

# Comparative Network Reconstruction using mixed integer programming

Evert Bosdriesz<sup>1</sup>, Anirudh Prahallad<sup>1</sup>, Bertram Klinger<sup>2,3</sup>, Anja Sieber<sup>2,3</sup>, Astrid Bosma<sup>1</sup>, René Bernards<sup>1</sup>, Nils Blüthgen<sup>2,3,4</sup> and Lodewyk F. A. Wessels<sup>1,5,\*</sup>

<sup>1</sup>Division of Molecular Carcinogenesis, The Oncode Institute, The Netherlands Cancer Institute, Amsterdam, The Netherlands, <sup>2</sup>Institute of Pathology, Charité Universitätsmedizin, Berlin, Germany, <sup>3</sup>IRI Life Sciences, Humboldt University of Berlin, Berlin, Germany, <sup>4</sup>Berlin Institute of Health, Berlin, Germany and <sup>5</sup>Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands

\*To whom correspondence should be addressed.

## Abstract

**Motivation:** Signal-transduction networks are often aberrated in cancer cells, and new anti-cancer drugs that specifically target oncogenes involved in signaling show great clinical promise. However, the effectiveness of such targeted treatments is often hampered by innate or acquired resistance due to feedbacks, crosstalks or network adaptations in response to drug treatment. A quantitative understanding of these signaling networks and how they differ between cells with different oncogenic mutations or between sensitive and resistant cells can help in addressing this problem.

**Results:** Here, we present Comparative Network Reconstruction (CNR), a computational method to reconstruct signaling networks based on possibly incomplete perturbation data, and to identify which edges differ quantitatively between two or more signaling networks. Prior knowledge about network topology is not required but can straightforwardly be incorporated. We extensively tested our approach using simulated data and applied it to perturbation data from a BRAF mutant, PTPN11 KO cell line that developed resistance to BRAF inhibition. Comparing the reconstructed networks of sensitive and resistant cells suggests that the resistance mechanism involves re-establishing wild-type MAPK signaling, possibly through an alternative RAF-isoform.

**Availability and implementation:** CNR is available as a python module at <https://github.com/NKI-CCB/cnr>. Additionally, code to reproduce all figures is available at <https://github.com/NKI-CCB/CNR-analyses>.

**Contact:** l.wessels@nki.nl

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Aberrations in cellular signal-transducing networks are one of the hallmarks of cancer (Hanahan and Weinberg, 2011), and many new anti-cancer drugs specifically target genes involved in signaling (Roberts and Der, 2007). Unfortunately, the effectiveness of targeted therapies is often limited by unexpected feedbacks or crosstalks (Klinger *et al.*, 2013; Prahallad *et al.*, 2012; Sun *et al.*, 2014a), and resistance typically emerges due to network adaptations that reactivate inhibited pathways (Ahronian *et al.*, 2015; Sun *et al.*, 2014b). A better and more quantitative understanding of the interactions in such networks, and how these are remodeled under drug exposure, could potentially aid in the design of effective treatments.

Several types of computational approaches have been developed to characterize signaling networks. Ordinary differential equation-based models give detailed, dynamic and quantitative descriptions of a system (Fey *et al.*, 2015). However, such models typically have many parameters that need to be estimated. Fitting these requires vast amounts of measurements and is computationally very expensive, limiting the scope to small systems of which the topology is already well known. Because of their relative simplicity, logic (or Boolean) models are suitable for larger systems (Saez-Rodriguez *et al.*, 2009), and have been used to predict synergistic drug interactions (Flobak *et al.*, 2015) and to identify differences in signaling network topologies between cell lines (Saez-Rodriguez *et al.*, 2011).

However, these models are not quantitative and require discretization of the data.

Modular response analysis (MRA) is a framework that strikes a balance between the level of detail and scope of a model (Bruggeman *et al.*, 2002; Kholodenko *et al.*, 2002). MRA is a mathematical framework that quantifies the interaction strengths between network nodes based on perturbation data. It only considers interactions between ‘modules’, characterized by a single output affecting other modules. In the context of signaling networks, this means that only the active form of the kinase needs to be considered. MRA is considerably simpler than ODE-based models while retaining the relevant information about how network nodes influence each other. In its original formulation, MRA and similar methodologies (Gardner *et al.*, 2003) require perturbations of all nodes, which is always not feasible, for instance, because not all proteins are druggable. This problem can be solved by a maximum-likelihood (Klinger *et al.*, 2013; Stelnic-Klotz *et al.*, 2012) or Bayesian (Halasz *et al.*, 2016; Santra *et al.*, 2013) reformulation of MRA, or similar methods (Korkut *et al.*, 2015), but these methods typically require a defined network topology as input.

While all existing approaches model a specific given signaling network, one is often specifically interested in the quantitative *differences* between signaling networks. For instance, how signaling networks in wild-type cells are transformed by oncogenic mutations or revealing resistance mechanisms that appear under drug exposure by comparing the treatment-naive and resistant networks. Here, we present Comparative Network Reconstruction (CNR), a reformulation of MRA as a Mixed Integer Quadratic Programming (MIQP) problem that allows for efficient network reconstruction and quantification of networks in multiple cell lines simultaneously, allowing the identification of the most relevant differences between them. As input, CNR uses possibly incomplete perturbation data in combination with annotation of the nodes that are directly affected by the perturbations. CNR does not require a predefined network topology, but prior information can be included straightforwardly.

We first extensively test CNR on simulated perturbation data, exploring how noise and incomplete perturbation data affect the quality of network reconstructions. We then demonstrate our approach using perturbation data from a BRAF mutant, PTPN11 KO colorectal cell line that acquired resistance resistance to BRAF inhibition. The resistant cells are less sensitive to BRAF inhibition, more sensitive to growth-factor stimulation, and have a negative feedback from ERK to MEK, suggesting that the resistance mechanism involves re-establishing wild-type MAPK signaling.

## 2 Materials and methods

### 2.1 Formulation of the Comparative Network

#### Reconstruction algorithm

Here, we briefly describe the CNR formalism. For more background information and details, please consult the [Supplementary information](#).

CNR is based on MRA (Kholodenko *et al.*, 2002), a mathematical framework to identify and quantify the interactions in a network based on perturbation experiments. MRA links the measured response  $R_{ij}$  (defined as the log2-fold change in the steady-state activity of node  $x_i$  after perturbation  $j$ ) to the unobserved physical interaction strengths  $r_{kl}$  (defined as the logarithmic partial derivative of  $x_k$  with respect to  $x_l$ ) and direct perturbation effects  $s_{mn}$  (the scaled direct effect of perturbation  $n$  on  $x_m$ ). The matrices of local

and global response coefficients and the direct perturbation effects are related through the equation

$$\mathbf{r} \cdot \mathbf{R} = -\mathbf{s}. \quad (1)$$

CNR aims to find a network model that solves these equations for multiple cell lines simultaneously while minimizing model complexity and the number of differences between the reconstructed networks of the different cell lines.

Formally, given  $N_n$  nodes in the network,  $N_p$  perturbations and  $N_c$  cell lines, and for  $i, j \in \{1, 2, \dots, N_n\}$ ,  $n \in \{1, 2, \dots, N_p\}$  and  $x \in \{1, 2, \dots, N_c\}$ , the MIQP problem reads as follows:

$$\begin{aligned} \text{minimize : } & \sum_{n,i,j,x} \{\epsilon_{in}^x + \eta \cdot I_{ij}^{\text{edge}} + \theta \cdot (I_{ij}^{\text{diff}} + I_{in}^{\text{diff}})\} \\ \text{subject to : } & \sum_{k=1}^{N_n} r_{ik}^x \cdot R_{kn}^x + s_{in}^x = \epsilon_{in}^x \quad \forall i, j, n, x \\ & I_{ij}^{\text{edge}} = 0 \Rightarrow r_{ij}^x = 0 \quad \forall i, j, x \\ & I_{ij}^{\text{diff}} = 0 \Rightarrow r_{ij}^x - r_{ij}^{\text{mean}} = 0 \quad \forall i, j, x \\ & I_{in}^{\text{diff}} = 0 \Rightarrow s_{in}^x - s_{in}^{\text{mean}} = 0 \quad \forall i, n, x \\ & r_{ij}^{\text{mean}} = \sum_x r_{ij}^x / N_c \quad \forall i, j \\ & s_{in}^{\text{mean}} = \sum_x s_{in}^x / N_c \quad \forall i, n \\ & I^{\text{edge}}, I^{\text{dif}}, I^{\text{diff}} \in \{0, 1\} \end{aligned}$$

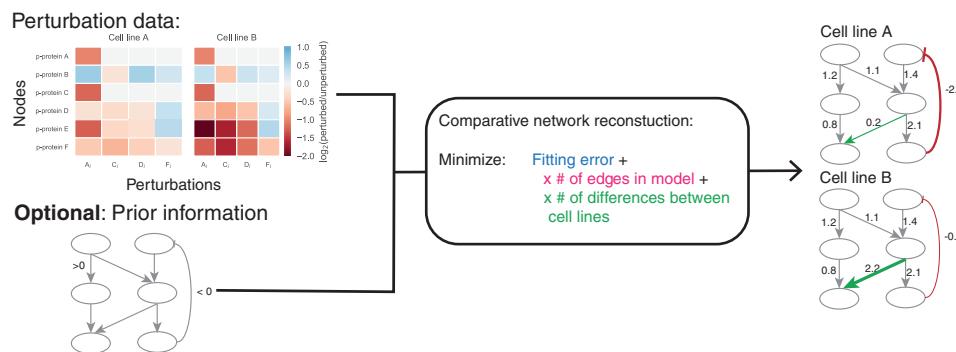
where the  $\epsilon$ s represents the degree to which Equation (1) is satisfied,  $\eta$  and  $\theta$  are hyper parameters to tune the degree of penalization on the number of edges and difference between cell lines, respectively.  $I_{ij}^{\text{edge}}$ ,  $I_{ij}^{\text{diff}}$  and  $I_{in}^{\text{diff}}$  are indicator variables indicating whether an edge is present or not, and whether and edge or perturbation strength is allowed to differ between cell lines, respectively.

