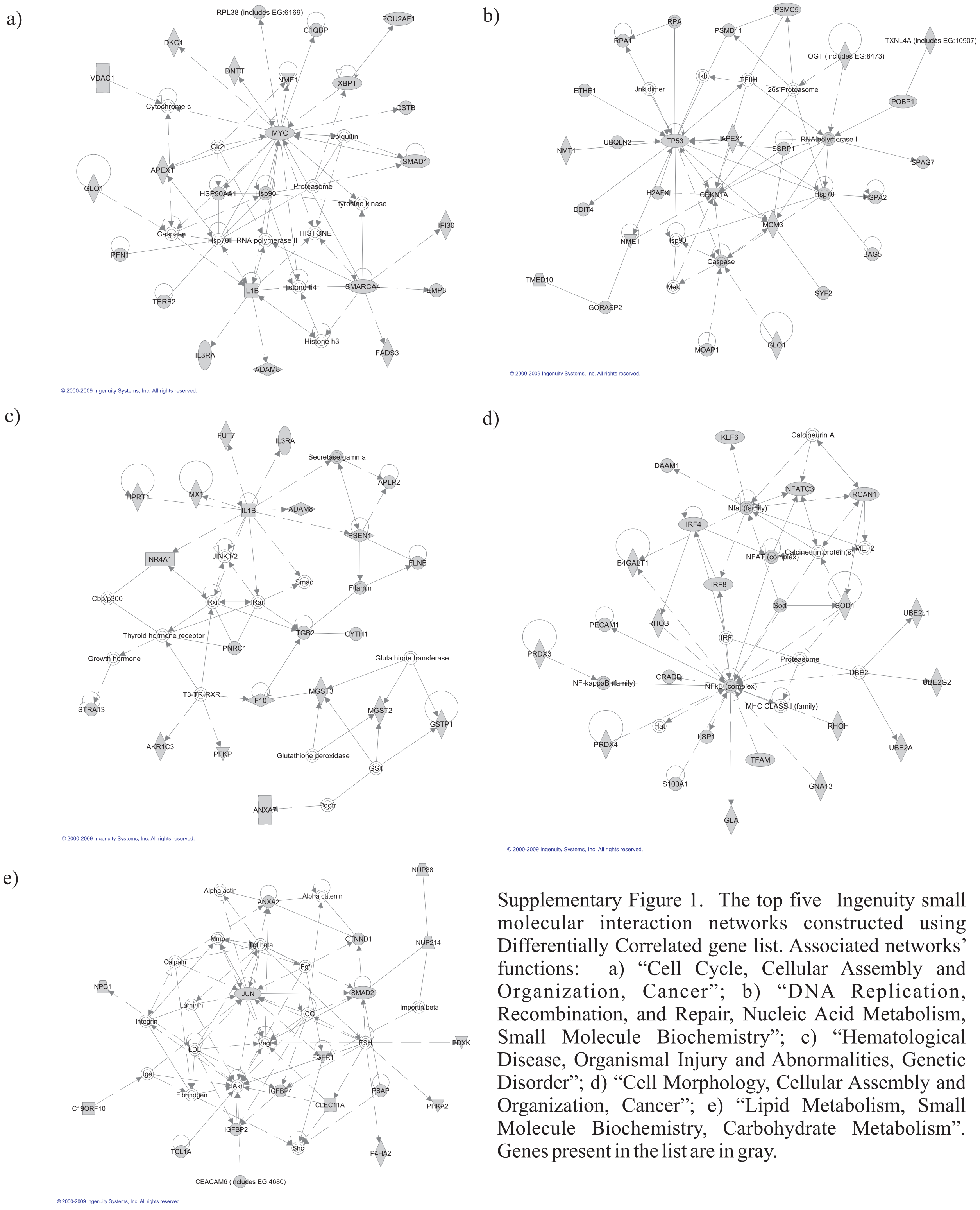
A New Gene Selection Procedure Based on the Covariance Distance: Supplementary Materials

