Using Copulas to Select Prognostic Genes in Melanoma Patients

Linda Chaba[a], John Odhiambo[a] and Bernard Omolo[b,*]

[a]Strathmore Institute of Mathematical Sciences, Strathmore University, Ole Sangale Road, Nairobi, Kenya

[b]Division of Mathematics and Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA

Abstract: Melanoma of the skin is the fifth and seventh most commonly diagnosed carcinoma in men and women, respectively, in the USA. So far, gene signatures prognostic for overall and distant metastasis-free survival, for example, have been promising in the identification of therapeutic targets for primary and metastatic melanoma. But most of these gene signatures have been selected using statistics that depend entirely on the parametric distributions of the data (e.g. *t*-statistics). In this study, we assessed the impact of relaxing the parametric assumptions on the power of the models used for gene selection. We developed a semi-parametric model for feature selection that does not depend on the distributions of the covariates. This copula-based model only assumed that the marginal distributions of the covariates are continuous. Simulations indicated that the copula-based model had reasonable power at various levels of the false discovery rate (FDR). These results were validated in a publicly-available melanoma dataset. Relaxing parametric assumptions on microarray data may yield procedures that have good power for differential gene expression analysis.

Keywords: Copula, False discovery rate, Melanoma, Microarray, Power.

1. BACKGROUND

Melanoma of the skin is among the most common cancer types in the United States. It is the fifth and seventh most commonly diagnosed carcinoma in men and women, respectively [1]. A major challenge with melanoma is the identification of therapeutic targets. Multi-gene signatures have shown promise in this regard and a number of these signatures have been developed within the last decade [2-9].

The development of such gene signatures require the use of statistical methods. A number of studies [7-9] used parametric methods based on the *t*-test with multiple corrections. One advantage of these methods is that they offer a straightforward approach to calculating *p*-values and confidence intervals. Moreover, for large samples, the distribution of the *t*-statistics is independent of the overall expression level of the gene. Unfortunately, for small sample sizes, the *t*-test based methods depend on strong parametric assumptions. These assumptions may be violated in practice, and so non-parametric methods have also been applied in some studies [3-5]. For these methods, the distribution of random errors are estimated without strong parametric assumptions. The significance analysis of microarrays (SAM) method [10], in particular, avoids high variance that results from estimating the variance of each gene separately. When the sample size is small, any method that reduces the variance in the estimates is more accurate. The non-parametric methods also have disadvantages which vary from one method to the other. For example, the Wilcoxson-test approach exhibits low power in the identification of differentially expressed genes [11]. For a detailed review of methods for finding differentially expressed genes, see [11-14].

Despite all the proposed methods mentioned above, there is no unanimous agreement on any particular gene selection method. Furthermore, most methods were developed for finding differentially expressed genes based on groups or classes of the samples (discrete covariates). However, there are many covariates of interest in microarrays that are continuous in nature.

This study proposed an algorithm for selecting genes associated with a continuous but non-clinical outcome based on a semi-parametric copula model. A copula can be loosely defined as a function that joins together the marginal distributions to the joint distribution. It is semi-parametric in the sense that no assumptions are made on the marginal distributions but the dependence parameter is assumed to come from a parametric family [15]. An advantage of the copula-based approach is its compatibility with any distribution function. This allows for the relaxation of the assumption of specific distributions, which is made in most of the existing methods. A possible setback of the copula approach for finding differentially expressed genes may be the use of an assumed copula. So far no method for selecting an optimum copula for analyzing gene expression data has been developed, hence the use of an assumed copula.

*Address correspondence to this author at the Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, South Carolina, USA; Tel: +1 864-503-5362; Fax: +1 864-503-5930; E-mail: bomolo@uscupstate.edu

E-ISSN: 1929-6029/17

Owzar *et al*. [16] had applied a copula-based approach to identify genes that are differentially expressed between stage I and III lung cancer patients, based on survival copulas and family-wise error rate (FWER) control. In contrast, our proposed algorithm is based on a continuous outcome from melanoma cell lines and controls for the false discovery rate (FDR), since the FWER is often too conservative [17].

The performance of our copula-based approach in terms of power was assessed via simulations. The method was then applied to a melanoma cell lines dataset to select genes that are correlated with the G_2 checkpoint function. The gene signature generated by the copula approach was then subjected to an independent primary melanoma dataset to determine if it is prognostic of 4-year distant metastasis-free survival in melanoma patients.