Prior information about network topology can be added as constraints of the form  $I_{ij}^{\text{edge}} = 0$  or  $I_{ij}^{\text{edge}} = 1$  for edges that are known to be absent or present, respectively. Similarly, information about the sign of an edge or perturbation can be added as a constraint of the form  $r_{ij}^x \geq 0$  or  $s_{in} \leq 0$ .

The optimization problem is solved using IBM ILOG CPLEX solver (Version 12.7.1), which is freely available for academic use. Importantly, the MIQP solver guarantees optimal solutions (within small numerical tolerances). In some cases, we also want to obtain close-to-optimal solutions. For this, we use the CPLEX solution-pool functionality.

### 2.2 Simulating perturbation data

To generate simulated perturbation data, we used an ODE model of EGFR signaling developed by Orton *et al.* (2009). We calculated the steady state by integrating the model over a suitably long time-interval. All proteins were in their inactive state in the initial conditions. Activating mutations in BRAF, RAS and EGFR were modeled by setting all these proteins to the active state in the initial conditions, and setting the conversion rate to the inactive state to zero. We simulated perturbations as knockdowns of a protein by reducing the total amount of that protein by 50%, and recalculating the steady state. We calculated the global response coefficient for protein  $i$  in response to perturbation  $p$  as  $R_{ip} = \log_2(X_{i,p}/X_{i,0})$ , where  $X_{i,p}$  and  $X_{i,0}$  are the perturbed and unperturbed steady-state concentrations of the active state of protein  $i$ , respectively. We added noise



**Fig. 1.** Schematic illustration of CNR. CNR requires perturbation data (log2-fold changes of node activity compared to a reference state) of one or multiple cell lines as input. Optionally, information on the network topology or edge signs can be included. Based on this data, an MIQP problem is formulated that aims to find a network model that fits the data while penalizing the number of edges in the network and the number of quantitative differences between the cell lines. The output is a network quantification and identification of which edges differ between the cell lines. Edge-weights indicate how strongly a target node responds to a change in the activity in the source node

to the global response coefficients by multiplying them with a random number drawn from a normal distribution with mean ( $\mu$ ) set to 1 and standard deviation ( $\sigma$ ) set to the noise level. To calculate the true local response coefficients for each interacting pair of proteins, we numerically calculated the partial derivatives of the steady-state concentration of the downstream active protein with respect to a perturbation in the concentration of the upstream active protein (by keeping the concentration of all other active proteins fixed). To obtain the true direct perturbation effects of knockdown a protein, we reduced its total amount by 50%, fixed the concentrations of all other proteins, and calculated the new steady-state concentration of that protein in its active state.

### 2.3 Cells and cell culture

All cell lines were derived from the BRAFV600E mutant VACO432 colorectal cancer cell line. Generation of the VACO432 PTPN11 KO clones is described by Prahallad *et al.* (2015). Persister cells were generated by prolonged culturing of VACO432 PTPN11 KO cells in the presence of 2  $\mu$ M Vemurafenib and selecting surviving colonies. Cell lines were cultured in RPMI supplemented with 10% fetal calf serum (FCS) 1% Glutamine and 1% Penicillin/Streptomycin (Gibco). The persister cell line was not maintained.

### 2.4 Perturbation experiments

Prior to the perturbation experiments, cells were synchronized by serum starvation during 24 hours. During this period, the persister cells were kept exposed to Vemurafenib. At  $t = -60$  minutes, cells were treated with an inhibitor (BRAF: Vemurafenib at 2  $\mu$ M, MEK: Selumetinib at 1  $\mu$ M, ERK: SCH772984 at 1  $\mu$ M, AKT: MK2206 at 1  $\mu$ M or PI3Ki: GDC0941 at 1  $\mu$ M; All SelleckChem). At  $t = 0$ , cells were stimulated with growth factor (EGF at 20 ng ml<sup>-1</sup>, HGF at 25 ng ml<sup>-1</sup> or NRG1 at 25 ng ml<sup>-1</sup>; all R&D systems) and at  $t = 30$  minutes were harvested by washing with ice-cold PBS and lysing with Bio-Plex Pro Cell Signaling Reagent Kit (Bio-Rad) lysis buffer. Cell lysate was analyzed with the Bio-Plex Protein Array system (Bio-Rad, Hercules, CA) according to the suppliers protocol as described previously (Klinger *et al.*, 2013). The following phospho-sites were measured: EGFR<sup>Y1068</sup>, MEK<sup>S217,S221</sup>, ERK1/2<sup>T202,Y204/T185,Y187</sup>, p90RSK<sup>S380</sup>, RPS6<sup>S235/S236</sup>, PI3K p85<sup>S15</sup>, AKT<sup>S473</sup>, mTOR<sup>S2448</sup> and GSK3A/B<sup>S21/S9</sup>.

## 3 Results

### 3.1 An efficient method for network reconstruction from perturbation data

We have developed CNR, a computational framework to (i) reconstruct and quantify interactions between signaling proteins based on potentially incomplete perturbation experiments; (ii) doing so in multiple cell lines simultaneously and (iii) with the possibility to use prior data on the network topology and edge signs (Fig. 1). CNR is a reformulation and extension of MRA (Khodenko *et al.*, 2002) as an MIQP problem.

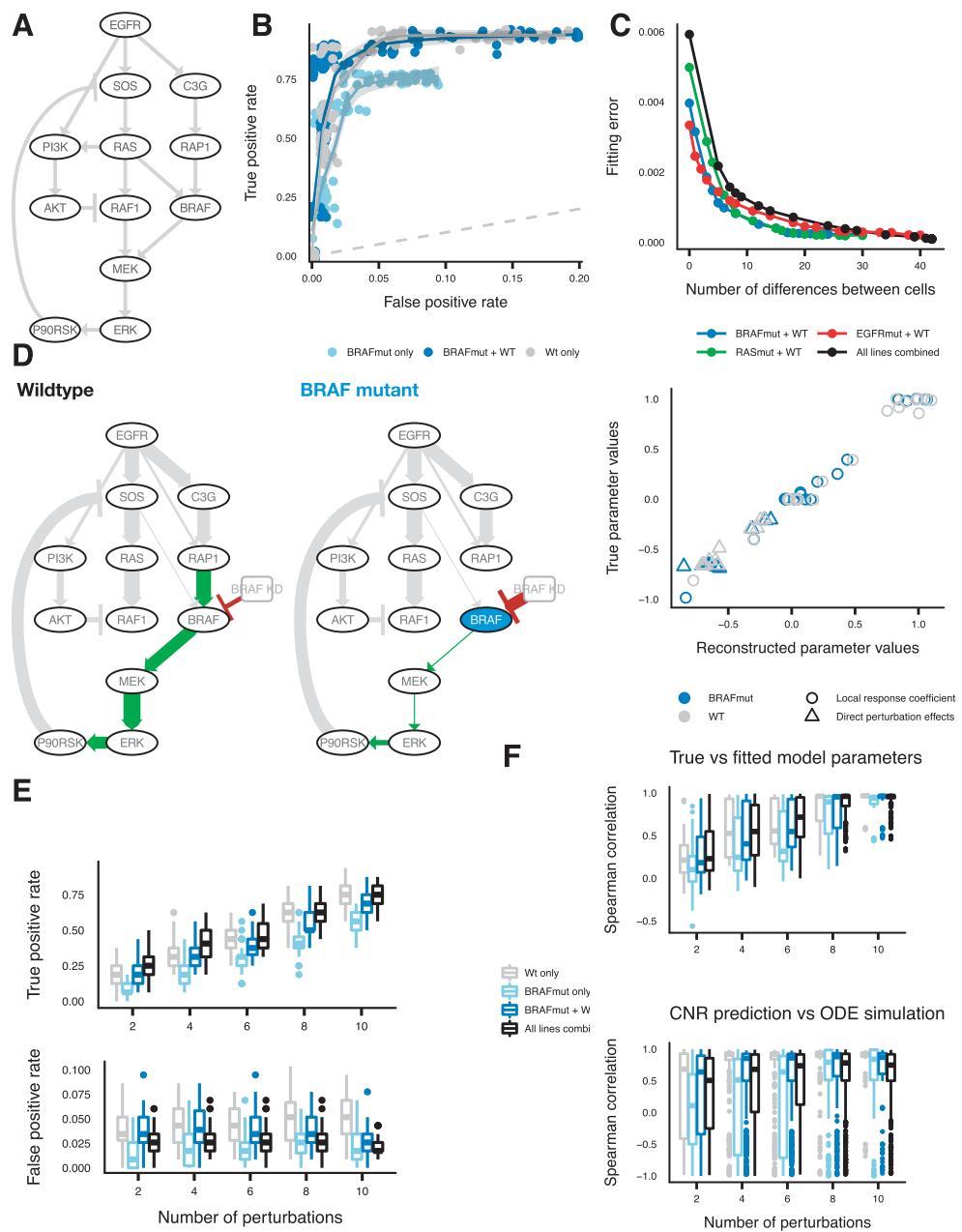
Reconstruction of a network using MRA requires each node of the network to be perturbed, because otherwise the problem is under-determined. This is typically not feasible, for instance due to cost constraints or simply because not all nodes can be perturbed by e.g. applying a drug. Furthermore, in the classical formulation, it is not possible to simultaneously reconstruct the networks of several cell lines, which complicates their comparison. CNR solves these problems.