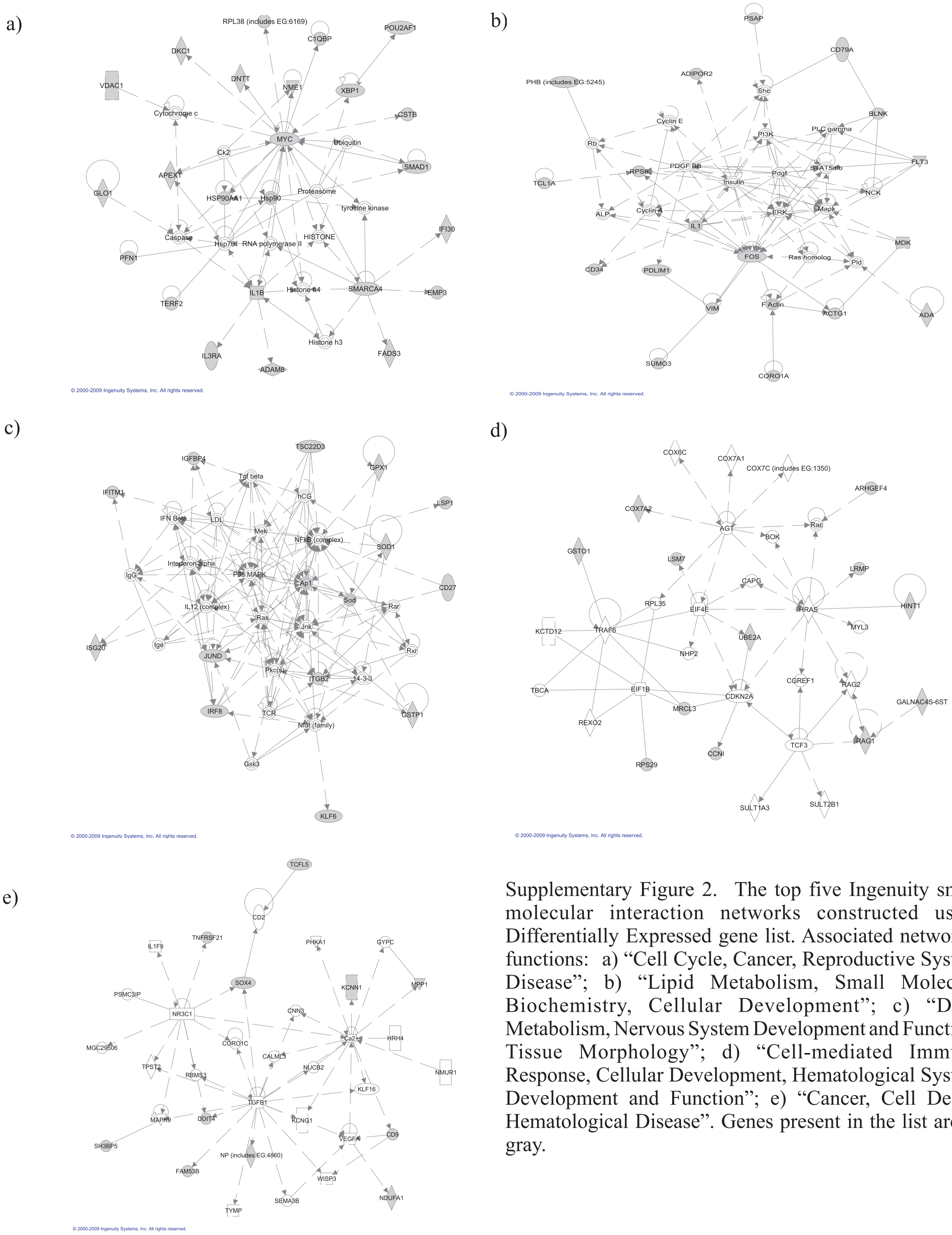
Rui Hu Xing Qiu Galina Glazko

September 24, 2009

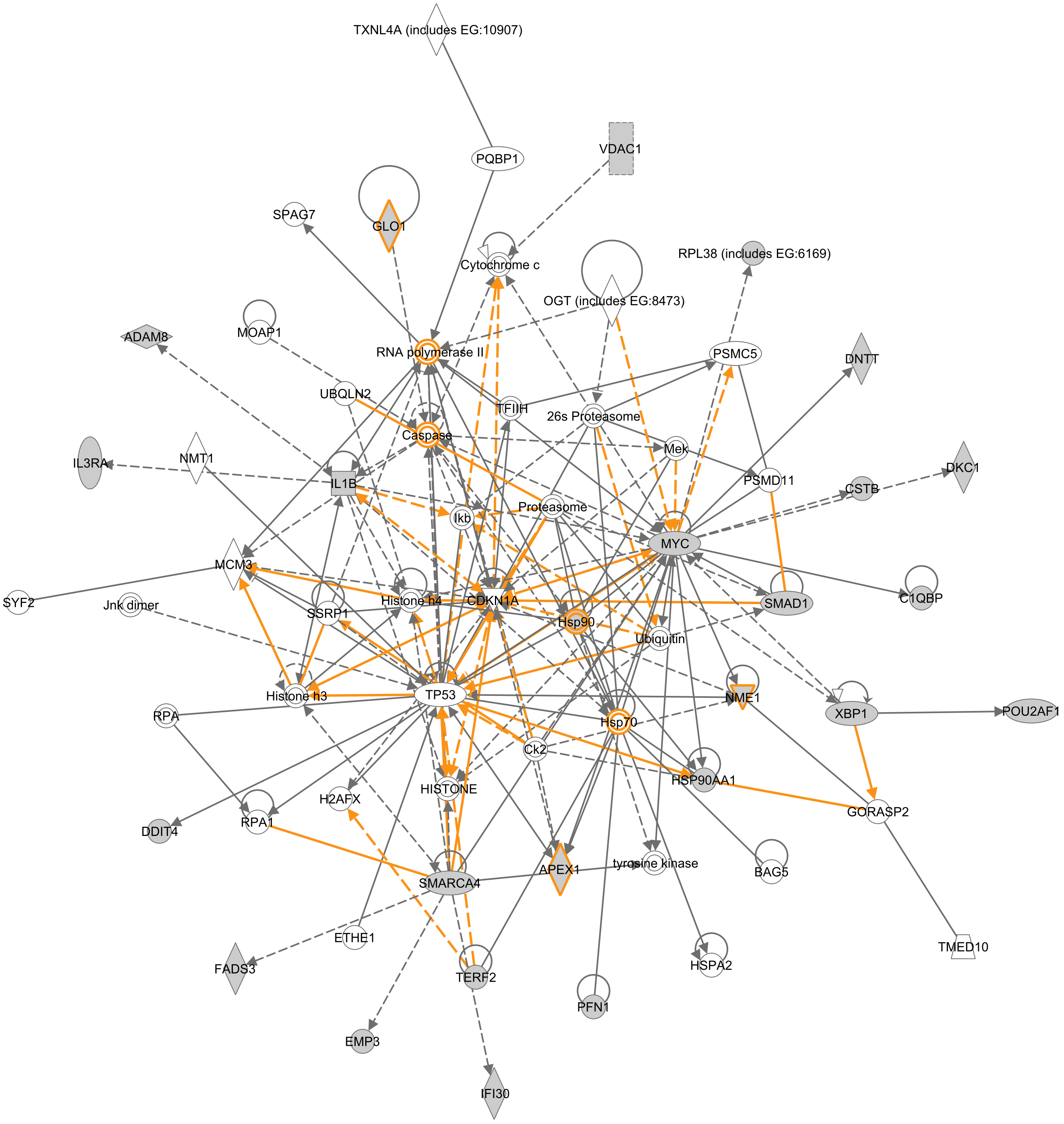
1 Tables and Figures



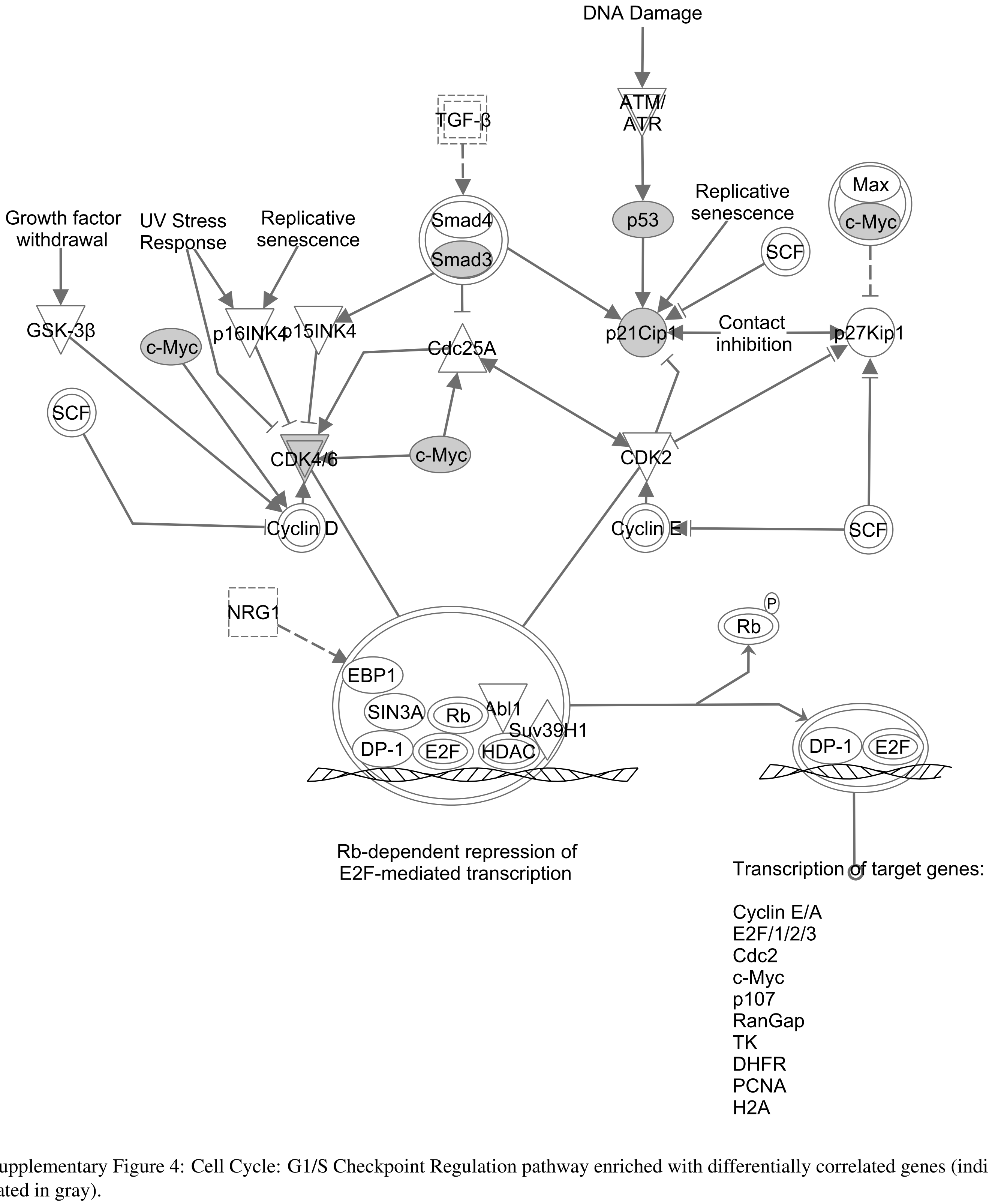
Supplementary Figure 1. The top five Ingenuity small molecular interaction networks constructed using Differentially Correlated gene list. Associated networks' functions: a) "Cell Cycle, Cellular Assembly and Organization, Cancer"; b) "DNA Replication, Recombination, and Repair, Nucleic Acid Metabolism, Small Molecule Biochemistry"; c) "Hematological Disease, Organismal Injury and Abnormalities, Genetic Disorder"; d) "Cell Morphology, Cellular Assembly and