Genomic regression analysis of coordinated expression

Ling Cai1,2, Qiwei Li2, Yi Du3, Jonghyun Yun4, Yang Xie2, Ralph J. DeBerardinis1 & Guanghua Xiao2

Co-expression analysis is widely used to predict gene function and to identify functionally related gene sets. However, co-expression analysis using human cancer transcriptomic data is confounded by somatic copy number alterations (SCNA), which produce co-expression signatures based on physical proximity rather than biological function. To better understand gene–gene co-expression based on biological regulation but not SCNA, we describe a method termed “Genomic Regression Analysis of Coordinated Expression” (GRACE) to adjust for the effect of SCNA in co-expression analysis. The results from analyses of TCGA, CCLE, and NCI60 data sets show that GRACE can improve our understanding of how a transcriptional network is re-wired in cancer. A user-friendly web database populated with data sets from The Cancer Genome Atlas (TCGA) is provided to allow customized query.
Highly coordinated expression of genes functioning in common processes is a widespread phenomenon in prokaryotes and eukaryotes. Spatial and temporal compartmentalization of gene expression in response to environmental cues allows cells to avoid futile reactions in metabolism, promotes efficiency in the stoichiometric assembly of macromolecular complexes, and reduces noise in signal transduction pathways, among other functions. Co-expression analysis is a widely adopted tool for functional prediction and identification of functionally related gene sets. Such analysis identifies associated genes based on their highly correlated expression profiles from high-throughput gene expression profiling data obtained by microarray or RNA-seq experiments.

Great efforts from multiple projects such as The Cancer Genome Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE) have been put forth to characterize cancer genomics from tumor tissues and cell lines. A rich resource of cancer transcriptome data has been generated from these studies. Many popular cancer data-mining platforms such as Oncomine, cBioPortal, and CCLE provide co-expression analysis tools based on these cancer transcriptome data. With these tools, users can search for genes co-expressed with a gene of interest in different tumor types or in cancer cell lines. However, results from co-expression analyses using standard methods are often hard to interpret because many top correlating genes arise from somatic copy number alteration (SCNA) rather than bona fide transcriptional regulation. Copy number alteration (CNA) is a common feature in cancer. Multiple studies have shown that gene expression levels are correlated with copy numbers in cancer cells as well as cultured human pluripotent stem cells. Notably, Fehrmann et al., who re-analyzed 77,840 expression profiles, found that 99% of all abundantly expressed human genes exhibited a positive correlation with DNA copy number. Dosage-compensation mechanisms, which ensure that genes are expressed at appropriate levels irrespective of their copy numbers, have not been found to exist for autosomes in humans. Hence, variation of RNA levels in cancer samples is a combined consequence of CNA and biological regulation of transcript production, processing and decay. Several groups have used copy number-adjusted expression values to understand transcriptional consequences of genomic aberrations. However, no method exists to remove the confounding effect of CNA in the analysis of gene–gene co-expression using cancer transcriptome data.

To better understand gene–gene co-expression based on biological regulation but not SCNA, here we describe a method named genomic regression analysis of coordinated expression (GRACE) to adjust for the effect of CNA from co-expression analysis. Through comprehensive analyses of genetics, genomics, proteomics, metabolomics, and drug response data from the public domain, we show that GRACE can improve our understanding of how a transcriptional network is re-wired in cancer. A user-friendly web database has also been built with data from multiple TCGA cohorts to allow for customized query.

**Results**

**SCNA in cancer cause co-expression of neighboring genes.** SCNA is frequently observed in cancer. In normal tissues where all cells are diploid, gene–gene correlations are usually a result of coordinated biological regulation to allow concerted expression of genes with related functions. In the presence of SCNA, the expression of genes becomes proportional to the copy number, while copy number changes in functionally related genes are often independent of each other. This results in a reduction in correlation between functionally related genes and an increase in correlation between neighboring genes situated in the same DNA segment that is affected by the same CNA event (Fig. 1a).

