PHENSIM: Phenotype Simulator

Salvatore Alaimo[1], Gioacchino P. Marceca[1], Alessandro La Ferlita[1,2], Oksana B. Serebrennikova[3], Philip N. Tsichlis[4], Alfredo Pulvirenti[1] and Alfredo Ferro[1*]

Affiliations:

[1]Bioinformatics Unit, Department of Clinical and Experimental Medicine, University of Catania, Viale A. Doria 6 c/o Department of Mathematics and Computer Science, I-95125, Catania, Italy

[2]Department of Physics and Astronomy, University of Catania, Viale A. Doria 6, I-95125, Catania, Italy

[3]Molecular Oncology Research Institute, Tufts Medical Center, 800 Washington Street, Boston, MA 02111

[4]Department of Cancer Biology and Genetics and the James Comprehensive Cancer Center, Ohio State University, Columbus, OH 43210

*To whom correspondence should be addressed:

Prof. Alfredo Ferro

Bioinformatics Unit

Department of Clinical and Experimental Medicine

University of Catania

Viale A. Doria 6

c/o Department of Mathematics and Computer Science

I-95125, Catania, Italy

Phone: +39 095 738 3087

Email: ferro@dmi.unict.it

Keywords: Pathway-based simulations, Pathway Analysis, Bioinformatics

Abstract

Motivation: Despite the unprecedented increase of our understanding of cell biology, connecting experimental data to the physiopathological status of cells and tissues under precise circumstances is still a challenge. This often results in difficulties during the design of validation experiments, which are usually labor-intensive, expensive to perform and hard to interpret.

Results: Here we propose PHENSIM, a systems biology approach, which can simulate the effects of activation/inhibition of one or multiple biomolecules on cell phenotypes by exploiting signaling pathways. Possible applications of our tool include prediction of the outcome of drug administration, knockdown experiments, gene transduction and exposure to exosomal cargo. Importantly, this method allows the user to make inferences on well-defined cell lines and includes pathway maps from three different model organisms. The basic assumption is that phenotypes can be described through changes in pathway activity. Results from our study show discrete prediction accuracy, highlighting the capabilities of this methodology.

Availability and Implementation: PHENSIM has been developed in Java and it is available, along with all data and source codes, at https://github.com/alaimos/phensim. A web-based user interface, developed in PHP is accessible at https://phensim.atlas.dmi.unict.it/.

Introduction

Cells of living organisms are continuously exposed to signals originating in both the extracellular and the intracellular microenvironment. These signals regulate multiple cellular functions, including gene expression, chromatin remodeling, DNA replication and repair, protein synthesis and cellular metabolism. The proper response to signals depends on expression, activation or inhibition of sets of interrelated genes/proteins, acting in a well-defined order within the context of well-defined vector-driven biological processes, aiming to reach specific endpoints. Such sub-cellular processes are referred to as biological pathways.[1]

In this context, the study of genome and transcriptome, the definition of protein–protein interaction networks and association studies between gene sets and defined molecular mechanisms in humans has produced valuable biological information. In particular, the development of RNA/protein profiling methods and high-throughput sequencing/detection techniques[2–4] has provided a powerful approach for discovering genetically-controlled biological mechanisms and for understanding the downstream effects of perturbations of these mechanisms in disease and in response to drugs.[3,5,6] However, despite the improvements in our understanding of cell biology, it is difficult to connect omics data to the physiopathological status of cells, tissues or organs under specific circumstances. In addition, studies addressing these issues are often labor-intensive and expensive to perform and produce huge datasets for analysis.

In recent times, systems biology computational approaches have emerged as efficient means capable of bridging the gap between experimental biology at the system-level and quantitative sciences, giving the opportunity to derive biological functions and properties starting from transcriptomic/proteomic data[7]. Such computational methods can be used as time- and cost-saving solutions, that allow efficient *in silico* predictions that have the potential to facilitate, or enhance, the design of molecular-pharmacological experiments[7,8] by helping scientists to discard inappropriate approaches. Obviously, the choice among different approaches depends on several factors, such as the question to be addressed and the type and availability of data to be analyzed.

Pathway analysis is a widely used systems biology approach for gaining insights on the biology of differentially expressed/activated genes and proteins. This approach has the advantage of reducing biological complexity by grouping wide sets of individual genes/proteins and placing them into pathways. By doing so, it has the potential to identify pathways whose activity differ between two conditions[9,10].

To address these challenges, we developed PHENSIM (PHENotype SIMulator), a web-based, user-friendly tool allowing phenotype predictions on selected cell lines or tissues in three different

organisms: *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. The main idea behind our method is the representation of the cellular phenotype through biological pathways and their relative endpoints. PHENSIM uses a randomized algorithm to compute the effect of dysregulated genes, proteins, microRNAs (miRNAs), and metabolites on KEGG pathways. The results of such calculations are summarized through an Activity Score, which represents an index of both the predicted effect of a gene dysregulation on a downstream node (up- or down-regulation) and its likelihood, compared with a proper null model. An Activity Score is also calculated at the pathway-level. Moreover, to achieve greater accuracy, PHENSIM performs all calculations in the KEGG meta-pathway, obtained by merging KEGG pathways after elimination of duplicates and disease pathways[11] (see Methods), and integrates information on miRNA-target and transcription factor (TF)-miRNA extracted from online public knowledge bases[11]. We implemented our tool as a freely accessible web application at the following URL: https://phensim.atlas.dmi.unict.it/

Methods

Overview of the method

PHENSIM is a randomized algorithm for computing the effect of (up/down) de-regulated genes, metabolites, or microRNAs on KEGG pathways. The results are synthesized as an Activity Score estimated for each element in a pathway. The sign of this score indicates the predicted effect on the node: positive for activation, negative for inhibition. Its value represents the log-likelihood of the effect compared to a null model. PHENSIM performs all calculations using the meta-pathway concept introduced in Alaimo et al.[27].

The algorithm underlying our method consists of 5 main phases: (i) meta-pathway construction; (ii) calculation of activation or inhibition probability for each meta-pathway node; (iii) estimation of likelihood ratios with respect to the null model; (iv) computation of the Activity Score; (v) p-value estimation. In the current version, our method extends KEGG pathways with information on validated miRNA-targets inhibitory interactions downloaded by miRTarBase (release 2.5)[28] and miRecords

(updated to April 2013)[29], and with information on TF-miRNAs interactions downloaded from TransmiR (release 1.0)[30].