Organization, Cancer"; e) "Lipid Metabolism, Small Molecule Biochemistry, Carbohydrate Metabolism". Genes present in the list are in gray.



Supplementary Figure 2. The top five Ingenuity small molecular interaction networks constructed using Differentially Expressed gene list. Associated networks' functions: a) "Cell Cycle, Cancer, Reproductive System Disease"; b) "Lipid Metabolism, Small Molecule Biochemistry, Cellular Development"; c) "Drug Metabolism, Nervous System Development and Function, Tissue Morphology"; d) "Cell-mediated Immune Response, Cellular Development, Hematological System Development and Function"; e) "Cancer, Cell Death, Hematological Disease". Genes present in the list are in gray.



Supplementary Figure 3: Two Ingenuity networks merged together. One network is constructed using differentially expressed (Cell Cycle, Cancer, Reproductive System Disease) genes (presented on Sup. Fig. 2a). Another network is constructed using differentially correlated (DNA Replication, Recombination, and Repair, Nucleic Acid Metabolism, Small Molecule Biochemistry) genes (presented on Sup. Fig. 1b). Orange lines mark known molecular interactions which become visible only after networks integration.



Supplementary Figure 4: Cell Cycle: G1/S Checkpoint Regulation pathway enriched with differentially correlated genes (indicated in gray).

