RESEARCH ARTICLE

Refining Pathways: A Model Comparison Approach

Giusi Moffa[1]*, Gerrit Erdmann[2], Oksana Voloshanenko[2], Christian Hundsrucker[1], Mohammad J. Sadeh[1], Michael Boutros[2], Rainer Spang[1]

1 Department of Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, Regensburg, Germany, **2** Division of Signaling and Functional Genomics, German Cancer Research Center (DKFZ) and Department of Cell and Molecular Biology, Faculty of Medicine Mannheim, Heidelberg University, Heidelberg, Germany

* giusi.moffa@gmail.com

Abstract

Cellular signalling pathways consolidate multiple molecular interactions into working models of signal propagation, amplification, and modulation. They are described and visualized as networks. Adjusting network topologies to experimental data is a key goal of systems biology. While network reconstruction algorithms like nested effects models are well established tools of computational biology, their data requirements can be prohibitive for their practical use. In this paper we suggest focussing on well defined aspects of a pathway and develop the computational tools to do so. We adapt the framework of nested effect models to focus on a specific aspect of activated Wnt signalling in HCT116 colon cancer cells: Does the activation of Wnt target genes depend on the secretion of Wnt ligands or do mutations in the signalling molecule β-catenin make this activation independent from them? We framed this question into two competing classes of models: Models that depend on Wnt ligands secretion versus those that do not. The model classes translate into restrictions of the pathways in the network topology. Wnt dependent models are more flexible than Wnt independent models. Bayes factors are the standard Bayesian tool to compare different models fairly on the data evidence. In our analysis, the Bayes factors depend on the number of potential Wnt signalling target genes included in the models. Stability analysis with respect to this number showed that the data strongly favours Wnt ligands dependent models for all realistic numbers of target genes.

OPEN ACCESS

Citation: Moffa G, Erdmann G, Voloshanenko O, Hundsrucker C, Sadeh MJ, Boutros M, et al. (2016) Refining Pathways: A Model Comparison Approach. PLoS ONE 11(6): e0155999. doi:10.1371/journal.pone.0155999

Editor: Hong Wang, Rutgers, the State Univesity of New Jersey, UNITED STATES

Received: August 20, 2015

Accepted: May 6, 2016

Published: June 1, 2016

Data Availability Statement: ArrayExpress database with accession number E-MTAB-651 http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-651/.

Funding: Work in the group of R.S. was supported by the NGFN-Project FKZ01GS08183 - Verbund Kolonkarzinom and the BioSysNet-Project - Bavarian Research Network for Molecular Biosystems. Work in the laboratory of M.B. was supported by grants from the BMBF (NGFN-plus 01GS08181-6) and BMBF GERF.

Introduction

Cellular signalling pathways like the Wnt pathway can be represented as networks modelling signal propagation, amplification and modulation of multiple molecular interactions. An important objective of systems biology is learning and adjusting the network topologies of pathways from experimental data.

During signal propagation, upstream components of the pathway control the activation of downstream components and, indirectly, the expression of target genes. Perturbations of

Competing Interests: The authors have declared that no competing interests exist.

signalling components affect the regulation of target genes and these gene expression changes reflect the topology of the pathway: Blocking an upstream component like a molecule of a receptor complex automatically blocks the activation of downstream components like signalling mediators or transcription factors. In fact, we use this concept as a definition of upstream/downstream relationships in pathways. With this definition, upstream components change the expression of more target genes then downstream components. Nested effects models (NEMs) implement this concept into an algorithm for inferring pathway topologies [1–3]. They have been successfully applied to analyse LPS-mediated signalling in Drosophila cells [1], B-cell receptor signalling in human BL2-cells [4], cellular decision making in early murin embryonic stem cells differentiation [3], the yeast mediator complex [5], rhinovirus infection mechanisms [6], or gene regulatory interaction networks in C. elegans [7].

A new perturbation data set can support the current working model of a pathway or it can falsify it. NEMs are statistical tools designed to analyze such data. They predict a topology that can be compared to the current model. To apply the basic NEM algorithm, we must specify all signalling components in a pathway, silence or inhibit all of them, monitor all target genes, and provide this data as input to the algorithm, which returns the best scoring pathway topology. There can be many obstacles to this enterprise: Lists of a pathway's signalling components and target genes are incomplete, or they differ between references. It might be too expensive or even impossible to perform all perturbation experiments. Even if we accomplish all experiments, the pathway with all its fine details can be so large that computing its topology becomes prohibitive. We argue that in these cases resolving the full pathway is too ambitious.

A working model is falsified, if a specific aspect of the topology like an edge is in clear conflict to data. For instance, a pathway model that does not include cross-talk between two parallel branches of a pathway, while the data cannot be explained without such cross-talk. The question whether there is cross talk reduces to a single bit of information: a *yes* or a *no*. The basic NEM algorithm will provide a complex topology. A topology that with limited input data is very questionable. However, if we combine nested effects models with statistical principles like model comparison and stability analysis we can address the questions directly focussing on the single bit of information desired. In a statistical case study on Wnt signalling in HCT116 colon cancer cells, we show how properties of the pathway's topology can be studied using only a few perturbation assays and easily feasible computations.

More than 85% [8] of sporadic colorectal cancers (CRC) harbour mutations in proteins which regulate the canonical Wnt signalling pathway, like the adenomatous polyposis coli gene (*APC*) or the *CTNNB1/β-catenin* gene. Upon activation of the pathway, the destruction complex dissociates, *β*-catenin accumulates and translocates to the nucleus. There *β*-catenin acts as a transcriptional coactivator for TCF transcription factors (including *TCF7L2*) leading to the regulation of target genes such as *AXIN2*. Many cells produce Wnt ligands that signal in an autocrine or paracrine manner. Wnt ligands secretion depends on the Wnt-specific secretion factor Evi/Wls. The cell line HCT116 has a mutated CTNNB1 gene, which encodes the signalling protein *β*-catenin. We focus on the following question: Does the activation of Wnt target genes still depend on the secretion of Wnt ligands (a) or do the CTNNB1 mutations make this activation independent from them (b)?

