scientific reports

Check for updates

OPEN

# Identifying intracellular signaling modules and exploring pathways associated with breast cancer recurrence

Xi Chen[1,2,5], Jinghua Gu[1,5], Andrew F. Neuwald[3], Leena Hilakivi-Clarke[4], Robert Clarke[4] & Jianhua Xuan[1]✉

**Exploring complex modularization of intracellular signal transduction pathways is critical to understanding aberrant cellular responses during disease development and drug treatment. IMPALA (Inferred Modularization of PAthway LAndscapes) integrates information from high throughput gene expression experiments and genome-scale knowledge databases to identify aberrant pathway modules, thereby providing a powerful sampling strategy to reconstruct and explore pathway landscapes. Here IMPALA identifies pathway modules associated with breast cancer recurrence and Tamoxifen resistance. Focusing on estrogen-receptor (ER) signaling, IMPALA identifies alternative pathways from gene expression data of Tamoxifen treated ER positive breast cancer patient samples. These pathways were often interconnected through cytoplasmic genes such as IRS1/2, JAK1, YWHAZ, CSNK2A1, MAPK1 and HSP90AA1 and significantly enriched with ErbB, MAPK, and JAK-STAT signaling components. Characterization of the pathway landscape revealed key modules associated with ER signaling and with cell cycle and apoptosis signaling. We validated IMPALA-identified pathway modules using data from four different breast cancer cell lines including sensitive and resistant models to Tamoxifen. Results showed that a majority of genes in cell cycle/apoptosis modules that were up-regulated in breast cancer patients with short survivals (< 5 years) were also over-expressed in drug resistant cell lines, whereas the transcription factors JUN, FOS, and STAT3 were down-regulated in both patient and drug resistant cell lines. Hence, IMPALA identified pathways were associated with Tamoxifen resistance and an increased risk of breast cancer recurrence. The IMPALA package is available at https://dlrl.ece.vt.edu/software/.**

A new direction[1,2] in the design of anti-cancer drug therapies is to "globally" target multiple genes involved in crosstalk among various cancer-associated signaling pathways[3] rather than the traditional approach of targeting a single molecular pathway. For example, BIRC5 intersects multiple pathways essential for cell proliferation, survival, and resistance to growth inhibition[3]. The goal is to identify anticancer drugs that interfere with multiple molecular targets in different subcellular compartments while minimizing damage to normal cells[1,4,5]. However, to be effective, such combinatorial drug design must address the complexity and heterogeneity inherent in most cancers, which, in turn, requires the development of systems biology tools to characterize multiple cancer-specific pathways and signaling networks[6]. Although there are computational methods for deciphering complex signal transduction pathways by integrating multi-platform genomic data with biological knowledge like GESA[7] and PARADIGM[8], their ability to discover novel pathway interactions is limited.

The current abundance of genome-wide protein–protein interaction (PPIs) data[9] provides an alternative source of information for signaling pathway identification, which typically has been formulated as a mathematical problem of reconstructing paths between source and target genes[10]. The main challenge for such methods—which include, for example, Netsearch[11], random color coding[12], integer linear programming (ILP)[10] and

[1]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, 900 North Glebe Road, Arlington, VA 22203, USA. [2]Center for Computational Biology, Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York, NY 10010, USA. [3]Institute for Genome Sciences and Department Biochemistry and Molecular Biology, University of Maryland School of Medicine, 670 W. Baltimore Street, Baltimore, MD 21201, USA. [4]Hormel Institute, University of Minnesota, 801 16th Ave NE, Austin, MN 55912, USA. [5]These authors contributed equally: Xi Chen and Jinghua Gu. ✉email: xuan@vt.edu

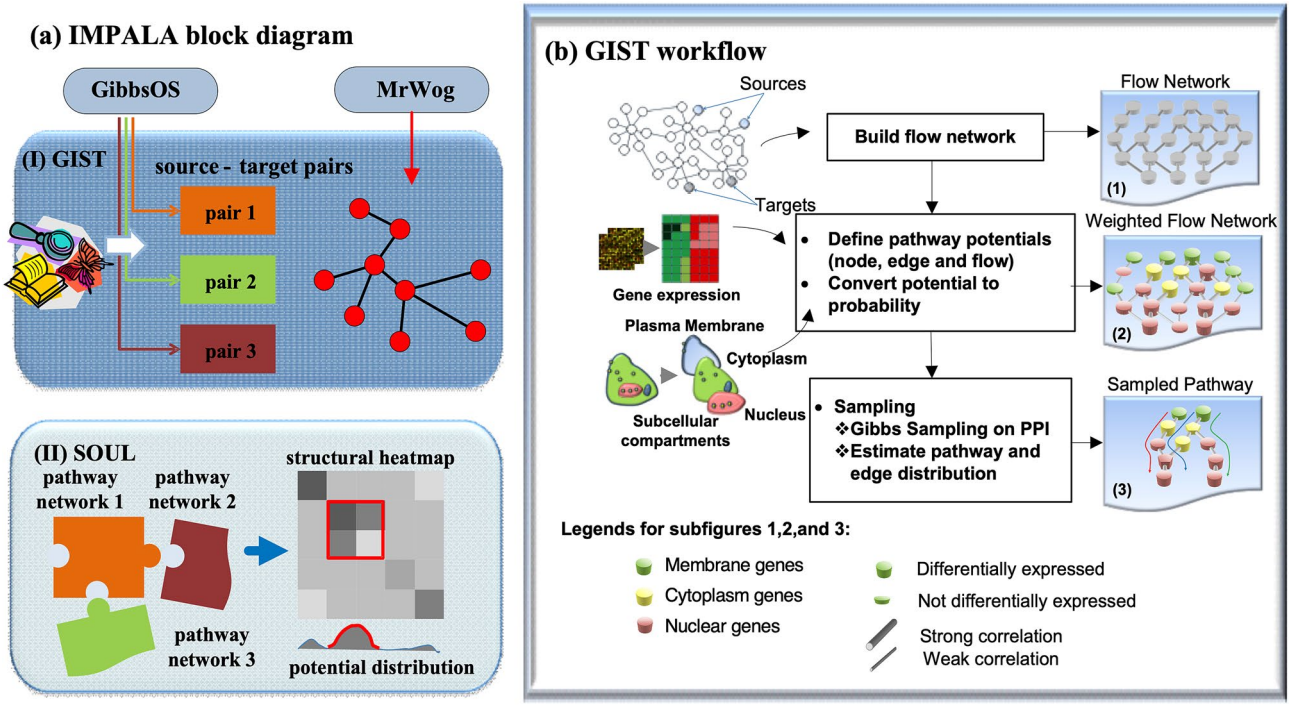


**Figure 1.** IMPALA block diagram and GIST workflow. (**a**) Key transcription factors and the candidate pathway landscape are identified using GibbsOS and MrWOG to pre-process gene expression and protein–protein interaction data (HPRD database). Then, IMPALA integrates gene expression and candidate pathways to identify aberrant signal pathway transduction using GIST (Gibbs sampler to Infer Signal Transduction) and pathway modules using SOUL (Structural Organization to Uncover pathway Landscape). (**b**) GIST integrates gene (node), gene–gene interaction (edge) and network flow potentials to build a weighted and directed Bayesian network and infers signaling directions between genes using Gibbs Sampling.