2. METHODS

2.1. Copula Density and Likelihood Function

A copula is a bivariate distribution with uniform marginals. By Sklar's theorem [18], for any distribution function, F, with marginals F_1 and F_2, there exists a copula, C, such that

$$F(x,y) = C[F_1(x_1),F_2(y)] \tag{1}$$

for $(x,y)'$ in the support of F. This result can be easily extended to multivariate distributions, to yield Sklar's theorem in *m*-dimensions, which we now state (without proof):

Let F be an *m*-dimensional distribution function with margins $F_1(x_1),...,F_m(x_m)$. Then there exists an *m*-copula, C, such that for all $\mathbf{x} = (x_1,x_2,...,x_m)' \in \bar{\mathbb{R}}$,

$$F(x_1,...,x_m) = C\left[F_1(x_1),...,F_m(x_m)\right]. \tag{2}$$

If $F_1,F_2,...,F_m$ are all continuous, then C is unique and can be expressed as

$$C(u_1,u_2,...,u_m) = F(F_1^{-1}(u_1),F_2^{-1}(u_2),...,F_m^{-1}(u_m)), \tag{3}$$

for any $\mathbf{u} = (u_1,u_2,...,u_m)' \in [0,1]^m$.

Upon differentiation, (2) becomes

$$f(x_1,x_2,...,x_m) = \frac{\partial^m C(F_1(x_1),...,F_m(x_m))}{\partial F_1(x_1)..\partial F_m(x_m)} \prod_{i=1}^{m} \frac{dF_i(x_i)}{dx_i}$$

$$= c(F_1(x_1),...,F_m(x_m))\prod_{i=1}^{m} f_i(x_i). \tag{4}$$

Here, f, c and f_i are the densities for F, C and F_i, $i = 1,2,...,m$, respectively. Let $u_i = F_i(x_i)$, $i = 1,2,...,m$. Then (4) becomes

$$L(\theta) = f(x_1,x_2,...,x_m) = c(u_1,u_2,...,u_m)\prod_{i=1}^{m} f_i(x_i). \tag{5}$$

One can fit a copula model by estimating its parameters using the maximum likelihood approach. In practice, it is more convenient to work with the logarithm of a likelihood function because it simplifies subsequent mathematical analysis. Since the logarithm is a monotonically increasing function, maximizing the log of a function is the same as maximizing the function itself. The log-likelihood is given as

$$\ell_n(\theta) = \sum_{j=1}^{n} log\, c(F_1(x_1),...,F_m(x_m)) + \sum_{j=1}^{n}\sum_{i=1}^{m} log(f_i(x_i)), \tag{6}$$

and the estimate of θ is

$$\hat{\theta} = \arg\max \ell_n(\theta). \tag{7}$$

This is equivalent to finding a solution to

$$\frac{1}{n}\frac{\partial}{\partial\theta}\ell_n(\theta). \tag{8}$$

Since the marginals are unknown, each $F_i(x_i)$ may be replaced with its marginal estimator $\hat{F}_i(x_i)$ to obtain $\hat{\theta}_i$. This approach is referred to as the canonical maximum likelihood estimation (CMLE) method [15]. Here, $\hat{F}_i(x_i)$ is given by

$$\hat{F}_i(x_i) = \frac{n}{n+1}\frac{1}{n}\sum_{j=1}^{n} I\left(X_i \le x_i\right), \tag{9}$$

where I is the indicator function. Rescaling the empirical distribution by $\frac{n}{n+1}$ avoids the potential unboundedness of $log[c(F_1(x_1),...,F_m(x_m)]$ as some of the $F_i(x_i)$'s tend to be one [15]. The corresponding pseudo-loglikelihood is given as

$$\ell_n^*(\theta) = \sum_{j=1}^{n} log\, c(\hat{F}_1(x_1),...,\hat{F}_m(x_m)) + \sum_{j=1}^{n}\sum_{i=1}^{m} log(f_i(x_i)), \tag{10}$$

and the estimate of θ is

$$\hat{\theta} \approx \arg\max \sum_{j=1}^{n} \log\, c(\hat{F}_1(x_1),...,\hat{F}_m(x_m)), \tag{11}$$

since the last summand in (10) does not depend on θ. Under suitable regularity conditions, $\hat{\theta}$ is consistent and is asymptotically normal [15]. In general, multivariate models do not have closed form estimators and, therefore, numerical methods are required in the estimation process [19].

2.2. Copula Model for Differential Gene Expression

We were interested in the pairwise correlation between each gene's expression profile and a quantitative outcome. Therefore, the copula of interest was the bivariate copula ($m=2$). Suppose a microarray experiment consists of n subjects/samples and G genes. Let $\mathbf{x_i} = (x_{1i},...,x_{ni})'$ be the gene expression profile for gene i and $\mathbf{y} = (y_1,...,y_n)'$ be a vector of the covariate of interest (quantitative trait). We wanted to find K genes that are correlated with $\mathbf{y}$, $0 < K < G$. That is, we were interested in determining whether, for each gene i, $\mathbf{x_i}$ and $\mathbf{y}$ are independent or not. The test for independence, thus, becomes testing for the null hypothesis

$$H_{0i} : Y \perp X_i \quad (X_i \text{ and } Y \text{ are independent}), \tag{12}$$

against the alternative hypothesis

$$H_{1i} : Y \not\perp X_i \quad (X_i \text{ and } Y \text{ are independent}). \tag{13}$$