We framed the binary question into two competing classes of pathway models: Models that depend on Wnt ligand secretion versus those that do not. The model classes translate into restrictions of the pathway topology space. Since Wnt dependent models are more flexible than Wnt independent models, we need to account for the extra complexity in our comparison of the two models. Bayes factors do so and fairly compare the data evidence for competing models. In our analysis, the Bayes factors depended on the number of potential Wnt signalling target genes included in the models. Stability analysis showed that the data strongly favours Wnt ligand dependent models for all (realistic) numbers of Wnt target genes.

For this analysis, we performed RNA interference experiments in HCT116 cells where we silenced just four genes: *EVI/WLS*, *APC*, *CTNNB1/β-catenin*, and *TCF7L2*. We monitored changes in target gene expression by RNAseq. Details about the experimental protocols are given in the appendix. This previously unpublished dataset is available at the ArrayExpress database with accession number E-MTAB-651. Supplementary material for full reproducibility of the analysis is available on github at https://github.com/annlia/featureNEM.

Methods

Restrictions on the topology of a signalling network can be sustained or rejected using Bayes factors on topology classes

Let f be an arbitrary feature of a topology, we denote by $\mathcal{C}_f$ the class of all topologies that exhibit feature f and by $\mathcal{C}_{\bar{f}}$ the complimentary class of topologies that do not have it. Bayes factors [9] are the classical Bayesian tools for model comparison, where we wish to quantify the evidence which the data $\boldsymbol{D}$ provides in favour of a model $\mathcal{M}_1$ relative to a different model $\mathcal{M}_0$. The Bayes factor B_{10} of model $\mathcal{M}_1$ versus $\mathcal{M}_0$ is defined as the ratio between the posterior and the prior odds

$$B_{10} = \frac{P(\mathcal{M}_1|\boldsymbol{D})/P(\mathcal{M}_0|\boldsymbol{D})}{P(\mathcal{M}_1)/P(\mathcal{M}_0)} \equiv \frac{P(\boldsymbol{D}|\mathcal{M}_1)}{P(\boldsymbol{D}|\mathcal{M}_0)}$$

which is the ratio between the models' marginal likelihoods. The Bayes factor provides more protection against over-fitting compared to log-likelihood ratios, since the parameters θ_m are integrated out rather than maximised

$$P(\boldsymbol{D}|\mathcal{M}_m) = \int_{\Theta_m} P(\boldsymbol{D}|\theta_m, \mathcal{M}_m)P(\theta_m|\mathcal{M}_m)\mathrm{d}\theta_m, \qquad m = 0, 1$$

If the models consist of topology classes: a class $\mathcal{C}_f$ of topologies with the feature f and its complement class $\mathcal{C}_{\bar{f}}$, the Bayes factor is

$$B_{f\bar{f}} = \frac{P(\boldsymbol{D}|\mathcal{C}_f)}{P(\boldsymbol{D}|\mathcal{C}_{\bar{f}})}$$

The models in each class will be sets of NEMs. To evaluate the Bayes factors we therefore need the mathematical treatment of their marginal likelihoods $P(\boldsymbol{D}|C_f)$, which we turn to next.

Quantifying the evidence for NEM structural features via Bayes factors

NEMs are probabilistic models for reverse engineering signalling pathways [1]. They represent pathways by directed graphs with distinct nodes for the signalling proteins (S-genes) and for the effected target genes (E-genes) that change expression in response to perturbations of S-genes. A NEM predicts that an E-gene changes expression when blocking the S-gene S, if and only if it is connected to S or to a descendent of S. Fig 1 gives examples for a nested effect model (A) and the expected data pattern it encodes (B).

In the Bayesian framework of [1] full network topologies are scored by their marginal posterior probability given the data $\boldsymbol{D}$, while integrating out the parameters which describe how E-genes attach to S-genes. Topologies with the same transitive closure cannot be distinguished on the basis of data because they are score equivalent [1]. Let $\Phi = (\phi_{ij})$ be the adjacency matrix of a