ResponseNet[13,14]—is inferring signaling directions between genes given non-directed PPI network information. Gitter et al. proposed to use maximum edge orientation (EO) on a PPI network to determine the most likely signaling directions that fulfil global optimality[15]. However, EO relies heavily on the assumption that most biological pathways are short (length < 5) in order to accommodate the requirement of exhaustive enumeration of possible pathways and fails to utilize important biological knowledge such as subcellular information. Hence, assigned signaling directions are usually difficult to interpret in a biological meaningful way. Furthermore, EO fails to jointly analyze individual pathways for structural or functional similarities, which are important for studying pathway crosstalk.

IMPALA (Inferring Modularization of PAthway LAndscape) integrates gene expression data and biological knowledge within a Bayesian framework to reconstruct aberrant pathway modules. IMPALA defines three potential functions representing gene expression, gene co-expression and prior network interactions. These functions, which jointly measure the aberrancy of individual pathways, are converted to probability distributions for pathway sampling. IMPALA estimates edge directions by aggregating pathway samples. To study crosstalk between multiple pathways, sampled pathways are clustered into interconnected modules based on structural similarities.

Here we use IMPALA to identify and explore estrogen-receptor (ER) signaling associated with Tamoxifen resistance in breast cancer and to build an aberrant pathway network connecting ER to transcription factors involved in cell proliferation and apoptosis. The identified pathway network was significantly enriched in ErbB, MAPK and JAK-STAT signaling components. Pathway clustering by IMPALA identified key functionally associated ER signaling, cell cycle and apoptosis modules with crosstalk. We validated the expression of module genes using breast cancer cell line models. Hence, IMPALA provides a novel and effective approach to investigate alternative pathways and pathway crosstalk in cancer cells.

## Results

**Identifying aberrant signaling pathway transduction in Tamoxifen-treated breast cancer patients.** IMPALA is a Bayesian approach to infer signaling pathway modules from gene expression data (Fig. 1). We applied IMPALA to a gene expression (microarray) dataset (termed Loi) including samples from Tamoxifen-treated ER positive breast cancer patients[16] and identified aberrant signal pathway transduction associated with Tamoxifen resistance. We normalized the data using PLIER (http://www.affymetrix.com), and then corrected the batch effects using ComBat[17]. A 5-year cut-off on distant-metastasis-free-survival (DMFS) was used to divide Loi samples into 'early recurrence' (DMFS ≤ 5 years) and 'late recurrence' (DMFS > 5 years) groups, yielding 88 and 92 samples, respectively.

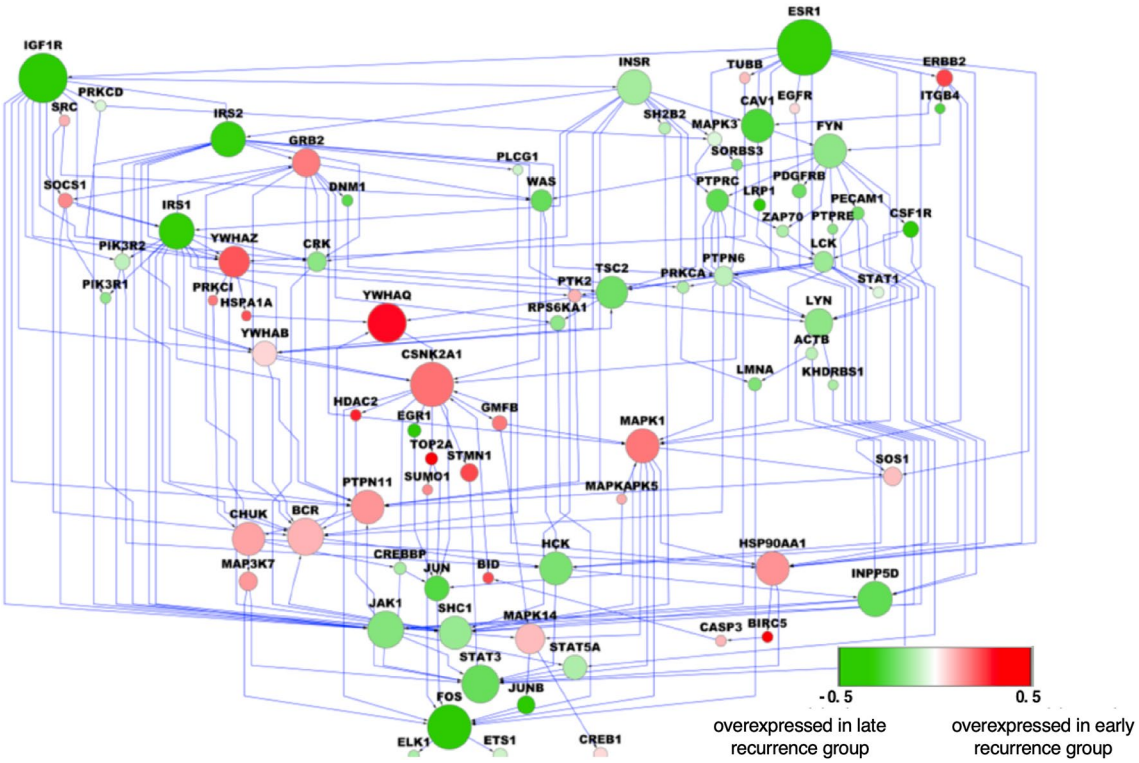


**Figure 2.** An ER signaling pathway network identified by IMPALA using Loi breast cancer gene expression data. The gene color represents the $\log_2(x)$-fold change of gene expression between early and late recurrence groups of patients in the Loi dataset (red: over-expressed in 'early recurrence' group; green: over-expressed in 'late recurrence' group). Gene's size is proportional to the probability (sampling frequency) estimated by GIST.

IMPALA utilizes two functional components: (1) Gibbs sampling to Infer Signal Transduction (GIST) and (2) Structural Organization to Uncover pathway Landscape (SOUL) (Fig. 1a). GIST reconstructs pathways (genes and directed interactions) related to ER signaling. Specifically, using MrWOG[18] a gene network was extracted from protein–protein interaction data to predict genes and interactions likely associated with ER signaling. Candidate pathways were constructed starting from the estrogen receptor ESR1 gene and targeting breast cancer-associated transcription factors, such as JUN, FOS, STAT1, STAT3, STAT5A, ELK1, and ETS1 (Target transcription factors were pre-identified by GibbsOS[19]; see Supplementary Tables S1 and S2). GIST uses a Bayesian framework to integrate candidate pathways with gene expression data and uses Gibbs sampling to iteratively infer signaling pathways (Fig. 1b).