CNR takes the log2-fold changes of signaling proteins in response to perturbations as input, together with annotations of which nodes are perturbed (Fig. 1). Optionally, information about the presence, absence or signs of interactions can also be supplied (Fig. 1). CNR then tries to find a network model that (i) fits the perturbations data; (ii) maintains the same topology for all cell lines; (iii) penalizes model complexity (i.e. the number of edges in the network) to prevent over-fitting and (iv) penalizes the number of edges that differ between cell lines to facilitate identification of the most relevant differences. The latter has the additional advantage of reducing the number of parameters, since most edges are described by the same parameter in all cell lines.

The output of CNR is a network quantification for each cell line, in combination with a quantification of the strengths of the effects of perturbations on their direct targets (Fig. 1). The interaction strength between a pair of nodes consisting of a source and target node is referred to as the local response coefficient. This coefficient is interpreted as the percentage change in the target node activity in response to a 1% change in the source node activity.

### 3.2 *In-silico* evaluation

To test our method, we simulated perturbation experiments using an ODE-based dynamic model of the EGFR signaling developed by Orton *et al.* (2009) (Supplementary Fig. S1). Following Orton *et al.*,



**Fig. 2.** Orton model simulation and CNR results. For all results presented here, 10% noise was added to the simulated data. (A) Network topology of the modules in the Orton model. (B) Receiver-operator-curves of network reconstructions using noisy data of the wild-type model, the BRAF model and both combined, with  $\theta = 0$ . (C) Fitting error versus number of differing edges between cell lines indicate that most edges can be set to the same value without affecting model fit. (D) Network reconstructions of the wild-type and BRAF mutant model. Left: Network topology. Edges that differ between the model-reconstructions are highlighted. Green edges are positive, red ones negative. Right: Correlation between reconstructed and actual model parameters. (E) True (top) and False (bottom) positive rates of network reconstruction as function of the number of perturbations in the input data. (F) Spearman correlation between predicted and simulated model parameters (top) and response for perturbations that were not used in reconstructing the network (bottom)

we also generated BRAF, KRAS or EGFR mutant versions of the model by forcing these proteins to all be in their active state. We simulated network perturbations by ‘knocking down’ all the relevant nodes in the network, one by one, and using the resulting measurements as input to CNR.

We tested the performance of CNR by comparing network reconstructions to the known model from which the data were simulated. Note that since CNR aims to quantify the effect of a perturbation on the steady-state concentration of the active form

of a protein only, the CNR network reconstruction will be much simpler than the original model. Nevertheless, the true interactions between active proteins are unambiguously defined (Fig. 2A). Specifically, we wanted to test how noise in the data, the penalty on differences between cell line, or incomplete data affect the ability to reconstruct or quantify the network interactions. In the following sections, unless stated otherwise, we did not employ any information about the model topology for the reconstructions, and we added 10% noise to the input data.

**3.2.1 The effect of measurement noise on topology reconstructions**  
 We first tested the ability of our method to reconstruct the network topology when the data is noisy. To this end, we performed reconstructions with different levels of noise added to the input data, for different values of the  $\eta$ -hyperparameter (using  $\theta=0$ ). By starting with a large  $\eta$  and gradually lowering it, we get solutions with an increasing number of edges. From these solutions we calculated the true-positive rate (number of true-positive edges/total number of real edges) and the false-positive rate (number of false-positive edges/total number of real-negative edges). Figure 2B shows the receiver-operating characteristic (ROC) curve for reconstructions of the BRAF mutant (light blue) and the wild-type (gray) cell lines, and simultaneous reconstruction of both together (dark blue), when 10% noise is added to the data. A true-positive rate of  $>0.8$  is attained while keeping the false-positive rate well below 0.05, so despite noisy data, a large majority of the edges can be correctly identified. (This is attained when using  $\eta=0.005$ .) The performance of the wild-type + BRAF-mutant and wild-type alone reconstructions are comparable, and better than that of the BRAF-mutant alone. This is presumably because due to constitutive BRAF-activation nodes downstream of BRAF are nearly unresponsive to perturbations upstream of BRAF, making these less informative for identifying downstream edges. Similar results were obtained for the other mutants (Supplementary Fig. S2). As expected, increasing the noise levels reduces the performance, but even at 100% noise, the reconstruction performs much better than randomly reconstructing the network, represented by the dotted gray line where the true-positive rate equals the false-positive rate (Supplementary Fig. S2). Interestingly, the performance of the BRAF and the RAS mutant cell line reconstructions suffer less from increasing noise than that of the wild-type.

**3.2.2 The effect of the penalty on cell line differences on model fit.**  
 Next, we first tested how the penalty on the differences between cell lines affects the model fit to the data by performing reconstructions using different values for the  $\theta$ -hyperparameter (using a fixed  $\eta=0.005$ ). We performed this analysis by pairing the wild-type with the BRAF, RAS and EGFR mutants, respectively. In addition, we performed the analysis for all cell lines combined (Fig. 2C). For each of these analyses, a model with only a small number of differences ( $\approx 10$ , attained when  $\theta=0.01$ ) between the cell lines fits the data nearly as well as a model where all edges differ between the cell lines. The number of edges that differ between cell lines can thus be strongly decreased without affecting the fitting error, which is what one would expect when comparing isogenic cell lines that only differ by a single mutation. Adding a non-zero  $\theta$ -value does not influence the ROC-curves (Supplementary Fig. S2).

### 3.2.3 Example reconstructions

Having established that CNR is capable of accurately reconstructing networks while strongly reducing the differences between cell lines, we next wanted to assess how interpretable the obtained differences between networks are. To this end, we reconstructed the wild-type and BRAF mutant models together, using  $\eta=0.005$  and  $\theta=0.01$ , and visualized the resulting networks (Fig. 2D, left). The thickness of the arrows represents the magnitude of the reconstructed local response coefficients. The edges that differ between the cell lines are highlighted in color. As expected, in the BRAF-mutant, BRAF activation is unresponsive to RAP1 and more responsive to knockdown of BRAF itself. Furthermore, all other differences between the cell lines are downstream of BRAF. (In the mutant, the nodes downstream of BRAF are less responsive to changes in their direct

upstream kinases due to saturation effects.) Similar results are obtained for the other mutants (Supplementary Fig. S3). This illustrates that a comparative reconstruction can identify where in the network two cell lines differ. Importantly, the reconstructed local response coefficients and direct perturbation effects correlate very well with their true value directly obtained from the Orton model (Fig. 2D, right). This indicates that CNR can be used to obtain quantitatively accurate model reconstructions from noisy data.

### 3.2.4 The effect of incomplete perturbation data

It is always not feasible to perturb each node in the network. To test how incomplete perturbation data affects the accuracy of network reconstruction, we performed reconstructions based on a reduced number of input perturbations. To this end, we randomly selected between 2 and 10 nodes to perturb, added 10% noise to the data, and performed the network reconstruction. (Because we expect the total fitting error to scale with the number of perturbations  $N_p$ , we scaled  $\eta$  and  $\theta$  using  $\eta=0.01 \cdot N_p/N_n$  and  $\theta=0.01 \cdot N_p/N_n$ , where  $N_n=12$ .) We repeated this 50 times for each number of perturbations. Figure 2E and Supplementary Figure S4 show the true- and false-positive rates of these reconstructions. As expected, the performance decreases with decreasing numbers of perturbations. However, with as few as two or four perturbations, we still attain performance levels well above random guessing. With low numbers of perturbations used as input, there is a clear benefit of combining multiple cell lines, as the panel combining all cell lines consistently has among the highest true-positive and lowest false-positive rates.

