Modeling oncology gene pathways network with multiple genotypes and phenotypes via a copula method

Le Bao, Zhou Zhu, and Jingjing Ye

Abstract— **Identification of interactions between molecular features (e.g. mutation, gene expression change) and gross phenotypes in diseases and other biological processes is one of the important challenges in genomic research. Popular approaches such as GSEA are limited to hypothesis tests of bivariate association. However, a specific phenotype is often dependent upon multiple molecular features. It is thus worth considering all possible interactions jointly for a more precise and realistic representation of the cellular network. In this article, a semiparametric copula model is developed to jointly model genotypes, pathways and phenotypes to accomplish this object. A two-step procedure for reconstruction of the network is described. Simulation studies indicate that the method is effective and accurate for the network reconstruction. Application using NCI60 cancer cell line data identifies several subsets of molecular features that jointly perform as the predictors of clinical phenotypes. The copula model is expected to have a broad impact on biomedical research, ranging from cancer treatment to disease prevention.**

I. INTRODUCTION

In the past decade, genome-wide molecular profiling using microarray and sequencing technologies has become a mainstay of biological research. The rapid accumulation of genomic data presents a grand challenge in gaining insights into the relationship between various molecular features (e.g. mutation, mRNA and miRNA expression change, CNV, DNA methylation, histone modification) and gross phenotypes (e.g. clinical outcome, drug response) from them. A number of novel statistical models have been developed for determining whether a specific group of genes has a coordinated association with a phenotype of interest, including the Gene Set Enrichment Analysis (GSEA) algorithm (Subramanian et al., (2005) [31]; Tian et al., (2005) [32]). GSEA starts with a predefined list of gene sets and assigns an enrichment score to every gene set, followed by a permutation analysis to measure the significance of gene set heterogeneity associated with the phenotype. Although these commonly used methods have been successfully applied in many basic science and clinical studies (Mootha et al., (2003) [22]; Sweet-Cordero et al., (2005) [36]; Alvarez et al., (2005) [1]), they remain limited to hypothesis tests of bivariate associations. In biological reality, however, a gross phenotype is rarely determined by one single gene/pathway but a combination of them.

To represent the cellular network more precisely, it is worth modeling all possible interactions jointly. In statistical literature, network based approaches and copula models are two mainstream techniques to describe the multivariate dependencies. Recent works of the reconstruction of gene networks from gene expression data include linear (Chen et al., (2006) [6]; Deng et al., (2005) [9]; Reverter and Chan, (2008) [28]; Li and Li, (2008) [20]) and nonlinear (Wei and Li, (2007) [37]) models. This article will focus on an alternative approach, copula models, which are a vigorously growing field. The applications of copula models in statistics are a rather modern phenomenon especially to some major topics in finance such as derivative pricing and quantitative risk management (Cherubini et al., (2004) [7]; McNeil et al., (2005) [21]). The word 'copula' originates from the Latin noun for a "link or tie" that connects two different things. An attractive feature of the copula models is its capability to specify the univariate marginal distributions and their associations separately.

Concepts of copula models have recently gained interests in bioinformatics. For instance, Owzar et al. (2007) [25] studied the bivariate association of a large panel of genes with the survival outcomes such as time-to-death or time-to-recurrence and detected the prognostic genes through hypothesis testing. Kim et al. (2008) [18] reconstructed the interaction of genes using the Farlie Gumbel Morgenstern copula. Yuan et al. (2008) [38] performed the gene copy family analysis via the Gaussian copula and the Student's t-copula, both of which dealt with continuous variables only. In this paper, we describe a semiparametric copula model via extended rank likelihood which allows us to estimate the dependence structure of multiple continuous variables, ordinal variables, or the mixture of those two types. Hence it provides more flexibility in modeling a wide variety of biomedical data. The proposed semiparametric copula model is expected to have a profound impact on biomedical research, ranging from cancer treatment to disease prevention.

The article is organized as the following. In Methods section, a semiparametric copula model is introduced. A two-step procedure for reconstruction of the network is described. The Results section describes a simulation study to validate the proposed semiparametric copula approach. The

Le Bao is with department of statistics, University of Washington, Seattle, WA 98195 USA (e-mail: lebao@stat.washington.edu).

Zhou Zhu is with structural and computational biology, Pfizer Global Research and Development, Pfizer Inc., San Diego, CA 92121 USA (e-mail: Zhou.Zhu@pfizer.com).

Jingjing Ye is with Global Non-clinical Statistics, Pfizer Global Research and Development, Pfizer Inc., 10777 Science Center Drive, San Diego, CA 92121 USA and is the corresponding author of the paper (phone: 858-622-8054; fax: 877-481-0846; e-mail: Jingjing.Ye@pfizer.com).

978-1-4244-2756-7/09/$25.00 ©2009 IEEE

model is further applied to a real data analysis utilizing NCI60 cell line data. Conclusions and Discussion sections are followed.

II. METHODS

Many measures of dependencies in bivariate distribution have been proposed. The first and most important of these is still the correlation coefficient which may be ascribed to Galton (1888) [11]. The correlation coefficient makes the implicit assumption that the two variables are jointly normally distributed. When this assumption is not justified, a non-parametric measure such as the Spearman's rank correlation (Spearman (1904) [33]) might be more appropriate. These rank-based correlations are robust to the scales on which the variables are measured. But they often involve ad-hoc methods for dealing with ties, and cannot be used for conditional inference. As described earlier, the bivariate correlation is utilized for modeling associations in GSEA (Subramanian et al., (2005) [31]; Tian et al., (2005) [32]) and many other researches.