A directed pathway network assembled by collapsing the top 200 GIST pathway samples is shown in Fig. 2. This reveals complex wiring of alternative pathways that are interconnected through frequently sampled cytoplasmic genes, such as IRS1/2, JAK1, YWHAZ, CSNK2A1, MAPK1 and HSP90AA1. Functional enrichment analysis using DAVID[20] returned, as significant, canonical insulin ($p$-value 2.4e−10), ErbB ($p$-value 4.0e−13), MAPK ($p$-value 5.1e−8), and JAK-STAT ($p$-value 2.0e−5) signaling pathways, each of which plays a key role in breast cancer[21]. We further examined the association of the pathway network with Tamoxifen recurrence by using the network to predict the survival of breast cancer patients based on a similar, but independently generated gene expression dataset (termed Symmans)[22]. Specifically, using the above ER signaling pathway network and the Loi gene expression data, we trained a NetSVM classifier[23] to group samples as early or late. Threefold cross-validation using Loi data returned the area under ROC curve (AUC) as 0.8. Applying the classifier to the Symmans dataset, which includes 103 patient samples, we obtained a prediction AUC of 0.79. Kaplan Meier analysis of Symmans data returned a hazard ratio of 3.26 ($p$-value = 0.016; Supplementary Fig. S2).

**Identifying pathway modules and crosstalk.** To study crosstalk between ER signaling and cancer cell proliferation, we further used GIST to identify cell cycle and apoptosis signaling modules (Supplementary Fig. S1). We used the SOUL component of IMPALA to analyze pooled samples from GIST and to investigate and assess the statistical significance of modules and crosstalk associated with ER, cell cycle, apoptosis signaling pathways, as shown in Fig. 3. SOUL hierarchically clustered sampled pathways based on gene overlap (Fig. 3a) and re-ordered the distribution of sampling frequency to be consistent with pathway clusters (Fig. 3b). Signaling modules were identified for each of four local peaks (modes) of the sample distribution, including two ER signaling modules (M1 and M2), one cell cycle module (M3) and one apoptosis module (M4). The specific genes in each module are listed in Supplementary Table S3. A pathway network of the four modules is shown in Fig. 3c. M1 is enriched with genes in response to hormones and also enriched with canonical MAPK and insulin signaling pathways. M2 corresponds to JAK-STAT signaling. The crosstalk between M3 and M4 is strong, as indicated

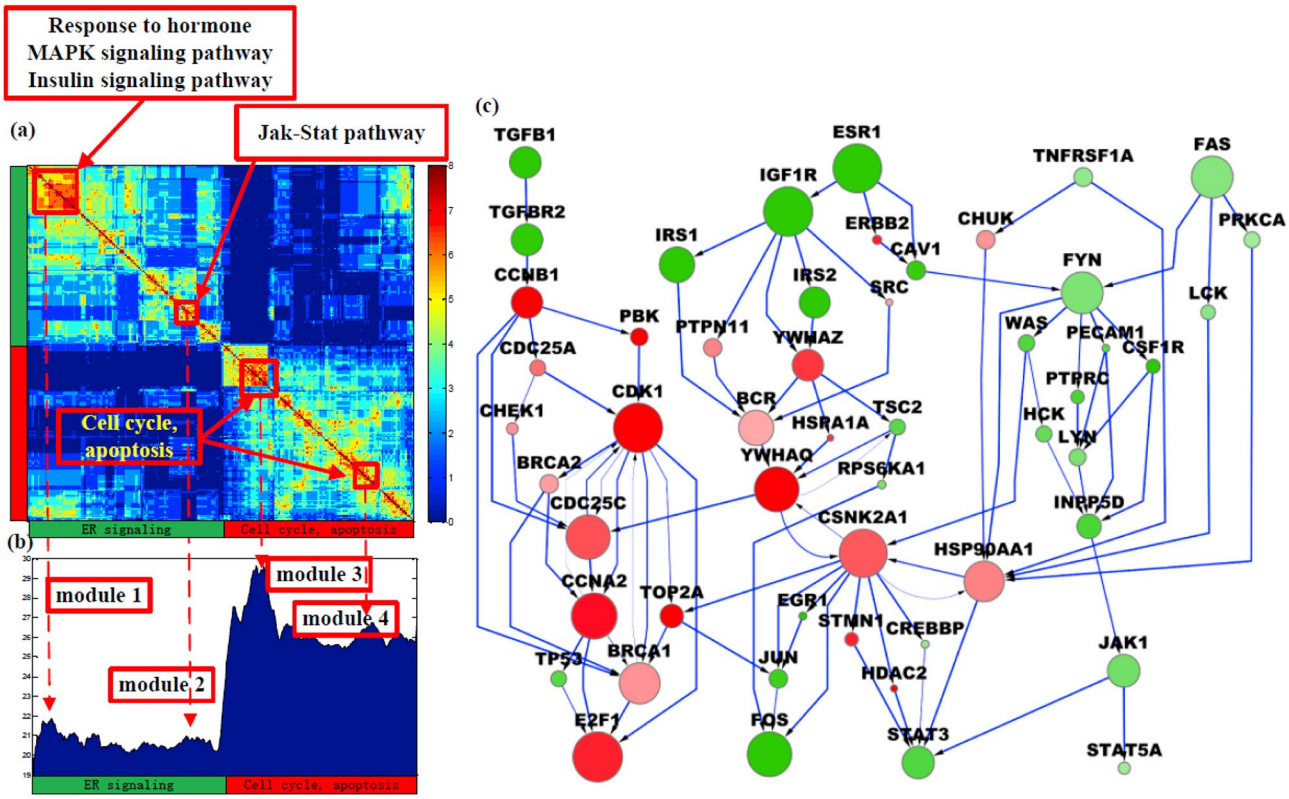


**Figure 3.** Pathway modules and crosstalk identified by IMPALA for the Loi dataset. (**a**) Pathway clustering based on gene similarity and gene functions in different clusters reveal the functional diversity of IMPALA-identified pathways. (**b**) Distribution of sampling frequency of pathways with peaks corresponding to major pathway clusters in (**a**). Four pathway modules were identified. (**c**) A combined pathway network consisting of the four modules with crosstalk.

by the pathway sample distribution. Although M4 contains genes functioning in apoptosis and cell death, it is also enriched with cell cycle genes, which suggests coupling of these cellular processes.

Genes upregulated in 'early recurrence' samples (survival ≤ 5 years) include signal transduction genes like YWHAQ, YWHAZ and PTPN11, the chaperone HSP90AA1, and STMN1, which functions in cytoskeletal rearrangements. HSP90AA1 is an intracellular gene that is actively expressed in breast cancer cells—high levels of which correlate with a low chance of survival[24]. Efficient progression through the cell cycle requires HSP90AA1[25]; when up-regulated in osteosarcoma it increases drug resistance by inducing autophagy and inhibiting apoptosis[26]. BRCA1 is a client gene of HSP90AA1, inhibition of which by 17-AAG Tanespimycin leads to degradation of BRCA1 via the ubiquitin–proteasome pathway. Subsequent loss of BRCA1 disrupts G2/M cell cycle checkpoint activation, resulting in mitotic catastrophe—an apoptosis-independent form of cell death caused by mechanical damage[27]. Thus, HSP90AA1 inhibition may promote survival in Tamoxifen-resistant tumors. STMN1 promotes catastrophes that ultimately lead to deregulation of the cell cycle, thereby hampering cell survival[28]. High STMN1 expression leads to shorter post-progression and overall survival in breast cancer patients[29], consistent with our finding that STMN1 is up-regulated among tumor samples in the 'early recurrence' group (labelled 'red' in Fig. 3c). CDK1 is an essential modulator of the initiation of and progression through mitosis, acting primarily through its interaction with CCNB1. CDK1 and CCNB1 help protect mitotic cells against extrinsic death stimuli[30]. Thus, increased expression of CDK1 in early recurrence breast cancer may explain Tamoxifen resistance by protecting tumor cells from antiestrogen-mediated cell death.

