Supplementary Materials for:

Network quantification of EGFR signaling unveils potential for targeted combination therapy

Bertram Klinger[1,2,*], Anja Sieber[1,*], Raphaela Fritsche-Guenther[1,*], Franziska Witzel[1,2],
Leanne Berry[3], Dirk Schumacher[1], Yibing Yan[4], Pawel Durek[1,2,5], Mark Merchant[3], Reinhold Schäfer[1,5], Christine Sers[1], Nils Blüthgen[1,2,**]

1 Laboratory of Molecular Tumour Pathology, Institute of Pathology, Charite Universitätsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany
2 Institute for Theoretical Biology, Humboldt University Berlin, Invalidenstr. 43, 10115 Berlin, Germany
3 Department of Translation Oncology, and
4 Oncology Biomarker Development, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA
5 Deutsches Konsortium f˙ur Translationale Tumorforschung, Germany
∗ Joint first authors
∗∗ Corresponding author: nils.bluethgen@charite.de

Content

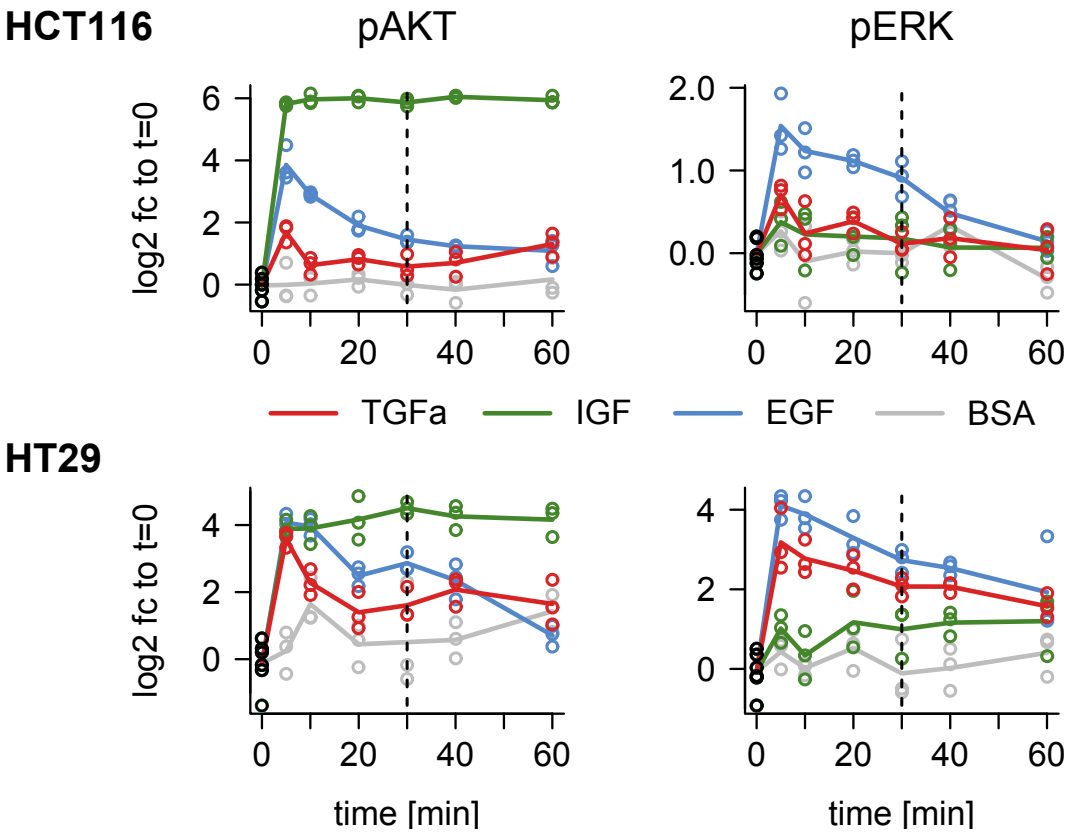


Figure S1. Phosphorylation kinetics after stimulation. Luminex measurements of HCT116 (upper row) and HT29 (lower row) treated for 5-60 min with the indicated ligands demonstrate for the relevant outputs that TGFα(red) as well as IGF (green) signalling is near post treatment steady state after about 30 min (dashed vertical line). The lines indicate the medium log2 fold change to the average control measurements at t=0 (black circles). ($n >= 3$).

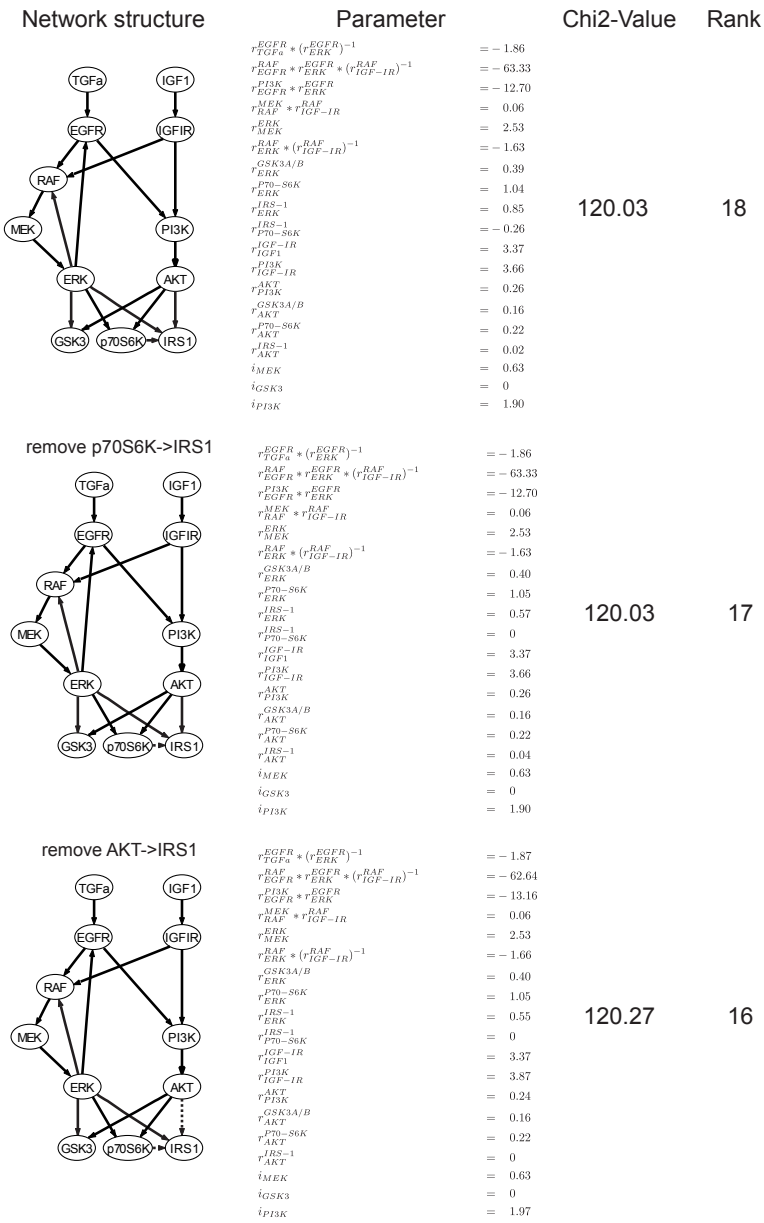
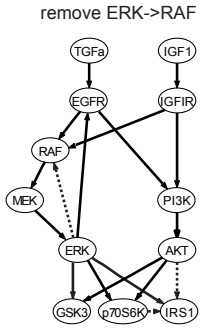
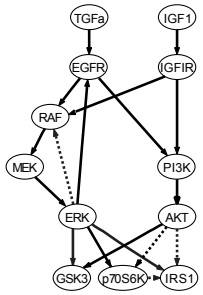