The impact of SCNA confounds the interpretation of co-expression genes (Fig. 1b). Here we provide some examples based on over one thousand tumor samples from TCGA Breast Invasive Carcinoma (BRCA) cohort. Significant focal and arm level SCNA events have been found for these samples. For example, chromosome 1p is frequently deleted, while chromosome 1q is frequently amplified (Fig. 1c). We examined the impact of SCNA on mRNA abundance in this cohort. A correlation heatmap revealed a strong positive correlation between expression levels and copy number values for genes in the same arm of chromosome 1 (Fig. 1d), and a similar pattern was also observed in other chromosomes. Importantly, because SCNA occurs in segments that cover multiple genes, RNA levels of a gene not only correlate with its own copy number values but also broadly correlate with the copy numbers of the physically neighboring genes, which causes the co-expression of neighboring genes (Fig. 1b, e). This bias toward positive correlation between RNA and copy number occurs for genes within the same chromosome but not for genes from different chromosomes (Fig. 1f).

**GRACE adjusts for effects of SCNA in co-expression analysis.** To adjust for the variation in gene expression contributed by SCNA, we fit a linear regression model using the copy number values as the predictor variable and RNA levels as the response variable. The residuals from this linear regression model represent variations in gene expression that could not be accounted for by SCNA. To apply this method, samples with both gene expression and copy number data were selected. Genes were filtered to remove under-expressed genes or genes with saturated copy number values so that both the RNA and copy number of the same gene could have an approximately linear relationship ("Methods"). The residuals for each gene were then calculated and used for subsequent co-expression analysis (Fig. 2a). Collectively, we refer to this method as genomic regression analysis of co-expression (GRACE). We use an autocorrelation plot to visualize the effect of our method in removing correlation from neighboring genes due to SCNA. Autocorrelation was calculated for all genes and averaged over all tumor samples or all normal samples. The average autocorrelation of gene expression in tumor samples is much higher than that of the normal samples presumably due to SCNA, and adjusting expression with copy number data brings down the gene–gene autocorrelation in tumor (Fig. 2b).

In Fig. 2c–h, we provide a specific example using EIF2D, a gene encoding eukaryotic translation initiation factor 2D, to compare the results from co-expression analysis with the standard method or GRACE using data from the TCGA BRCA cohort.

EIF2D is located on chromosome 1q32.1. In TCGA BRCA tumor samples, the gene expression levels of EIF2D positively correlate with its copy number levels (Fig. 2a) as a result of the frequent amplification events of chromosome 1q (Fig. 1a). The residuals from regressing gene expression levels of EIF2D on the copy number levels of EIF2D were calculated to represent the variations in EIF2D transcript levels after adjusting for its copy number levels (Fig. 2a). The standard method of co-expression analysis calculates expression correlation among genes based on gene expression levels and sorts the genes by strength of positive correlation, and the resulting top 10 EIF2D co-expressing genes are all from chromosome 1q, neighboring the EIF2D locus (Fig. 2c). In contrast, using our method GRACE, which calculates correlation based on copy number-adjusted gene expression levels, the resulting top 10 EIF2D co-expressing genes are from various chromosomes and are almost exclusively ribosomal protein genes involved in translational processes similar to EIF2D (Fig. 2d). An extended view of the top 200 EIF2D co-expressing genes was provided in Fig. 2e. 

**Analysis of gene co-expression in BRCA**

In the BRCA cohort, two neighboring genes, EIF2D and EIF2C, were co-expressed with 99% of all abundantly expressed human genes (Fig. 2f). This bias is due to SCNA because the expression of EIF2C is negatively correlated with the expression of its neighboring gene, EIF2D (Fig. 2e). Using our method, GRACE, we can remove this bias and find the top 1000 co-expressed genes that are not influenced by SCNA (GRACE score). In contrast, using the standard method without correcting for SCNA, we find a similar pattern (GRACE score). We found that 95% of the top 1000 co-expressed genes after removing SCNA were not in the top 1000 co-expressed genes when using the standard method. The top 1000 co-expressed genes after removing SCNA were found to be almost exclusively ribosomal protein genes involved in translational processes similar to EIF2D (Fig. 2d). An extended view of the top 200 EIF2D co-expressing genes was provided in Fig. 2e.

**Analysis of gene co-expression in other tumor types**

In addition to BRCA, we applied GRACE to other cancer types including lung, breast, prostate, colon, and brain. The results were similar to those in BRCA, with the top 1000 co-expressed genes after removing SCNA being almost exclusively ribosomal protein genes involved in translational processes similar to EIF2D (Fig. 2d).
**Fig. 1** Increased correlations between genes from neighboring loci as a result of copy number variation. 