To launch a simulation, the method requires a set of nodes (at least one) together with their relative deregulations as input values. For simplicity, we first define the case of independent input elements – i.e. pairs of input nodes with no directed path between them. However, we also report an efficient technique to deal with dependent input elements without any impact on the performance of the methodology. Besides this minimum requirement, the method provides: (i) the possibility to input additional information, such as lists of non-expressed genes, allowing one to gain higher accuracy, (ii) two different simulation modalities (simple and advanced), and (iii) support for different organisms, such as *H. sapiens*, *M. musculus* and *R. norvegicus*. Our tool has been implemented as an easy-to-use web application available at: https://phensim.atlas.dmi.unict.it/. All data and source codes generated or analyzed during this study are available at: https://github.com/alaimos/phensim.

Building the meta-pathway

To achieve greater accuracy, a simulation algorithm should take into account pathway crosstalk. A simple approach to this problem has been previously described[27]. All KEGG pathways are merged to build a single directed network using a two-steps procedure. First, all disease pathways are removed. Next, all nodes and edges are merged, removing duplicates. As a result, a KEGG meta-pathway will be generated, ideally representing the human cell environment. The meta-pathway is then annotated with the aforementioned experimentally validated miRNA-target interactions.

Since PHENSIM runs a perturbation analysis on the meta-pathway to estimate outcomes likelihoods, we use the same previously described approach[27] to perform a computationally efficient analysis. More precisely, we sort all nodes according to an approximated topological ordering, using a DFS-like (Depth-First Search) algorithm. The perturbation is therefore computed on each node by following such an ordering, caching each call results, allowing a reduction of the number of recursive calls and improving performance.

Computing Probabilities

To compute the activity of a node in a pathway caused by de-regulation, *PHENSIM* considers each node (X_i) as a discrete random variable that can take three possible values: activated (1), inhibited (-1), and unchanged (0).

Since the probability distribution of each variable is unknown, *PHENSIM* estimates it by simulating random variations of the input and computing the probable state of a node using a topological pathway perturbation analysis. Let $E = \{X_{j_1} = x_{j_1}, \dots, X_{j_n} = x_{j_n}\}$ be the input set of up/down-regulated nodes, where $x_{j_k} \in \{1, -1\}$ for $1 \leq k \leq n$. To compute the state of any node X_i in the meta-pathway, *PHENSIM* estimates $\Pr(X_i = x_i|E)$, where $x_i \in \{-1,0,1\}$. This is achieved by sampling the space of feasible Log-Fold-Changes (*LFC*) for each input node and running *MITHrIL*[11] diffusion algorithm to estimate its perturbation $\mathcal{P}_E(X_i, t)$. In detail, given a node $X_{j_k} \in E$, its Log-Fold-Change for the *t*-th step of the simulation is sampled as:

$$LFC_E(X_{j_k}, t) \sim \begin{cases} \mathfrak{U}(\mathbb{R}^+) & if\ X_{j_k} = 1 \\ \mathfrak{U}(\mathbb{R}^-) & if\ X_{j_k} = -1 \end{cases} \quad (1)$$

where $\mathfrak{U}$ is the uniform distribution. Next, we estimate node perturbation as:

$$\mathcal{P}_E(X_i, t) = LFC_E(X_i, t) + \sum_{X_u \in U(X_i)} \frac{w(X_u, X_i)}{\sum_{d \in D(X_u)} w(X_u, X_d)} \mathcal{P}_E(X_u, t) \quad (2)$$

where $U(X)$ and $D(X)$ are the set of upstream and downstream nodes of X, respectively, and w is a weight establishing the type of interaction between a pair of nodes. Therefore, we can evaluate the state of node X_i at *t*-th step of the simulation as:

$$S_E(X_i, t) = \begin{cases} 1 & if\ \mathcal{P}_E(X_i) > \epsilon \\ -1 & if\ \mathcal{P}_E(X_i) < -\epsilon \\ 0 & otherwise \end{cases}, \quad (3)$$

where ϵ is a user-defined inactivity threshold. By repeating the previous process T times, the probability $\Pr(X_i = x_i|E)$ is estimated as:

$$\Pr(X_i = x_i|E) = \frac{\#S_E(X_i, t) = x_i}{T} \quad , \quad (4)$$

By iterating the above procedure on randomized variants of E, the above argument can be used to approximate the probability distribution of the null model. More precisely, let $E_R = \{X_{r_1} = x_{j_1}, \dots, X_{r_n} = x_{j_n}\}$ be a randomization of the input. Through equations 1, 2, 3 and 4 it is possible to estimate the probability of node alterations in the null model, $\Pr(X_i = x_i|E_R)$.

Computing Activity scores

PHENSIM summarizes the activity of a node X_i, given the input E, in an *activity score,* $Ac_E(X_i)$. Such a value has a dual function: the sign indicates predicted activity (positive means activation, negative inhibition), the value represents the log-likelihood of such an observation with respect to the null model. Therefore, to compute this value, *PHENSIM* needs to determine a log-likelihood ratio, $\mathcal{L}(E|X_i = x_i)$, for each possible outcome x_i of a node. Let $\Pr(X_i = x_i|E)$ and $\Pr(X_i = x_i|E_R)$, be the probabilities that X_i state is x_i in the input and null model, respectively. We can determine the log-likelihood ratio as:

$$\mathcal{L}(E|X_i = x_i) = \log\frac{\Pr(X_i = x_i|E)}{\Pr(X_i = x_i|E_R)}. \quad (5)$$

The activity score for node X_i can be determined as:

$$Ac_E(X_i) = \begin{cases} \mathcal{L}(E|X_i = 1) & if\ \mathcal{L}(E|X_i = 1) > \max\big(\mathcal{L}(E|X_i = -1), \mathcal{L}(E|X_i = 0)\big) \\ -\mathcal{L}(E|X_i = -1) & if\ \mathcal{L}(E|X_i = -1) > \max\big(\mathcal{L}(E|X_i = 1), \mathcal{L}(E|X_i = 0)\big) \\ 0 & if\ \mathcal{L}(E|X_i = 0) > \max\big(\mathcal{L}(E|X_i = -1), \mathcal{L}(E|X_i = 1)\big) \end{cases} \quad (6)$$

Computing p-values

Determining the relevance of an Activity Score is an important step to filter PHENSIM output. In principle an Activity Score is relevant if the likelihood ratio between user input and null model is sufficiently high. To evaluate this, we can calculate a p-value that compares the states distributions of a node in both models.