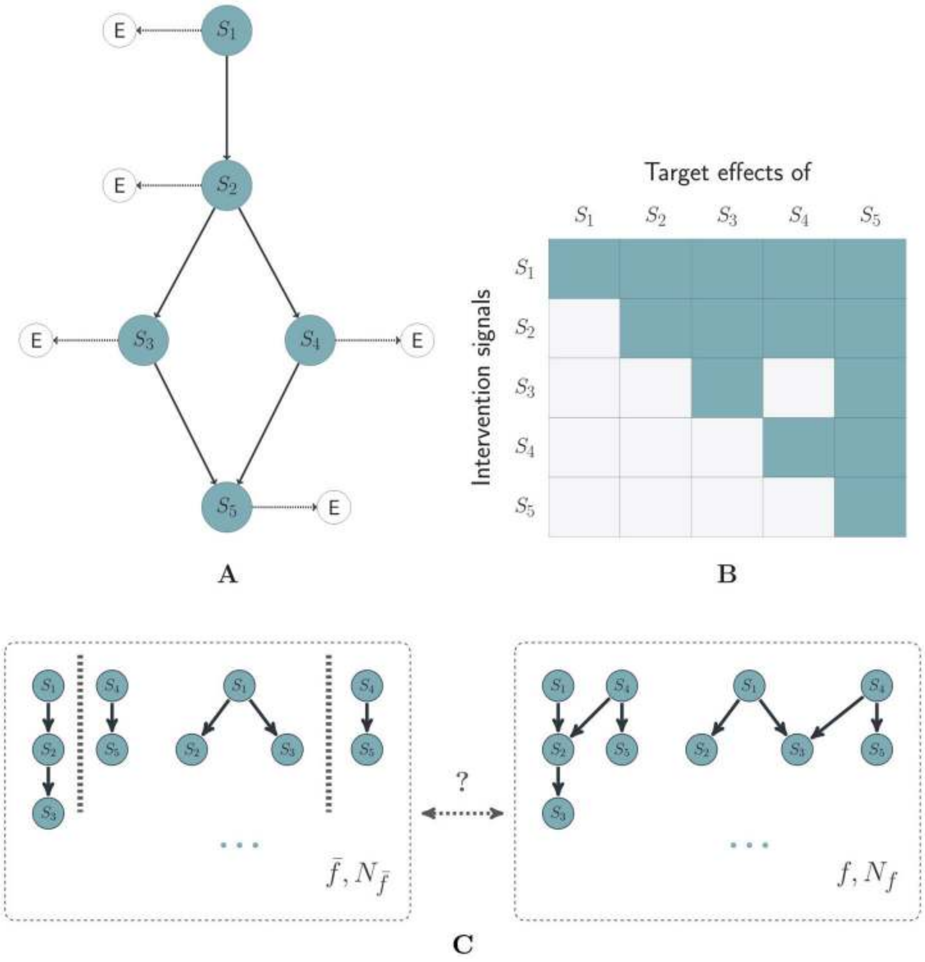

Fig 1. Nested effect models and structural features. An example of a NEM network, its expected observation pattern and structural features. **A** shows a signalling network including S- and E-genes. **B** shows the corresponding expected downstream effects of E-genes that are attached to S-genes (columns), if the S-genes in the rows are blocked. **C** shows classes of topologies that share a specific feature. In this case the distinguishing feature is the existence of an edge that connects the sub-networks formed by (S_1, S_2, S_3) and (S_4, S_5). The topologies on the right have this feature while those on the left do not.

doi:10.1371/journal.pone.0155999.g001

transitively closed network, $\theta = (\theta_l)$ be the attachment parameters, such that $\theta_l = k$ if E_l is attached to S_k, and $\boldsymbol{D} = (d_{lk})$ a data matrix whose elements quantify the effect observed for E_l when 'blocking' S_k. The network posterior probability is

$$P(\Phi|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\Phi)P(\Phi)}{P(\boldsymbol{D})}$$

The marginal likelihood $P(\boldsymbol{D}|\Phi)$ is obtained by summing over θ

$$P(\boldsymbol{D}|\Phi) = \sum_{\theta} P(\boldsymbol{D}|\theta, \Phi)P(\theta|\Phi) \quad (1)$$

where $P(\theta|\Phi)$ is a prior distribution for θ and is assumed to be independent of the network structure Φ, such that it can be written as $P(\theta)$.

Defining $\boldsymbol{D}_l$ as the l-th row of the data matrix, or in other words the vector of observations corresponding to E-gene E_l, the likelihood decomposes as

$$P(\boldsymbol{D}|\Phi,\Theta)=\prod_{l=1}^{L}P(\boldsymbol{D}_l|\Phi,\theta_l),\quad P(\boldsymbol{D}_l|\Phi,\theta_l)=\prod_{k=1}^{K}P(d_{lk}|\Phi,\theta_l) \tag{2}$$

Different likelihood models $P(d_{lk}|\Phi, \theta_l)$ exist for discrete [1] and continuous data [10]. Expression profiling data is better described as continuous, so we chose a Bayesian version of the continous data likelihood. Expression changes of E-genes were provided as posterior probabilities p_{lk}, where l refers to an E-gene and k to a perturbation. We used the Bayesian linear modelling implemented in the `limma` package [11–13] to calculate these posterior probabilities. Finally, network topologies are scored by the marginal likelihood in Eq (1) where the terms in the likelihood Eq (2) are defined as

$$P(d_{lk})=\begin{cases} p_{lk} & \text{if } \Phi \text{ predicts an effect}\\ 1-p_{lk} & \text{otherwise}\end{cases}$$

Fig 1 shows an example of two classes of topologies $\mathcal{C}_f$ and $\mathcal{C}_{\bar{f}}$ that are distinguished by a feature f. Let N be the total number of admitted topologies on the set of S-genes and N_f and $N_{\bar{f}}$ the numbers of topologies in $\mathcal{C}_f$ and $\mathcal{C}_{\bar{f}}$ respectively. The Bayes factor for class $\mathcal{C}_f$ versus $\mathcal{C}_{\bar{f}}$ can be written as

$$\frac{P(\boldsymbol{D}|\mathcal{C}_f)}{P(\boldsymbol{D}|\mathcal{C}_{\bar{f}})}=\frac{\sum_{n=1}^{N}P(\boldsymbol{D},\boldsymbol{\Phi}_n|\mathcal{C}_f)}{\sum_{n=1}^{N}P(\boldsymbol{D},\boldsymbol{\Phi}_n|\mathcal{C}_{\bar{f}})}=\frac{\sum_{n=1}^{N}P(\boldsymbol{D}|\boldsymbol{\Phi}_n,\mathcal{C}_f)P(\boldsymbol{\Phi}_n|\mathcal{C}_f)}{\sum_{n=1}^{N}P(\boldsymbol{D}|\boldsymbol{\Phi}_n,\mathcal{C}_{\bar{f}})P(\boldsymbol{\Phi}_n|\mathcal{C}_{\bar{f}})} \tag{3}$$

Given the topology Φ_n the marginal likelihood of the data $\boldsymbol{D}$ is independent of the class C_m to which the topology belongs

$$P(\boldsymbol{D}|\boldsymbol{\Phi}_n,\mathcal{C}_m)\equiv P(\boldsymbol{D}|\boldsymbol{\Phi}_n)$$

for all $n \in \{1, \ldots, N\}$ and for $m = f,\bar{f}$. The prior distribution $P(\boldsymbol{\Phi}_n|\mathcal{C}_m)$ of the network topologies is assumed to be uniform within each class

$$P(\boldsymbol{\Phi}_n|\mathcal{C}_m)=\frac{1}{N_m}I_{\{\boldsymbol{\Phi}_n\in\mathcal{C}_m\}}$$

for $m = f,\bar{f}$, with I the indicator function. The Bayes Factor then reduces to

$$\frac{P(\boldsymbol{D}|\mathcal{C}_f)}{P(\boldsymbol{D}|\mathcal{C}_{\bar{f}})}=\frac{N_{\bar{f}}}{N_f}\frac{\sum_{\boldsymbol{\Phi}_n\in\mathcal{C}_f}P(\boldsymbol{D}|\boldsymbol{\Phi}_n)}{\sum_{\boldsymbol{\Phi}_n\in\mathcal{C}_{\bar{f}}}P(\boldsymbol{D}|\boldsymbol{\Phi}_n)} \tag{4}$$

where we can see that the class $\mathcal{C}_f$ is penalised by the ratio $N_{\bar{f}}/N_f$ if it allows for more topologies than $\mathcal{C}_{\bar{f}}$.