**a** Schematic diagram illustrating an example of copy number alteration confounding co-expression analysis. In this example, three subunits of a protein complex are encoded by three genes from different chromosomes. In normal tissues, coordinated transcriptional regulation ensures equal amounts of three subunits are produced to facilitate stoichiometric assembly of the complex. Transcript levels of these three genes will therefore be highly correlated. In tumor samples, the third gene is amplified together with a neighboring oncogene with unrelated functions. Consequently, correlation of the third gene with the other two related genes decreases, whereas its correlation with the unrelated oncogene neighbor increases. 

**b** Schematic diagram delineating correlation relationships between copy numbers of neighboring genes, between RNA levels of neighboring genes and between copy number and RNA levels of the same gene as a result of CNA events in cancer. 

**c** Relative copy number for genes on Chromosome 1 in 1075 tumor samples from TCGA breast cancer cohort. Genes are ordered by location on chromosome, and borders of cytobands from p and q arms are marked on top. P arm is frequently deleted and q arm is frequently amplified. 

**d** Copy number-RNA Pearson correlation matrix for genes from chromosome 1. Arm-level copy-number alteration events from chromosome 1 results in positive correlation between copy number and RNA for genes from the same arm. 

**e** RNA-RNA Pearson correlation matrix for genes from chromosome 1. Co-amplification or co-deletion of genes located in the same segment result in increased positive correlation between neighboring genes. 

**f** Copy number-RNA Pearson correlation matrix for all genes. Genes are ordered by location in chromosomes, and names of chromosomes are given on top. Increased positive correlation is observed only for genes from the same chromosome or chromosomal arm but not for genes from different chromosomes.
Fig. 2 Genomic regression analysis of co-expression (GRACE) corrects for correlation bias due to copy number variation. a Relative copy number levels vs. RNA levels of EIF2D for all 1075 tumor samples. The residuals from the linear regression model fitted with all 1075 tumor samples were marked by purple lines. b Average autocorrelation of neighboring genes in TCGA BRCA samples. Correlations between neighboring genes separated by 0–40 genes in between were calculated as autocorrelation for each sample. The average of autocorrelations was taken for 1075 tumor samples and 112 normal samples. The autocorrelation of neighboring genes based on copy number-adjusted expression levels variation (tumor residuals) was markedly reduced compared to autocorrelation calculated as autocorrelation for each sample. c Top 10 EIF2D correlated genes by standard method. d Top 10 EIF2D correlated genes by GRACE. e EIF2D correlated genes by standard method with a p value of 0.10. f Enrichment of KEGG_RIBOSOME genes from the top 100 EIF2D correlated genes by standard method with a p value of 0.00. g EIF2D correlated genes by standard method with a p value of 9.0e−143. h EIF2D correlated genes by GRACE from Chr1q. i Relative frequency distribution of chromosomal neighbors in top 10 co-expressing genes for all genes. GRACE markedly reduced the number of chromosomal neighbors in the top 10 co-expressing genes. j, k Kernel density estimation plots that visualize the distribution of pooled Spearman rank correlation coefficients for top 10 co-expressing genes from the same chromosome (j) or not from the same chromosome (k). Compared to the standard method, GRACE decreased intra-chromosomal gene–gene correlation and increased inter-chromosomal gene–gene correlation. All analyses are based on TCGA BRCA data.
genes by the standard method or by GRACE again shows that genes identified by the standard method are heavily biased toward genes in the physical proximity of EIF2D, while genes identified by GRACE bear more biological relevance to the function of EIF2D in protein translation (Fig. 2e,f,h).

Following this specific example, we also performed a systematic comparison between GRACE and the standard methods using the TCGA BRCA data. For every gene, we counted how many of its top 10 co-expressing genes were from the same chromosome on which this gene is located. This count of chromosomal neighbor genes ranges from 0 to 10. The relative frequency distributions of these counts are given in Fig. 2i. With the standard method, over 25% of genes have all of the top 10 co-expressing genes coming from the same chromosome, while this number is reduced to <2% with GRACE. GRACE has largely reduced the number of physically neighboring genes in the top 10 co-expressing genes. We have also calculated such distribution using normal samples from the TCGA BRCA cohort, which are presumably all diploid cells. Furthermore, we computed the simulated distribution assuming all genes are randomly distributed across 23 chromosomes (the Y chromosome is excluded from the TCGA BRCA data). The normal samples have more co-expressing genes from the same chromosome than the simulated distribution, which agrees with the notion that gene order in the human genome is not completely random and certain clusters of genes with functional relevance are co-expressed.