Therefore, let $E[S_E(X_i,t)]$ and $V[S_E(X_i,t)]$ be the expected value and variance of the states distributions computed for node X_i. These values can be estimated using the sample probabilities of equation 4 as:

$$E[S_E(X_i,t)] = \sum_{x_i \in \{-1,0,1\}} \Pr(X_i = x_i|E) \cdot x_i \quad (7)$$

and

$$V[S_E(X_i,t)] = \sum_{x_i \in \{-1,0,1\}} \Pr(X_i = x_i|E) \cdot (x_i - E[S_E(X_i,t)])^2 \quad (8)$$

These two values can be corrected to obtain unbiased estimators using the Welford's method[31] for the expected value, and Bessel correction for the variance. In details, the unbiased estimator of the mean is obtained as:

$$E_U[S_E(X_i,t)] = E[S_E(X_i,t)] + \frac{1}{T}\sum_{x_i \in \{-1,0,1\}}(x_i - E[S_E(X_i,t)]) \, , \quad (9)$$

where T is the number of input assignments used to compute the probabilities. The unbiased estimator of the variance is computed as:

$$V_U[S_E(X_i,t)] = \frac{T}{T-1} V[S_E(X_i,t)] \, . \quad (10)$$

The same computation can be done for the null model. Finally, the two distributions can be compared using an unpaired heteroscedastic two-samples T-test.

Dealing with dependent nodes

Equation 1 requires that all input nodes be independent of each other, thus allowing to uniformly assign values in the set of possible Log-Fold-Changes. Two nodes are considered independent if no direct path links them. Indeed, if two dependent nodes are in the input set, it is possible that the expression of the upstream node will influence the downstream one. Consequently, a uniform distribution would not be appropriate. To overcome this limitation, we can modify the pathways and the input set to avoid changes in equation 1.

Let E be the simulation input set and let $X_{i_1}, \dots, X_{i_t}$ be a subset of E composed by dependent nodes. We can create a novel node X^* directly connected to $X_{i_1}, \dots, X_{i_t}$. For each edge $X^* \rightarrow X_{i_k}$, its weight, which establishes the interaction type, can be assigned as $w(X^*, X_{i_k}) = x_{i_k}$ where x_{i_k} is the direction of the deregulation we wish to simulate. Therefore, we can build a new input set E^*, where all nodes are independent, as $E^* = \{X^* = 1\} \cup E \setminus \{X_{i_1}, \dots, X_{i_t}\}$.

This new set can be used to approximate simulated expressions, taking dependencies into account, without estimating how such dependencies alter Log-Fold-Changes distribution. A detailed graphical representation of the process is depicted in figure S5.

Results

To assess the potential of PHENSIM, we performed four different simulations. The aim of these simulations was to address whether PHENSIM predicts the resulting phenotype in cells treated with (i) metformin, (ii) everolimus (RAD001), (iii) exosomal vesicles derived from acute myeloid leukemia (AML) cells and (iv) TNFα/siTPL2, a combination that results in synthetic lethality of a subset of human cancer cell lines. We then examined the ability of PHENSIM to correctly predict the activity status of both individual genes/proteins and signaling pathways by comparing PHENSIM predictions with prior experimental data. To evaluate the overall accuracy of our tool, we identified nodes that are upregulated, downregulated, or non-altered with respect to known biology outcomes,

and we calculated the specificity and sensitivity as measures of performance (table S4). *Sensitivity* measures the proportion of altered (up- or downregulated) nodes that are correctly predicted by our method. *Specificity* measures the proportion of non-altered genes that are correctly identified as such. Based on this, PHENSIM showed average *sensitivity* 0.74 and *specificity* 0.35.

Simulation #1: anti-cancer effects of metformin

Metformin is known to uncouple the electron transport chain in the mitochondria by targeting Complex I. This results in the downregulation of the cellular levels of ATP (adenosine triphosphate) and the activation of LKB1 (liver kinase B1) and AMPK (5' AMP-activated protein kinase). AMPK phosphorylates TSC2 (tuberous sclerosis complex 2), and stabilizes the TSC1/TSC2 complex, an inhibitor of the mTORC1 (mammalian target of rapamycin complex 1) complex. In addition, AMPK phosphorylates Raptor, a fundamental regulatory subunit of mTORC1, and this inhibits the recruitment of its targets p70S6K (ribosomal protein S6 kinase B1) and 4EBP1 (eukaryotic translation initiation factor 4E binding protein 1).[12] In parallel with this, metformin also downregulates insulin and IGF1 (insulin-like growth factor 1), the MAPK (mitogen-activated protein kinase) pathway, glycolysis and the TCA (tricarboxylic acid) cycle[12] (fig. 1).

