# SUPPLEMENTAL METHODS

**Transduced cell array validation using publically available microarray data**

Microarray data were downloaded from the GEO database using the ArrayExpress library[1]. GSE18938[2], a human exon array, HuGene-1.0st v1, was rma background corrected[3], quantile normalized[4] and probeset summarized with the oligo package[5]. One condition was removed due to large residual values, high polyA spike values, and lack of clustering with the rest of conditions using principal component analysis (MCF10A cells treated with EGF at day 5). Genes were considered significant if the contrast between the average value of normalized intensity for EGF treated cells versus the untreated MCF10A cells was statistically significant (p-value $\leq$ 0.001) at an absolute FC $\geq$3. Contrasts were calculated for each individual time point (i.e., difference between EGF condition and control after 3 days) as well as for the average of all the time points using the limma package[6].

Explain v3.0 from Biobase (www.biobase-international.com) identified the most likely TFs that might regulate the expression of the significant genes with P-Match[7]. P-Match predicts transcription factor binding sites in DNA sequences combining pattern matching and weight matrix approaches using TRANSFAC database. No significant genes in each temporal set, FC $\leq$ 1.001 and p-value $\geq$ 0.5, were utilized as a background set, and the search was performed between -1000 to 500 base pairs (bps) with respect to the TSS. P-Match output that provided the most likely TFs that might regulate the significantly changed genes was further filtered to remove position weighted matrices (PWMs) whose ratio between the significant and background genes was less than 1, the p-value for that ratio was less than 0.1 and matched promoter sequence p-value was less than 0.2.

The PWMs associated with the most likely TFs that regulate those significant genes according to P-Match were contrasted against the PWMs associated with the most likely TFs that bind to each of the experimental TFr employed and that have a core score $\geq$0.9 and matrix score $\geq$0.9 (Supplementary File 1). From these comparisons, we determined whether the activity of any TFr could be altered by any of the active TF. Additionally, we specified the TFr sensitivity to measure the activity of the TF for which the TFr was originally constructed or if it was instead reporting the activity of any other TFs. Sensitivity and specificity of TRACER were calculated assuming the computational results from the transcriptomic measurements as reference. Final results are in Supplementary Figure 2.

**Networks for transcriptional activity arrays**

Networks for tanscriptional activity cell arrays, NTRACER, aim to identify the dynamics of cellular processes that control an observed phenotype using dynamic measurements of TFr activity. NTRACER combines prior knowledge and an ensemble of diverse inference methods to determine the possible relationships between the given cellular inputs and the TFrs. The dynamic network is currently modeled as a three-level Boolean model, although other paradigms are possible. The most likely connections present at each time are later established by minimizing the difference between the experimental data and the model, penalizing network complexity. NTRACER employs normalized activity data from the significant TFrs as input. The computational pipeline involves three main steps: i) statistical analysis to identify significant changes in the TFr activity data, ii) generation of an initial network topology, and iii) network identification. An overview of the pipeline is in Supplementary Fig. 6.

TRACER provides data with two unique aspects, TF activity and dynamic data over several days, which motivated the development of NTRACER. NTRACER is based on a modification of CellNOptR[8, 9] to accommodate the unique aspects of TRACER data. CellNOptR starts with an initial network topology, coming exclusively from prior knowledge obtained from biological databases or from literature search, which is optimized to minimize the differences between the experimental data and the results of a two-level Boolean model of such a network while penalizing complexity. CellNOptR was conceived originally for analyzing and modelling steady-state phosphorylation data. Therefore, CellNOptR was modified to take into consideration the uniqueness of dynamic TF activity data:

a) Indirect connections: connections between nodes of the TF regulatory network imply an indirect interaction of the two TFs. Examples of this indirect interaction are that TF A binds to the promoter region of the gene that encodes TF B, or TF A binds to the promoter region of a gene whose production modulates the activity of TF B. TRACER measures the activity of the two TFs, yet cannot distinguish between the different indirect pathways of interaction. The available prior knowledge is in the form of protein-DNA interactions, and does not account for all possible indirect connections. Thus, inference methods were employed to identify these connections and add them to the initial prior knowledge network.