The copula models have attracted substantial attention in recent literature on multivariate observations modeling. An old mathematical result known as Sklar's Theorem states that every multivariate probability distribution can be represented by its univariate marginal distribution and a copula (Nelsen (1999) [23]). Copula models specify the joint distribution in two stages and this enables one to treat the univariate marginal distributions and the intrinsic joint behavior separately. The copula parameters capture the intrinsic dependence among the marginal variables, and they can be estimated by (i) the traditional parametric method with assumptions on marginal and copula families; (ii) the semiparametric method that treats the univariate marginal distributions as unknown functions (Genest, et. al. (1995) [13]); (iii) the complete nonparametric method, for example a multivariate kernel density estimate. An extensive simulation study (Kim et al. (2007) [17]) showed that the semiparametric method is more robust against misspecification of the marginal distributions than full parametric methods. Therefore, the parameters estimated by the semiparametric method are preferred.

To provide the inferences for copula parameters, Hoff (2007) [16] developed a method of semiparametric inference for copula models via a type of rank likelihood function for the association parameters. The proposed extended rank likelihood does not make any assumptions on the marginal distributions of the data, and therefore is appropriate for the analysis of mixed data (continuous and discrete) with arbitrary marginal distributions. Semiparametric estimation for dependence parameters of multivariate Gaussian copula is available in the R package "sbgcop" via a straightforward Gibbs sampling algorithm. The R package "sbgcop" can be downloaded at: http://cran.r-project.org/web/packages/sbgcop/index.html. The multivariate dependencies in this research then utilize the semiparametric method to estimate the copula parameters and the inferences are made based on Hoff (2007) [16]'s approach.

The detailed approach is described as the following. Let y_{ij} denote the i^{th} observation of the j^{th} variable, j=1, …, p, where p is the number of variables, i=1,…, n, where n is the number of observations. The variables are gross phenotypes, genotypes and pathway gene expressions in this research. Let z_j be the j^{th} p-dimensional latent variable. The semiparametric Gaussian copula model could be expressed as:

$$z_1,...z_n \mid C \sim \text{i.i.d. multivariate normal } (0, C)$$

$$y_{ij} = F_j^{-1}(\Phi(z_{ij}))$$

where C is a $p \times p$ correlation matrix, Φ is the cumulative distribution function of standard normal distribution and F_j could be arbitrary distribution functions, j=1,…, p.

We want to make the inference about the correlation matrix C, without making assumptions on F_j's. The latent z's are not observed, but the observed y's can provide some information about z. Since F_j's are non-decreasing, observing $y_{i_1 j} < y_{i_2 j}$ indicates that $z_{i_1 j} < z_{i_2 j}$. Using this knowledge, given observed y, the latent z must lie in the set D:

$$D = \{Z \in R^{n \times p} : \max\{z_{kj} : y_{kj} < y_{ij}\} < z_{ij} < \min\{z_{kj} : y_{ij} < y_{kj}\}\}$$

Take the occurrence of $Z \in D$ as our data and make inference based on the likelihood:

$$\Pr(Z \in D \mid C) = \int_D P(Z \mid C) dZ$$

Bayesian inference of C could be made by finding the posterior distribution of C: $P(C \mid Z \in D) \propto P(C)P(Z \in D \mid C)$. The Markov chain could be constructed by Gibbs sampling in the case of Gaussian model with semi-conjugate priors. For detailed prior and posterior distributions, please refer to section 3.2 of Hoff (2007) [16].

Since the copula model jointly estimates all possible associations of latent normal variables, the conditional correlation could be derived from the regression coefficients, which perhaps are more interesting for multivariate data. Because all of the latent variables Z are normally distributed, given the correlation matrix C, the regression coefficients of the j^{th} variable can be simply calculated as $C[j,-j]\ C^{-1}[-j,-j]$, j=1,…, p. To construct the network of mixed phenotypes, genotypes and gene

expressions, a backward elimination procedure is applied to identify the set of statistically significant (p-value<0.05) predictors for the j^{th} variable. The implementation is straightforward given the posterior samples of copula covariance matrix C:

(i) For the j^{th} variable, calculate the regression coefficients as $C[j,-j]\ C^{-1}[-j,-j]$, j=1,.., p;

(ii) Remove the corresponding row and column for the predictor whose coefficient is most likely to be zero based on the posterior distribution;

(iii) Iterate (i) and (ii) until all remaining predictors whose coefficients have 95% posterior probabilities of not being equal to 0.

By iterating the backward elimination procedure over j, j=1,…, p, a network is obtained. The directional connection i → j in the network indicates the i^{th} variable is a non-negligible predictor for the j^{th} variable, i, j=1, …, p. Note that i → j does not imply the i^{th} variable should be the cause of the j^{th} variable.

Summarizing the above procedures together, the two-step procedure to reconstruct the network is developed by estimating the copula parameters and their inferences first and then applying the backward elimination process. Two common tools for convergence diagnostic, Raftery-Lewis diagnostics (Raftery and Lewis (1992) [26]; (1995) [27]) and Gelman-Rubin's R statistic (Gelman and Rubin (1992) [12]; Brooks and Gelman (1997) [4]), are used to approximate the number of iterations required in MCMC. Both tools are available in R package CODA.

III. Results

A. Simulations

First, a set of simulations is performed to evaluate the performance of the proposed two-step method. To choose the appropriate ranges of the association between genotypes and gross phenotypes, a real example of the NCI60 data is utilized to help determine the ranges. The NCI60 data is a set of 59 human cancer cell lines derived from diverse tissues: brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate and skin. NCI60 project was described in detail in Ross et. al., (2000) [30] and the data are publicly available for download from http://genome-www.stanford.edu/nci60/. Genotypes in the NCI60 data are gene mutation status, coded as binary variables with 0 as wild type and 1 as mutated. Gross phenotypes in the data are cancer types, which are binary outcomes as well, with 1 denoted as the specified cancer and 0 as other types of cancer. After screening the associations of pair-wise outcomes between mutation status and cancer types, two pairs are selected: (i) BRAF mutation status and Melanoma (ME) have the highest linear association of 0.76, representing highly correlated genotype and phenotype; (ii) PTEN mutation status and cancer of central nervous system (CNS) have linear association 0.33, representing moderately correlated ones.