Figure S2. Model reduction example. Stepwise reduction of the model and the corresponding identifiable parameter combinations exemplified by the cell line HT29. First column denotes the network structure with dashed lines indicating removed links. Second column depicts the parameterization (natural logarithm) after fitting the model to the structure. Thereby the first parameter values are taken from the best fit of 10000 random initializations. For all further reduced models the best inital fit is given as starting point (except for the removed link which is kept at zero). the last two columns show the chi-squared value and the number of identifiable parameters respectively.



$r_{TGFa}^{EGFR} * (r_{ERK}^{EGFR})^{-1} = -1.82$
$r_{EGFR}^{RAF} * r_{ERK}^{EGFR} * (r_{IGF-IR}^{RAF})^{-1} = -73.81$
$r_{EGFR}^{PI3K} * r_{ERK}^{EGFR} = -15.69$
$r_{RAF}^{MEK} * r_{IGF-IR}^{RAF} = 0.06$
$r_{MEK}^{ERK} = 2.29$
$r_{ERK}^{RAF} * (r_{IGF-IR}^{RAF})^{-1} = 0$
$r_{ERK}^{GSK3A/B} = 0.40$
$r_{ERK}^{P70-S6K} = 1.05$
$r_{ERK}^{IRS-1} = 0.55$
$r_{P70-S6K}^{IRS-1} = 0$
$r_{IGF1}^{IGF-IR} = 3.37$
$r_{IGF-IR}^{PI3K} = 3.95$
$r_{PI3K}^{AKT} = 0.24$
$r_{AKT}^{GSK3A/B} = 0.16$
$r_{AKT}^{P70-S6K} = 0.21$
$r_{AKT}^{IRS-1} = 0$
$i_{MEK} = 0.56$
$i_{GSK3} = 0$
$i_{PI3K} = 2.01$

121.77 15



$r_{TGFa}^{EGFR} * (r_{ERK}^{EGFR})^{-1} = -1.81$
$r_{EGFR}^{RAF} * r_{ERK}^{EGFR} * (r_{IGF-IR}^{RAF})^{-1} = -76.34$
$r_{EGFR}^{PI3K} * r_{ERK}^{EGFR} = -13.64$
$r_{RAF}^{MEK} * r_{IGF-IR}^{RAF} = 0.06$
$r_{MEK}^{ERK} = 2.29$
$r_{ERK}^{RAF} * (r_{IGF-IR}^{RAF})^{-1} = 0$
$r_{ERK}^{GSK3A/B} = 0.40$
$r_{ERK}^{P70-S6K} = 1.15$
$r_{ERK}^{IRS-1} = 0.55$
$r_{P70-S6K}^{IRS-1} = 0$
$r_{IGF1}^{IGF-IR} = 3.37$
$r_{IGF-IR}^{PI3K} = 3.26$
$r_{PI3K}^{AKT} = 0.29$
$r_{AKT}^{GSK3A/B} = 0.16$
$r_{AKT}^{P70-S6K} = 0$
$r_{AKT}^{IRS-1} = 0$
$i_{MEK} = 0.56$
$i_{GSK3} = 0$
$i_{PI3K} = 1.92$

123.50 14

Figure S2. (continued) Model reduction example.

