PLOS | ONE

RESEARCH ARTICLE

Function of Cancer Associated Genes Revealed by Modern Univariate and Multivariate Association Tests

Malka Gorfine[2☯]*, Boaz Goldstein[1☯], Alla Fishman[1], Ruth Heller[3], Yair Heller[4], Ayelet T. Lamm[1]*

1 Faculty of Biology, Technion- Israel Institute of Technology, Technion City, Haifa 3200003, Israel, **2** Faculty of Industrial Engineering and Management, Technion- Israel Institute of Technology, Technion City, Haifa 3200003, Israel, **3** Department of Statistics and Operations Research, Tel Aviv University, Ramat Aviv, Tel Aviv 6997801, Israel, **4** Tel Aviv, Israel

☯ These authors contributed equally to this work.
* ayeletla@tx.technion.ac.il (ATL); gorfinm@ie.technion.ac.il (MG)

Citation: Gorfine M, Goldstein B, Fishman A, Heller R, Heller Y, Lamm AT (2015) Function of Cancer Associated Genes Revealed by Modern Univariate and Multivariate Association Tests. PLoS ONE 10(5): e0126544. doi:10.1371/journal.pone.0126544

Academic Editor: Lin Chen, The University of Chicago, UNITED STATES

Received: September 27, 2014

Accepted: April 3, 2015

Published: May 12, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was funded by National Institutes of Health (grant P01CA53996 to MG), The Israeli Centers of Research Excellence (I-CORE) program, (Center No. 1796/12 to ATL), The Israel Science Foundation (grant No. 644/13 to ATL). ATL is a Taub fellow - supported by the Taub Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Copy number variation (CNV) plays a role in pathogenesis of many human diseases, especially cancer. Several whole genome CNV association studies have been performed for the purpose of identifying cancer associated CNVs. Here we undertook a novel approach to whole genome CNV analysis, with the goal being identification of associations between CNV of different genes (CNV-CNV) across 60 human cancer cell lines. We hypothesize that these associations point to the roles of the associated genes in cancer, and can be indicators of their position in gene networks of cancer-driving processes. Recent studies show that gene associations are often non-linear and non-monotone. In order to obtain a more complete picture of all CNV associations, we performed omnibus univariate analysis by utilizing dCov, MIC, and HHG association tests, which are capable of detecting any type of association, including non-monotone relationships. For comparison we used Spearman and Pearson association tests, which detect only linear or monotone relationships. Application of dCov, MIC and HHG tests resulted in identification of twice as many associations compared to those found by Spearman and Pearson alone. Interestingly, most of the new associations were detected by the HHG test. Next, we utilized dCov's and HHG's ability to perform multivariate analysis. We tested for association between genes of unknown function and known cancer-related pathways. Our results indicate that multivariate analysis is much more effective than univariate analysis for the purpose of ascribing biological roles to genes of unknown function. We conclude that a combination of multivariate and univariate omnibus association tests can reveal significant information about gene networks of disease-driving processes. These methods can be applied to any large gene or pathway dataset, allowing more comprehensive analysis of biological processes.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Copy number variations (CNV) are a part of normal Human genetic variability. Tens of thousands of CNVs have been reported in the Database of Genomic Variants (DGV) based on healthy control samples [1,2]. However, CNVs are also a significant component of variation in disease risk and occurrence of many diseases and disorders, including cancer, HIV infection, autism, and psychiatric diseases [3–5]. In cancer, CNV is one of the most important somatic aberrations found [6]. Nowadays CNV analysis has become a central part of cancer research and many studies concentrate on detecting CNVs in the human genome in normal and diseased tissues and cells. ([7,8], DGV (http://projects.tcag.ca/variation)). In clinics a growing number of CNV are used for diagnostics and personalized therapy.

While individual CNVs can be detected by fluorescent in situ hybridization (FISH), whole genome CNV detection requires microarray-based comparative genomic hybridization (array CGH) or next generation sequencing (NGS) platforms [6]. These platforms generate very high volumes of data, making the analysis very challenging. One major task of CNV data analysis is identifying and characterizing associations between CNVs and diseases, which may potentially be driven by biologically relevant mechanisms [9–11].

Several association studies have been performed for the purpose of linking CNVs to diseases [7,8,12]. For example, Stamoulis et al. [11] focused on monotone relationships between CNV within and across chromosomes; Bussey et al. [12] looked at Pearson's correlation between CNV and gene expression levels. While most studies associated CNV with gene expression profile, very few, if any, attempts have been made to associate between CNVs of different genes detected in diseased tissue, even though the identification of associations between genes is extremely important for understanding basic biological processes and modeling gene regulatory networks. In this work we undertook such an approach to analyze cancer related CNV data. The rationale was that since CNV formation is part of carcinogenesis, associations between CNVs of genes would be indicative of their roles in carcinogenesis. Additionally, identification of these associations might enable building a gene network of disease driving processes.

To date, the most commonly used association tests are based on Pearson's or Spearman's correlation coefficient. Pearson's test is sensitive to the linear component in a relationship between two variables, while Spearman's test detects monotone relationships, such as a sigmoid. Hence, both tests are not able to detect non-monotone relationships such as U-shaped, ellipse, sinusoid, etc. Recent studies show that gene associations are often non-linear and non-monotone [13–15]; therefore in order to obtain a complete unbiased picture of all gene associations one must apply other statistical methods.

Recently, several statistical tests for detecting any type of relationships, including non-monotone ones, were proposed. In particular, Szekely et al. [16,17] suggested a test, named dCov, based on distance covariance and distance correlation; Reshef et al. [18] presented a test based on a novel measure of dependence—the maximal information coefficient (MIC); and Heller et al. [19] proposed a test based on ranks of distances, named HHG. Extensive simulation studies comparing between HHG, dCov, MIC, Spearman and Pearson have been performed [13,19]. Their main conclusions were that HHG is typically more powerful than dCov and dCov is usually more powerful than MIC in non-monotone settings.

In addition to their being univariate analysis tools capable of identification of a broad range of association types, dCov and HHG are also applicable for multivariate analysis, i.e., testing for dependence between the variables X and Y, when X and Y are vectors rather than single variables. Thus these tests can be used for identifying associations between pathways or between a gene and a pathway, even when the sample size is much smaller than the dimension of either X or Y.

The second aim of this work was demonstrating the effectiveness of association tests which are also capable of detecting non-monotone relationships, such as dCov, MIC and HHG for analyzing whole genome association data. For this purpose we utilized these tests alongside the standard Spearman and Pearson test in the analysis of CNV data derived from 60 human cancer cell lines (NCI-60) [12]. We have found that the application of tests capable of detecting any type of relationships, such as dCov and HHG, for univariate analysis, results in identification of twice as many associations compared to those found by Spearman and Pearson alone. Most of the new associations were detected by the HHG test. Moreover, multivariate analysis by means of dCov and HHG was able to associate between genes of unknown function from our dataset and basic biological pathways, providing a clue to possible biological functions of these genes.