b) Non-linear relationships: due to the nature of the data, TF activity, the TF regulatory networks encompass multiple processes (e.g., transcription, translation, phosphorylation) that are represented by the edges between nodes. Each process can be non-linear and thus their combination might be expected to be non-linear as well. Therefore inference methods that can handle non-linear interactions, such as mutual information or Bayesian network based methods, are better suited for identifying edges in non-linear processes. Depending on the time resolution, a non-linear process can be considered as linear, and therefore, adding linear methods, such as PLSR, might capture those connections as well.

c) Network motifs: the type of network motifs in TF activity networks are not known a priori as TRACER has not been previously employed. Multiple motifs are possible such as fan in, fan out, or feed-forward motifs. Community inference challenges[10] have indicated that an inference method may have a preference for certain types of motifs. Thus employing multiple inference methods is aimed at providing an unbiased view of the TF activity network structure.

d) Dynamic TF activity data: each time step is modelled as a three-level Boolean paradigm to account for the possibility of the TF activity to be above or below the control treatment (i.e., no treatment). In other words, the three-level Boolean model allows the determination of the possibility that the different study treatments may have activation or inhibitory effects in the TF with respect to the control. Each time step network is related with the previous one as the latter structure is used as the initial guess for the former time step during the optimization process.

Combining prior knowledge with connections obtained from a union of diverse inference methods provided an initial network to optimize using CellNOptR. This initial network contains all experimental prior knowledge about possible protein-DNA interactions

between TFs as well as possible connections between the activities of the TF reporters obtained from inference methods. Inference methods contain multiple false positive connections. Furthermore, not all the connections previously established in the literature are present, as the connections in the literature were established from a variety of cell types[11]. Thus, multiple false positive edges are present in the prior knowledge for our specific system (i.e., ErbB2). Using a structure optimization procedure that penalized complexity such as CellNOptR allows the removal of edges that are not in agreement with the observed experiments, and therefore, it removes false positive edges coming from prior knowledge as well as inference methods. Overall, the envisioned computational pipeline (Supplementary Fig.6) was developed to identify highly robust and consistent results in the final networks by protecting against the erroneous identification of edges that could result from noisy data. Robustness and consistency was accomplished by: i) data pre-filtering to remove highly non-significant reporters; ii) inherence robustness of inference methods against noise. Inference methods are not likely to identify connections between two TFrs if their values are very noisy; iii) NTRACER has a mechanism to remove edges, and edges involving noisy data would be highly unlikely and thereby removed; iv) bootstrapping techniques and sample permutation were employed to select only robust results.

*Prior Knowledge:* The initial network topology is originated from an equally weighted number of prior knowledge sources and inference methods. Prior knowledge information includes directed human protein-DNA interactions, either proximal or distal regulation with respect to their TSSs, obtained from TRANSFAC 2011.2 as of February 2012, IPA as of March 2012, and GeneGO v6.9 as of March 2012. Databases were merged based on Entrez ID and later transformed back into symbols using Gene Info from the NCBI database as of March 4th, 2012 (ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/). For those interactions whose effects were unknown, both an arbitrary activation effect and an inhibitory effect were given to the connection. Complexes were split into their individual proteins. Additionally, prior knowledge information also includes directed signed connections between external stimuli, i.e., EGF, or lapatinib, and each TFr. These connections were identified using as a template the ErbB family networks from GeneGO (MetaCore version 6.10) by determining all the shortest pathways between each of the stimuli given to the 10A/ErbB2 cells (i.e., EGF, DA, lapatinib) and the different TFrs measured in TRACER (Supplementary Fig. 9). Finally, a conversion list is created to match the names of the experimental reporters to the possible TFs that they represent in the current study. This is necessary as some reporters represent a family of TFs, such as GATA or they can be the result of a complex, AP1, for instance.