For multiple genes, (12) is tested simultaneously and so the hypothesis of interest becomes

$$H_0 : Y \perp X_i \text{ for all } i = \bigcap_{i=1}^{G} H_{0i} \tag{14}$$

vs.

$$H_1 : Y \not\perp X_i \text{ for some } i = \bigcup_{i=1}^{G} H_{1i}. \tag{15}$$

In terms of copulas, assume that for each gene i, the joint distribution of Y and X_i is generated by a parametric copula $C(u_1,u_2;\theta_i)$ such that

$$H_i(y,x) = C[F(y), F_i(x); \theta_i], \tag{16}$$

where $H_i(y,x)$, $F(y)$ and $F_i(x)$ are the CDFs of (Y,X_i), Y and X_i respectively. Here $u_1 = F(y)$, $u_2 = F_i(x)$ and θ_i is the dependence parameter. Equation (14) and (15) now becomes

$$H_0 : \bigcap_{i=1}^{G} \left[C(u_1,u_2;\theta_i) = u_1 u_2 \text{ for all } (u_1,u_2)^T \in [0,1]^2 \right], \tag{17}$$

vs.

$$H_1 : \bigcup_{i=1}^{G} \left[C(u_1,u_2;\theta_i) \neq u_1 u_2 \text{ for some } (u_1,u_2)^T \in [0,1]^2 \right]. \tag{18}$$

A normal copula, for instance, attains independence when θ_i = 0. In this case, the global hypothesis to test for the dependence in terms of θ_i is expressed as

$$H_0 : \bigcap_{i=1}^{G} (\theta_i = 0) \;\; vs. \;\; H_1 : \bigcup_{i=1}^{G} (\theta_i \neq 0). \tag{19}$$

2.3. Hypothesis Testing

To test (19), we needed the distribution of $\hat{\theta}_i$ under the null hypothesis. Rather than assume a parametric distribution for the null hypothesis, we used a permutation resampling based approach [20]. For a given α, a gene is differentially expressed if its *p*-value is less than $< \alpha$. To adjust for multiple comparisons, the false discovery rate (FDR) approach [21] was used. The global null hypothesis (19) is rejected if at least one of its components (H_{0i}) is rejected, based on the estimated FDR values.

2.4. Copula Algorithm for Identifying DEGs

Our copula-based algorithm for finding differentially expressed genes (DEGs) can be summarised as follows:

1. Estimate θ_i using the CMLE method. In the CMLE approach, no assumption is made on the marginal distribution. The marginal distribution for each gene, $F_i(x_i)$ and a quantitative outcome, $F(y)$, are replaced with their estimators $\hat{F}_i(x_i)$ and $\hat{F}(y)$, respectively, to obtain $\hat{\theta}_i$.

$$\hat{\theta}_{\mathbf{i}} \approx \arg\max \sum_{j=1}^{n} \log c(\hat{F}_i(x_i), \hat{F}(y)). \tag{20}$$

A detailed explanation of the CMLE method is provided in the Supplementary Materials (B).

2. Find gene-specific *p*-values (unadjusted *p*-values) using the permutation based resampling method. See Supplementary Materials (C) for details.

3. Apply the FDR approach to control for type I error. See Supplementary Materials (D) for details.

4. A gene is differentially expressed if its estimated FDR (estimated *q*-value) is less than some specified value $\alpha \in$ [0,1].

An R code for implementing the algorithm is available from the authors upon request.

2.5. Simulations

Twelve simulation scenarios were considered in evaluating the performance of the proposed copula method in terms of power. Let n and G denote the number of samples and genes, respectively. Further, let D denote the number of genes assumed to be truly differentially expressed. Then $(G-D)$ genes are assumed to be non-differentially expressed. The gene expression data matrix, $\mathbf{X}$, is a $G \times n$ matrix of log2-ratios. We can write $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$, where $\mathbf{X_1}$ and $\mathbf{X_2}$ are $D \times n$ and $(G-D) \times n$ matrices, respectively. We set $D \in (50,100,200)$, $n \in (20,35,50,100)$ and G to be 1000. We generated the $(1000-D)$ genes from the standard normal distribution. To generate the D genes, we used the standard normal distribution in conjunction with the Cholesky decomposition [22] of their correlation matrix as follows:

1. Generate an unstructured correlation matrix $\mathbf{\Omega}$. $\mathbf{\Omega}$ is a $(D+1)\times(D+1)$ matrix that has $(i,j)^{th}$ element given by $\omega_{i,j} = \text{corr } (x_i, x_j)$.

2. Find the Cholesky factor, $\mathbf{A}$, of $\mathbf{\Omega}$ such that $\mathbf{\Omega} = \mathbf{AA}'$.

3. Let $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I}_n), i = 1,2,...,(D+1)$.

4. $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{D+1})'$.

5. $\mathbf{X}_{D+1} = \mathbf{AZ}$.