Next, we compared the strength of correlation among genes from the same chromosome and among genes from different chromosomes. As expected, the overall correlation among genes from the same chromosome is reduced with GRACE compared to the standard method (Fig. 2j). On the other hand, correlations among genes from different chromosomes are improved with GRACE compared to the standard method (Fig. 2k). This increase in significance of inter-chromosomal gene–gene co-expression by GRACE is likely a result of SCNA noise removal in co-expression analysis and could help us better understand the biological regulation of transcript levels in the cancer context. Similar results are also observed with the CCLE cell line data and METABRIC discovery set data (Supplementary Fig. 1).

GRACE facilitates discovery of tumor-specific co-expression. Tissue-specific gene networks are useful in capturing tissue-specific functional interactions. But in cancer, the transcription network could be rewired to support cancer-specific needs. Here with two specific examples, we show that GRACE can help us better understand how coordinated gene expression is rewired from normal to tumor tissues. Supplementary Data 1 include the top results from systematic identification of tumor-unique- or normal-unique-coexpressing gene enrichment in gene families from HUGO Gene Nomenclature Committee (HGNC) classification or canonical pathway gene sets from the Molecular Signatures Database (MSigDB) collection.

In the first example, we examined the difference in co-expressing genes for Poly (ADP-ribose) polymerase-2 (PARP2) in tumor tissues vs. normal tissues (Fig. 3a–c). PARP2 is a member of the PARP family. The founding member PARP1 is a cancer therapeutic target famous for its role in detection of DNA damage and recruitment of various proteins for DNA repair. PARP inhibitors that target both PARP1 and PARP2 have been used to treat BRCA1- and BRCA2-deficient tumors and the synergism in conjunction with DNA damaging agents is also being tested in various clinical trials. Despite their similar role in DNA repair, PARP2 lacks the N-terminal DNA-binding zinc fingers in PARP1. With the TCGA BRCA tumor samples, the top PARP2 co-expressing genes found by the standard method are mostly located near the PARP2 locus at chromosome 14q11.2, whereas the top co-expressing genes called by GRACE are from various chromosomes and many are involved in cell cycle processes (Fig. 3a). The top PARP2 co-expressing genes found using the normal samples, however, are far less related to cell cycle. Only two of the top 100 PARP2 co-expressing genes in BRCA normal samples belong to the "REACTOME CELL CYCLE" gene set, while close to half of the top 100 PARP2 co-expressing genes in BRCA tumor samples called by GRACE belong to this gene set (Fig. 3b). This difference between PARP2 co-expressing genes in tumor samples and in normal samples is consistently observed over multiple TCGA cohorts (Fig. 3b). Moreover, genes belonging to the C2H2-type zinc fingers gene family from the HGNC database are frequently found in the top 100 PARP2 co-expressing genes from normal tissues of several TCGA cohorts (Fig. 3c). This difference in the function of PARP2 co-expressing genes between tumor samples and normal samples suggests that PARP2 might be highly involved in DNA damaging repairs in tumor because of the genomic instability in cancer, while in normal tissues it might be involved in processes that also require C2H2-type zinc finger proteins, such as transcriptional regulation. The levels of PARP2 transcripts are similar in normal and tumor samples, whereas PARP1 has on average a twofold increase in tumor samples compared to normal samples (Supplementary Fig. 2a, b). In addition, PARP1 co-expressing genes from normal tissues are not enriched in C2H2-type zinc finger genes, suggesting that unlike PARP2, PARP1 is not functionally related with the C2H2-type zinc finger transcription factor (Supplementary Fig. 2d).

In the second example, we examined the co-expressing genes for CCT4. CCT4 encodes for a subunit of the chaperonin-containing TCP1 complex (CCT). The CCT complex assists the folding of newly translated peptides, including actin and tubulin. As expected, CCT4 co-expresses with genes encoding other subunits of the CCT complex in tumor, and this is better detected by GRACE than by the standard method (Fig. 3d, e). Interestingly, CCT4 has a much tighter correlation with ribosomal genes in normal samples in several TCGA cohorts (Fig. 3d–f). While the co-expression of CCT4 and ribosomal proteins might suggest better coupling of peptide folding and translation processes in normal tissue, it is also possible that monomeric CCT4 could engage in functions independently of the CCT oligomer.

