BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 15 2012, pages 2045–2051
doi:10.1093/bioinformatics/bts318

Systems biology

Advance Access publication May 30, 2012

Identification of functional CNV region networks using a CNV-gene mapping algorithm in a genome-wide scale

Chihyun Park[1,†], Jaegyoon Ahn[1,†], Youngmi Yoon[2] and Sanghyun Park[1,*]

[1]Department of Computer Science, Yonsei University, South Korea, Seoul 120-749 and [2]Department of Computer Engineering, Gachon University, Incheon 406-709, South Korea

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Identifying functional relation of copy number variation regions (CNVRs) and gene is an essential process in understanding the impact of genotypic variations on phenotype. There have been many related works, but only a few attempts were made to normal populations.

Results: To analyze the functions of genome-wide CNVRs, we applied a novel correlation measure called Correlation based on Sample Set (CSS) to paired Whole Genome TilePath array and messenger RNA (mRNA) microarray data from 210 HapMap individuals with normal phenotypes and calculated the confident CNVR–gene relationships. Two CNVR nodes form an edge if they regulate a common set of genes, allowing the construction of a global CNVR network. We performed functional enrichment on the common genes that were *trans*-regulated from CNVRs clustered together in our CNVR network. As a result, we observed that most of CNVR clusters in our CNVR network were reported to be involved in some biological processes or cellular functions, while most CNVR clusters from randomly constructed CNVR networks showed no evidence of functional enrichment. Those results imply that CSS is capable of finding related CNVR–gene pairs and CNVR networks that have functional significance.

Availability: http://embio.yonsei.ac.kr/∼Park/cnv_net.php.

Contact: sanghyun@cs.yonsei.ac.kr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on October 31, 2011; revised on May 11, 2012; accepted on May 23, 2012

1 INTRODUCTION

Copy number variation (CNV) is one type of human genomic structural variation and is recognized as a major source of human genetic variability, occupying a larger proportion of the genome than single nucleotide polymorphisms (Levy *et al.*, 2007). Specially, there have been several studies to find the impact of CNV on gene expression phenotypes. Much of the variation in messenger RNA (mRNA) transcript levels may be compensated for by regulatory networks, but understanding how genetic variants affect gene expression has provided an essential framework and model for elucidating the causes of various types of phenotypic variation including diseases.

Research investigations on the impact of CNV on phenotypes are generally association studies using various types of gene expression and genotyped data from specific disease samples versus samples with normal phenotypes. Berger *et al.* (2006) proposed an approach to detect variation patterns by integrating gene expression and copy number data from breast cancer samples. Perry *et al.* (2007) found that copy number of CNVR that harbors salivary amylase gene (AMY1) is correlated positively with salivary amylase protein level. Lee *et al.* (2008) found CNV–gene relationships through analysis on CNV and a set of genes which have similar patterns in breast cancer samples. Menezes *et al.* (2009) analyzed the role of CNV in the development of breast cancer in conjunction with genes in the whole genomic area. Junnila *et al.* (2010) profiled gastric cancer cell lines and found that 256 differentially expressed genes are located in the recurrent region of gains and losses. These results imply that those genes are important gastric cancer genes. Generally, these studies focus on the abnormal copy number variation regions (CNVRs) in specific disease samples. Also, many of these focus on CNVs that harbor genes (Berger *et al.*, 2006; Junnila *et al.*, 2010). The global impact of DNA copy numbers on the transcriptome has not been studied, especially for people with normal phenotypes (Henrichsen *et al.*, 2009).

A representative investigation of the impact of CNVs on the phenotypes of normal people in terms of gene expression was made by Stranger *et al.* (2007a), who inferred relationships between DNA copy numbers and gene expression levels of normal phenotyped 210 HapMap samples, using a linear regression model. However, they focused on the genes that lie in or close to CNVRs (distance to a CNVR $<$ 2 Mbp). Considering that there are many more CNVRs that do not harbor genes, Stranger *et al.* (2007a) detected only a small portion of the relationships between genotype and phenotype.

Recent work by Klijn *et al.* (2010) aimed to identify the collaborating CNVs that are responsible for hematopoietic tumorigenesis. They first found gains and losses that are commonly detected in the majority of hematopoietic tumor samples. Then, they constructed gain or loss networks by functional analysis of the genes that were harbored in gain or loss pairs. CNVRs in the central part of the gain or loss networks harbored the well-studied cancer-related genes. Like Stranger *et al.* (2007a) and many other studies, they also focused on CNVRs that harbor genes or gene regulatory regions. However, it is notable that they observed several CNVRs that contributed to the same phenotypic traits by using the concept of a CNVR network.