The methods presented here can be useful in many other settings which require detection of associations of genes and pathways, such as reconstruction of networks and pathways—an important task in systems biology [20]. This study demonstrates that by using these methods researchers can uncover more associations of various types, and thus have a broader picture at their disposal when attempting to study biological phenomena.

Results

Identification of Gene-by-Gene Associations

In order to find associations between cancer-related CNVs, we used CNV data obtained by an array CGH from 60 human cancer cell lines (the NCI-60; [12]). Within the CGH array we selected clones that have known gene symbols and, for consistency, no missing values in any cell line. The result contained 99 genes. In addition to the traditional association tests, Spearman and Pearson, we applied three tests, dCov, MIC and HHG, which are also capable of detecting non-monotone relationships. An association was considered significant if the FDR-adjusted p-value was less than 0.05 using the Benjamini-Hochberg procedure [21]. Out of 4851 pairwise comparisons, Pearson or Spearman detected 254 significant associations, dCov detected 256, MIC detected 157 and HHG detected 400 significant associations (see Fig 1, Table 1, S1 Fig, and S1 Table for detailed results). Comparison of the three tests capable of detecting any type of relationships, namely dCov, MIC and HHG, revealed that they share 139 common significant results. Furthermore, 44 associations were found significant only by dCov; 11 only by MIC and 183 only by HHG (S1 Fig, top-right). Comparing Pearson and Spearman with dCov and HHG revealed that 29 significant associations were discovered solely by Pearson or Spearman, only 10 solely by dCov while 184 were discovered solely by HHG (Fig 1).

Of the number of significant statistical associations found by dCov, MIC or HHG, but not by Pearson or Spearman, the number found by HHG was exceptionally large. Specifically, while the number of significant associations shared by Pearson or Spearman and HHG is 190, Pearson and Spearman missed 210 associations found by HHG, whereas HHG missed only 64 associations found by Pearson or Spearman. In the above analysis, we combined Pearson's and Spearman's results that had adjusted p-value less than 0.05 as if they were a single method, even though this gives then an advantage compared to other methods. Given this, it is all the more interesting that HHG found 57% more associations then Pearson and Spearman. We therefore conclude that analysis based on the traditional Pearson and Spearman association tests could miss a significant proportion of all possible associations between genes.

In order to demonstrate the biological relevance of the associations detected by HHG we took a closer look at the detected associated gene pairs. One example of an association found only by HHG is the association between the genes LYN and CTSB (Fig 2). LYN encodes a non-receptor tyrosine-protein kinase, a regulator of many signal transduction pathways, while

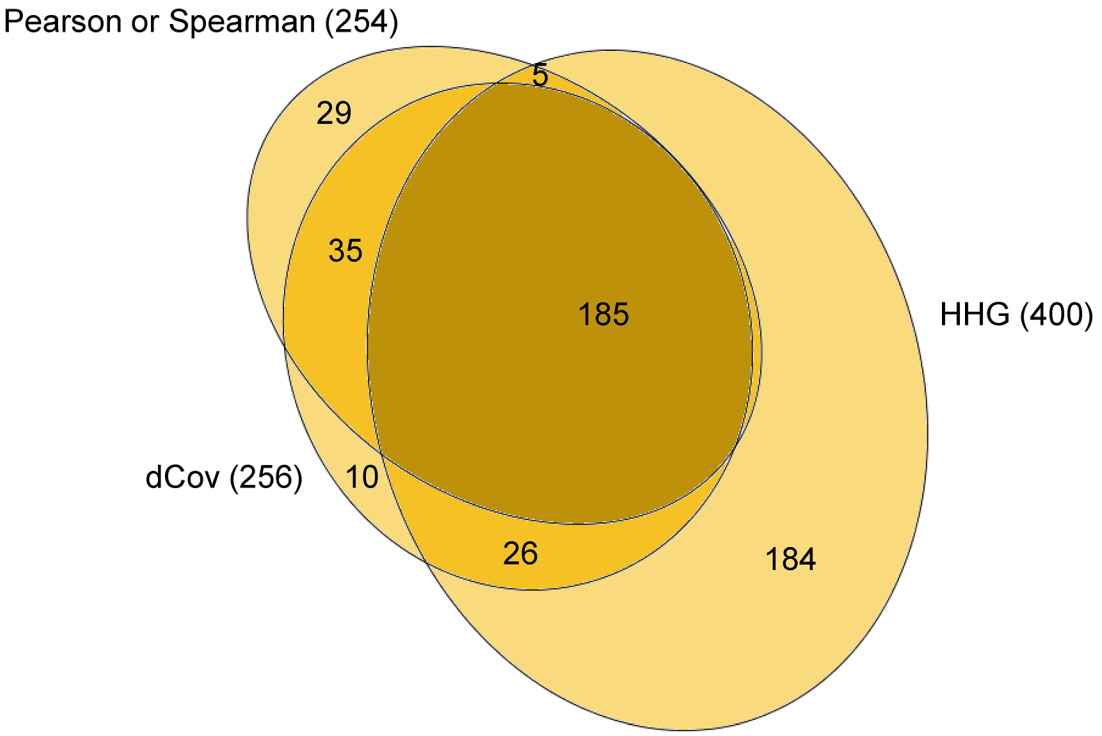

Fig 1. Euler diagram of the significant discoveries found by Pearson or Spearman, dCov and HHG. MIC was excluded due to the small number of significant findings provided by this method. The area of each oval represents the number of significant tests of each method, and intersections (emphasized by different colors) represent common discoveries. Evidently, Pearson or Spearman, dCov and HHG share 185 discoveries; 184 tests were significant by HHG but not by Pearson, Spearman or dCov; 10 tests were significant by dCov and not by Pearson, Spearman or HHG; 29 tests were significant by Pearson or Spearman but not by dCov or HHG; dCov and HHG share 26 discoveries; Pearson or Spearman and dCov share 35 discoveries; and Pearson or Spearman and HHG share only 5 discoveries.

doi:10.1371/journal.pone.0126544.g001

CTSB encodes cathepsin B, a thiol protease participating in intracellular degradation and turnover of proteins. No direct biological interactions between these two proteins are known, however they both interact directly with a third protein, Sphingosine kinase 1 (SPHK1). SPHK1 catalyzes the phosphorylation of sphingosine to form sphingosine-1-phosphate (S1P), a key sphingolipid signaling molecule involved in cell growth, survival, differentiation and motility. Interaction between LYN and SPHK1 is essential for the activation of SPHK1 [22]. On the other hand, interaction between Cathespin B and SPHK1 has been shown to down-regulate SPHK1 levels *in vivo* [23] and to cleave it *in vitro* [24]. This experimental data demonstrates that the association between LYN and CTSB identified by HHG is indeed biologically relevant. Moreover, the existence of the association between CNV of LYN and CTSB points to LYN-SPHK1 and CTSB- SPHK1 interactions as being important for carcinogenesis.