Which of the two classes $\mathcal{C}_f$ and $\mathcal{C}_{\bar{f}}$ describes the data better? If $\mathcal{C}_f$ is larger then $\mathcal{C}_{\bar{f}}$ does the data support this extra complexity of the class? The Bayes factor helps to answer exactly these questions. If the Bayes factor exceeds 1, the evidence favours including feature f in the working model of a pathway, otherwise f should not be part the model. More refined inference comes from including prior beliefs on the two model classes and considering the ratio of posterior probabilities.

Results

Perturbations of Wnt signalling in colon cancer HCT116 cells

We here describe a case study of a focused pathway analysis based on NEMs. To this end, we investigate canonical Wnt signalling in colon cancer HCT116 cells. These cells carry a heterozygous one amino acid (δSERer45 [14, 15]) deletion in the β-catenin protein [16] which has been associated with constitutively active canonical Wnt signalling [15]. Recently, it was proposed that independent of constantly active mutation of β-catenin upstream components (like Evi/Wls), Lrp5/6 and Dvls regulate the canonical Wnt pathway [17].

We depleted signalling molecules at different levels of the Wnt pathway using RNA interference (RNAi). Specifically, we silenced EVI/WLS, APC, CTNNB1/β-catenin, and TCF7L2 by RNAi. Changes in the expression of Wnt target genes were monitored using RNAseq. In addition HCT116 cells were treated with a non-targeting siRNA as a control. All experiments were replicated twice, with the exception of TCF7L2, for which only one biological replicate was available. In addition, two samples of HCT116 cells were treated with a nonsense siRNA and profiled as control. Silencing efficiency of the genes was confirmed by qPCR (data not shown) and showed efficiencies above 80% (Fig 2A). On average 37 million sequence reads were generated per knock-down. Reproducibility of the log-cpm value between the two replicates of the same experiment is shown in Fig 2B for the EVI/WLS knock-down. Reproducibility of log-cpm values translates into reproducibility of log fold changes of cpm values, when comparing perturbations to controls (Fig 2C). Similar reproducibility could also be observed for the other experiments. The fold changes also matched the qPCR knock-down results. For example, RNAi depletion of CTNNB1/β-catenin led to a log fold change of the CTNNB1/β-catenin gene of -3.98, and a log fold change of -3.71 in the Wnt target gene AXIN2, whereas non-related genes such as ACTB were not affected (S1 Fig). The RNAseq data also confirmed the presence of the mutated CTNNB1/β-catenin allele, which expressed at higher levels than the wild-type allele (S1 Fig). Indeed, we could not detect any phosphorylation of S45 which is deleted in the mutated allele, suggesting that in HCT116 mutated CTNNB1/β-catenin is more abundant than in the wild-type.

We preprocessed the count data using the `voom` function of the `limma` package [13] and calculated fitted Bayesian linear models to the data using the `limma` package [11]. From these models we calculated the posterior probability that the expression of a gene is affected by the

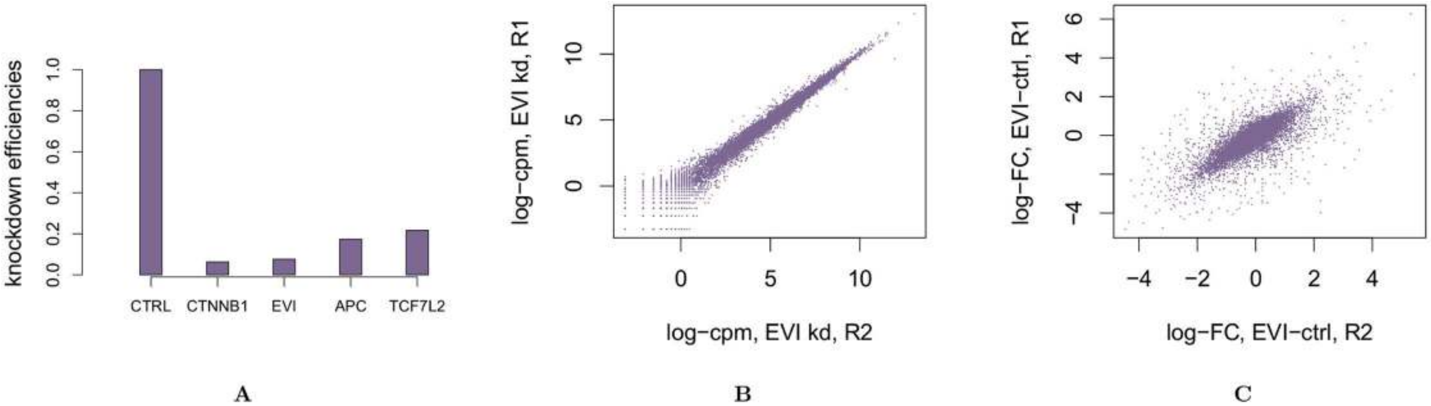

Fig 2. Experiments' efficiency and reproducibility. Panel **A** reports the relative expression of the direct siRNA targets with respect to the control in the corresponding experiment. log-cpm measured in the two replicates of the EVI silencing experiment are shown in panel **B**. Panel **C** shows the log fold changes between the cpm measured in two pairs of experiment, namely EVI silencing and CTRL.

