Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression

Highlights
- Identification of the most relevant Smad complexes in liver-derived cells
- Assessment of the contribution of the Smad complexes on target gene expression
- Link between Smad protein abundance, complex formation, and gene expression
- Increased Smad abundance and Smad2 phosphorylation in hepatocellular carcinoma

In Brief
Transforming growth factor β (TGF-β) leads to the phosphorylation of Smad proteins and thereby facilitates the formation of different trimeric Smad complexes. By combining quantitative mass spectrometry with mathematical modeling, the identities of the formed trimeric Smad complexes are resolved and the link of these transcription factors with target gene expression is established. This approach allows predicting based on gene expression data that in hepatocellular carcinoma the abundance of Smad proteins and their phosphorylation is elevated, which was experimentally validated.

Authors
Philippe Lucarelli, Marcel Schilling, Clemens Kreutz, ..., Wolf D. Lehmann, Jens Timmer, Ursula Klingmüller

Correspondence
u.klingmueller@dkfz.de

Lucarelli et al., 2018, Cell Systems 6, 75–89
January 24, 2018 © 2017 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2017.11.010
Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression

Philippe Lucarelli,1,13 Marcel Schilling,1,13 Clemens Kreutz,2,13 Artyom Vlasov,1 Martin E. Boehm,1,4 Nao Iwamoto,1 Bernhard Steiert,2,13 Susen Lattermann,1 Marvin Wäscher,1 Markus Stepnath,1 Matthias S. Matter,1,4 Mathias Heikenwalder,6 Katrin Hoffmann,7 Daniela Deharde,8 Georg Damm,8,9 Daniel Seehofer,8,9 Maria Muciek,10 Norbert Gretz,10 Wolf D. Lehmann,2 Jens Timmer,2,11 and Ursula Klingmüller1,12,14, *

1Division Systems Biology of Signal Transduction, German Cancer Research Center (DKFZ), INF 280, 69120 Heidelberg, Germany
2Institute of Physics, University of Freiburg, 79104 Freiburg, Germany
3Center for Biological Systems Analysis, University of Freiburg, 79104 Freiburg, Germany
4Molecular Structure Analysis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
5Institute of Pathology, University of Basel, 4003 Basel, Switzerland
6Division Chronic Inflammation and Cancer, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
7Department of General and Transplantation Surgery, Ruprecht Karls University Heidelberg, 69120 Heidelberg, Germany
8Department of General-, Visceral- and Transplantation Surgery, Charité University Medicine Berlin, 13353 Berlin, Germany
9Department of Hepatobiliary Surgery and Visceral Transplantation, University of Leipzig, 04103 Leipzig, Germany
10Medical Research Center, Medical Faculty Mannheim, University of Heidelberg, 68167 Mannheim, Germany
11BIOSS Centre for Biological Signalling Studies, University of Freiburg, 79104 Freiburg, Germany
12Translational Lung Research Center (TLRC), Member of the German Center for Lung Research (DZL), 69120 Heidelberg, Germany
13These authors contributed equally
14Lead Contact

*Correspondence: u.klingmueller@dkfz.de

https://doi.org/10.1016/j.cels.2017.11.010

SUMMARY

Upon stimulation of cells with transforming growth factor β (TGF-β), Smad proteins form trimeric complexes and activate a broad spectrum of target genes. It remains unresolved which of the possible Smad complexes are formed in cellular contexts and how these contribute to gene expression. By combining quantitative mass spectrometry with a computational selection strategy, we predict and provide experimental evidence for the three most relevant Smad complexes in the mouse hepatoma cell line Hepa1-6. Utilizing dynamic pathway modeling, we specify the contribution of each Smad complex to the expression of representative Smad target genes, and show that these contributions are conserved in human hepatoma cell lines and primary hepatocytes. We predict, based on gene expression data of patient samples, increased amounts of Smad2/3/4 proteins and Smad2 phosphorylation as hallmarks of hepatocellular carcinoma and experimentally verify this prediction. Our findings demonstrate that modeling approaches can disentangle the complexity of transcription factor complex formation and its impact on gene expression.

INTRODUCTION

Transforming growth factor β (TGF-β) is a pleiotropic factor with multiple functions for which the underlying mechanisms are only partially understood. The most prominent intracellular mediators of TGF-β signaling are the Smad proteins. The receptor Smad proteins, Smad2 and Smad3, interact with the common Smad (Smad4), form trimeric complexes, and translocate to the nucleus, where they interact with other proteins including other transcription factors or crosstalk with other signaling pathways (Feng and Derynck, 2005). In addition, the amount, composition, or dynamics of Smad complex formation could differ. Therefore, systematic studies are required to decipher the contribution of individual Smad complexes to gene expression.