Another example for an association found only by HHG is the association between the genes CDKN1A and TKT (Fig 2). CDKN1A codes for CDK-interacting protein 1 (p21), a potent cyclin-dependent kinase inhibitor that regulates cell cycle progression through the G1/S checkpoint. TKT codes for Transketolase, a central enzyme of the Pentose phosphate pathway. The association between CDKN1A and TKT detected by HHG reflects in fact a relationship between the pathways these two genes belong to. Following cell cycle progression from G1 towards the S phase, there is an up-regulation of the Pentose phosphate pathway, which is responsible for production of ribose-5-phosphate (R5P), needed for the synthesis of nucleotides and nucleic acids [25]. All the genes in the examples above are located on different

Table 1. Summary of the significant discoveries (after adjusting for multiple testing) found by Pearson or Spearman, dCov, MIC and HHG.

Pearson or Spearman (254)	dCov (256)	MIC (157)	HHG (400)	Number of discoveries
V	X	X	X	29
X	V	X	X	9
X	X	V	X	10
X	X	X	V	**178**
V	V	X	X	220
V	X	V	X	140
V	X	X	V	190
X	V	V	X	140
X	V	X	V	211
X	X	V	V	145
V	V	V	X	139
V	V	X	V	185
V	X	V	V	138
X	V	V	V	139
V	V	V	V	138

V and X, respectively, indicate whether the method is included or excluded in each comparison. For example, line 1 of the table indicates that 29 tests were found significant only by Pearson or Spearman; line 4 shows that 178 tests were found significant only by HHG; and the last line implies that Pearson or Spearman, dCov, MIC and HHG share 138 common significant findings.

doi:10.1371/journal.pone.0126544.t001

chromosomes or far away from each other on the same chromosome; hence physical proximity cannot explain the CNV-based associations.

Identification of gene function using multivariate association tests

Detection of associations between pairs of genes by univariate analysis is a good start towards deriving biological information from CNV data, as shown above. However, when dealing with a large number of genes, the function and a relation to biological pathways of many genes are often unknown. Finding associations with known genes may shed light on their possible function, but multivariate analysis could provide additional important information. Therefore, we applied the multivariate tests for dependence between several genes of unknown function in our dataset and known pathways, using dCov and HHG multivariate tests. Specifically, of the 99 genes in our dataset, twelve genes have no known function or relation to a biological pathway (Fig 3), as determined by using KEGG pathway ([26,27]; http://www.genome.jp/kegg/tool/map_pathway1.html). To detect their associations with known pathways, we first assigned the rest of the genes to pathways based on KEGG pathway mapper (S2 Table), and then selected eight experimentally proven biological pathways containing at least five genes from our dataset (Fig 3). In addition, the apoptosis pathway, being one of the basic cancer related mechanisms, was included in our study even though only two genes from our dataset have been assigned to it. Next, we tested for associations between each gene-pathway pair among those twelve genes and nine pathways. We applied dCov and HHG which were, of the tests we used above, the only two tests capable of multivariate analysis, i.e., testing for association between vectors (more details are available in the Materials and Methods Section). In total, 108 tests were performed with each method and a test result was considered significant if its FDR-adjusted p-value was less than 0.05 using the Benjamini-Hochberg procedure [21]. Of the twelve genes, six genes showed significant associations to pathways (Fig 3A and S3 Table).

