Short Communication **Open Access**

Measuring Inequalities in Gene Co-expression Networks of HIV-1 Infection Using the Lorenz Curve and Gini Coefficient

Chuang Ma[1,2], Sheng-He Huang[1,3]* and Yanhong Zhou[3]

[1]Saban Research Institute of Children's Hospital Los Angeles and the University of Southern California, Los Angeles, USA
[2]School of Plant Sciences, University of Arizona, Tucson, AZ, USA
[3]Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

The Gini methodology is a family of mathematical models that describe various relations in or between variables [1,2]. The basic concept of Gini methodology is the Gini coefficient (also known as Gini index, or Gini ratio), which measures the inequality of a distribution (e.g., income) with values ranged from 0 (complete equality) to 1 (complete inequality), has been popularly used in economics for quantifying the income inequality in a country [3,4]. Due to the superiority of analyzing data with normalized and non-normalized distribution [2], Gini coefficient and the derived statistical algorithms have been extended to apply in disciplines as diverse as social science, chemistry and engineering. Recently, the Gini methodology has also been introduced to biology for inferring transcription regulation relationships from gene expression data [5], and for exploring the symbiosis and pathogenesis of human immunodeficiency virus type 1 (HIV-1) infection [6].

HIV-1 is a virus that can cause acquired immunodeficiency syndrome (AIDS), leading to thousands of death per year in the world due to the lack of effective vaccines and cure. As one of powerful systems biology approaches, gene co-expression networks (GCNs) have been recently applied to investigate the molecular mechanisms of HIV-1 infection by organizing genes into a network, in which two genes with similar expression patterns are connected by an edge [6-8]. An in-depth statistical analysis of HIV-related network properties will be helpful to discover new biomarkers and signatures of HIV-1 infection. Here we applied the Gini methodology to explore inequalities in GCNs constructed with 943 genes differentially expressed in human lymphatic tissues of uninfected subjects and infected patients at different stages of HIV-1 infection (the acute, the asymptomatic, and the AIDS stages). More details about the microarray data generation and normalization, and the selection of differentially expressed genes can be found in Xu et al. [9]. To construct GCNs, the similarities of expression patterns between two genes were measured with Pearson correlation coefficient (PCC). Two genes were connected in the GCNs if the significance level (*p*-value) of PCC is lower than 0.05. The *p*-values were estimated with permutation method by shuffling gene expression data in the microarray dataset.

Application of Gini Co-efficient to Quantify the Connectivity In-equality (CI) in the GCN

Despite the connectivity distributions give some insights about how genes are connected in the GCN (Figure 1A), they fails to quantify the characterization of the connectivity in the whole network, leading to the difficulty of comparing two GCNs constructed for different biological conditions. Here the connectivity inequality (CI) is introduced to consider the distribution of connectivity of genes in the whole network with the Gini coefficient algorithm. The CI can be graphically represented with the Lorenz curve, which is a two-dimensional plot of the cumulative fraction of the number of genes in the network versus the cumulative fraction $L(p)$ of total connectivity from these genes. The Lorenz curve more close to the diagonal line indicate that genes in the network are more equally connected. The Gini coefficient is equal to

one minus twice the area under the Lorenz curve, and can be computed with the formula [10]:

$$CI = \frac{\sum_{i=1}^{n}(2i-n-1)X_{(i)}}{(n-1)\sum_{i=1}^{n}X_{(i)}}$$, where n is the number of genes in the

network, $X_{(i)}$ is the ith value of connectivity sorted in increasing order, $0 \leq X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. We observed that the Lorenz curve from the GCN at the AIDS stage is markedly deviated from the diagonal line that those from the other three GCNs (Figure 1B). At the same time, the Gini coefficient from the GCN at the AIDS stage is much higher than those from the other three GCNs. These results indicate that dramatic changes of transcriptional regulation at the last stage of HIV infection.

Application of Gini Coefficient to Estimate the Contribution of Positive and Negative Connectivity to the Connectivity Inequality (CI)