doi:10.1371/journal.pone.0155999.g002

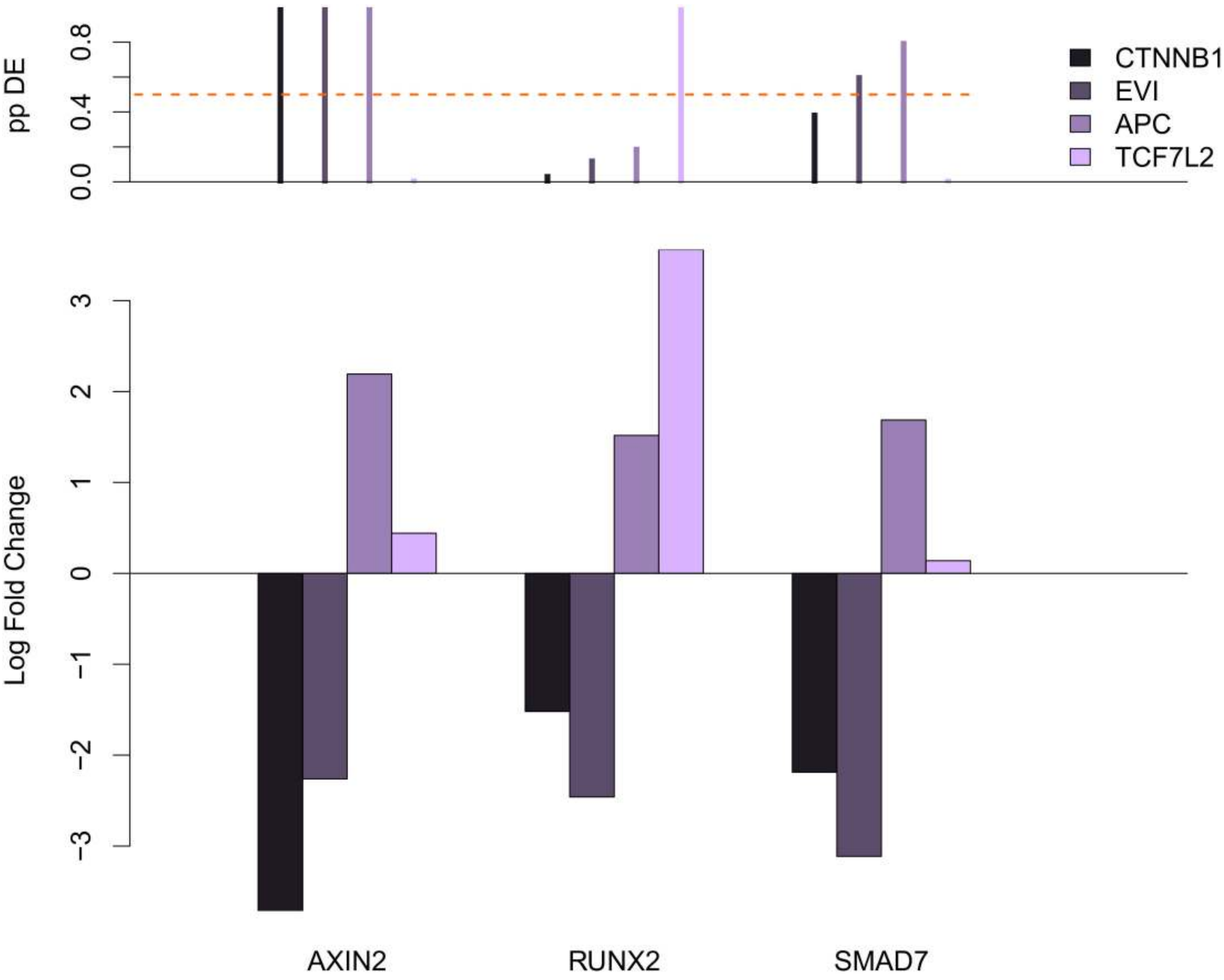

Fig 3. Expression of Wnt target genes. Comparison of gene expression profiles after siRNA mediated knock-down of Wnt pathway components show high similarity between *EVI*/*WLS* and downstream positive pathway regulators. The top panel shows the posterior probability of differential expression in the experiments, with the dashed line marking probability.5. There is strong evidence for differential expression of the canonical target gene AXIN2 when silencing *CTNNB1*/*β-catenin*, *EVI*/*WLS* and *APC*. RUNX2 and SMAD7 show a similar pattern of opposite regulation after knock-down of *APC* and *CTNNB1*/*β-catenin*, but while for SMAD7 the posterior probability of differential expression are still above a half for the *EVI*/*WLS* and *APC* knock-down, the probabilities for RUNX2 are below, which translates into reduced evidence for reproducibility of the fold changes.

doi:10.1371/journal.pone.0155999.g003

knock-down again using `limma`. For the direct targets of the siRNAs the posterior probabilities of change were all estimated as effectively 1. A probability of nearly 1 was also obtained for *AXIN2* under depletion of *CTNNB1*/*β-catenin*, while a probability well below.01 was estimated for the non Wnt target *ACTB* under the same intervention. Fold changes and corresponding posterior probabilities for a number of well known Wnt targets are shown in Fig 3.

Transcriptome wide downstream effects are outlined in Fig 4 which shows a heat map of posterior probabilities for all four knock-downs and 2978 genes. Red corresponds to high probabilities of expression change and blue to virtually zero probability. All 2978 genes showed a high probability in at least one condition. The genes to the left of the green line reacted to both

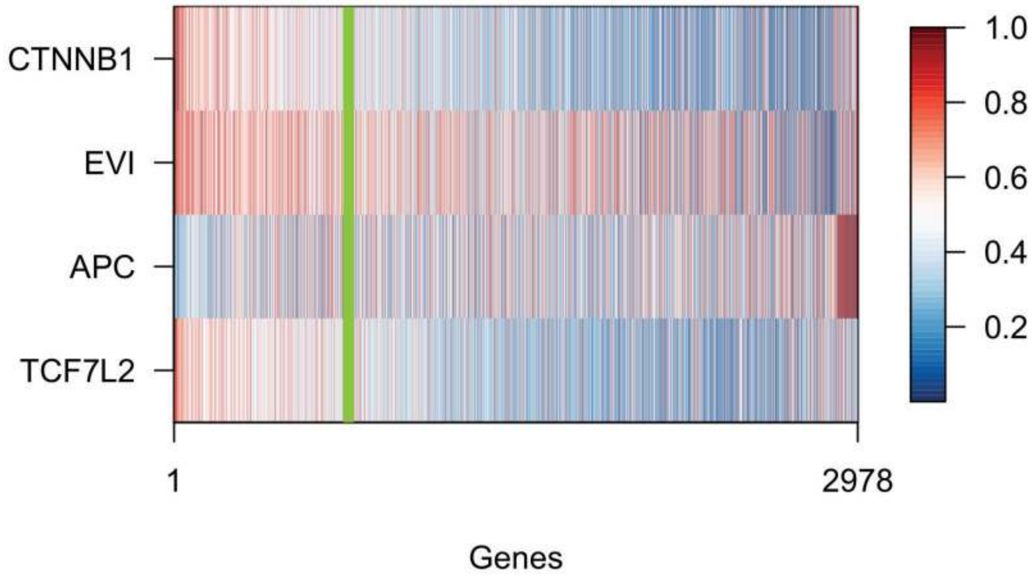

Fig 4. Posterior probabilities of differential expression. Heat-map of the posterior probabilities of differential expression in the silencing experiments. Only genes which have a posterior probability larger than .5 in at least one of the knock-down experiments are shown. The green line leaves about 750 genes on its left. The pattern there shows that the majority of those genes respond not only to intervention on *CTNNB1/β-catenin*, but also to *EVI/WLS*.