Let Y_1 and Y_2 denote the genotype and phenotype respectively. From the information in the NCI60 data, two scenarios are generated to represent their associations: one is 0.76 and the other is 0.33. Since genome-wide gene expression data have been collected for the NCI60 cell lines, two gene sets X_1 and X_2 are generated corresponding to Y_1 and Y_2, respectively. Several factors are taken into consideration: gene sets size, association of gene set and the binary outcome, and association between gene sets. For gene set size, each gene set is assumed to contain 10, 30, or 100 genes. 10% or 50% of the genes in a gene set are simulated based on the binary outcome: if the outcome (phenotype or genotype) is positive (1), the gene expression levels follow a normal distribution with mean of 5 and variance of 1; otherwise the standard normal distribution is used. The rest of genes (90% or 50%) are considered to be noise which is simulated under the standard normal distribution regardless of the outcomes. The proportion of genes that are simulated based on the outcomes indicates medium or high correlation between the gene set and the outcome. The expression level of the gene set is then summarized as the sum of expression levels of all gene members in the set. Finally, the association between gene sets X_1 and X_2 is controlled by the proportion of overlapped genes. None or half of the noise genes are shared between X_1 and X_2, implying low or high correlation between the two gene sets. A diagram of the simulated network setting is presented in Figure 1.

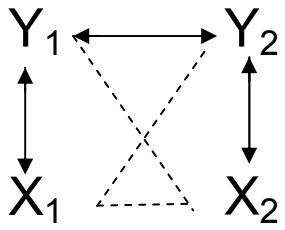

Figure 1: Diagram of the simulation setting. Y_1 and Y_2 are genotype and phenotype, respectively. X_1 and X_2 are two gene sets. The solid line with arrows in the graph indicates direct association. The dash line in the graph indicates indirect association. The link between X_1 and X_2 can be solid or dash depending on the setting. For simplicity, only dash line is presented in the graph.

For each simulation setting, 100 datasets are generated. After applying the two-step procedure, the percentages of cases that one variable is considered as significant predictor of the other are summarized in Table 1. Since the results are same for both directions of XY connection, e.g.

$X_1 \rightarrow Y_1$ and $Y_1 \rightarrow X_1$, only that for one of the directions is shown in Table 1. The posterior means of the marginal correlations are shown in Table 2. The tables are organized by gene set sizes. Within each gene set size, *P* denotes the proportion of genes in the set that is simulated based on the binary outcome; *r* is the linear correlation between two outcomes; overlap = T means gene sets X_1 and X_2 share 50% of the noise genes, while overlap=F means they share none.

From Table 2, it can be seen that in all simulated settings, the marginal correlation for each pair of variables are statistically significant (2.5^{th} and 97.5^{th} quartiles do not contain zero). Most of the marginal correlations are greater than 0.79, except the $X_1 \sim X_2$ correlation where the gene sets contain 10 genes with one gene set correlates with the phenotype and the outcomes association is weak. When we investigate the conditional correlations calculated by the two-step procedure in Table 1, the conclusions fit the simulated settings much better. First, the network distinguishes the direct connections $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$ from the indirect connections $X_2 \rightarrow Y_1$ and $X_1 \rightarrow Y_2$. When *XY* association is strong (*P* = 0.5 or gene set size is 30 or 100), $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$ are correctly identified with probability greater than 0.98; $X_2 \rightarrow Y_1$, and $X_1 \rightarrow Y_2$ are misspecified with probability less than 0.03. When *XY* association is weak (*P* = 0.1 or small gene set), the direct connections are still superior to the indirect connections, but their differences are sensitive to the number of genes in the gene set. Note that for larger gene sets, the noise could be evened out, thus resulting in a stronger *XY* correlation. Second, the $Y_1 \sim Y_2$ connections are identified as a consequence of either strong *YY* correlation (*r* = 0.76) or a weak *XY* correlation (*P* = 0.1 or small gene set), and they are always dominant to the $X_1 \sim X_2$ connections even under weak *YY* correlation and strong *XY* correlation scenarios. Note that the directional network we have constructed allows asymmetric connections, so the case that Y_1 performs as a predictor of Y_2 does not imply Y_2 is necessarily a predictor of Y_1. Third, sharing the same genes among gene sets (overlap = T) has little impact on the $X_1 \sim X_2$ associations, as long as those shared genes are independent of outcomes Y_1 and Y_2. As expected, the $X_1 \sim X_2$ correlation is only present with a strong $Y_1 \sim Y_2$ correlation. We thus conclude the proposed two-step procedure is able to separate direct and indirect associations and is consistent with the simulation settings.

B. NCI60 analysis

The method is further tested on a "real-world" analysis of the NCI60 data. To model the proposed method, genotypes of KRAS, BRAF and TP53 mutation status are included. Phenotypes considered are BR (breast cancer), CNS (central nervous system), CO (colon cancer), LE (leukemia), ME (melanoma), LC (lung cancer), OV (ovarian cancer) and RE (renal carcinoma). Both genotypes and phenotypes are binary outcomes. Gene expression data from mRNA profiling are also considered. Canonical pathways are downloaded from Broad Institute MSigDB database http://www.broad.mit.edu/gsea/msigdb/downloads.jsp [3]. The gene expression data are grouped based on gene sets of these 637 pathways using Entrez GeneID as identifiers. The enrichment scores of the pathways, as developed in GSEA (Subramanian et al., (2005) [31]), are first calculated for an initial screen. For each binary outcome, gene sets with top 10 enrichment scores are collected. By ignoring those related to metabolism in consideration of computational time, we focus on the resulting 44 signaling pathways for subsequent network building. The expression levels of the pathways are further calculated as summation of the individual gene levels associated with the pathways.

