Title

Selection of personalized patient therapy through the use of knowledge-based computational models that identify tumor-driving signal transduction pathways

Authors

Wim Verhaegh¹, Henk van Ooijen¹, Márcia A. Inda¹, Pantelis Hatzis²*, Rogier Versteeg³, Marcel Smid⁴, John Martens⁴, John Foekens⁴, Paul van de Wiel¹, Hans Clevers², Anja van de Stolpe¹

¹Molecular Diagnostics, Philips Research, Eindhoven, The Netherlands

²Hubrecht Institute, Utrecht, The Netherlands

³Human Genetics, AMC Amsterdam, Amsterdam, The Netherlands

⁴Medical Oncology, Erasmus MC, Rotterdam, The Netherlands

*Pantelis Hatzis is now with BSRC Alexander Fleming, Vari, Greece

Keywords

Signal transduction pathway, computational model, targeted therapy, gene expression, companion diagnostics, breast cancer, colon cancer, medulloblastoma, hepatoblastoma, Wnt, estrogen receptor pathway.

Financial support

This work is supported by NanoNextNL, a micro and nanotechnology consortium of the Government of the Netherlands and 130 partners (www.nanonext.nl).

Corresponding author


Conflict of interest

The authors disclose no potential conflicts of interest.
Abstract

Increasing knowledge concerning signal transduction pathways as drivers of cancer growth has elicited the development of ‘targeted drugs’ which inhibit aberrant signaling pathways. They require a companion diagnostic test which identifies the tumor-driving pathway, however, currently available tests like ER protein expression for hormonal treatment of breast cancer, do not reliably predict therapy response, at least in part because they do not adequately assess functional pathway activity.

We describe a novel approach to predict signaling pathway activity, based on knowledge-based Bayesian computational models, which interpret quantitative transcriptome data as the functional output of an active signaling pathway, by using expression levels of transcriptional target genes. Following calibration on only a small number of cell lines or cohorts of patient data, they provide a reliable assessment of signaling pathway activity in tumors of different tissue origin. As proof of principle, models for the canonical Wnt and ER (estrogen receptor) pathways are presented, including initial clinical validation on independent datasets from various cancer types.

Major findings

As expected, the Wnt pathway was predicted inactive in normal colon samples and active in 97% of tested colon adenomas and carcinomas (known to be Wnt driven), and in all tested medulloblastomas containing an activating beta-catenin mutation. Furthermore, in primary liver cancer, Wnt activity was predicted in 56% of samples containing a beta-catenin mutation, against 18% without documented mutations. In breast cancer, Wnt activity was predicted in 30% of basal type breast cancers versus 7% of other subtypes, confirming expectations based on clinical research [1].

The ER pathway model predicted inactivity in practically all tested samples, except for 39% of ER-positive breast cancer samples. Furthermore, ER pathway activity was associated with increased disease-free survival compared to patients in which the pathway was predicted inactive, even though the model was not trained for this purpose.

Clinical implementation of our models is expected to enable a more informed choice of therapy and improved prediction of targeted therapy response. Furthermore, it may help to focus search
for tumor-driving genomic defects in whole genome sequencing data, to only those genes involved in the aberrantly active pathway(s).

Quick Guide to Equations and Assumptions

Major assumptions

The key assumption in the Bayesian pathway model we present is that functional activity of a signaling pathway is determined by activity of its respective transcription complex, while the latter can be inferred from mRNA expression data of its transcriptional target genes.

Furthermore, we assumed that the oncogenic signaling pathway driving tumor growth is not transiently and dynamically activated, but long term or even irreversibly. Hence the model was developed for interpretation of a static cellular condition, and complex dynamic pathway features were not incorporated.

Finally, to handle the noisy characteristics of mRNA expression data from microarrays, as well as additional unknown biological factors, we modeled signaling pathways in a probabilistic manner, using Bayesian networks.

Equations

The main equations related to our Bayesian models describe probabilistic relations between different types of nodes. Conditional probability tables describing the relation between transcription complex node TC and target gene nodes \( TG_i \) are given by equation (1), and between each target gene node \( TG_i \) and its respective probeset nodes \( PS_{i,j} \) by equations (2a) and (2b) in the supplementary material.

Introduction

Knowledge on intracellular signal transduction pathways governing cancer cell behavior and controlling cell division is rapidly increasing. This development has elicited a paradigm shift towards development of a whole new category of ‘targeted drugs’, aiming to target the aberrant
signaling pathway which drives tumor growth in the individual cancer patient [2]. In contrast to conventional chemotherapy, targeted therapy requires a highly personalized approach to treatment choice, in principle based on predicting treatment response prior to administering the drug or drug combination of choice.