$\mathbf{X}_{D+1}$ is the gene expression matrix for D genes assumed to be differentially expressed and a covariate $\mathbf{y}$. $\mathbf{y}$ can take any of the $D+1$ row vectors from the matrix $\mathbf{X}_{D+1}$. $\mathbf{X}_1$ is therefore a submatrix of $\mathbf{X}_{D+1}$ with dimensions $D \times n$.

The developed copula method was applied to the 12 simulated datasets to evaluate its power in identifying DEGs. We transposed $\mathbf{X}$ in the analysis, so that $\mathbf{X}'$ had genes on the columns and samples on the rows. We followed the procedure in section 2.4 to identify DEGs at different estimated FDR values. A normal copula was assumed. See Supplementary Materials (A) for the description of the normal copula. Power was calculated as the ratio of the number of correctly identified differentially expressed genes, true positives (TP), to the total number of actual DEGs, D. Thus,

$$\text{Power} = \frac{TP}{D} \tag{21}$$

2.6. Application

The proposed copula model was applied to a publicly available melanoma dataset. This dataset contained gene expression data (raw intensities) on 54 cell-lines (35 melanoma cell lines and 19 normal human melanocytes (NHMs), each with 45,015 probes. Only the melanoma cell lines were analyzed. The raw data was median-normalized and log2-transformed. Multiple probes were reduced to one per gene by using the most variable probe(set)–measured by interquartile range (IQR)–across arrays. Filtration and normalization of the gene expression data is implemented using BRB Array Tools software [23]. A gene was filtered out if less than 20% of its expression data values had at least 1.5-fold change in either direction from the genes median value. Genes with more than 50% missing data across all its samples were also filtered out. There were 3,860 genes available for subsequent analysis.

We used the G_2 checkpoint function to quantify the biological process in melanoma progression. Omolo *et al*. [8] found the G_2 checkpoint function to be prognostic for the development of distant metastasis, hence its choice for this study. The G_2 checkpoint is a position of control in the cell cycle that delays or arrests mitosis when DNA damage by radiation is detected. It prevents cells with damaged DNA from entering mitosis, thereby providing the opportunity for repair and stopping the proliferation of damaged cells. The G_2 checkpoint function in melanoma cell-lines were scored as a ratio of mitotic cells in 1.5 Gy ionizing radiation (IR)-treated cultures in comparison to their sham-treated control (i.e. IR to sham ratio) [8].

A normal copula was assumed for the analysis of the melanoma dataset. Such an assumption was previously made in [16] for lung cancer. Genes that were correlated with the G_2 checkpoint function were selected based on their estimated FDR values. To check the biological significance of the G_2 signature generated by the copula method, we used the independent dataset in [2] to identify genes that could predict a patient's risk (low/high) for developing distant metastasis within 4 years of primary diagnosis. The supervised principal component method [24] was used to separate the samples into high/low risk group. The procedure was implemented by BRB-ArrayTools software [23].

3. RESULTS AND DISCUSSION

Figure **1** shows the top differentially expressed genes for each of the simulated datasets, while Table **1** provides the number of DEGs obtained at various

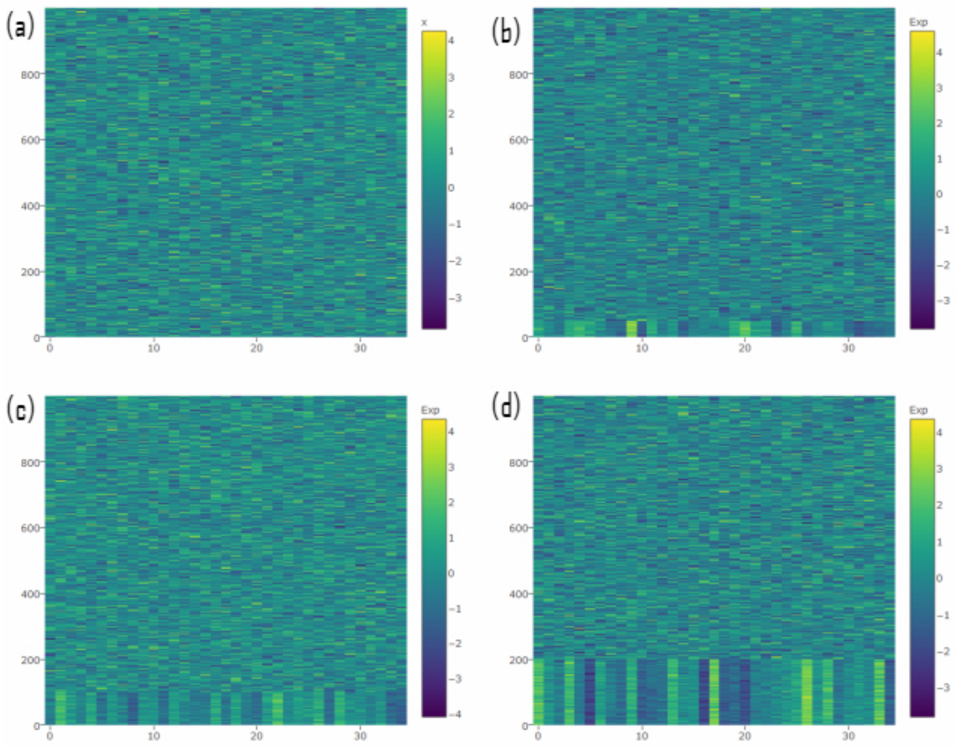

Figure 1: Heatmaps using simulated data. Each simulated dataset contains 1000 genes and $n = 35$ samples. (**a**). No assumption is made on the number of DEGs, D; (**b**). $D = 50$; (**c**). $D = 100$; and (**d**). $D = 200$. The top DEGs are at the bottom of the list of 1000 genes.