The two-step procedure is applied to the processed data. A total of 55 variables is included in the model. An inverse-Wishart prior distribution is set with identity matrix and *p*+2=57 degrees of freedom. The Gibbs sampling scheme is iterated 25,000 times with parameter values saved every 5 scans, resulting in 5,000 samples of C for posterior analysis. The convergence of Markov chains is checked based on two diagnostics. For Raftery diagnostics, the number of iterations required to estimate the quartile *q* within an accuracy of +/- *r* with probability *p* is calculated. Separate calculations are performed for each variable within each chain. Among all correlation parameters, the maximum number of iterations required to make the 2.5^{th} and 97.5^{th} quartiles within accuracy of 0.01 with probability of 0.95 is 9,860, which indicates 25,000 iterations are sufficient for the convergence. The Gelman-Rubin's diagnostics is also performed based on 5 parallel chains, which are of length 5,000 and start with different initial values. The Gelman-Rubin's R for elements of C has the maximum value 1.38, indicating very good convergence.

The network based on the copula correlations and backward elimination procedure is then constructed. The two-dimensional plot of the network is presented in Figure 2 using the R package "statnet", an integrated set of tools for the representation, visualization, analysis and simulation of network data (Handcock et. al., (2008) [15]). Figure 2 shows the network structure of the pathways that are connected with at least one of the outcomes, and hence some of the pathways' interactions are not presented.

Unlike the Pearson's linear correlation, which relies on the assumption of normality, and the Spearman's rank correlation, which is completely non-parametric, the copula correlation lies between those two extremes and reveals the

dependence of latent variables Z, which follows standard normal distribution. Hence the comparison of copula correlations appears more valid for mixed types of data. Moreover, the uncertainty of copula correlations could be estimated by the Bayesian procedure. The estimated copula correlations are listed in Table 3, where the copula correlations whose posterior probability regions (based on 2.5^{th} and 97.5^{th} quartiles) do not contain zero are highlighted.

The copula analysis of the NCI60 dataset reveals some interesting biological insights into the pathogenesis of multiple cancer types. For instance, we find melanoma to be jointly determined by BRAF mutation and NDK-dynamin pathway statuses. The BRAF gene encodes a serine/threonine kinase which plays an important role in the initiation and progression of melanoma. Large-scale sequencing effort has identified high frequency (>60%) of activating mutations of the BRAF gene in human melanomas (Davies et al. (2002) [8]). NDK, short for nucleoside diphosphate kinase and also known as Nm23, is a tumor suppressor gene (Rosengard et al. (1989) [29]; Stahl et al. (1991) [34]). Its mutation has been reported in melanoma cells of high metastatic potential (Hamby et al. (1995) [14]). Our observation here of the statistical dependency between NDK-dynamin endocytosis pathway and melanoma from mRNA profiling data supports a model in which NDK may function as a tumor suppressor by facilitating the downregulation of activated growth factor receptors via endocytosis (Futter et al. (1996) [10]; Krishnan et al. (2001) [19]). In addition to uncover relationships between molecular features and clinical phenotypes, the copula method also identifies directional interactions among molecular features, including the link from p53 mutation to CDC25 pathway. The CDC25 phosphatase mediates cell cycle progression into mitosis and is a key player in the G2 checkpoint in response to DNA damage (Nilsson and Hoffman (2000) [24]). It has recently been shown to be a novel target for transcriptional downregulation by tumor suppressor protein p53 and this repression can contribute to p53-dependent cell cycle arrest (St. Claire et al. (2004) [35]). Meanwhile, it is also worth noting that the copula model fails to establish a few "expected" linkages such as the link between ERBB2 pathway and breast cancer (BR) as ERBB2/HER2 is often disregulated in BR patients. While this may be partially explained by the biological and technical noises in the genomic datasets, accuracy in pathway composition as well as the lack of contrasting normal/non-disease samples in the NCI60 dataset, further investigation and improvement on the methodology are warranted.

IV. Conclusions and Discussion

With the rapid advancements in genome-wide molecular profiling technologies, there is an urgent need to develop computationally feasible models and methods for identifying associations within these high-dimensional datasets as well as with respect to disease phenotypes. The copula model has been successfully used in the areas of finance and social science, and this article represents one of the pioneering efforts in adapting it to genomic analysis. The semiparametric copula approach, proposed in Hoff (2007) [16], estimates the associations of mixed continuous gene expressions and discrete outcomes jointly via extended rank likelihood. It provides a gain in robustness because the inferences no longer rely on assumptions of the marginal distributions. This approach simplifies the estimation problem by eliminating a potential high-dimensional set of nuisance parameters of the marginal distributions and makes the conditional inference straightforward with the normal copula. Inferences on the scale of original data can be obtained with a posterior predictive distribution, which is derived by plugging in the empirical univariate marginal distributions as described in Section 4.3 of Hoff (2007) [16]. A two-step procedure for reconstruction of the network involving multiple pathways, genotypes and phenotypes is described in this work: (i) estimate the copula correlations and the inferences based on Hoff (2007) [16], (ii) determine the best subset of predictors for each variable by backward elimination. In simulation experiments, the proposed model yields the results that are consistent with the simulation settings and is able to distinguish direct from indirect associations. Moreover, the model offers the advantage of asymmetricity on the built network. When the model is applied to the analysis of a cancer cell line dataset (NCI60), interesting associations are found between molecular features and clinical phenotypes, including some that are further supported by existing data in the literature. Additionally, biologically relevant directional interactions between molecular features are identified as well.

In this article, the expression levels of gene sets are studied. For further studies and variations of the method, it is possible to investigate the expressions on individual gene level. An analysis on data consisting of normal individuals and diseased patients may be more interesting to validate the method. Readers should not confuse the cancer types, which are a set of binary outcomes, with a decomposition of a categorical variable. The decomposed binary variables from the categorical outcome cannot form a proper copula because their associations are determined by the marginal distribution of the categorical outcome. Although the method can be improved in several aspects as discussed, it is considered a useful tool for biomedical research, ranging from cancer treatment to disease prevention.