Numerous intragenic mutations and homozygous deletions of Smad2 and Smad4, as well as downregulation of Smad3 mRNA, were reported for different forms of carcinoma (Levy and Hill, 2006). While attempts were made to quantify Smad expression...
levels (Dzieran et al., 2013), potential alterations in the abundance of Smad proteins in liver cancer tissue and hepatoma cell lines compared with primary hepatocytes have not been addressed. So far the impact of TGF-β stimulation was only examined in proteome-wide studies that analyzed changes in proteins and phosphorylation sites (Ali and Molloy, 2011; D’Souza et al., 2014). However, comprehensive information on the abundance and the degree of phosphorylation of Smad proteins is currently not available.

TGF-β-induced signal transduction has been approached by mathematical modeling that addressed the dynamics of ligand-receptor interaction, identified the role of negative feedbacks, and provided insights into nuclear-cytoplasmic shuttling (Schmi erer et al., 2008; Zi et al., 2011). However, most of these mathematical models were primarily based on literature knowledge, and only few of these studies included experimental data. None of the previously published mathematical models accounted for the composition of the trimeric Smad complexes and the specific link to Smad complex-mediated gene expression.

Combining mass spectrometric data and mathematical modeling as utilized for the analysis of mechanisms governing dimerization of phosphorylated Stat5 (Boehm et al., 2014), could provide valuable information on Smad complex formation. In such an approach, proteomics provides data on protein abundance, while mathematical modeling provides a tool for an unbiased selection strategy for the identification of transcription factor complexes that are present in a given cell type. Since the number of candidate models grows exponentially with the number of model parameters, finding the exact solution of such a model selection task is very challenging. We developed a method to distinguish non-essential from essential model parameters by combining nonlinear mathematical modeling with L1 regularization (Merkle et al., 2016; Steiert et al., 2016). The L1 regularization approach can be employed to investigate which model reactions are required to describe experimental data. Therefore the L1 regularization approach could be utilized to statistically assess which individual reaction parameters leading to complex formation are necessary and sufficient and thereby identify essential protein complexes.

Here we combine quantitative experimental techniques with mathematical modeling approaches to resolve complexity in Smad complex formation and to establish a quantitative link to transcriptional activities in hepatoma cell lines and primary hepatocytes.

RESULTS

Abundance and Interactions of Smad Proteins

To examine the TGF-β-induced formation of Smad complexes, we quantified the amount of Smad proteins in unstimulated Hepa1-6 cells using antibodies that specifically recognize Smad2, Smad3, or Smad4 as well as an antibody with equal affinity to Smad2 and Smad3 (Figure S1A). Immunoprecipitation (IP) and quantitative immunoblotting (IB) experiments in combination with recombinant proteins (Figure S1B) revealed 825,000 ± 74,000 Smad2 and 402,000 ± 113,000 Smad4 molecules per cell in Hepa1-6 cells (Figure 1A). To define the relative abundance of Smad2 and Smad3, we combined isoform-independent Smad2/3 IP with quantitative mass spectrometry (Boehm et al., 2014). The results revealed a ratio of approximately 10:1 between Smad2 and Smad3 that was unaffected by TGF-β treatment, with a total amount of 83,000 ± 6,600 Smad3 molecules per Hepa1-6 cell (Figure 1B).

To analyze the dynamics of TGF-β-induced phosphorylation of Smad2 and Smad3 in Hepa1-6 cells, we stimulated cells with 1 ng/mL TGF-β for up to 10 hr and performed IP experiments followed by mass spectrometry. These measurements enabled us to distinguish non-phosphorylated Smad2 and Smad3 (nSmad2 and nSmad3), Smad2 and Smad3 phosphorylated at the most C-terminal serine residue (pSmad2 at Ser467; pSmad3 at Ser425), and Smad2 and Smad3 phosphorylated at the two most C-terminal serine residues (ppSmad2 at Ser465 and Ser467; ppSmad3 at Ser423 and Ser425). At t = 0 min almost all Smad2 and Smad3 molecules were non-phosphorylated (Figure 1C). Upon TGF-β stimulation, the amount of nSmad2 decreased until 60 min and increased at later time points. 60 min after TGF-β stimulation, 5% of total Smad2 was present as pSmad2. The abundance of ppSmad2 rapidly increased, with a peak at 60 min after TGF-β stimulation at which the amount corresponded to 60% of total Smad2, followed by a decrease to basal levels after 10 hr. Similar dynamics were observed for nSmad3, pSmad3, and ppSmad3.