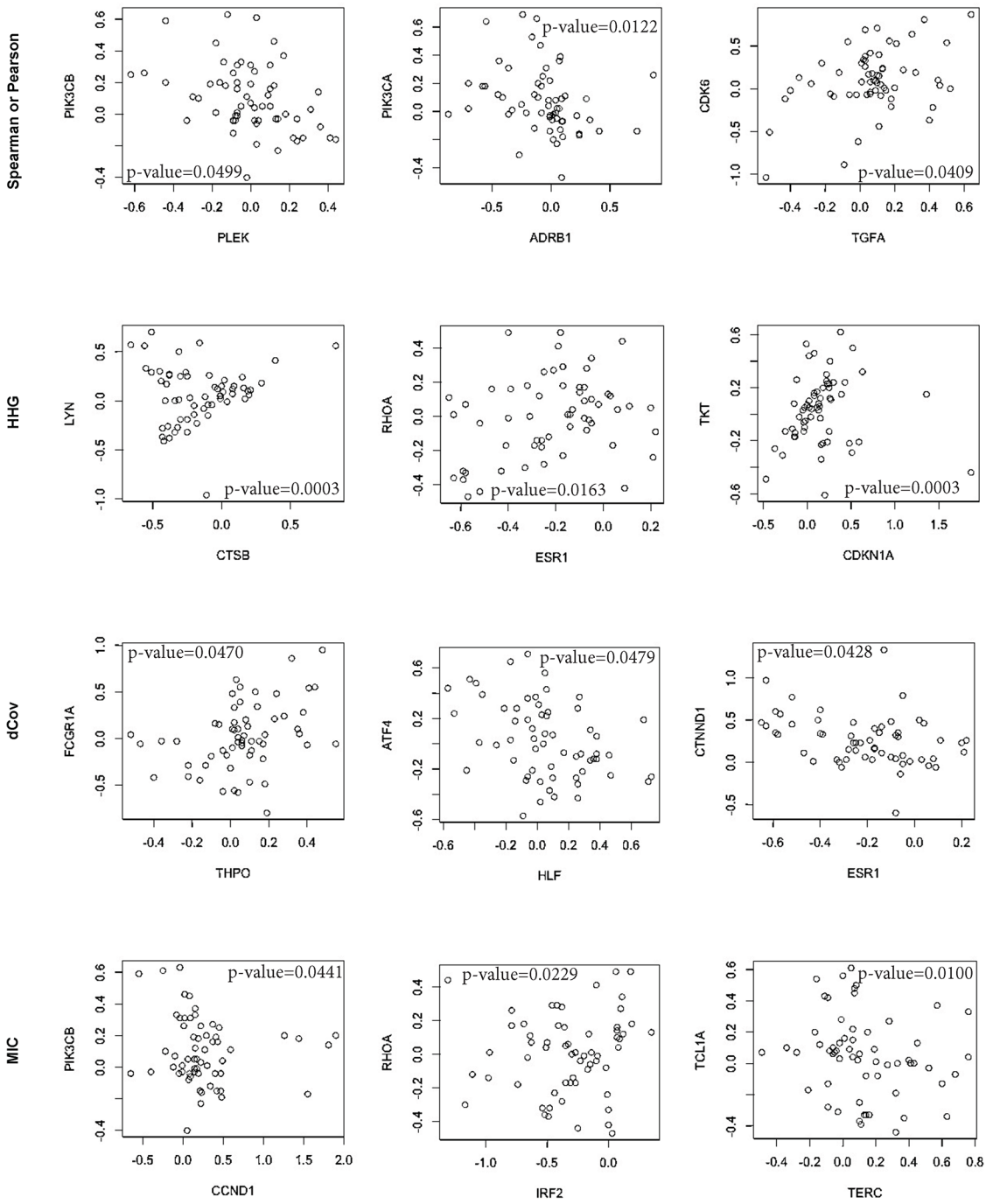

Fig 2. Example of significant relationships. First line consists of three findings discovered only by Spearman or Pearson; second, only by HHG; third, only by dCov; and fourth, only by MIC. P-values (after adjusting for multiple testing) are denoted in each plot.

doi:10.1371/journal.pone.0126544.g002

Two genes, LRRC32 and SPI1, were found to be associated with most of the pathways, suggesting they might be signal transduction intermediates, regulating downstream targets belonging to these pathways. These findings are in agreement with the results of the univariate analysis, which significantly associated both genes with serine/threonine kinase PAK1 and SPI1 gene also with HRAS, a GTPase of RAS family. Indeed, according to KEGG pathway

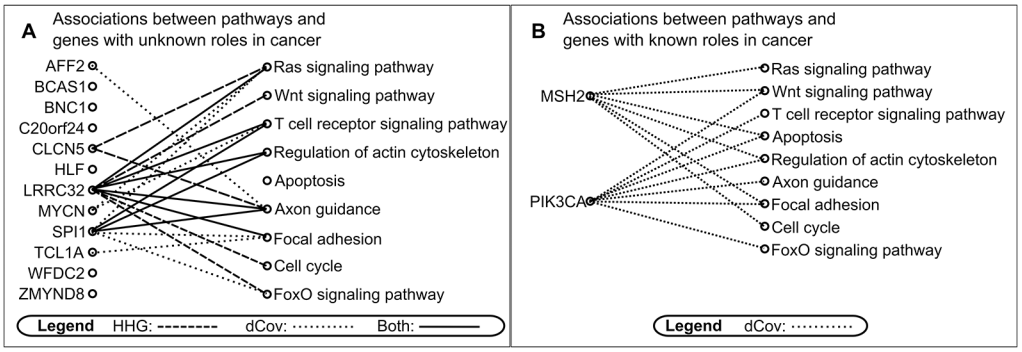

Fig 3. Bipartite graph displaying gene-to-pathway associations, as determined by HHG and dCov. In panels A and B, genes (on the left) and pathways (on the right) were analyzed for association by HHG and dCov. Significant associations (after adjusting for multiple testing) are linked by lines: dashed for HHG, dotted for dCov, and solid for both. A) Significant associations between genes with unknown function and cancer related pathways. Associations found by dCov and HHG are marked. B) Significant associations between genes with known function and cancer related pathways. Only associations found by dCov are shown as no significant associations were found by HHG.

doi:10.1371/journal.pone.0126544.g003

mapper PAK1 and HRAS belong to most of the pathways with which LRRC32 and SPI1 were found to be associated. Moreover, both PAK1 and HRAS are involved in transduction of proliferation signals and their miss-regulation leads to abnormal signal transduction and cancer [28,29]. Thus, while a univariate analysis could find association between genes of unknown function and individual genes with known function, the above multivariate analysis could point out their associations with biological processes.

The four remaining associated genes, AFF2, CLCN5, MYCN, and TCL1A, were found to be associated each to one or two specific pathways suggesting they constitute downstream effectors in these pathways (see examples below). No associations were found between the other six genes and any of the pathways.

In the multivariate analysis applied above to genes of unknown function, dCov and HHG discovered similar number of significant multivariate relationships, 15 by dCov, and 13 by HHG, while 8 were detected by both methods. Therefore our analysis did not reveal any clear evidence of superiority of one method over the other in this specific application.

In addition to the multivariate analysis applied to genes of unknown role in cancer, we picked two genes from the dataset, PIK3CA and MSH2, which have established biological function and do not belong to any of the eight pathways according to KEGG, and performed gene-pathway multivariate tests of association by dCov and HHG, similar to those performed above for genes of unknown function. While dCov found 13 significant results, HHG found none (Fig 3B and S4 Table).

The associations, detected by dCov, between MSH2 and cell cycle, apoptosis, focal adhesion, RAS, WNT and actin pathways are consistent with its function in DNA mismatch repair and its connection to cell division [31]. Similarly, associations between PIK3CA, and the following pathways: apoptosis, actin, Focal adhesion, FoxO signaling, T cell receptor signaling, Axon guidance and Wnt (Fig 3B and S4 Table) are supported by vast biological data [32–35]. The relation of PIK3CA to these pathways, as well as its pivotal role in human cancers, is a consequence of it being a key player in activation of signaling cascades involved in cell growth, survival, proliferation, motility and morphology [36]. The discrepancy in the current results of dCov and HHG (Fig 3B) is due to the linear nature of the relationship between these genes and the pathways, and the fact that the strength of HHG is in finding non-monotone relationships. For example, dCov discovered significant association between PIK3CA and the Axon guidance pathway. Looking back at the univariate analysis (S1 Table) we see that PIK3CA was found to

