearly the operating characteristics will largely depend on the data. In practice one can try a few threshold values (e.g., 0.3, 0.5, and 0.7) and compare the findings. It would be desirable to have a data-adaptive procedure to select this threshold, such as one similar to the statistical significance threshold criteria Ip [9]; this is still an open problem at this point.

III. Validation Inference

The PheGCESs are constructed by a search specific for "high" correlations with EPs significantly associated the phenotype. A statistical concern of this search from the machine learning standpoint is "overtraining" that may lead to a high number of false positive findings. This can be addressed by a validation inference internally [6], or externally on an independent validation dataset. We now describe an external validation procedure below; first some additional notation.

Let N_k be the number of EPs and $J_1, J_2, \dots, J_{Nk}$ be the indices (identifiers) of the EPs in the kth PheGCES. Let D_j^T be the direction of association between the phenotype and the *j*th EP from the training dataset, $j=J_1, J_2,\dots, J_{Nk}$.

Algorithm 2: Validation P value of PheGCES

1. For every EP, test its association with the phenotype on the validation dataset; obtain a phenotype-association P value for each EP on the validation set.
2. Sort the EPs according to the above-obtained P values in ascending order.
3. For the *k*th PheGCES, let D_j^V be the direction of association between the phenotype and the *j*th EP on the validation dataset, and R_j be the *j*th EP's rank on the order determined in Step 2, for $j=J_1, J_2,\dots, J_{Nk}$. Compute the PheGCES rank score statistic
$$S_k = -\sum_{j=J_1\dots J_{Nk}} I(D_j^T D_j^V > 0)\log\left(\frac{R_j}{m+1}\right),$$
where $I()$ is the indicator function taking value 1 if the condition inside the parentheses is true, 0 otherwise; and m is the total number of EPs.
4. The validation P value of the *k*th PheGCES is determined by $P_k^V=1-F(S_k)$, with the cumulative distribution function F described below.
5. Repeat Steps 3 and 4 for each PheGCES.

Now each PheGCES has a validation P value (Step 4) measuring the statistical evidence for it is association with the phenotype on the validation set. The indicator function $I()$ in the statistic S_k insists that an EP in a PheGCES contributes supporting evidence for validation only when its directions of association with the phenotype are the same on both the training and validation sets. The larger is the statistic, the stronger evidence for positive validation. The validation call for a PheGCES can be made based on its validation P value and the usual 5% significance level, or more conservatively after an adjustment for multiple tests.

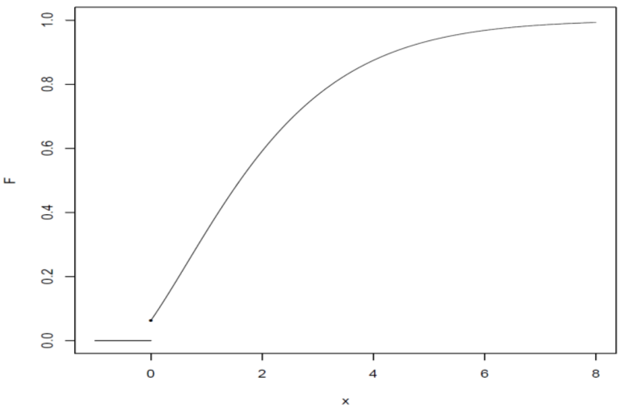
The validation P value is computed for testing the statistical null hypothesis that none of the EPs in the PheGCES is associated with the phenotype in the validation set. Under this null hypothesis the two directions of association agrees by chance; so the indicator $I()$ follows the Bernoulli(0.5) distribution; and the ranks R_j are distributed uniformly over $\{1,2,\dots,m\}$. When m is large (typically $>10^5$) the cumulative distribution function of S_k under the null hypothesis is approximately

$$F(x)=\begin{cases} 0, & x<0 \\ 0.5^{N_k}, & x=0 \\ 0.5^{N_k}+\sum_{t=1}^{N_k} b(t;N_k,0.5)G(x;t,1), & x>0 \end{cases}$$

where $b(\cdot;N,p)$ is the Binomial(N,p) probability mass function, and $G(\cdot;a,b)$ is the cumulative distribution function of the Gamma(a,b) distribution. This is the $F()$ function used in Step 4. Note $F()$ has a jump at 0 and for x>0 it is a mixture of Gamma distributions (Figure 1, for N_k=4).