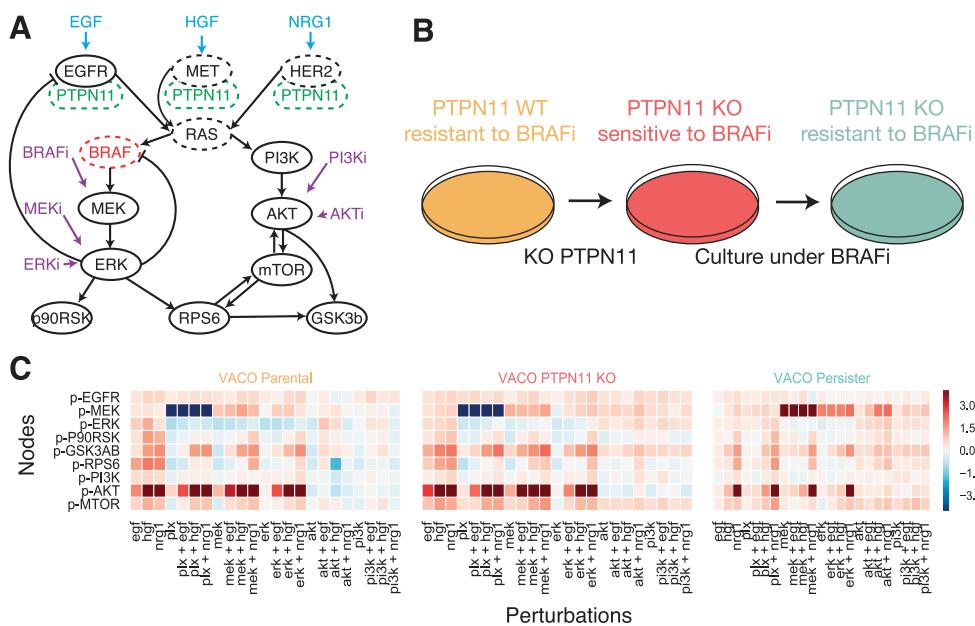
### 3.2.5 Predicting the effect of new perturbation

Sometimes one is more interested in network quantification and prediction of the effect of new perturbations from a systems were the topology is known, rather than identifying which edges are present. We therefore next tested how well CNR is capable of quantifying the model parameters—local response coefficients and direct perturbation effects—based on incomplete perturbation data. We again randomly selected between 2 and 10 perturbations, added 10% noise to the data, performed reconstructions this time with the correct network topology supplied, and compared the fitted parameters with the true parameters directly obtained from the Orton model (Fig. 2F and Supplementary Fig. S5, top). With as few as 6 perturbations, the median Spearman correlation is roughly 0.5, and for 8 perturbations it approaches 1.

We then used the obtained network-quantifications to predict the response to perturbations that were not used in quantifying the network. To quantify the predictive accuracy of these networks, we calculated the Spearman correlation between the predicted and true response for each perturbation. In calculating the correlation, we excluded the perturbed node itself since the direct-perturbation effect is not part of the prediction. Perturbations where none of the nodes respond are also excluded. Figure 2F and Supplementary Figure S5 (bottom) show that for as few as four or six perturbations, the median Spearman correlation between the predictions and the simulated responses is very close to 1. Interestingly, while the performance in edge reconstruction of the BRAF and RAS mutants were slightly better than those of the wild-type and EGFR mutants, their quantitative predictions are slightly worse (Supplementary Fig. S5).

## 3.3 Network quantification of PTPN11 KO cells with and without acquired resistance to BRAF inhibition

To test the ability of CNR in a real-world setting, we set out to identify the cause of resistance to targeted anti-cancer drugs by elucidating how the signaling network changed in resistant cells.



**Fig. 3.** Perturbation experiments in PTPN11 WT, KO and BRAF inhibitor resistant cells. **(A)** Schematic of role of PTPN11 in signal transduction. **(B)** Overview of the cell-lines used. VACO432 cells are insensitive to BRAF inhibition, but can be sensitized by knocking out PTPN11. Prolonged culturing VACO432 PTPN11 KO cells in the presence of BRAF inhibitor gives rise to resistant VACO432 PTPN11 KO cells. **(C)** Results of the perturbation experiments. The color scale indicates log2-fold change relative to unstimulated, uninhibited controls

Specifically, we wanted to understand how a PTPN11 KO colorectal cancer (CRC) cell line harboring an activating BRAF<sup>V600E</sup> mutation become resistant to BRAF inhibition with Vemurafenib.

### 3.3.1 PTPN11 KO and sensitivity to BRAF inhibition

Normally, BRAF<sup>V600E</sup> mutated CRC cells are unresponsive to BRAF inhibition due to a feedback loop that activates the receptor tyrosine kinase (RTK), EGFR, upon BRAF inhibition (Klinger *et al.*, 2013; Prahallas *et al.*, 2012). Recently, Prahallas *et al.* showed that knockout of PTPN11—which is required to transduce signals from RTKs to the MAPK pathway—prevents this feedback loop from reactivating the MAPK signaling pathway (Prahallas *et al.*, 2015) (Fig. 3A). However, when PTPN11 KO cells are cultured in the presence of a BRAF inhibitor, they eventually become resistant to this inhibitor.

### 3.3.2 Generation of persisters

To test how resistance to the BRAF inhibition in PTPN11 KO CRC cells emerges, we generated ‘persister’ cells by culturing a PTPN11 KO clone of the BRAF<sup>V600E</sup> mutated CRC cell line VACO432 in the presence of the BRAF inhibitor Vemurafenib. After roughly 2 weeks, persister cells emerged that were able to grow in the presence of Vemurafenib at a growth rate comparable to that of the parental line in the absence of drug (Fig. 3B).

### 3.3.3 Perturbation experiments

We systematically perturbed VACO432 parental, PTPN11 KO and the derived persister cells with combinations of growth factors and inhibitors targeting proteins in the PI3K-AKT and MAPK pathway, and measured the steady-state responses of the phosphorylation status of main proteins in the pathways (Fig. 3A and C). The log2-fold changes, relative to their unperturbed (i.e. uninhibited and unstimulated) state, were used as input to the CNR optimization problem. As the MAPK and AKT pathways are relatively well characterized, we did not focus on the reconstruction of the pathway topology.

Instead, we restricted the model to consist of known interactions and feedbacks obtained from the literature (Fig. 3A).

### 3.3.4 Modeling of perturbations

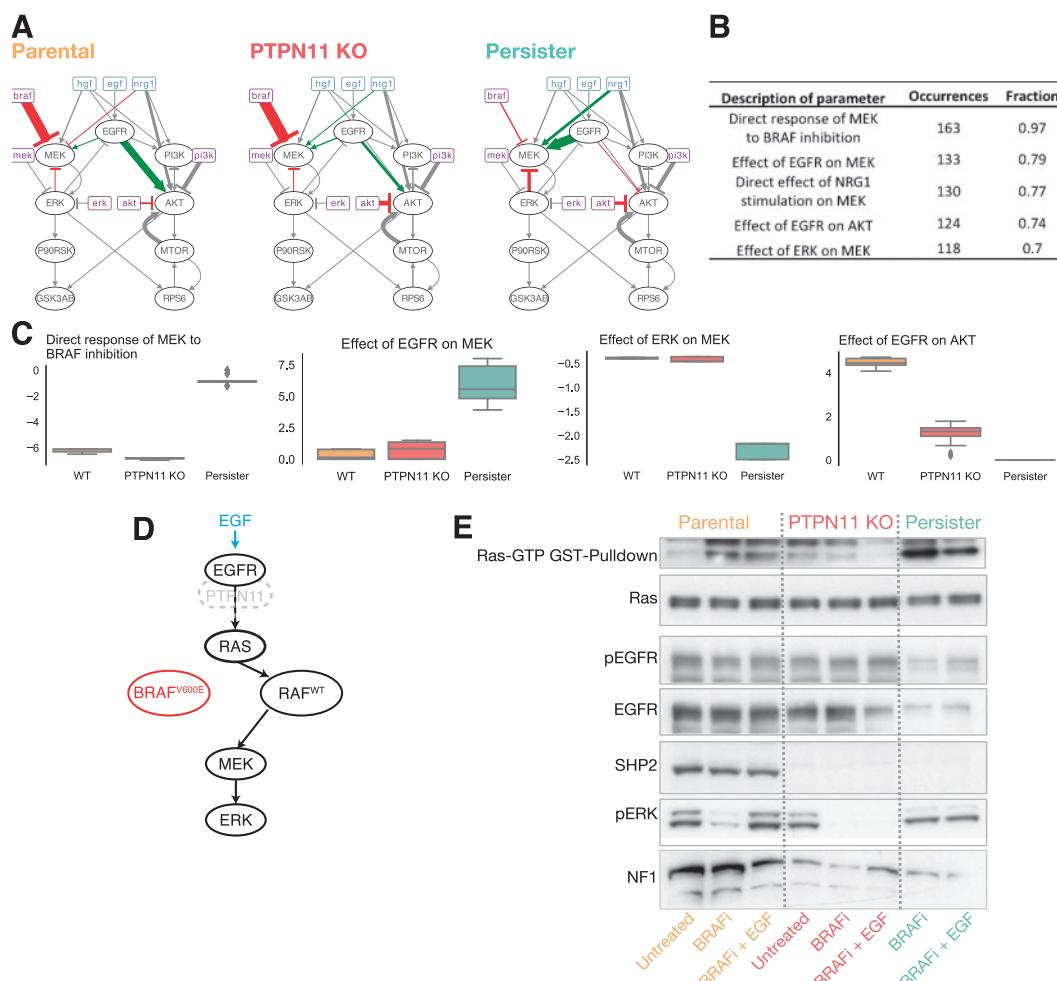
Some kinase inhibitors prevent activation of their target by blocking the phosphorylation of the kinase itself, whereas other work by preventing the kinase activity of the phosphorylated protein. We model the latter type of inhibitor as perturbing the substrates of its direct target. Specifically, the MEK and PI3K inhibitors are preventing the kinase activity of their targets and are thus modeled to perturb ERK and AKT, respectively. Since we did not measure BRAF, we cannot model the direct effect of BRAF inhibition on BRAF, but instead model it as a perturbation of MEK. Similarly, we did not measure activity of HER2 and MET, so we model HGF and NRG1 stimulation as affecting MEK, PI3K and AKT directly.

