A systems biology approach for pathway level analysis

Sorin Draghici,[1,2,4] Purvesh Khatri,[2] Adi Laurentiu Tarca,[1,2,3] Kashyap Amin,[2] Arina Done,[2] Calin Voichita,[2] Constantin Georgescu,[2] and Roberto Romero[3]

[1,2,3]Karmanos Cancer Institute, Wayne State University, Detroit, Michigan 48202, USA; [2]Department of Computer Science, Wayne State University, Detroit, Michigan 48202, USA; [3]Perinatology Research Branch, NIH/NICHD, Detroit, Michigan 48201, USA

A common challenge in the analysis of genomics data is trying to understand the underlying phenomenon in the context of all complex interactions taking place on various signaling pathways. A statistical approach using various models is universally used to identify the most relevant pathways in a given experiment. Here, we show that the existing pathway analysis methods fail to take into consideration important biological aspects and may provide incorrect results in certain situations. By using a systems biology approach, we developed an impact analysis that includes the classical statistics but also considers other crucial factors such as the magnitude of each gene's expression change, their type and position in the given pathways, their interactions, etc. The impact analysis is an attempt to a deeper level of statistical analysis, informed by more pathway-specific biology than the existing techniques. On several illustrative data sets, the classical analysis produces both false positives and false negatives, while the impact analysis provides biologically meaningful results. This analysis method has been implemented as a Web-based tool, Pathway-Express, freely available as part of the Onto-Tools (http://vortex.cs.wayne.edu).

[Supplemental material is available online at www.genome.org.]

Together with the ability of generating a large amount of data per experiment, high-throughput technologies also brought the challenge of translating such data into a better understanding of the underlying biological phenomena. Independent of the platform and the analysis methods used, the result of a high-throughput experiment is, in many cases, a list of differentially expressed genes. The common challenge faced by all researchers is to translate such lists of differentially expressed genes into a better understanding of the underlying biological phenomena and, in particular, to put this in the context of the whole organism as a complex system. In 2002, a computerized analysis approach using the Gene Ontology (GO) was proposed to deal with this issue (Khatri et al. 2002; Draghici et al. 2003). This approach takes a list of differentially expressed genes and uses a statistical analysis to identify the GO categories (e.g. biological processes, etc.) that are over- or under-represented in the condition under study. Given a set of differentially expressed genes, this approach compares the number of differentially expressed genes found in each category of interest with the number of genes expected to be found in the given category just by chance. If the observed number is substantially different from the one expected just by chance, the category is reported as significant. A statistical model (e.g. hypergeometric) can be used to calculate the probability of observing the actual number of genes just by chance, i.e., a *P*-value. Currently, there are over 20 tools using this over-representation approach (ORA) (Khatri and Draghici 2005). In spite of its wide adoption, this approach has a number of limitations related to the type, quality, and structure of the annotations available. An alternative approach considers the distribution of the pathway genes in the entire list of genes and performs a functional class scoring (FCS), which also allows adjustments for gene correlations (Goeman et al. 2004; Pavlidis et al. 2004). Arguably the state of the art in the FCS category, the Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005; Tian et al. 2005), ranks all genes based on the correlation between their expression and the given phenotypes, and calculates a score that reflects the degree to which a given pathway *P* is represented at the extremes of the entire ranked list. The score is calculated by walking down the list of genes ordered by expression change. The score is increased for every gene that belongs to *P* and decreased for every gene that does not. Statistical significance is established with respect to a null distribution constructed by permutations.

Both ORA and FCS techniques currently used are limited by the fact that each functional category is analyzed independently without a unifying analysis at a pathway or system level (Tian et al. 2005). This approach is not well suited for a systems biology approach that aims to account for system level dependencies and interactions as well as to identify perturbations and modifications at the pathway or organism level (Stelling 2004). Several pathway databases such as KEGG (Ogata et al. 1999), BioCarta (http://www.biocarta.com), and Reactome (Joshi-Tope et al. 2005) currently describe metabolic pathway and gene signaling networks offering the potential for a more complex and useful analysis. A recent technique, ScorePage, has been developed in an attempt to take advantage of these types of data for the analysis of metabolic pathways (Rahnenfuhrer et al. 2004). Unfortunately, no such technique currently exists for the analysis of gene signaling networks. All pathway analysis tools currently available use one of the ORA approaches above and fail to take advantage of the much richer data contained in these resources. GenMAPP/MAPPfinder (Doniger et al. 2003; Dahlquist et al. 2002) and Gene-Sifter use a standardized Z-score. PathwayProcessor (Grosu et al. 2002), PathMAPA (Pan et al. 2003), Cytoscape (Shannon et al. 2003), and PathwayMiner (Pandey et al. 2004) use Fisher's

[4]Corresponding author.
E-mail sod@cs.wayne.edu; fax (313) 577-0868.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6202607.

exact test. MetaCore uses a hypergeometric model, while ArrayXPath (Chung et al. 2004) offers both Fisher's exact test and a false discovery rate (FDR). Finally, VitaPad (Holford et al. 2004) and Pathway Studio (Nikitin et al. 2003) focus on visualization alone and do not offer any analysis.

The approaches currently available for the analysis of gene signaling networks share a number of important limitations. First, these approaches consider only the set of genes on any given pathway and ignore their position in those pathways. This may be unsatisfactory from a biological point of view. If a pathway is triggered by a single gene product or activated through a single receptor and if that particular protein is not produced, the pathway will be greatly impacted, probably completely shut off. A good example is the insulin pathway (http://www.genome.ac.jp/KEGG/pathway/hsa/hsa04910.html). If the insulin receptor (*INSR*) is not present, the entire pathway is shut off. Conversely, if several genes are involved in a pathway but they only appear somewhere downstream, changes in their expression levels may not affect the given pathway as much.

Second, some genes have multiple functions and are involved in several pathways but with different roles. For instance, the above *INSR* is also involved in the adherens junction pathway as one of the many receptor protein tyrosine kinases. However, if the expression of *INSR* changes, this pathway is not likely to be heavily perturbed because *INSR* is just one of many receptors on this pathway. Once again, all these aspects are not considered by any of the existing approaches.

Probably the most important challenge today is that the knowledge embedded in these pathways about how various genes interact with each other is not currently exploited. The very purpose of these pathway diagrams is to capture some of our knowledge about how genes interact and regulate each other. However, the existing analysis approaches consider only the sets of genes involved on these pathways, without taking into consideration their topology. In fact, our understanding of various pathways is expected to improve as more data are gathered. Pathways will be modified by adding, removing or redirecting links on the pathway diagrams. Most existing techniques are completely unable to even sense such changes. Thus, these techniques will provide identical results as long as the pathway diagram involves the same genes, even if the interactions between them are completely redefined over time.