FIGURE I

The probability distribution to compute validation P value



IV. An Application to Childhood ALL

Childhood Acute Lymphoblastic Leukemia (ALL) is the most common pediatric cancer. One of the strongest clinical prognostic factors is the Minimal Residual Disease (MRD; [13]) at the end of remission induction. Biologically MRD is an indicator of *in vivo* resistance to chemotherapy. Gene expression profiling studies for MRD have revealed genes in the cell cycle and proliferation pathways [14]–[15]. To illustrate the above method we now present a PheGCES analysis for MRD. *De novel* leukemic blasts of 288 ALL patients were assayed using Affymetrix U133A GeneChip®. Quality MRD data were available on 189 patients. Although MRD data were incomplete on the remaining 99 patients, quality clinical outcome data (time to relapse, other adverse events, or the last follow up) were available. Because MRD is tightly related to outcome [13], the 99-patient cohort formed a natural validation set with time to relapse as a validation phenotype. With MRD as a binary variable (positive or negative at the end of remission induction) PheGCES detection was performed on the 189-patient training set (Algorithm 1): Differential expression of each of the m≈22,000 EPs (probesets) between MRD-positive and MRD-negative status was tested using the Wilcoxon rank-sum test (Steps 1 and 2); The adaptive statistical significance threshold criteria Ip [9] was applied to select top probesets, 223 probesets were selected (Step 3); using these top probesets, Spearman's correlation, and ρ=0.8 (a stringent threshold), 188 PheGCES were found (Step 4). Out of these 188 PheGCES, 26 contained more than 1 probsets. In the validation inference (Algorithm 2), association between risk of ALL relapse and gene expression was tested using a hazard rate regression model [16] (Steps 1 and 2); validation P value of each PheGCES was computed on the validation set (Steps 3, 4 and 5), where for each probeset direction of expression-MRD association on the training set was the sign of the median expression in the MRD-positive group minus that in the MRD-negative group, and direction of the expression-relapse association on the validation set was the sign of the hazard regression coefficient. After the conservative Bonferroni adjustment for 188 tests, the validation P value of PheGCES #65 and #70 reached the 10% statistical significance level (Table 1).

TABLE I
PheGCES result of the application to childhood ALL

PheGCES valid. P val. (Bonf. Adj.)	Probe-set	P val. MRD	P val. Relapse	Dir. MRD	Dir. Relapse
#65 0.0003	212021_s_at	0.0009	0.0048	-1	-1
	212023_s_at	0.0039	0.0028	-1	-1
	212020_s_at	0.0046	0.0000	-1	-1
	212022_s_at	0.0304	0.0069	-1	-1
#70 0.0751	204026_s_at	0.0010	0.0010	-1	-1
	210559_s_at	0.0024	0.0860	-1	-1

PheGCES #65 contained 4 probesets of the *MKI67* gene which encode a nucleotide and ATP binding protein in the cell cycle regulation pathway; lower expression was related to positive MRD and higher risk of relapse. Two of the 4 probesets were not among the top probesets determined by the Ip criteria, but they were strongly related to the risk of relapse in the validation cohort. PheGCES #70 contained a probeset of *ZWINT* and one of *CDC2*; the former is on the cell cycle regulation pathway and the latter encodes a nucleotide binding protein involved in cell division and regulation of cell cycle; again lower expression of these genes was related to positive MRD and higher risk of relapse. It is conceivable that all these are (possibly distant) genes on the cell cycle regulation and proliferation pathways supporting the leukemic cells' survival when the cells were exposed to cytotoxic agents.

Another dimension reduction technique for EP-phenotype association analysis is the random coefficient (effect) regression model [17]. Applying this method to test the global association of the 22,000 EPs with the MRD, we obtained weak statistical significance P=0.0998.

V. Discussion and concluding Remarks

We have developed the Phenotype-Driven Dimension Reduction (PhDDR) approach to integrated genomic analysis and illustrated its usage in gene co-expression analysis with an application to childhood ALL. When a study centers at a phenotype, PhDDR can achieve very effective dimension reduction in a way pertinent to the study's biological context. We have also developed an inference procedure for external validation of the findings.