Acknowledgment

The authors are grateful to Xiaoyue Niu for bringing us the idea of the semiparametric copula model. The authors thank Xiaoyue and Paul Rejto for helpful discussions, and are grateful to the editor and three anonymous referees whose detailed comments have helped improving the manuscript.

References

[1] Alvarez, J., Febbo, P., Ramaswamy, S., Loda, M., Richardson, A. and Frank, D. (2005) Identification of a genetic signature of

activated signal transducer and activator of transcription 3 in human tumors. *Cancer Research*, 65, 5054-5062.
[2] Bild, A., Yao, G., Chang, J., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J., Berchuck, A., Olson, J., Marks, J., Dressman, H., West, M. and Nevins, J. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439, 353-357.
[3] Broad Institute, curated gene set pathways [Online]. Available: http://www.broad.mit.edu/gsea/msigdb/downloads.jsp.
[4] Brooks, SP. and Gelman, A. (1997) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.
[5] Butts., CT. (2007). Network: A Package for Managing Relational Data in R. *Journal of Statistical Software,* 24 (2), 1-36.
[6] Chen, XW., Anantha, G. and Wang, X. (2006) An effective structure learning method for constructing gene networks. *Bioinformatics* 22 (11), 1367-1374.
[7] Cherubini, U., Luciano, E. and Vecchiato, W. (2004) Copula methods in Finance. *Wiley, New York.*
[8] Davies, H, Bignell, GR, Cox, C, Stephens, P, Edkins, S, Clegg, S, Teague, J, Woffendin, H, Garnett, MJ, Bottomley, W, Davis, N, Dicks E, Ewing, R, Floyd, Y, Gray, K, Hall, S, Hawes, R, Hughes, J, Kosmidou, V, Menzies, A, Mould, C, Parker, A, Stevens, C, Watt, S, Hooper, S, Wilson, R, Jayatilake, H, Gusterson, BA, Cooper, C, Shipley, J, Hargrave, D, Pritchard-Jones, K, Maitland, N, Chenevix-Trench, G, Riggins, GJ, Bigner, DD, Palmieri, G, Cossu, A, Flanagan, A, Nicholson, A, Ho, JW, Leung, SY, Yuen, ST, Weber, BL, Seigler, HF, Darrow, TL, Paterson, H, Marais, R, Marshall, CJ, Wooster, R, Stratton, MR, Futreal, PA (2002) Mutations of the BRAF gene in human cancer. *Nature*, 417(6892), 949-954.
[9] Deng, X., Geng, H. and Ali., H. (2005) EXAMINE: a computational approach to reconstructing gene regulatory networks. *Biosystems*, 81, 125–136.
[10] Futter CE, Pearse, A, Hewlett, LJ, Hopkins, CR (1996) Multivesicular endosomes containing internalized EGF-EGF receptor complexes mature and then fuse directly with lysosomes. *The Journal* of *Cell Biology*, 132(6), 1011-1023.
[11] Galton, F. (1888) Co-relations and their measurement, chiefly from anthropological data. *Proceedings of The Royal Society London*, 45, 135-145.
[12] Gelman, A. and Rubin, DB. (1992) Inference from iterative simulation using multiple sequences, *Statistical Science*, 7, 457-511.
[13] Genest, C., Ghoudi, K., Rivest, L.-P. (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* 82 (3), 543-552.
[14] Hamby, CV, Mendola, CE, Potla, L, Stafford, G, Backer, JM (1995) Differential expression and mutation of NME genes in autologous cultured human melanoma cells with different metastatic potentials. *Biochemical and Biophysical Research Communications*, 211(2), 579-585.
[15] Handcock, M., Hunter D., Butts, CT., Goodreau, SM., Morris, M. (2007) **statnet**: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software,* 24 (1), 1-11.
[16] Hoff, P., (2007) Extending the rank likelihood for semiparametric copula estimation. *Annuls of Applied Statistics,* 1 (1), 265-283.
[17] Kim, G., Silvapulle, M.J., Silvapulle, P. (2007) Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis* 51 2836-2850.
[18] Kim, J.-M., Jung, Y.-S., Sungur, E.A., Han, K.-H., Park, C., and Sohn, I. (2008) A copula method for modeling directional dependence of genes. *BMC Bioinformatics*, 9, 225.
[19] Krishnan, KS, Rikhy, R, Rao, S, Shivalkar, M, Mosko, M, Narayanan, R, Etter, P, Estes, PS, Ramaswami, M (2001) Nucleoside diphosphate kinase, a source of GTP, is required for dynamin-dependent synaptic vesicle recycling. *Neuron*, 30(1), 197-210.
[20] Li, C. and Li, H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24 (9), 1175-1182.
[21] McNeil, AJ, Frey, R. and Embrechts, P. (2005) Quantitative risk management: concepts, techniques, and tools. *Princeton University Press*.
[22] Mootha, V., Lindgren, C., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M., Patterson, N., Mesirov, J., Golub, T., Tamayo, P., Spiegelman, B., Lander, E., Hirschhorn, J., Altshuler, D. and Groop, L. (2003) Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34, 267-273.
[23] Nelsen, RB. (1999) An introduction to copulas. *Springer, New York*.
[24] Nilsson, I and Hoffman, I (2000) Cell cycle regulation by the Cdc25 phosphatase family. *Progress in Cell Cycle Research*, 4:107-114.
[25] Owzar, K., Jung, S.-H., and Sen, P.K. (2007) A Copula Approach for Detecting Prognostic Genes Associated With Survival Outcome in Microarray Studies. *Biometrics*, 63, 1089-1098.
[26] Raftery, AE. and Lewis, SM. (1992) One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7, 493-497.
[27] Raftery, AE. and Lewis, SM. (1995) The number of iterations, convergence diagnostics and generic Metropolis algorithms in Practical Markov Chain Monte Carlo. *Chapman and Hall, London.*
[28] Reverter, A. and Chan, EKF. (2008) Combining partial correlation and an information theory approach to the reversed-engineering of gene co-expression networks. *Bioinformatics,* 24(21):2491-2497.
[29] Rosengard, AM, Krutzsch, HC, Shearn, A, Biggs, JR, Barker, E, Margulies, IM, King, CR, Liotta, LA, Steeg, PS (1989) Reduced Nm23/Awd protein in tumour metastasis and aberrant Drosophila development. *Nature*, 342(6246), 177-180.
[30] Ross, DT, Scherf, U, Eisen, MB, Perou, CM, Rees, C, Spellman, P, Iyer, V, Jeffrey, SS, Van de Rijn, M, Waltham, M, Pergamenschikov, A, Lee, JC, Lashkari, D, Shalon, D, Myers, TG, Weinstein, JN, Botstein, D, Brown, PO (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3), 227-235. NCI60 data are downloadable from: http://genome-www.stanford.edu/nci60/.
[31] Subramanian, A., Tamayo, P., Mootha, VK., Mukherjee, S., Ebert, BL., Gillette, MA., Paulovich, A., Pomeroy, SL., Golub, TR., Lander, ES., Mesirov, JP. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102, 15545–15550.
[32] Tian, L., Greenberg, SA., Kong, SW., Altschuler, J., Kohane, IS., Park, PJ. (2005) Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102, 13544–13549.
[33] Spearman, C. (1904) The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101
[34] Stahl, JA, Leone, A, Rosengard, AM, Porter, L, King, CR, Steeg, PS (1991) Identification of a second human nm23 gene, nm23-H2. *Cancer Research*, 51(1), 445-449.
[35] St. Claire, S, Giono, L, Varmeh-Ziaie, S, Resnick-Silverman, L, Liu, WJ, Padi, A, Dastidar, J, DaCosta, A, Mattia, M, Manfredi, JJ (2004) DNA damage-induced downregulation of Cdc25C is mediated by p53 via two independent mechanisms: one involves direct binding to the cdc25C promoter. *Molecular Cell*, 16(5), 725-736.
[36] Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J., Ladd-Acosta, C., Mesirov, J., Golub, T. and Jacks, T. (2005) An oncogenic KRAS2 expression signature identi_ed by cross-species gene-expression analysis. *Nature Genetics*, 37, 48-55.
[37] Wei, Z. and Li, H. (2007) A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23 (12), 1537-1544.
[38] Yuan, A., Chen, G., Zhou, Z., Bonney, G., and Rotimi, C. (2008) Gene Copy Number Analysis for Family Data Using Semiparametric Copula Model. *Bioinformatics and Biology Insights*, 2, 349-361.