Finally, up to now the expression changes measured in these high-throughput experiments have been used only to identify differentially expressed genes (ORA approaches) or to rank the genes (FCS methods), but not to estimate the impact of such changes on specific pathways. Thus, ORA techniques will see no difference between a situation in which a subset of genes is differentially expressed just above the detection threshold (e.g., twofold) and the situation in which the same genes are changing by many orders of magnitude (e.g., 100-fold). Similarly, FCS techniques can provide the same rankings for entire ranges of expression values, if the correlations between the genes and the phenotypes remain similar. Even though analyzing this type of information in a pathway and system context would be extremely meaningful from a biological perspective, currently there is no technique or tool able to do this.

We propose a radically different approach for pathway analysis that attempts to capture all aspects above. An impact factor (*IF*) is calculated for each pathway incorporating parameters such as the normalized fold change of the differentially expressed genes, the statistical significance of the set of pathway genes, and the topology of the signaling pathway. We show on a number of real data sets that the intrinsic limitations of the classical analysis produce both false positives and false negatives while the impact analysis provides biologically meaningful results.

Impact analysis

Our goal is to develop an analysis model that would require both a statistically significant number of differentially expressed genes *and* biologically meaningful changes on a given pathway. In this model, the *IF* of a pathway P_i is calculated as the sum of two terms:

$$IF(P_i) = \log\left(\frac{1}{p_i}\right) + \frac{\sum_{g\in P_i} |PF(g)|}{|\overline{\Delta E}| \cdot N_{de}(P_i)}. \quad (1)$$

The first term is a probabilistic term that captures the significance of the given pathway P_i from the perspective of the set of genes contained in it. This term captures the information provided by the currently used classical statistical approaches and can be calculated using either an ORA (e.g., z-test [Doniger et al. 2003], contingency tables [Pan et al. 2003; Pandey et al. 2004], etc.), a FCS approach (e.g., GSEA; Mootha et al. 2003; Subramanian et al. 2005) or other more recent approaches (Robinson et al. 2004; Breslin et al. 2005; Tian et al. 2005). The p_i value corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed, when the null hypothesis is true. The results presented here were obtained using the hypergeometric model (Tavazoie et al. 1999; Draghici et al. 2003) in which p_i is the probability of obtaining at least the observed number of differentially expressed gene, N_{de}, just by chance.

The second term in Equation 1 is a functional term that depends on the identity of the specific genes that are differentially expressed as well as on the interactions described by the pathway (i.e., its topology). In essence, this term sums up the absolute values of the perturbation factors (PFs) for all genes g on the given pathway P_i. The PF of a gene g is calculated as follows:

$$PF(g) = \Delta E(g) + \sum_{u\in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)}. \quad (2)$$

In this equation, the first term captures the quantitative information measured in the gene expression experiment. The factor $\Delta E\ (g)$ represents the signed normalized measured expression change of the gene g determined using one of the available methods (Quackenbush 2001; Churchill 2002; Draghici 2002; Yang and Speed 2002). The second term is a sum of all PFs of the genes u directly upstream of the target gene g, normalized by the number of downstream genes of each such gene $N_{ds}(u)$, and weighted by a factor β_{ug}, which reflects the type of interaction: $\beta_{ug} = 1$ for induction, $\beta_{ug} = -1$ for repression. (In KEGG, which is the source of the pathways used here, this information about the type of interaction is available for every link between two genes in the description of the pathway topology.) US_g is the set of all such genes upstream of g. The second term here is similar to the PageRank index used by Google (Page et al. 1998), only we weight the downstream instead of the upstream connections (a Web page is important if other pages point to it, whereas a gene is important if it influences other genes).

Under the null hypothesis, which assumes that the list of differentially expressed genes only contains random genes, the likelihood that a pathway has a large *IF* is proportional to the

