TEGS-CN: A Statistical Method for Pathway Analysis of Genome-wide Copy Number Profile

Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (A)

TEGS-CN: A Statistical Method for Pathway Analysis of Genome-wide Copy Number Profile

Yen-Tsung Huang[1], Thomas Hsu[2] and David C. Christiani[3,4]

[1]Department of Epidemiology, Brown University, Providence, RI. [2]Program in Biology, Brown University, Providence, RI. [3]Departments of Environmental Health and Epidemiology, Harvard School of Public Health, Boston, MA. [4]Massachusetts General Hospital/Harvard Medical School, Boston, MA.

ABSTRACT: The effects of copy number alterations make up a significant part of the tumor genome profile, but pathway analyses of these alterations are still not well established. We proposed a novel method to analyze multiple copy numbers of genes within a pathway, termed Test for the Effect of a Gene Set with Copy Number data (TEGS-CN). TEGS-CN was adapted from TEGS, a method that we previously developed for gene expression data using a variance component score test. With additional development, we extend the method to analyze DNA copy number data, accounting for different sizes and thus various numbers of copy number probes in genes. The test statistic follows a mixture of X^2 distributions that can be obtained using permutation with scaled X^2 approximation. We conducted simulation studies to evaluate the size and the power of TEGS-CN and to compare its performance with TEGS. We analyzed a genome-wide copy number data from 264 patients of non-small-cell lung cancer. With the Molecular Signatures Database (MSigDB) pathway database, the genome-wide copy number data can be classified into 1814 biological pathways or gene sets. We investigated associations of the copy number profile of the 1814 gene sets with pack-years of cigarette smoking. Our analysis revealed five pathways with significant *P* values after Bonferroni adjustment ($<2.8 \times 10^{-5}$), including the PTEN pathway (7.8×10^{-7}), the gene set up-regulated under heat shock (3.6×10^{-6}), the gene sets involved in the immune profile for rejection of kidney transplantation (9.2×10^{-6}) and for transcriptional control of leukocytes (2.2×10^{-5}), and the ganglioside biosynthesis pathway (2.7×10^{-5}). In conclusion, we present a new method for pathway analyses of copy number data, and causal mechanisms of the five pathways require further study.

KEYWORDS: copy numbers, pathway analyses, gene set analyses, variance component test, cancer genomics

SUPPLEMENT: Array Platform Modeling and Analysis (A)

CITATION: Huang et al. TEGS-CN: A Statistical Method for Pathway Analysis of Genome-wide Copy Number Profile. *Cancer Informatics* 2014:13(S4) 15–23 doi: 10.4137/CIN.S13978.

RECEIVED: April 13, 2014. **RESUBMITTED:** June 5, 2014. **ACCEPTED FOR PUBLICATION:** June 6, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: YTH received funding from Salomon Research Fund and Brown University. DCC received NIH grants CA092824, CA074386, and CA090578. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Yen-Tsung_Huang@brown.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

In the United States, lung cancers, 85% of which are non-small-cell lung carcinomas, are the second most common type of cancer and the leading cause of cancer-related death.[1] A well-documented and heavily studied risk factor for lung cancers is cigarette smoking.[2] Tobacco usage has been shown to mirror mortality rates caused by lung cancer.[3] Although there has been observation of carcinogens from cigarette smoking causing damage to the lungs via direct DNA interference, the exact mechanism through which smoking causes genomic damage is not completely understood.[4]

Copy number alteration is one of the leading causes of the variation in genomic DNA between humans.[5] Similar to single-nucleotide polymorphisms (SNPs), the alterations result in repeated or deleted sequences, although, unlike SNPs these repetitions or deletions can code for entire genes instead of individual bases. There is existing evidence to suggest that copy number alterations have a significant effect on the body's ability to regulate and combat tumor gene expression.[6] We have reported the effects of smoking on copy number alterations,[7] but have focused on probing copy numbers by the genomic location rather than by its biological function as a gene or a pathway. This study aims to incorporate more biological relevance to our study of copy number alterations by using pathway analysis. Pathway analysis has largely been used as a means of testing for expression data, but the use of the technique to study copy number variations is novel and is the main motivation behind this study.

The objective of this study was to use variance component tests on a gene set to analyze the methods through which smoking may cause tumorigenesis. Pack-years was chosen as the means of quantifying smoking habits, as it was a reliable measure of the total dose in the patient that accounted for both frequency and total time. The information is then analyzed using a proprietary statistical model called the Test for the Effect of a Gene Set with Copy Number data (TEGS-CN), which was adapted from our previous method, TEGS, a testing procedure for a multivariate linear regression model using permutation and scaled X^2 approximations.[8]

Methods

Lung cancer dataset. As reported in our previous study,[7] 264 snap-frozen tumor samples from non-small-cell lung cancer patients were collected from Massachusetts General Hospital, Boston, MA, and the National Institute of Occupational Health, Oslo, Norway. Detailed information regarding cigarette smoking and other demographic information was collected by trained research assistants following a modified version of the American Thoracic Society's standard respiratory questionnaire. Written consent was obtained from all patients and the study was approved by the institutional review boards from Massachusetts General Hospital, the Harvard School of Public Health, the Norwegian Data Inspectorate, and the Local Regional Committee for Medical Ethics. From these 264 subjects, genome-wide DNA copy numbers were measured using Affymetrix 250 K Nsp SNP array. From the array, copy numbers from a total of 262,264 probes were recorded. The raw copy numbers were preprocessed and normalized using dCHIP algorithm[9] with a reference panel consisting of blood DNA or DNA from adjacent normal tissues collected from lung cancer patients. The preprocessed copy numbers were then standardized for the analysis of TEGS-CN.