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

When a CNVR regulates a gene that is on the CNVR itself or that is located near it, we call this type of CNVR–gene regulation *cis*-regulation. In contrast, when a regulated gene location is independent of the CNVR's location, this type of CNVR–gene regulation is called *trans*-regulation. In research elucidating the impact of CNVR on gene expression phenotypes, restricting CNVR–gene relationships into *cis*-regulations is problematic because there are many CNVRs that have neither genes nor known regulation regions. When these CNVRs are strongly suspected to have some functional role, *trans*-regulation as well as *cis*-regulation between CNVRs and genes should be investigated.

A naïve approach to detecting both *cis*- and *trans*-regulations is to measure the correlation between the estimated copy number values of a CNVR and the expression values of a gene using Pearson's correlation coefficient (PCC) or its variants. However, just measuring correlation might lead to many potentially false positives, because this approach has two problems. First, both Whole Genome TilePath (WGTP) array and gene expression data are very noisy. Second, regulation between a CNV and a gene may be observed only under specific conditions or in samples in which some molecular functions or biological processes are taking place. To reduce the false rates of detection of *cis*- and *trans*-regulation, we developed a novel correlation measure named CSS, which effectively relieves the noise of WGTP and gene expression data by applying k-means clustering and also addresses the second problem by using a set of conditions which have similar intensities or expression patterns. In brief, CSS makes use of two vectors: one from the copy numbers of the CNVR A and the other from the expression values of gene B. CSS uses one additional vector from gene C that is connected to gene B through protein–protein interaction (PPI). The rationale of exploiting PPI is that two interacting proteins have a high possibility of being involved in the same biological function in the PPI network. If the expression values of genes that are densely connected to the inspected gene are also correlated to the copy number values of CNVR, the relationship between the inspected gene and a CNVR is more convincing.

Using the WGTP array data and paired mRNA expression data profiled from 210 HapMap samples from individuals with normal phenotypes (Altshuler *et al.*, 2005), we applied CSS to identify significant CNVR–gene relationships. Rather than considering a single CNVR and a single gene, we considered the possibility that many CNVRs are affecting a set of genes that are involved in the same functions. We found that several CNVRs were related to the common set of genes. Using this common gene information, we constructed the functional CNVR network. Our CNVR network is not cancer specific, as in Klijn *et al.* (2010) rather, it contains information about more fundamental biological processes and cellular functions that take place in people with normal phenotypes.

To validate our CNVR network, we performed functional enrichment on the common genes that were *cis*- and *trans*-regulated with CNVRs clustered together in the network. As a result, we confirmed that most CNVR clusters were reported to be involved in some biological processes or cellular functions in the Gene Ontology (GO) database. In contrast, the functional enrichment tests of most CNVR clusters from the network that were constructed using PCC, as well as those from three types of randomly constructed CNVR networks, failed. These results imply that CSS is capable of finding related CNVR–gene pairs, and that our CNVR network has functional significance.

2 METHODS

To construct a genome-wide functional CNVR network, we first find correlated CNVR and gene pairs from preprocessed WGTP array and mRNA microarray data. Two CNVR nodes that are correlated to a common set of genes comprise an edge of the CNVR network.

2.1 Data description

We downloaded WGTP array data from Redon *et al.* (2006). Data are represented by log2 ratio values from comparative genomic hybridization of all 210 unrelated HapMap individuals (Altshuler *et al.*, 2005) from four populations (60 Utah residents with ancestry from northern and western Europe; 45 Han Chinese from Beijing; 45 Japanese from Tokyo and 60 Yoruba from Ibadan, Nigeria) against a common reference individual, NA10851, on an array comprising 26 574 large insert clones. For analysis, we used 867 CNVRs that were identified by Redon *et al.* (2006); for each CNVR, we averaged the log2 ratio values of probes that lie in this CNVR. CNVRs from sex chromosomes were excluded because of their imbalance in males and females.

Transcript levels had been determined in previous studies (Stranger *et al.*, 2007a, b) using Illumina's commercial whole genome expression array. Data were interrogated in cell lines from the same set of 210 HapMap individuals. Of the 47 294 transcripts that were interrogated, we used 19 664 transcripts that can be mapped to human gene symbol space. The gene expression data can be downloaded from the Gene Expression Omnibus with the submission number series GSE6536.