Table 1: The number of connections identified via copula correlations out of 100 repeated simulated networks. *P* is the proportion of genes that is generated from the mixture Gaussian distribution corresponding to the outcomes. *r* is the linear correlation between two binary outcomes. Overlap = T means that gene sets X_1 and X_2 share 50% of the genes that are generated from standard Gaussian distribution. $X_1 \rightarrow Y_1$ indicates that X_1 is a statistically significant predictor for Y_1.

gene set size = 10			X_1->Y_1	X_2->Y_1	X_2->Y_2	X_1->Y_2	Y_2->Y_1	Y_1->Y_2	X_2->X_1	X_1->X_2
P=0.5	*r*=0.76	overlap=F	100	0	100	0	35	59	10	11
P =0.5	*r*=0.76	overlap=T	100	0	100	0	53	50	12	19
P =0.5	*r*=0.33	overlap=F	99	2	100	1	0	0	0	0
P =0.5	*r*=0.33	overlap=T	98	1	100	1	0	0	0	2
P =0.1	*r*=0.76	overlap=F	40	18	41	18	100	99	8	14
P =0.1	*r*=0.76	overlap=T	39	18	43	22	100	100	8	8
P =0.1	*r*=0.33	overlap=F	71	20	94	14	47	50	7	10
P =0.1	*r*=0.33	overlap=T	66	24	93	11	51	52	8	11
gene set size = 30			X_1->Y_1	X_2->Y_1	X_2->Y_2	X_1->Y_2	Y_2->Y_1	Y_1->Y_2	X_2->X_1	X_1->X_2
P=0.5	*r*=0.76	overlap=F	100	0	100	0	34	53	10	9
P =0.5	*r*=0.76	overlap=T	100	0	100	0	36	52	12	14
P =0.5	*r*=0.33	overlap=F	100	3	100	0	1	1	0	1
P =0.5	*r*=0.33	overlap=T	99	3	100	0	1	0	0	2
P =0.1	*r*=0.76	overlap=F	89	6	93	6	82	85	15	18
P =0.1	*r*=0.76	overlap=T	87	5	88	7	87	89	25	20
P =0.1	*r*=0.33	overlap=F	90	9	100	14	12	14	11	13
P =0.1	*r*=0.33	overlap=T	94	6	100	17	8	17	12	11
gene set size = 100			X_1->Y_1	X_2->Y_1	X_2->Y_2	X_1->Y_2	Y_2->Y_1	Y_1->Y_2	X_2->X_1	X_1->X_2
P=0.5	*r*=0.76	overlap=F	100	0	100	0	34	51	13	12
P =0.5	*r*=0.76	overlap=T	100	0	100	0	46	51	7	16
P =0.5	*r*=0.33	overlap=F	100	3	100	1	1	0	1	2
P =0.5	*r*=0.33	overlap=T	99	2	100	1	4	0	0	1
P =0.1	*r*=0.76	overlap=F	100	0	100	0	42	46	9	11
P =0.1	*r*=0.76	overlap=T	100	0	100	0	37	54	11	14
P =0.1	*r*=0.33	overlap=F	96	0	100	3	3	0	1	2
P =0.1	*r*=0.33	overlap=T	100	1	100	1	0	0	0	0