Molecular Signature Pathway/Gene Set Database. Pathways were compared with the existing data curated by the Broad Institute Gene Set Enrichment Analysis Molecular Signatures Database (MSigDB) for matching and gene identification purposes.[10] We required a cutoff of at least 15 valid probes per gene pathway from our dataset in order to consider a given pathway in our analysis. This cutoff served as a means of avoiding excessively short pathways that would potentially bias our results. In total, of the 1892 pathways in our dataset, a total of 1814 underwent analysis and 78 failed to reach the cutoff threshold.

Test for the effect of a gene set with copy number data. *Model.* Suppose that there are n subjects ($n = 264$ in the lung cancer dataset) and subject i has P DNA copy numbers of J genes, $\boldsymbol{Y}_{\mathrm{i}1}$, $\boldsymbol{Y}_{\mathrm{i}2}$, …, $\boldsymbol{Y}_{\mathrm{iJ}}$, where each gene j has p_j copy number probes, $\boldsymbol{Y}_{ij}^T = \left(Y_{ij1}, \ldots, Y_{ijpj}\right)$ and $P = \sum_{j=1}^{J} p_j$. In our model for pathway analysis of copy numbers, the outcomes indicate the P copy number values of the J genes in a gene set, and x_{i} is an independent variable, the pack-years of cigarette smoking for subject i. We consider the multivariate linear model:

$$Y_{ijk} = \alpha_{jk} + x_i\beta_{jk} + \varepsilon_{ijk}, \tag{1}$$

where i=1,…,n, j=1,…,J; and k=1,…,p_j; the errors $\varepsilon_{ij}^T = \left(\varepsilon_{ij1}, \ldots, \varepsilon_{ijp_j}\right)$ are assumed to be independent across different subjects and follow an arbitrary distribution with mean 0 and true covariance Σ_j, which is often unknown, and a_{jk} is the average copy number of probe k at gene j for those with x=0. As copy numbers are the read from copy number probes, we may use copy numbers and copy number probes interchangeably. Covariates can be incorporated in the model (1) by expanding α_{jk} to be $\sum_{l=1}^{L} \alpha_{jkl} z_{il}$, where L is the number of covariates plus one (ie, the intercept), z_{il} is the covariate l of subject i, z_{i1} is 1, and α_{jkl} is the regression coefficient of the covariate l for the copy number k of gene j. Model (1) can be written in matrix notation by stacking data of n subjects and p_j copy numbers from the gene j as:

$$Y_j = \boldsymbol{J}_j \boldsymbol{\alpha}_j + X_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where $\boldsymbol{Y}_j = \left(\boldsymbol{Y}_{1j}^T, \ldots, \boldsymbol{Y}_{nj}^T\right)^T$ is an np_j × 1 vector, $\boldsymbol{\epsilon}_j^T = \left(\boldsymbol{\epsilon}_{1j}^T, \ldots, \boldsymbol{\epsilon}_{nj}^T\right)$, $\boldsymbol{J}_j^T = \left(\boldsymbol{I}_{pj}, \ldots, \boldsymbol{I}_{pj}\right)$, $\boldsymbol{X}_j^T = \left(x_1\boldsymbol{I}_{pj}, \ldots, x_n\boldsymbol{I}_{pj}\right)$, $\boldsymbol{\alpha}_j^T = \left(\alpha_{j1}, \ldots, \alpha_{jp_j}\right)$, $\boldsymbol{\beta}_j^T = \left(\beta_{j1}, \ldots, \beta_{jp_j}\right)$.

Multi-locus copy number analysis: a variance component score test. *Testing procedure for p_j copy numbers in gene j* We have developed an algorithm termed TEGS for pathway analyses of multiple gene expressions from mRNA expression array data.[8] In the following, we show how to adapt the TEGS to analyze copy number data. For testing the effect of cigarette smoking on copy numbers of a gene, the null hypothesis of interest is that cigarette smoking x has no association with the p_j copy numbers of gene j in a pathway, or equivalently,

$$H_o : \boldsymbol{\beta}_j = 0.$$

It has been shown that the traditional multivariate tests such as likelihood ratio tests have limited power because the number of copy numbers (ie, p_j) in a pathway is large and copy numbers nearby share high correlation. To overcome this problem, we resort to an empirical Bayes approach by assuming the regression coefficients for the gene j, $\boldsymbol{\beta}_j$ follow an arbitrary common distribution with mean 0 and variance τ_j. The resulting model (2) hence becomes a linear mixed model.[11] The null hypothesis $H_o : \boldsymbol{\beta}_j = 0$ is thus equivalent to the null hypothesis for the variance component,

$$H_o : \tau_j = 0. \tag{3}$$

To test for the null (3), one can develop a score test for the variance components.[12] Specifically, it can be shown that the score for variance component τj has the expression:

$$\left(Y_j - J_j\boldsymbol{\alpha}_j\right)\sum_{nj}^{-1} X_j X_j^T \sum_{nj}^{-1}\left(Y_j - J_j\boldsymbol{\alpha}_j\right) - tr\left(\sum_{nj}^{-1} X_j X_j^T\right),$$

where $\boldsymbol{\Sigma}_{nj} = diag\left(\boldsymbol{\Sigma}_{j,\ldots,}\boldsymbol{\Sigma}_j\right)$ is an $np_j \times np_j$ block diagonal matrix. Since the second term is a constant, we use the first term of the score for τj to construct the test statistic, which is a nice quadratic form of the copy numbers $\boldsymbol{Y}_j$ and involves the true covariance $\boldsymbol{\Sigma}_j$:

$$Q_{Tj} = \left(Y_j - J_j\boldsymbol{\alpha}_j\right)^T \boldsymbol{\Sigma}_{nj}^{-1} X_j X_j^T \boldsymbol{\Sigma}_{nj}^{-1}\left(Y_j - J_j\boldsymbol{\alpha}_j\right).$$