To determine to which extent Smad2 and Smad3 engage in complex formation, Hepa1-6 cells were stimulated with 1 ng/mL TGF-β for 60 min. The total amount of Smad2 and Smad3, their phosphorylation status, and the amount of the co-immunoprecipitated Smad2, Smad3, and Smad4 were quantified by mass spectrometry (Figure 1D). Only doubly phosphorylated Smad3 co-immunoprecipitated with Smad2 (Figure 1D, left panel) and the amount of co-immunoprecipitated Smad3 (28,000 ppSmad3 molecules/cell) was low compared with the amount of Smad2 (825,000 molecules/cell). Only ppSmad2 was co-immunoprecipitated with Smad3 (Figure 1D, right panel). Out of 550,000 ppSmad2 molecules per Hepa1-6 cell, only about 4% formed a complex with Smad3. In Smad2 and Smad3 IPs, we could detect co-immunoprecipitation (colP) of Smad4. The Smad4 amount detected after Smad2 IP was approximately 35-fold higher than after Smad3 IP (69,000 and 2,000 molecules/cell, respectively), suggesting a higher abundance of formed Smad2: Smad4 complexes compared with Smad3:Smad4 interactions. These results revealed that only doubly phosphorylated Smad2 and Smad3 molecules formed complexes. Despite the high degree of Smad2 and Smad3 phosphorylation, only few Smad2 molecules interacted with Smad3. On the other hand, around 30% of all Smad3 molecules are associated with Smad2 upon stimulation with TGF-β. We also examined the Smad complex formation in a time- and TGF-β dose-dependent manner in Hepa1-6 cells (Figures S1C and S1D). We concluded that the efficacy of Smad complex formation may be highly dependent on the molecular ratio of the individual Smad proteins.

Identification of the Most Relevant Smad Complexes

Theoretically, a large number of trimeric Smad2, Smad3, and Smad4 complexes with different composition could be formed. Trimeric complexes (k = 3) with nSmad2, pSmad2, ppSmad2,
nSmad3, pSmad3, ppSmad3, and Smad4 (n = 7) would result in 84 possible complexes based on Equation 1. Since only ppSmad2, ppSmad3, and Smad4 substantially engage in complex formation (n = 3), this number is reduced to 10. As we observed major differences in the total amount of Smad2 and Smad3 proteins, it is possible that only a much smaller number of Smad complexes occurs in Hepa1-6 cells.

To disentangle the combinatorial complexity of Smad complex formation, we combined quantitative experiments with mathematical modeling. The established mathematical model consists of 31 mass-action kinetic reactions and 16 dynamical parameters that describe TGF-β receptor activation and Smad complex formation (Figure 2A). In the model, the formation of each of the ten possible trimeric Smad complexes consisting of ppSmad2, ppSmad3, and Smad4 was characterized by a complex-specific association rate (kon). The dissociation of each trimeric Smad complex was dependent on the dephosphorylation of ppSmad2 and ppSmad3 in the heterotrimeric Smad complexes or on the dissociation of the homotrimeric Smad4 complex (Figure 2A). The mathematical model with ten complexes was capable of describing the time and dose dependency of Smad complex formation (Figure S1E).

To identify the most relevant Smad complexes in Hepa1-6 cells, we employed a data-based model selection approach (Merkle et al., 2016; Steiert et al., 2016) to eliminate complexes not required to explain the experimental data. For this purpose, we added an L1 regularization term to the kon parameters favoring a minimal number of distinct complexes in the mathematical model. For statistical assessment, we calculated the profile likelihoods for the ten kon parameters of the considered trimeric Smad complexes (Raue et al., 2009), indicating the parameter ranges that are compatible with the experimental data. For seven of the ten kon parameters (Figure 2B), the best parameter estimation value (red asterisk) was compatible with \( 10^{-14} \), which is equivalent to zero. For the other three kon parameters, the best parameter estimation value was significantly different from zero. These results suggested that only three complexes ppSmad2:ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4, and ppSmad2:ppSmad3:Smad4 are necessary to describe the experimental data. The reduced mathematical model comprising only these three Smad complexes (Figure 2C) was
able to describe the experimental data (Figure 2D) to a similar extent as the comprehensive mathematical model considering all ten possible trimeric Smad complexes. The goodness-of-fit of the reduced model was assessed by the chi-square statistics \( \chi^2 = \sum (y_i - f_i)^2 / \sigma_i^2 \), which increases by 0.71 after removing seven complexes from 743.56 to 744.28 for 342 data points, supporting the model reduction. Thus, the model proposes ppSmad2:ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4, and ppSmad2:ppSmad3:Smad4 as the most relevant TGF-β-induced Smad complexes in Hepa1-6 cells.
Experimental Validation of the Model-Predicted Smad Complexes