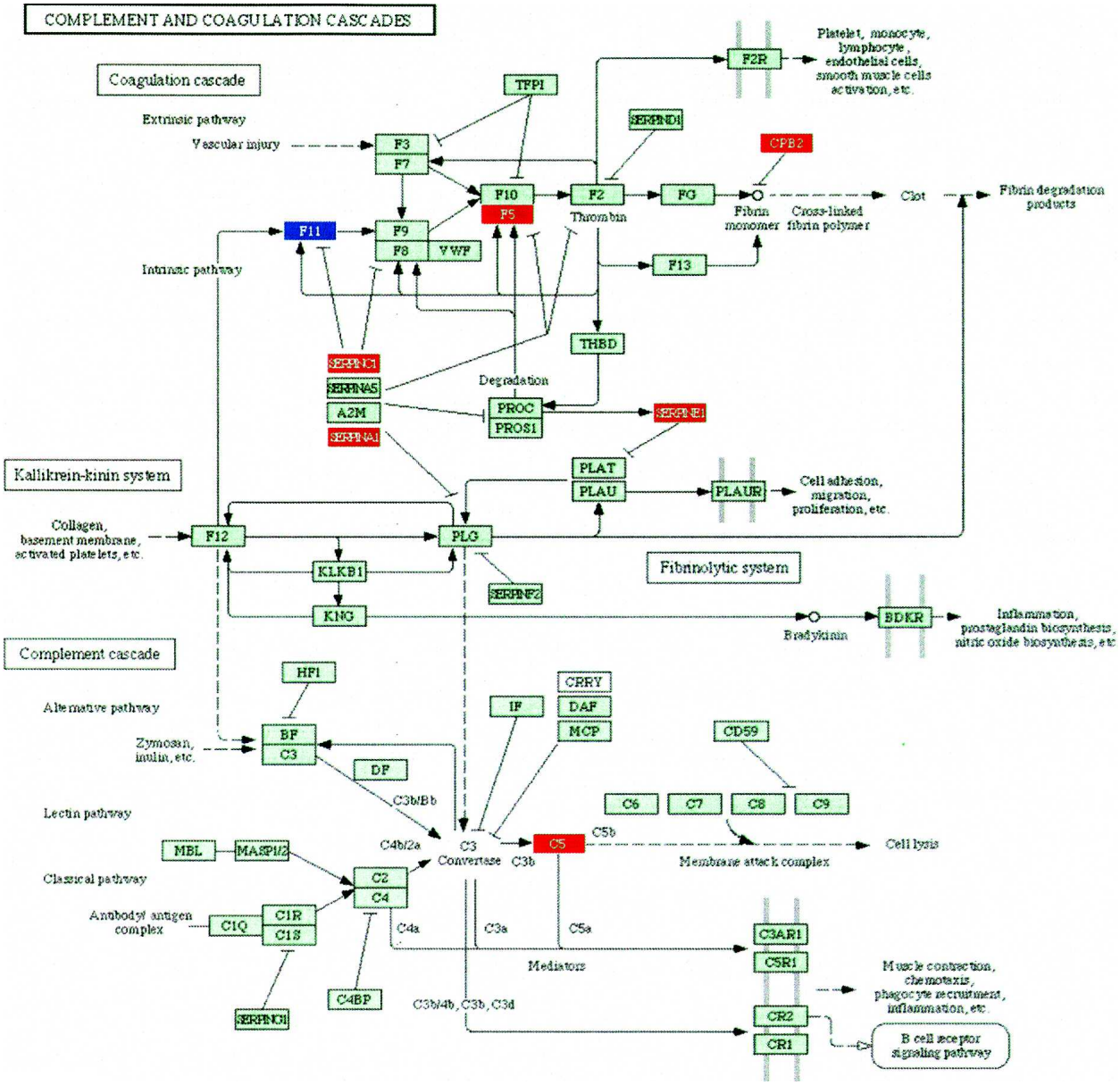

Figure 1. The complement and coagulation cascade as affected by treatment with palmitate in a hepatic cell line. There are seven differentially expressed genes (red, up-regulated; blue, down-regulated) out of 69 total genes. All classical ORA models would give any other pathway with the same proportion of genes a similar *P*-value, disregarding the fact that six out of these seven genes are involved in the same region of the pathway, closely interacting with each other. Both ORA and GSEA would yield exactly the same significance value to this pathway even if the diagram were to be completely redesigned by future discoveries. In contrast, the impact factor can distinguish between this pathway and any other pathway with the same proportion of differentially expressed gene, as well as take into account any future changes to the topology of the pathway.

number of such "differentially expressed" genes that fall on the pathway, which in turn is proportional to the size of the pathway. Thus, we need to normalize with respect to the size of the pathway by dividing the total perturbation by the number of differentially expressed genes on the given pathway, $N_{de}(P_i)$. Furthermore, various technologies can yield systematically different estimates of the fold changes. For instance, the fold changes reported by microarrays tend to be compressed with respect to those reported by RT-PCR (Canales et al. 2006; Draghici et al. 2006). In order to make the *IF*s as independent as possible from the technology, and also comparable between problems, we also divide the second term in Equation 1 by the mean absolute fold change ΔE, calculated across all differentially expressed genes. Assuming that there are at least some differentially expressed genes anywhere in the data set, both ΔE and $N_{de}(P_i)$ are different from zero so the second term is properly defined. (If there are no differentially expressed genes anywhere, the problem of finding the impact on various pathways is meaningless.)

It can be shown that the *IF*s correspond to the negative log of the global probability of having both a statistically significant number of differentially expressed genes and a large perturbation in the given pathway. *IF* values, *if*, will follow a $\Gamma(2,1)$ distribu-

tion from which P-values can be calculated as $P = (if + 1) \cdot e^{-if}$ (for details, see Supplemental materials).

The impact analysis proposed here extends and enhances the existing statistical approaches by incorporating the novel aspects discussed above. For instance, the second term of the gene perturbation (in Equation 2) increases the PF scores of those genes that are connected through a direct signaling link to other differentially expressed genes (e.g., the PFs of *F5* and *F11* in Fig. 1 are both increased because of the differentially expressed *SERPINC1* and *SERPINA1*). This will yield a higher overall score for those pathways in which the differentially expressed genes are localized in a connected subgraph, as in this example. Interestingly, when the limitations of the existing approaches are forcefully imposed (e.g., ignoring the magnitude of the measured expression changes or ignoring the regulatory interactions between genes), the impact analysis reduces to the classical statistics and yields the same results. For instance, if there are no perturbations directly upstream of a given gene, the second term in Equation 2 is zero and the PF reduces to the measured expression change ΔE, which is the classical way of assessing the impact of a condition upon a given gene. A more detailed discussion of various particular cases is included in the Supplemental materials.

Results

We have used this pathway analysis approach to analyze several data sets. A first such set includes genes associated with better survival in lung adenocarcinoma (Beer et al. 2002). These genes have the potential to represent an important tool for the therapeutic decision, and if the correct regulatory mechanisms are identified, they could also be potential drug targets. The expression values of the 97 genes associated with better survival identified by Beer and colleagues were compared between the cancer and healthy groups. These data were then analyzed using a classical ORA approach (hypergeometric model), a classical FCS approach (GSEA), and our impact analysis. Figure 2 shows a comparison between the results obtained with the three approaches.

From a statistical perspective, the power of both classical techniques appears to be very limited. The corrected P-values do not yield any pathways at the usual 0.01 or 0.05 significance levels, independently of the type of correction. If the significance levels were to be ignored and the techniques used only to rank the pathways, the results would continue to be unsatisfactory. According to the classical ORA analysis, the most significantly affected pathways in this data set are *prion disease*, *focal adhesion*, and *Parkinson's disease*. In reality, both prion and Parkinson's diseases are pathways specifically associated to diseases of the central nervous system and are unlikely to be related to lung adenocarcinomas. In this particular case, prion disease ranks at the top only due to the differential expression of *LAMB1*. Since this pathway is rather small (14 genes), every time any one gene is differentially expressed, the hypergeometric analysis will rank it highly. A similar phenomenon happens with Parkinson's disease, indicating that this is a problem associated with the method rather than with a specific pathway. At the same time, pathways highly relevant to cancer such as *cell cycle* and *Wnt signaling* are ranked in the lower half of the pathway list. The most significant pathways reported as enriched in cancer by GSEA (Subramanian et al. 2005) are cell cycle, *Huntington's disease*, *DRPLA*, *Alzheimer's disease*, and Parkinson's disease (see Fig. 2). Among these, only cell cycle is relevant, while Huntington's, Alzheimer's and Par-

A

Pathway name	ORA (hypergeometric)		
	p-value	FDR	Bonferroni
Prion disease	0.149649	0.627132	1
Focal adhesion	0.155424	0.627132	1
Parkinson's disease	0.164842	0.627132	1
Dentatorubropallidoluysian atrophy	0.179767	0.627132	1
Calcium signaling pathway	0.262884	0.627132	1
Alzheimer's disease	0.277100	0.627132	1
Apoptosis	0.283744	0.627132	1
TGF-beta signaling pathway	0.303663	0.627132	1
Huntington's disease	0.327491	0.627132	1
Toll-like receptor signaling pathway	0.330069	0.627132	1
Wnt signaling pathway	0.369145	0.637613	1
Regulation of actin cytoskeleton	0.439390	0.695701	1
MAPK signaling pathway	0.560814	0.762988	1
Phosphatidylinositol signaling system	0.572396	0.762988	1
Adherens junction	0.602359	0.762988	1
Complement and coagulation cascades	0.680333	0.766820	1
Cell cycle	0.686102	0.766820	1
Cytokine-cytokine receptor interaction	0.820650	0.866242	1
Neuroactive ligand-receptor interaction	0.972996	0.972996	1