We also downloaded 194 988 human PPIs from the I2D database on October 2010, which includes known, experimental and predicted PPIs for humans, as well as those of five other organisms. The proteins in these PPIs were mapped into gene symbols using UniPROT. After removing duplicated PPIs and PPIs that contain proteins that are not mapped to a gene symbol, we obtained 108 544 PPIs.

2.2 Finding correlated CNVR–gene pairs

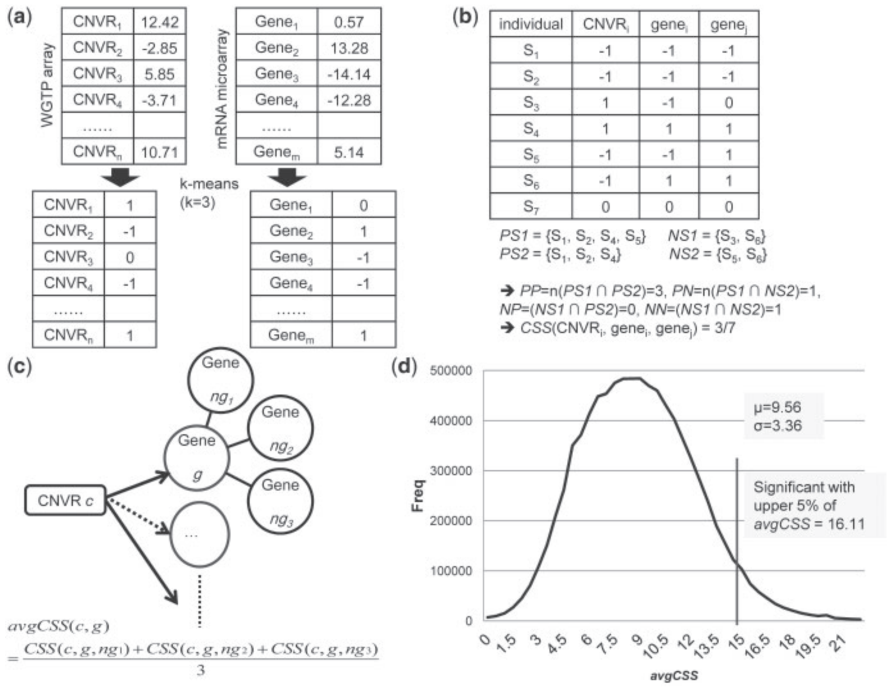
Each individual of the 210 HapMap samples has both WGTP intensities for all probes and gene expression values in log2 ratio, which means that each CNVR and each gene has a vector with 210 values.

To deal with the noise of the WGTP and gene expression data, we preprocess the data using k-means clustering algorithm with $k=3$. First, we average the log2 ratio values of probes that are covered by each CNVR detected by Redon *et al.* (2006) for each individual. We define c_{ij} as the average copy number value of the ith CNVR of the jth individual. Likewise, g_{ij} is defined as the expression value of the ith gene of the jth individual. Vector c_j and g_j are defined as the c_{ij}s and g_{ij}s of all CNVRs and all genes, respectively. Applying k-means to each c_j and g_j, the high values are marked with 1, low values are marked with -1 and intermediate values are marked with 0. Examples of flagging datasets are illustrated in Figure 1a.

Given CNVR c, gene g and genes $\{ng_i\}$ that are connected to g through PPIs (Fig. 1c), PS1, NS1, PS2 and NS2 are defined as follows:

PS1 = {samples that are marked with 1 or -1 in both c and g}.

NS1 = {samples that are marked with 1 in c and -1 in g or that are marked with -1 in c and 1 in g}.



$CNVR_1$	12.42
$CNVR_2$	-2.85
$CNVR_3$	5.85
$CNVR_4$	-3.71
......	
$CNVR_n$	10.71

$Gene_1$	0.57
$Gene_2$	13.28
$Gene_3$	-14.14
$Gene_4$	-12.28
......	
$Gene_m$	5.14

$CNVR_1$	1
$CNVR_2$	-1
$CNVR_3$	0
$CNVR_4$	-1
......	
$CNVR_n$	1

$Gene_1$	0
$Gene_2$	1
$Gene_3$	-1
$Gene_4$	-1
......	
$Gene_m$	1

individual	$CNVR_i$	$gene_i$	$gene_j$
S_1	-1	-1	-1
S_2	-1	-1	-1
S_3	1	-1	0
S_4	1	1	1
S_5	-1	-1	1
S_6	-1	1	1
S_7	0	0	0