We found ESR1 and IGF1R to be overexpressed in the 'late recurrence' group ('green' hub genes in Fig. 3c). Crosstalk between the IGF and ER signaling pathways is well known[31]. TSC2 is a negative regulator of mTOR, which in turn inhibits autophagy. Although cellular stress from therapeutic drugs can induce cell death via autophagy, lysosomal degradation or prolonged stress[32] can sustain long-term survival or dormancy by enabling autophagy of some tumor cells[33].

**Validating pathways and modules using Symmans breast cancer gene expression data.** To validate the robustness of IMPALA for characterizing networks associated with Tamoxifen resistance in breast cancer, we applied it to the Symmans dataset[22] (Tamoxifen treated breast cancer gene expression (microarray) dataset; consisting of 47 'early recurrence' and 56 'late recurrence' samples based on a 5-year DMFS cutoff). Source receptor genes were the same as selected for the Loi data analysis, while target transcription factors were identified using GibbsOS for ER signaling, cell cycle, and apoptosis (Supplementary Table S4). Pathway networks of the top GIST-sampled pathways for ER signaling and for cell cycle and apoptosis are shown in Fig. 4 and in

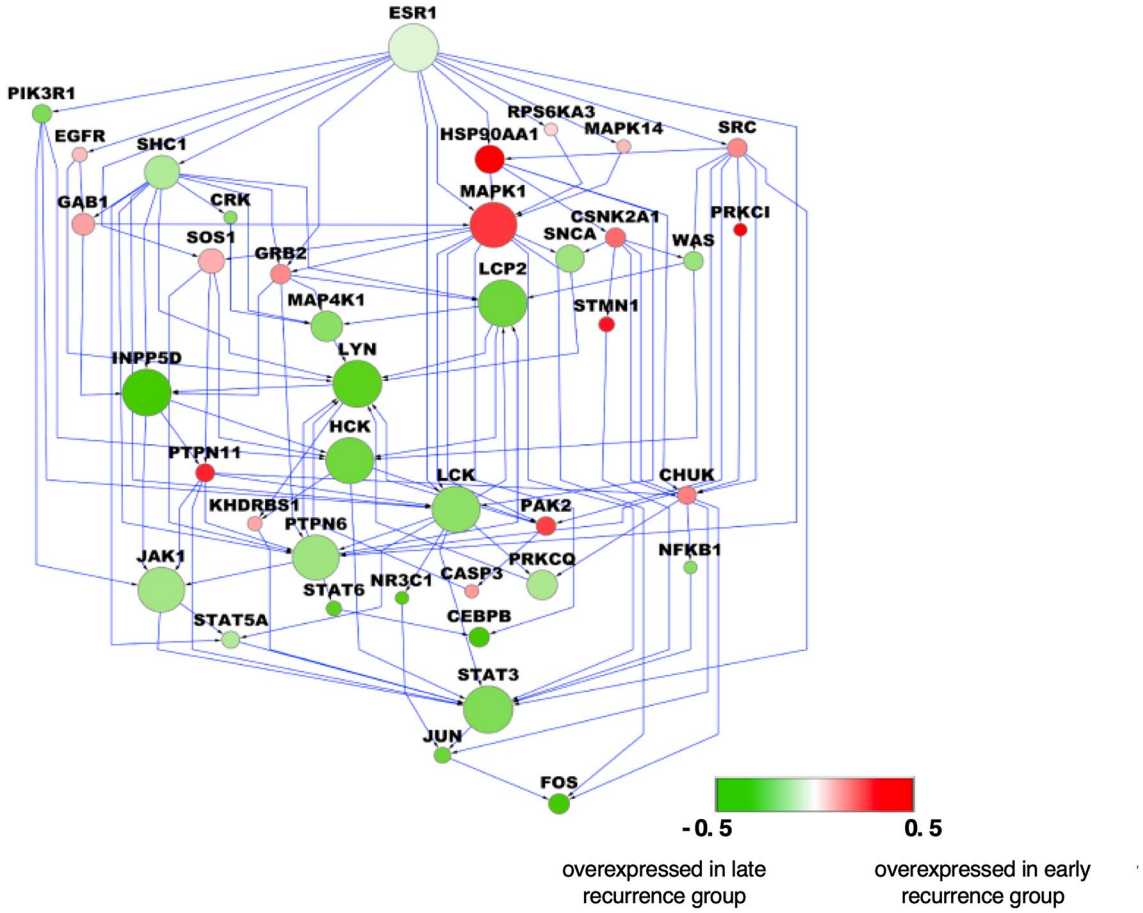


**Figure 4.** An ER signaling pathway network identified by IMPALA using Symmans data. Gene colors represent the log2 fold change of gene expression between 'early recurrence' and 'late recurrence' patients in the Symmans dataset (red: over-expressed in 'early recurrence' group; green: over-expressed in 'late recurrence' group). Gene size is proportional to the probability (sampling frequency) estimated by GIST.

Supplementary Fig. S3, respectively. The similarity to genes in the Loi-based pathway networks for ER, cell cycle, and apoptosis signaling were 73%, 53% and 54%, respectively.

SOUL identified the four pathway modules (M1-M4) shown in Fig. 5. Specific genes in each module are listed in Supplementary Table S5. Again, we observed signal transductions from the membrane through cytoplasmic genes MAPK1, HSP90AA1, and CSNK2A1 to the nuclear transcription factors. In M1, signal pathways started from IGFR1 and INSR, passed through cytoplasmic signaling hubs SRC, CHUK, and HSP90AA1, and converged to the same targets within the nucleus. In M2 and M4, signal transduction took diverse pathways between membrane receptors and JAK-STAT activation. Signaling could be initiated by ESR1 via canonical members of the JAK-STAT pathway (PIK3R1, SOS1, and PTPN6), by various membrane receptors (INSR, EGFR), or by death receptors (FAS, TNFRSF1A) through PTPN6, SHC1, or LYN. Although M3 genes are mostly shared with M2 and M4, they form an alternative pathway for cell cycle progression genes (CDC2 and E2F1). Based on IMPALA pathway analyses of both the Loi and Symmans datasets, we conclude that HSP90AA1, CSNK2A1, and MAPK1 play key topological roles in intracellular signal transduction initiated by plasma membrane genes or canonical death receptors to regulate the cell cycle and apoptosis.