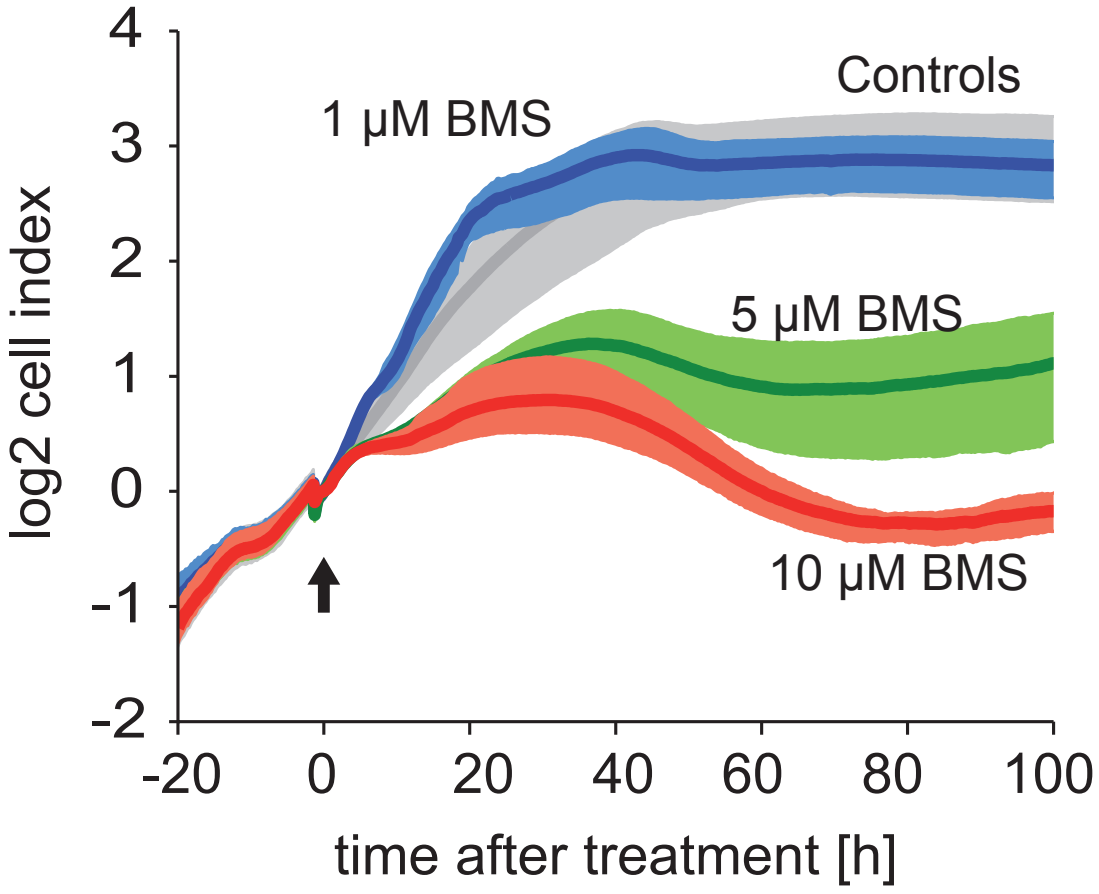


Figure S3. Growth curves for different concentrations of IKK inhibitor BMS34554. Xcelligence measurements of HCT116 cell growth in response to 3 different concentrations of BMS345541 or its solvent control DMSO (grey) at time 0 (arrow) grown in full serum. Shown is the maximal range of $n >= 3$ independent wells along with the mean (thick line). Note that for concentration of 10μM cells even reach a lower cell index than at time 0 indicating cell death.

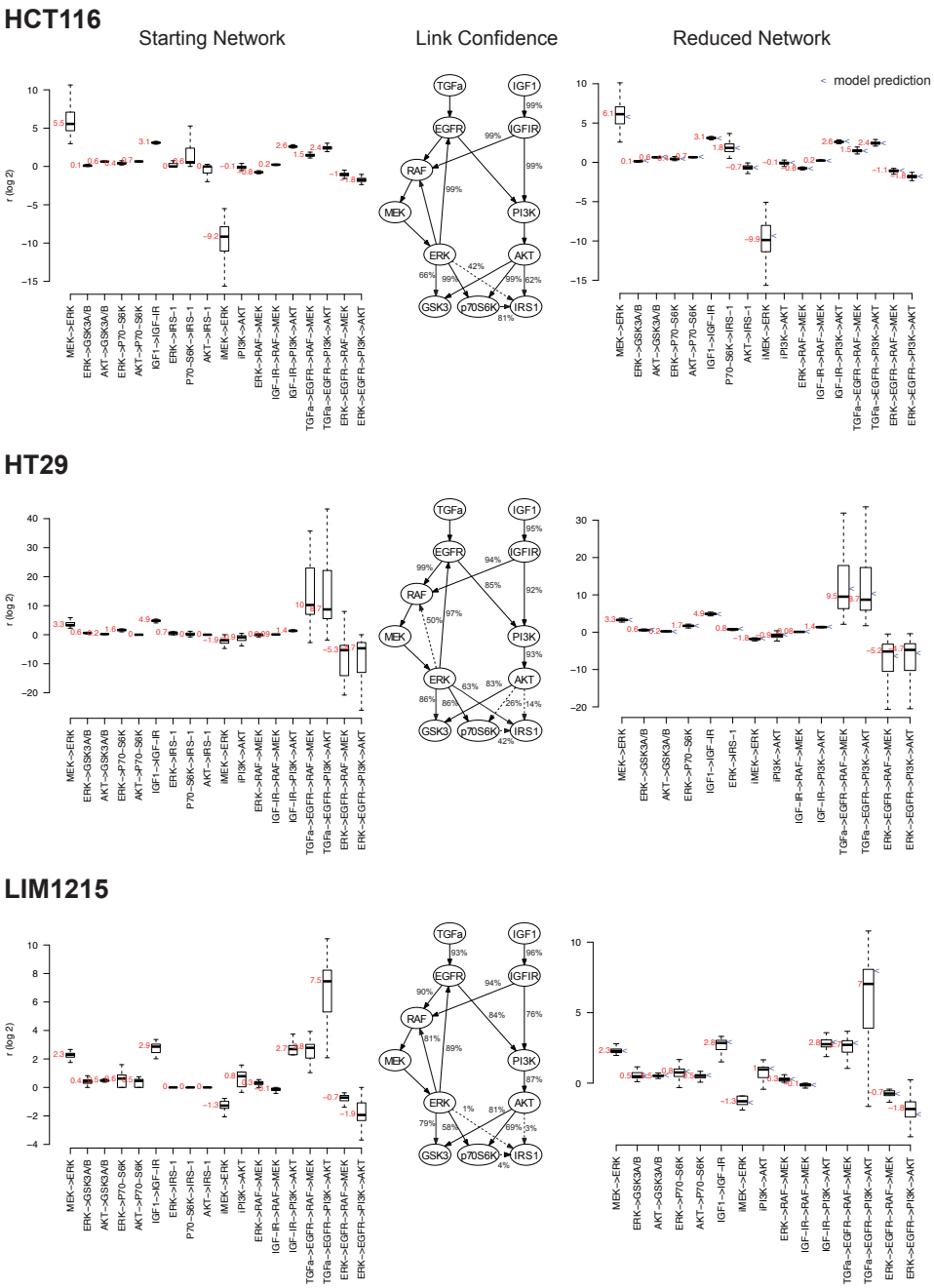


Figure S4. Parameter robustness. Bulk parameter distribution for originally 100 data sets simulated from the original data by adding random noise taking either the starting network structure (left) or the reduced structure (right). Red numbers indicate the median value and the blue arrow on the right boxplot indicates the parameter estimation from the original model. The networks contain the percentage of non-zero links found in the starting network simulations if less than 100%. The correlation for removed links (dashed lines) and lower confidence is high. Outliers corresponding to fits with chi-squared values higher than 10 times the median chi-sqared value were removed.