*Inference methods and initial network topology:* Normalized TFr activity data were interpolated using natural spline interpolation, doubling the number of data points. Interpolated and normalized TFr activity from the untreated samples (control) was subtracted from the rest of the treatments to eliminate the variation from cellular processes that were not activated or repressed by the given treatment. As currently no method is able to determine all the possible types of biological connections and each inference method tends to better predict certain motifs over others[12, 13], different inference methods were studied: partial least squares regression (PLSR[14, 15]), mutual information methods (ARACNE[16], CLR[17], MRNET[17]) and Bayesian networks (BANJO[18]).Therefore, NTRACER inferred all the possible connections between the stimuli and the TFr and between TFrs based on the provided data and generated a final inference network that combines equally all the results from the different inference methods.

Partial least squares regression (PLSR) was employed to infer connections between the stimuli or inputs and the TFrs. In general, PLSR consists of the decomposition of the matrix X that contains the predictors, in our case the TFs, and Y, which holds the responses, in two matrices that are called scores (T and U) and loading matrices (P and Q) in the following manner:

$$X = TP^T + E$$
$$Y = UQ^T + F$$

where E and F are the residuals or the errors terms. The relationship between X and Y is then in the form:

$$Y = TBQ^T + F$$

where B is the regression coefficient matrix of the Y and X score matrices

$$U = TB$$

PLSR was applied to mean-center and unit variance scaled data, using the pls package in R. Differences between two consecutive time treatment scaled values were regressed using PLSR on the initial temporal responses of the TFrs, for three interpolated time points, conditions at which the stimuli (i.e, EGF, DA, lapatinib) were believed to directly affect them. Connections were considered significant if their weight load for their first component was greater than 0.15 or 0.14, for the tissue formation or therapeutic effects, respectively. The directionality of the interaction is given by the sign of the loading. The selected cut-off corresponds with the mean weight loading to remove very unlikely connections between the treatments and the TFrs. To identify the TFrs that most likely affected other constructs, differences between two consecutive temporal TFr scaled activities were regressed with respect to the previous time point of the other constructs for all the experimental time points using PLSR. Connections were considered significant if their load for their first component was greater than 0.3 or 0.28, for the tissue formation or therapeutic effects, respectively, which allows selecting only less than 2% of the possible TFr-TFr connections to assure choosing only a few strong edges. The cut-offs between treatments and TFrs were purposely selected lower to compensate for the limited amount of information available in the literature describing the studied stimuli-TF interactions.

Mutual information (MI) methods were only considered to determine the interactions between stimuli or inputs and TFrs. Connections between the stimuli and TFrs were determined using three well-known MI methods, ARACNE , CLR and MRNET, from the minet package in R, employing the same mutual information matrix (MIM). The sign of the interaction between the stimuli and the

TFrs was determined by the initial slope over time for each stimulus. MIM was estimated with the Schurmann-Grassberger estimate of the entropy by equal frequency for discretization of the data. Default parameters were used otherwise. Only positive scoring results for each of the methods, ARACNE, CLR and MRNET, were included as edges in the MI inference network.

Finally, Bayesian networks were obtained using BANJO, http://www.cs.duke.edu/~amink/software/banjo/. BANJO was run assuming that all the data were independent due to the large experimental frequency used; therefore a static Bayesian network approach was chosen, and no prior knowledge was provided. No parents were allowed for any of the stimuli, and all the data were discretized in three intervals except for the stimuli, only in two levels, 0 or 1. Simulated annealing with random local moves was the choice for the searching strategy with the default parameters and a maximum parent size of 5. Due to the heuristic nature of simulated annealing that does not guarantee the global minimum, Banjo was run 20 times. An ensemble of non-redundant interactions from the top network of each run was created. Only edges present in at least 90% of the runs were deemed significant to include only highly robust results. Interaction signs were given by the influence score. Both activation and inhibition connections were included if the influence score was 0. All the inference networks were combined together and finally merged with the simplified prior knowledge network.