As the number of copy numbers in a gene, p_j can be large, the true covariance matrix for $\boldsymbol{\epsilon}_{ij}$, $\boldsymbol{\Sigma}_j$ may not be easily estimated. We have shown in Huang and Lin[8] that one can replace the true covariance matrix by a working covariance $\boldsymbol{V}_{nj}$ with the resulting test statistic for gene j:

$$Q_j = \left(Y_j - J_j\boldsymbol{\alpha}_j\right)^T V_{nj}^{-1}\boldsymbol{X}_j \boldsymbol{X}_j^T V_{nj}^{-1}\left(Y_j - J_j\boldsymbol{\alpha}_j\right), \tag{4}$$

where $\boldsymbol{\alpha}_j$ can be estimated from the null model: $\boldsymbol{Y}_j = \boldsymbol{J}_j\boldsymbol{\alpha}_j + \boldsymbol{\epsilon}_j$. The null distribution of Q_j has been shown to follow a mixture of χ^2 distributions, which can be approximated with the inversion of characteristic function.[13] Because the number of subjects (n = 264) may not be large enough relative to the number of copy numbers within a gene P_j, we utilize a permutation procedure with scaled χ^2 approximation[14] to calculate the P value. We have shown through numerical simulations and real data analyses that the procedure is robust to the different choices of working covariances, protects type I error rate and outperforms other methods such as gene set enrichment analysis (GSEA).[8,10]

For implementation, we regressed DNA copy numbers and pack-years of cigarette smoking on the covariates, including age and gender, and the residuals of the regression models then became the input of our test statistic as the adjusted DNA copy numbers $\boldsymbol{Y}_j$ and smoking pack-years $\boldsymbol{X}_j$. This partial regression technique can avoid repeated fitting null model in each permutation and save the computation cost.

Testing procedure for *P* copy numbers from all *J* genes in a pathway. We have shown above that the TEGS can be adapted to perform joint analyses of multiple copy numbers in a gene j. But it requires additional development prior to analyzing all P copy numbers from J genes in a pathway. Model (1) can be written in matrix notation by stacking data of n subjects and total P copy numbers from J genes as:

$$\boldsymbol{Y} = \boldsymbol{J}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}^T = \left(\boldsymbol{Y}_1^T,\ldots,\boldsymbol{Y}_n^T\right.$ is an $nP \times 1$ vector, $\boldsymbol{Y}_i^T = \left(\boldsymbol{Y}_{i1}^T,\ldots,\boldsymbol{Y}_{iJ}^T\right)$, $\boldsymbol{\epsilon}^T = \left(\boldsymbol{\epsilon}_1^T,\ldots,\boldsymbol{\epsilon}_n^T\right)$, $\boldsymbol{\epsilon}_i^T = \left(\boldsymbol{\epsilon}_{i1}^T,\ldots,\boldsymbol{\epsilon}_{iJ}^T\right.$, $\boldsymbol{J}^T = \left(\boldsymbol{I}_P,\ldots,\boldsymbol{I}_P\right)$, $\boldsymbol{X}^T = \left(x_1\boldsymbol{I}_P,\ldots,x_n\boldsymbol{I}_P\right)$, $\boldsymbol{\alpha}^T = \left(\boldsymbol{\alpha}_1^T,\ldots,\boldsymbol{\alpha}_J^T\right)$, $\boldsymbol{\beta}^T = \left(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_J^T\right)$. The null hypothesis that there does not exist any association between the P copy numbers in the pathway and the smoking pack-years can be expressed as:

$$H_0 : \boldsymbol{\beta} = 0.$$

We may follow the same development for the gene j by assuming that all elements in $\boldsymbol{\beta}$ follow the same arbitrary distribution with mean 0 and variance τ:

$$H_0 : \tau = 0. \tag{5}$$

This is a valid test, but has the disadvantage that the gene with more copy numbers is more likely to dominate the test. Therefore, we have to adjust for the fact that genes have various number of copy number probes and make the signals from genes with different sizes comparable. To achieve this, we modify the above null hypothesis as:

$$H_0 : \tau_j = w_j\tau = 0, \text{for } j = 1,\ldots,J. \tag{6}$$

One can choose different weighting schemes to up-weight the gene with less copy numbers and down-weight the gene with more copy numbers. For example, w_j can be the inverse of the number of copy number probes within gene j; or with equal weighting, the null reduces to the hypothesis (5). We propose to weight by the variability of Q_j, with wj being the inverse of the variance for Q_j.

$$w_j^{-1} = \text{Var}\left(Q_j\right) = 2\text{tr}\left(\boldsymbol{V}_{nj}^{-1}\boldsymbol{X}_j\boldsymbol{X}_j^T\boldsymbol{V}_{nj}^{-1}\boldsymbol{\Sigma}_{nj}\boldsymbol{V}_{nj}^{-1}\boldsymbol{X}_j\boldsymbol{X}_j^T\boldsymbol{V}_{nj}^{-1}\boldsymbol{\Sigma}_{nj}\right), \tag{7}$$