Applying GRACE in multi-dimensional data analysis. Cancer cell lines are widely used as in vitro models for biological research. Many cancer cell lines have been extensively profiled for their copy number, transcriptome, proteome, metabolome and drug sensitivity, among others. Because SCNA also cause a distortion of RNA levels in these cancer cell lines from what would normally occur for a diploid cell in the same biological context, we exploited the use of residuals (RES) from regressing RNA expression levels on copy number values to identify genes with copy number-adjusted RNA expression correlated with orthogonial molecular features. We then compared them with the standard approach using unadjusted RNA expression.

In the first example, we looked at the top 100 genes with RES or RNA levels correlated with the protein levels of Myc measured by reverse phase protein array (RPPA) in NCI-60 cell lines. Myc is a frequently amplified oncogene located on chromosome 8q. One of its critical roles in tumorigenesis is to broadly regulate transcription, in particular the ribosome biogenesis program. We found a higher prevalence of genes belonging to a previously characterized Myc target gene set or a ribosome gene set from the top 100 Myc RPPA correlated genes by RES than by RNA
Fig. 3 Examples of co-expressing genes differing in tumor vs. normal tissues. a, Top 10 PARP2 co-expressing genes from tumor (by standard method or GRACE) or normal tissue samples. Top PARP2 co-expressing genes by standard method in tumors are dominated by genes from 14q, whereas top PARP2 co-expressing genes by GRACE in tumors contain many known cell cycle genes (genes marked in orange are from gene set “REACTOME_CELL_CYCLE” pertinent to the known function of PARP2 in DNA repair). Top PARP2 co-expressing genes from normal tissues, however, do not contain cell cycle genes (genes marked in magenta belong to gene set “Zinc fingers, C2H2-type”). b–f, Examples of co-expressing genes differing in tumor vs. normal tissues from different TCGA cohorts. Analyses are based on TCGA BRCA data.
This result suggests that using copy number-adjusted RNA expression improves the correlation between transcription factor protein abundance and its transcriptional target levels (Fig. 4a, b). Spearman correlation coefficient at false discovery rate of 0.05 based on the Benjamini & Hochberg method was marked by an orange dashed line.

In the second example, we used the metabolomics data of NCI-60 cell lines and compared genes correlated with the levels of TCA cycle intermediates based on RNA or RES. Genes negatively correlate with drug maximal effect level (\(A_{\text{max}}\)) (\(q<0.001\) by Pearson correlation) and significantly correlated with TCA cycle intermediates. While the results are all significant, for most metabolites, correlation based on RES outperformed RNA (Fig. 4c), suggesting the association between the metabolic status of the cell and the transcriptional program of oxidative phosphorylation is better detected when the noise from SCNA is reduced.

Lastly, we correlated gene expression to drug sensitivity using data from CCLE\(^4\), using RNA or RES. Genes with negative correlation to the maximal effect level of each drug were selected based on a cut-off of adjusted \(p\) value less than 0.001. Overall, correlation with RES identified more significant genes than with RNA (Fig. 4d), and this often resulted in more significant results.
from functional enrichment analysis. For instance, paclitaxel, a drug that targets tubulin to interfere with mitotic spindle assembly, had more significant association with cell cycle-related gene sets based on correlation by RES compared to RNA (Fig. 4e), supporting the notion that cells actively engaged in proliferation could be more susceptible to drugs blocking cell division. This suggests that the transcriptional profile associated with drug sensitivity phenotype is better revealed without influence from SCNA.

For all three sets of orthogonal molecular data, we found that the number of significantly associated genes detect by the RES method was significantly larger than by the RNA method (Supplementary Data 2). One-sided p values from Wilcoxon signed rank test are 7e−4, 4e−9, and 1e−4 for RPPA, metabolome, and drug sensitivity data respectively).

The GRACE web database. To introduce our method to the cancer research community, we have constructed a user-friendly open-access web database named GRACE (https://grace.biohpc.swmed.edu) using data sets from TCGA that cover over 20 disease cohorts. GRACE provides two analysis options, "RNA Copy Number Scatter Plot" and "Co-expression Analysis". In "RNA Copy Number Scatter Plot", users can visualize the relationship between the RNA expression and copy number for a gene of interest in tumor samples of a specific TCGA disease cohort (Supplementary Fig. 4). In "Co-expression Analysis", users can identify co-expressing genes for a specific gene in tumor or normal samples from a specific TCGA disease cohort. The standard method and GRACE are both provided for co-expression analysis of tumor samples. In addition, we also provide follow-up functional enrichment analysis of the co-expressing genes. Users may choose from different gene set databases to run the customized enrichment analysis. From the resulting enriched gene set table, users can select different gene sets to view the localization of genes from the chosen gene set in the co-expressing gene table (Fig. 5).