The anticipated increasing availability of targeted drugs [3] stresses the need for reliable companion diagnostics to predict therapy response, for which identification of the tumor-driving signaling pathway and the underlying defect that causes its aberrant activation is of high importance [4]. Unfortunately, currently available tests often lack predictive value with respect to targeted therapy response. In general these tests demonstrate (over-)expression of key proteins in signaling pathways of interest, e.g. ER or HER2 in breast cancer, or identify DNA mutations (e.g. in the PIK3CA gene) or structural changes (like HER2 gene amplification) in genes encoding for proteins that directly influence signaling pathway activity. However, though associated with response to targeted drugs (hormonal therapy and trastuzumab in example cases ER and HER2, respectively) such tests do not provide conclusive information on the functional activity status of the associated signaling pathways [5].

Around ten major oncogenic signaling pathways play a role in tumor growth and metastasis: Estrogen and Androgen Receptor pathways (ER, AR), PI3K (activated by multiple growth factor receptors, like HER2, EGFR), canonical Wnt, Notch, Hedgehog, TGFbeta, NFkappaB, VEGF, RAS/MAPK/ERK, and FGF signaling pathways [6]; see also supplementary Section 1. Whole genome and transcriptome analysis methods, like DNA and RNA sequencing and microarray technologies, are in principle capable of producing all data necessary to extract information on signaling pathway activity from cancer tissue samples, on the premise that adequate software is available to interpret the highly complex data [7]. This has proven to be a tremendous challenge, and most efforts are directed towards identification of genotypic changes, instead of using the transcriptome to provide information on the functional phenotype of the cancer cell – which is determined in concert by both genotype and micro-environment of cancer cells [8].

We describe the development and partial clinical validation of a new type of knowledge-based probabilistic computational modeling framework for oncogenic cell signaling pathways, which enables functional assessment of pathway activity in individual tissue samples based on quantitative transcriptome data as input. The models have been built using Affymetrix HG-U133Plus2.0 data, but can be calibrated to other quantitative mRNA data formats like RNA
sequencing or other microarray types. First models of canonical Wnt and ER pathways are presented as proof of principle for prediction of pathway activity, therapy response and prognosis in cancer patients.

Materials and Methods

Development of Bayesian models for signal transduction pathways

Our signal transduction pathway modeling approach is based on inferring pathway activity from the expression profile of its target genes using probabilistic Bayesian network inference. Bayesian networks were built using the Bayes Net Toolbox for MATLAB, as detailed in the supplementary information. The Bayesian network structure used as a basis for our modeling approach (Figure 1) is a simplified model of the transcriptional program of a cellular signal transduction pathway, consisting of three types of nodes: (a) transcription complex, (b) target genes and (c) microarray probesets corresponding to target genes. The model describes (i) how expression of target genes depends on transcription complex activation, and (ii) how probeset intensities in turn depend on expression of the respective target genes. For the latter, probeset intensities are taken from fRMA preprocessed Affymetrix HG-U133Plus2.0 microarrays, widely available from Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo) and ArrayExpress (www.ebi.ac.uk/arrayexpress); an overview of used datasets is provided in supplementary Table S3.

As our pathway models are a simplification of signaling pathway biology and as biological measurements are typically noisy, we opted for a probabilistic approach, meaning that relationships (i) between transcription complex and target genes and (ii) between target genes and their respective probesets, are described in probabilistic terms. Furthermore, we assumed that the oncogenic signaling pathway driving tumor growth is not transiently and dynamically activated, but long term or even irreversibly. Hence the model was developed for interpretation of a static cellular condition, and complex dynamic pathway features were not incorporated.

Once the Bayesian network model has been built and calibrated for a particular signaling pathway (sketched below and further detailed in the supplementary information), the model can
be used on microarray data of a new tumor sample by entering probeset measurements as observations in the bottom layer, and inferring backwards in the model the activity probability of the transcription complex. This latter probability is hence the primary read-out used to indicate pathway activity, which can be translated into odds of the pathway being active by taking the ratio of the probability of being active versus inactive (i.e. odds are given by $p/(1-p)$ if $p$ is the predicted probability of being active).