As discussed above, the estimation for the covariance for P_j copy numbers may not be stable. We approximated $\boldsymbol{\Sigma}_{nj}$ by only taking the diagonal component of the sample covariance matrix or adding the fifth percentile of the variances to the diagonal elements to stabilize the covariance matrix. The resulting test statistic follows a similar expression as (4):

$$Q\text{pathway} = \left(\boldsymbol{Y}_w - \boldsymbol{J}\alpha\right)^T \boldsymbol{V}_n^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{V}_n^{-1}\left(\boldsymbol{Y}_w - \boldsymbol{J}\alpha\right), \tag{8}$$

where $\boldsymbol{Y}_w^T = \left(\boldsymbol{Y}_{w1}^T,\ldots,\boldsymbol{Y}_{wn}^T\right)$ is an $nP \times 1$ vector, $\boldsymbol{Y}_{wi}^T = \left(\widehat{w}_1^{-1/2}\boldsymbol{Y}_{i1}^T,\ldots,\widehat{w}_j^{-1/2}\boldsymbol{Y}_{iJ}^T\right)$. Note that $\boldsymbol{V}_n$ in (8) needs to be a correlation matrix rather than a covariance matrix since the variance have been accounted for in the weighting scheme. Again, we are able to approximate the distribution of Q_{pathway} through permutation and calculate the P value by comparing the distribution and the observed value of Q_{pathway}. We can also approximate the distribution with scaled χ^2 distribution by matching the first two moments.[8] We term this new procedure for copy number analyses as TEGS-CN.

Examples of working covariance $\boldsymbol{Vn}$ include (1) working independence, which assumes that the genes are independent in a gene set and (2) unstructured sample covariance. The unstructured sample covariance is estimated using the residuals obtained by performing separate simple linear regression of individual copy numbers on smoking pack-years in (1). When the total number of copy number probes in a gene set, P, is large and the sample size, n, is small, the standard empirical unstructured sample covariance estimator is unstable. We hence stabilize it using a ridge estimator by adding the fifth percentiles of sample variances to the diagonal of the empirical covariance estimator. Other working covariance structures have been compared and discussed in TEGS.[8]

Simulation Studies. To mimic the real copy number data, we based our simulation on the copy number data of the 264 lung cancer tumors. We simulated the data from a gene set, type 3 secretin system, which contained 22 genes and 58 copy number probes. We then randomly selected a proportion π of the 58 copy numbers to be causal copy numbers $\boldsymbol{Y}^*$ and simulated pack-years of cigarette smoking X^* according to a regression model:

$$X_i^* = \boldsymbol{Y}_i^{*T}\boldsymbol{\beta} + \in_\iota^*$$

where $i = 1,\ldots,264, \boldsymbol{Y}_i^{*T} = \left(Y_{1,i}^*,\ldots,Y_{J^*,i}^*\right), \boldsymbol{\beta}^T = \left(\beta_1,\ldots,\beta_{J^*}\right), J^*$ is a rounded integer of 58π, $\in_i^* \sim N\left(0,35.9\right)$, and for simplicity, $\beta_1 = \ldots = \beta_{J^*} = \beta$.

We conducted three sets of simulation studies to evaluate the performance of TEGS-CN. In the first set of simulation, we fixed the proportion of causal copy numbers, π to be 0.2 and varied the magnitude of the CN-smoking association β from 0 to 5 (Fig. 1A). For each parameter configuration, we simulated 1,000 datasets, and estimated the size and the power of TEGS-CN as the proportion of 1000 $P < 0.05$. We studied tests using both the working independence and the sample correlation, and P values were calculated using permutation and scaled χ^2 approximation. In the second set of simulation, we fixed the magnitude of non-zero β's to be 1 and varied the proportion of causal ones among the 58 copy numbers, π from 0 to 1 (Fig. 1B).

In the third set of simulation, we fixed the proportion of causal copy numbers, π, to be 0.2, similar to the first set of simulation. However, instead of randomly assigning the causal copy numbers, we assign them to smaller genes (Fig. 2). As a result, the proportion of causal genes was 0.36 (8 genes) among the 22 genes, although the proportion of causal copy numbers was 0.2 among the 58 probes. In this setting, we compared the size and the power of TEGS and TEGS-CN, both with working independence, to illustrate the importance of weighting in TEGS-CN.

Results

Simulation Studies. The size of TEGS-CN was well protected at $P < 0.05$ under both working independence and unstructured sample correlation: 0.047 (0.049 using scaled χ^2 approximation) for working independence and 0.043 (0.049) for sample correlation. The power increased with the magnitude of CN-smoking association (Fig. 1A) as well as the proportion of non-zero CN-smoking association (Fig. 1B). The statistical powers from permutation and scaled χ^2 approximation were very similar. TEGS-CN with working independence performed better when the signals were weak or sparse, whereas TEGS-CN with working sample correlation performed better when the signals were strong or dense. We noted that the difference between the two working correlation structures is that TEGS-CN with sample correlation is able to borrow information from the neighbor copy number probes. When the CN-smoking association is weak or sparse, the additional information from neighbor copy numbers may introduce unnecessary noises rather than significant signals. For the genome-wide scan, we chose to use working independence since the overall signals across the genome may not be strong. However, one can use sample correlation for candidate gene sets if strong or dense signals for certain gene sets are plausible assumptions.

To illustrate the importance of weighting in TEGS-CN and its novelty over the original TEGS, we compared their performance when the association signals are sparse but enriched in eight smaller genes out of the total 22 genes within a gene set. In Figure 2, the numerical simulation revealed that TEGS-CN were consistently more powerful than TEGS. Through weighting in (7), we were able to account for the number of copy number probes within a gene and to make the effects contributed from genes with different sizes comparable. In contrast, TEGS treated each copy number probes equally, which implicitly up-weighted the larger genes and down-weighted the smaller genes. Despite the different performance under the alternative, it is noteworthy that both TEGS-CN and TEGS had well-protected type I error under the null, 0.053 (0.053 using scaled χ^2 approximation) for TEGS-CN, and 0.058 (0.058 using approximation) for TEGS. As mentioned in the Methods section, both are valid tests without inflation of type I error under the null; however, TEGS is likely to be subject to bias due to various sizes of the genes within a gene set.