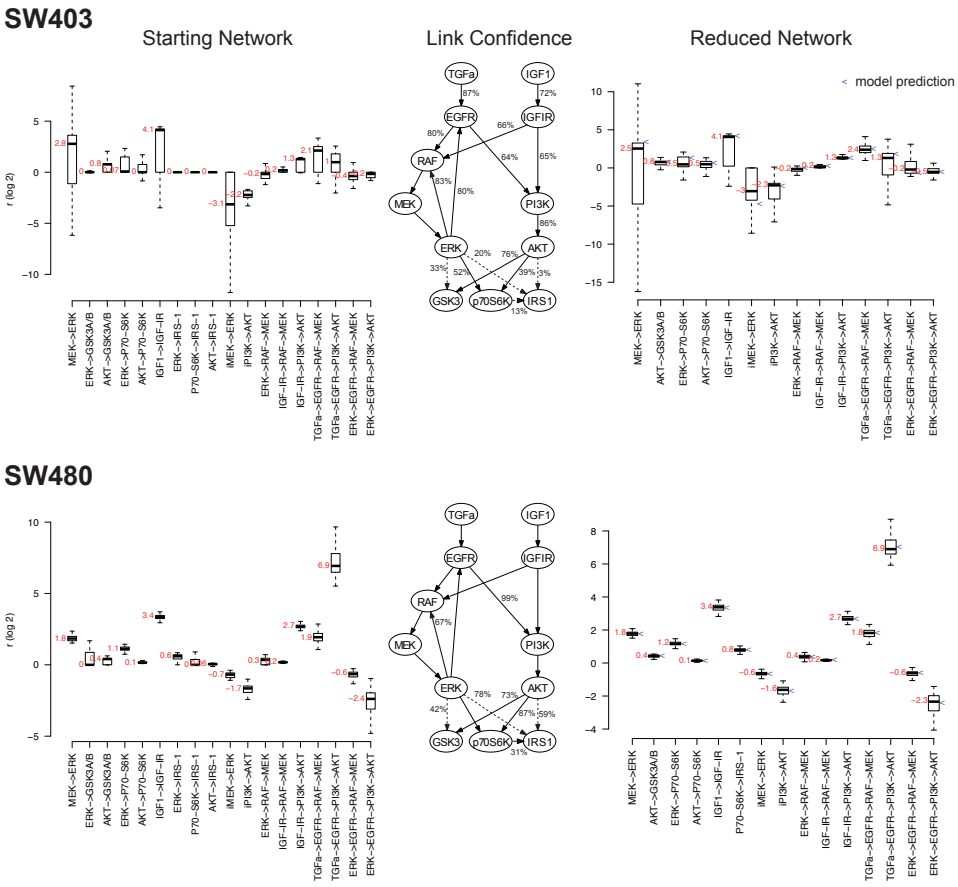


Figure S4. (continued) Parameter robustness.

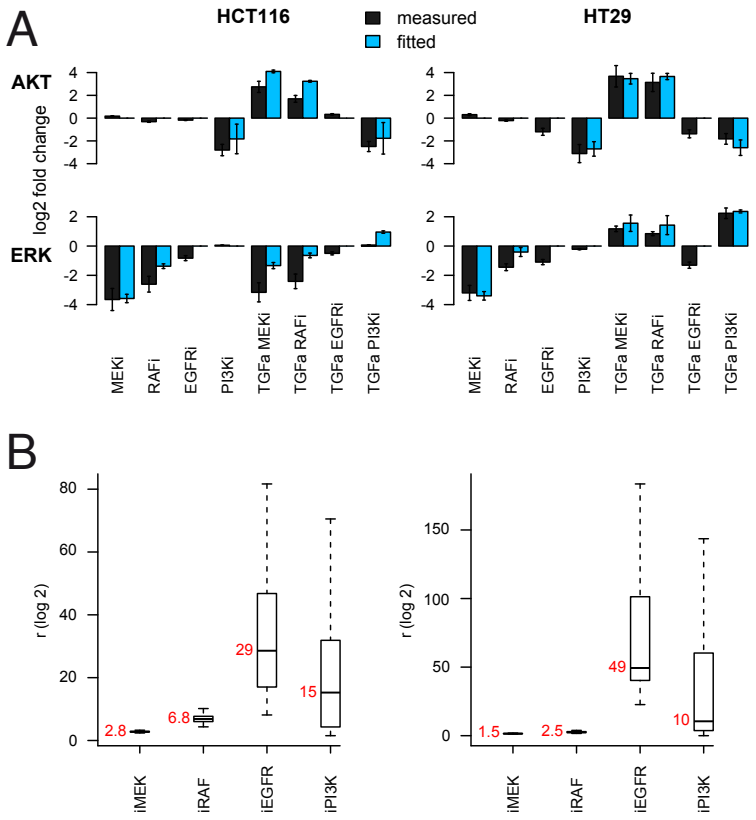


Figure S5. Altered parameters for prediction of selected inhibitor combinations. A Data (black bars) used to determine the parameters for inhibitors against RAF, EGFR and MEK and PI3K. RAF (Sorafenib) and EGFR (Gefitinib) inihibitors where newly introduced to the model and fitted (blue bars). Additionally AZD6244 and PI3K were re-fitted, see also FIG. 6A in the main manuscript. Fitting was done by determining the maximum likelihood fit for the four parameters, leaving all other parameters unchanged. To obtain confidence intervals for the fitted parameters, we used the fitted values of these parameters for all model fits obtained from data with simulated noise (see also Fig. S6). The standard deviation for the fitted values is indicated with error bars. For the data, the average of two replicates is shown together with error bars indicating the standard deviation calculated by the error model from these replicate measurements. **B** Log-2 parameters for the inhibitors against MEK, RAF, EGFR and PI3K obtained by the fit discribed in **A** used to predict double perturbations. Distribution was derived by starting with the parameters derived for the initial screen that were 100 times noised.