B

Enriched in cancer

Pathway Name	NOM p-val	FDR q-val	FWER p-val
Cell cycle	0.038	0.118	0.140
Huntington's disease	0.074	0.217	0.546
Dentatorubropallidoluysian atrophy (DRPLA)	0.149	0.291	0.751
Alzheimer's disease	0.189	0.344	0.877
Parkinson's disease	0.373	0.485	0.984
Adherens junction	0.583	0.651	0.998
Wnt signaling pathway	0.861	0.785	1

Enriched in normal

Pathway Name	NOM p-val	FDR q-val	FWER p-val
MAPK signaling pathway	0.007	0.170	0.361
Apoptosis	0.019	0.175	0.304
Complement and coagulation cascades	0.037	0.255	0.298
Phosphatidylinositol signaling system	0.189	0.343	0.823
Regulation of actin cytoskeleton	0.010	0.356	0.223
Focal adhesion	0.160	0.384	0.817
Cytokine-cytokine receptor interaction	0.241	0.420	0.910
Toll-like receptor signaling pathway	0.330	0.451	0.963
Calcium signaling pathway	0.308	0.489	0.960
Prion disease	0.474	0.563	0.986
TGF-beta signaling pathway	0.631	0.699	0.998
Neuroactive ligand-receptor interaction	0.947	0.957	1

C

Pathway name	Impact Factor			
	IF	p-value	FDR	Bonferroni
Cell cycle	19.26	8.76E-08	1.66E-06	1.66E-006
Focal adhesion	7.414	0.005072	0.048180	0.0956831
Wnt signaling pathway	6.780	0.008840	0.055988	0.1679642
Dentatorubropallidoluysian atrophy	5.535	0.025788	0.122495	0.4899810
Huntington's disease	4.543	0.058985	0.203925	1
Apoptosis	4.407	0.065921	0.203925	1
Regulation of actin cytoskeleton	4.246	0.075130	0.203925	1
TGF-beta signaling pathway	3.511	0.134730	0.319984	1
Complement and coagulation cascades	3.161	0.176357	0.354145	1
Adherens junction	2.953	0.206279	0.354145	1
Alzheimer's disease	2.752	0.239378	0.354145	1
Parkinson's disease	2.631	0.261455	0.354145	1
Toll-like receptor signaling pathway	2.576	0.272054	0.354145	1
Prion disease	2.572	0.272839	0.354145	1
Calcium signaling pathway	2.538	0.279588	0.354145	1
Cytokine-cytokine receptor interaction	2.353	0.318815	0.366952	1
Phosphatidylinositol signaling system	2.311	0.328326	0.366952	1
MAPK signaling pathway	2.205	0.353353	0.372984	1
Neuroactive ligand-receptor interaction	0.576	0.885936	0.885936	1

Figure 2. A comparison between the results of the classical probabilistic approaches (*A*, hypergeometric; *B*, GSEA) and the results of the pathway impact analysis (*C*) for a set of genes differentially expressed in lung adenocarcinoma. The pathways marked with green are considered most likely to be linked to this condition in this experiment. The ones in red are unlikely to be related. The ranking of the pathways produced by the classical approaches is very misleading. According to the hypergeometric model, the most significant pathways in this condition are: prion disease, focal adhesion, and Parkinson's disease. Two out of these three are likely to be incorrect. GSEA yields cell cycle as the most enriched pathway in cancer, but three out of the four subsequent pathways are clearly incorrect. In contrast, all three top pathways identified by the impact analysis are relevant to the given condition. The impact analysis is also superior from a statistical perspective. According to both hypergeometric and GSEA, no pathway is significant at the usual 1% or 5% levels on corrected P-values. In contrast, according to the impact analysis, the cell cycle is significant at 1%, and focal adhesion and Wnt signaling are significant at 5% and 10%, respectively.