Analysis of cigarette smoking and copy numbers in Non-small cell lung cancer (NSCLC) tumor genome. As discussed in the Methods section, we applied our proposed algorithm to analyze the genome-wide copy number data of non-small-cell lung cancer, studying the association of copy number profile with cigarette smoking. The demographic and

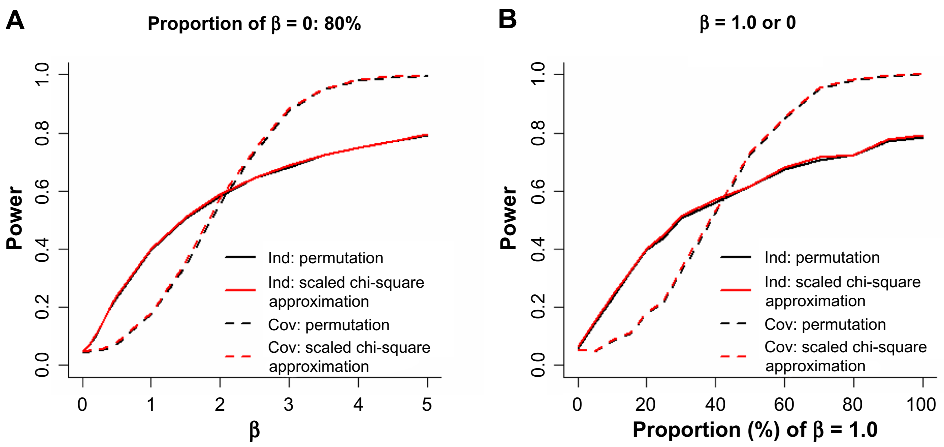

Figure 1. Power curves of TEGS-CN varying the magnitude of associations β (**A**) and the proportion of non-zero associations (**B**).

clinical characteristics of 264 patients are presented in Table 1. We divided the total participants into 132 heavy smokers and 132 light- or non-smokers by the median value (35.6) of the pack-years of cigarette smoking collected by a modified standard American Thoracic Society questionnaire.

Different pathways contain different numbers of genes that DNA copy numbers were measured from the array. The distribution of the number of genes in a pathway is shown in Figure 3A. The average number of genes per pathway was 39.7, while the median was 19 (interquartile range (IQR): 10–39.8). The smallest pathway was one gene long, while the largest (STEMCELL_NEURAL_UP) was 1201 genes long. Genes also contain different numbers of copy number probes: the largest number of probes per gene is 600 (*CSMD1*) and the smallest is 1, with average number being 8.8, median being 3, and IQR being 1–8. We also present the distribution of the number of copy number probes per pathway in Figure 3B.

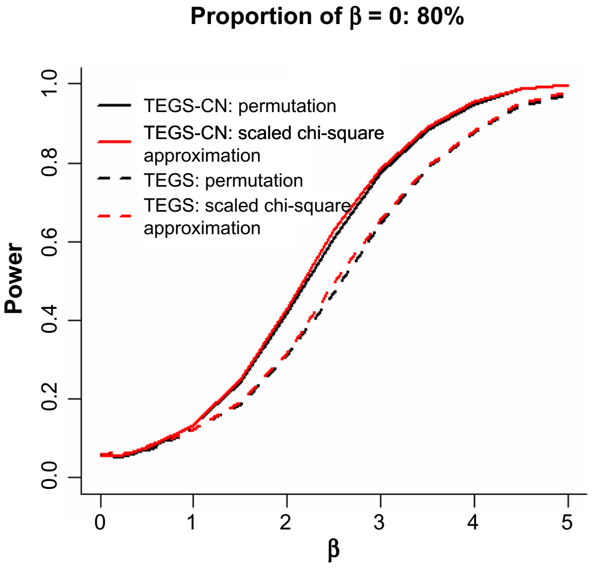

Figure 2. Power curves of TEGS-CN and TEGS, with CN-smoking associations occurring in small genes.

The average number of probes per pathway was 365.5, and the median (IQR) was 146 (60–344.8). The largest pathway (ALZHEIMERS_DISEASE_DN) includes 10,060 copy number probes, whereas the smallest includes one probe. The large number of genes or copy number probes and the modest sample size (n = 264) represented the nature of large *P* and small *n* for the analytic problem. Moreover, the various numbers of copy number probes per gene demonstrated the necessity of weighting the signals from different genes with various sizes, as the null hypothesis (6).

We utilized the TEGS-CN algorithm to investigate the association of copy numbers of non-small-cell lung cancer with the pack-years of cigarette smoking. After running the pathways through the TEGS-CN analysis tests, it was found that of the 1814 pathways, 53 pathways had *P* values less than 1×10^{-3}, 14 pathways had *P* values less than 1×10^{-4}, and 5 pathways were under the Bonferroni cutoff at $0.05/1814 = 2.8 \times 10^{-5}$. Because of the large number of copy number probes per pathway, we introduced a working independence structure to save computation burden. Table 2 shows the top five pathways, including the PTEN pathway (PTENPATHWAY: $P = 7.8 \times 10^{-7}$), the gene set up-regulated in heat shock experiment (HEATSHOCK_YOUNG_UP: $P = 3.6 \times 10^{-6}$), the gene set down-regulated in rejection of kidney transplantation (FLECHNER_PBL_KIDNEY_TRANSPLANT_REJECTED_VS_OK_DN: $P = 9.2 \times 10^{-6}$), the gene set involved in transcriptional regulation of leukocytes (SCHRAETS_MLL_TARGETS_UP: $P = 2.2 \times 10^{-5}$), and the ganglioside biosynthesis pathway (GANGLIOSIDE_BIOSYNTHESIS: $P = 2.7 \times 10^{-5}$).