Table 1: DEGs at FDR Level between 0.001 and 0.2 on Twelve Simulated Datasets each with 1000 Genes. Sample Size was Set at $n \in \{20,35,50,100\}$. Number of Significant Genes were Set to be 50, 100 and 200.

n	D	Estimated FDR Threshold					
		0.001	**0.01**	**0.025**	**0.05**	**0.1**	**0.2**
20	50	30	30	30	44	49	59
	100	79	92	95	103	117	132
	200	137	184	197	209	222	251
35	50	49	49	52	53	54	62
	100	99	101	101	105	110	121
	200	201	201	204	209	222	258
50	50	50	50	51	51	53	61
	100	100	101	102	106	112	124
	200	202	205	210	216	223	249
100	50	50	50	50	50	55	58
	100	100	101	101	107	111	136
	200	201	201	208	212	226	259

levels of estimated FDR. From Table **1**, it can be seen that as the estimated FDR values increases, more genes are identified as being differentially expressed. For example, for $n = 35$, $D = 50$ and $FDR = 0.05$, 49 genes are identified and at $FDR = 0.1$ for the same n and D, 52 genes are identified. The same pattern is seen for the other values of n. This shows that the proposed algorithm is stable as the known DEGs selected at a lower FDR threshold are all contained in the genes selected at a higher FDR threshold.

Table **2** shows the power of the copula method at different estimated FDR levels for four sample sizes: n = 20, 35, 50 and 100. The results show that the power

Table 2: Power at FDR Level between 0.001 and 0.2 on Six Simulated Datasets each with 1000 Genes. Sample Size was Set at $n \in \{20,35,50,100\}$. Number of Significant Genes were Set to be 50, 100 and 200.

n	D	Estimated FDR Threshold					
		0.001	0.01	0.025	0.05	0.1	0.2
20	50	0.58	0.58	0.58	0.80	0.88	0.96
	100	0.79	0.92	0.95	0.98	0.99	1.00
	200	0.69	0.92	0.98	1.00	1.00	1.00
35	50	0.98	0.98	1.00	1.00	1.00	1.00
	100	0.99	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00
50	50	1.00	1.00	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00
100	50	1.00	1.00	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00
	200	1.00	1.00	1.00	1.00	1.00	1.00

of the copula method is sensitive to low sample sizes. For example, the power is 0.58 when $n = 20$ at $D = 50$ but increases to 1 for the same value of D as n increases to 100. For the sample size of at least 35, the least value of the power observed from the analysis is 0.98. This shows that the copula method is quite powerful in finding differentially expressed genes. The copula approach is also robust to sample size, especially as the number of known DEGs increases.

Even though Owzar *et al*. [16] and our study both assessed the power of the test, a direct comparison of the conclusion on the power of the test may not be appropriate. Owzar *et al*. [16] assessed the power by using a constant sample size ($n = 50$) and constant number of DEGs ($D = 5$) but varied the expected proportion of censored observations, correlation between the gene expression profiles and correlation between the prognostic genes and the time-to-event variable. They concluded that power was affected by the expected proportion of censored observations and that an increase in the expected proportion of censored observations lead to a decrease in the power of the test. Our study assessed the power of the test by varying the number of genes that were assumed to be differentially expressed, the sample size and the estimated FDR thresholds. Our approach is less conservative compared to Owzar *et al*. [16] and hence could be more appealing in its own right.

When applied to the cell lines dataset, the copula method identified 9 genes at FDR < 0.01 and 25 genes at FDR < 0.2. Table **3** lists the genes that are correlated with G_2 checkpoint function at FDR < 0.2. Annotation of the 25 genes was performed using the Database for Annotation, Visualization, and Integrity Discovery (DAVID) [25]. Seven of them did not have DAVID identifiers and were hence “unknown". We compared our results and the results presented in [8]. They found 165 genes that were correlated with G_2 checkpoint function. The 165 unique genes were generated by two methods: a Bayesian approach and the quantitative trait analysis (QTA). The overlapping genes between our 25 genelist and their 165 genelist were *ZNF711, DGKE* and *ARNTL2* (Figure **2**). It is noteworthy that the QTA method applied in [8] did not adjust for multiplicity. Therefore, a direct comparison of the two genelists may not be appropriate.