Target gene selection

For optimal performance, the Bayesian network models should contain (only) direct target genes of the respective pathways. Unfortunately, pathway databases such as KEGG (www.genome.jp/kegg) and Biocarta (www.biocarta.com) are fairly incomplete and inconsistent on this aspect [9]. Hence, we manually selected target genes based on extensive scientific evidence for each individual gene being a direct target gene of the respective transcription complex, including promotor region motif analysis, transcription factor binding experiments and differential expression analysis. For Wnt, extensive research at the Hubrecht Institute over the past decades [10–13] culminated in a list of 34 ‘bona fide’ target genes listed in supplementary Table S1. For ER, we extensively investigated available literature, as detailed further in the supplementary Section 5, yielding the 27 genes in Table S2. These numbers of genes are on the one hand low enough to give specific results, but large enough to get robust models.

Model calibration

The probabilistic relations in the Bayesian network models need to be made quantitative, to allow for quantitative probabilistic reasoning. To improve generalization behavior across tissue types, we manually set parameters describing the probabilistic relationships (i) between transcription complex and target genes, as detailed in the supplementary information. Parameters describing relationships (ii) between target genes and their respective probesets were calibrated on experimental data. For the latter, we either used microarray data from cell line experiments with defined active and inactive pathway settings (Wnt and ER pathway), or from patient samples with known pathway activity status (Wnt pathway only).
Statistical tests

Statistical tests, generation of Kaplan-Meier curves and other graphics were performed using R [14]. Generally, one-sided tests were used because the expected sign of a relation is known.

Results

Wnt pathway model, initial validation

Two instances of the Wnt model were created, a first one for initial proof of concept using cell line data for calibration, and a second one using a larger calibration dataset with patient data, with the advantage of better reflecting the variation encountered across patient samples.

The first Wnt model was calibrated on data from twelve samples of a Wnt abrogation experiment on an LS174T colon cancer cell line (GSE18560) [15], of which six have an active Wnt pathway and six an inactivated Wnt pathway. For initial proof of concept, this model was tested on a dataset with 32 normal colon tissue samples and 32 colon adenoma samples from patients (GSE8671) [12]. The Wnt pathway is thought to be active in colon adenoma and inactive in normal intestinal tissue [16], and Figure 2A shows that this is almost perfectly predicted by our model. Although two of the 32 adenoma samples are predicted to have an inactive Wnt pathway, if the threshold is set at odds of 1:1, the difference with normal colon samples is highly distinctive. Note that reported odds get as large as a million to one, which is due to the model using 34 genes (83 probesets in total), so although one gene may give quite noisy information, combining 34 genes gives quite confident predictions.

Because the first calibration dataset has limited diversity, and ground truth information on Wnt pathway activity is in principle known for normal colon and adenoma, a second model was calibrated using the 32 normal colon samples and 32 colon adenoma samples from GSE8671 [12], and used this model for the rest of the Wnt experiments reported in this paper.

Wnt pathway in colon cancer

In general, in colon tumors APC tumor suppressor activity is absent due to loss of functional APC alleles, which is associated with active Wnt signaling, providing an excellent opportunity for clinical evaluation of the Wnt model [16]. In dataset GSE20916 (Figure 2B) [17], all cancer
and microdissected adenoma samples obtained through colonoscopy, as well as most (32 out of 36) surgically resected colon carcinoma samples were predicted by our model to have an active Wnt pathway (97%, n=101), while all normal colon tissue samples (n=44) were predicted Wnt inactive. The four Wnt-inactive surgical colon cancer samples may be explained by cancer tissue heterogeneity and abnormal gene promoter methylation associated with more advanced cancer, or the surgery-associated sampling procedure may have resulted in mRNA degradation and unreliable microarray results [18,8]. Such factors may interfere with the expected Wnt target gene mRNA profile and reduce sensitivity of our current model. Indeed and illustratively, when our model was applied to the transcriptome of colon cancer cell line HCT116, in which specific Wnt target genes are methylated [19], the model predicted an inactive Wnt pathway (odds 8:1, data not shown).

Above results provide evidence that the model can identify active versus inactive Wnt pathway state in tumors arising from the colon epithelial cell type used to develop and calibrate the model.

Use of the Wnt model in tumors of other tissue origin

The aim of our models is to enable wide diagnostic usage across tumors of different cellular origins. While direct target genes are transcribed by induced activity of one or more pathway-specific transcription factors binding to their respective gene response elements, indirect target genes are more likely to depend on additional cellular proteins for their transcription, increasing the likelihood of cell type-specific effects on expression regulation. For this reason, gene selection used to build the models focused on direct pathway target genes. To evaluate this premise of relative tissue-type independent functioning, we subsequently tested the models on other tumor types.