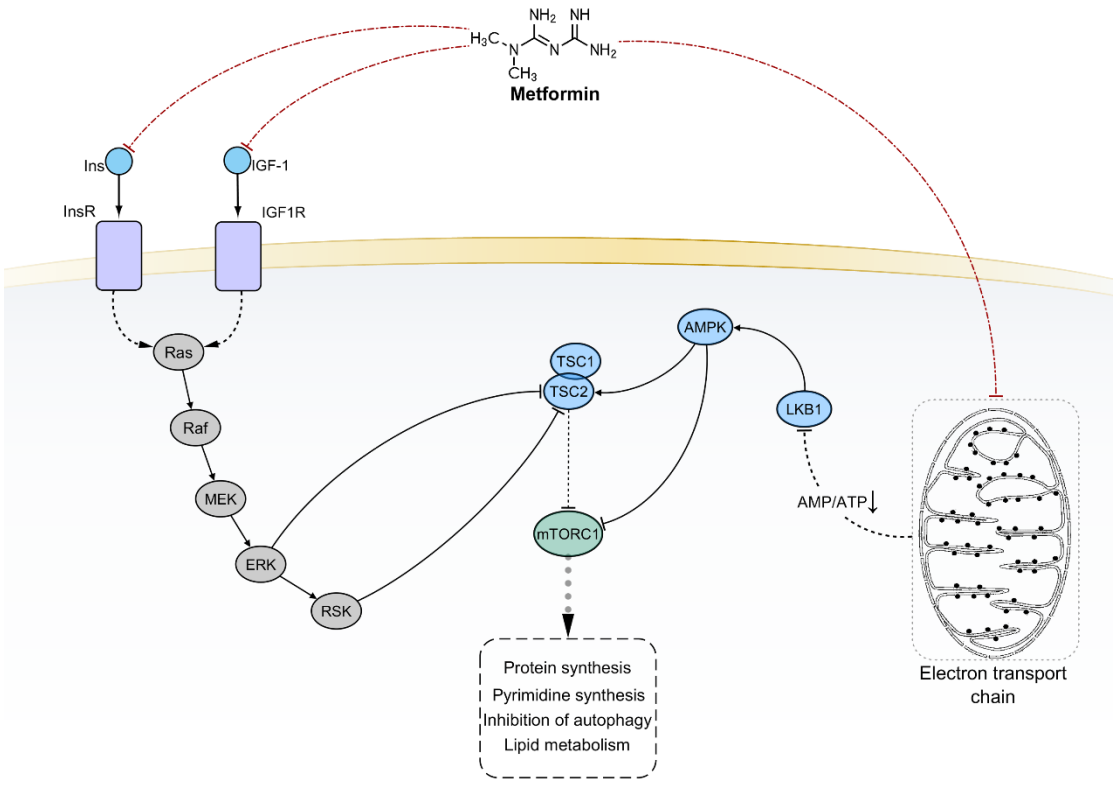

Figure 1. Current model of metformin-mediated pharmacological effects. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions taken into account for PHENSIM prediction.

Based on this information, we launched PHENSIM by simulating the simultaneous upregulation of LKB1 and the downregulation of both insulin (Ins) and IGF1. PHENSIM can contextualize simulations by giving a set of non-expressed/active genes in a specific cellular setting. In this simulation we did consider all genes active, since we wanted to determine all possible effects of metformin in any cellular setting. As expected, PHENSIM returned significant downregulation of Insulin and mTOR signaling (pathway activity score = -2.2981 and -2.3491, respectively; p-value < 0.0001). All isoforms of PI3Ks (phosphoinositide 3-kinases) and AKT (serine/threonine protein kinase Akt) as well as the metabolite PIP3 (phosphatidylinositol (3,4,5)-trisphosphate) were also downregulated (e.g. node activity score = -3.3524 and -2.3567 for PIK3CA and AKT1, respectively;

p-value < 0.0001). The negative regulation of mTOR (node activity score = -3.2626; p-value < 0.0001) resulted in the activation of the *repressor of translation initiation 4EBP* (node activity score = 3.1847; p-value < 0.0001) and the inhibition of downstream nodes involved in protein synthesis (fig. S1a). The TCA cycle was also inhibited (pathway activity score = -3.0033; p-value < 0.0001), along with MAPK and NF-κB (nuclear factor kappa B) signaling (pathway activity score = -2.0641 and -1.7574; p-value < 0.0001). For each of these three pathways, a number of downregulated enzymes and metabolites was predicted, in full agreement with data from literature[12] (fig. S1b). However, PHENSIM also predicted the inhibition of p53 signaling, and the lack of changes in cytokine gene expression, which is contrary to prior literature.[12] In the context of simulation #1, PHENSIM showed sensitivity 0.84 and specificity 0 (table S1).

Simulation #2: RAD001 effects on the mammary tissue

Everolimus (RAD001, Afinitor®), an analog of rapamycin, is an allosteric inhibitor of mTORC1, a master regulator of multiple metabolic pathways. mTORC1 promotes protein synthesis by: (a) stimulating ribosome biogenesis via phosphorylation and inhibition of the RNA Polymerase III repressor MAF1;[13] (b) by phosphorylating the p70S6K and 4EBP1 and modulating the activity of their downstream targets;[14] (c) by regulating nucleocytoplasmic RNA transport.[13,14] In addition, mTORC1 stimulates pyrimidine biosynthesis and lipid biosynthesis.[14] Finally, mTORC1 phosphorylates ULK1 (unc-51 like autophagy activating kinase 1) and DAP (death-associated protein) inhibiting autophagy.[15] All these functions of mTORC1 are reversed by everolimus and other mTORC1 inhibitors[16,17] (fig. 2).

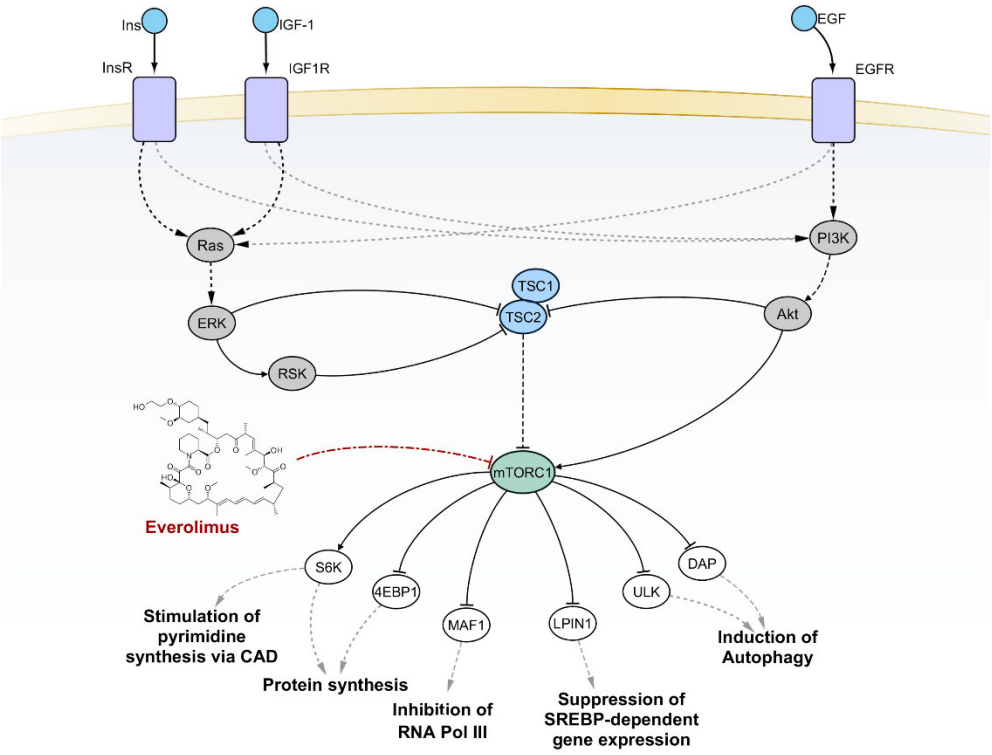

Figure 2. mTORC1 and its downstream signaling pathways. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions taken into account for PHENSIM prediction.