PS1 = {S_1, S_2, S_4, S_5} *NS1* = {S_3, S_6}
PS2 = {S_1, S_2, S_4} *NS2* = {S_5, S_6}

➔ *PP*=n(*PS1* ∩ *PS2*)=3, *PN*=n(*PS1* ∩ *NS2*)=1,
NP=(*NS1* ∩ *PS2*)=0, *NN*=(*NS1* ∩ *NS2*)=1
➔ *CSS*($CNVR_i$, $gene_i$, $gene_j$) = 3/7

$$avgCSS(c,g) = \frac{CSS(c,g,ng_1)+CSS(c,g,ng_2)+CSS(c,g,ng_3)}{3}$$

Fig. 1. Calculating the correlation between CNVR c and gene g. (**a**) Each table represents vectors of one individual. (**b**) The table represents the three vectors for a specific CNVR and genes. (**c**) Gene g is connected to ng1, ng2 and ng3 through PPIs. (**d**) avgCSS values of all possible CNVR–gene associations form a normal distribution. The significant avgCSS threshold is 16.11

PS2 = {samples that are marked with 1 or −1 in both c and ng_i}.

NS2 = {samples that are marked with 1 in c and −1 in ng_i, or that are marked with −1 in c and 1 in ng_i}.

And PP, PN, NP and NN are defined as follows:

PP = n(PS1 ∩ PS2).

PN = n(PS1 ∩ NS2).

NP = n(NS1 ∩ PS2).

NN = n(NS1 ∩ NS2).

PP means the number of samples that show positive correlation on c and g, and also positive correlation on c and g_i. Likewise, PN stands for the number of samples that show positive correlation on c and g but negative correlation on c and ng_i. Example for calculating PS1, NS1, PS2, NS2, PP, PN, NP and NN is shown in Figure 1b. PP, PN, NP and NN do not behave in a similar fashion. In some CNV–gene links, PP can be a major measure for calculating avgCSS, and in other links, PN can be major. However, all these measures are equally meaningful, and it is not proper to give weight to some measures. The avgCSS between c and g is defined by the following formula:

$$\text{CSS}\ (c,\ g,\ ng_i) = \frac{\max(\text{PP,PN,NP,NN})}{\text{number of sample}},$$
$$\text{avgCSS}\ (c,g) = \text{average}(\text{CSS}\ (c,\ g,\ ng_i)).$$

Example for calculating CSS and avgCSS is illustrated in Figure 1c.

The CSS value increases as the number of individuals with similar expression patterns in both a CNVR and a gene increases. Therefore, CSS can be used in situations where regulation between a CNVR and a gene may be observed under specific conditions. Moreover, CSS increases the confidence of correlation by considering the set of genes in the second level that are directly interacted with the specific gene, g, by PPI.

As shown in Figure 1d, the frequencies of the avgCSS values of all possible CNVR–gene associations form a normal distribution (tested with Shapiro–Wilk test (Shapiro and Wilk, 1965), P-value > 0.05) with a mean of 9.59 and standard deviation of 3.36. A CNVR–gene association is significant if its avgCSS is greater than the significant avgCSS threshold of 16.11, which is the value for the upper 5% of avgCSSs.

2.3 Construction of CNVR networks

Two CNVRs that share a similar set of genes can be said to be functionally related. The degree of functional relationship should be proportional to the weight of an edge of the CNVR network. Given two CNVRs c_1 and c_2, the weight of edge (c_1, c_2) is calculated using the following formula:

$$\text{weight}\ (c_1, c_2) = \sum_{g_i \in \text{intersections}} \text{harmonic mean}\ (A,\ B) \times \frac{|\text{intersections}|}{|\text{unions}|},$$

where A = avgCSS (c_1, g_i), B = avgCSS (c_2, g_i) and intersections indicate genes that are commonly related to both CNVRs and unions indicate genes that are related to more than one CNVR. We use harmonic mean instead of arithmetic mean because genes that are strongly related to only one CNVR are not good for intersections.

Table 1. Summary and functional analysis of our CNVR network

Scale (%)	No. of connected nodes	No. of edges	Average degree[a]	Average degree[b]	No. of clusters (=a)	No. of enriched clusters (=b)	Success ratio (=b/a, %)	No. of non-connected nodes (=c)	No. of enriched non-connected nodes (=d)	Success ratio (=d/c, %)
0.1	36	294	0.767	16.333	18	18	100.00	731	393	53.76
0.2	55	588	1.533	21.382	32	32	100.00	712	374	52.53
0.5	83	1469	3.831	35.398	31	31	100.00	684	346	50.58
1	134	2938	7.661	43.851	126	126	100.00	633	297	46.92
2	185	5876	15.322	63.524	208	198	95.19	582	251	43.13
5	275	14 689	38.302	106.829	269	189	70.26	492	187	38.01
10	365	29 377	76.602	160.970	583	267	45.80	402	139	34.58
20	469	58 753	153.202	250.546	777	511	65.77	298	75	25.17

[a]Calculated using 833 nodes.
[b]Calculated without unconnected nodes.