Liver cancer

In hepatocellular carcinoma and hepatoblastoma, heterozygous somatic mutations (or deletions) in the 3rd coding exon (codons 32,33,34,37,41,45) of the beta-catenin (CTNNB1) gene have been frequently identified, resulting in substitution (or loss) of an amino acid, which may be associated with aberrant activation of the Wnt pathway [20-22].
Dataset GSE9843 [23] contains 91 hepatocellular liver cancer tissue samples, of which 27 contain a beta-catenin gene mutation, 60 possess wild-type beta-catenin, and four are unknown. Furthermore, 31 of the 91 samples scored positive for nuclear beta-catenin staining, 55 negative, and five unknown. Interestingly, correlation between beta-catenin mutation status and nuclear staining is not significant in this dataset (OR=2.3, one-sided Fisher’s exact test p=0.07), illustrating the difficulty to get ground truth information on Wnt pathway activity in these samples, due to lack of a reliable test.

Chiang et al. [23] applied hierarchical clustering based on mRNA microarray data of the 91 samples of GSE9843, yielding five groups labeled ‘unannotated’, ‘polysomy chr7’, ‘inflammation’, ‘proliferation’ and ‘CTNNB1’. Despite its label, the latter group of 24 samples only contains 16 of the 27 samples with a beta-catenin mutation, and 14 of the 31 samples with a positive staining. The results of our Wnt pathway model on this dataset are shown in Figure 3A. In the CTNNB1 group, 83% (20 of 24) of the samples are predicted to have an active Wnt pathway, versus 26% (6 of 23) in the proliferation group and 0% in the other three groups. In addition, while beta-catenin mutation status and staining were not significantly correlated, we found a significant correlation of Wnt pathway activity with both beta-catenin mutation status and beta-catenin staining (OR=5.4 and 6.1, respectively, and one-sided Fisher’s exact test p=6.9e-4 and 3.4e-4, respectively; detail in supplementary Section 6).

Analysis of data from a second dataset of hepatocellular carcinoma samples (GSE6764) [24] with unknown mutational status yielded an increased incidence of Wnt activity in patients with relatively more malignant tumors, as 7 of the 27 early HCC, advanced HCC and very advanced HCC samples were predicted to have an active Wnt pathway, compared to none in the very early HCC and non-malignant sample groups (Figure 3B, one-sided Fisher’s exact test p=4.5e-4).

**Medulloblastoma**

In medulloblastoma, a subset of tumors is known to possess an activating mutation in the beta-catenin gene [25]. We applied our model in a blinded manner to the medulloblastoma dataset from Kool et al. (n=62, GSE10327) [25], and successfully identified all samples with a Wnt pathway-activating beta-catenin mutation (Figure 4). In another medulloblastoma dataset (n=40, GSE12992) [26], the Wnt model also correctly identified the four samples with a driving beta-
catenin mutation against 36 without (data not shown). These two datasets show perfect performance of the model in this tumor type with 100% sensitivity and 100% specificity.

Breast cancer

In patients with breast cancer, direct Wnt activating gene mutations are generally not present, except for a few rare metaplastic breast cancer cases [27]. Rather, activation of the Wnt pathway has been indicated by circumstantial evidence in triple negative or basal type breast cancer patients, most likely induced by interaction between the cancer cell and its microenvironment [28].

Two datasets were used for analysis of Wnt pathway activity within breast cancer subtypes: GSE12276 [29] and GSE21653 [30]; see Figures 5A and 5B. Using mRNA data, patient samples in these datasets were subtyped according to Desmedt [31] and Perou [32], respectively. Despite the difference in subtyping approach, Wnt pathway activity was very comparable, with an active Wnt pathway in 30% (21 of 65 and 21 of 75, respectively) of basal type cancer samples, versus only 7% (15 of 139 and 9 of 191, respectively) in other breast cancer subtype samples (one-sided Fisher’s exact test p=2.7e-4 and p=4.8e-7, respectively).

Taken together, the above results provide initial evidence that the Wnt model performs well on a variety of tumors without requiring additional training steps on the different cell types of origin.

ER Pathway model, initial validation

The ER pathway model was calibrated on data from eight samples of the breast cancer cell line MCF7, of which four were deprived from estrogen, and four were stimulated with 25nM E2 (GSE8597) [33], yielding an inactive and an active ER pathway, respectively. Estradiol concentrations are typically around 0.5 nM in normal breast tissue, but elevated in breast cancer to around 2 nM [34]. As a result, the 25nM model may be slightly less sensitive than desired, but still useful for a first analysis.