### 3.3.5 Network reconstruction

Figure 4A shows the network reconstruction of the three cell lines, and how they differ. We allowed for six differences between networks as it gives a good balance between model-fit and model-complexity (Supplementary Fig. S6). As expected, the response of MEK activation to BRAF inhibition is much smaller in the persisters compared to the parental and the PTPN11 KO cells. Furthermore, in the persisters the feedback from ERK to MEK is stronger and MEK is more responsive to receptor stimulation. Finally, the effect of EGFR on AKT is weaker in both the PTPN11 KO and the persister cells, indicating that PTPN11 might be involved in signaling to the AKT pathway as well as the MAPK pathway.

### 3.3.6 Robustness of network reconstruction

To test how robust these results are, we generated a large pool of solutions by varying the allowed number of differences between 0 and 16, and also considering close-to-optimal solution (with an objective value at most 1.5 times the best possible given the number of differences between cell lines). We counted how often each edge



**Fig. 4.** Model reconstruction of sensitive and persister VACO432 cells. **(A)** Network reconstruction from perturbation data. **(B)** Overview of parameters that differ in many of the solutions. **(C)** Boxplots of selected parameters indicating that the differences are quantitatively similar in different solutions. **(D)** Hypothesis: Resistance is due to re-establishment of normal MAPK signaling. **(E)** Consistently, pERK is activated and GTP constitutively loaded with GTP

differed between cell lines in the 168 solutions in the solution-pool. There are a limited number of edges that occur in the majority of these solutions (Fig. 4B). These are (i) the response of MEK to BRAF inhibition; (ii) the effects of EGFR on MEK and AKT; (iii) the feedback from ERK to MEK and (iv) the responsiveness of MEK to growth-factor stimulation. Furthermore, the often-occurring differences were also quantitatively consistent in most solutions (Fig. 4C), and in agreement with the observations we made earlier: (i) MEK activity in persisters is less responsive to BRAF inhibition; (ii) more responsiveness to growth factor stimulations; (iii) persisters have a negative feedback from ERK to MEK and (iv) the connection between EGFR and AKT is stronger in the (PTPN11 WT) parental cells. This indicates that these results are robust and do not depend on the detailed settings in the network reconstruction.

### 3.3.7 Interpretation

The CNR results suggest a model in which the persister cells became resistant by re-establishing normal (RAF-wild-type) MAPK signaling, by, for instance, using an alternative RAF-isoform (Fig. 4D). Such a model explains the reduced response of MEK to Vemurafenib, which specifically targets BRAF<sup>V600E</sup> and not wild-

type RAF. It explains the stronger negative feedback from ERK to MEK, which is mediated through RAF and thus weak if BRAF is constitutively active. It also explains the increased responsiveness of MEK to EGFR and other growth factors, which are also mediated through RAF and thus weak if BRAF is constitutively active. Consistent with this hypothesis, pERK is activated in the persisters and RAS is constitutively loaded with GTP (Fig. 4E). This hypothesis suggests that the persister cells might be sensitive to pan-RAF inhibition.

## 4 Discussion

Understanding which targeted therapies are effective in what context, and how cells adapt to become resistant, requires a quantitative understanding of signaling networks and how they differ between cells. In this study we developed CNR, a computational method to reconstruct signaling networks based on perturbation data. CNR focuses on identifying quantitative differences between a set of cell lines (or other experimental models), as this is often the relevant biological question. Using simulated data we showed that CNR works well with noisy and incomplete data.

We applied CNR to perturbation data we generated from BRAF<sup>V600E</sup> mutated, PTPN11 KO colorectal cancer cells that were

made resistant to the BRAF inhibitor Vemurafenib through continuous exposure to sub-lethal doses of the drug. The network reconstructions suggested a reversion to RAF wild-type-like MAPK signaling in the resistant cells that would likely make them sensitive to pan-RAF inhibition. This hypothesis is consistent with the observation that both *in vitro* and in patients resistance to combined BRAF/EGFR inhibition is mediated by alterations that reactivate MAPK pathway activity (Ahronian *et al.*, 2015; Hazar-Rethinam *et al.*, 2018). It also raises a number of other interesting questions, such as how RAF can be activated and respond to growth factor stimulation in the absence of PTPN11.

CNR is formulated as an MIQP problem. Theoretically, such problems are NP-hard, and the search space increases exponentially with the number of nodes in the network. In practice the MIQP problem associated with CNR can typically be efficiently solved. For example, a typical reconstruction for the 12-node Orton model is optimized in under a second on a standard laptop computer. Parameter-setting that lead to highly connected networks are somewhat slower (Supplementary Fig. S7), but since most proteins interact with a limited set of other proteins, biological networks are sparse and the highly connected solutions are not relevant. In addition, the run time can be strongly reduced by incorporating prior information, for instance by providing a starting network with well-known interactions and allowing for a limited number of additional edges.

Some other quantitative network reconstruction approaches, notably those based on maximum-likely optimization (Dorel *et al.*, 2018; Klinger *et al.*, 2013) and Bayesian statistics (Halasz *et al.*, 2016; Jastrzebski *et al.*, 2018; Thijssen *et al.*, 2018), are statistically more rigorously defined than CNR. However, the main advantages of CNR compared to these approaches is its flexibility regarding prior knowledge of the network topology or edge signs. CNR does not require such knowledge, but it can straightforwardly be incorporated when it is available. An additional benefit of CNR is the efficiency, making it scalable to much larger networks than other quantitative approaches.

In this work we focused on signaling networks, but CNR can be applied to any perturbation dataset were the targets of the perturbations are known. We used a L0-penalty for the differences between networks (an edge is either different or not) because such results are convenient to interpret as only a subset of edges need to be inspected. However, the penalty can be replaced with an L1 or L2-penalty (The absolute or square of the difference between two edges, respectively), which in some cases better resembles the biological problem, and is expected be faster to solve.

To conclude, CNR provides a flexible and efficient way to reconstruct networks from perturbation data. Using simulations, we demonstrated its ability to accurately reconstruct networks when data is noisy and incomplete. Application to a real-world problem showed its robustness and potential in formulating hypotheses on biological problem.

## Funding

This project was supported by Nederlands Organisatie voor Wetenschappelijk Onderzoek (NWO), under the frame of ERACoSysMed-1, the ERA-Net for Systems Medicine in clinical research and medical practice.

*Conflict of Interest:* none declared.