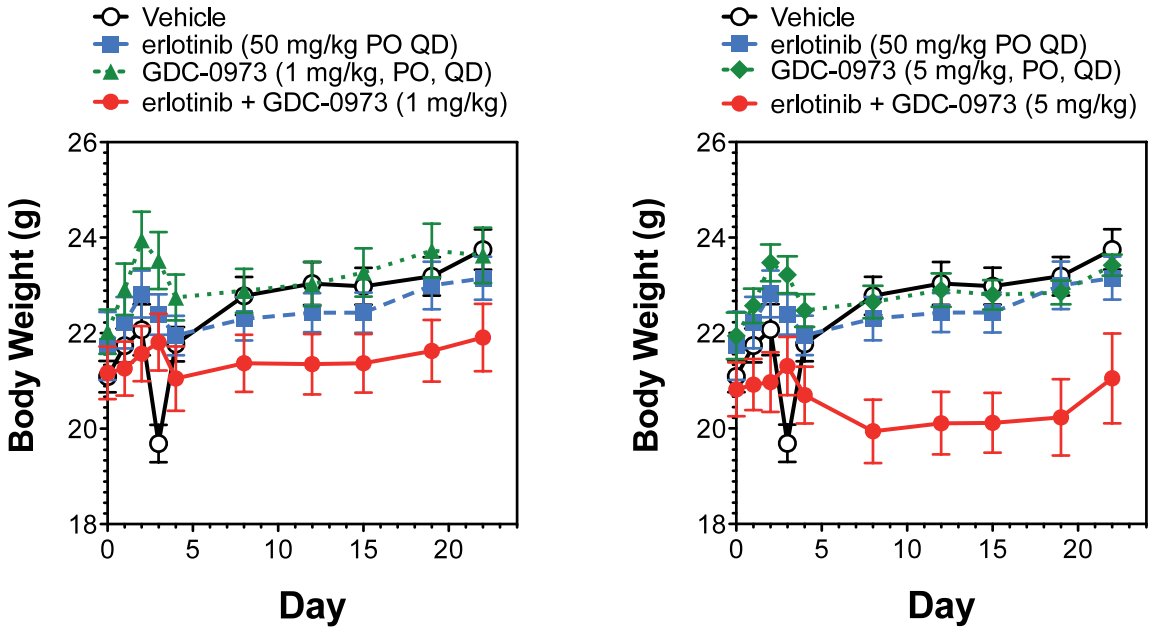


Figure S6. Body weigth of DLD-1 xenograft animal models during drug treatment. Body weight of DLD-1 xenografted nude mice treated with the indicated drugs (described in Fig. 6 in the main manuscript). Mice show no body weight loss during the treatment, however during combinatorial treatment no weight gain can be observed.

1 Modeling perturbation of signaling networks with MRA

To model the effect of a perturbation at one node on the other nodes in the network, we apply the methodology of Modular Response Analysis (MRA). In the following we briefly review this methodology, and introduce extensions that are required to apply it in the context of our study. To illustrate the derived equations and approaches, we consider three relatively simple networks, depicted in Fig. S7-S9, and in addition to general derivations, we derive explicit equations for those. Readers that are interested in the detailed derivation of MRA may refer to the work of Boris Kholodenko and coworkers [1, 2] and our previous work [3]. An illustration of the parameters and their identifiable combinations for the networks describing HT29 cells can be found in Fig. S5.

Definition and Derivation

MRA is an analysis of the steady-state responses to parameter perturbations that considers the first, linear terms in the Taylor expansion of the dependence of steady states on parameter. In the following, the vector $\mathbf{x}$ describes the steady-state of variables of an ODE system, with each of them depending on a subset of paramters stored in the parameter vector p. If we now choose the subset of parameters such that each x_i depends only on the parameter p_i directly, then we can choose for simplicity a parameterisation such that $\frac{p_j}{x_i}\frac{\partial x_i}{\partial p_j} = \delta_{ij}$. Once a single parameter is perturbed by an infinitesimal amount, the response of the network can be calculated by:

$$\frac{p_j}{x_i}\frac{dx_i}{dp_j} = \frac{p_j}{x_i}\frac{\partial x_i}{\partial p_j} + \sum_{k\neq i}\frac{x_k}{x_i}\frac{\partial x_i}{\partial x_k}\frac{p_j}{x_k}\frac{dx_k}{dp_j}\,. \tag{1}$$

Since per definition $\frac{p_j}{x_i}\frac{\partial x_i}{\partial p_j} = \delta_{ij}$, one can rewrite this equation as:

$$-\delta_{ij} = -\frac{p_j}{x_i}\frac{dx_i}{dp_j} + \sum_{k\neq i}\frac{x_k}{x_i}\frac{\partial x_i}{\partial x_k}\frac{p_j}{x_k}\frac{dx_k}{dp_j} \tag{2}$$

MRA provides interpretation and names for the scaled derivatives. First, the total derivatives are called global response coefficents,

$$R_j^k = \frac{p_j}{x_k}\frac{dx_k}{dp_j}\,. \tag{3}$$

These coefficients describe the response of the entire network to a particular perturbation, when the network is allowed to relax into a new steady state. The entries of R_j^k relate to measurable quantities, such as log fold-changes of expression or activity of the steady state after perturbation. Second, MRA defines local response coefficients. In contrast to

the global response matrix, it provides information about the direct interactions between variables. These can typically not be measured and need to be inferred. It is defined as

$$r_k^i = \begin{cases} \frac{x_k}{x_i}\frac{\partial x_i}{\partial x_k} & \text{for } i \neq k \\ -1 & \text{otherwise} \end{cases} , \tag{4}$$

and can be interpreted as the normalized Jacobian matrix of the underlying ODE system. Using these definitions, one can rewrite equation 2 as:

$$\sum_k r_k^i R_j^k = -\delta_{ij} \, . \tag{5}$$

Thus, the rows of the matrix $\mathbf{r}$ and the columns of $\mathbf{R}$ are orthogonal for $i \neq j$, and we can write this down in a matrix formalism as:

$$\mathbf{rR} = -I \, , \tag{6}$$

By inverting matrix $\mathbf{r}$, this can be rewritten as:

$$\mathbf{R} = -\mathbf{r}^{-1} \tag{7}$$

If one perturbs the system at a single or several nodes, one can define a perturbation vector, Δp, which contains non-zero entries only at nodes that are directly perturbed. The reaction (in terms of logarithmic fold-change) of all nodes in the network to this perturbation can then be calculated by multiplying the global response matrix $\mathbf{R}$ or the inverse of the local response matrix $-\mathbf{r^{-1}}$ with this perturbation vector. Thus, the global response of node k to a perturbation in node j is defined as:

$$R_j^k = (-\mathbf{r}^{-1})_j^k \, \Delta p_j \tag{8}$$

In Figure S7 the relationship of local and global response matrix is exemplified on a three-tiered network with a feedback from node C to A. This feedback now causes the global response matrix to be fully connected although only three non-zero entries are present in the local response matrix. This means when perturbing any node the steady state of all nodes will be changed. Since only the global response matrix but not the local response matrix can be measured, an intuitive derivation of the direct links will be hardly possible when feedbacks structures occur.