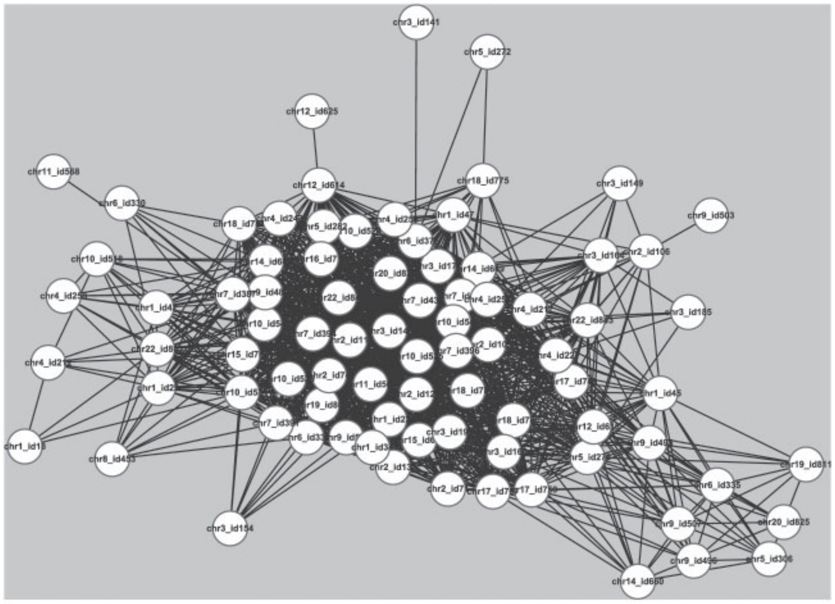


Fig. 2. CNVR network with weight scale = 0.5. Unconnected nodes were excluded when visualized

3 RESULTS

3.1 CNVR networks with different scales

CNVR networks were constructed using 0.1, 0.2, 0.5, 1, 2, 5, 10 and 20% of edges (CNVR pairs) after the edges, we sorted in descending order according to weight. As previously mentioned in Section 2.3, we calculated the weights between all CNVR pairs to construct the complete graph. However, although we calculated the weights for all CNVR pairs, the edges having high weights are more important than the others and we focused only on selected edges. Of 867 CNVRs identified by Redon *et al.* (2006), the number of CNVRs that are matched to the raw WGTP array is 833. Among 833 CNVRs, some CNVRs did not have any genes satisfying avgCSS threshold (see Section 2). Thus, we reduced the selection to 767 CNVRs. Details of each network are presented in Table 1. 'Average degree' in Table 1 was computed using 833 CNVRs. 'Average degree' in Table 1 was calculated using only CNVRs that are connected with each other. The number of clusters means the number of identified communities in the network for each scale. The number of enriched clusters indicates the number of the clusters that are successfully enriched using GO database. The number of non-connected nodes indicates the number of CNVRs that have no connection with other CNVRs. The number of enriched non-connected nodes denotes the number of those CNVRs that are successfully enriched using GO database.

Unlike in a PPI network or genetic interaction (GI) network (Lin *et al.*, 2010), many CNVRs are unconnected. This implies that many CNVRs may function alone, unlike proteins or genes. The average degree and distribution of degree of CNVR networks when considering all CNVRs were similar to those in PPI or GI networks (Gursoy *et al.*, 2008; Kar *et al.*, 2009.

However, when unconnected CNVRs were excluded, the average degree was far greater than that of a PPI or GI network. The fact that CNVRs with more than one connection form very densely connected CNVR networks can be also observed in Figure 2, which was visualized using Cytoscape (Shannon *et al.*, 2003).