## References

- Ahronian,L.G. *et al.* (2015) Clinical acquired resistance to raf inhibitor combinations in braf-mutant colorectal cancer through mapk pathway alterations. *Cancer Discov.*, 5, 358–367.
- Bruggeman,F.J. *et al.* (2002) Modular response analysis of cellular regulatory networks. *J. Theor. Biol.*, 218, 507–520.
- Dorel,M. *et al.* (2018) Modelling signalling networks from perturbation data. *Bioinformatics*, doi: 10.1093/bioinformatics/bty473.
- Fey,D. *et al.* (2015) Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.*, 8, ra130–ra130.
- Flobak,Ö. *et al.* (2015) Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS Comput. Biol.*, 11, e1004426.
- Gardner,T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301, 102–105.
- Halasz,M. *et al.* (2016) Integrating network reconstruction with mechanistic modelling to predict cancer therapy. *Sci. Signal.*, 9, ra114.
- Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, 144, 646–674.
- Hazar-Rethinam,M. *et al.* (2018) Convergent therapeutic strategies to overcome the heterogeneity of acquired resistance in BRAFV600E colorectal cancer. *Cancer Discov.*, 8, 417–427.
- Jastrzebski,K. *et al.* (2018) Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines. *Cancer Res.*, 78, 4396–4410.
- Kholodenko,B.N. *et al.* (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. USA*, 99, 12841–12846.
- Klinger,B. *et al.* (2013) Network quantification of EGFR signaling unveils potential for targeted combination therapy. *Mol. Syst. Biol.*, 9, 673.
- Korkut,A. *et al.* (2015) Perturbation biology nominates upstream–downstream drug combinations in RAF inhibitor resistant melanoma cells. *Elife*, 4, 1–31.
- Orton,R.J. *et al.* (2009) Computational modelling of cancerous mutations in the EGFR/ERK signalling pathway. *BMC Syst. Biol.*, 3, 100.
- Prahallad,A. *et al.* (2012) Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483, 100–103.
- Prahallad,A. *et al.* (2015) PTPN11 is a central node in intrinsic and acquired resistance to targeted cancer drugs. *Cell Rep.*, 12, 1978–1985.
- Roberts,P.J. and Der,C.J. (2007) Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, 26, 3291–3310.
- Saez-Rodriguez,J. *et al.* (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.*, 5, 331.
- Saez-Rodriguez,J. *et al.* (2011) Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res.*, 71, 5400–5411.
- Santra,T. *et al.* (2013) Integrating Bayesian variable selection with Modular Response Analysis to infer biochemical network topology. *BMC Syst. Biol.*, 7, 57.
- Stelnicek-Klotz,I. *et al.* (2012) Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS. *Mol. Syst. Biol.*, 8, 601.
- Sun,C. *et al.* (2014a) Intrinsic resistance to mek inhibition in kras mutant lung and colon cancer through transcriptional induction of erbB3. *Cell Rep.*, 7, 86–93.
- Sun,C. *et al.* (2014b) Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. *Nature*, 508, 118–122.
- Thijssen,B. *et al.* (2018) Delineating feedback activity in the mapk and akt pathways using feedback-enabled inference of signaling activity. *bioRxiv*.

# Supplementary Information to “Comparative network reconstruction identifies resistance mechanisms to targeted cancer treatment”

Evert Bosdriesz<sup>1</sup>, Anirudh Prahallad<sup>1</sup>, Bertram Klinger<sup>2,3</sup>, Anja Sieber<sup>2,3</sup>, Nils Blüthgen<sup>2,3,4</sup>, and Lodewyk FA Wessels<sup>1,5,\*</sup>

<sup>1</sup>*Division of Molecular Carcinogenesis, The OncoCode Institute, The Netherlands Cancer Institute, Amsterdam, the Netherlands*

<sup>2</sup>*Institute of Pathology, Charité Universitätsmedizin, Berlin, Germany*

<sup>3</sup>*IRI Life Sciences, Humboldt University of Berlin, Berlin, Germany*

<sup>4</sup>*Berlin Institute of Health, Berlin, Germany*

<sup>5</sup>*Faculty of EEMCS, Delft University of Technology, Delft, the Netherlands*

\*Corresponding author: l.wessels@nki.nl

## Supplementary Information

### Basics of Modular Response Analysis

Consider a network consisting of  $N$  nodes, the state of which is specified by  $\mathbf{x} = \{x_1, \dots, x_N\}$ . These nodes can be the proteins involved in a signaling network, in which case their state is the concentration of phosphorylated protein. More generally, these nodes are modules of arbitrary complexity, which may involve many cellular components connected by intra-modular reaction. However, by assumption only a single element of the module serves as output. The steady state of the system is defined by:

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{p}) = \mathbf{0}, \quad (\text{S1})$$

where  $\mathbf{p}$  is vector of parameters. Modular Response Analysis (MRA) [1, 2] aims to reconstruct an interaction network based on measuring the steady state responses to (many) perturbations.

A perturbation in  $p_j$  can affect  $x_i$  directly, but also indirectly through network effects. Upon a infinitesimal perturbation in  $p_j$ , the following relation holds for each node:

$$\frac{p_j dx_i}{x_i dp_j} = \frac{p_j \partial x_i}{x_i \partial p_j} + \sum_{k \neq i} \frac{x_k}{x_i} \frac{\partial x_i}{\partial x_k} \cdot \frac{p_j dx_k}{x_k dp_j} \quad (\text{S2})$$

Hence, an experiment of  $M$  perturbations yield  $M \times N$  relations.

In MRA, all terms in equation (S2) have a name and an interpretation. The global response coefficient

$$R_{in} \equiv \frac{p_n}{x_i} \frac{dx_i}{dp_n} \quad (\text{S3})$$

is the response of the node  $i$  to a perturbation in a parameter  $n$ . This is typically the quantity that can be measured. The local response coefficient

$$r_{ij} \equiv \frac{x_j}{x_i} \frac{\partial x_i}{\partial p_n} \quad (\text{S4})$$

represents the direct effect of a change in  $x_j$  on  $x_i$ . By definition,  $r_{ii} \equiv -1$ . The elements of the perturbation matrix

$$s_{in} \equiv \frac{p_n}{x_i} \frac{\partial x_i}{\partial p_n} \quad (\text{S5})$$

quantify the direct response of node  $i$  to a perturbation in  $p_j$ .

Equation (S2) can be reformulated in matrix notation to yield:

$$\mathbf{r} \cdot \mathbf{R} = -\mathbf{s} \quad (\text{S6})$$

where  $\mathbf{r}$  is an  $N \times N$  matrix (with the local response coefficients as its entries) and  $\mathbf{R}$  and  $\mathbf{s}$  are a  $N \times M$  matrices.  $\mathbf{r}$  (and to a lesser extend  $\mathbf{s}$ ) are the quantities of interest because they represent the direct interaction. However, these but can not be measured directly because a node cannot be manipulated in isolation. The goal of MRA then, is to find these based on the measured  $\mathbf{R}$ .

The idea behind classical MRA is to perturb each node independently, meaning all nodes are perturbed and that each perturbation affects only one node. In this case, it is possible to select a parameterization such that  $s_{in} = \delta_{in}$  (where  $\delta$  is the Kronecker delta). The solution to the problem than simply is:

$$\mathbf{r} = -\mathbf{R}^{-1} \quad (\text{S7})$$

## Formulation of Comparative Network Reconstruction

For the purpose of comparing cell lines (or other experimental models), there are two main problems with MRA. The first is that it is typically not feasible to perturb all nodes in a network, which makes the problem under-determined. The second is that it requires fitting a new model for each cell line, which prevents ‘sharing’ data between cell lines and thus might incorrectly identify differences that are the result of e.g. measurement errors.

Our strategy to solve this is to reformulate MRA as a Mixed Integer Quadratic Programming (MIQP) problem. MIQP is a form of constrained, non-linear optimization where the objective function is a convex, quadratic function while the constraints are linear, and some variables are integers.

The idea is to look for set of models — one for each cell line — that:

- fits the perturbation data while
- adding a sparsity constraint by penalizing the number of edges in the network, and
- penalizing edges that differ between cell lines.

The fit is found by minimizing  $\mathbf{r} \cdot \mathbf{R} + \mathbf{s}$ . The penalties on the number of, and differences between, edges are implemented using so-called indicator constraints. These are binary variables that can code logical relationship. (Because these indicator constraints are integers, this is a MIQP problem instead of simply a QP one.) For each potential edge  $r_{ij}$  in the network we add to two indicators,  $I_{ij}^{\text{edge}}$  coding for the presence/absence of the edge, and  $I_{ij}^{\text{diff}}$  coding for whether it is allowed to differ between cell lines. If  $I_{ij}^{\text{diff}} = 0$ , the value of corresponding local response coefficient  $r_{ij}$  is constrained to 0 in all cell lines. Similarly, if  $I_{ij}^{\text{diff}} = 0$ ,  $r_{ij}$  is constrained to have the same value in all cell lines. The penalties are implemented by adding (weighted) sums over the indicator constraints to the objective functions.

Formally, the MIQP problem reads as follows:

$$\begin{aligned}
\text{minimize : } & \sum_{n \in \text{perturbations}} \sum_{i,j \in \text{nodes}} \sum_{x \in \text{cell lines}} \epsilon_{in}^x {}^2 + \eta \cdot I_{ij}^{\text{edge}} + \theta \cdot (I_{ij}^{\text{diff}} + I_{in}^{\text{sdiff}}) \\
\text{subject to : } & \sum_{k \in \text{nodes}} r_{ik}^x \cdot R_{kn}^x + s_{in}^x = \epsilon_{in}^x \quad \forall i, j, n, x \\
& I_{ij}^{\text{edge}} = 0 \Rightarrow r_{ij}^x = 0 \quad \forall i, j, x \\
& I_{ij}^{\text{diff}} = 0 \Rightarrow r_{ij}^x - r_{ij}^{\text{mean}} = 0 \quad \forall i, j, x \\
& I_{in}^{\text{sdiff}} = 0 \Rightarrow s_{in}^x - s_{in}^{\text{mean}} = 0 \quad \forall i, n, x \\
& r_{ij}^{\text{mean}} = 1 / N_{\text{cell lines}} \sum_x r_{ij}^x \quad \forall i, j \\
& s_{in}^{\text{mean}} = 1 / N_{\text{cell lines}} \sum_x s_{in}^x \quad \forall i, n \\
& I^{\text{edge}}, I^{\text{dif}}, I^{\text{sdiff}} \in \{0, 1\} \\
& i, j \in \{\text{nodes}\}, \quad n \in \{\text{perturbations}\}, \quad x \in \{\text{cell lines}\}
\end{aligned} \tag{S8}$$

where  $\eta$  and  $\theta$  are hyper-parameters to tune the degree of penalization of on the number of edges and difference between edges, respectively.