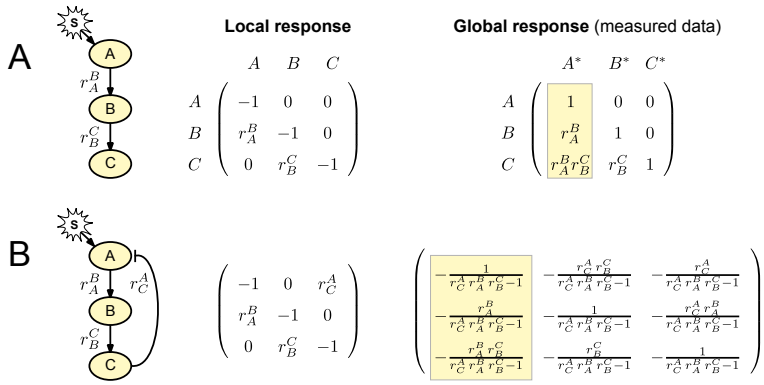


Figure S7. Visualizing the basic principles of MRA. (A) The left graph shows a linear network consisting of three nodes. In the middle the weights of the network are represented in the local response matrix **r** with the ingoing node indicated in the column and the outgoing node indexed in the row. On the right the observable response of the network called global response matrix is shown. In this simple network only the effect of a perturbation in A* on C is not executed directly and the intuitive derivation of the underlying local response matrix is possible (i.e. multiply along the path). Asteris denote directly perturbed nodes. **(B)** Shows how the local and global response matrix changes when a feedback from C to A is added to the network from **(A)**. Although the local response matrix is still sparse the global response matrix, which represents the actual measurement results is completely filled. Thus addition of a feedback makes an intuitive deduction of the underlying network extremly difficult. Note when only perturbing at node A (as indicated in the network) only the first column of **R** can be measured (marked in yellow). To obtain the complete matrix every node has to be perturbed systematically and the response of all nodes measured. Note that perturbation strength was set to 1 and such omitted in these examples.

Integration of different perturbation types

We now consider the effect of two principle perturbation types commonly applied on signaling pathways: ligand-driven stimulation and small-molecule inhibitor-mediated inhibition.

Stimulation of node j by a ligand can be simulated by encoding the strength in Δp_j as in equation 8. However, for our further analysis it proved crucial to keep perturbation vectors binary and to transfer any non binary information to the local response matrix. Consequently, we use local response coefficients for the ligands, set the perturbation to 1, and fit the corresponding values in **r** to yield an estimate of the induced fold change at the receptor level (see Fig. S8).

To be able to include inhibition by small-molecule inhibitors, two effects have to be

considered: (i) the inhibition negatively affects the innate signal sent out by the targeted molecule and (ii) it reduces the ability of the inhibited node to relay upstream signals to downstream nodes. The first effect is modelled by a negative perturbation of the targeted node m specified in a seperate perturbation vector Δp^{inh} with $\Delta p_m^{inh} \in (-\infty, 0]$. The other effect, the reduced relay of the incoming signal, is modeled by multiplying an inhibitory term $i_m \in [0..1]$ to all outgoing local response coefficients. This yields a modified matrix $\tilde{\mathbf{r}}$. The strength of i_m can be calculated from the corresponding entry in Δp^{inh}. Since Δp_m^{inh} is defined as the logarithm of the inhibitor-induced fold change on downstream pathways, the relationship of these two coupled inhibitor effects is defined as

$$i_m = \exp(\Delta p_m^{inh}) \quad . \tag{9}$$

The response of node k to a combined perturbation consisting of a stimulus at node j and an inbibition at node m can thus be modeled by

$$R_{j,m}^k = (\tilde{\mathrm{r}}^{-1})_j^k \, \Delta p_j + (\mathrm{r}^{-1})_m^k \, \Delta p_m^{inh} \quad . \tag{10}$$

Thereby the coefficient i_m included in $\tilde{r}$ dampens the stimulation and Δp_m^{inh} reduces the basal activity of the node or, in the case of an negative effect on a downstream node, relieves the basal repression (Fig. S8). Note that not all repressed nodes exhibit basal activity ,e.g. receptors are usually inactive when not externally stimulated. For this case the second term in equation 10 is zero.

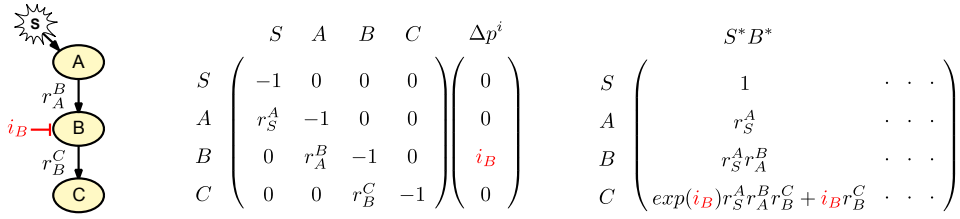


Figure S8. Integration of simultaneous stimulation and inhibition in MRA. LEFT: The network from Fig. S7A with the additional feature that node B is inhibited (marked in red). MIDDLE: Local response matrix and the inhibitory perturbation vector. Note that we have now introduced a new node S whose coefficient r_S^A contains the perturbation strength of stimulus S and therefore the corresponding stimulatory perturbation vector entry is set to 1. RIGHT: First column of the global response matrix for simultaneous stimulation and inhibition. The global response from stimulus S to node C is now a sum of the stimulatory pathway damped by the inhibitor and the reduction of the basal activity of r_B^C is realized by multiplication of i_B.

Removing non-identifiable parameters

For each perturbation and for each measured node, the model prediction (in terms of log-fold change of the unperturbed steady state) can be written as a function of the negative inverse of the matrix r and the parameters for inhibition. As already apparent from the examples in Fig S7 and S8, the equations for the response of the network typically contains products of entries of r along paths in the network (see also [4]). We call these products the path strengths P and define them as

$$P_l = \prod_{(j,k) \text{ edges in path l}} r_j^k \quad . \tag{11}$$

Depending on the experimental design (i.e. which nodes are perturbed, which nodes are measured etc), it is often the case that some r_k's only occur in combination in these paths, never alone.

For example in the network in Fig S9 the node B is not measured and signals down to two nodes. Therefore r_A^B occurs only in combination as can be seen at the entries of the measured entries in the global response matrix. Since more parameters than entries in R have to be fitted, an infinite number of parameter realizations can be found. This might prolong fitting procedures and can in some cases even lead to suboptimal parameter values. In addition the fitted values of those non-identifiable parameters are themselves meaningless as they can be only interpreted in combination. One intuitive of several (mathematically equaly valid) solutions to circumvent this, would be to re-define a new module consisting of A+B with one ingoing and outgoing node thereby effectively fitting the identifiable parameter combinations $r_A^B r_B^C$ and $r_A^B r_B^D$.