Differentially correlated genes		Differentially expressed genes	
Associated Network Functions	Score	Associated Network Functions	Score
Cell Cycle, Cellular Assembly and Organization, Cancer	49	Cell Cycle, Cancer, Reproductive System Disease	47
DNA Replication, Recombination, and Repair, Nucleic Acid Metabolism, Small Molecule Biochemistry	39	Lipid Metabolism, Small Molecule Biochemistry, Cellular Development	33
Hematological Disease, Organismal Injury and Abnormalities, Genetic Disorder	32	Drug Metabolism, Nervous System Development and Function, Tissue Morphology	21
Cell Morphology, Cellular Assembly and Organization, Cancer	31	Cell-mediated Immune Response, Cellular Development, Hematological System Development and Function	20
Lipid Metabolism, Small Molecule Biochemistry, Carbohydrate Metabolism	26	Cancer, Cell Death, Hematological Disease	19

Supplementary Table 1: Different biological networks, found in DC and DE gene lists.

2 The Covariance Distance

The covariance distance between two genes x_i^c and x_j^c is defined as follows (notations are defined in the main text):

$$d_{ij}^c = \widehat{\sigma}(x_i^c - x_j^c)$$

We claim this statistic is the sample counterpart of an L^2 distance defined on a Hilbert space of random variables.

Recall that X_i^c is the random variable of the expression level associated with gene i in phenotype c, $i = 1, \ldots, m$, and $c = A, B$. Let $(\Omega, \mathcal{F}, P)$ be the probability space on which X_i^cs are defined and $L^2(\Omega, \mathrm{d}P)$ be the random variables with finite variance (which implies finite second order moment). It is well known that the following equivalence classes in $L^2(\Omega, \mathrm{d}P)$ form a Hilbert space $\mathcal{H}$.

1. Two random variables X and Y are said to be equivalent if they differ by a nonrandom constant with probability 1: $\exists a \in \mathbb{R}$, s.t. $P\{X - Y = a\} = 1$. It is easy to show that this is indeed an equivalence relation on $L^2(\Omega, \mathrm{d}P)$. Denote $[X]$ as the equivalent class containing X.

2. The set of all equivalent classes form a linear space with the following addition and scalar multiplication operations:

$$[X] + [Y] = [X + Y], \quad k[X] = [kX].$$

3. The covariance function can serve as the inner product on this linear space: $\langle [X],\ [Y] \rangle = \mathrm{cov}([X], [Y])$.

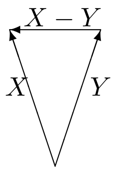
This inner product induces a norm (length) and a distance function:

$$\|[X]\| = \sqrt{\mathrm{cov}([X], [X])} = \sigma([X]), \quad \rho([X], [Y]) = \|[X] - [Y]\| = \sigma([X] - [Y]).$$

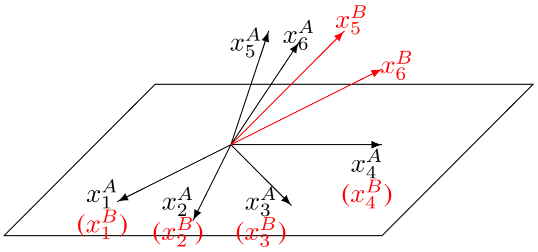
Clearly, the covariance distance is the sample counterpart of the distance function induced by the covariance inner product.

Gene expressions can be considered as the vectors in this Hilbert space $\mathcal{H}$. Figure 5 shows a graphical rendition of two such vectors in $\mathcal{H}$. The (population) covariance distance between X and Y is the length of $X - Y$.

Figure 6 depicts a more realistic situation: 1. six genes are expressed under biological conditions A(black) and B(red); 2. only genes 5 and 6 change their associations (in terms of the covariance distance) with other genes. The relational changes of genes 5 and 6 are reflected by the changes of the covariance distance between genes.



Supplementary Figure 5: Correlation distance.



Supplementary Figure 6: Genes under two different conditions.

3 The N-statistic

We choose a multivariate nonparametric N-distance with Euclidean kernel as a measure of the distance between two random vectors. Denote the random vectors in groups A and B by $\mathbf{D}^A$ and $\mathbf{D}^B$, respectively. Given n_s realizations of these two vectors $\mathbf{D}_k^A$ and $\mathbf{D}_k^B$ $(1 \leqslant k \leqslant n_s)$, the sample N-distance between these two random vectors is defined as follows:

$$\begin{aligned} N \quad = \quad & \frac{2}{n_s^2} \sum_{k=1}^{n_s} \sum_{l=1}^{n_s} L(\mathbf{D}_k^A, \mathbf{D}_l^B) \\ & - \frac{1}{n_s^2} \sum_{k=1}^{n_s} \sum_{l=1}^{n_s} L(\mathbf{D}_k^A, \mathbf{D}_l^A) \\ & - \frac{1}{n_s^2} \sum_{k=1}^{n_s} \sum_{l=1}^{n_s} L(\mathbf{D}_k^B, \mathbf{D}_l^B), \end{aligned}$$

where $L(x, y) = \|x - y\| = \sqrt{\sum_{s=1}^{d} (x_s - y_s)^2}$ is the kernel defined by Euclidean distance with vector dimension d.