For the top five pathways, we further performed gene analyses where we analyzed multiple copy numbers in a gene to study the association with cigarette smoking (Tables 3–7). Among the 10 genes in PTENPATHWAY, the association of *PTK2* copy numbers with cigarette smoking was highly significant; the association of *MAPK1*, *SOS1*, and *PIK3R1* were marginally significant ($P < 0.1$). For the gene sets for heat shock up-regulation (HEATSHOCK_YOUNG_UP), *DYNLL1* seemed to drive the

Table 1. Characteristics of the 264 non-small-cell lung cancer patients.

	TOTAL	HEAVY SMOKERS*	LIGHT- OR NON-SMOKER*
Sample Size	264	132	132
Male (%)	61.4	67.4	55.3
Age			
Mean +/- SD	67.4 +/- 8.3	67.7 +/- 8.0	67.1 +/- 8.5
Cigarette Smoking in Pack-Years			
Median +/- IQR	35.6 +/- 38.72	58.8 +/- 40.5	19.8 +/- 18.7
Clinical Stage			
Stage 1 (%)	73.1	73.5	72.7
Stage 2 (%)	17.0	15.9	18.2
Stage 3 or 4 (%)	9.08	10.1	7.6
Cigarette Smoking Status			
Never Smoked (%)	6.8	0	13.6
Ex-Smokers (%)	48.5	49.2	47.7
Current Smokers (%)	44.7	50.8	38.6
Adenocarcinoma (%)	66.3	64.4	68.2

Notes: *Heavy smokers are defined as pack-years of cigarette smoking >35.6 (the median of the smoking pack-years in 264 subjects), and light- or non-smokers are those with ≤35.6 smoking pack-years.

significance of the gene set. Of the 31 genes involved in kidney transplantation rejection, there were five significant genes with $P < 0.05$ and three marginally significant genes with P value between 0.05 and 0.1. Of the 21 genes involved in transcriptional regulation of leukocytes (SCHRAETS_MLL_TARGETS_UP), there were seven significant genes with $P < 0.05$. Finally, there were three highly significant genes out of the six involved in ganglioside biosynthesis.

Discussion

Copy number alterations have been reported to correspond to the majority of the variation in gene expression of tumor genome.[6] There is a pressing need to understand copy number alterations in relation to phenotypic trait such as cigarette smoking or clinical outcome in the context of biological pathway or gene set. Thus, in this study, we develop a new method for pathway analyses of copy number data, TEGS-CN. The method is adapted from our previous algorithm TEGS, a testing procedure for gene set or pathway analyses. Although TEGS has been developed for pathway analyses of gene expression data, it may not be readily applicable to copy number analysis due to the various numbers of copy number probes in genes. Without proper adjustment in TEGS, larger genes with more copy number

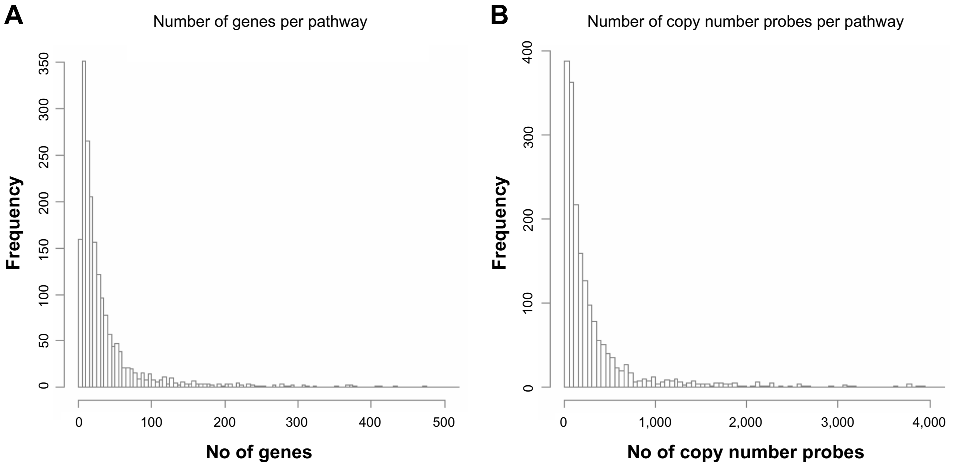

Figure 3. Distribution of numbers of genes (**A**) and copy number probes (**B**) in pathways. Note that the number of genes (**A**) and copy number probes (**B**) per pathway was truncated at 500 and 4000, respectively, due to skewness of the distribution, and the entire range was described in the text.

Table 2. The five pathways with *P* value <0.05 after Bonferroni adjustment.

GENE PATHWAY (MSigDB ID)	NUMBER OF GENES	NUMBER OF COPY NUMBER PROBES	NOMINAL P-VALUES
PTENPATHWAY	10	54	7.8×10^{-7}
HEATSHOCK_YOUNG_UP	5	21	3.6×10^{-6}
FLECHNER_PBL_KIDNEY_TRANSPLANT_REJECTED_VS_OK_DN	31	103	9.2×10^{-6}
SCHRAETS_MLL_TARGETS_UP	21	120	2.2×10^{-5}
GANGLIOSIDE_BIOSYNTHESIS	6	37	2.7×10^{-5}

probes may dominate the signal for the gene set and bias the results. Therefore, we extended the algorithm by introducing a general weighting scheme to make the information from genes with different sizes comparable. We showed in simulation that with proper weighting for each copy number probes to make each gene comparable, TEGS-CN outperformed TEGS in terms of the statistical power (Fig. 2). The general weighting is not only limited to adjusting for gene size but can also be used to incorporate other parameters, for example, prior information of a gene in relation to the phenotypic trait of interest. Moreover, the TEGS-CN can also be used to analyze other genomic data such as DNA methylation array.