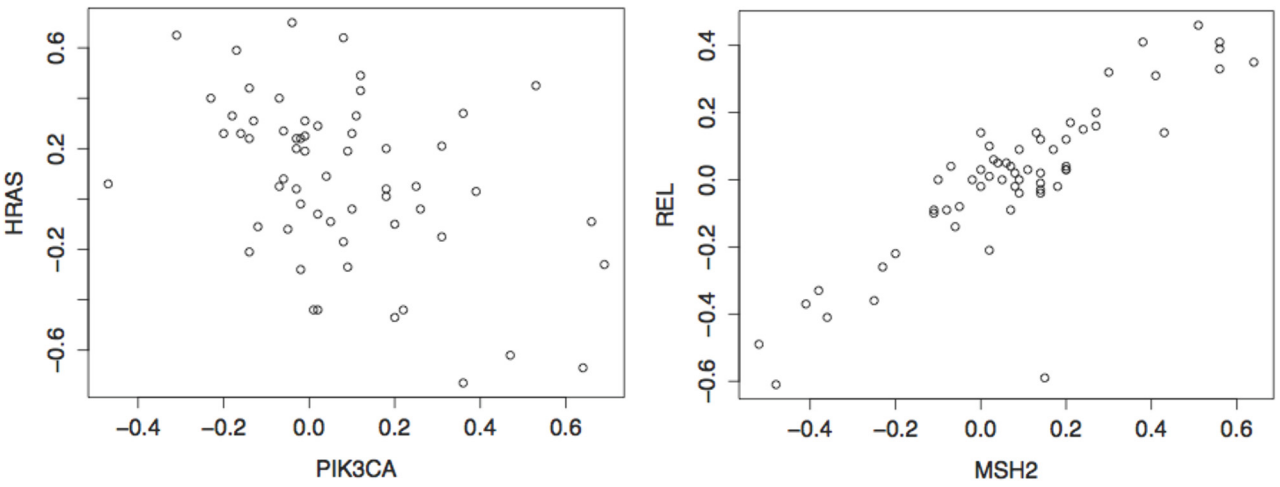

Fig 4. Indication for the linear associations that explains the difference between dCov and HHG in the multivariate analysis with known genes. Scatter plots of PIK3CA versus HRAS (left panel) and MSH2 versus REL (right panel).

doi:10.1371/journal.pone.0126544.g004

be significantly associated with HRAS, which belongs to the Axon guidance pathway, and this association was also found by Pearson or Spearman. Such results indicate strong linear relationship between PIK3CA and HRAS (Fig 4). Similarly, the association found by dCov, but not by HHG, between MSH2 and the Ras signaling pathway can be explained by the significant association found by Pearson or Spearman between MSH2 and gene REL, which belongs to this pathway (S1 Table, and Fig 4). It is expected that known relationships between genes discovered by laboratory methods (such as co-IP) or by bioinformatic analysis of high-throughput data based on classic linear or monotone oriented methods will be strongly biased towards linear or monotone relationships.

Collectively, these results provide a proof of concept for the ability of multivariate analysis to reveal biologically relevant gene-pathway associations.

Discussion

In this work we undertook a novel approach to whole genome CNV analysis, with the goal being identification of associations between CNV of different genes (CNV-CNV) across 60 human cancer cell lines. We used modern association tests that can detect non-linear and non-monotone associations and applied them in univariate settings, in attempt to identify gene-gene associations. We also used them in multivariate settings, in attempt to identify associations of genes of unknown function with established cancer-related pathways.

Collectively, our univariate analysis demonstrates that associations between CNV of genes found by HHG reflect true biological processes. This suggests that univariate analysis by means of statistical tests which target only linear or monotone associations might result in many biologically important findings remaining unrevealed. Additionally, in this dataset, the superiority of the HHG test over the other tests capable of detecting non-monotone relationships is obvious.

In the multivariate setting, the difference between the highly associated genes (LLRC32 and SPI1) and the other four associated genes is an example of how multivariate analysis can hint at the position of a gene within a pathway. Applied to a larger dataset and combined with univariate analysis, this analysis would allow even more refined positioning of a gene within a pathway.

Six genes did not associate with any of the pathways. This can be due to several reasons; one of them is the limited number of biological pathways with which the genes of unknown function were associated, as a consequence of a limited number of genes (99) with complete CNV data in the database used for this study. Another reason might be the limited biological data reported in KEGG, however this situation is anticipated to improve dramatically in the near future due to continuous accumulation of data from systems biology studies.

In case of LRRC32 and SPI1 discussed above, the univariate and multivariate results complement each other as these genes were found to be associated with pathways by the multivariate analysis and to the specific members of these pathways by the univariate analysis. However it is important to note that this is not a general rule. As a multivariate test of independence identifies dependency between two vectors, while a univariate method only loops over pairs of components and tests for dependency between each pair of variables. Therefore, it is possible to obtain non-significant univariate tests but a significant multivariate test for the same dataset. In fact there is a possibility of no association between any two individual genes and yet of a multivariate association with the full pathway. This can occur due to the combined effects of the variables in the multivariate test. For example, AFF2 was found to be significantly associated with the axon guidance pathway (adjusted p-value = 0.022) by multivariate analysis while no significant associations between AFF2 and any of the genes constituting the axon guidance pathway were found by the univariate analysis. This might be the result of weak associations between AFF2 and pathway members, or alternatively due to a strong association with a pathway member that was not included in the data. In any case, the discovered multivariate analysis gene-pathway association could not have been deduced based on the univariate analysis results.

In the opposite case, two genes, A and B, may be associated by univariate analysis, while no association between gene A and the pathway gene B belongs to is found by multivariate analysis. For example CLCN5 was found by the univariate analysis to be associated with MET and BCL2, both of which belong to the Focal adhesion pathway, which was not associated with CLCN5 by multivariate testing. A multivariate analysis did reveal, however, associations between CLCN5 and the Axon guidance and RAS pathways (Fig 3). Both of these pathways contain MET, the only pathway member found to be associated with CLCN5 by the univariate analysis. Such results are expected since MET is a Receptor Tyrosine Kinase, transducing signals from outside the cell, and thus is at the very start of many pathways, whereas BCL2 is a terminal protein in many pathways. This means that a univariate association with them is not strong enough to detect a pathway association. Corroboration that CLCN5 CNVs are associated with the Axon guidance pathway comes from the observation that 65.9% of central nervous system cancers have a loss of one or two copies of the CLCN5 gene (COSMOS, [30]).

These examples demonstrate the possible advantage of multivariate tests of independence over univariate tests when the goal is finding a relationship between a gene and a group of genes, such as a pathway, or finding an association between two groups of genes (e.g. two pathways). In general, in order to obtain a complete picture, both association tests types should be applied.

The dCov and the HHG tests are permutation tests, and the computation of many such tests can be computationally challenging. Distribution-free univariate tests of a flavor similar to HHG were recently introduced in [37]. These tests can be useful alternatives to the HHG test when a large number of univariate tests are simultaneously examined.

In summary, our results indicate: (1) Multivariate analysis is a very useful tool for ascribing biological roles to genes of unknown function; (2) Univariate omnibus analysis, i.e. using tests that detect all types of relationships, could uncover many new important associations that can not be detected by the common linear and monotone association tests; (3) The HHG test

outperformed all the other tests in finding univariate associations; And most importantly, (4) Using a combination of multivariate and univariate associations tests can reveal significant information about gene networks and, in the current context, about cancer-driving processes.

Materials and Methods

CNV databases

Comparative genomic hybridization (CGH) data of a panel of 60 human cancer cell lines (the NCI-60) was obtained from [12,38]. The CGH contains 349 clones. After excluding clones with missing values and clones with unknown gene symbols, our analysis was performed on a set of 99 CGH clones, representing 99 genes. S5 Table contains aCGH raw data from NCI-60.