Table 2: The marginal correlations averaged over 100 simulated datasets. $X_1 \sim Y_1$ shows the averaged marginal correlation between X_1 and Y_1.

gene set size = 10			X_1~Y_1	X_2~Y_1	X_2~Y_2	X_1~Y_2	Y_1~Y_2	X_1~X_2
P=0.5	*r*=0.76	overlap=F	1.00	1.00	1.00	1.00	1.00	0.99
P =0.5	*r*=0.76	overlap=T	1.00	1.00	1.00	1.00	1.00	0.99
P =0.5	*r*=0.33	overlap=F	1.00	0.84	1.00	0.89	0.89	0.82
P =0.5	*r*=0.33	overlap=T	1.00	0.79	1.00	0.91	0.87	0.85
P =0.1	*r*=0.76	overlap=F	1.00	0.99	1.00	0.99	1.00	0.84
P =0.1	*r*=0.76	overlap=T	1.00	0.99	1.00	0.99	1.00	0.83
P =0.1	*r*=0.33	overlap=F	1.00	0.81	0.99	0.91	0.98	0.69
P =0.1	*r*=0.33	overlap=F	1.00	0.79	0.99	0.88	0.98	0.67
gene set size = 30			X_1~Y_1	X_2~Y_1	X_2~Y_2	X_1~Y_2	Y_1~Y_2	X_1~X_2
P=0.5	*r*=0.76	overlap=F	1.00	1.00	1.00	1.00	1.00	0.99
P =0.5	*r*=0.76	overlap=T	1.00	1.00	1.00	1.00	1.00	0.99
P =0.5	*r*=0.33	overlap=F	1.00	0.83	1.00	0.90	0.89	0.83
P =0.5	*r*=0.33	overlap=T	1.00	0.86	1.00	0.92	0.91	0.86
P =0.1	*r*=0.76	overlap=F	1.00	1.00	1.00	1.00	1.00	0.98
P =0.1	*r*=0.76	overlap=T	1.00	1.00	1.00	1.00	1.00	0.98
P =0.1	*r*=0.33	overlap=F	1.00	0.84	1.00	0.93	0.95	0.82
P =0.1	*r*=0.33	overlap=F	1.00	0.84	1.00	0.94	0.95	0.79
gene set size = 100			X_1~Y_1	X_2~Y_1	X_2~Y_2	X_1~Y_2	Y_1~Y_2	X_1~X_2
P=0.5	*r*=0.76	overlap=F	1.00	1.00	1.00	1.00	1.00	0.99
P =0.5	*r*=0.76	overlap=T	1.00	1.00	1.00	1.00	1.00	0.99
P =0.5	*r*=0.33	overlap=F	1.00	0.83	1.00	0.90	0.90	0.84
P =0.5	*r*=0.33	overlap=T	1.00	0.86	1.00	0.92	0.91	0.86
P =0.1	*r*=0.76	overlap=F	1.00	1.00	1.00	1.00	1.00	0.99
P =0.1	*r*=0.76	overlap=T	1.00	1.00	1.00	1.00	1.00	0.99
P =0.1	*r*=0.33	overlap=F	1.00	0.84	1.00	0.93	0.90	0.85
P =0.1	*r*=0.33	overlap=F	1.00	0.84	1.00	0.91	0.91	0.84