4 Testing Differential Correlation by Likelihood Ratio Test

Suppose we have two genes x_1 and x_2. Assume that the joint distribution of them is $N\left((\mu_1,\mu_2),\ \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)\right)$, where ρ takes two possible values: $H_0\ :\ \rho=\rho_0$ and $H_1\ :\ \rho=\rho_0+\delta$. Since we are interested in the small change of ρ, we assume that the difference δ is relatively small.

For these two genes, denote their expression levels of n subjects by x_{1j} and x_{2j} ($1\leqslant j\leqslant n$), respectively. Their log-likelihood functions are

$$\ell(\rho|x_1,x_2)=-n\log(2\pi)-\frac{n}{2}\log(1-\rho^2)-\frac{1}{2}\sum_{j=1}^{n}(x_{1j}-\mu_1,x_{2j}-\mu_2)\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}^{-1}\begin{pmatrix}x_{1j}-\mu_1\\ x_{2j}-\mu_2\end{pmatrix}.$$

The log-likelihood ratio test statistic takes the form

$$T=\sum_{j=1}^{n}(x_{1j}-\mu_1,x_{2j}-\mu_2)\begin{pmatrix}2\rho_0 & -\rho_0^2-1\\ -\rho_0^2-1 & 2\rho_0\end{pmatrix}\begin{pmatrix}x_{1j}-\mu_1\\ x_{2j}-\mu_2\end{pmatrix}. \tag{1}$$

This is because

$$\begin{aligned}2(\ell(\rho_0|x_1,x_2)-\ell(\rho_0+\delta|x_1,x_2)) &= C+\sum_{j=1}^{n}(x_{1j}-\mu_1,x_{2j}-\mu_2)\left(\begin{pmatrix}1 & \rho_0+\delta\\ \rho_0+\delta & 1\end{pmatrix}^{-1}-\begin{pmatrix}1 & \rho_0\\ \rho_0 & 1\end{pmatrix}^{-1}\right)\begin{pmatrix}x_{1j}-\mu_1\\ x_{2j}-\mu_2\end{pmatrix}\\ &= C+D\sum_{j=1}^{n}(x_{1j}-\mu_1,x_{2j}-\mu_2)\begin{pmatrix}2\rho_0+\delta & -\rho_0^2-\delta\rho_0-1\\ -\rho_0^2-\delta\rho_0-1 & 2\rho_0+\delta\end{pmatrix}\begin{pmatrix}x_{1j}-\mu_1\\ x_{2j}-\mu_2\end{pmatrix}\\ &\approx C+D\sum_{j=1}^{n}(x_{1j}-\mu_1,x_{2j}-\mu_2)\begin{pmatrix}2\rho_0 & -\rho_0^2-1\\ -\rho_0^2-1 & 2\rho_0\end{pmatrix}\begin{pmatrix}x_{1j}-\mu_1\\ x_{2j}-\mu_2\end{pmatrix}.\end{aligned}$$

where C and D are constants which do not depend on the observation terms $x_{1\cdot}$ and $x_{2\cdot}$.

According to (1), when ρ is close to 0, T is approximately $\propto\sum_{j=1}^{n}(x_{1j}-\mu_1)(x_{2j}-\mu_2)$ which is equivalent to the sample correlation coefficient. In other words, if we assume genes are uncorrelated, the sample correlation coefficient is the most power test statistic for testing small change (δ term) of the correlation coefficient due to the Neyman-Pearson lemma.

On the other hand, when ρ is close to 1, T is approximately $\propto\sum_{j=1}^{n}(x_{1j}-x_{2j}-\mu_1+\mu_2)^2$, which is equivalent to the covariance distance. I.e., when genes are highly positively correlated, the covariance distance, rather than the sample correlation coefficient, is the most power test statistic for testing small change (δ term) of the correlation coefficient.

Based on the real data analysis, we observe that most pairwise intergene correlation coefficients are much closer to one than to zero. Therefore it is no surprise that the **TCDV** method out-performs the **CV** method.