*Determination of TF network evolution over time*: An initial network topology was generated from the combination of prior knowledge and inference methods. Not all the possible connections between the stimuli and TF reporters have been exhaustively explored experimentally, and some interactions are cell dependent, they might not be present in all cell types[11, 19]. This deficiency was compensated by including inference methods that could identify novel connections between the stimuli and the TFrs and between the TFrs based on TRACER data. Similarly, inference methods are not perfect as they usually present many false positives and false negatives[12, 13]. Finally, the processes studied with TRACER are dynamic in nature, so the connections between TFrs and stimuli will change over time as the cell evolves to achieve its final phenotype. Therefore, to overcome the limitation of prior knowledge and ensemble methods and yet identify the dynamical interactions, a structure optimization approach was selected as a solution. The structure optimization consisted of identification of the edges that were more likely to be present at each experimental point measured. As the sampling frequency was much longer than the usual time scales for signaling and translation, days versus minutes, each experimental data was assumed to be under a pseudo-steady state. The network was modeled as a three-level Boolean paradigm, where the stimuli and TFrs were the nodes and the edges were the gates. The best structures or networks were those in which the difference between the model simulations and the experimental data was minimized. To achieve this final goal, we developed a modification of the original CNO[8], currently implemented as an R package in Bioconductor, CellNOptR[9] , to handle the temporal TFr activity data.

Significantly changed TFrs with respect to the control treatment were subsequently studied. Normalized TFr activity with respect to the control and initial experimental time were discretized in three levels, 1, 0 and -1, indicating whether the TFr activity was significantly above, not significant or significantly below the control, respectively. The discretization was based on the statistical results when comparing the dynamic changes of each TFrs with respect to the starting experimental time. For instance, if the normalized TFr activity with respect to the control treatment and initial experimental time for a TFr, for a given treatment, repeat and time set, was positive and p-value of the corresponding comparison was deemed significant, the discretize value was assigned as 1. Similarly, if the normalized TFr activity with respect to the control treatment and initial experimental time, for a given TFr, treatment, repeat and time set, was negative and p-value of the corresponding comparison was considered significant, the discretize value was assigned as -1. For all the comparisons that were not significant, the values where set to 0. The cut-off p-value was 0.15 and 0.05 for the tissue formation cases and the therapeutic treatments respectively. If bootstrapping or permutation options were selected, the discretized data were randomly shifted to create a new sample (see next section for more detail). Discretized data were formatted adequately to be used within the CellNOptR framework. A generic inhibitory mechanism (InhM) that can potentially affect to all the TFrs in the network was also incorporated to account for all the possible processes that might not be modeled just by TF-TF interactions, such as endocytosis, apoptosis, or dephosphorylation.

The PKN was first simplified by removing all the interactions that were neither between stimuli and significant TF reporters nor between significant TF reporters. This simplified network was merged with the inferred network. New connections were added between InhM and each of the significant TFrs . We modified the CellNOptR functions so that the initial topology was transformed into a three-level Boolean model. Subsequently, the initial Boolean model was further expanded by generating all the inhibitory AND gates from all the stimuli and InhM to all the significant TFrs.

The structure optimization was performed using a genetic algorithm (GA) that aimed to determine the simplest ensemble of edges that produced the minimum difference between the discretized experimental data and the simulation results of the three-level Boolean model that only contained those edges. We aimed to minimize the difference between the (median) experimental and simulated values, penalizing complex networks, long term use of stimuli connections and unknown inhibitory mechanisms to determine the optimal network per time interval. The function that the GA aimed to minimize was the network score defined as:

$$Score = \frac{1}{N}\left(\sum_{i=1}^{NC}(x_B - x_i)^2 + NA_{Pen}(N - NC)\right) + \frac{1}{NInp}\left(size_{Pen}NSig + Stim_{Pen}SP^{(OrdT-1)}NStim + InhM_{Pen}size_{Pen}NInhM\right)$$

here $N$ is the total number of experimental observations; $NC$ is the total number of simulations in which the Boolean model converged; $x_B$ represents the simulation results from the Boolean model; $x_i$ denotes the discretized experimental results; *NAPen* is the penalties assigned to the number of non-converged simulations; *sizePen*, *StimPen* and *InhMPen*are the penalties for the total number of edges coming from TFrs, stimuli and InhM, respectively; *NInp* represents the total number of edges in the network; *NSig*, *NStim* and *NInhM* are the total number of edges originated from TFrs, stimuli and InhM, respectively; *SP* is the stimuli policy increased to penalize the appearance of long term stimuli edges and *OrdT* indicates the order of the experimental time whose structure is being optimized. Note that this scoring function is the same as the one originally proposed in CellNOptR with the addition of two new terms to account for

dynamic TF activity data. Adding a penalty to the use of long term stimuli-originated edges guarantees that the GA will search for networks that predict the observed experimental values employing TFr originated edges, especially after the cells have been exposed to the stimuli for some time. Finally, the inhibitory originated edge penalty helps to minimize their use in the final network to explain the experimental data when there are other more complex paths that could produce the same results.