**Validating pathway gene expression in breast cancer cell line models.** We used in vitro breast cancer cell line models to validate the expression of genes in aberrant pathway modules identified by IMPALA. Four MCF7 derived cell models were included in the analysis: MCF7-STR, MCF7RR-STR, LCC1, and LCC2[34]. MCF7RR-STR and LCC2 are Tamoxifen resistant, whereas MCF7-STR and LCC1 are sensitive. As shown in Fig. 6, 20 genes from IMPALA-identified pathway modules exhibited consistent expression patterns between patient data and cell line data. ER signaling genes, such as STMN1, PBK, CCNB1 and HSP90AA1, were over-expressed in early recurrence/resistant groups, whereas IRS1, IRS2, IGF1R and TSC2 were overexpressed in the 'late recurrence/drug-sensitive' groups. The cell cycle/apoptosis genes BRCA1, BRCA2, CCNA2, E2F1, CDC25A, CDC25C, TOP2A, CDC2, and CHUK were up-regulated in the 'early recurrence' group and also in the Tamoxifen resistant cell lines, whereas the transcription factors JUN, FOS, and STAT3 were down-regulated. Gene expression for in vitro cell lines identified from Loi and Symmans datasets are shown in Supplementary

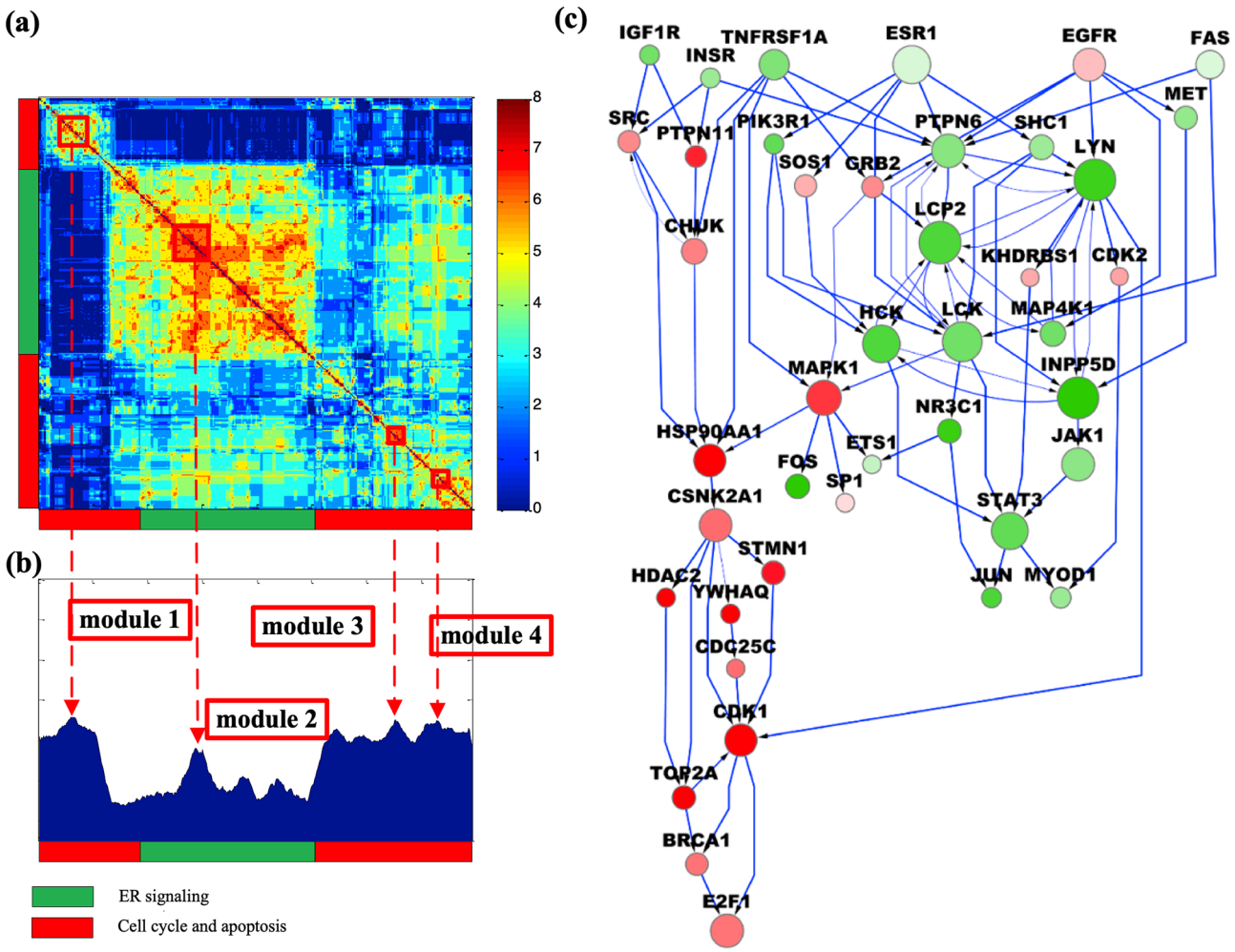


**Figure 5.** Pathway modules and crosstalk identified by IMPALA using the Symmans data. (**a**) Pathway clustering based on gene similarity and gene functions in different clusters reveal the functional diversity of IMPALA-identified pathways. (**b**) Distribution of sampling frequency of pathways with peaks corresponding to major pathway clusters in (a). Four pathway modules were identified. (**c**) A combined pathway network consisting of the four modules with crosstalk.

Figures S4 and S5, respectively. The concordance between patient and cell line data demonstrates the association of IMPALA identified pathways with Tamoxifen resistance and with increased breast cancer recurrence.

## Discussion

IMPALA characterizes intracellular signal transduction pathways by integrating multi-platform data and by identifying crosstalk among pathways. Using this approach, we identified breast cancer-associated aberrant pathways by integrating breast cancer gene expression data with protein-DNA and protein–protein interaction data, and with published information regarding signaling pathways.

IMPALA has several notable advantages over existing methods. First, GIST allows users to incorporate the subcellular location of genes in order to focus on signal transduction components in the nucleus, the cytoplasm, or the plasma membrane. Second, most existing methods either fail to assign signaling directions between genes or else infer signaling direction in an ad hoc manner. GIST assigns a posterior probability for each signaling direction, thereby estimating a degree of confidence. Third, SOUL models network components as structurally related modules to better identify local modules within a large-scale pathway landscape. This identifies overlap between modules, which corresponds to crosstalk between pathways.

Unravelling signaling pathways from complex molecular networks in cancer cells is challenging[35]. Here, IMPALA revealed that breast cancer-associated pathway modules are structurally interconnected with cross-talk between ER signaling, cell cycle and apoptosis pathways, thereby imparting tamoxifen resistance. And, by characterizing the pathway landscape, IMPALA systematically categorized complex pathway interactions into within-module and between-module interactions. This echoes the increasing emphasis among researchers on networks, rather than pathways, as a reflection of the complex and integrated nature of molecular signaling.