Here we wanted to launch PHENSIM by simulating the inhibition of mTORC1. Unfortunately, simulating mTORC1 inhibition was not feasible because KEGG does not distinguish the mTOR node in mTORC1 from the one included in mTORC2. To overcome such limitation, we created a dummy element representing the drug, and connected it with the well-known downstream targets of mTORC1: p70S6K (p70S6Ka and p70S6Kb), 4EBP and ULK1. The setting we used were downregulation for the first two and upregulation for ULK1. Then we uploaded a list of non-expressed genes in breast tissue to simulate the effects of the drug on such a tissue. Following this, we launched the simulation, which predicted that factors associated with RNA transport would be downregulated, while factors involved in autophagy would be upregulated. The simulation when

launched in this setting showed that RNA transport and mTOR signaling pathways exhibits the lowest activity scores (-2.9570 and -2.3177, respectively; p-value < 0.0001) (fig S2a), as factors involved in RNA transport and protein synthesis, such as eIF4E, eIF4B, eIF4A (eukaryotic translation initiation factor 4E, 4B and 4A) and S6 (ribosomal protein S6), were downregulated (node activity score = -3.3524, -4.6565, -4.6052 and -4.6565, respectively; p-value < 0.0001) (fig S2b). Upregulation of the autophagy (pathway activity score = 1.4966, p-value < 0.0001) was consequent to alterations in ULK1 phosphorylation levels, which were preset by us. However, PHENSIM failed in predicting the deregulation of p21 (cyclin-dependent kinase inhibitor 1), cyclin D and NF-kB.[17] In simulation #2, PHENSIM showed sensitivity 0.25 and specificity 1 (table S2).

Simulation #3: effects of exosomal vesicles on hematopoietic stem/progenitor cells (HSPCs) in the bone marrow (BM)

Exosomes derived from AML blasts contain complex cargoes which function via paracrine mechanisms to modulate the properties of both the tumor cells themselves and the BM niche. Several microRNAs have been shown to be selectively incorporated in these exosomes, including miR-150 and miR-155.[18,19] One of the targets of these microRNAs is the transcription factor c-MYB, which is downregulated in tumor cells exposed to the exosomes.[19] Additional targets include c-KIT, DNMT1, Lymphoid Cell Helicase (HELLS), PAICS, an enzyme involved in purine biosynthesis, TAB2, and others. The downregulation of these molecules compromises hematopoiesis via stroma-independent mechanisms. However the cargo of AML cell-derived exosomes also targets mesenchymal stromal progenitors, inhibiting the production of hematopoietic stem cell supporting factors such as CXCL12 (C-X-C motif ligand 12), KITL (c-Kit ligand) and IGF1 and interfering with both hematopoiesis and osteogenesis[20] (fig. 3).

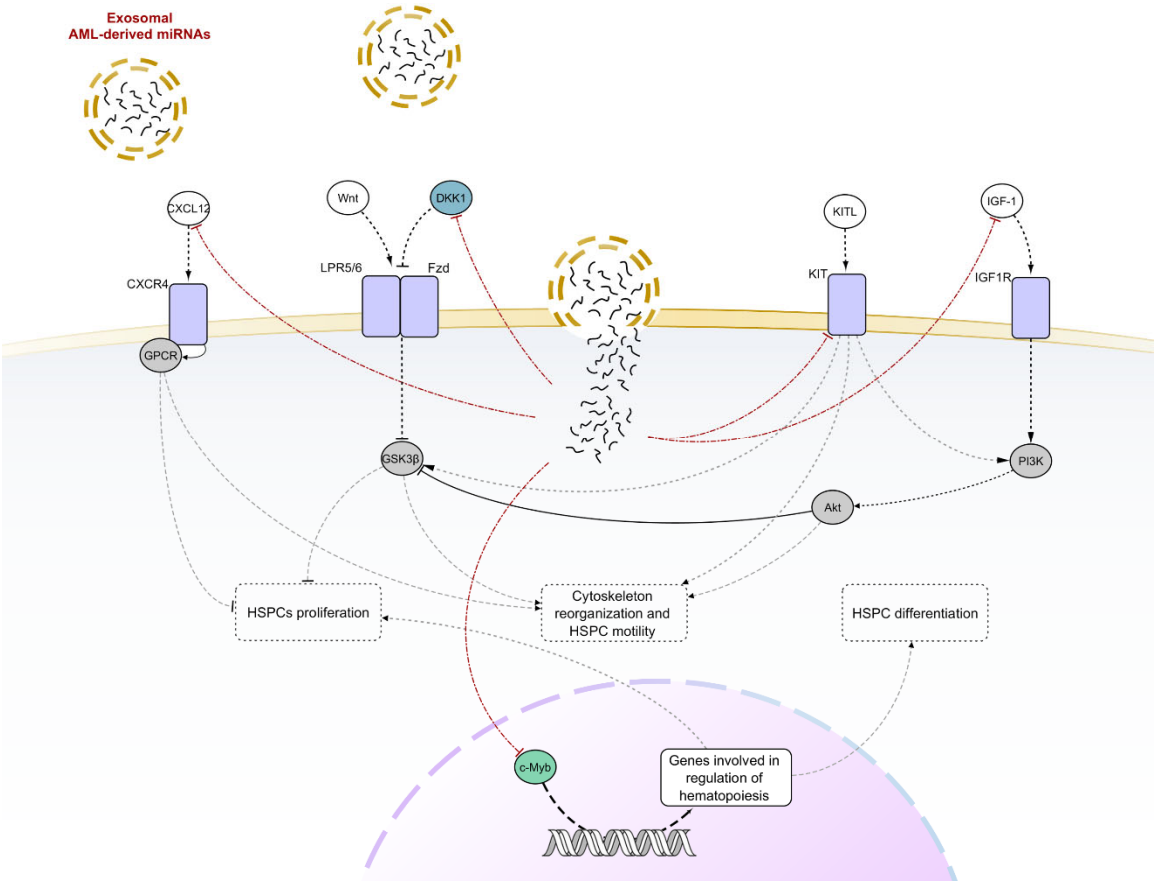

Figure 3. A reconstructed model showing cellular components involved in hematopoiesis and motility of HSPCs and their downregulation mediated by exosomal-miRNAs derived from AML cells. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions taken into account for PHENSIM prediction.