An initial validation of the ER pathway model was performed on MCF7 breast cancer cell lines with and without a knockdown of the gene encoding for ER (from datasets GSE10890 and GSE37820, only published at GEO). All knockdown samples (n=8) were predicted to have an
inactive ER pathway, while all other MCF7 samples (n=28) were predicted to have an active ER pathway (data not shown). In addition, in the cancer cell line encyclopedia (GSE36133) [35], in all 861 cancer cell lines other than breast cancer the ER pathway was predicted to be inactive (data not shown), indicating a very high specificity of the current model for ER pathway activity in breast cancer. Furthermore, running the pathway model on replicate experiments from datasets E-MTAB-37 [36] and GSE23593 [37] showed good reproducibility of predictions (supplementary Table S4 and Figure S1).

Next, we applied the ER pathway model to the same two patient breast cancer datasets used above: GSE12276 [29] and GSE21653 [30]; Figures 5C and 5D, respectively. The figures show an active ER pathway in 41% (38 of 102 and 61 of 138, respectively) of luminal type patients, versus only 4% (3 of 102 and 7 of 128, respectively) in other breast cancer subtype samples (one-sided Fisher’s exact test p=1.3e-10 and p=3.7e-14, respectively). The five HER2 type patients with predicted active ER pathway had scored positive for ER and/or PR by IHC staining. All basal type breast cancer samples were predicted to have an inactive ER pathway.

If patient samples are grouped according to ER IHC status (Figures 5E and 5F), we observed in datasets GSE12276 and GSE21653 an active ER pathway in 39% (27 of 88 and 67 of 150, respectively) of ER positive tumors, versus practically none (1 of 77 and 0 of 113, respectively) of the ER negative tumors.

**ER pathway activity and tamoxifen sensitivity**

To link ER pathway activity to tamoxifen sensitivity, we analyzed dataset GSE21618 [38], containing samples from tamoxifen sensitive and resistant MCF7 breast cancer cell lines (all ER positive) treated with estradiol. In particular, we took the samples that had first been deprived of estrogen, and next been stimulated for up to 48 hours. Figure 6 shows the resulting probability of ER pathway activity as a function of stimulation time. Clearly, tamoxifen sensitive cell lines quickly respond to estrogen stimulation, with probabilities steeply increasing towards 1, while tamoxifen resistant cell lines respond to a lesser extent.
Initial assessment of prognostic value of the ER pathway model

Although the pathway models have been developed and trained to assess pathway activity in order to predict therapy response, we also tested to what extent they can have prognostic value. To this end, survival time analysis was performed on a data set of 164 ER positive breast cancer patients that all received (only) adjuvant tamoxifen treatment for five years (Figure 7; combined datasets GSE6532 & GSE9195) [39,40]. The analysis was restricted to the first five years only, as tamoxifen treatment was limited to five years. As expected, patients with an active ER pathway have a better survival prognosis on tamoxifen treatment than patients for which the ER pathway is predicted inactive (one-sided logrank test p=0.034).
These results indicate that the ER pathway model may also have clinical utility in assessing prognosis in individual patients with breast cancer, even though it has not been developed for this purpose.

Discussion

With a few exceptions, e.g. HER2 and ER protein staining in breast cancer, most companion diagnostic assays to predict therapy response focus on identification of a tumor-specific genetic defect, associated with activation of a specific oncogenic signaling pathway [41,5,42]. If the test result is positive, the associated signaling pathway is assumed to be active and driving tumor growth. While a recent study by MD Anderson provides proof of principle that mutation-based identification of the tumor-driving signaling pathway improves therapy choice, it also illustrates that such a DNA-based companion diagnostic approach is unlikely to provide the complete answer as in this study therapy response increased from five to only 27 percent [42]. Indeed, it is clear that signaling pathway activation status (the functional ‘phenotype’ of the cell) is determined not only by errors in the cancer cell (epi-)genome, but to a large extent by interactions between the cancer cell and its microenvironment [2,33,43]. To assess the phenotype of the cell, in addition to the genotype, we presented a method to interpret cancer tissue transcriptome data as direct quantitative ‘output’ of active signal transduction pathway(s), using knowledge-based Bayesian models. The pathway ‘output’ is represented by transcribed pathway-specific target genes, which need to be known to create the models. So far, the data input for the models is from Affymetrix HG-U133Plus2.0 microarrays, but an important advantage of this
type of knowledge-based models is that they can be easily calibrated to other input modalities, such as other array types, RNA sequencing or dedicated multiplex PCR assays. We provided proof of principle that the models when trained on a limited number of cell line samples, or if available patient samples, already perform very well and robustly identify active oncogenic signal transduction pathway(s) in individual tissue samples obtained from a variety of malignancies.

For the Wnt pathway, our pathway model analysis results for adenoma, colon carcinoma and medulloblastoma were in full concordance with existing evidence on Wnt pathway activation in these tumor types [16,25,43].