Table 2. Functional analysis of the 14th cluster of the 0.5% CNVR network detected by the clustering algorithm (Ahn *et al.*, 2010)

GO term ID	*P*-value	GO term description
GO:0044428	0.0001	Nuclear part
GO:0005515	0.0001	Protein binding
GO:0032991	0.0001	Macromolecular complex
GO:0048522	0.0001	Positive regulation of cellular process
GO:0009987	0.0001	Cellular process
GO:0044260	0.0001	Cellular macromolecule metabolic process

In addition, we performed an experiment with more recent set of CNVRs from Conrad *et al*. (2010). We constructed CNVR networks with our method using 0.1, 0.2, 0.5, 1 and 2% of highly weighted edges. And then we clustered the nodes and performed functional analysis for the networks in the same way we had performed. The result of the functional enrichment is in Table S1 in the Supplementary Material. We could not find clusters from 5, 10 and 20% scaled network because the clustering algorithm we used failed to find clusters from those large-sized networks within the practical time. We can confirm that all CNVR networks were successfully enriched, which shows our method performs well for other CNVR data.

3.2 Functional analysis of CNVR network

For a more detailed analysis of the CNVR networks, we clustered the CNVR nodes using the network clustering algorithm (Ahn *et al.*, 2010), which hierarchically detects the clusters in the network while allowing overlapping nodes (CNVRs) between them. When the weight scale of the network was ⩾20, the number of clusters decreased. The suspected reason is that false-positive CNVR–CNVR edges make the whole CNVR network too dense.

Each CNVR cluster has common associated genes. For each CNVR cluster, genes that were associated with more than (n/2) CNVRs in the cluster of n CNVRs were extracted. We enriched these gene sets with the GO database using the FuncAssociate (Berriz *et al.*, 2003). The FuncAssociate is a web-based tool that reports all GO terms that are enriched with statistical significance ($\alpha = 0.0005$) for a given gene set. If no GO term is reported, the gene set is said to fail to be enriched or validated. We confirmed that the gene sets from all scales were successfully enriched with more than one GO term, indicating that the CNVRs in each CNVR cluster cooperate with each other in some biological processes or cellular functions. One of the example CNVR clusters from the CNVR network of Figure 2 is composed of 13 CNVRs, chr14_669, chr11_565, chr19_808, chr7_396, chr17_749, chr3_195, chr6_376, chr20_822, chr4_217, chr5_282, chr10_524, chr18_775 and chr4_250. These have 1228 common associated genes. Table 2 shows the GO enrichment results of those 1228 genes. We can infer that these CNVRs seem to be involved in protein complex formation processes, which contribute to intracellular transport. We provide the list of CNVRs and CNVR pairs, the list of CNVR clusters for each scale and the list of the associated genes for each CNVR, at http://embio.yonsei.ac.kr/∼Ahn/cnv_net.php.

Although detecting the collaboration of CNVRs is one purpose of constructing a CNVR network, its other purpose is the filteration of false-positive gene–CNVR associations through common gene extraction during the network construction. In other words, genes that are associated by only one CNVR are less convincing than genes that are associated by two or more CNVRs, and these singly associated genes are eliminated.

3.3 Comparison between CNVR networks based on avgCSS and those based on PCC

To validate that our association measure, CSS, was superior in identifying the correlations between CNVR and gene expression, we constructed a CNVR network using PCC (PCC network) and compared the functional enrichment of that and our CNVR network. This was the core process to calculate the association between the CNVRs and genes because we constructed the CNVR network based on the correlations for each CNVR–gene pair. All other experimental conditions were the same as in the previous situations. We applied the same significance level, upper 0.5%, for the threshold of PCC and calculated weights between all pairs of CNVRs. We also constructed eight networks with the same range of scales that we applied to our CNVR network. Unlike our CNVR network in Table 1, the numbers of clusters in the PCC network were larger and the densities of networks with connected nodes were two or four times lower. Also, most of the clusters of the PCC network were not functionally enriched. Table 3 shows the summary and results of functional analysis of PCC network. Unconnected CNVRs were also not functionally enriched. This implies that a PCC network may not include the biological meaning, in contrast to the results of our CNVR network.

3.4 Validation of CSS by random tests

Previously, we successfully validated our CNVR network through a functional enrichment test. If a randomly constructed CNVR network failed to be validated, we claim that our methods produce a CNVR network that has significant biological meaning.

Second, we generated 833 false CNVRs by simulating the distribution of hybridization intensities, N(2.72256E-14, 12) and calculated CNVR–gene associations for all possible pairs of false CNVR. As the log2 ratio of false CNVRs followed the distribution of real data, most of them were neutral. The remaining steps were the same as those of our method. Table 5 shows the results of functional analysis. The numbers of connected nodes and edges were very similar to those in our CNVR network for each scale because we used the same measure, CSS, to identify associations and the remaining steps were the same. However, the number of clusters and the ratio of enriched clusters definitely differed on all scales from those of our network. This experimental result indicated that the network constructed using neutral regions instead of copy numbered areas did not reflect biological function.