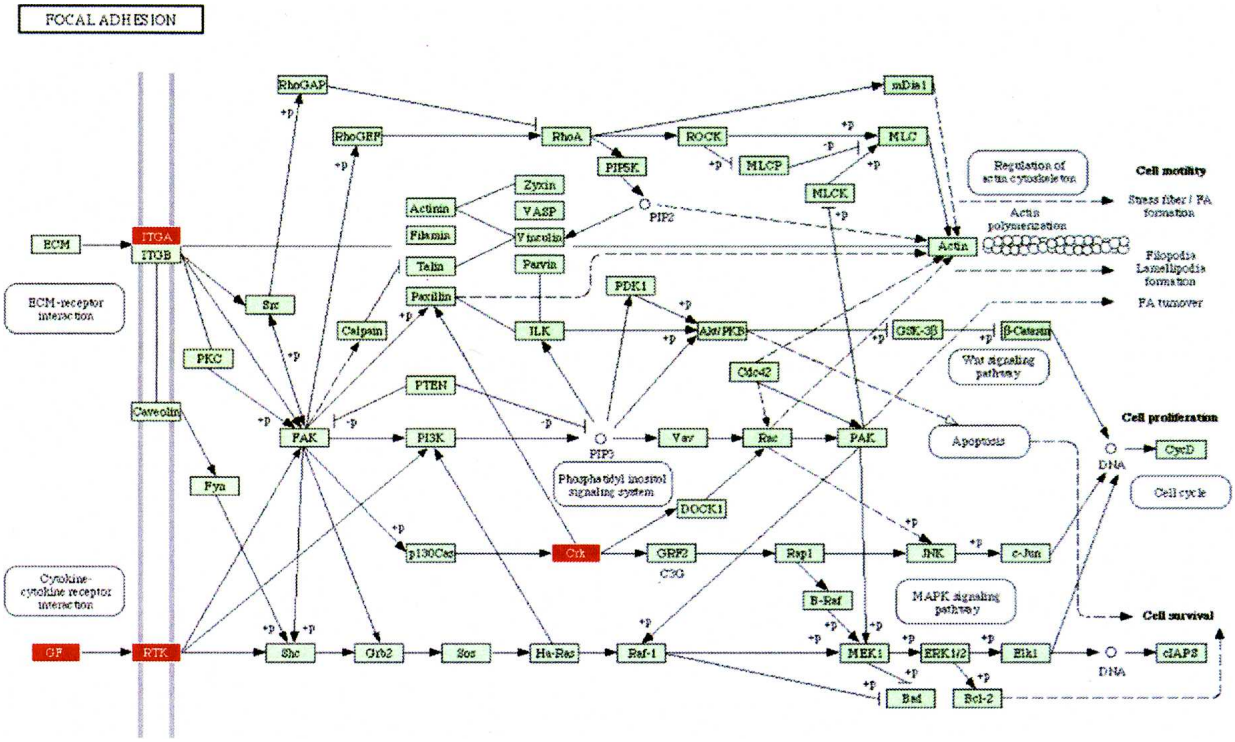

Figure 3. The focal adhesion pathway as impacted in lung adenocarcinoma vs. normal. In this condition, both *ITG* and *RTK* receptors are perturbed, as well as the *VEGF* ligand. Because these three genes appear at the very beginning and affect both entry points controlling this pathway, their perturbations are widely propagated throughout the pathway and this pathway appears as highly impacted. All classical approaches completely ignore the positions of the genes on the given pathways and fail to identify this pathway as significant.

kinson's diseases are clearly incorrect. However, although ranked first, cell cycle is not significant in GSEA, even at the most lenient 10% significance and with the least conservative correction.

In contrast, the impact analysis reports cell cycle as the most perturbed pathway in this condition and also as highly significant from a statistical perspective ($P = 1.6 \times 10^{-6}$). Since early articles on the molecular mechanisms perturbed in lung cancers (Slebos and Rodenhuis 1989; Nau et al. 1985) until the most recent articles on this topic (Panani and Roussos 2006; Coe et al. 2006), there is a consensus that the cell cycle is highly deranged in lung cancers. Moreover, cell cycle genes have started to be considered both as potential prognostic factors and therapeutic targets (Vincenzi et al. 2006). The second most significant pathway as reported by the impact analysis is focal adhesion. An inspection of this pathway (shown in Fig. 3) shows that in these data, both *ITG* and *RTK* receptors are perturbed, as well as the *VEGF* ligand. Because these three genes appear at the very beginning and affect both entry points controlling this pathway, their perturbations are widely propagated throughout the pathway. Furthermore, the *CRK* oncogene was also found to be upregulated. Increased levels of *CRK* proteins have been observed in several human cancers, and over-expression of *CRK* in epithelial cell cultures promotes enhanced cell dispersal and invasion (Rodrigues et al. 2005). For this pathway, the impact analysis yields a raw *P*-value of 0.005, which remains significant even after the FDR correction ($P = 0.048$), at the 5% level. In contrast, the ORA analysis using the hypergeometric model yields a raw *P*-value of 0.155 (FDR corrected to 0.627), while the GSEA analysis yields a raw *P*-value of 0.16 (FDR corrected to 0.384). For both techniques, not even the raw *P*-values are significant at the usual levels of 5% or 10%. This is not a mere accident but an illustration of the intrinsic limitations of the classical approaches. These approaches completely ignore the position of the genes on the given pathways, and therefore, they are not able to identify this pathway as being highly impacted in this condition. Note that any ORA approach will yield the same results for this pathway for any set of four differentially expressed genes from the set of genes on this pathway. Similarly, GSEA will yield the same results for any other set of four genes with similar expression values (yielding similar correlations with the phenotype). Both techniques are unable to distinguish between a situation in which these genes are upstream, potentially commandeering the entire pathway as in this example, or randomly distributed throughout the pathway.

The third pathway as ranked by the impact analysis is Wnt signaling (FDR corrected $P = 0.055$, significant at 10%). The importance of this pathway is well supported by independent research. At least three mechanisms for the activation of Wnt signaling pathway in lung cancers have been recently identified: (1) over-expression of Wnt effectors such as *Dvl*, (2) activation of a non-canonical pathway involving *MAPK* (previously known as *JNK*), and (3) repression of Wnt antagonists such as *WIF* (Mazieres et al. 2005). Mazieres and colleagues also argue that the blockade of Wnt pathway may lead to new treatment strategies in lung cancer.

In the same data set, Huntington's disease, Parkinson's disease, prion disease, and Alzheimer's disease have low *IF*s (corrected *P*-values of >0.20), correctly indicating that they are unlikely to be relevant in lung adenocarcinomas.

A second data set includes genes identified as being associated with poor prognosis in breast cancer (van't Veer et al. 2002).

A

Pathway name	ORA (hypergeometric)		
	p-value	FDR	bonferroni
Cell cycle	3.1E-07	2.8E-06	2.765E-06
MAPK signaling pathway	0.02513	0.11309	0.2261834
Parkinson's disease	0.10752	0.32255	0.9676532
Cytokine-cytokine receptor interaction	0.24736	0.47992	1
Focal adhesion	0.29628	0.47992	1
Calcium signaling pathway	0.37158	0.47992	1
Regulation of actin cytoskeleton	0.40691	0.47992	1
TGF-beta signaling pathway	0.42660	0.47992	1
Neuroactive ligand-receptor interaction	0.58749	0.58749	1

B

Enriched in poor prognosis

Pathway Name	NOM p-val	FDR q-val	FWER p-val
Ubiquitin mediated proteolysis	0.031	0.113	0.111
Prion disease	0.352	0.570	0.802
Alzheimer's disease	0.279	0.625	0.683
Tight junction	0.848	0.749	0.974
Parkinson's disease	0.638	0.795	0.958

Enriched in good prognosis

Pathway Name	NOM p-val	FDR q-val	FWER p-val
Notch signaling pathway	0.082	0.277	0.636
Neuroactive ligand-receptor interaction	0.050	0.280	0.542
Adherens junction	0.136	0.400	0.829
Wnt signaling pathway	0.058	0.410	0.534
Circadian rhythm	0.078	0.582	0.960
Complement and coagulation cascades	0.232	0.638	0.997
Apoptosis	0.212	0.691	0.996
MAPK signaling pathway	0.046	0.693	0.479
Amyotrophic lateral sclerosis	0.244	0.738	0.993
Jak-STAT signaling pathway	0.913	0.952	1
Dentatorubropallidoluysian atrophy	0.792	0.987	1
Cytokine-cytokine receptor interaction	0.913	0.987	1
Calcium signaling pathway	0.522	1	1
Focal adhesion	0.556	1	1
Regulation of actin cytoskeleton	0.575	1	1
Phosphatidylinositol signaling system	0.735	1	1
TGF-beta signaling pathway	0.815	1	1
Cell cycle	0.859	1	1
Huntington's disease	0.885	1	1