Our method (TEGS-CN) analyzes the copy number as a continuous measure. If either amplifications or deletions of copy number have a significant relation to the phenotypic trait, our method is able to detect the association. If one would like to analyze the copy number by categorizing into discrete classes (amplified, neutral, and deleted copy numbers), additional development is required to incorporate such a kind of analyses. However, collapsing the continuous copy numbers into categorical classes loses information and therefore decreases the statistical power. On the other hand, TEGS-CN is an omnibus test, which can detect if there is any effect of a gene set (see null hypothesis (6)) but cannot distinguish where the effect is and the directionality of the effect. Once identifying candidate gene sets or genes, one can study the dose–response relationship for each copy number, to further investigate which kind of alterations is related to the phenotype.[7]

The top pathways have been shown to be biologically significant in cancer. The PTEN pathway, which gave us the highest *P* value among all the pathways, is a well-known cancer-related pathway. Typically, *PTEN* suppresses tumors by inducing apoptosis in abnormal cells, but in many cancers, including non-small-cell lung cancer, it becomes expressed in reduced copy numbers, which leads to excessive cell proliferation.[15] One of the genes expressing high *P* values is *PTK2*, which controls cellular adhesion and mobility. Down-regulation of this gene means less metastasis in cancer cells as they no longer move freely.[16]

The following three pathways that were discovered were a variety of different immunological pathways. Each relates in some way to different cellular functions of the human body that functions as anti-cancer or responses to toxic exposures and organ transplants. The HEATSHOCK_YOUNG_UP pathway is a pathway that becomes up-regulated in lymphocytes after heat shock in younger individuals when compared with the lymphocyte reaction in older individuals. It is considered to be an indicator of immune response strength and recovery, as its regulation reflects the gradual degradation of the recuperative ability with age.[17] Among the genes in the heat shock pathway, the one with the most significant *P* value was *DYNLL1*. This gene codes for a light chain in the protein dynein, and as such influences a wide variety of cellular processes.[18] The FLECHNER_PBL_KIDNEY_TRANSPLANT_REJECTED_VS_OK_DN pathway is

Table 3. *P* values of genes in PTENPATHWAY.

GENES	WORKING INDEPENDENCE	SAMPLE COVARIANCE
PTK2	6.52×10^{-7}	0.0592
SOS1	0.0391	0.0131
PIK3R1	0.0726	0.0606
MAPK1	0.0891	0.135
PIK3CA	0.106	0.106
ITGB1	0.239	0.255
CDKN1B	0.295	0.295
ILK	0.270	0.270
GRB2	0.457	1.65×10^{-4}
PTEN	0.539	0.539

Table 4. *P* values of genes in HEATSHOCK_YOUNG_UP.

GENES	WORKING INDEPENDENCE	SAMPLE COVARIANCE
DYNLL1	1.42×10^{-6}	5.92×10^{-7}
CHD1L	0.214	0.492
TRA2A	0.230	0.177
WDR89	0.360	0.360
ANXA1	0.767	0.404

Table 5. *P* values of genes in FLECHNER_PBL_KIDNEY_TRANSPLANT_REJECTED_VS_OK_DN.

GENES	WORKING INDEPENDENCE	SAMPLE COVARIANCE
ZNF148	9.86×10^{-5}	0.245
SYNPO	0.00064	0.00064
IK	0.00588	0.00588
ANAPC13	0.0363	0.0363
SETD1A	0.0453	0.0495
MYD88	0.0519	0.0519
SFRS3	0.0733	0.0897
GOLGA1	0.0987	0.113
CD46	0.104	0.0810
STK38	0.203	0.203
RIOK3	0.206	0.1623
GSPT1	0.208	0.1477
ITGB1	0.239	0.255
LMO1	0.287	0.288
CDKN1B	0.295	0.295
SFRS2B	0.335	0.335
MAPK14	0.367	0.0444
RABGGTB	0.372	0.261
GPR56	0.381	0.187
PTGER2	0.438	0.438
GTF2B	0.438	0.438
PCBP2	0.470	0.470
TMEM63A	0.488	0.488
TRAF3IP3	0.506	0.0411
ACO2	0.524	0.548
CHAF1A	0.554	0.554
BICD2	0.555	0.576
RALBP1	0.607	0.653
TES	0.634	0.101
CTCF	0.656	0.656
TMF1	0.700	0.429

a pathway relating to immunosuppression. It has been studied in patient reactions to kidney transplants, principally how the body mounts acute rejection or has no reaction.[19] Two genes had particularly strong results in this pathway. The gene *ZNF148* regulates transcription of a variety of genes, one of which is stromelysin, a protein thought to be active in tumor initiation.[20] *SYNPO* is a gene that controls synaptodin, which relates to actin and cell motility.[21] SCHRAETS_MLL_TARGETS_UP is a pathway that controls transcriptional regulation of human leukocytes. Although the normal function of the pathway is not well understood, it is known that translocations in this pathway will result in acute leukemia.[22] *ENPP2* is the autotaxin enzyme, which increases tumor cell motility via the creation of lysophosphatidic acid (LPA).[23]

Table 6. *P* values of genes in SCHRAETS_MLL_TARGETS_UP.