In the GCN, the connectivity of a gene is composed with positive and negative connectivity, which present the connection to other genes with positive and negative PCC values, respectively. The contribution of positive and negative connectivity to the overall inequality of connectivity in the network is defined based on the decomposition of Gini coefficient (1):
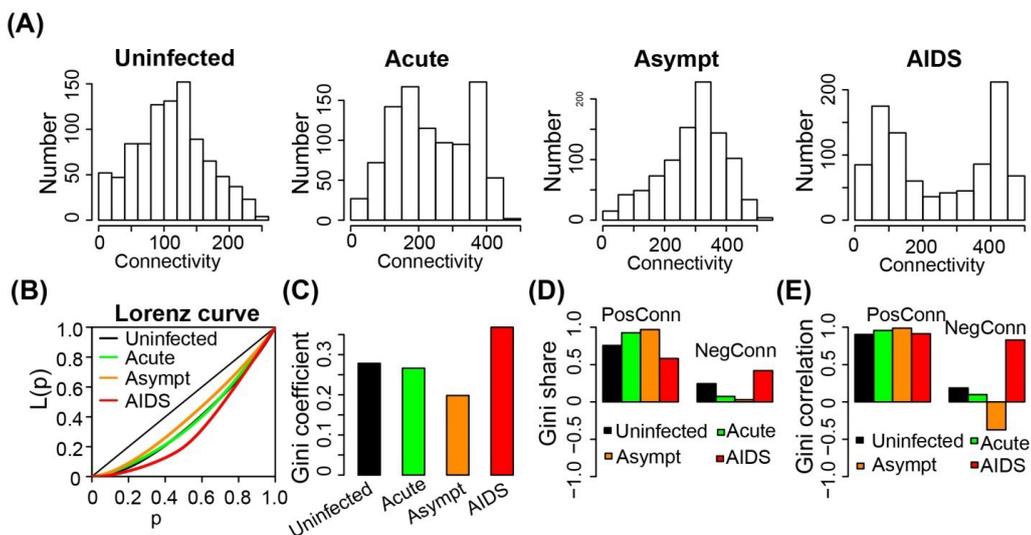
$CI = S_p \tau(X_p, X)CI_p + S_n \tau(X_n, X)CI_n$, where CI and CI_n are the inequality of positive and negative connectivity, respectively. S_p and $S_n (S_n = 1 - S_p)$ are two Gini share measures represent the percentages of positive and negative connection in the whole network, respectively. $\tau(X_p, X)$ and $\tau(X_n, X)$ are two Gini correlation coefficients ranged from -1 to 1, indicating the contribution of positive and negative connectivity to the CI, respectively. As shown in Figure 1D, Gini share of positive connectivity in four networks are remarkably higher than that of negative connectivity, indicating the positive regulation is the dominant relation in the network for uninfected subjects and patients at different stages of HIV-1 infection. Interestingly, the positive regulations were enhanced at the first two stages of HIV-1 infection. In contrast, The negative regulations at the AIDS stage were enhanced. From the HIV uninfected to the AIDS stage, the Gini correlation of

Figure 1: The connectivity in equality in the geneco-expression networks (GCNs) of HIV-1 infection.
(A) The connectivity distributions of GCNs for uninfected subjects and patients at different stages of HIV-1 infection.
(B) The Lorenz curves of connectivity distributionin four GCNs. prepresents the cumulative fraction of the number of genes in the network, L(p) denotes the cumulative fraction of connectivity from these genes.
(C) Gini coefficient of connectivity infour GCNs.
(D) Gini share of positive and negative connectivity in four GCNs.
(E) Gini correlation of positive and negative connectivity infour GCNs.
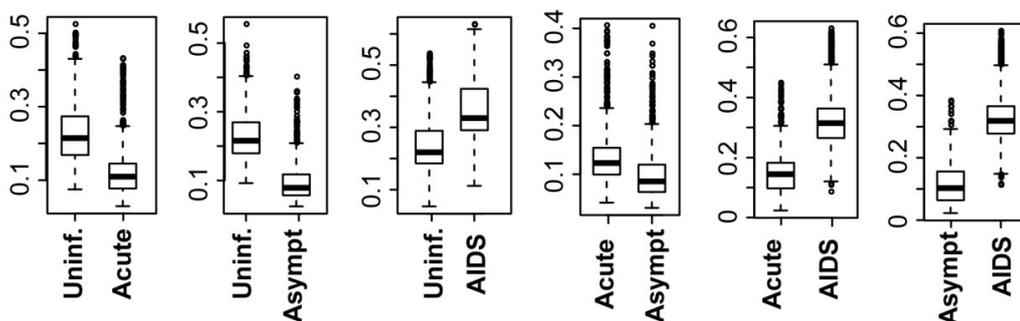
Figure 2: The inequality of edge weights in geneco-expression networks of HIV-1 infection."Uninf." represent networks for uninfected subjects.

negative connectivity is changed more significantly than that of positive connectivity (Figure 1E), indicating that positive and negative co-expression associations might play different roles in the pathogenesis of HIV infection.