To determine whether PHENSIM can make the correct predictions in this model, we launched PHENSIM by simulating the uptake of the eight most representative miRNAs (miR-150, -155, -146a, -191, -221, -99b, -1246 and let-7a) included in AML-derived exosomes by hematopoietic stem cells.[18] As in simulation #1, we did not input any non-expressed gene, since our objective was to evaluate the general impact exerted by AML-related exosomal miRNAs on HSPCs homing in the BM. The simulation returns osteoclast differentiation and cytokine-cytokine receptor interaction pathways as

inhibited (pathway activity score = -2.4196 and -2.5975, respectively; p-value < 0.0001) (fig. S3a-b). In full agreement with the literature, some genes involved in modulation of normal hematopoiesis, like CXCL12 and IGF1, were returned downregulated (node activity score = -5.2005 and -2.8102, respectively; p-value < 0.0001).[20] Similarly, c-MYB, which is involved in HSPC differentiation and proliferation, was also returned downregulated[19] (node activity score = -4.1044; p-value < 0.0001) (fig. S3c). However, PHENSIM failed to predict the upregulation of DKK1 and the downregulation of KITL and IL-7.[20] Since the information on the effects of AML-derived exosomal miRNAs on recipient cells in the BM was limited, we were not able to calculate sensitivity/specificity for this simulation.

Simulation #4: testing TNFα/siTPL2-dependent synthetic lethality on a subset of human cancer cell lines

TNFα (tumor necrosis factor alpha), a type II transmembrane protein, is a member of the tumor necrosis factor cytokine superfamily and has an important role in innate immunity and inflammation. Although it has the potential to induce cell death, most cells are protected by a variety of mechanisms. In a recent paper[22], Serebrennikova et al. showed that one of the checkpoints of TNFα-induced cell death is TPL2 (MAP3K8), a MAP3 kinase that is known to have an important role in immunity, inflammation and oncogenesis. The knockdown of TPL2 resulted in the downregulation of miR-21 and the upregulation of its target CASP8 (caspase-8). This combined with the downregulation of the caspase-8 inhibitor cFLIP (FADD-like IL-1β-converting enzyme inhibitory protein), resulted in the activation of caspase-8 by TNFα and the initiation of apoptosis (fig. 4). The activation of caspase-8 also promotes the activation of the mitochondrial pathway of apoptosis, although some molecules such as BIM_L (Bcl-2-like protein 11, isoform L), which are also involved in the activation of the mitochondrial pathway, may be activated via caspase-8-independent mechanisms. An important upstream regulator of this pathway is NF-κB. The activation of ERK (MAPK1/2), JNK (c-Jun N-terminal kinase) and p38MAPK, the activation of AKT and the phosphorylation of GSK3 (glycogen

synthase kinase 3) at Ser9/21 are also inhibited by the knockdown of TPL2. However, their inhibition does not appear to have a role in the initiation of TNFα/siTPL2-induced apoptosis. It is worthy to notice that the activation of the apoptotic (caspase-8-dependent) pathway in TNFα/siTPL2 treated cells was observed in some, but not all cancer cell lines, suggesting that correct prediction will depend on whether the data analyzed by PHENSIM are derived from sensitive or resistant cells.

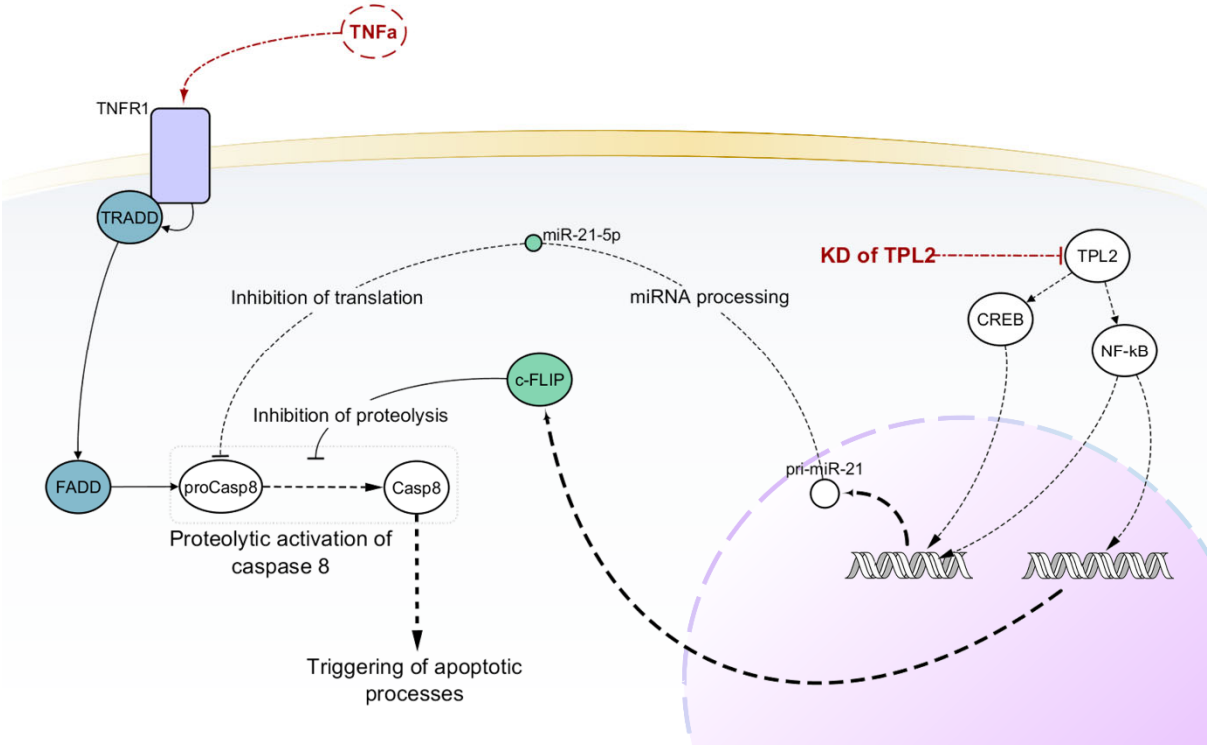

Figure 4. Generalized model showing molecular mechanisms underlying the TNFα/siTPL2-dependent synthetic lethality. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions taken into account for PHENSIM prediction.