An important benefit of formulating MRA as an MIQP problem is that prior knowledge can be incorporated straightforwardly by additional constraints.

- Prior information about the network topology can be incorporated by adding constraints on the  $I^{\text{edge}}$ . For instance, addition of  $I_{ij}^{\text{edge}} = 0$  to equation (S8) indicates that there is no interaction between node  $i$  and  $j$ .
- Prior information on the sign of an interaction can be incorporated by adding constraints on local response coefficients. For instance, if a feedback from node  $j$  to node  $i$  is known to be negative, this is encoded by adding  $r_{ij}^x < 0$  to equation (S8). Similarly, a perturbation that is known to be inhibitory can be encoded as  $s_{in} < 0$ .
- Prior information about edges that differ between cell lines can be incorporated by adding constraints of the form  $I_{ij}^{\text{diff}} = 0$  to equation (S8).
- Finally, a maximum on total number of edges or differences between edges can be added by simply adding a constraint of the form  $\sum I^{\text{edge}} \leq N^{\text{edge}}$  and  $\sum I^{\text{diff}} \leq N^{\text{dif}}$ .

## Supplementary figures

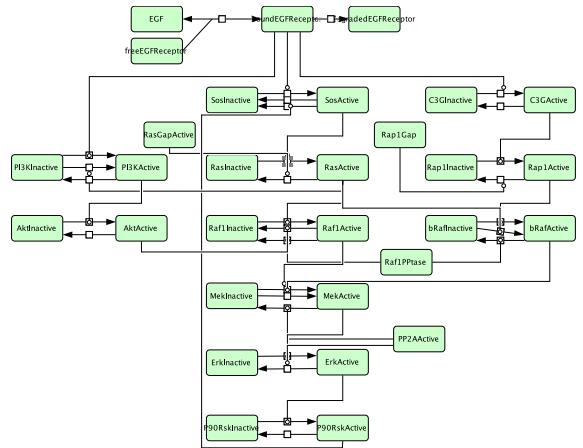
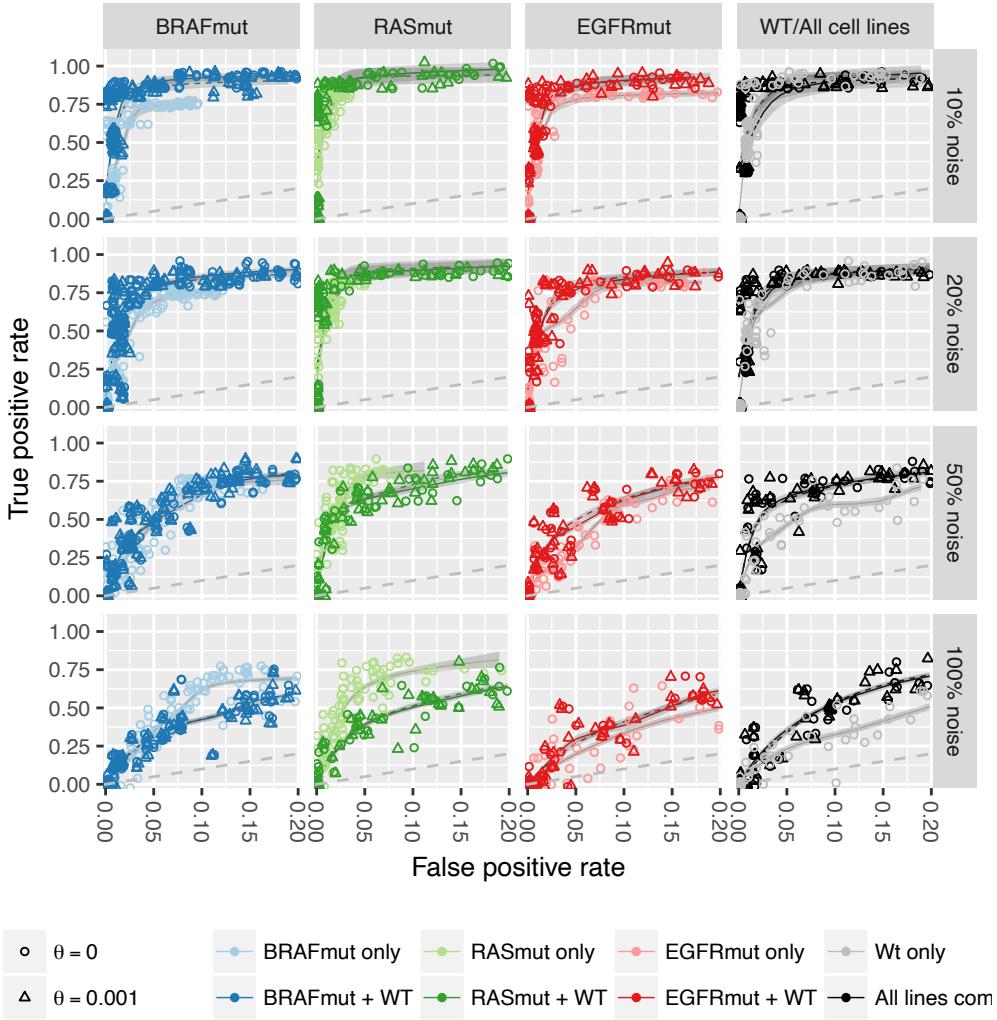


Figure S1: Schematic of the ODE model used to simulate the perturbation.



**Figure S2: Receiver-operator characteristic curves of different cell-line panels with different levels of noise added to the data.** By scanning over the  $\eta$ -hyperparameter, reconstructions with different numbers of edges were obtained, leading to different true and false positive rates. Reconstructions were performed on either a mutant cell line alone, or together with the wild-type cell line. Additionally, reconstructions on all cell lines simultaneously was performed, and compared to reconstruction on the wild-type cell-line alone. For each noise-level, the scan over the  $\eta$ -hyperparameter was repeated 10 times. For each reconstruction, the  $\theta$ -hyperparameter, penalizing the differences between cell lines, was set to either 0 or 0.001. Each point represents one network reconstruction. Lines are estimated by fitting a smooth, monotonically increasing curve through the points using the R scam package [3]. Solid curves represent  $\theta = 0$ , dashed curves  $\theta = 0.001$ , but because these nearly overlap, they are hard to distinguish. The shaded regions indicate standard error of the means. Note that the false positive rate is only shown until 0.2. For visualization purposes, the points were slightly jittered.