## Application of Gini Coefficient to Measure the Inequality of Edge Weights in GCNs

Besides the connectivity, the edge weights (i.e., correlation values) in GCNs were also changed during the HIV-1 infection. For a given gene i, the changes in the correlation strengths can be calculated using the differential co-expression (dC) measure with the formula [12]: $dC(i) = \dfrac{\sqrt{\sum_{j=1}^{m}(C_{ij}^{1}-C_{ij}^{0})}}{\sqrt{m}}$ where gene i connects m genes in two networks, C_{ij}^{1} and C_{ij}^{0} represent the correlation values between gene i and j in two networks, respectively. In this study, we observed that there were differences in the inequality of edge weights between GCNs of HIV-1 infection (Figure 2). At the acute and asymptomatic

stages of HIV-1 infection, the edge weights are more equal than those in network for uninfected subjects. However, the edge weights become dramatically unequal in network for patients at the AIDS stage (Figure 2). On this basis, a novel measure "delta Gini" was introduced to consider the differences in the inequality of edge weights between two networks. Although the delta Gini and dC were significantly correlated in most network comparisons (except AIDS vs. Uninfected) (Figure 3), the delta Gini provided additional information about the changes of edge weights between two networks. First, the delta Gini is ranged from -1 to 1, with positive value indicating the inequality of edge weights is increased and negative values indicating the inequality of edge weights is decreased. Second, the delta Gini is valuable to identify candidate biomarkers of HIV-1 with low rank of dC values. For instance, MRC1 is a mannose receptor interacting with several HIV proteins to promote viral spread [13-15], and has a delta Gini value of -0.44 (rank=2) and a dC value of 0.96 (rank=173) while comparing networks constructed for patients at the AIDS stage and for uninfected subjects. Similarly, PPFIBP1, which plays roles in HIV-1 replication, also has a high rank of delta Gini (value=-0.42; rank=3) but a low
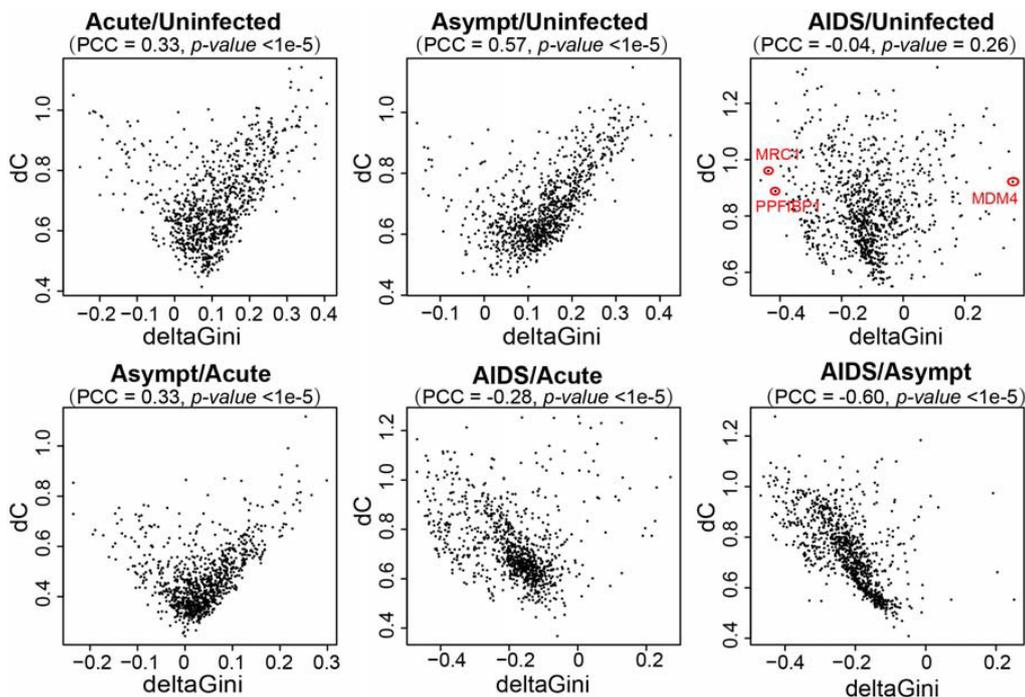
Figure 3: The dot plots of delta Gin iand dC for comparing GCNs of HIV-1 infection. Each dot represents a gene."PCC" denotes the Pearson correlation coefficient. The p-value of PCC was calculated using the "cor.test" function in R programming language.