Univariate analysis

Association analysis was performed on the 99 clones based on their copy number in each of the 60 cell lines from NCI-60. We tested all possible pair-wise associations among the 99 clones, generating 4851 pairs. We used the following tests of independence: (i) test based on Pearson correlation coefficient [39] (ii) test based on Spearman rank correlation coefficient [40] (iii) distance covariance (dCov) [16,17]; (iv) maximal information coefficient (MIC) [18]; and (v) a test based on ranks of distances (HHG) [19]. For each method we adjusted for multiple comparisons by FDR of Benjamini and Hochberg [21], and a test result was considered as significant if its adjusted p-value was less than or equal 0.05.

In the following we provide a summary of the tests. Assume we have N independent observations (X_i, Y_i), $i = 1, \ldots, N$, from the joint distribution of (X, Y), $X, Y \in R$ and our goal is to test whether there is a relationship between X and Y.

i. Pearson correlation coefficient. The sample Pearson correlation coefficient, denoted by r_p, is given

$$r_p = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{(N-1)S_X S_Y}$$

where $S_X^2 = \sum_{i=1}^{N}(X_i - \bar{X})^2/(N-1)$ and S_Y^2 is defined similarly based on $Y_1,\ldots,Y_N$. The value of r_p is between -1 and 1. r_p equals 1 or -1 corresponds to data points lying exactly on a line. A value of 0 implies that there is no linear correlation between X and Y. If (X, Y) follows the bivariate normal distribution, under the null hypothesis of no linear relationship between X and Y (i.e. the true correlation coefficient equals 0), $r_p\sqrt{(N-2)/(1-r_p^2)}$ follows a Student's t distribution with $N - 2$ degrees of freedom [39]. This Student's t distribution also holds approximately, if the distribution of (X, Y) is not normal but the sample size is large enough. We applied this test by using the function cor.test with parameter method = 'pearson' in the package *stats* of R (http://www.r-project.org).

ii. Spearman correlation coefficient. Spearman correlation coefficient, denoted by r_s, is defined similarly to r_p but instead of using the observed values their ranks are used [40]. In case of tied values, a rank equal to the average of their positions in the ascending order of the values is assigned. A value of 1 or -1 for r_s corresponds to the case in which X and Y are perfect monotone functions of each other. Under the null hypothesis of no monotone relationship between the variables and large sample size, $r_s\sqrt{(N-2)/(1-r_s^2)}$ follows a Student's t distribution with $N - 2$ degrees of freedom [40]. We applied this test by using the function cor.test with parameter method = 'spearman' in the package *stats* of R (http://www.r-project.org).

iii. The dCov test. The distance covariance test [16,17] uses all pairwise Euclidean distances $a_{ij} = |X_i - X_j|$ and $b_{ij} = |Y_i - Y_j|$, $i, j = 1,\ldots,N$. Then, the resulting two distance matrices are centered by

$$A_{ij} = a_{ij} - \frac{1}{N}\sum_{i=1}^{N} a_{ij} - \frac{1}{N}\sum_{j=1}^{N} a_{ij} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}$$

and

$$B_{ij} = b_{ij} - \frac{1}{N}\sum_{i=1}^{N} b_{ij} - \frac{1}{N}\sum_{j=1}^{N} b_{ij} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}.$$

The sample distance covariance, is defined as the average of the componentwise product matrix of the two centered distance matrices: $\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}B_{ij}/N^2$, and is the test statistic for testing the null hypothesis of independence between X and Y. The value of the population distance covariance of X and Y equals zero if and only if they are independent. The dCov test is implemented in the R package *energy* as a permutation test (http://www.r-project.org).

iv. The MIC test. The test of MIC [18] is based on the discrete version of the mutual information

$$\sum_{x}\sum_{y} \Pr(X = x, Y = y)\log\left\{\frac{\Pr(X = x, Y = y)}{\Pr(X = x)\Pr(Y = y)}\right\}.$$

If a relationship exists between two variables, then a grid can be drawn on the scatter plot of the two variables that partitions the data to encapsulate that relationship. Specifically, consider all grids G partitioning the X-values and Y-values into x and y bins, respectively. Let $I(G; x, y)$ be the empirical mutual information of a grid G with x and y bins, such that the probability distribution functions are replaced by the fraction of observations falling in that cell.

Their aim was to use as test statistic

$$M = \max_{(x,y)}\left\{\frac{\max_G I(G; x, y)}{\log\min\{x, y\}}\right\}.$$

In practice, the MIC test statistic is based on a dynamic programming algorithm that only approximates $\max_G I(G; x, y)/\log\min\{x, y\}$, and the outer maximization step in M is over (x, y) such that $xy < N^{0.6}$. The MIC test statistic, under the null hypothesis of independence, depends only on the ranks of the data, therefore look-up tables of the quantiles of the null distribution were generated for various sample sizes. The code for applying the MIC test and the look-up tables are available at the MINE website: exploredata.net.

v. The HHG test. The rationale of the HHG [19] test is the observation that if X and Y are associated, closeness in the X-values tends to give rise to closeness in Y-values. The test is based on the pairwise distances between the sample values of X and Y respectively,$\{d_X(X_i, X_j): i, j \in \{1,\ldots, N\}\}$, $\{d_Y(Y_i, Y_j): i, j \in \{1,\ldots, N\}\}$. The only restriction on the distance metrics $d_X(\cdot,\cdot)$ and $d_Y(\cdot,\cdot)$ is that they are determined by norms. For simplicity of notation we consider here identical norm distances for X and Y, denoted by $d(\cdot,\cdot)$. Consider two fixed observations i and j, and for $m = 1,\ldots,N$ let $\delta_m(X_i, X_j) = I\{d(X_i, X_m) \leq d(X_i, X_j)\}$ and $\mu_m(Y_i, Y_j) = I\{d(Y_i, Y_m) \leq d(Y_i, Y_j)\}$, where $I\{A\}$ is an indicator function that equals 1 when A is true, and 0 otherwise. For each fixed i and j, a 2×2 contingency table is constructed based on δ_m and μ_m, $m = 1,\ldots, N$, $m \neq i, j$, with entries $A_{11}(i, j)$, $A_{12}(i, j)$, $A_{21}(i, j)$, $A_{22}(i, j)$, where