From Tables 4 and 5, we can say that CNVR networks constructed using random CNVR–gene selection or false CNVRs failed to be validated while CNVR networks constructed by our approach were validated. This implies that CSS is able to detect significant CNVR–gene associations, and our weight scheme is capable of constructing the correct CNVR network.

To further show the significance of CSS, we performed the permutation test. First, we divided CNVR–gene pairs into two groups, A and B, based on avgCSS threshold. Pairs above threshold are labeled as group A and pairs below threshold are labeled as

Table 3. Summary and functional analysis of a CNVR network constructed using PCC instead of CSS

Scale (%)	No. of connected nodes	No. of edges	Average degree[a]	Average degree[b]	No. of clusters (=a)	No. of enriched clusters (=b)	Success ratio (=b/a, %)	No. of non-connected nodes (=c)	No. of enriched non-connected nodes (=d)	Success ratio (=d/c, %)
0.1	165	347	0.833	4.206	134	0	0.00	668	0	0.00
0.2	234	694	1.666	5.932	215	1	0.47	599	0	0.00
0.5	317	1733	4.161	10.934	432	0	0.00	516	0	0.00
1	422	3466	8.322	16.427	487	0	0.00	411	0	0.00
2	526	6931	16.641	26.354	660	0	0.00	307	0	0.00
5	680	17 327	41.601	50.962	844	1	0.12	153	0	0.00
10	801	34 653	83.200	86.524	1016	0	0.00	32	0	0.00
20	832	69 306	166.401	166.601	1184	0	0.00	1	0	0.00

[a]Calculated using 833 nodes.
[b]Calculated without unconnected nodes.

Table 4. Summary and functional analysis of CNVR networks generated using random CNVR–gene selection

Scale (%)	No. of connected nodes	No. of edges	Average degree[a]	Average degree[b]	No. of clusters (=a)	No. of enriched clusters (=b)	Success ratio (=b/a, %)	No. of non-connected nodes (=c)	No. of enriched non-connected nodes (=d)	Success ratio (=d/c, %)
0.1	0	0	0.000	0.000	N/A	N/A	N/A	833	1	0.12
0.2	0	0	0.000	0.000	N/A	N/A	N/A	833	1	0.12
0.5	0	0	0.000	0.000	N/A	N/A	N/A	833	1	0.12
1	0	0	0.000	0.000	N/A	N/A	N/A	833	1	0.12
2	0	0	0.000	0.000	N/A	N/A	N/A	833	1	0.12
5	4	5	0.012	2.500	1	0	0.00	829	1	0.12
10	13	21	0.050	3.231	7	0	0.00	820	1	0.12
20	29	106	0.255	7.310	11	0	0.00	804	1	0.12

[a]Calculated using 833 nodes.
[b]Calculated without unconnected nodes.

Table 5. Summary and functional analysis of CNVR networks constructed with false CNVRs

Scale (%)	No. of connected nodes	No. of edges	Average degree[a]	Average degree[b]	No. of clusters (=a)	No. of enriched clusters (=b)	Success ratio (=b/a, %)	No. of non-connected nodes (=c)	No. of enriched non-connected nodes (=d)	Success ratio (=d/c, %)
0.1	37	295	0.708	15.946	10	0	0.00	768	106	13.80
0.2	53	589	1.414	22.226	26	2	7.69	752	105	13.96
0.5	82	1470	3.529	35.854	25	2	8.00	726	102	14.05
1	138	2939	7.056	42.594	125	18	14.40	673	97	14.41
2	187	5877	14.110	62.856	317	46	14.51	627	90	14.35
5	277	14 690	35.270	106.065	159	7	4.40	540	82	15.19
10	363	29 378	70.535	161.862	212	40	18.87	456	66	14.47
20	465	58 754	141.066	252.705	416	57	13.70	357	54	15.13

[a]Calculated using 833 nodes.
[b]Calculated without unconnected nodes.

group B. Second, we reconstructed groups A' and B' using randomly chosen pairs from A and B, and calculate a difference between the mean of avgCSS values in A' and one in B'. The second procedure was iterated 999 999 times, and we display the distribution of the statistics in Supplementary Figure S1. H_0 is that our CNVR–gene pairs are not significant. P-value for difference between means of A and B (statics for observed data) is 100E-06. Therefore, H_0 was rejected.