To experimentally validate the model-predicted Smad complexes, we combined sequential IP experiments using lysates of Hepa1-6 cells with detection by quantitative IB. To confirm the presence of the model-predicted ppSmad2:ppSmad3:ppSmad3 complex, we examined a TGF-β-dependent, but Smad4-independent, interaction of Smad2 with Smad3. We depleted Smad4 by three repetitive IP experiments from lysates of Hepa1-6 cells that had been treated with TGF-β or were left unstimulated. Depletion of Smad4 from the lysate was confirmed by IB, showing that the Smad4 signal was reduced to background. The Smad4-depleted lysates were exposed to Smad3 by IB, showing the Smad3 signal was reduced to background. The Smad4-depleted lysates were exposed to Smad3 IP, and co-precipitated Smad2 was detected by quantitative IB. Even in the absence of Smad4, Smad2 associates with IP, and co-precipitated Smad2 was detected by quantitative ground. The Smad4-depleted lysates were exposed to Smad3 by IB, showing that the Smad4 signal was reduced to background. Depletion of Smad4 from the lysate was confirmed by IB, showing that the Smad4 signal was reduced to background. The Smad4-depleted lysates were exposed to Smad3 IP, and co-precipitated Smad2 was detected by quantitative IB. Even in the absence of Smad4, Smad2 associates with Smad3, but only in lysates of TGF-β-stimulated cells in which most of Smad2 and Smad3 are doubly phosphorylated (Figure 3A), thereby supporting the model-predicted ppSmad2:ppSmad3:ppSmad3 complex. Analogous experiments were performed to validate the ppSmad2:Smad4:Smad4 complex. Smad3 was depleted from lysates of TGF-β-stimulated or unstimulated Hepa1-6 cells, again reducing the Smad3 signal to background levels. The Smad3-depleted lysates were subjected to Smad4 IP, and the analysis of Smad2 by IB showed the TGF-β-dependent interaction of Smad2 with Smad4 (Figure 3B), providing experimental evidence for the ppSmad2:ppSmad4:Smad4 complex.

To verify the ppSmad2:ppSmad3:Smad4 complex, we immunoprecipitated Smad3 from lysates of Hepa1-6 cells, which were stimulated with TGF-β or were left untreated. Immunoprecipitated proteins bound to the beads were dissociated by the addition of an excess of the Smad3 blocking peptide. The resulting supernatants were used for Smad3 IP, and the detection of Smad2 by IB confirmed the colP of Smad2 (Figure 3C). Since only doubly phosphorylated Smad2 and Smad3 engage in complex formation (Figure 1D), these results verify the TGF-β-dependent formation of a trimeric complex between ppSmad2, ppSmad3, and Smad4.

Model-Based Link of Smad Complexes with Gene Expression Dynamics

The impact of individual trimeric Smad transcription factor complexes on the expression dynamics of specific genes is unknown. To establish time-resolved expression profiles of TGF-β genes, RNA was extracted from TGF-β-treated and untreated Hepa1-6 cells over time and subjected to microarray analysis. To define transcripts with similar gene expression kinetics, k-means clustering was performed. We identified 12 clusters with distinct dynamic patterns of gene expression, each containing 20–50 genes (Figure 3D). The transcripts of clusters 1, 2, and 3 showed transient expression kinetics with maximally increased gene expression at around 3 hr. Clusters 4, 5, and 8 displayed sustained dynamics, and clusters 6 and 7 were characterized by a transient kinetics with a peak at 60 min. In contrast to these positively regulated clusters, gene expression in clusters 9 to 12 was downregulated by TGF-β.