Discussion
Amplification of oncogenes and deletion of tumor suppressor genes commonly take place in the process of tumorigenesis. However, SCNA are often large-scale events that broadly affect many passenger genes besides the driver genes and thus negatively impact the fitness of the cells with increased replication stress, greater burden on the protein quality-control system, altered metabolism, etc. Measurement of copy number changes in cancer are critical in deciphering the driver events in cancer and explaining the disease phenotypes; in addition, passenger events from SCNA could also be exploited as specific vulnerabilities that could represent alternative therapeutic targets for cancer. But from a standpoint of understanding transcriptional regulation in cancer, SCNA could be a big source of noise. Our work has revealed that teasing apart the contribution of SCNA from transcriptome data could improve the biological interpretation of co-expression analysis as well as the correlation between gene expression data and orthogonal molecular features. The benefits may be more obvious for genes severely affected by CNAs and for genes that are more tightly transcriptionally co-regulated with other genes. Our method can also help to discern the differences in transcriptional co-regulation between tumor and normal tissues. For example, we showed that whereas both PARP1 and PARP2 co-express with cell cycle genes in tumor samples consistent with their known roles in DNA repair, only PARP2 strongly co-expresses with many C2H2-type zinc finger genes in normal tissues. It is possible that while the inhibition of PARP1 by PARP inhibitors may provide therapeutic efficacy in cancer, the simultaneous inhibition of PARP2 may result in toxicity via transcriptional dysregulation in normal tissues.

Our method does have several limitations. First, residuals are calculated by regressing RNA expression levels of a gene on its copy number levels. This simultaneous use of two types of data requires the exclusion of samples or genes available to only one type. Second, both data sets need to be in linear range in order for the linear modeling to be accurate. In some cases like the TCGA
BRCA copy number data, we found that copy number measurements were saturated for some highly amplified genes such as ERBB2, a gene encoding a receptor tyrosine kinase that is highly amplified in breast cancer (Supplementary Fig. 3a). These important genes unfortunately need to be filtered out. Genes with many undetected values also need to be excluded, and this is more common in microarray data, perhaps due to the lower sensitivity compared to RNA-seq.

We also noticed that GRACE did not fully correct the bias toward calling physical neighbors in the co-expression analysis (Fig. 2b, i, Supplementary Fig. 1a, d). This could be due to an over-simplified linear model. In our current study, we use preprocessed data from previous publications. It would be interesting to assess the impact of different data acquisition and processing procedures on this method.

Finally, the dynamics of gene–gene correlation or anti-correlation are much more significant in normal samples compared to tumor samples (Supplementary Fig. 3b, c). Although our method corrects for noise introduced by SCNA, there is little complexity to the tumor transcriptional profile (Supplementary Fig. 3d). There are still many challenges in untangling the dysregulated transcriptome in cancer and understanding the origin of cancer-specific transcriptional changes.

Methods
Gene co-expression analysis adjusted for copy number value. In this study, we propose a practical but powerful method to adjust the effect of copy number on gene expression data in cancer samples. Let $Y_i$ and $X_i$ denote random variables for the gene expression and the copy number variation (CNV) of gene $i$, respectively. Without loss of generality, we assume these variables have zero mean and unit variance. We consider the following regression models describing the relation between two genes $j$ and $h$:

$$Y_j = \beta_{jh} X_j + \alpha_{jh} Y_h + \epsilon_j$$

where $\epsilon_j$ and $\epsilon_h$ are Gaussian noise with variance $\sigma^2_j$ and $\sigma^2_h$, respectively. The regression coefficient $\alpha_{jh}$ describes the association of gene expression between genes $j$ and $h$, adjusting for the CNV effect of gene $j$ and $h$. Our goal is to detect genes with high absolute values of $\alpha_{jh}$ to identify co-expressed genes. However, the identification based on these two coefficients is computationally intensive, as it requires inference on the above regression models for all pairs of genes simultaneously. Instead, our proposed method uses the residual from a simpler regression model $(Y_j - \beta_{jh} X_j)$, and correlations between each pair of residuals are used to detect pairs of genes with high values of $|\alpha_{jh}|$ and $|\alpha_{jh}|$. We did simulation studies to compute the calculated coefficients using our approximation method with the true correlation coefficients conditional on the expression of all other genes. The simulation results show that our approach provides a good approximation of the true correlation values under most scenarios (Please see the simulation studies in Supplementary Note.).