The copula genelist was also subjected to class prediction of the G_2 checkpoint function, using the least absolute shrinkage and selector operator (LASSO) method [26]. Results show that the genelist could predict well ($R^2 = 0.311$, *p* = 0.00494). We further subjected our genelist to a survival risk prediction (SRP) analysis to assess its biological importance. Our list generated 4 prognostic genes, which shows a significant separation of the samples into low and high risk group (χ^2 = 5.9, *p* = 0.0147) (Figure **3**). Similar results are reported in [8], using their 32 prognostic genes (χ^2 = 5.6, *p* = 0.018). Our list of 25 genes performed better in SRP than a randomly selected 25 genes from the 3860 genes (χ^2 = 0.1, *p* = 0.655). However, unsupervised hierarchical clustering indicated no significant (*p* = 0.317) separation of the samples for distance metastasis-free survival (Figure **3**).

Only one gene, *ZNF711*, was found to overlap between the two sets of Cox genes. This gene,

Table 3: List of Genes that are Correlated with the G_2 Checkpoint Function as Selected by the Copula Approach at FDR < 0.2.

Agilent ID	Symbol	Gene Name
A_23_P14612	FGF7	fibroblast growth factor 7 (FGF7)
A_23_P153964	INHBB	inhibin beta B subunit (INHBB)
A_23_P203115	TMEM25	transmembrane protein 25 (TMEM25)
A_23_P211631	FBLN1	fibulin 1 (FBLN1)
A_23_P214080	EGR1	early growth response 1 (EGR1)
A_23_P217297	ZNF711	zinc finger protein 711 (ZNF711)
A_23_P364504	ERFE	Erythroferrone (ERFE)
A_23_P369328	C10orf35	chromosome 10 open reading frame 35 (C10orf35)
A_23_P389250	Smco2	single-pass membrane protein with coiled-coil domains 2 (SMCO2)
A_23_P393034	HAS3	hyaluronan synthase 3 (HAS3)
A_23_P69537	NMU	neuromedin U (NMU)
A_24_P130952	MLK4	mixed lineage kinase 4 (MLK4)
A_24_P196665	GNGT1	G protein subunit gamma transducin 1 (GNGT1)
A_24_P20814	KHDC1L	KH domain containing 1 like (KHDC1L)
A_32_P209230	CITED4	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 4 (CITED4)
A_32_P232559	PRKCQ-AS1	PRKCQ antisense RNA 1 (PRKCQ-AS1)
A_32_P399546	ARNTL2	aryl hydrocarbon receptor nuclear translocator like 2 (ARNTL2)
A_32_P540991	DGKE	diacylglycerol kinase epsilon (DGKE)
A_23_P153958	Unknown	Unknown
A_32_P134427	Unknown	Unknown
A_32_P154726	Unknown	Unknown
A_32_P190343	Unknown	Unknown
A_32_P227158	Unknown	Unknown
A_32_P874394	Unknown	Unknown
A_32_P30874	Unknown	Unknown

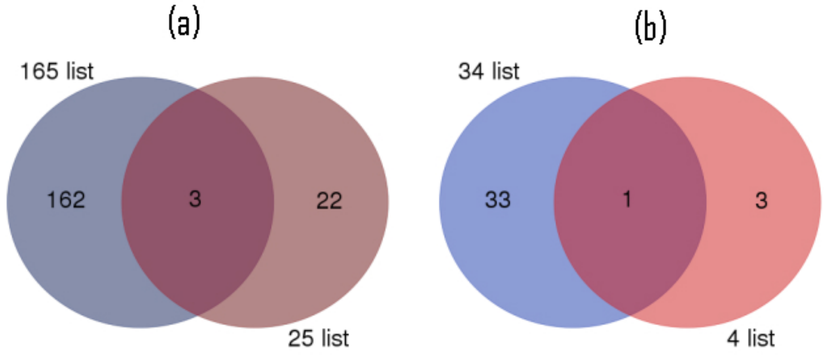

Figure 2: Venn diagrams of genes from different genelists. (**a**). Intersection of the copula genelist and the 165 genes in Omolo *et al*. (2013). (**b**). Intersection of Cox genes, 4 from the copula genelist and 34 from the 165 genes in Omolo *et al*. (2013).

however, has not been previously reported in relation to melanoma development. It lies in a region of the X-chromosome which has been associated with mental retardation [27].