For other tumors such as primary liver and breast cancer, no easy ‘ground truth’ with respect to activity status of the Wnt pathway is available, but this pathway is likely to play a role in at least a number of cases [1,43,44,45]. In agreement, our model identified an active Wnt pathway in the majority of liver cancer samples with, and a minority without, a beta-catenin mutation. Identification of an active Wnt pathway in the absence of a beta-catenin mutation may be due to the presence of other pathway-activating mutations in genes like APC, Axin1 and Axin2 [46], or by paracrine Wnt activation [47,48]. Indeed a significant correlation was found between model-predicted Wnt pathway activity and staining of nuclear beta-catenin, presumably the active form of the Wnt pathway transcription factor. On the other hand, the presence of a beta-catenin mutation does not necessarily mean that the pathway is activated, and in the cases in which the model did not detect Wnt activation despite a beta-catenin mutation, another pathway may have taken the lead in tumor growth, for example induced by the microenvironment.

In breast cancer, Wnt activity was detected by the model in around one third of triple negative or basal type samples, which agrees with available evidence on a role for Wnt activity in this breast cancer subtype [1]. In only very rare cases of breast cancer a potentially Wnt activating gene mutation has been found, suggesting that Wnt activity is most likely induced by paracrine interactions between cancer cells and their microenvironment. With a number of Wnt targeting drugs in the pipeline of pharmaceutical companies, development of a reliable test to identify Wnt pathway activity in this cancer subtype with highly unfavorable prognosis is considered high.
priority [49, 44, 50], as beta-catenin staining is not reliable enough to indicate Wnt pathway activity [50]. Analysis of transcriptome data by our model to is expected to provide information on Wnt activity in breast cancer, but final validation will require a clinical trial with an appropriate Wnt inhibiting drug.

Other approaches have been directed towards assessing the phenotype of cancer by analyzing its transcriptome. To deduce pathway activity from tissue transcriptome data, most pathway analysis approaches use pathway information from databases such as KEGG (www.genome.jp/kegg) and Biocarta (www.biocarta.com) that mainly list genes encoding signaling proteins. Furthermore, they invariably extrapolate quantitative mRNA levels to levels of corresponding signaling proteins, followed by a search for a role of the transcript-encoded protein in a signaling pathway, which is subsequently defined as an active pathway [43, 51, 44]. This approach is intrinsically flawed for pathway activation analysis since induction of an mRNA transcript coding for a component of a signaling pathway is not reliably correlated to the actual translated protein level, and even less to the activation status of the encoded signaling protein which requires additional post-translational protein modifications.

As expected, our approach leads to a different interpretation result of mRNA profiling data than conventional pathway analysis approaches. For example, we have run gene set enrichment analysis (GSEA) [52] using the accompanying curated canonical pathways, to identify the pathways differentially activated between the normal colon and colon adenoma samples from dataset GSE8671 [12], but none of the Wnt-related pathways was identified with a significant p-value after correction for multiple testing (see Table S5 in the supplementary information). Furthermore, another common pathway analysis approach, as presented by Skrzypczak et al. [17], in which first a list of differentially expressed genes is determined and next pathway sets are analyzed for over-representation of this gene list, did not identify the Wnt pathway as significantly different between normal colon and colon neoplasms – while our results were highly convincing as to Wnt pathway activation status.

From the results on tamoxifen sensitive and resistant cell line data it is inferred that the ER model can successfully detect ER activity in breast cancer cells. Being positive for the estrogen
receptor (ER) as measured by IHC or microarray analysis appears to be a necessary but not sufficient condition for ER pathway activity as assessed by the model. Since the model was trained on data from cell line experiments performed with a relatively high dose of estradiol (25nM), it cannot be excluded that the current model lacks to some extent in sensitivity to detect all samples with an active ER pathway. However, the finding that only a subgroup of ER positive patients seemed to have an active ER pathway agrees with the common clinical observation that a number of ER positive patients are primary resistant to hormonal therapy. Moreover, the model identified ER positive patients treated with hormonal adjuvant therapy but with an inactive ER pathway as having a worse prognosis. This is in agreement with the concept that patients with the ER pathway driving tumor growth are more likely to benefit from hormonal adjuvant therapy.

With respect to the prediction of hormonal therapy response in breast cancer patients, Symmans et al. [53] have described an mRNA profile of 165 ER-related genes, called the SET index. In contrast to the ER target genes in our model, genes underlying the SET index were not selected based on evidence for them being target genes of the ER transcription factor, but on increased expression levels found in ER positive patients. The SET index was shown to identify patients with node-negative disease which have a good prognosis when treated with adjuvant hormonal therapy, as well as patients treated with neoadjuvant chemotherapy who have a high risk at relapse when subsequently treated with adjuvant hormonal treatment. Its predictive value may be partly contributed to the incorporation of six of the ER target genes used in our model.