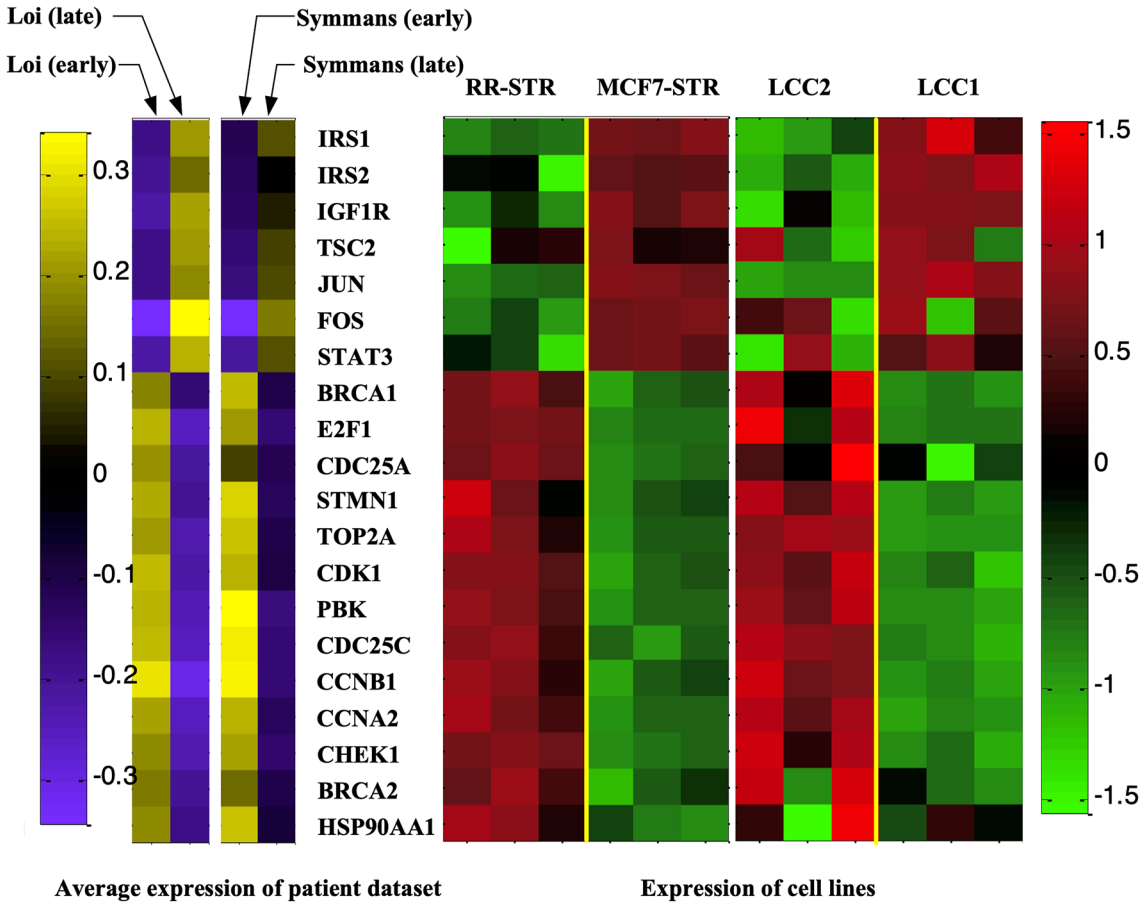


**Figure 6.** Cell line validation for identified pathway genes from patient datasets. The left panel shows the average $\log_2$ expression of selected pathway genes. The right panel shows the $\log_2$ expression of two cell line studies: (MCF7-STRP vs. MCF7RR-STRP and LCC1 vs. LCC2). Seven genes (IRS1, IRS2, IGF1R, TSC2, JUN, FOS, STAT3) are consistently over-expressed in the 'early recurrence' patient samples and sensitive human breast cancer cell lines. The remaining genes, which mainly relate to cell cycle and apoptosis, are over-expressed in the resistant groups.

## Methods

IMPALA applies GIST to identify signaling pathways by integrating gene expression data with protein–protein interactions (PPIs), and SOUL to explore the pathway landscape for pathway module and crosstalk identification.

**Identifying source and target genes for pathway exploration.** To build the candidate pathway landscape, we pre-selected the source and target genes for each signaling pathway. Specifically, we selected ESR1 for ER signaling, membrane receptors and the growth factors EGFR, TGFB1, IGF1R, INSR, FGFR1 for cell cycle, and canonical death receptors IL1R1, FAS, and TNFRSF1A for apoptosis. Based on literatures, we selected transcription factors associating to breast cancer recurrence as pathway targets. Categorized transcription factors selected for ER signaling, cell cycle, and apoptosis are listed in Supplementary Table S1. To refine the candidate target genes, we applied GibbsOS[36] to the Loi and Symmans datasets, respectively, and selected transcription factors significantly associated with the survival difference, as listed in Supplementary Tables S2 and S4.

**Building the candidate pathway landscape using MRWOG.** To build a candidate pathway landscape, we used MRWOG[18] to pre-screen human PPIs for an ER-related, Tamoxifen resistant sub-network. An ESR1-centered PPI subnetwork including 2326 genes (all genes within a two-step distance from ESR1) was selected.

**The GIST algorithm.** To infer signal directions between genes, GIST constructs a flow network of a given pathway length between source and target genes. To weight the flow network, node (gene), edge (interaction) and flow (network) potentials are defined for individual pathways. GIST converts the three potentials into a joint probability distribution so that samples of candidate pathways can be drawn probabilistically. Signaling pathway directions were inferred by aggregating the pathways samples and then selecting the interconnected linear pathways with the largest potentials.

We define a vector $\boldsymbol{\theta}_{1\times L} = \{\theta_1, \theta_2, \ldots, \theta_L\}$ to represent a linear pathway with length $L$ genes, where $\theta_i$ is a categorical variable representing the $i$th gene in the pathway. $\theta_1$ and $\theta_L$ are the source and target genes, respectively. Let $\Omega_i$ denote the domain of $\theta_i$ and we have $\Omega_1 \bigcap \Omega_2 \bigcap \cdots \bigcap \Omega_L \subseteq \Omega$, where the full domain $\Omega$ denotes the whole set of genes in the PPI dataset. Given gene expression data $\mathbf{X}_{n\times m}$, which includes $n$ genes and $m$ samples with two conditions (to study aberrant signal pathway transduction between conditions), we derive gene potential $\mathrm{V}_1(\theta_i;\ \mathbf{X})$, defined as the sum of pathway gene differential expression z-scores between the two types[37]; edge potential $\mathrm{V}_2(\theta_i,\ \theta_{i+1};\ \mathbf{X})$, defined as the sum of z-scores calculated from the statistical significance of Pearson's correlation between interacting genes[38]; and flow potential $\mathrm{V}_3(\boldsymbol{\theta})$, defined as a proportionally score reflecting the concordance between a pathway and prior information regarding cellular location[39]. Derivations of the three potentials are provided in the Supplementary Methods.