Edges are represented as a string of 0, non-present, or 1, present edges. The initial present edges in the structure were randomly selected. In order to improve the computational demand, the initial structure was further updated. If contradictory edges were present, i.e., A -> B and A -| B or A -> B and A+InhM -| B, NTRACER removes both of them. If redundant edges A -| B and A+InhM -| B are present in the initial structure, NTRACER removes the one with the higher penalty for that time point.

The gates results, truth tables, were updated differently from the previous version of CellNOptR to handle three-levels in the gate output as well as the dynamic TFrs data (Supplementary Table 5). The gate outcome depends on:

a) The gate type that could contain an activation or inhibition input with single or multiple edges combined in AND or OR gates.

b) The input state indicates the input capability to modify other TFrs. The input state is active or inactive if the discretized TFr activity is equal to 1 or -1, respectively. If the discretized TFr activity is 0, but at least one condition for that given TFr is below the control, the input state is considered active. In other words, having at least one condition below control is an indication that the TFr in the untreated condition was active.

c) Input prior change indicates whether the TFr has been previously modified by the given condition or treatment, either in the current time interval or in a prior one. For instance, if the discretized dynamic trend of a TFr activity is (0, 0, 1, 0, -1), the prior input change vector would be (0, 0, 1, 1, 1). The TFr discretized dynamic trend has not been modified until the third time point, and we therefore consider that before the third time point the TFr is not capable to alter the response of any other downstream TFrs. Afterwards, the TFr is capable to modify any other TFr condition as long as the former TFr in an active state. For this example, the TFr input state is (0, 0, 1, 1, 0) and thus the TFr has a transforming capability only at the third and fourth time points.

d) Total discretized activity change for a given TFr gate depends on the effects (i.e, inhibitory, -1, or activation, 1) and total number of inputs. The total discretized activity change is the summation of all its inputs (1 for activation, -1 for inhibition). If the result of the summation was zero, the TFr is assumed to not have changed, and the previous time point value remained. If it is greater or less than 1 or -1, the value is set at 1 or -1 respectively.

e) Output prior level is added to the total discretized activity change. The output prior level is an input to the model based on the experimental discretized TFr activity data. In our example, the output prior change will be (0, 0, 1, 0). If the model predicts an increase in activity of 1 for the transition between time one (0) to time two (0), NTRACER will add to the prior output level (0) the model prediction (1), providing a final gate output of 1. Note that this will be penalized as it differs from the experimental results.

The Boolean model simulation was performed iteratively until all the levels of each TFr at each condition remained unchanged or the maximum number of iterations was reached. The different structures were ranked based on their scores and the next generation of structures was produced for a given elitism, probability of mutation, and selective pressure. Once the maximum number of generations was achieved, the simulation results were used as the initial values for the next time interval. In order to reduce the computational time, a two-level factorial design with a central point was conducted to determine the parameters that yielded the lowest score for the same number of generations. The results of the two-level factorial design were fitted to a quadratic square model with first order interactions. Significant parameters were obtained using the Akaike information criterion and those with p-value > 0.1 were filtered. Optimized parameters are given in Supplementary Table 5. Finally, as GA is a heuristic search algorithm and does not guarantee the global minimum, a total of 500 runs with different starting initial network guesses was the selected strategy to avoid being confined to a local minimum. An ensemble of the top 1% lower score networks was the reported dynamic consensus network that combines optimal and several suboptimal structures.

**Testing for edge significance in the final dynamic consensus networks**

In order to identify the significance of each edge present in the consensus networks, 1000 bootstrap samples were generated by randomly selecting the same sample size for each time point, treatment and TFr with replacement, independently of the origin of the array. These samples were used as input data for NTRACER and final dynamic networks were generated for each of the bootstrapping samples. The number of times that a given edge was present in each network for a specific time interval in the bootstrapping results divided by the total number of samples (n=1000) provided a confidence score or probability for each dynamic edge (Supplementary File 2).