We hypothesize that in ER positive patients with an inactive ER pathway other pathways like Wnt or Hedgehog, associated with more aggressive behavior and worse outcome, may actually have been driving tumor growth [54,2,55]. This is in agreement with the reported decline in SET index with advancing pathologic cancer stage, despite ER positivity, suggesting decreasing tumor dependency on an active ER pathway [53]. Second, a high gene expression grade index, developed in a similar way as the SET index, also identifies subpopulations of ER positive breast cancer with unfavorable prognosis [56]. Taken together, these results strongly suggest that conventional quantification of nuclear ER immunohistochemistry staining is not sufficiently specific in detecting functional ER pathway activity. According to pathologist guidelines, ER
activity in a breast cancer sample is inferred from the presence of positive ER staining, with a minimum of 1% of ER positive tumor nuclei as a threshold level. Such staining assays have an estimated 20% error rate due to multiple factors interfering with reliability, among which non-standardized staining procedure and subjective interpretation [57,43,50]. In contrast, microarray assays are quite standardized, and the model-based analysis uses a number of ER target genes interpreted in a weighted manner to calculate a probability of pathway activation, instead of ER as single variable as is the case in IHC testing. These conditions are promising for a more robust prediction of ER pathway activity.

Clinical utility of pathway models

The fact that the Bayesian network models are knowledge- and not data mining-based has several advantages. First, the models appeared to be well applicable to data analysis of multiple unrelated tumor types and this property allows use of the models for diagnostics purpose in cancers of different cell types of origin. Nevertheless, it is expected that the models can be further optimized with respect to sensitivity and specificity by including target genes that are specifically expressed in the tissue type of origin of the tumor. Such adaptation could, for example, entail adjusting the conditional probabilities and/or adding new nodes for novel target genes. Another advantage is the relatively easy translation of the model to another data format, such as Illumina DASL microarrays, RNA sequencing or PCR-based testing. Finally the remarkable reproducibility of the pathway model results across multiple (mostly public) datasets of a specific tumor type, generated at completely different hospitals and locations, demonstrates the robustness of this modeling approach for individual patient diagnostic use. The pathway model series will be extended to include all major oncogenic signaling pathways, ultimately providing a multi-pathway analysis suite for identification of both the major active pathway, as well as potential underlying resistance pathways, applicable to tissue samples from a number of tumor types. Upon further clinical validation, currently under way, the expected main clinical utility of the described pathway models lies in therapy response prediction and monitoring in neoadjuvant and metastatic settings.
References


Figure legends

Figure 1. The structure of the Bayesian networks used to model the transcriptional program of signaling pathways.

Figure 2. Results from Wnt model analysis of microarray data from samples from patients with colon adenoma and carcinoma. For each sample a bar is plotted, indicating the odds of the Wnt pathway being active ("on") vs. inactive ("off") on a logarithmic scale. At the top of the graph the odds are 1 million to 1 that the pathway is active, at the lowest point they are 1 million to 1 that the pathway is inactive.
A. Results of Wnt model analysis on normal colon samples and colon adenoma samples (GEO dataset GSE8671) [12], calibrated on colon cancer cell lines.
B. Results of Wnt model analysis of GEO dataset GSE20916 [17] using the Wnt model calibrated on the dataset of normal colon and colon adenoma (shown in A). Analyzed samples were colon tumors (n=101) obtained by colonoscopy (dark blue), microdissected adenoma (orange), microdissected colon carcinoma (purple) and colon carcinoma obtained by surgery (dark green); corresponding control intestinal tissue samples (n=44) consist of microdissected distal normal colon tissue (yellow and light blue) and normal colon tissue obtained by colonoscopy (red) or surgery (green).

Figure 3. Results from Wnt model analysis of tissue samples from primary liver cancer.
A. Patients with hepatocellular carcinoma (n=91, dataset GSE9843) [23], labeled 'unannotated' (red), 'polysomy chr 17' (yellow), 'inflammation' (green), 'proliferation' (light blue) and 'CTNNB1' (dark blue).
B. Patients (n=69, dataset GSE6764, beta-catenin mutational status unknown) [24] with normal liver (n=10, red), cirrhotic liver tissue (n=13, yellow and green), low-grade (n=10, light blue) and high-grade (n=7, dark blue) dysplastic liver tissue, and hepatocellular carcinoma (n=8 very early HCC, orange; n=10 early HCC, purple; n=7 advanced HCC, dark green; n=10 very advanced HCC, pink).
The y-axis shows the odds of the Wnt pathway being active ('on') vs. inactive ('off'), on a logarithmic scale. Each tissue sample result is represented by a bar.