GENES	WORKING INDEPENDENCE	SAMPLE COVARIANCE
MGLL	5.40×10^{-5}	0.00177
RSPO2	7.98×10^{-5}	0.0395
HSPB8	0.000145	9.61×10^{-5}
ENPP2	0.00199	0.000263
TGFBI	0.00952	0.0954
THBS2	0.0243	0.00311
IL1RN	0.0272	0.0265
CTSH	0.0529	0.0559
CAP1	0.0869	0.106
ARHGDIB	0.161	0.171
COL6A3	0.215	0.00658
MVK	0.281	0.233
CDKN1B	0.295	0.295
GATA6	0.466	0.466
GSTA4	0.476	0.425
ACADM	0.491	0.0850
FOXC2	0.492	0.492
DFFB	0.625	0.625
LIMK1	0.686	0.686
CD53	0.775	0.405
PITX2	0.994	0.994

GANGLIOSIDE_BIOSYNTHESIS is a pathway that controls the formation of gangliosides in the human body. Gangliosides serve as differentiating surface markers in cells, controlling cell growth, development, and apoptosis. It is also known that certain cancers result in the expression of modified gangliosides or an overall reduced expression. Two gangliosides created in this pathway that appeared significant in our data were *ST8SIA1* and *ST3GAL5*. *ST8SIA1* codes for an enzyme that produces the GD3 ganglioside, a ganglioside known to be involved in cell adhesion and malignant growth.[24] *ST3GAL5* codes for an enzyme that produces the GM3 ganglioside, which participates in differentiation, morphology, proliferation, and adhesion.[25]

Table 7. *P* values of genes in GANGLIOSIDE_BIOSYNTHESIS.

GENES	WORKING INDEPENDENCE	SAMPLE COVARIANCE
ST3GAL1	0.000441	0.118
ST8SIA1	0.000489	0.176
ST3GAL5	0.00527	0.00578
ST3GAL4	0.0721	0.0725
ST6GALNAC2	0.402	0.316
ST3GAL2	0.740	0.740

Author Contributions

Conceived and designed the experiments: YTH, DCC. Analyzed the data: YTH, TH. Wrote the first draft of the manuscript: YTH, TH. Contributed to the writing of the manuscript: YTH, TH, DCC. Agree with manuscript results and conclusions: YTH, TH, DCC. Jointly developed the structure and arguments for the paper: YTH, TH. Made critical revisions and approved final version: YTH, TH, DCC. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin*. 2009;59(4):225–49.
2. Bach PB. Smoking as a factor in causing lung cancer. *JAMA*. 2009;301(5): 539–41.
3. Dubey S, Powell CA. Update in lung cancer 2008. *Am J Respir Crit Care Med*. 2009;179(10):860–8.
4. Alavanja MC. Biologic damage resulting from exposure to tobacco smoke and from radon: implication for preventive interventions. *Oncogene*. 2002;21(48):7365–75.
5. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437–55.
6. Li Q, Seo JH, Stranger B, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*. 2013;152(3):633–41.
7. Huang YT, Lin X, Liu Y, et al. Cigarette smoking increases copy number alterations in nonsmall-cell lung cancer. *Proc Natl Acad Sci USA*. 2011;108(39):16345–50.
8. Huang YT, Lin X. Gene set analysis using variance component tests. *BMC Bioinformatics*. 2013;14:210.
9. Zhao X, Li C, Paez JG, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*. 2004;64(9):3060–71.
10. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545–50.
11. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88:9–25.
12. Lin X. Variance component testing in generalised linear models with random effects. *Biometrika*. 1997;73:309–26.
13. Davies R. Algorithm AS 155: the distribution of a linear combination of chi-square random variables. *Appl Stat*. 1980;29:323–33.
14. Satterthwaite F. An approximation distribution of estimates of variance components. *Biometrics*. 1946;2:110–4.
15. Chalhoub N, Baker SJ. PTEN and the PI3-kinase pathway in cancer. *Annu Rev Pathol*. 2009;4:127–50.
16. Andre E, Becker-Andre M. Expression of an N-terminally truncated form of human focal adhesion kinase in brain. *Biochem Biophys Res Commun*. 1993;190(1):140–7.
17. Visala Rao D, Boyle GM, Parsons PG, Watson K, Jones GL. Influence of ageing, heat shock treatment and in vivo total antioxidant status on gene-expression profile and protein synthesis in human peripheral lymphocytes. *Mech Ageing Dev*. 2003;124(1):55–69.
18. Pfister KK, Fisher EM, Gibbons IR, et al. Cytoplasmic dynein nomenclature. *J Cell Biol*. 2005;171(3):411–3.
19. Flechner SM, Kurian SM, Head SR, et al. Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. *Am J Transplant*. 2004;4(9):1475–89.
20. Tommerup N, Vissing H. Isolation and fine mapping of 16 novel human zinc finger-encoding cDNAs identify putative candidate genes for developmental and malignant disorders. *Genomics*. 1995;27(2):259–64.
21. Mundel P, Heid HW, Mundel TM, Kruger M, Reiser J, Kriz W. Synaptopodin: an actin-associated protein in telencephalic dendrites and renal podocytes. *J Cell Biol*. 1997;139(1):193–204.
22. Schraets D, Lehmann T, Dingermann T, Marschalek R. MLL-mediated transcriptional gene regulation investigated by gene expression profiling. *Oncogene*. 2003;22(23):3655–68.
23. Houben AJ, Moolenaar WH. Autotaxin and LPA receptor signaling in cancer. *Cancer Metastasis Rev*. 2011;30(3–4):557–65.
24. Haraguchi M, Yamashiro S, Yamamoto A, et al. Isolation of GD3 synthase gene by expression cloning of GM3 alpha-2,8-sialyltransferase cDNA using anti-GD2 monoclonal antibody. *Proc Natl Acad Sci USA*. 1994;91(22):10455–9.
25. Ishii A, Ohta M, Watanabe Y, et al. Expression cloning and functional characterization of human cDNA for ganglioside GM3 synthase. *J Biol Chem*. 1998;273(48):31652–5.