In the course of our modeling procedure nodes of the network were automatically removed. Thereby structural non-identifiabilities will be occuring and disappearing urged

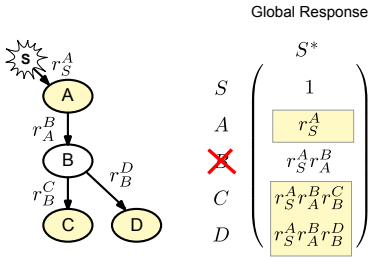


Figure S9. Structural non-identifiability. LEFT: 4-node network stimulated at A. Nodes with a yellow background indicate measured nodes. RIGHT: Global responses to the stimulus S^* with the stimulus strength included as a response coefficient. Note that due to the response of node B not being measured the 4 parameters have to be fitted from 3 entries introducing non-identifiabilities.

us to find a means to detect those structures and to fit only identifiable parameter combinations. We therefore developed a methodology to systematically reduce these so-called structural non-identifiables. For this, we first re-parameterized the equation for the response in terms of the path strengths:

$$R_j^k = f(P)_j^k \, \Delta p_j \, . \tag{12}$$

Then, we expressed the path strength of each P_l as sums of r's by taking the logarithm:

$$\log P_l = \sum_{j,k} a_j^k \log r_j^k \, . \tag{13}$$

With a_j^k standing for the exponent of r_j^k for path P_l (usually 0 if the path does not go from edge j tp k, and 1 otherwise). This is a set of linear equations, which can be represented in matrix form:

$$0 = [\mathbf{A}, -\mathbf{I}] \times (\log r_1, \ldots \log r_M, \log P_1, \ldots \log P_N)^T \, , \tag{14}$$

where $\mathbf{A}$ contains the a_j^k's and $\mathbf{I}$ is the identity matrix whose size equals the total number of paths. By applying the Gaussian elimination to $[\mathbf{A}, -\mathbf{I}]$, one arrives at the reduced row echelon form. The resulting matrix can be separated into 4 parts:

$$\mathbf{G} = \left(\begin{array}{c|c} A' & G_1 \\ \hline 0 & G_2 \end{array} \right) \, . \tag{15}$$

The number of rows in A' equals the number of independent parameters. Together with G_1 it provides rules how to construct them, and G_2 identifies shared or equal paths.

For the simple example given in Fig. S9. we can define $\mathbf{A}$ from the global response to stimulus S as

$$\mathbf{A} = \left(\begin{array}{c|cccc} & r_S^A & r_A^B & r_B^C & r_B^D \\ \hline P_1 & 1 & 0 & 0 & 0 \\ P_2 & 1 & 1 & 1 & 0 \\ P_3 & 1 & 1 & 0 & 1 \end{array} \right) \, . \tag{16}$$

After applying the Gaussian elimination algorithm one arrives at:

$$\mathbf{G} = \left(\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{array} \right) \, . \tag{17}$$

Since in this example the rank of $\mathbf{A}$ is equal to its number of rows, the resulting $\mathbf{G}$ contains only the upper two parts A' and G_1. A' suggests to reparameterize the response coefficients into 3 combinations k:

$$\begin{aligned} k_1 &= r_S^A = P_1 \\ k_2 &= r_A^B r_D^B = \frac{P_3}{P_1} \\ k_3 &= \frac{r_B^C}{r_D^B} = \frac{P_2}{P_3} \, . \end{aligned}$$

For the fitting procedure the paths will now be taken as parameters to fit the global response matrix to the experimental data. After the fitting procedure is finished the best fitting paths can then be used to reconstruct the identifiable combination of local response coefficients r_i from the paths.

As long as the entries in the global response matrix consist of pure products $R_j^k = P_l$ as shown in the example in Fig. S9. However through feedbacks (Fig. S7B), inhibitors acting downstream of a stimulus (Fig. S8) or branching paths that are converging again, sums are introduced which will result in one R_j^k being represented by more than one P_l. The experiments should therefore be designed such, that all paths are identifiable. Next to combinations of perturbations always the effect of single perturbations should be measured and feedback structures and branch-merge patters should be broken by perturbations within. Following these rules, non-identifiable paths will not occur during the fitting procedure.

As long as links of the network are only removed and not added as in our work, this approach is valid if the rules were applicable for the starting network. However if one plans to randomy add links an introduction of non-identifiable paths is likely to occur. One can simply extend our methodology to even this approach by applying an additional non-identifiability analysis on the path level with respect to the measured global responses. An example of identifiabile parameter combinations for the models describing the HT29 cells during model prooning is shown in Fig. S5.

Parameter estimation

After reducing the number of parameters such that the model becomes structurally identifiable, we could set up a strategy to estimate these parameters from the data. For the fitting procedure, we used a maximum likelihood approach. For each measurement, the logarithm of the likelihood for given P and $\mathbf{\Delta p}$ is (without prefactors):

$$L(\mathbf{P}, \mathbf{\Delta p}|\mathbf{R}) \approx -\chi^2 = \sum_{i=1}^{N} \left(\frac{R_j^k - f(P)_j^k \, \Delta p_j}{\epsilon_j^k} \right)^2 \tag{18}$$

Thereby entries of $\mathbf{P}$ can only be estimated numerically using the above defined log-likelihood function.

Until this point, we have assumed that the network structure is known, i.e. which of the entries of $\mathbf{r}$ are zero. This is typically the case for very well studied signaling pathways. However, if the network structure is unclear or may require estimation, one can iteratively modify the structure. In this study we started with an detailed literature model which was then reduced. Alterations of the network structure was judged based on whether they contributed to the fit or not. More specificially, when removing links, it is assumed that the fit in terms of the likelihood worsens due to the removed complexity

of the model. Thus, a test was chosen that quantifies whether the observed change in likelihood is because the less complex model (i) is not able to fit the data, or (ii) can be termed as minimal model that is required for explaining the data. While there are criteria that penalize model complexity based on information theory, we decided to use the likelihood ratio test that provides a p-value for each model refinement.