doi:10.1371/journal.pone.0155999.g004

CTNNB1/β-catenin and *EVI/WLS* knock-downs. If we considered them all as bona fide Wnt target genes, the heat map would already give compelling evidence that the activation of Wnt target genes in fact depends on Evi/Wls and thus on Wnts secretion. Interestingly, most of these genes also responded to blocking TCF7L2 but only few of them to APC. To the right of the green line are some genes that are blue in the *EVI/WLS* row but red in the *CTNNB1/β-catenin* row. They respond to *CTNNB1/β-catenin* but not *EVI/WLS*. Nevertheless, they are a small minority.

A focused analysis of Wnt signalling in HCT116 cells using Bayes factors

We now examine the descriptive arguments of the previous section with sound statistical inference. Our leading question is: Does the activation of Wnt target genes in HCT116 cells depend on the secretion of Wnt ligands (a) or do the CTNNB1 mutations make this activation independent from them (b)? And to answer this question we rely on perturbation data for four proteins *EVI/WLS*, *APC*, *CTNNB1/β-catenin*, and *TCF7L2*. The mutated gene CTNNB1 encodes for *CTNNB1/β-catenin*. Hence, the first two proteins (*EVI/WLS*, *APC*) operate upstream of the mutated protein, the third protein is mutated, and the last protein (*TCF7L2*) operates downstream of the mutated protein. All four proteins together constitute only a tiny fragment of the Wnt pathway.

We represented the two competing models (a) and (b) by classes of topologies. For the Wnt independent topologies (b) we requested that there are no edges connecting (*EVI/WLS* and *APC*) with (*CTNNB1/β-catenin* and *TCF7L2*). In contrast, for the Wnt dependent topologies (a) we requested that there must be at least one edge between the two groups of proteins. There were more topologies in the Wnt dependent class making the dependent model more complex and more flexible. A Wnt dependent topology fitted the data best. But was this simply because there are more of them, or did the data justify this additional model complexity? We had translated our initial biological question into a statistical model comparison problem. We chose Bayes factors to address it. The computational implications were already addressed in the method section.

A practical difficulty emerged when it came to naming all Wnt target genes. In our model they played the role of E-genes. Some canonical Wnt target genes, like AXIN2, SMAD7 and MYC, are well established; but there is no consensus on a complete list of Wnt targets. Most likely, the Wnt pathway regulates different target genes in different cellular systems. We decided to use our own RNAseq data to screen for potential Wnt targets in HCT116 cells. To this end, we ranked all genes by the maximum posterior probability across the four perturbations. For a cut-off λ we included all genes with a score above λ. This reduced the E-gene selection problem to a parameter calibration problem. Recalling Eq (2) we saw that determining the number of E-genes was crucial because this number drives the right hand product of conditional probabilities. So, how many Wnt target genes are there in HCT116 cells? The Nusse lab currently names 30 genes as Wnt targets in human colon cancer on their *The Wnt Homepage* resource http://web.stanford.edu/group/nusselab/cgi-bin/wnt/target_genes. All of them are backed up by publications. This list does not claim to be complete. In a more data driven approach we relied on our posterior probabilities and cut the list at a threshold of 95%. This gave us 1483 potential target genes with high support from data. In fact, determining the correct number of Wnt target genes in HCT116 cells might be a very hard problem, and not even well defined. Fortunately, stability analysis showed that we do not need an exact solution.

We run the Bayes factor analysis for the full range of cut-offs between 0 and 1. Fig 5B plots the log Bayes factor against the number of E-genes used to calculate it. In the context of our initial question: Does the activation of Wnt target genes in HCT116 cells still depend on the secretion of Wnt ligands (a) or do the CTNNB1 mutations make this activation independent from them (b)? Positive log Bayes factors are evidence in favor of (a) while negative values support (b). We considered values above 10 or below -10 as strong evidence as they correspond to odds ratios exceeding 1:1000. If we used the 1483 E-genes corresponding to a 95% posterior probability cut-off we reached log Bayes factors much higher than 10 (dashed yellow line). If we used only the top 200 or the top 30 genes the evidence in favour of a Wnt secretion dependent model was even stronger. In fact, any model based on less then 2200 potential target genes endorsed Wnt secretion dependence (dashed purple line). This number greatly exceeds any number of direct Wnt target genes reported so far. With even lower cut-offs, we believe that noise dominated the models. In summary, although we do not know how many Wnt target genes there are in HCT116 cells, we could confidently answer our leading question. Our answer is (a): The activation of Wnt target genes in HCT116 cells depends on the secretion of Wnt ligands. All models with a realistic number of E-genes consistently answered (a).

Code and data for reproducing the plot of Fig 5B are provided as supplementary material.