Finally, we compared those dynamic edge probabilities generating from the networks obtained from the bootstrapping samples with the ones generated by a random model. 1000 samples were created by randomly shifting the data for each TFr, so that data correlation could be eliminated. The permutation samples were given to NTRACER and as before, dynamic networks were generated for each permutation sample. The probability of each dynamic edge in the permutation samples were calculated as above.

An edge was deemed significant if it was more probable to be observed compared with the random model for a given time (Supplementary Figs. 7 and 8). In practice, for each experimental condition, we selected three times the value at which the bootstrapping edges become more probable than the random edges (note that it is interval dependent). Finally, no statistical significant edges were removed from the consensus graphs. P-values for each edge were calculated based on the area under the permutation curve for the same probability as the given edge according to the bootstrapping runs (Supplementary File 2).

**Dynamic network target validation in human breast cancer tumors and lapatinib treated BT474 and SKBR3 cell line**

*22* ErbB2 positive and 58 triple negative Agilent microarrays were downloaded from TGCA (http://cancergenome.nih.gov/). Transcriptomic data were background corrected, using the normexp method, 50 units were added to avoid the presence of negative values and were further quantile normalized using the limma package[6]. A total of 143 genes are differentially expressed genes (FC>1.2 and p-value<0.01) between the ErbB2 positive genes and triple negative breast cancer tumors using limma[6]. Possible TFs that could

regulate those genes and are direct targets of ErbB2 overexpression based on the dynamic network (AP-1, STAT, SRF, E2F and YY families) were explored.

Two additional experiments of lapatinib treated BT474, BT474-J4 and SKBR3 cell lines were employed in the validation studies (E-GEOD-16179 and E-MEXP-440). The entire set of raw microarrays are not available for E-MEXP-440, so the significant genes obtained by O'Neil et al.[20] were used in that case (see the reference for details on the analysis). E-GEOD-16179 arrays were background corrected and normalized using the rma algorithm[3] from the affy package[21]. Probes below the background were filtered and genes were deemed significant if their expression level with respect to the DMSO control was statistically significant at a FC≥1.3 and an uncorrected p-value ≤ 0.1. Possible TFs that could regulate those significant genes and are direct targets of lapatinib overexpression based on the dynamic network (ELK-1, RAR, GATA and P53 families) were explored.

TF gene targets were identified in three manners. From experimentally validated targets obtained from GeneGO (MetaCore, Thomson Reuters), a list of more than 7000 interaction was compiled for the above TF families. Computationally predicted targets were extracted by exploring the promoter regions of the entire human genome, NCBI36/hg18, (from the Regulatory Sequence Analysis Tools, http://rsat.ulb.ac.be/) and the consensus mammalian promoter regions[22] between -2000 to 2000 from the TSS. Mammalian consensus and human promoter regions were investigated using MATCH[23] and FIMO[24] at three cut-off levels. For MATCH, a core score equal to 1 was always required and three matrix scores were considered 0.95, 0.99 and 0.999. In the case of FIMO, uncorrected p-values of $10^{-4}$, $10^{-5}$ and $10^{-6}$ were studied. None of the two methods can handle gaps, so consensus mammalian promoter sequences were fragmented into continuous regions before the analysis. All vertebrate PWMs for the considered TF families from TRANSFAC were used (as downloaded in 08/01/2013). Three lists for the mammalian consensus and human promoter regions respectively were created that identified genes that could potentially be regulated by a given TF according to both MATCH and FIMO at 0.95 and $10^{-4}$, 0.99 and $10^{-5}$ and 0.999 and $10^{-6}$ cut-off..