$A_{11}(i,j) = \sum_{m=1,m\neq i,j}^{N} \delta_m(X_i, X_j)\mu_m(Y_i, Y_j)$, A_{12}, A_{21}, A_{22} are similarly defined, $A_{.k} = A_{1k} + A_{2k}$ and $A_{k.} = A_{k1} + A_{k2}$, k = 1,2. Then, the HHG test statistic is defined as $T = \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} S(i,j)$ where $S(i,j)$ is either Pearson's chi squared test statistic based on the contingency table given i and j, namely,

$$S_p(i,j) = \frac{(N-2)\{A_{12}(i,j)A_{21}(i,j) - A_{11}(i,j)A_{22}(i,j)\}}{A_{1.}(i,j)A_{.1}(i,j)A_{2.}(i,j)A_{.2}(i,j)},$$

or the log-likelihood ratio statistic,

$$S_{LR}(i,j) = 2\sum_{k=1,2}\sum_{l=1,2} A_{kl}(i,j)\log\left\{\frac{(N-2)A_{kl}(i,j)}{A_{k.}(i,j)A_{.l}(i,j)}\right\}.$$

In case of zero margin in the contingency table, $S_P(i,j) = 0$, and for $S_{LR}(i,j)$, a term is zero if $A_{kl}(i,j) = 0$. The sampling distribution of the test statistic under the null hypothesis is computed based on values of T under random shuffling of the indices of X. The p-value of this permutation test is computed by ranking the observed test statistic among the shuffled test statistics. The current analysis is based on Euclidean distances and the Pearson's chi squared test statistic. The HHG test is implemented in the R package *HHG* (http://www.r-project.org).

Multivariate analysis

We grouped some of the CGH genes into nine different pathways. We used KEGG Mapper—Search Pathway [26,27] to map genes into pathways and chose only pathways that are not specific to cancer. In addition, we selected pathways with at least five genes. We included the apoptosis pathway even though it has only two genes because of the importance of this pathway in cancer. A separate similar analysis was conducted with MSH2 and PIK3CA genes, which have known function and established biological role in cancer. The aim of the multivariate analysis was to test whether there is an association between any of the genes and the pathways. For the multivariate analysis, we used the dCov and HHG tests, as they are multivariate consistent tests against all alternatives. Pearson, Spearman and MIC are not applicable in multivariate settings. Under the multivariate setting, we let X and Y be random vectors of lengths p and q, respectively, and F_X, F_Y, F_{XY} denote the respective multivariate cumulative distribution functions of X, Y and (X,Y). Under the null hypothesis the vectors X and Y are independent, namely, H_0: $F_{XY} \neq F_X F_Y$, and under the alternative the vectors are dependent, H_1: $F_{XY} \neq F_X F_Y$. The dCov and HHG tests were applied using the Euclidean norm as described above. It should be noted that dCov and HHG are applicable under any dimensions p and q, even for p or q that are greater than the sample size N.

Supporting Information

S1 Fig. Summary of data analysis by area-proportional Euler diagram. The area of each oval represents the number of significant tests found by each method, and intersections (emphasized by different colors) represent common discoveries. The numbers represent the number of significant tests at 0.05 significance level after FDR multiplicity correction.
(PDF)

S1 Table. Adjusted p-values calculated by multiple statistical methods. All the genes in our dataset were tested against each other for association. Each pair of genes was tested with Pearson, Spearman, MIC, dCov and HHG, yielding five p-values. The statistically significant

(adjusted p-value<0.05) results are marked in yellow.
(XLSX)

S2 Table. The pathways used for the multivariate analysis, along with the genes they consist of. A list of pathways obtained from KEGG. Alongside each pathway are the genes from our dataset included within the KEGG pathway. The pathways chosen were those containing at least five genes from our list, or apoptosis—which we considered exceptionally interesting.
(XLSX)

S3 Table. Adjusted p-values for testing for association between unknown genes and pathways by dCov and HHG. The pathways in the table were curated from KEGG. A row in the table represents an association test between a pathway and a gene. Each pathway and gene pair appears together with the dCov and HHG adjusted p-values that result from the test. Statistically significant p-values are marked in yellow for HHG and green for dCov. For each gene we tested for association between the gene and the genes in the pathway, and placed the ones yielding statistically significant result (adjusted p-value<0.05) in either HHG or dCov.
(XLSX)

S4 Table. Adjusted p-values for testing for association between known genes and pathways by dCov and HHG. The pathways in the table were curated from KEGG. A row in the table represents an association test between a pathway and a gene. Each pathway and gene pair appears together with the dCov and HHG adjusted p-values that result from the test. Statistically significant p-values are marked in yellow for HHG and green for dCov. For each gene we tested for association between the gene and the genes in the pathway, and placed the ones yielding statistically significant result (adjusted p-value<0.05) in either HHG of dCov.
(XLSX)

S5 Table. aCGH Raw data from NCI-60 that was used in this paper.
(TXT)

Acknowledgments

We thank Itai Yanai for critical reading of the manuscript and Mahlet G Tadesse for introducing us the NCI-60 dataset.

Author Contributions

Conceived and designed the experiments: MG BG ATL. Analyzed the data: MG BG ATL. Contributed reagents/materials/analysis tools: MG RH YH. Wrote the paper: MG BG ATL AF.

References

1. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science (New York, NY). 2005; 307(5714):1434–40. doi: 10.1126/science.1101160
2. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 2007; 39(7 Suppl):S16–21. doi: 10.1038/ng2028 PMID: 17597776
3. Malhotra D, Sebat J. CNVs: Harbinger of a Rare Variant Revolution in Psychiatric Genetics. Cell. 2012; 148(6):1223–41. doi: 10.1016/j.cell.2012.02.039 PMID: 22424231
4. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006; 439(7078):851–5. doi: 10.1038/nature04489 PMID: 16482158
5. Cappuzzo F, Hirsch FR, Rossi E, Bartolini S, Ceresoli GL, Bemis L, et al. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. J Natl Cancer Inst. 2005; 97(9):643–55. doi: 10.1093/jnci/dji112 PMID: 15870435