This approximation greatly improves the computation efficiency.

To calculate the correlation of the residuals, we regress the gene expression levels on the copy number values for each gene across different samples and take the residuals for subsequent co-expression analysis. Specifically, let $y_{ij}$ denote the gene expression of gene $j$ in patient $i$, and $x_{ij}$ denote the corresponding copy number value. For each gene $j$, we fit a simple linear regression model: $y_{ij} = b_{0j} + b_{1j} x_{ij} + \epsilon_{ij}$, and determine the fitted values $b_{0j}$ and $b_{1j}$ for the generic coefficients $b_{0j}$ and $b_{1j}$. Finally, the residual $\epsilon_{ij} = y_{ij} - b_{0j} - b_{1j} x_{ij}$ is calculated and used as the copy number-adjusted gene expression value (for gene $j$ and patient $i$), as seen in the following co-expression analysis.

For standard co-expression analysis using the TCGA data, under-expressed genes with 0 values in over 10% of the total samples were filtered out. For GRACE using the TCGA data, we noticed the copy numbers for some highly amplified genes had saturated detection in the GISTIC copy number data. Hence, in addition to removal of the under-expressed genes as above, we also removed genes with saturated copy number values in over 5% of the total samples.

Functional enrichment analysis was performed based on the hypergeometric test. Multiple comparisons correction was performed using the Benjamin–Hochberg procedure with a corrected false discovery rate (FDR) cut-off of 0.05. To find genes associated with drug sensitivity, too many genes were identified at the FDR cut-off of 0.05, so in order to find the most important drug response associated genes, we applied a more stringent cut-off of FDR<0.001.

Spearman rank correlation was used for assessing correlation between gene expression levels. Comparisons made between the standard method and our GRACE method are based on data from the same samples, i.e., samples without an available copy number will also be excluded from the standard analysis, whereas we include all available samples for the standard method in the GRACE web database.

Statistical analysis and data visualization were carried out using R.

Correlation with orthogonal data sets. In the CCLE data, we did not notice any saturated copy numbers. However, under-detection of gene expression was more prevalent, so we filtered out under-expressed genes with 0 values in over 50% of the total samples. For NCI-60 data, we filtered out genes with missing values in copy number data or gene expression data.

Spearman rank correlation was used for assessing correlation between protein levels, metabolite levels and gene expression levels, whereas Pearson correlation was used for correlation between drug sensitivity and gene expression levels.

Implementation of the GRACE web database. From the GRACE web database, users are prompted to input the gene name, choose the type of analysis, pick a TCGA cohort and select their preferred sample and method in a stepwise manner. Based on the input from each step, the options without available data will be disabled from the succeeding steps. The design of the analysis configuration and relational database follows the flow chart shown in Supplementary Fig. 5. The web interface was implemented in Javascript.

Data availability. All the data used in this study are from publically available data sets:

- For TCGA data sets, we downloaded the normalized RNA-seq data processed by RSEM (RNA-Seq by Expectation Maximization) method and copy number data processed by GISTIC2 from GDC Firehose (http://firebrowse.org/) from analysis run ’02 April 2015’.[35]. For CCLE data sets, we downloaded gene-centric RNA-normalized mRNA microarray expression data, DNA copy number data and pharmacological profiling data from the CCLE website (http://www.broadinstitute.org/ccle)36. For NCI-60 data sets, we downloaded the five-platform gene transcript-processed RNA data set and combined aGHT-processed DNA copy number data from the CellMiner database (http://discover.nci.nih.gov/cellminer/loadDownload.do)37, 38, we also downloaded the reverse protein lysate data set and Metabolon metabolomics data set from the NCI DTP Molecular Target program (https://wiki.nci.nih.gov/display/NCIDTDTPData/Molecular+Target+Data).