4. CONCLUSION

In this study, we have proposed a copula-based algorithm for finding differentially expressed genes

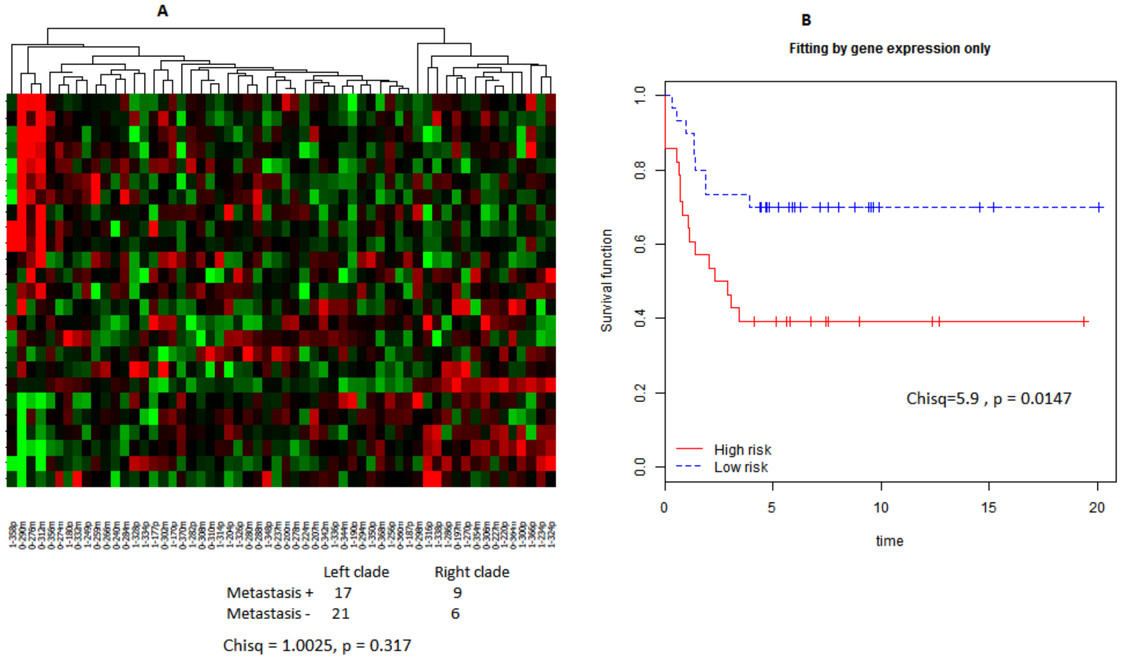

Figure 3: Kaplan-Meier survival plot and heatmap for the copula gene signature. A. Unsupervised hierarchical cluster of 58 samples using the 25 copula genes that were correlated with G2 checkpoint function. The classification of the samples yielded non-significant results (χ^2 = 1.0025, *p* = 0.317). **B**. The separation of the two groups in the Kaplan-Meier survival plot was significant (χ^2 = 5.9, *p* = 0.0147).

when the outcome of interest is continuous. The bivariate normal copula was employed in the analysis. We have shown the potential of the proposed copula-based approach in finding genes that are correlated with quantitative outcome in melanoma studies. The main focus was on the assessment of the power of the copula method in selecting genes that are correlated with quantitative outcome while controlling for FDR. Simulations indicated that the copula-based model had reasonable power at various levels of the FDR. Our approach is flexible as no parametric assumption is made on the marginal distributions, except that they are continuous. Relaxing parametric assumptions on microarray data may yield procedures that have good power for selecting differentially expressed genes. Although the copula model was applied to microarray data generated from the Agilent platform (dual-channel), it can be adopted for data from single-channel platforms (e.g. Affymetrix) as well.

A possible limitation of our study was the assumption of the normal copula for the analysis. This implied that the dependence structure between the G2 checkpoint function and the expression level for each gene in the analysis were all identical, which could result in loss of power if the assumption is not true. An alternative approach would be to model the pairs using different copulas. Thus, a test for choosing an optimal copula would have to be performed for each pair. Goodness-of-fit test methods maybe useful in this regard [28, 29].

New technologies such as RNA-sequencing (RNA-seq) are quickly replacing microarray technology. Current methods being developed for differential gene expression analysis are focusing on RNA-seq data. However, RNA-seq is more costly than microarrays [2]. Microarrays can still provide reliable and sensitive results and are quick and easy to work with. Therefore, new methods for analysing data from microarrays are still needed.

ACKNOWLEDGEMENTS

We are indebted to Dr. William K. Kaufmann for the continuous and survival outcome data used in this study. This work was partially supported by a grant from the Simons Foundation (# 282714 to BO). LC was supported by a scholarship from the African Union.

SUPPLEMENTAL MATERIALS

The supplemental materials can be downloaded from the journal website along with the article.