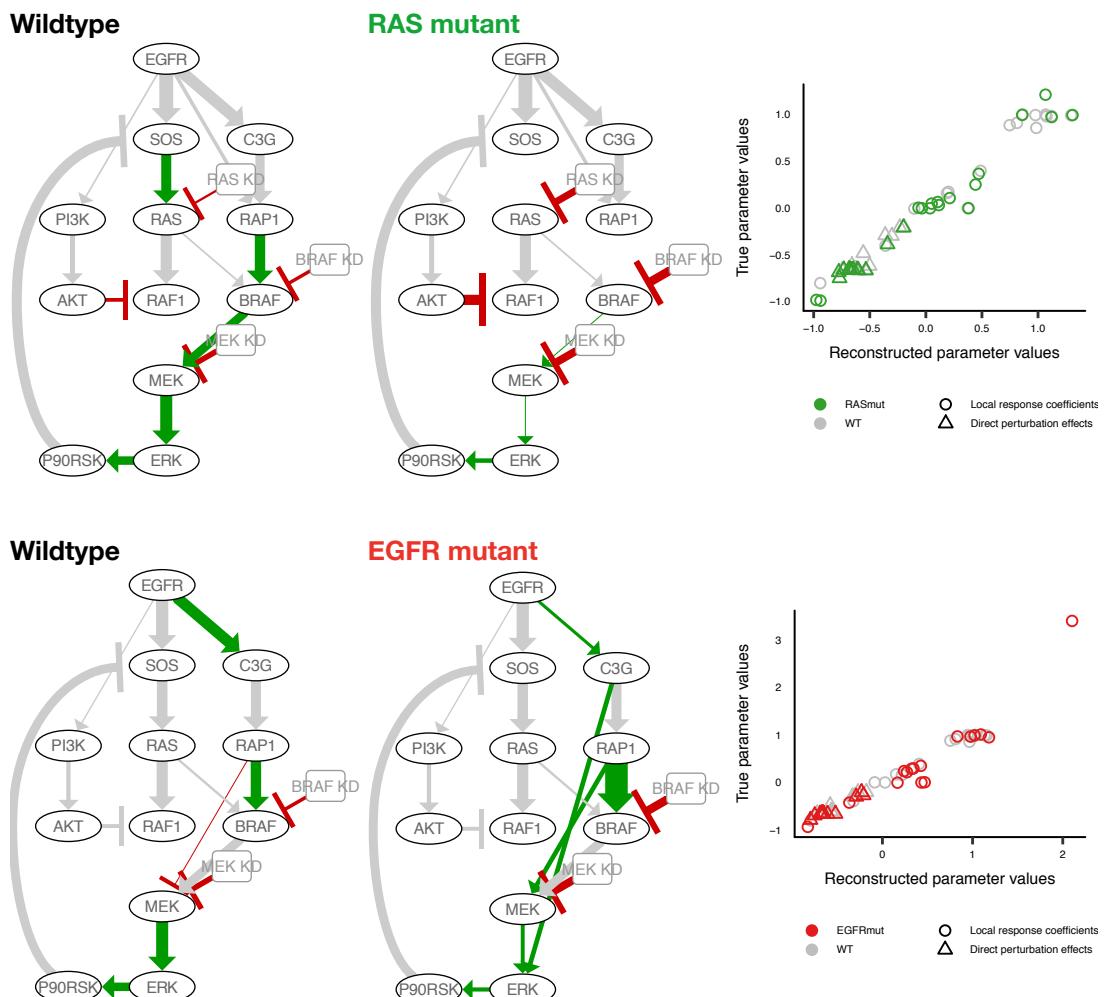
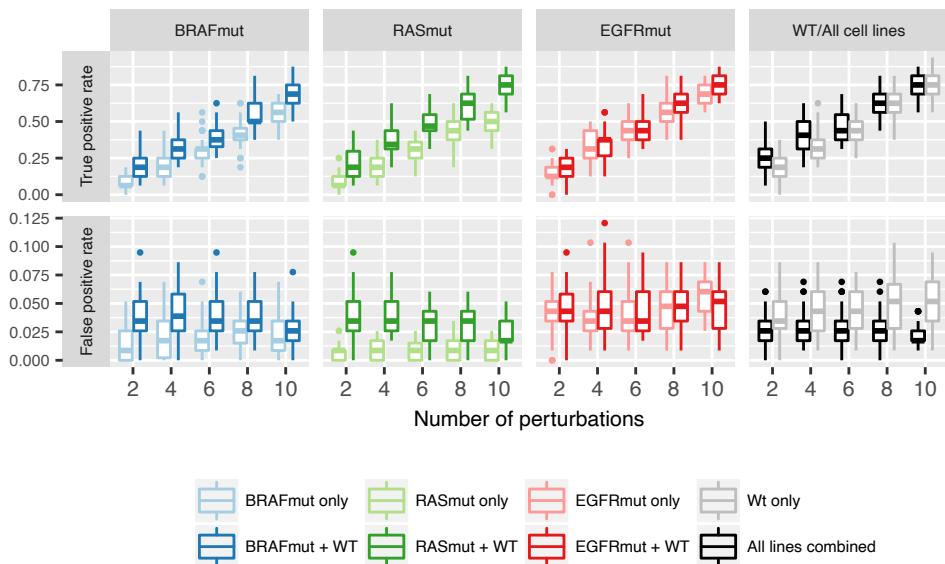
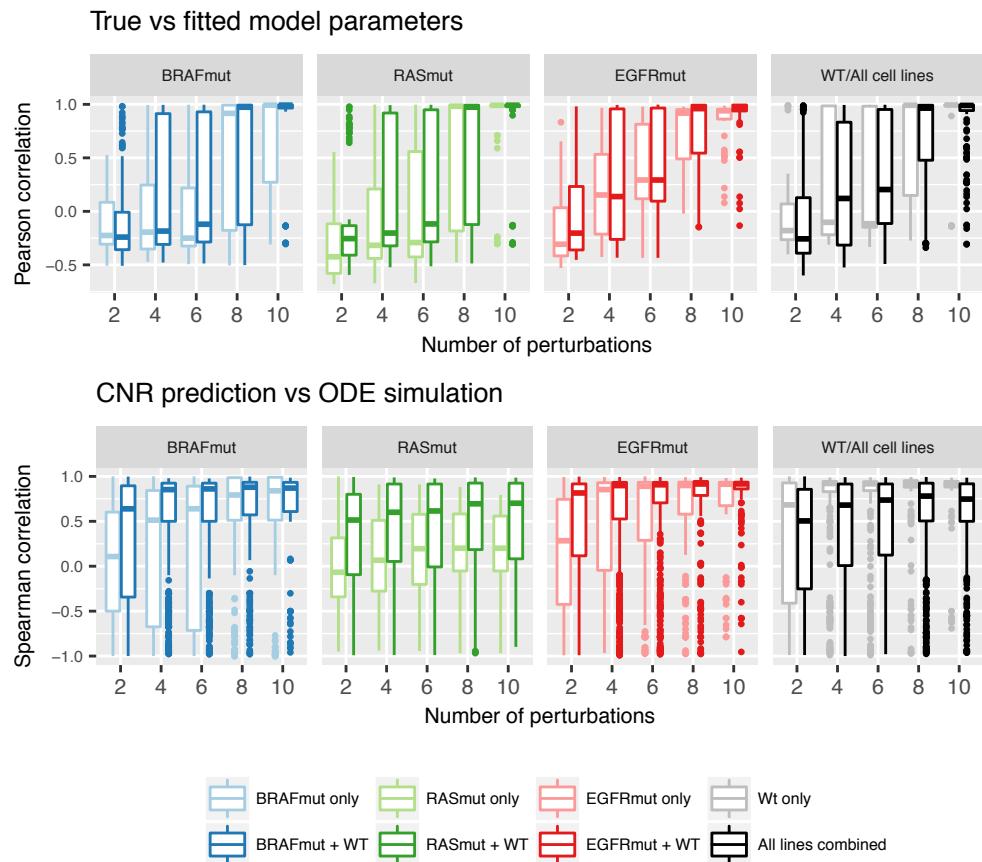


Figure S3: Network reconstruction of RAS-WT and EGFR-WT cell line panels.



**Figure S4: True and false positive rates of edge prediction for different number of perturbations used for network reconstruction.** Reconstructions were performed using different numbers of randomly selected perturbations. For each reconstruction, 10% noise was added to the original global response matrix and a fixed  $\eta$  and  $\theta$  were used. To ensure a fair comparison, the same perturbations and noisy global response matrices were used for each panel. The procedure was repeated 10 times for each number of different perturbations.



**Figure S5: Spearman correlation between predicted and simulated model parameters (top) and response to perturbations that were not used in reconstructing the network (bottom).** Network quantifications were performed using different number of randomly selected perturbations. The correct network topology was provided as input. For each reconstruction, 10% noise was added to the original global response matrix and a fixed  $\eta$  and  $\theta$  were used. The obtained network quantifications were compared to the true model parameters (top). The y-axis shows the Spearman-correlation between the reconstructed and the true local response coefficients and direct perturbation effects. The network quantification were then used to predict the response to perturbations not used in the reconstruction (bottom). The y-axis shows the Spearman-correlation between the predicted and the true response. For each perturbation, the perturbed node is excluded from the correlations since the direct-perturbation effect is not part of the prediction. Perturbations were none of the nodes respond are also excluded.

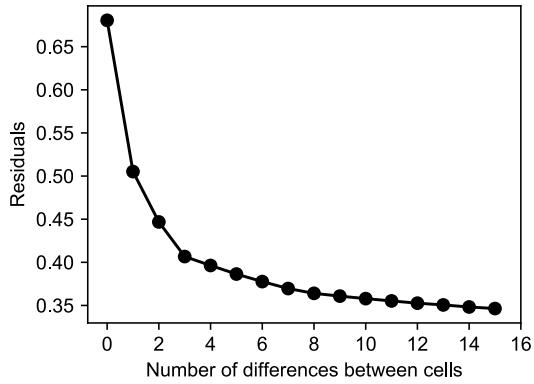


Figure S6: Relation between fitting error and allowed number of differences between cell lines for the VACO432 cell line network reconstructions.

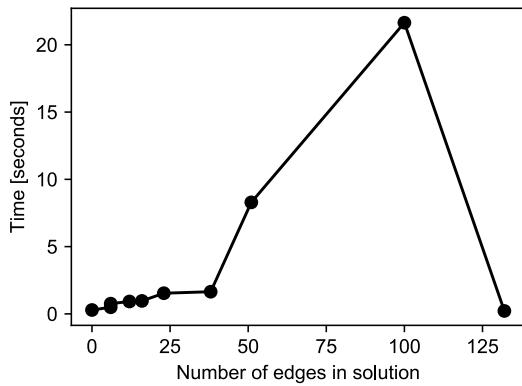


Figure S7: Time to solve the CNR problem as function of the number of edges in the solutions. Denser solutions cost more time. However, as the network becomes almost fully connected, the optimization time starts to decrease because the search-space decreases.

## References

1. Bruggeman, F. J., Westerhoff, H. V., Hoek, J. B. & Kholodenko, B. N. Modular Response Analysis of Cellular Regulatory Networks. *Journal of Theoretical Biology* **218**, 507–520 (Oct. 2002).
2. Kholodenko, B. N. *et al.* Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12841–6 (Oct. 2002).
3. Pya, N. & Wood, S. N. Shape constrained additive models. *Statistics and Computing* **25**, 543–559 (May 2015).