Figure 4. Results from Wnt pathway model analysis of samples from patients with medulloblastoma (n=62, GSE10327) [25], ordered as: samples expressing retinal differentiation genes, either high (red) or low (yellow), samples with a mutation in SHH (light blue) or CTNBB1 (dark blue), and rest (green).
The y-axis shows the odds of the Wnt pathway being active ('on') vs. inactive ('off'), on a logarithmic scale. Each tissue sample result is represented by a bar.

Figure 5. Results from Wnt (A,B) and ER (C-F) pathway model analysis on mRNA microarray data from samples of patients with breast cancer. The y-axis shows the odds of the respective pathway being active ('on') vs. inactive ('off'), on a logarithmic scale. Each tissue sample result
is represented by a bar.

A,C. Odds of Wnt (A) and ER (C) pathway activity in a dataset with breast cancer samples (n=204, GSE12276) [29] subtyped according to the module approach from Desmedt et al. [31] as luminal A (green), luminal B (dark blue), HER2 (orange) and basal (red). All patients within this dataset suffered a relapse (median time to recurrence: 21 months, range: 0 – 115 months). The ordering of samples within each group is different in the two graphs; only five of the 204 samples have both an active Wnt and ER pathway.

B,D. Odds of Wnt (B) and ER (D) pathway activity in a dataset with breast cancer samples (n=266, GSE21653) [30] subtyped according to Perou’s subtyping scheme [32], as luminal A (green), luminal B (dark blue), HER2 (orange), basal (red) and normal-like (light blue). The ordering of samples within each group is different in the two graphs; only four of the 266 samples have both an active Wnt and ER pathway.

E,F. Results from the ER pathway model analysis on mRNA microarray data from samples of patients with breast cancer from datasets GSE12276 (E, n=204) [29] and GSE21653 (F, n=266) [30]. Samples are grouped here according to ER IHC status, as ER negative (red), ER positive (blue) or unknown (orange).

Figure 6. Prediction of ER pathway activity by the ER model in tamoxifen sensitive and tamoxifen resistant MCF7 cell lines over time. The cell lines have first been deprived of estrogen, and next stimulated for up to 48 hours with estrogen. The vertical axis shows the predicted probability that the ER pathway is active. Microarray measurements were taken from dataset GSE21618 [38]. Individual samples are plotted as points; the drawn lines are trend lines.

Figure 7. Kaplan-Meier curves showing prognostic value of ER pathway activity in patients with ER positive breast cancer from datasets GSE6532 and GSE 9195, all treated with adjuvant tamoxifen for five years (n=164) [39,40]. The grey line represents patients in which the model predicted an active ER pathway (odds > 1:1); the black line represents cases in which the model predicted an inactive ER pathway (odds ≤ 1:1).
Figure 1
1.5 column
Black-and-white

(a) Transcription complex
(b) Target genes
(c) Probesets

TC

TG1

TG2

TGn

PS1a
PS1b
PS1c
PS2a
PSna
PSnb
Figure 2
1.5 column

Color

A

B

Odds Wnt on/off

10^6 : 1

10^3 : 1

1 : 1

1 : 10^3

1 : 10^6

normal (microdiss.)
adenocarcinoma (col.)

adenoma (microdiss.)
carcinoma (microdiss.)
dist. normal (microdiss.)
carcinoma (surgery)

1 : 10^6

1 : 10^3

1 : 1

1 : 10^3

1 : 10^6

normal (microdiss.)
adenocarcinoma (col.)

adenoma (microdiss.)
carcinoma (microdiss.)
dist. normal (microdiss.)
carcinoma (surgery)
Figure 3
1.5 column
Color
Figure 4
1.5 column
Color

- nrl_high
- shh
- nrl_low
- wnt
- rest

Odds Wnt on/off

10^6 : 1
10^3 : 1
1 : 1
1 : 10^3
1 : 10^6
1 : 10^6
Figure 6
1 column
Black-and-white

![Graph showing the relationship between P(ER active) and hours stimulated with E2 for Tamoxifen resistant and sensitive cells.](image-url)
Figure 7
1 column
Black-and-white
Selection of personalized patient therapy through the use of knowledge-based computational models that identify tumor-driving signal transduction pathways

Wim Verhaegh, Henk van Ooijen, Marcia A. Inda, et al.

Cancer Res  Published OnlineFirst April 2, 2014.