REFERENCES

[1] Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin 2017; 67: 7-30. https://doi.org/10.3322/caac.21387

[2] Winnepenninckx V, Lazar V, Michiels S, Dessen P, Stas M, Alonso SR, *et al.* Gene Expression Profiling of Primary Cutaneous Melanoma and Clinical Outcome. J Natl Cancer Inst 2006; 98: 472-482. https://doi.org/10.1093/jnci/djj103

[3] Mandruzzato S, Callegaro A, Turcatel G, Francescato S, Montesco MC, Chiarion-Sileni V, *et al.* A gene expression signature associated with survival in metastatic melanoma. J Transl Med 2006; 4: 50. https://doi.org/10.1186/1479-5876-4-50

[4] John T, Black MA, Toro TT, Leader D, Gedye CA, Davis ID, *et al.* Predicting Clinical Outcome through Molecular Profiling in Stage III Melanoma. Clin Cancer Res 2008; 14: 5173-5180. https://doi.org/10.1158/1078-0432.CCR-07-4170

[5] Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu YL, Adams S, *et al.* Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. Proc Natl Acad Sci USA 2009; 106: 20429-20434. https://doi.org/10.1073/pnas.0905139106

[6] Jonsson G, Busch C, Knappskog S, Geisler J, Miletic H, Ringnr M, *et al.* Gene Expression Profiling Based Identification of Molecular Subtypes in Stage IV Melanomas with Different Clinical Outcome. Clin Cancer Res 2010; 16: 3356-3367. https://doi.org/10.1158/1078-0432.CCR-09-2509

[7] Carson C, Omolo B, Chu H, Zhou Y, Sambade MJ, Peters EC, *et al.* A prognostic signature of defective p53-dependent G1 checkpoint function in melanoma cell lines: A signature of defective p53 function in melanoma. Pigment Cell Melanoma Res 2012; 25: 514-526. https://doi.org/10.1111/j.1755-148X.2012.01010.x

[8] Omolo B, Carson C, Chu H, Zhou Y, Simpson DA, Hesse JE, *et al.* A prognostic signature of G2 checkpoint function in melanoma cell lines. Cell Cycle 2013; 12: 1071-1082. https://doi.org/10.4161/cc.24067

[9] Kaufmann WK, Carson CC, Omolo B, Filgo AJ, Sambade MJ, Simpson DA, *et al.* Mechanisms of chromosomal instability in melanoma: Chromosomal Instability in Melanoma. Environ Mol Mutagen 2014; 55: 457-471. https://doi.org/10.1002/em.21859

[10] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001; 98: 5116-5121. https://doi.org/10.1073/pnas.091062498

[11] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002; 18: 1454-1461. https://doi.org/10.1093/bioinformatics/18.11.1454

[12] Chaba L, Odhiambo J, Omolo B. Evaluation of Methods for Gene Selection in Melanoma Cell Lines. Int J Stats Med Res 2017; 6: 1-9. https://doi.org/10.6000/1929-6029.2017.06.01.1

[13] Bandyopadhyay S, Mallik S, Mukhopadhyay A. A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. IEEE/ACM Trans Comput Biol Bioinformatics 2014; 11: 95-115. https://doi.org/10.1109/TCBB.2013.147

[14] Bair E. Identification of significant features in DNA microarray data: Feature selection in DNA microarray data. Wiley Interdiscip Rev Comput Stat 2013; 5: 309-325. https://doi.org/10.1002/wics.1260

[15] Genest C, Ghoudi K, Rvest LP. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 1995; 82(3): 543-552. https://doi.org/10.1093/biomet/82.3.543

[16] Owzar K, Jung SH, Sen PK. A Copula Approach for Detecting Prognostic Genes Associated With Survival Outcome in Microarray Studies. Biometrics 2007; 63: 1089-1098. https://doi.org/10.1111/j.1541-0420.2007.00802.x

[17] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol 1995; 57: 289-300. Available from: http://www.jstor.org/stable/2346101.

[18] Sklar. Fonctions de r'epartition 'a n dimensions et leures marges. Publications de l'Institut de Statistique de L'Universit'e de Paris 1959; 8: 229-231.

[19] Joe H. Asymptotic efficiency of the two-stage estimation method for copula-based models. J Multivar Anal 2005; 94: 401-419. https://doi.org/10.1016/j.jmva.2004.06.003

[20] Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons 1993; vol. 279.

[21] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100: 9440-9445. https://doi.org/10.1073/pnas.1530509100

[22] Golub GH, Van Loan CF. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press; 1996. Available from: https://books.google.co.ke/books?id=mlOa7wPX6OYC.

[23] Simon R, Lam A, Li MC, Ngan M, Menenzes S, Zhao Y. Analysis of gene expression data using BRB-Array Tools. Cancer Inform 2007; 3: 11.

[24] Bair E, Tibshirani R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLoS Biol 2004; 2. https://doi.org/10.1371/journal.pbio.0020108

[25] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 2009; 4: 44-57. https://doi.org/10.1038/nprot.2008.211

[26] Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996; 58: 267-288.

[27] Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, *et al.* A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet 2009; 41: 535-543. https://doi.org/10.1038/ng.367

[28] Genest C, Quessy JF, Remillard B. Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation. Scand J Statist 2006; 33: 337-366. https://doi.org/10.1111/j.1467-9469.2006.00470.x

[29] Berg D. Copula goodness-of-fit testing: an overview and power comparison. Euro J Financ 2009; 15: 675-701. https://doi.org/10.1080/13518470802697428

Received on 31-01-2017 Accepted on 06-03-2017 Published on 04-08-2017

https://doi.org/10.6000/1929-6029.2017.06.03.3