To launch the simulation, we set TPL2 as downregulated and TNFα as upregulated. Since our goal was to simulate the outcome of such treatment in six different cell lines, i.e. HeLa, HCT116, U2-OS, CaCo-2, RKO and SW480, we launched six different simulation. Each simulation had a different list of non-expressed genes, one for each cell-line. Among these tumor cell lines, only HeLa, HCT116,

U2-OS were sensitive to treatment with TNFα/siTPL2. At the end of the computations, PHENSIM could not predict upregulation of caspase-8 and miR-21 for any of the six cell lines. However, PHENSIM was able to predict the upregulation of cFLIP in the resistant cell lines. This could be the result of missing information in KEGG pathways, as further detailed in the discussion section. Also, PHENSIM did not predict any alteration for MLC1 (Mcl-1 apoptosis regulator) and XIAP (X-linked inhibitor of apoptosis) nodes, but it predicted the upregulation of the apoptosis inhibitors BCL2 and BCL-XL, and the downregulation of the inducer of mitochondrial apoptosis BAX (fig. S4a-d). Although this result contrasted our expectations, it was confirmed by results obtained by the previously mentioned experimental study[22], which suggested that the change in the expression of these molecules was due to the activation of feedback mechanisms. Interestingly, this result was obtained only for four out of six cancer cell lines, of which three were sensitive (HeLa, HCT116 and U2-OS) and one was resistant (CaCo-2). PHENSIM did not report significant alteration of BCL2/BCL-XL or BAX in the case of RKO or SW480, both resistant. In addition, phosphorylated ERK, MEK, JNK and p38 as well as cIAP2 (baculoviral IAP repeat containing 2) activity were returned strongly downregulated for all of six the cell lines, as confirmed by the experimental data (fig S4a-b). In the context of simulation (4), PHENSIM showed sensitivity 0.83 and specificity 0.14 (table S3).

Discussion

Although a variety of pathway-based analytical strategies are currently available, the basic principle underlying these methods widely relies on the use of proper statistical testing aiming to identify dysregulated biological functions under a certain condition. Contrary to single gene/protein analysis approaches, the vast majority of pathway analysis methods have been purposely structured with the intention to manage and analyze massive omics data[10,23]. Here, the main idea is whether the set of differentially regulated pathways and nodes returned by the analytical process is capable to summarize the resulting phenotype.

In the present paper, we have introduced PHENSIM, a flexible, user-friendly pathway analysis method functioning as a phenotype simulator. Differently from most of the previous pathway analysis approaches, PHENSIM is independent of fold-change values as it has been developed mainly with the purpose to predict the effects of one or multiple molecular deregulations on cell/tissue phenotype, instead of analyzing wide datasets of omics data. Thus, we propose PHENSIM as an easy-to-use, supportive pathway-based method that can make predictions on the biological outcome of experiments targeting the expression of activity of signaling processes.

To test the potential of our tool, we tested it by simulating four distinct scenarios consisting in drug administration to cultured cells (simulations #1 and #2), effects of exosomal-derived miRNAs in recipient cells (simulation #3), and the combined targeting of two signaling molecules, which is known to induce synthetic lethality in a subset of cell lines (simulation #4). After comparison the literature data and PHENSIM predictions were in almost full agreement with simulation #1 and in partial agreement with the three remaining simulations, showing a discrete degree of accuracy in terms of sensitivity and specificity.

Discrepancies with baseline data outlined some limitations in the predictive potential of our method. However, since pathway analysis relies on prior knowledge about how genes, proteins and metabolites interact with each other, we hypothesize that such a negative outcome is at least in part due to incompleteness of the existing knowledges employed in the study. Indeed, since the biological pathways reported on current databases still remain largely fragmented because of gaps in our knowledge, calculations based on them will inevitably produce less reliable results[10]. One example of this is provided by mTORC1 downstream signaling. It is known that mTORC1 promotes protein synthesis by phosphorylating p70SK and 4EBP, and by stimulating ribosome biogenesis via inhibitory phosphorylation of the RNA Polymerase III repressor MAF1.[13] Concerning mTORC1-induced pyrimidine biosynthesis, it is well known that its stimulation is due to p70S6K-mediated phosphorylation of the multifactorial enzyme CAD (carbamoyl-phosphate synthetase 2, aspartate

transcarbamylase, and dihydroorotase) as well as to upregulation of 5-phosphoribosyl-1 pyrophosphate (PRPP), a product of the pentose-phosphate pathway that functions as an allosteric activator of CAD[24,25]. Since signaling pathways provided by KEGG do not consider such interactions, it was not possible for our tool to predict any perturbations for these biological processes, thus causing a loss of predictive accuracy. Similar observations can be made concerning the upregulation of caspase-8 for any of the three siTPL2/TNFα-sensitive cancer cell lines by our method, although we were able to identify indirect evidences of such activity. On the other hand, the correct predictions obtained for autophagy, RNA transport and mTOR signaling in simulation #2, as well as those obtained for genes related to the mitochondrial apoptotic pathway and limited to siTPL2/TNFα-sensitive cell lines in simulation #4 suggest that, *provided with the right information, PHENSIM may be able to significantly improve its sensibility and specificity*.

A further limitation for pathway analysis methods consists in the inability of current knowledge bases to contextualize gene expression and pathway activation in a cell- and condition-specific manner.[10] Two important consequences of this omission are that (i) the annotation of genes in a given pathway includes both genes encoding core components of the pathway, as well as genes encoding proteins that may have a role in the regulation of the pathway under specific experimental condition. Such genes, distantly involved in the regulation of the pathway, may be relevant only in some cells, exposed to unique microenvironment. (ii) Pathways, as currently defined, do not consider isoforms of a given protein encoded by different genes or differently processed mRNAs derived from a single gene. This is a significant limitation since such isoforms may have unique and sometimes opposite signaling properties. By developing a strategy that allows to remove non-expressed genes from the computation, we offer to the user the possibility to contextualize predictions in a cell- or tissue-dependent manner. In conjunction with this, the integration of KEGG pathways with information coming from posttranscriptional regulators such as miRNAs increased the accuracy of the results and leads to considerable improvements of predictions[26]. Moreover, using the KEGG meta-pathway

approach, instead of single disjointed pathways, partially addresses the issue of pathways independence[10].