No.	Uninfected/Acute	Uninfected/Asympt	Uninfected/AIDS	Acute/Asympt	Acute/AIDS	Asympt/AIDS
1	NM_002604	NM_006275	BE856807	Z00008	N45309	AU118882
2	J04162	BC025250	NM_002438	AF234255	NM_002612	U52913
3	AW169973	AI743534	NM_003622	NM_014668	AL554245	NM_002612
4	AB040957	AW169973	AL574194	BE886225	AI803181	NM_003622
5	AI760944	AK055572	BF592034	AW294722	S67238	AK095698
6	AI650364	NM_002759	NM_016164	L21961	AK095698	AK000776
7	NM_006482	AI650364	AI655611	BI598831	AA541622	NM_016164
8	AK000776	AU146963	BC038422	AL049337	AW027333	AL574194
9	NM_000570	AI554909	NM_017631	BF673779	BF340228	AA053711
10	AI148006	AI926479	AI803181	U52914	M_014710	U52914

Table 1: List of top 10 genes with largest changes of edge weights between two compared networks.

rank of dC (value=0.89; rank=283). MDM4 is another representative example showing a positive and high-ranked delta Gini (value=0.36; rank=22), but a low-ranked dC (value=0.92; rank=235). This gene was recently demonstrated to be a direct calpain substrate playing roles in the HIV-induced neuronal damage [16]. The detailed values of delta Gini and dC for all comparisons of GCNs of HIV-1 infection were listed in Supplemental Table 1.

These results indicate that Gini algorithm would be a complementary approach to dC for comparing the differences between two GCNs.

References

1. Schechtman E, Yitzhaki S (1999) On the proper bounds of the Gini correlation. Economics Letters 63: 133-138.

2. Yitzhaki, S. (2003). Gini's mean difference: a superior measure of variability for non-normal distributions. METRON-International Journal of Statistics LXI: 32.

3. Mikkelson GM, Gonzalez A, Peterson GD (2007) Economic inequality predicts biodiversity loss. PLoS One 2: e444.

4. Mussard, S, SavardL (2012) The Gini multi-decomposition and the role of Gini'stransvariation: application to partial trade liberalization in the Philippines. Applied Economics 44: 1235-1249.

5. Ma C, Wang XF (2012) Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis. Plant Physiol 160: 192-203.

6. Ma C, Xin MM, Feldmann KA, Wang XF (2014) Machine learning-baseddifferential network analysis: a study of stress-responsive transcriptomes in Arabidopsis thaliana. Plant Cell.

7. Ma C, Zhou Y, Huang SH (2011) Inequalities and duality in gene coexpression networks of HIV-1 infection revealed by the combination of the double-connectivity approach and the Gini's method. J Biomed Biotechnol 2011: 926407.

8. Levine AJ, Horvath S, Miller EN, Singer EJ, Shapshak P et al. (2013) Transcriptome analysis of HIV-infected peripheral blood monocytes: Gene transcripts and networks associated with neurocognitive functioning. J Neuroimmunol 265: 96-105.

9. Xu WW, Han MJ, Chen D, Chen L, Guo Y, et al. (2013) Genome-wide search for the genes accountable for the induced resistance to HIV-1 infection in activated CD4+ T cells: apparent transcriptional signatures, co-expression networks and possible cellular processes. BMC Med Genomics 6: 15.

10. Li Q, Smith AJ, Schacker TW, Carlis JV, Duan L, et al. (2009) Microarray

analysis of lymphatic tissue reveals stage-specific, gene expression signatures in HIV-1 infection. J Immunol 183: 1975-1982.

11. Glasser GJ (1962) Variance formulas for the mean difference and coefficient of concentration. Journal of the American Statistical Association 57: 648-654.

12. Liu BH, Yu H, Tu K, Li C, Li YX, et al. (2010) DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics 26: 2637-2638.

13. Lai J, Bernhard OK, Turville SG, Harman AN, Wilkinson J, et al. (2009) Oligomerization of the macrophage mannose receptor enhances gp120-mediated binding of HIV-1. J Biol Chem 284: 11027-11038.

14. Vigerust DJ, Egan BS, Shepherd VL (2005) HIV-1 Nef mediates post-translational down-regulation and redistribution of the mannose receptor. J Leukoc Biol 77: 522-534.

15. Caldwell RL, Lane KB, Shepherd VL (2006) HIV-1 Tat interaction with cyclin T1 represses mannose receptor and the bone morphogenetic protein receptor-2 transcription. Arch Biochem Biophys 449: 27-33.

16. Colacurcio DJ, Yeager A, Kolson DL, Jordan-Sciutto KL, Akay C (2013) Calpain-mediated degradation of MDMx/MDM4 contributes to HIV-induced neuronal damage. Mol Cell Neurosci 57: 54-62.