C

Pathway name	Impact Factor			
	IF	p-value	FDR	Bonferroni
Cell cycle	18.8	1.3E-07	1.2E-06	1.19E-006
Focal adhesion	7.06	0.00692	0.03112	0.0622412
TGF-beta signaling pathway	6.56	0.01075	0.03225	0.0967557
MAPK signaling pathway	5.40	0.02886	0.06493	0.2597164
Regulation of actin cytoskeleton	4.49	0.06180	0.11125	0.5562285
Parkinson's disease	3.12	0.18207	0.23946	1
Cytokine-cytokine receptor interaction	3.09	0.18624	0.23946	1
Neuroactive ligand-receptor interaction	2.87	0.21942	0.24685	1
Calcium signaling pathway	2.44	0.30047	0.30047	1

Figure 4. A comparison between the results of the classical (ORA) probabilistic approach (*A*), GSEA (*B*), and the results of the pathway impact analysis (*C*) for a set of genes associated with poor prognosis in breast cancer. The pathways marked with green are well supported by the existing literature. The ones in red are unlikely to be related. After correcting for multiple comparisons, GSEA fails to identify any pathway as significantly impacted in this condition at any of the usual significance levels (1%, 5%, or 10%). The hypergeometric model pinpoints cell cycle as the only significant pathway. Relevant pathways such as focal adhesion, TGF-beta signaling, and MAPK do not appear as significant from a hypergeometric point of view. While agreeing on the cell cycle, the impact analysis also identifies the three other relevant pathways as significant at the 5% level.

Figure 4 shows a comparison among the classical hypergeometric approach, GSEA, and the pathway impact analysis. Based on these data, GSEA finds no significantly impacted pathways at any of the usual 5% or 10% levels. In fact, the only FDR-corrected value below 0.25, in the entire data set is 0.11, corresponding to the *ubiquitin mediated proteolysis* pathway. Furthermore, GSEA's ranking does not appear to be useful for these data, with none of the cancer-related pathways being ranked toward the top. The most significant signaling pathway according to the hypergeometric analysis, cell cycle, is also the most significant in the impact analysis. However, the agreement between the two approaches stops here. In terms of statistical power, according to the classical hypergeometric model, there are no other significant pathways at either 5% or 10% significance on the corrected *P*-values. If we were to ignore the usual significance thresholds and only consider the ranking, the third highest pathway according to the hypergeometric model is Parkinson's disease. In fact, based on current knowledge, Parkinson's disease is unlikely to be related to rapid metastasis in breast cancer. At the same time, the impact analysis finds several other pathways as significant. For instance, focal adhesion is significant with an FDR-corrected *P*-value of 0.03. In fact, a link between focal adhesion and breast cancer has been previously established (Golubovskaya et al. 2002; van Nimwegen and van de Water 2006). In particular, *PTK2* (previously known as *FAK*), a central gene on the focal adhesion pathway, has been found to contribute to cellular adhesion and survival pathways in breast cancer cells, which are not required for survival in nonmalignant breast epithelial cell (Beviglia et al. 2003). Recently, it has also been shown that Doxorubicin, an anti-cancer drug, caused the formation of well-defined focal adhesions and stress fibers in mammary adenocarcinoma MTLn3 cells early after treatment (van Nimwegen et al. 2006). Consequently, the *PTK2/PI-3 kinase/PKB* signaling route within the focal adhesion pathway has been recently proposed as the mechanism through which Doxorubicin triggers the onset of apoptosis (van Nimwegen et al. 2006).

TGF-beta signaling ($P = 0.032$) and *MAPK* ($P = 0.064$) are also significant. Both fit well with previous research results. TGF-beta1, the main ligand for the TGF-beta signaling pathway, is known as a marker of invasiveness and metastatic capacity of breast cancer cells (Todorovic-Rakovic 2005). In fact, it has been suggested as the missing link in the interplay between estrogen receptors and *ERBB2* (previously known as *HER-2*; human epidermal growth factor receptor 2) (Todorovic-Rakovic 2005). Furthermore, plasma levels of TGF-beta1 have been used to identify low-risk postmenopausal metastatic breast cancer patients (Nikolic-Vukosavljevic et al. 2004). Finally, MAPK has been shown to be connected not only to cancer in general but to this particular type of cancer. For instance the proliferative response to progestin and estrogen was shown

to be inhibited in mammary cells microinjected with inhibitors of MAP kinase pathway (Chen et al. 2001). Also, it is worth noting the gap between the *P*-values for *regulation of actin cytoskeleton* ($P = 0.111$), which may be relevant in cancer, and the next pathway, Parkinson's disease ($P = 0.239$), which is irrelevant in this condition.

A third data set involves a set of differentially expressed genes obtained by studying the response of a hepatic cell line when treated with palmitate (Swagell et al. 2005). Figure 5 shows the comparison between the classical statistical analysis (ORA) and the pathway impact analysis. (The GSEA analysis requires expression values for all genes; since this experiment was performed with a custom array and not all values are publicly available, GSEA could not be applied here.) The classical statistical analysis yields three pathways significant at the 5% level: *complement and coagulation cascades*, focal adhesion, and MAPK. The impact analysis agrees on all these but also identifies several additional pathways. The top four pathways identified by the impact analysis are well supported by the existing literature. There are several studies that support the existence of a relationship between different coagulation factors, present in the complement and coagulation cascades pathway, and palmitate. Sanders et al. (1999), for instance, demonstrated that a high palmitate intake affects factor VII coagulant (*FVIIc*) activity. Interestingly, Figure 1 shows not only that this pathway has a higher than expected proportion of differentially expressed genes but also that six out of seven such genes are involved in the same region of the pathway, suggesting a coherently propagated perturbation. The focal adhesion and tight junction pathways involve cytoskeletal genes. Swagell et al. (2005) considered the presence of the cytoskeletal genes among the differentially expressed genes as very interesting and hypothesized that the down-regulation of these cytoskeletal genes indicates that palmitate decreases cell growth. Finally, the link between *MAPK* and the palmitate was established by Susztak et al. (2005), who showed that *p38* MAP kinase is a key player in the palmitate-induced apoptosis.