Discussion

In the context of a statistical case study we illustrated here how NEM modelling can be combined with model comparison and stability analysis to analyse complex pathways with limited data. We developed the computational implications of Bayes factor analysis in the context of NEM modelling with restricted topology classes. In the case study we were confronted with the statistical question on whether a perturbation data set justified the refinement of an existing simpler pathway model. By allowing for additional edges the pathway model became more complex, and naturally fitted data better. The Bayes factors penalized the additional model

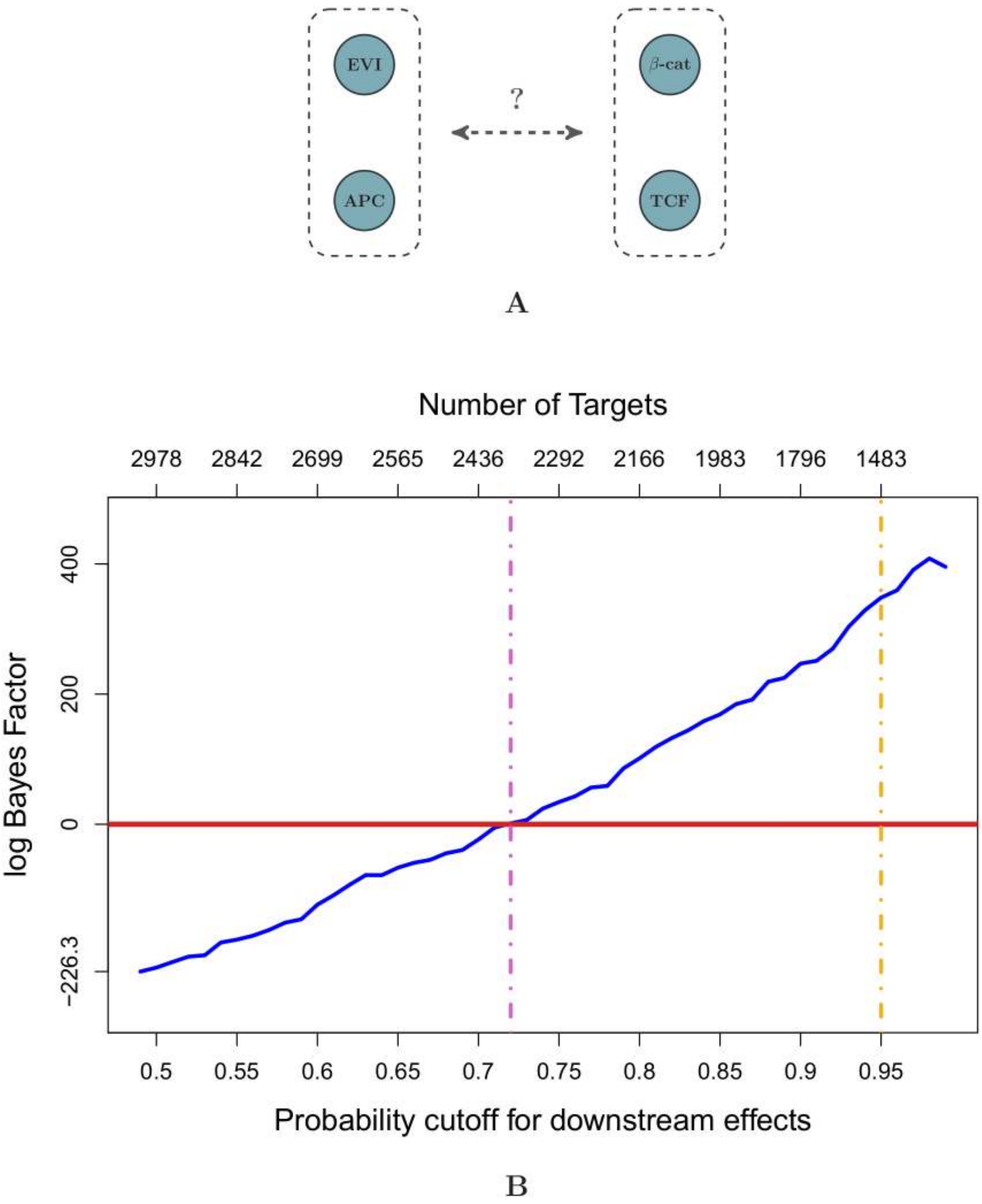

Fig 5. Model comparison for network structural features. Panel **A** is a schematic representation of the two topology classes compared. Panel **B** summarizes the main result of our statistical case study. Log Bayes Factors were obtained from NEMs including different numbers of potential target genes (x-axis, top), which are included according to a cut-off on the posterior probability that a gene is affected for at least one perturbation. Positive log Bayes factor reflect evidence in favour of a Wnt secretion dependent model. The dashed purple line indicates the smallest cut-off that still favours this topology class. A cut-off of 95% posterior probability is marked by the yellow dashed line.

doi:10.1371/journal.pone.0155999.g005

complexity but nevertheless strongly endorsed the more complex model. A complication emerged when we had to decide which and how many genes to include in the analysis. We overcame the problem by stability analysis. No matter how many genes we included, model comparison gave the same answer: The more complex model is appropriate.

We believe that strategies similar to that used in the case study can greatly improve the scope of problems in which NEM modelling can be applied. First, it is common that pathway features rather then complete pathways are under study. Second, it is also common that perturbation data is limited and does not cover all components of a pathway. Third, models of full pathways involve many parameters that need to be jointly estimated from data requiring very high numbers of repeated experiments to control the variance of the estimators. The variance of a single estimated binary parameter is easier to control. To do so, Bayesian model comparison is our method of choice. It can be used to focus analysis on a defined aspect of a pathway, it does not require data covering the full pathway, and, as shown in our case study, it provides strong collective evidence by joining individual evidence from many target genes even with small data sets.

To date, model comparison is not the standard approach to network analysis in biology. Most algorithms bank on a maximum likelihood approach, or on heuristic scoring systems. Maximum likelihood is prone to over-fitting. There is an eminent danger of deriving pathway models that are too complex. In the context of Gaussian graphical models this has been addressed for example by likelihood penalisation [18], or shrinkage [19]. For NEM no such approach existed to date. Bayes factors address the over-fitting problem. Eq (4) highlights the intrinsic penalty for larger topology classes.

In our understanding, all current working models of signalling pathways are abstractions that cannot live up to the true complexity of cellular processes. There is always data that cannot be explained by them. Pathway models focus on the dominant mechanisms of a pathway, those that are indispensable to understand it. If new data emerges that is in contradiction to an established pathway model, there is the inevitable question whether the new data provides sufficient evidence justifying a refinement of the pathway model. Especially, because every refinement comes with increasing complexity, that makes a pathway more difficult to understand. Model comparison addresses exactly this problem.