In conclusion, PHENSIM showed discrete accuracy in most applications and was able to predict the effects of several biological events starting from the analysis of their impact on KEGG. We assume that several discrepancies can be traced to the incompleteness of knowledge in KEGG pathways and/or to the lack of appropriate cell- and condition-specific information. Such an incompleteness can be partially addressed through a manual annotation of the pathways with the missing elements and links, including miRNA-target and TF-miRNA interactions. We should add that PHENSIM is limited to the simulation of changes in the expression and/or activity of signaling molecules. It is not suitable to simulate genetic aberrations, unless such alterations are known to directly affect the expression and/or activity of such molecules. Despite these limitations, our approach shows appreciable utility in the experimental field, as a tool for the prioritization of experiments with greater chances of success.

Funding

ALF is supported by the PhD fellowship on Complex Systems for Physical, Socio-economic and Life Sciences funded by the Italian MIUR "PON RI FSE-FESR 2014-2020". SA, AF and AP have been partially supported by the research project "Marcatori molecolari e clinico-strumentali precoci, nelle patologie metaboliche e cronico-degenerative" founded by the Department of Clinical and Experimental Medicine of University of Catania. SA, AF and AP have been partially supported by the MIUR PON research project BILIGeCT "Liquid Biopsies for Cancer Clinical Management". AP has also been partially supported by the italian MIUR FFABR grant.

Author contributions

SA, AF, and AP conceived the work. AF and AP coordinated the research. SA, GPM, and ALF wrote the paper. SA developed the system. SA, GPM, ALF run the computational experiments. OBS and

PNT provided biological insights and data. All authors analyzed the data and reviewed the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Ramanan, V. K., Shen, L., Moore, J. H. & Saykin, A. J. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* **28**, 323–332 (2012).

2. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA* **8**, e1364 (2017).

3. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

4. Picotti, P. *et al.* High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nat. Methods* **7**, 43–46 (2010).

5. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15**, 256–278 (2014).

6. Giacomini, K. M. *et al.* Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nat. Rev. Drug Discov.* **16**, 1 (2017).

7. Wang, R.-S., Maron, B. A. & Loscalzo, J. Systems medicine: evolution of systems biology from bench to bedside. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **7**, 141–161 (2015).

8. Kirchmair, J. *et al.* Predicting drug metabolism: experiment and/or computation? *Nature Reviews Drug Discovery* **14**, 387–404 (2015).

9. Jin, L. *et al.* Pathway-based Analysis Tools for Complex Diseases: A Review. *Genomics, Proteomics & Bioinformatics* **12**, 210–220 (2014).

10. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).

11. Alaimo, S. *et al.* Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget* **7**, 54572–54582 (2016).

12. Gong, J. *et al.* The expanding role of metformin in cancer: an update on antitumor mechanisms and clinical development. *Target. Oncol.* **11**, 447–467 (2016).

13. Michels, A. A. MAF1: a new target of mTORC1. *Biochem. Soc. Trans.* **39**, 487–491 (2011).

14. Ben-Sahra, I. & Manning, B. D. mTORC1 signaling and the metabolic control of cell growth. *Curr. Opin. Cell Biol.* **45**, 72–82 (2017).

15. Eisenberg-Lerner, A., Bialik, S., Simon, H.-U. & Kimchi, A. Life and death partners: apoptosis, autophagy and the cross-talk between them. *Cell Death Differ.* **16**, 966–975 (2009).

16. Hurvitz, S. A. *et al.* In vitro activity of the mTOR inhibitor everolimus, in a large panel of breast cancer cell lines and analysis for predictors of response. *Breast Cancer Research and Treatment* **149**, 669–680 (2015).

17. Lui, A., New, J., Ogony, J., Thomas, S. & Lewis-Wambi, J. Everolimus downregulates estrogen receptor and induces autophagy in aromatase inhibitor-resistant breast cancer cells. *BMC Cancer* **16**, 487 (2016).

18. Hornick, N. I. *et al.* Serum Exosome MicroRNA as a Minimally-Invasive Early Biomarker of AML. *Sci. Rep.* **5**, 11295 (2015).

19. Hornick, N. I. *et al.* AML suppresses hematopoiesis by releasing exosomes that contain microRNAs targeting c-MYB. *Sci. Signal.* **9**, ra88 (2016).

20. Kumar, B. *et al.* Acute myeloid leukemia transforms the bone marrow niche into a leukemia-permissive microenvironment through exosome secretion. *Leukemia* **32**, 575–587 (2018).

21. Lapid, K. *et al.* GSK3β regulates physiological migration of stem/progenitor cells via cytoskeletal rearrangement. *J. Clin. Invest.* **123**, 1705–1717 (2013).

22. Serebrennikova, O. B. *et al.* The combination of knockdown and TNFα causes synthetic lethality via caspase-8 activation in human carcinoma cell lines. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14039–14048 (2019).

23. Ramanan, V. K., Shen, L., Moore, J. H. & Saykin, A. J. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* **28**, 323–332 (2012).

24. Ben-Sahra, I., Howell, J. J., Asara, J. M. & Manning, B. D. Stimulation of de novo pyrimidine synthesis by growth signaling through mTOR and S6K1. *Science* **339**, 1323–1328 (2013).

25. Robitaille, A. M. *et al.* Quantitative phosphoproteomics reveal mTORC1 activates de novo pyrimidine synthesis. *Science* **339**, 1320–1323 (2013).

26. Alaimo, S., Micale, G., La Ferlita, A., Ferro, A. & Pulvirenti, A. Computational Methods to Investigate the Impact of miRNAs on Pathways. *Methods Mol. Biol.* **1970**, 183–209 (2019).

27. Alaimo, S., Marceca, G. P., Ferro, A. & Pulvirenti, A. Detecting Disease Specific Pathway Substructures through an Integrated Systems Biology Approach. *Noncoding RNA* **3**, (2017).

28. Hsu, S.-D. *et al.* miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **39**, D163–9 (2011).

29. Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* **37**, D105–10 (2009).

30. Wang, J., Lu, M., Qiu, C. & Cui, Q. TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.* **38**, D119–22 (2010).

31. Welford, B. P. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics* **4**, 419–420 (1962).