We also tested the robustness of our method by comparing number of enriched clusters and enriched non-connected nodes generated with half of the transcript data, and less. Supplementary Figure S2 shows the proportions of enriched non-connected nodes with same scales. Approximately 64.18% of the enriched clusters were detected even when only 10% of the samples were used with scale of 0.1%. It performed comparably well with 25% of samples. Supplementary Figure S3 shows the proportions of enriched non-connected nodes with same scales. Fifty percent of samples approximately give comparable success ratio to whole samples in finding functionally enriched clusters and nodes. These two tests indicate that our method is robust to the number of samples used.

4 DISCUSSIONS

We showed that the ratio of functionally enriched CNVR clusters of the CNVR network constructed using CSS was far greater than that of PCC, in Section 3. Two major different characteristics of CSS and PCC appear to make the difference of performance. First, correlation is calculated using some group of samples that have similar patterns in CSS, while PCC is calculated using all the samples of the dataset. As there is no guarantee that functional relationship is active in all conditions or samples, our approach seems to have definite strength.

Second, CSS measures the correlation between a CNVR and a gene by integrating the PPIs, while PCC needs no PPI. Given CNVR *c* and gene *g*, correlation between *c* and *g* is not reliable because both arrayCGH log ratio values and gene expression values are noisy, and not all samples show the significant correlation between *c* and *g*, as we have explained above. To correct the correlation value, we introduce other genes that are believed to be functionally linked, in other words, linked through PPI to *g*. Correlations between *c* and each of other genes can correct the significance level of the correlation between *c* and *g*.

Of course, integrating the PPIs has obvious weakness. If the PPIs that are relevant to some biological functions in the protein network are sparse, CNVR–gene relations that are related to those functions would be incorrect. Moreover, PPI datasets have been known to suffer from many false positives. However, PPIs datasets are becoming more complete and accurate, and we can expect that the weakness of CSS can be overcome to some degree.

In addition, we analyzed more detailed functionality of some CNVR clusters we found, referring to the literature. The analysis can be found in Supplementary File.

5 CONCLUSIONS

The main contribution of our study is that we constructed functional CNVR networks for people with normal phenotypes using a novel correlation measure called CSS and the weighting scheme. Notably, our CNVR networks also contain CNVRs that are close to or that harbor some genes. The *trans*-regulations between those CNVRs and the genes allow our CNVR network to function on a genome-wide basis. We showed that our CNVR networks have biological meaning, which means that we were able to identify several CNVRs that commonly contribute to some biological functions that are not dependent on specific conditions or diseases. We expect our methods can also be successfully applied to the construction of the disease-specific CNVR networks and that our methods will contribute to various studies regarding *trans*-regulation.

ACKNOWLEDGEMENT

We appreciate Prof. Sung-Bong Yang of Yonsei University for his careful advice.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0002887)

Conflict of Interest: none declared.

REFERENCES

Ahn,Y.Y. *et al.* (2010) Link communities reveal multiscale complexity in networks. *Nature*, **466**, 761–764.

Altshuler,D. *et al.* (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Berger,J.A. *et al.* (2006) Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 2–16.

Berriz,G.F. *et al.* (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Conrad,D.F. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Gursoy,A. *et al.* (2008) Topological properties of protein interaction networks from a structural perspective. *Biochem. Soc. Trans.*, **36**, 1398–1403.

Henrichsen,C.N. *et al.* (2009) Copy number variants, diseases and gene expression. *Hum. Mol. Genet.*, **18**, R1–R8.

Junnila,S. *et al.* (2010) Genome-wide gene copy number and expression analysis of primary gastric tumors and gastric cancer cell lines. *BMC Cancer*, **10**, 73.

Kar,G. *et al.* (2009) Human cancer protein–protein interaction network: a structural perspective. *PLoS Comput. Biol.*, **5**, e1000601.

Klijn,C. *et al.* (2010) Identification of networks of co-occurring, tumor-related DNA copy number changes using a genome-wide scoring approach. *PLoS Comput. Biol.*, **6**, e1000631.

Lee,H. *et al.* (2008) Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes. *Bioinformatics*, **24**, 889–896.

Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, 2113–2144.

Lin,A. *et al.* (2010) A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res.*, 20, 1122–1132.

Menezes,R.X. *et al.* (2009) Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*, **10**, 203.

Perry,G.H. *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.*, **39**, 1256–1260.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shapiro,S.S. and Wilk,M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Stranger,B.E. *et al.* (2007a) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Stranger,B.E. *et al.* (2007b) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.