Identification of the most likely active TFs were calculated using a hypergeometric test for all experimentally and computationally obtained targets, and a z-score test for the computationally acquired targets[25]. The gene background list to compare against was the non-significant genes. 1000 bootstrapping samples were selected of size identical to the number of significant genes. Mean, median, and mode values for p-values and z-scores were recorded. Results from the three different methods were consolidated using a meta-analysis approach for the same type of experiment (i.e, E-MEXP-440 results and BT474 from E-GEOD-16179 were combined using the meta-analysis method), in which z-scores were previously translated into p-values:

$$\chi_{2k} = -2\sum_{i=1}^{k} ln(p_i)$$

were $p_i$ are the p-values and $\chi_{2k}$ is the corresponding chi-square value.. Median values were reported due to the skew of the bootstrapping results. Only results from the more conservative cut-offs, 0.999 for MATCH and $10^{-6}$ for FIMO were deemed to be representative, due to the limited number of false positives (Supplementary Files 3 and 4).

**References**


1. A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra and S. A. Sansone, *Nucleic Acids Res*, 2003, 31, 68-71.
2. C. R. Pradeep, W. J. Kostler, M. Lauriola, R. Z. Granit, F. Zhang, J. Jacob-Hirsch, G. Rechavi, H. B. Nair, B. T. Hennessy, A. M. Gonzalez-Angulo, R. R. Tekmal, I. Ben-Porath, G. B. Mills, E. Domany and Y. Yarden, *Oncogene*, 2012, 31, 907-917.
3. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, *Biostatistics*, 2003, 4, 249-264.
4. B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed, *Bioinformatics*, 2003, 19, 185-193.
5. B. S. Carvalho and R. A. Irizarry, *Bioinformatics*, 2010, 26, 2363-2367.
6. G. K. Smyth, *Stat Appl Genet Mol Biol*, 2004, 3, Article3.
7. D. S. Chekmenev, C. Haid and A. E. Kel, *Nucleic Acids Res*, 2005, 33, W432-437.
8. J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt and P. K. Sorger, *Molecular systems biology*, 2009, 5, 331.
9. C. Terfve, T. Cokelaer, D. Henriques, A. MacNamara, E. Goncalves, M. K. Morris, M. van Iersel, D. A. Lauffenburger and J. Saez-Rodriguez, *BMC systems biology*, 2012, 6, 133.
10. D. Marbach, J. C. Costello, R. Kuffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, D. Consortium, M. Kellis, J. J. Collins and G. Stolovitzky, *Nat Methods*, 2012, 9, 796-804.
11. S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein and J. A. Stamatoyannopoulos, *Cell*, 2012, 150, 1274-1286.
12. D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano and G. Stolovitzky, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, 107, 6286-6291.
13. D. Marbach, J. C. Costello, R. Kuffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky and D. Consortium, *Nature Methods*, 2012, 9, 796-+.
14. P. Geladi and B. R. Kowalski, *Anal Chim Acta*, 1986, 185, 1-17.
15. P. K. Kreeger, *Sci Signal*, 2013, 6, tr7.
16. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera and A. Califano, *BMC Bioinformatics*, 2006, 7 Suppl 1, S7.
17. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, *PLoS Biol*, 2007, 5, e8.
18. J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink and E. D. Jarvis, *Bioinformatics*, 2004, 20, 3594-3603.
19. N. M. Luscombe, M. M. Babu, H. Y. Yu, M. Snyder, S. A. Teichmann and M. Gerstein, *Nature*, 2004, 431, 308-312.
20. F. O'Neill, S. F. Madden, S. T. Aherne, M. Clynes, J. Crown, P. Doolan and R. O'Connor, *Molecular Cancer*, 2012, 11.
21. L. Gautier, L. Cope, B. M. Bolstad and R. A. Irizarry, *Bioinformatics*, 2004, 20, 307-315.
22. X. H. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis, *Nature*, 2005, 434, 338-345.
23. A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis and E. Wingender, *Nucleic Acids Research*, 2003, 31, 3576-3579.
24. C. E. Grant, T. L. Bailey and W. S. Noble, *Bioinformatics*, 2011, 27, 1017-1018.
25. S. J. H. Sui, J. R. Mortimer, D. J. Arenillas, J. Brumm, C. J. Walsh, B. P. Kennedy and W. W. Wasserman, *Nucleic Acids Research*, 2005, 33, 3154-3164.