GIST integrates the three potentials into a pathway energy function as follows:

$$\mathrm{U}(\boldsymbol{\theta};\ \mathbf{X}) = \sum_{i=1}^{L} \mathrm{V}_1(\theta_i;\ \mathbf{X}) + \sum_{i=1}^{L-1} \mathrm{V}_2(\theta_i,\ \theta_{i+1};\ \mathbf{X}) + \mathrm{V}_3(\boldsymbol{\theta}).$$

Due to the large number of genes and their interactions, finding the optimal solution of Eq. (1) is a NP hard problem. Therefore, we convert the optimization task into a distribution learning problem as show in Eq. (2) and used Gibbs sampling to search for the optimal solution.

$$\begin{aligned} P(\boldsymbol{\theta};\ \mathbf{X}) &= \frac{1}{Z}\cdot e^{\frac{-\mathrm{S}(\boldsymbol{\theta};\ \mathbf{X})}{T}} = \frac{1}{Z}\cdot e^{\frac{\mathrm{U}(\boldsymbol{\theta};\ \mathbf{X})}{T}} \\ &= \frac{1}{Z}\cdot \exp\left(\frac{\sum_{i=1}^{L} \mathrm{V}_1(\theta_i;\ \mathbf{X}) + \sum_{i=1}^{L-1} \mathrm{V}_2(\theta_i,\ \theta_{i+1};\ \mathbf{X}) + \mathrm{V}_3(\boldsymbol{\theta})}{T}\right), \end{aligned} \tag{2}$$

where $Z = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} e^{\frac{1}{T}\mathrm{U}(\boldsymbol{\theta};\mathbf{X})}$ is a partition function and $T$ is the "temperature" that controls the shape of the distribution. GIST samples pathway genes iteratively from a conditional distribution as $\theta_i^{(t+1)} \sim P(\theta_i|\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)},\ \theta_{i+1}^{(t)},\ \ldots \theta_L^{(t)};\ \mathbf{X})$. In each iteration, it probabilistically samples $\theta_i$ conditioned on the other, currently assigned genes $\theta_{-i}$ in the pathway. After the sampler appears to have converged to a stationary distribution, GIST accumulates samples from this conditional distribution to approximate the posterior distribution. Details about GIST sampling are provided in Supplementary Methods, Figures S7 and S8.

After 10,000 iterations, GIST pools the pathway samples and then estimates edge directions. We introduce a binary variable $e_{i,j}$ to denote the signaling direction from gene $\omega_i \in \Omega$ to gene $\omega_j \in \Omega$. The probability of $e_{i,j}$ is estimated as follows:

$$p^*_{i,j} = P(e_{i,j} = 1) = \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} P(e_{i,j} = 1|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{3}$$

where $P(e_{i,j} = 1|\boldsymbol{\theta}) = 1$ if $e_{i,j}$ corresponds to a connected edge in pathway $\boldsymbol{\theta}$; otherwise it equals 0. Using Eq. (3), GIST models each directed edge as a Bernoulli random variable with success rate $p_{i,j}$. It performs both forward and reverse searching so that the probabilities of edge direction from gene $i$ to gene $j$ and its reverse direction are both estimated (Supplementary Methods, Fig. S6). If $p_{i,j}$ is close to 1, the signal flows from gene $\omega_i$ to gene $\omega_j$ with high confidence, while $p_{i,j}$=0.5 indicates a lack of confidence in the direction of signal flow.

**The SOUL algorithm.** SOUL post-processes distributions of GIST pathway samples to reconstruct the overall landscape. Given thousands of genes, the pathway sample distribution can be multi-modal and some hub genes (i.e., those involved in multiple pathways more often than others) could bias the sample distribution. Instead of directly ranking pathways based on their GIST sampling frequency, SOUL first clusters pathway samples based on their structural similarities using hierarchical clustering, resulting in a re-organized pathway topological pattern visualized as a pathway structural heatmap (as in Fig. 3a). Next, SOUL re-orders the pathway sampling frequencies to be consistent with pathway clusters (as in Fig. 3b). Finally, it identifies high-confidence pathway modules from local peaks in the pathway sampling frequency distribution.

**IMPALA performance evaluation on simulated data.** We evaluated the performance of GIST for pathway identification on simulated datasets generated by two different pathway structures: type I, corresponding to alternative pathways between a single source gene and a single target gene; and type II, corresponding to multiple pathways with crosstalk among multiple sources and targets (Supplementary Fig. S9). PPI data from the HPRD database[40,41] and canonical pathways from the KEGG database[42] were used to simulate pathways that include 261 genes and 998 interactions for type I, and 266 genes and 1026 interactions for type II. We added noise to gene expression data (Gaussian distributed noise with zero-mean and variance varying from 0.2 to 0.8, compared to the gene expression data) and to simulated pathway networks (false gene interactions varying from 10 to 50%, compared to the 'true' interactions).

Supplementary Figures S10–S13 and Tables S6 and S7 summarize the performance of IMPALA versus three competing algorithms: random color coding[12], edge orientation[15], and integer linear programming (ILP)[10]. Note that we only applied ILP to pathway gene identification because ILP does not infer signaling directions. IMPALA consistently obtained comparable or better performance in all cases. When the level of noise was set to 0.2 (20% false interactions in the network), IMPALA gained about a 16% increase in precision for type I pathway gene

identification, and an even larger improvement of 24% for edge identification. Similarly, for type II GIST achieved about a 15% increase in average precision for gene identification, and a 17% increase for edge identification.

Received: 12 September 2020; Accepted: 18 November 2020
Published online: 11 January 2021