In our case study on the Wnt pathway we could sustain the proposition that the activation of Wnt target genes in colon cancer cells depends on Wnts secretion, at least for HCT116 cells. This result has been previously reported and endorsed by a series of biochemical experiments [17]. Our modelling adds a new aspect to the previous work. We analyse Wnt signalling controlled gene regulation on a transcriptome wide level. Surprisingly, we observed more than a thousand genes that changed expression to knock-downs of Wnt signalling components at a 95% posterior probability level. This included both activated and repressed genes suggesting that the Wnt pathway directly or indirectly affects many cellular processes, maybe even more than known today.

A limitation of the analysis in our study is that conclusions can only be drawn for the investigated HCT116 cells lines and not for colon cancer in general. We cannot exclude effects from the culturing or that there are colon cancers where the Wnt pathway operate in a Wnt secretion independent mode. To further uncover the important mechanisms in the Wnt pathway and colon tumour development, a comparative study of a diverse set of tumorous, as well as non-tumorous, colon primary cells would therefore be highly valuable. From such data, one could then refine the pathway inference and draw more generalised conclusions again utilising the general framework of network model comparison developed in this manuscript.

Supporting Information

S1 Code. Reproducible analysis. For the sake of allowing full reproducibility the complete code for the analysis presented in the manuscript is available on github at https://github.com/annlia/featureNEM.
(PDF)

S1 Protocols. Protocols. Details about the experimental protocols are given in the supplementary material.
(PDF)

S1 Fig. Sequence reads. Examples about the distribution of the sequence reads, with a quantification of the expression of both β-catenin alleles (mutated and wild-type) in HCT116 cells are reported in the supplementary material.
(PDF)

S2 Fig. Phosphorylation. Results from Western blot analysis are reported in the supplementary material.
(PDF)

S1 Alternative Analysis. No-CONAN analysis. Details about the No-CONAN approach.
(PDF)

Acknowledgments

Funding: Work in the group of R.S. was supported by the NGFN-Project FKZ01GS08183—Verbund Kolonkarzinom and the BioSysNet-Project—Bavarian Research Network for Molecular Biosystems. Work in the laboratory of M.B. was supported by grants from the BMBF (NGFN-plus 01GS08181-6) and BMBF GERF.

Author Contributions

Conceived and designed the experiments: MB RS. Performed the experiments: OV GE. Analyzed the data: GM CH. Wrote the paper: GM RS. Developed the statistical method: GM RS MJS.

References

1. Markowetz F, Bloch J, Spang R. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics. 2005; 21(21):4026–4032. doi: 10.1093/bioinformatics/bti662 PMID: 16159925
2. Markowetz F, Kostka D, Troyanskaya OG, Spang R. Nested effects models for high-dimensional phenotyping screens. Bioinformatics. 2007; 23(13):i305–i312. doi: 10.1093/bioinformatics/btm178 PMID: 17646311
3. Anchang B, Sadeh MJ, Jacob J, Tresch A, Vlad MO, Oefner PJ, et al. Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. Proceedings of the National Academy of Sciences. 2009; 106(16):6447–6452. doi: 10.1073/pnas.0809822106
4. Pirkl M, Hand E, Kube D, Spang R. Analyzing synergistic and non-synergistic interactions in signalling pathways using Boolean Nested Effect Models. Bioinformatics. 2015; p. btv680.
5. Niederberger T, Etzold S, Lidschreiber M, Maier KC, Martin DE, Fröhlich H, et al. MC EMiNEM Maps the Interaction Landscape of the Mediator. PLoS Comput Biol. 2012; 8(6):e1002568. doi: 10.1371/journal.pcbi.1002568 PMID: 22737066
6. Siebourg-Polster J, Mudrak D, Emmenlauer M, Rämö P, Dehio C, Greber U, et al. NEMix: Single-cell nested effects models for probabilistic pathway stimulation. PLoS Comput Biol. 2015; 1(2):e1004078. doi: 10.1371/journal.pcbi.1004078
7. Mac Neil L, Pons C, Arda EH, Giese GE, Myers CL, Walhout A. Transcription Factor Activity Mapping of a Tissue-Specific In Vivo Gene Regulatory Network. Cell Systemsl. 2015; 11(4):152–162. doi: 10.1016/j.cels.2015.08.003
8. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. doi: 10.1038/nature11252
9. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995; 90 (430):773–795. doi: 10.1080/01621459.1995.10476572

10. Fröhlich H, Fellmann M, Sültmann H, Poustka A, Beissbarth T. Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data. Bioinformatics. 2008; 24(22):2650–2656. doi: 10.1093/bioinformatics/btm634 PMID: 18227117

11. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer; 2005. p. 397–420.

12. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. 2004; 3. doi: 10.2202/1544-6115.1027 PMID: 16646809

13. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014; 15:R29. doi: 10.1186/gb-2014-15-2-r29 PMID: 24485249

14. Ilyas M, Tomlinson I, Rowan A, Pignatelli M, Bodmer W. β-Catenin mutations in cell lines established from human colorectal cancers. Proceedings of the National Academy of Sciences. 1997; 94 (19):10330–10334. doi: 10.1073/pnas.94.19.10330

15. Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, et al. Activation of β-catenin-Tcf signaling in colon cancer by mutations in β-catenin or APC. Science. 1997; 275(5307):1787–1790. doi: 10.1126/science.275.5307.1787 PMID: 9065402

16. Sekine S, Shibata T, Sakamoto M, Hirohashi S. Target disruption of the mutant beta-catenin gene in colon cancer cell line HCT116: preservation of its malignant phenotype. Oncogene. 2002; 21(38):5906. doi: 10.1038/sj.onc.1205756 PMID: 12185590

17. Voloshanenko O, Erdmann G, Dubash TD, Augustin I, Metzig M, Moffa G, et al. Wnt secretion is required to maintain high levels of Wnt activity in colon cancer cells. Nature communications. 2013; 4. doi: 10.1038/ncomms3610 PMID: 24162018

18. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The annals of statistics. 2006; p. 1436–1462. doi: 10.1214/009053606000000281

19. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology. 2005; 4(1). PMID: 16646851