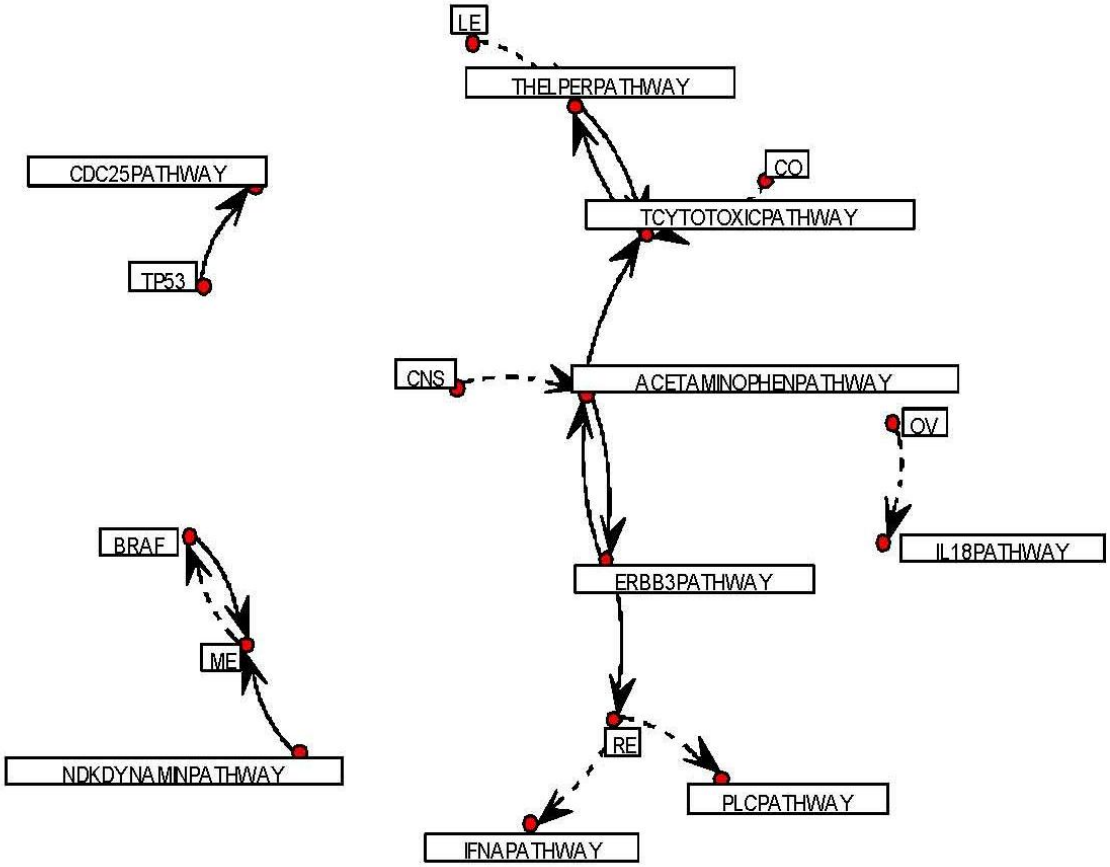

Figure 2: The plot of network based on NCI60 data, which presents the interactions between phenotype, genotypes of mutations and pathways associations. The phenotypes include BR, CNS, CO, LE, ME, LC, OV and RE cancers. The genotypes of mutations include KRAS, BRAF, and TP53 mutations. The arrows of i → j indicates the i^{th} variable is a non-negligible predictor for the j^{th} variable. The solid line is the direction from pathways or mutations to phenotypes and the dash line is the direction from phenotypes to pathways or mutations.

Table 3: Copula correlation among variables of interest: posterior probability regions (based on 2.5th and 97.5th quartiles) do not cover zero are highlighted in Bold typeface. Correlations of 17 variables out of 55 are listed and those 17 variables are the ones plotted in Figure 2. First 8 variables (BRAF to RE) are categorical outcome variables and the rest are continuous variables with pathway names.

Copula	BRAF	TP53	CNS	CO	LE	ME	OV	RE	ACETAMINOPHEN
BRAF	1.00	-0.08	-0.04	0.16	-0.12	**0.46**	-0.15	-0.09	-0.12
TP53	-0.08	1.00	0.06	0.16	0.14	-0.20	-0.03	-0.16	-0.02
CNS	-0.04	0.06	1.00	-0.16	-0.14	-0.01	-0.08	0.04	**-0.32**
CO	0.16	0.16	-0.16	1.00	-0.17	-0.02	0.04	-0.08	0.03
LE	-0.12	0.14	-0.14	-0.17	1.00	-0.13	-0.15	-0.31	0.00
ME	**0.46**	-0.20	-0.01	-0.02	-0.13	1.00	-0.06	-0.07	-0.09
OV	-0.15	-0.03	-0.08	0.04	-0.15	-0.06	1.00	0.03	0.14
RE	-0.09	-0.16	0.04	-0.08	-0.31	-0.07	0.03	1.00	0.12
ACETAMINOPHEN	-0.12	-0.02	**-0.32**	0.03	0.00	-0.09	0.14	0.12	1.00
CDC25	0.06	**0.44**	0.06	0.26	0.25	-0.22	-0.18	-0.33	**-0.27**
ERBB3	-0.09	-0.04	0.04	0.14	**-0.40**	-0.14	0.15	**0.47**	0.19
IFNA	0.07	-0.10	0.12	-0.23	-0.11	0.02	-0.14	**0.53**	0.06
IL18	-0.25	-0.11	-0.09	0.08	-0.06	-0.18	**0.34**	0.06	0.14
NDKDYNAMIN	-0.18	0.08	**0.38**	-0.04	**-0.35**	-0.23	-0.03	**0.38**	-0.15
PLC	0.03	0.14	0.00	-0.19	**0.45**	0.05	-0.11	**-0.50**	-0.13
TCYTOTOXIC	-0.28	0.03	0.00	**-0.45**	**0.62**	-0.18	-0.17	-0.15	0.00
THELPER	-0.28	0.04	-0.01	**-0.45**	**0.62**	-0.18	-0.17	-0.16	0.01

Copula	CDC25	ERBB3	IFNA	IL18	NDKDYNAMIN	PLC	TCYTOTOXIC	THELPER
BRAF	0.06	-0.09	0.07	-0.25	-0.18	0.03	-0.28	-0.28
TP53	**0.44**	-0.04	-0.10	-0.11	0.08	0.14	0.03	0.04
CNS	0.06	0.04	0.12	-0.09	**0.38**	0.00	0.00	-0.01
CO	0.26	0.14	-0.23	0.08	-0.04	-0.19	**-0.45**	**-0.45**
LE	0.25	**-0.40**	-0.11	-0.06	**-0.35**	**0.45**	**0.62**	**0.62**
ME	-0.22	-0.14	0.02	-0.18	-0.23	0.05	-0.18	-0.18
OV	-0.18	0.15	-0.14	**0.34**	-0.03	-0.11	-0.17	-0.17
RE	-0.33	**0.47**	**0.53**	0.06	**0.38**	**-0.50**	-0.15	-0.16
ACETAMINOPHEN	**-0.27**	0.19	0.06	0.14	-0.15	-0.13	0.00	0.01
CDC25	1.00	**-0.25**	-0.18	-0.19	-0.01	**0.29**	0.01	0.00
ERBB3	**-0.25**	1.00	**0.34**	0.15	**0.39**	**-0.51**	**-0.29**	**-0.29**
IFNA	-0.18	**0.34**	1.00	-0.11	**0.32**	**-0.28**	0.06	0.05
IL18	-0.19	0.15	-0.11	1.00	0.01	-0.21	-0.05	-0.05
NDKDYNAMIN	-0.01	**0.39**	**0.32**	0.01	1.00	**-0.37**	-0.20	-0.21
PLC	0.29	**-0.51**	**-0.28**	-0.21	**-0.37**	1.00	**0.38**	**0.39**
TCYTOTOXIC	0.01	**-0.29**	0.06	-0.05	-0.20	**0.38**	1.00	**0.95**
THELPER	0.00	**-0.29**	0.05	-0.05	-0.21	**0.39**	**0.95**	1.00