## References

1. Kang, B. H. *et al.* Combinatorial drug design targeting multiple cancer signaling networks controlled by mitochondrial Hsp90. *J. Clin. Investig.* **119**, 454–464. https://doi.org/10.1172/JCI37613 (2009).
2. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847. https://doi.org/10.1038/ng.3593 (2016).
3. Altieri, D. C. Survivin, cancer networks and pathway-directed drug discovery. *Nat. Rev. Cancer* **8**, 61–70. https://doi.org/10.1038/nrc2293 (2008).
4. Kang, B. H. & Altieri, D. C. Compartmentalized cancer drug discovery targeting mitochondrial Hsp90 chaperones. *Oncogene* **28**, 3681–3688. https://doi.org/10.1038/onc.2009.227 (2009).
5. Rajendran, L., Knolker, H. J. & Simons, K. Subcellular targeting strategies for drug design and delivery. *Nat. Rev. Drug Discov.* **9**, 29–42. https://doi.org/10.1038/nrd2897 (2010).
6. Melas, I. N. *et al.* Identification of drug-specific pathways based on gene expression data: Application to drug induced lung injury. *Integr. Biol. (Camb)* **7**, 904–920. https://doi.org/10.1039/c4ib00294f (2015).
7. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550. https://doi.org/10.1073/pnas.0506580102 (2005).
8. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-245. https://doi.org/10.1093/bioinformatics/btq182 (2010).
9. Szklarczyk, D. *et al.* STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613. https://doi.org/10.1093/nar/gky1131 (2019).
10. Zhao, X. M., Wang, R. S., Chen, L. & Aihara, K. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res.* **36**, e48. https://doi.org/10.1093/nar/gkn145 (2008).
11. Steffen, M., Petti, A., Aach, J., D'Haeseleer, P. & Church, G. Automated modelling of signal transduction networks. *BMC Bioinform.* **3**, 34 (2002).
12. Scott, J., Ideker, T., Karp, R. M. & Sharan, R. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol.* **13**, 133–144. https://doi.org/10.1089/cmb.2006.13.133 (2006).
13. Lan, A. *et al.* ResponseNet: Revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res.* **39**, W424-429. https://doi.org/10.1093/nar/gkr359 (2011).
14. Yeger-Lotem, E. *et al.* Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* **41**, 316–323. https://doi.org/10.1038/ng.337 (2009).
15. Gitter, A., Klein-Seetharaman, J., Gupta, A. & Bar-Joseph, Z. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res.* **39**, e22. https://doi.org/10.1093/nar/gkq1207 (2011).
16. Loi, S. *et al.* Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* **9**, 239. https://doi.org/10.1186/1471-2164-9-239 (2008).
17. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127. https://doi.org/10.1093/biostatistics/kxj037 (2007).
18. Wang, C. From network to pathway: Integrative network analysis of genomic data. *Virginia tech PhD dissertation* (2011).
19. Stecklein, S. R. *et al.* BRCA1 and HSP90 cooperate in homologous and non-homologous DNA double-strand-break repair and G2/M checkpoint activation. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13650–13655. https://doi.org/10.1073/pnas.1203326109 (2012).
20. da Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. https://doi.org/10.1038/nprot.2008.211 (2009).
21. Eroles, P., Bosch, A., Perez-Fidalgo, J. A. & Lluch, A. Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Cancer Treat. Rev.* **38**, 698–707. https://doi.org/10.1016/j.ctrv.2011.11.005 (2012).
22. Symmans, W. F. *et al.* Genomic index of sensitivity to endocrine therapy for breast cancer. *J. Clin. Oncol.* **28**, 4111–4119. https://doi.org/10.1200/JCO.2010.28.4273 (2010).
23. Chen, L., Xuan, J., Riggins, R. B., Clarke, R. & Wang, Y. Identifying cancer biomarkers by network-constrained support vector machines. *BMC Syst. Biol.* **5**, 161. https://doi.org/10.1186/1752-0509-5-161 (2011).
24. Liu, K. *et al.* BJ-B11, an Hsp90 inhibitor, constrains the proliferation and invasion of breast cancer cells. *Front. Oncol.* **9**, 1447. https://doi.org/10.3389/fonc.2019.01447 (2019).
25. Pfeiffer, J., Tarbashevich, K., Bandemer, J., Palm, T. & Raz, E. Rapid progression through the cell cycle ensures efficient migration of primordial germ cells—The role of Hsp90. *Dev. Biol.* **436**, 84–93. https://doi.org/10.1016/j.ydbio.2018.02.014 (2018).
26. Xiao, X. *et al.* HSP90AA1-mediated autophagy promotes drug resistance in osteosarcoma. *J. Exp. Clin. Cancer Res.* **37**, 201. https://doi.org/10.1186/s13046-018-0880-6 (2018).
27. Fragkos, M. & Beard, P. Mitotic catastrophe occurs in the absence of apoptosis in p53-null cells with a defective G1 checkpoint. *PLoS ONE* **6**, e22946. https://doi.org/10.1371/journal.pone.0022946 (2011).
28. Cassimeris, L. The oncoprotein 18/stathmin family of microtubule destabilizers. *Curr. Opin. Cell Biol.* **14**, 18–24. https://doi.org/10.1016/s0955-0674(01)00289-7 (2002).
29. Obayashi, S. *et al.* Stathmin1 expression is associated with aggressive phenotypes and cancer stem cell marker expression in breast cancer patients. *Int. J. Oncol.* **51**, 781–790. https://doi.org/10.3892/ijo.2017.4085 (2017).
30. Matthess, Y., Raab, M., Sanhaji, M., Lavrik, I. N. & Strebhardt, K. Cdk1/cyclin B1 controls Fas-mediated apoptosis by regulating caspase-8 activity. *Mol. Cell Biol.* **30**, 5726–5740. https://doi.org/10.1128/MCB.00731-10 (2010).
31. Fagan, D. H., Uselman, R. R., Sachdev, D. & Yee, D. Acquired resistance to tamoxifen is associated with loss of the type I insulin-like growth factor receptor: Implications for breast cancer treatment. *Cancer Res.* **72**, 3372–3380. https://doi.org/10.1158/0008-5472.CAN-12-0684 (2012).
32. Mizushima, N., Levine, B., Cuervo, A. M. & Klionsky, D. J. Autophagy fights disease through cellular self-digestion. *Nature* **451**, 1069–1075. https://doi.org/10.1038/nature06639 (2008).
33. Clarke, R. *et al.* Endoplasmic reticulum stress, the unfolded protein response, autophagy, and the integrated regulation of breast cancer cell fate. *Cancer Res.* **72**, 1321–1331. https://doi.org/10.1158/0008-5472.CAN-11-3213 (2012).
34. Clarke, R., Leonessa, F., Welch, J. N. & Skaar, T. C. Cellular and molecular pharmacology of antiestrogen action and resistance. *Pharmacol. Rev.* **53**, 25–71 (2001).
35. Hill, S. M. *et al.* Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods* **13**, 310–318. https://doi.org/10.1038/nmeth.3773 (2016).

36. Gu, J. *et al.* Robust identification of transcriptional regulatory networks using a Gibbs sampler on outlier sum statistic. *Bioinformatics* **28**, 1990–1997. https://doi.org/10.1093/bioinformatics/bts296 (2012).
37. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18**(Suppl 1), S233–S240. https://doi.org/10.1093/bioinformatics/18.suppl_1.s233 (2002).
38. Fieller, E. C., Hartley, H. O. & Pearson, E. S. Tests for rank correlation coefficients. *Biometrica* **44**, 470–481 (1957).
39. Gu, J. *et al.* GIST: A Gibbs sampler to identify intracellular signal transduction pathways. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2434–2437**, 2011. https://doi.org/10.1109/IEMBS.2011.6090677 (2011).
40. Mathivanan, S. *et al.* Human Proteinpedia enables sharing of human protein data. *Nat. Biotechnol.* **26**, 164–167. https://doi.org/10.1038/nbt0208-164 (2008).
41. Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinform.* **7**(Suppl 5), S19. https://doi.org/10.1186/1471-2105-7-S5-S19 (2006).
42. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114. https://doi.org/10.1093/nar/gkr988 (2012).


## Acknowledgements

This work is supported by National Institutes of Health (NIH) [CA149653, CA164384, CA149147 and GM125878].

## Author contributions

J.X. conceived the idea of the method. J.G. and X.C. implemented the algorithm and performed the experiments. J.G., X.C. and J.X. wrote the manuscript. A.F.N., L.H.-C., R.C. and J.X. revised the manuscript. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79603-5.

**Correspondence** and requests for materials should be addressed to J.X.

**Reprints and permissions information** is available at www.nature.com/reprints.

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