A

Pathway name	ORA (hypergeometric)		
	p-value	FDR	Bonferroni
Complement and coagulation cascades	1.26958E-07	2.28525E-06	2.28525E-06
Focal adhesion	4.03691E-05	0.000363322	0.000726643
MAPK signaling pathway	0.000523961	0.003143765	0.009431295
TGF-beta signaling pathway	0.011698758	0.052644412	0.210577648
Toll-like receptor signaling pathway	0.018714569	0.067372448	0.336862241
Calcium signaling pathway	0.024575814	0.072600598	0.442364654
Tight junction	0.028233566	0.072600598	0.508204185
Wnt signaling pathway	0.050174237	0.100857467	0.903136270
Phosphatidylinositol signaling system	0.058285692	0.100857467	1
Prion disease	0.060516063	0.100857467	1
Jak-STAT signaling pathway	0.061635119	0.100857467	1
Apoptosis	0.106427143	0.146873866	1
Cell cycle	0.106427143	0.146873866	1
Regulation of actin cytoskeleton	0.115415266	0.146873866	1
Alzheimer's disease	0.122394888	0.146873866	1
Huntington's disease	0.146968097	0.165339109	1
Neuroactive ligand-receptor interaction	0.233787848	0.247540075	1
Cytokine-cytokine receptor interaction	0.429908167	0.429908167	1

B

Pathway name	Impact Factor			
	IF	p-value	FDR	Bonferroni
Complement and coagulation cascades	19.374	7.85335E-08	1.41360E-06	1.44761E-06
Focal adhesion	13.791	1.51580E-05	1.36422E-04	3.01180E-04
MAPK signaling pathway	9.475	8.03922E-04	0.004823531	0.014470593
Tight junction	7.128	0.006521277	0.029345745	0.117382981
TGF-beta signaling pathway	6.868	0.008187095	0.029473543	0.147367717
Toll-like receptor signaling pathway	6.391	0.012391594	0.037174781	0.223048688
Calcium signaling pathway	5.774	0.021048873	0.052496861	0.378879719
Apoptosis	5.653	0.023331938	0.052496861	0.419974887
Regulation of actin cytoskeleton	5.225	0.033492741	0.066985482	0.602869334
Jak-STAT signaling pathway	4.983	0.041004319	0.073807774	0.738077735
Wnt signaling pathway	4.313	0.071158653	0.116441431	1
Phosphatidylinositol signaling system	3.975	0.093427025	0.133344438	1
Prion disease	3.937	0.096304316	0.133344438	1
Huntington's disease	3.839	0.104111596	0.133857767	1
Alzheimer's disease	3.387	0.148324272	0.171694058	1
Cell cycle	3.350	0.152616940	0.171694058	1
Neuroactive ligand-receptor interaction	2.414	0.305405348	0.323370368	1
Cytokine-cytokine receptor interaction	2.208	0.352624224	0.352624224	1

Figure 5. A comparison between the results of the classical probabilistic approach (*A*) and the results of the impact analysis (*B*) for a set of genes found to be differentially expressed in a hepatic cell line treated with palmitate. Green pathways are well supported by literature evidence, while red pathways are unlikely to be relevant. The classical statistical analysis yields three pathways significant at the 5% level: complement and coagulation cascades, focal adhesion, and MAPK. The impact analysis agrees on these three pathways but also identifies several additional pathways. Among these, *tight junction* is well supported by the literature.

Conclusions

A statistical approach using various models is commonly used in order to identify the most relevant pathways in a given experiment. This approach is based on the set of genes involved in each pathway. We identified a number of additional factors that may be important in the description and analysis of a given biological pathway. Based on these, we developed a novel impact analysis method that uses a systems biology approach in order to identify pathways that are significantly impacted in any condition monitored through a high-throughput gene expression technique. The impact analysis incorporates the classical probabilistic component but also includes important biological factors that are not captured by the existing techniques: the magnitude of the expression changes of each gene, the position of the differentially expressed genes on the given pathways, the topology of the pathway that describes how these genes interact, and the type of signaling interactions between them. The results obtained on several independent data sets show that the proposed approach is very promising. This analysis method has been implemented as a Web-based tool, Pathway-Express, freely available as part of the Onto-Tools (http://vortex.cs.wayne.edu).

Acknowledgments

This material is based upon work supported by the following grants: NSF DBI-0234806, CCF-0438970, 1R01HG003491-01A1,

1U01CA117478-01, 1R21CA100740-01, 1R01NS045207-01, 5R21EB000990-03, 2P30 CA022453-24. Onto-Tools currently runs on equipment provided by Sun Microsystems EDU 7824-02344-U and by NIH(NCRR) 1S10 RR017857-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NIH, DOD, or any other of the funding agencies.

References

Beer, D.G., Kardia, S.L., Huang, C.-C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8:** 816–824.

Beviglia, L., Golubovskaya, V., Xu, L., Yang, X., Craven, R.J., and Cance, W.G. 2003. Focal adhesion kinase N-terminus in breast carcinoma cells induces rounding, detachment and apoptosis. *Biochem. J.* **373:** 201–210.

Breslin, T., Krogh, M., Peterson, C., and Troein, C. 2005. Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics* **6:** 163. doi: 10.1186/1471-2105-6-163.

Canales, R.D., Luo, Y., Willey, J.C., Austermiller, B., Barbacioru, C.C., Boysen, C., Hunkapiller, K., Jensen, R.V., Knight, C.R., Lee, K.Y., et al. 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24:** 1115–1122.

Chen, Z., Gibson, T.B., Robinson, F., Silvestro, L., Pearson, G., Xu, B., Wright, A., Vanderbilt, C., and Cobb, M.H. 2001. MAP kinases. *Chem. Rev.* **101:** 2449–2476.

Chung, H.-J., Kim, M., Park, C.H., Kim, J., and Kim, J.H. 2004. ArrayXPath: Mapping and visualizing microarray gene-expression data with integrated biological pathway resources using scalable vector graphics. *Nucleic Acids Res.* **32**:W460–W464. doi: 10.1093/nar/gkh476.

Churchill, G.A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **32 (Suppl. S):** 490–495.

Coe, B.P., Lockwood, W.W., Girard, L., Chari, R., Macaulay, C., Lam, S., Gazdar, A.F., Minna, J.D., and Lam, W.L. 2006. Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br. J. Cancer* **94:** 1927–1935.

Dahlquist, K., Salomonis, N., Vranizan, K., Lawlor, S., and Conklin, B. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* **31:** 19–20.

Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., and Conklin, B.R. 2003. MAPPfinder: using Gene Ontology and GenMAPP to create a global gene expression profile from microarray data. *Genome Biol.* **4:** R7. doi:10.1186/gb-2003-4-1-r7.

Draghici, S. 2002. Statistical intelligence: Effective analysis of high-density microarray data. *Drug Discov. Today* **7:** S55–S63.

Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., and Krawetz, S.A. 2003. Global functional profiling of gene expression. *Genomics* **81:** 98–104.