In order to avoid local minima, we tried to obtain an estimate near the global optimimum (Monte Carlo 10000 realisations) for the starting network. Subsequently we applied local error minimization by the Levenberg-Marquard algorithm to all networks reduced by one link starting with the parameter set derived from the initial Monte Carlo simulation. The links were then reduced in a greedy hill climbing algorithm until the network can not be further reduced (see Fig. S10). Finally, after the final network structure was derived, identifiable response coeficients and combinations thereof where calculated from the fitted path parameters using the relationship derived from the Gaussian elimination matrix G.

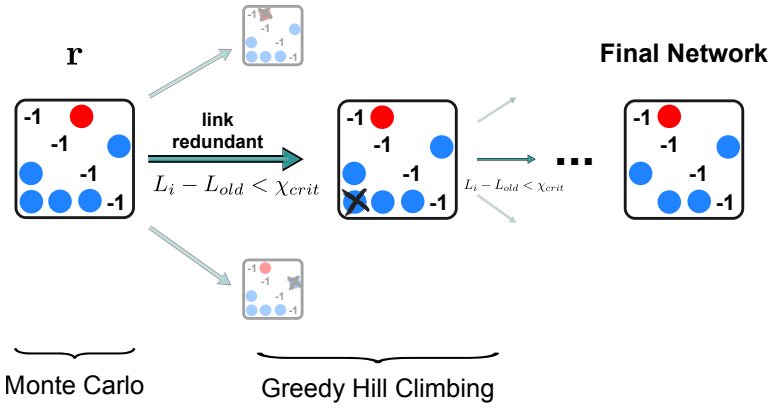


Figure S10. Parameter fitting procedure. After deciding for a suitable starting network the non-identifiable parameters are thoroughly searched for an optimum in a Monte Carlo simulation (iterations=10^4). The likelihood of the best fit L_{old} is then compared to the best fit of all network realizations containing one node less. Thereby the new fit starts with the old parameter setting and is allowed to alter it to the best nearby optimum. If the difference is below the threshold ($p = 0.05$ with χ^2 distribution under consideration of the number of dfs (usually 1)) the reduced network is taken as the new starting network that is compared to models with yet another link removed. This parameter reduction by the so called greedy hill climbing procedure is continued until all $L_i - L_{old} > \chi_{crit}$. For all parameter estimations the Levenberg-Marquardt algorithm was employed.

Error model derivation

The standard deviations of the data used for model fitting were taken from an error model that we trained on replicate measurements.

In particular, we used an error model consiting of additive and multiplicative error. The muliplicative error was calculated by averaging the Coefficients of Variations (CV) of replicate measurements, and the additive component was taken from background measurements. Subsequently, we generated for each antibody k and each cell line one error model:

$$\text{error}_k = \text{background}_k + \text{CV}_k \cdot \text{signal}_k \tag{19}$$

In case signals were below twice the background, they were discarded in estimating the CV. If less than two different replicate measurements remained to calculate the CV we set the CV to a value of 30%. To prevent overfitting the lower threshold of the CV was set to 10%.

2 Detection and annotation of Single Nucleotide Variations(SNVs)

For SNV detection we performed variant calling by the IonTorrent Suite, Version 2.1 Variant Caller (Life Technologies). Heterozygous positions with minor allele frequencies below 10% were assumed as homozygous. Accordingly the minor allele was removed from further consideration. In addition we also removed SNVs in non-coding regions, synonymous SNVs, as the evidence for the SNVs to be true or evidence for an impact on cancer development was reasonably low. A non-natural SNV was defined as having impact on cancer, if the mutation was predicted to be deleterious by at least three predictors: SIFT [5], PolyPhen [6], LRT [7], Mutation-Taster [8] or predicted to cause a Nonsense Mediated mRNA Decay [9]. A SNV was defined as natural variant if the allele was found in the 1000 Genome Project [10].

The vcf-files and a list of SNVs with an extended annotation of the SNV are available ST1. Mutation patterns were compared to verified mutations as published by the Cancer Genome Project (`http://www.sanger.ac.uk/genetics/CGP/`). The annotations were obtained from UCSC [11], HGNC [12], COSMIC v.55 [13], dbSNP v.135 [14], Uniprot release 2011_05 [15]) and OMIN (`http://omim.org/`). Pathway annotations were obtained from GO-Ontology [16], KEGG [17], Pathway Interaction Database [18] and Reactome [19], as available by December 2010. Natural allele frequencies were taken from the 1000 Genomes Project [10], while the number of tumor-patients caring the respective mutations was obtained from publicly availably sequencing data available from The Cancer Genome Atlas (TCGA; `http://cancergenome.nih.gov/`). Prediction of the impact of a SNV on protein function is based on precompiled data from dbNSFP [20]. Nonsense Mediated mRNA Decay was predicted according to criteria proposed by Nagy and Maquat [9].

References

1. Bruggeman F, Westerhoff H, Hoek J (2002) Modular response analysis of cellular regulatory networks. Journal of theoretical ... .

2. Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag E, Westerhoff HV, et al. (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. Proc Natl Acad Sci USA 99: 12841–6.

3. Stelniec-Klotz I, Legewie S, Tchernitsa O, Witzel F, Klinger B, et al. (2012) Reverse engineering a hierarchical regulatory network downstream of oncogenic kras. Molecular Systems Biology 8: 601.

4. Kholodenko BN, Hoek JB, Westerhoff HV, Brown GC (1997) Quantification of information transfer via cellular signal transduction pathways. FEBS Lett 414: 430–4.

5. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. Nat Protoc 4: 1073–81.

6. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–9.

7. Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. Genome Res 19: 1553–61.

8. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) Mutationtaster evaluates disease-causing potential of sequence alterations. Nat Methods 7: 575–6.

9. Nagy E, Maquat LE (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects rna abundance. Trends in biochemical sciences 23: 198–9.

10. Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–73.

11. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The ucsc genome browser database: update 2011. Nucleic Acids Res 39: D876–82.

12. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA (2011) genenames.org: the hgnc resources in 2011. Nucleic Acids Res 39: D514–9.

13. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 39: D945–50.

14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbsnp: the ncbi database of genetic variation. Nucleic Acids Res 29: 308–11.

15. Consortium U (2012) Reorganizing the protein space at the universal protein resource (uniprot). Nucleic Acids Res 40: D71–5.

16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–9.

17. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) Kegg for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109–14.

18. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) Pid: the pathway interaction database. Nucleic Acids Res 37: D674–9.

19. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619–22.

20. Liu X, Jian X, Boerwinkle E (2011) dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. Hum Mutat 32: 894–9.