Other dimension reduction techniques used in genomics include for example sparse canonical correlation [18]; but in order for this type of method computationally feasible, often great simplifications, such as assuming only diagonal covariance matrix, has to be made. Comparing to the existing methods [17] – [18], the PhDDR approach relies on fewer statistical assumptions, is more pertinent to the biological context defined by the phenotype, and more computationally efficient by avoiding manipulations of exceedingly large matrices.

Acknowledgment

We thank our colleagues at St. Jude Children's Research Hospital, Dr. Dario Campana and Ms. Elaine Coustan-Smith for the interesting application and the MRD data, Dr. James Downing for the gene expression data, and Dr. Stan Pounds for helpful discussions. We thank the referees for their constructive critiques that helped improve the paper much.

References

[1] R. Xulvi-Brunet, and H. Li, "Co-expression networks: graph properties and topological comparisons," *Bioinformatics,* vol. 26, pp. 205–214, 2010.

[2] W. Zhao, P. Langfelder, T. Fuller; J. Dong, A. Li, and S. Hovarth, "Weighted gene coexpression network analysis: state of the art," *J. Biopharm Statist,* vol. 20, pp. 281–300, 2010.

[3] J. Schafer, and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics,* vol. 21, pp. 754–764, 2005.

[4] B. Zhang, and S. Horvath, "A general framework of weighted gene co-expression network analysis," *Statist. Appl. Genetics Mol. Biol.* Vol. 4, Article 17, 2005.

[5] A. D. Gordon, *Classification*, 2nd ed. New York, NY: Chapman & Hall/CRC, 1999, ch. 2–3.

[6] C. Cheng, "Internal validation inferences of significant genomic features in genome-wide screening," *Comput. Statist. Data Anal.,* vol. 53, pp. 788–800, 2009.

[7] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist, "Detecting differential gene expression with a semiparametric hierarchical mixture method," *Biostatistics*, vol. 5, pp. 155–176, 2004.

[8] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to ionizing radiation response," *Proc. Natl. Acad. Sci. USA,* vol. 98, pp. 5116–5121, 2001.

[9] C. Cheng, S. Pounds, J. M. Boyett, D. Pei, M-L Kuo, and M. F. Roussel, "Statistical significance threshold criteria for analysis of microarray gene expression data," *Statist. Appl. Genetics Mol. Biol.,* vol. 3, Article 36, 2004.

[10] A. Reiner, D. Yekutieli, and Y Benjamini, "Identifying differentially expressed genes using false discovery rate controlling procedures," *Bioinformatics,* vol. 19, pp. 368–375, 2003.

[11] S. Pounds, and C. Cheng, "Robust estimation of the false discovery rate," *Bioinformatics,* vol. 22, pp. 1979–1987, 2006.

[12] D. B. Allison, and G. L. Gadbury, "A mixture model approach for the analysis of microarray gene expression data," *Comput. Statist. Data Anal.,* vol. 39 pp. 1–20, 2002.

[13] E. Coustan-Smith, J. Sancho, F. G. Behm, M. L. Hancock, B. I. Razzouk, R. C. Ribeiro, G. K. Rivera, J. E. Rubnitz, J. T. Sandlund, C-H Pui, and D. Campana, "Prognostic importance of measuring early clearance of leukemic cells by flow cytometry in childhood acute lymphoblastic leukemia," *Blood,* vol. 100, pp. 52–58, 2002.

[14] C. Flotho, E. Coustan-Smith, D. Pei, S. Iwamoto, G. Song, C. Cheng, C-H Pui, J. R. Downing, and D. Campana, "Genes contributing to minimal residual disease in childhood acute lymphoblastic leukemia: prognostic significance of *CASP8AP2*," *Blood,* vol. 108, pp. 1050–1067, 2006.

[15] C. Flotho, E. Coustan-Smith, D. Pei, C. Cheng, G. Song, C-H Pui, J. R. Downing, and D. Campana, "A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukemia", *Blood,* vol. 110 pp. 1271–1277, 2007.

[16] J. P. Fine, and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *J. Am. Statist. Assoc.,* vol. 94 pp. 496–509, 1999.

[17] J.J. Goeman, S.A. van de Geer, F. de Kort, and H.C. van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20 pp. 93–99, 2004.

[18] D.M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10 pp.515-534, 2009.