Draghici, S., Khatri, P., Eklund, A.C., and Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22:** 101–109.

Goeman, J.J., van de Geer, S.A., de Kort, F., and van Houwelingen, H.C. 2004. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* **20:** 93–99.

Golubovskaya, V., Beviglia, L., Xu, L.H., Earp, H.S., Craven, R., and Cance, W. 2002. Dual inhibition of focal adhesion kinase and epidermal growth factor receptor pathways cooperatively induces death receptor-mediated apoptosis in human breast cancer cells. *J. Biol. Chem.* **277:** 38978–38987.

Grosu, P., Townsend, J.P., Hartl, D.L., and Cavalieri, D. 2002. Pathway processor: A tool for integrating whole-genome expression results into metabolic networks. *Genome Res.* **12:** 1121–1126.

Holford, M., Li, N., Nadkarni, P., and Zhao, H. 2004. VitaPad: Visualization tools for the analysis of pathway data. *Bioinformatics* **21:** 1596–1602.

Joshi-Tope, G., Gillespie, M., Vasrik, I., D'Eustachio, P., Schmidt, E., de Bone, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., et al. 2005. Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* **33:** D428–D432. doi: 10.1093/nar/gki072.

Khatri, P. and Draghici, S. 2005. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics* **21:** 3587–3595.

Khatri, P., Draghici, S., Ostermeier, G.C., and Krawetz, S.A. 2002. Profiling gene expression using Onto-Express. *Genomics* **79:** 266–270.

Mazieres, J., He, B., You, L., Xu, Z., and Jablons, D.M. 2005. Wnt signaling in lung cancer. *Cancer Lett.* **222:** 1–10.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. 2003. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34:** 267–273.

Nau, M.M., Brooks, B.J., Battey, J., Sausville, E., Gazdar, A.F., Kirsch, I.R., McBride, O.W., Bertness, V., Hollis, G.F., Minna, J.D., et al. 1985. L-myc, a new myc-related gene amplified and expressed in human small cell lung cancer. *Nature* **318:** 69–73.

Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. 2003. Pathway studio—The analysis and navigation of molecular networks. *Bioinformatics* **19:** 2155–2157.

Nikolic-Vukosavljevic, D., Todorovic-Rakovic, N., Demajo, M., Ivanovic, V., Neskovic, B., Markicevic, M., and Neskovic-Konstantinovic, Z. 2004. Plasma TGF-beta1-related survival of postmenopausal metastatic breast cancer patients. *Clin. Exp. Metastasis* **21:** 581–585.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27:** 29–34.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the Web. Technical report. Stanford University, Palo Alto, CA.

Pan, D., Sun, N., Cheung, K.-H., Guan, Z., Ma, L., Holford, M., Deng, X., and Zhao, H. 2003. PathMAPA: A tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for *Arbidopsis*. *BMC Bioinformatics* **4:** 56. doi: 10.1186/1471-2105-4-56.

Panani, A.D. and Roussos, C. 2006. Cytogenetic and molecular aspects of lung cancer. *Cancer Lett.* **239:** 1–9.

Pandey, R., Guru, R.K., and Mount, D.W. 2004. Pathway Miner: Extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* **20:** 2156–2158.

Pavlidis, P., Qin, J., Arango, V., Mann, J.J., and Sibille, E. 2004. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.* **29:** 1213–1222.

Quackenbush, J. 2001. Computational analysis of microarray data. *Nat. Rev. Genet.* **2:** 418–427.

Rahnenfuhrer, J., Domingues, F. S., Maydt, J., and Lengauer, T., 2004. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.* **3:** article16. http://www.bepress.com/sagmb/vol3/iss1/art16/

Robinson, P.N., Wollstein, A., Bohme, U., and Beattie, B. 2004. Ontologizing gene-expression microarray data: Characterizing clusters with gene ontology. *Bioinformatics* **20:** 979–981.

Rodrigues, S., Fathers, K., Chan, G., Zuo, D., Halwani, F., Meterissian, S., and Park, M. 2005. CrkI and CrkII function as key signaling integrators for migration and invasion of cancer cells. *Mol. Cancer Res.* **3:** 183–194.

Sanders, T.A., de Grassi, T., Miller, G.J., and Humphries, S.E. 1999. Dietary oleic and palmitic acids and postprandial factor VII in middle-aged men heterozygous and homozygous for factor VII R353Q polymorphism. *Am. J. Clin. Nutr.* **69:** 220–225.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowskis, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:** 2498–2504.

Slebos, R.J. and Rodenhuis, S. 1989. The molecular genetics of human lung cancer. *Eur. Respir. J.* **2:** 461–469.

Stelling, J. 2004. Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.* **7:** 513–518.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102:** 15545–15550.

Susztak, K., Ciccone, E., McCue, P., Sharma, K., and Bttinger, E.P. 2005. Multiple metabolic hits converge on CD36 as novel mediator of tubular epithelial apoptosis in diabetic nephropathy. *PLoS Med.* **2:** e45. doi: 10.1371/journal.pmed.0020045.

Swagell, C., Henly, D., and Morris, C.P. 2005. Expression analysis of a human hepatic cell line in response to palmitate. *Biochem. Biophys. Res. Commun.* **328:** 432–441.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* **22:** 281–285.

Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., and

Park, P.J. 2005. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci.* **102:** 13544–13549.
Todorovic-Rakovic, N. 2005. Tgf-beta 1 could be a missing link in the interplay between er and her-2 in breast cancer. *Med. Hypotheses* **65:** 546–551.
van Nimwegen, M.J. and van de Water, B. 2006. Focal adhesion kinase: A potential target in cancer therapy. *Biochem. Pharmacol.* **73:** 597–609.
van Nimwegen, M.J., Huigsloot, M., Camier, A., Tijdens, I.B., and van de Water, B. 2006. Focal adhesion kinase and protein kinase b cooperate to suppress doxorubicin-induced apoptosis of breast tumor cells. *Mol. Pharmacol.* **70:** 1330–1339.
van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveenothers, A.T., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530–536.
Vincenzi, B., Schiavon, G., Silletta, M., Santini, D., Perrone, G., Di Marino, M., Angeletti, S., Baldi, A., and Tonini, G. 2006. Cell cycle alterations and lung cancer. *Histol. Histopathol.* **21:** 423–435.
Yang, Y.H. and Speed, T. 2002. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **3:** 579–588.

Received December 11, 2006; accepted in revised form June 28, 2007.

A systems biology approach for pathway level analysis

Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, et al.

Genome Res. published online September 4, 2007
Access the most recent version at doi:10.1101/gr.6202607