6. Valsesia A, Mace A, Jacquemont S, Beckmann JS, Kutalik Z. The growing importance of CNVs: new insights for detection and clinical interpretation. Front Genet. 2013; 4. doi: 10.3389/fgene.2013.00092
7. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007; 39:S37–S42. doi: 10.1038/ng2080 PMID: 17597780
8. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: New insights in genome diversity. Genome Res. 2006; 16(8):949–61. doi: 10.1101/gr.3677206 PMID: 16809666
9. Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet. 2013; 206(12):432–40. doi: 10.1016/j.cancergen.2013.11.002 PMID: 24405614
10. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget. 2013; 4(11):1868–81. PMID: 24240121
11. Stamoulis C, Betensky RA. A novel signal processing approach for the detection of copy number variations in the human genome. Bioinformatics (Oxford, England). 2011; 27(17):2338–45. doi: 10.1093/bioinformatics/btr402 PMID: 21752800
12. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Molecular cancer therapeutics. 2006; 5(4):853–67. doi: 10.1158/1535-7163.MCT-05-0155 PubMed PMID: 16648555; PubMed Central PMCID: PMC2733874.
13. de Siqueira Santos S, Takahashi DY, Nakata A, Fujita A. A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief Bioinformatics. 2013. doi: 10.1093/bib/bbt051
14. Kumari S, Nie J, Chen H-S, Ma H, Stewart R, Li X, et al. Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. PLoS ONE. 2012; 7(11). doi: 10.1371/journal.pone.0050411
15. Guo X, Zhang Y, Hu W, Tan H, Wang X. Inferring Nonlinear Gene Regulatory Networks from Gene Expression Data Based on Distance Correlation. PLoS ONE. 2014; 9(2). doi: 10.1371/journal.pone.0087446
16. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. The Annals of Statistics. 2007; 35(6):2769–94.
17. Székely GJ, Rizzo ML. Brownian distance covariance. The annals of applied statistics. 2009; 3 (4):1236–65.
18. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011; 334(6062):1518–24. doi: 10.1126/science.1205438 PMID: 22174245
19. Heller R, Heller Y, Gorfine M. A consistent multivariate test of association based on ranks of distances. Biometrika. 2013; 100(2):503–10.
20. Sedaghat N, Saegusa T, Randolph T, Shojaie A. Comparative study of computational methods for reconstructing genetic networks of cancer-related pathways. Cancer informatics. 2014; 13(Suppl 2):55–66. doi: 10.4137/CIN.S13781 PubMed PMID: 25288880; PubMed Central PMCID: PMC4179645.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289–300.
22. Urtz N, Olivera A, Bofill-Cardona E, Csonga R, Billich A, Mechtcheriakova D, et al. Early Activation of Sphingosine Kinase in Mast Cells and Recruitment to FcepsilonRI Are Mediated by Its Interaction with Lyn Kinase. Mol Cell Biol. 2004; 24(19):8765–77. doi: 10.1128/MCB.24.19.8765–8777.2004 PMID: 15367693
23. Taha TA, Kitatani K, Bielawski J, Cho W, Hannun YA, Obeid LM. Tumor necrosis factor induces the loss of sphingosine kinase-1 by a cathepsin B-dependent mechanism. J Biol Chem. 2005; 280 (17):17196–202. doi: 10.1074/jbc.M413744200 PMID: 15710602
24. Taha TA, El-Alwani M, Hannun YA, Obeid LM. Sphingosine kinase-1 is cleaved by cathepsin B in vitro: Identification of the initial cleavage sites for the protease. FEBS Lett. 2006; 580(26):6047–54. doi: 10.1016/j.febslet.2006.09.070 PMID: 17064696
25. Vizán P, Alcarraz-Vizán G, Díaz-Moralli S, Solovjeva ON, Frederiks WM, Cascante M. Modulation of pentose phosphate pathway during cell cycle progression in human colon adenocarcinoma cell line HT29. Int J Cancer. 2009; 124(12):2789–96. doi: 10.1002/ijc.24262 PMID: 19253370
26. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28 (1):27–30. PMID: 10592173

27. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014; 42(Database issue):D199–205. doi: 10.1093/nar/gkt1076 PMID: 24214961

28. Ong CC, Jubb AM, Zhou W, Haverty PM, Harris AL, Belvin M, et al. p21-activated kinase 1: PAK'ed with potential. Oncotarget. 2011; 2(6):491–6. PMID: 21653999

29. Grabocka E, Pylayeva-Gupta Y, Jones MJK, Lubkov V, Yemanaberhan E, Taylor L, et al. Wild-type H- and N-Ras promote mutant K-Ras-driven tumorigenesis by modulating the DNA damage response. Cancer Cell. 2014; 25(2):243–56. doi: 10.1016/j.ccr.2014.01.005 PMID: 24525237

30. Franchitto A, Pichierri P, Piergentili R, Crescenzi M, Bignami M, Palitti F. The mammalian mismatch repair protein MSH2 is required for correct MRE11 and RAD51 relocalization and for efficient cell cycle arrest induced by ionizing radiation in G2 phase. Oncogene. 2003; 22(14):2110–20. doi: 10.1038/sj.onc.1206254 PMID: 12687013

31. Saijilafu, Hur E-M, Liu C-M, Jiao Z, Xu W-L, Zhou F-Q. PI3K–GSK3 signalling regulates mammalian axon regeneration by inducing the expression of Smad1. Nat Commun. 2013; 4. doi: 10.1038/ncomms3690

32. Perry JM, He XC, Sugimura R, Grindley JC, Haug JS, Ding S, et al. Cooperation between both Wnt/β-catenin and PTEN/PI3K/Akt signaling promotes primitive hematopoietic stem cell self-renewal and expansion. Genes Dev. 2011; 25(18):1928–42. doi: 10.1101/gad.17421911 PMID: 21890648

33. Tenbaum SP, Ordóñez-Morán P, Puig I, Chicote I, Arqués O, Landolfi S, et al. β-catenin confers resistance to PI3K and AKT inhibitors and subverts FOXO3a to promote metastasis in colon cancer. Nat Med. 2012; 18(6):892–901. doi: 10.1038/nm.2772 PMID: 22610277

34. Wang J, Kuropatwinski K, Hauser J, Rossi MR, Zhou Y, Conway A, et al. Colon carcinoma cells harboring PIK3CA mutations display resistance to growth factor deprivation induced apoptosis. Molecular cancer therapeutics. 2007; 6(3):1143–50. doi: 10.1158/1535-7163.MCT-06-0555 PMID: 17363507

35. Karakas B, Bachman KE, Park BH. Mutation of the PIK3CA oncogene in human cancers. Br J Cancer. 2006; 94(4):455–9. doi: 10.1038/sj.bjc.6602970 PMID: 16449998

36. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10:Unit-10.1. doi: 10.1002/0471142905.hg1011s57

37. Heller R, Heller Y, Kaufman S, Brill B, Gorfine M. Consistent distribution-free K-sample and independence tests for univariate random variables. arXiv:14106758 [stat]. 2014.

38. Massion PP, Kuo WL, Stokoe D, Olshen AB, Treseler PA, Chin K, et al. Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: implications of the phosphatidylinositol 3-kinase pathway. Cancer research. 2002; 62(13):3636–40. PubMed PMID: 12097266.

39. Edwards AL. The Correlation Coefficient. An Introduction to Linear Regression and Correlation. CA: W.H. Freeman; 1976. p. 33–46.

40. Lehmann EL, D'Abrera HJM. Nonparametrics: Statistical Methods Based on Ranks (rev. ed.) Prentice-Hall. Englewood Cliffs, NJ. 1998:292,300,23.