The workflow we developed in this study allowing users to perform gene–gene co-expression study using TCGA data can be accessed through: https://grace.biohpc.swmed.edu/

Received: 25 October 2016 Accepted: 13 November 2017
Published online: 19 December 2017

References
1. Niehrs, C. & Pollet, N. Synexpression groups in eukaryotes. *Nature* **402**, 483–487 (1999).
2. Tu, B. P., Kudlicki, A., Rowicka, M. & McKnight, S. L. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* **310**, 1152–1158 (2005).
3. Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14**, 1083–1094 (2004).
4. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
5. The Cancer Genome Atlas Research Network et al. *Nat. Genet.* **45**, 1113–1120 (2013).
6. Rhodes, D. R. et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
7. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the BioPortal. *Sci. Signal.* **6**, p1 (2013).
8. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
9. Pollack, J. R. et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* **99**, 12963–12968 (2002).
10. Tsafir, D. et al. Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res.* **66**, 2129–2137 (2006).
11. Ben-David, U. et al. Aneuploidy induces profound changes in gene expression, proliferation and tumorigenicity of human pluripotent stem cells. Nat. Commun. 5, 4825 (2014).

12. Fehrman, R. S. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet. 47, 115–125 (2015).

13. Tang, Y. C. & Amon, A. Gene copy-number alterations: a cost-benefit analysis. Cell 152, 394–405 (2013).

14. Antosiewicz, A. et al. Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. PLoS ONE 8, e53014 (2013).

15. Ding, J. et al. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat. Commun. 6, 8554 (2015).

16. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).

17. Dmitriev, S. E. et al. GTP-independent tRNA delivery to the ribosomal P-site by a novel eukaryotic translation factor. J. Biol. Chem. 285, 26779–26787 (2010).

18. Michalak, P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91, 243–248 (2008).

19. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352 (2012).

20. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. 47, 569–576 (2015).

21. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 43, D1079–D1085 (2015).

22. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005).

23. Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H. & Poirier, G. G. PARP inhibition: PARP1 and beyond. Nat. Rev. Cancer 10, 293–301 (2010).

24. Aire, M. R. et al. Polycomb-mediated DNA damage-dependent poly(ADP-ribose) polymerase. J. Biol. Chem. 274, 17860–17868 (1999).

25. Najafabadi, H. S. et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. 33, 555–562 (2015).

26. Yebenes, H., Mesa, P., Munoz, I. G., Montoya, G. & Valpuesta, J. M. Chaperonins: two rings for folding. Trends Biochem. Sci. 36, 424–432 (2011).

27. Spies, M., Echbarthi, M., Svanstrom, A., Karlsson, R. & Grantham, J. Over-expression analysis of all eight subunits of the molecular chaperone CCT in mammalian cells reveals a novel function for CCTdelta. J. Mol. Biol. 427, 2757–2764 (2015).

28. van Riggelen, J., Yettl, A. &Felsher, D. W. MYC as a regulator of ribosome biogenesis and protein synthesis. Nat. Rev. Cancer 10, 301–309 (2010).

29. Schuhmacher, M. et al. The transcriptional program of a human B cell line in response to Myc. Nucleic Acids Res. 29, 397–406 (2001).

30. Nijhawan, D. et al. Cancer vulnerabilities unveiled by genomic loss. Cell 150, 842–854 (2012).

31. Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenic agreement between two cancer cell line data sets. Nature 528, 84–87 (2015).

32. Broad Institute TCGA Genome Data Analysis Center (Broad Institute of MIT and Harvard, 2015).

33. Shankavaram, U. T. et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genom. 10, 277 (2009).

34. Reinhold, W. C. et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. 72, 3499–3511 (2012).

Acknowledgements
The authors would like to thank the BioHPC team in UT Southwestern Medical Center for help with web deployment and providing computational resources. We also thank Fang Huang from the DeBerardinis lab and Xiaowei Zhan from the QBRC for constructive suggestions for the GRACE web database. Jessie Norris for proofreading the manuscript, Xiaowei Zhan and Tao Wang for critical reading of the manuscript. This study was supported by the American Association for Cancer Research (AACR) Basic Cancer Research Fellowship (15-40-01-CRL) awarded to L.C.; and grants from the National Cancer Institute (CA220449 to R.J.D., CA172211-01 to G.X., P50CA70967 to Y. X. and L.C.) and Cancer Prevention Research Institute of Texas (RP130272 to R.J.D., RP120732 to Y.X.).

Author contributions
G.X., R.J.D., and Y.X. supervised the project. L.C. and G.X. conceived the method. L.C. designed and performed the analyses, interpreted the results and developed the web application with advice from G.X. and Y.X. Y.D. deployed the web application. J.Y., Q.L., and G.X. wrote the mathematical expression of the method. L.C. drafted the article. G.X., Y.X., and R.J.D. critically edited the article.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02181-0.

Competing interests: The authors declare no competing financial interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.