In each cluster, we selected a representative gene that was previously linked to TGF-β-induced Smad2/3 signaling. We selected Ski (Luo et al., 1999) for cluster 1, Skil (SnoN) (Stroschein et al., 1999) for cluster 2, Dnmt3a (Domingo-Gonzalez et al., 2015) for cluster 3, Sox4 (Qin et al., 2009) for cluster 4, Jun (Koinuma et al., 2009) for cluster 5, Smad7 (Lebrun et al., 1999) for cluster 6, Klf10/Tieg1 (Dosen-Dahl et al., 2008) for cluster 7, Bmp4 (Greber et al., 2007) for cluster 8, Cxcl5 with its human ortholog CXCL8/IL8 (Ge et al., 2010) for cluster 9, Dusp5 (Tao et al., 2016) for cluster 10, Tgfα (Nozato et al., 2003) for cluster 11, and Pak4 (Stockert et al., 2011) for cluster 12. To verify that these genes are bona fide TGF-β target genes, we analyzed the expression of these genes in Hepa1-6 cells upon stimulation with 1 ng/mL TGF-β in the presence or absence of the selective TGF-β receptor inhibitor SB-431542. SB-431542 selectively inhibits ALK4, ALK5, and ALK7, and thereby impairs canonical Smad-mediated TGF-β signaling, whereas non-canonical TGF-β signaling is not affected (Inman et al., 2002). Our results demonstrated that SB-431542 reduces the TGF-β-induced upregulation of the genes of clusters 1 to 8 as well as the TGF-β-mediated downregulation of the genes of clusters 9 to 12 (Figure S2A).

To quantitatively link the dynamics of TGF-β-induced formation of Smad complexes to gene expression, we established an integrative mathematical model that extends our reduced mathematical model to downstream transcriptional regulation. Since the specific connection between the considered Smad complexes and target gene expression was not known, we constrained the potential regulatory mechanisms in the integrative mathematical model as little as possible to allow for positive and negative regulation of each complex on every target gene, as well as for a gene-specific turnover (Figure 3E).

Figure 3. Experimental Evidence for the Predicted Smad Complexes and Identification of Distinct Clusters of TGF-β-Regulated Genes

The (A) ppSmad2:ppSmad3:ppSmad3, (B) ppSmad2:Smad4:Smad4, and (C) ppSmad2:ppSmad3:Smad4 complexes were examined in Hepa1-6 cells stimulated with 1 ng/mL TGF-β for 60 min, or unstimulated, lysed, and subjected to IP and IB. Lysates were depleted of Smad4 by three sequential IPs and used for Smad3 IP, which was analyzed first by a Smad2 IB and second by a Smad3 IB. Experiments were performed in biological triplicates and means and SD are shown. Lysates were depleted of Smad3 by three sequential IPs and used for Smad4 IP, which was analyzed first by a Smad2 IB and second by a Smad4 IB. Experiments were performed in biological triplicates and means and SD are shown. Lysates were used for Smad3 IP and proteins were dissociated from beads. The supernatants were subjected to Smad4 IP and the associated Smad2 was detected by IB. IPs were confirmed by IB. Experiments were performed in biological triplicates and means and SD are shown. Microarray-based gene expression analysis of Hepa1-6 cells treated (red) or untreated (gray) with 1 ng/mL TGF-β for up to 10 hr was performed (n = 2) and divided into 12 groups by k-means clustering. Continuous lines, estimated dynamics; dashed lines, cluster average. Representative TGF-β target genes in each cluster were linked to the reduced pathway model, establishing an integrative mathematical model. Each complex could have an activating (green) or an inhibitory (red) effect on each target gene with a gene-specific turnover rate. Gene activation dynamics of the 12 target genes were validated by qRT-PCR. Dots, experimental data (n = 3); continuous lines, model simulations; shaded area, estimated error.
We determined by qRT-PCR the TGF-β-induced expression of the 12 clusters in Hepa1-6 cells (Figure 3F). The integrative mathematical model was able to describe the expression dynamics of the 12 representative TGF-β target genes and captured up- and downregulated as well as transient and sustained gene expression. The model feature that facilitated the description of transient or sustained gene expression was a gene-specific mRNA turnover parameter. We calculated the gene-specific half-life based on this parameter (Figure S2B). The model predicted a fast turnover for Jun, Smad7, Klf10, and Bmp4 (half-life < 10 min), an intermediate turnover for Skil, Dusp5, and Pdk4 (half-life between 10 and 100 min), and a slow turnover for the five remaining genes (half-life >100 min). These model-predicted values showed good agreement (Figure S2D) with the mRNA half-life experimentally determined by the addition of Actinomycin D (Figure S2C).

These results indicate that the knowledge on the TGF-β-induced dynamics of the ppSmad2:ppSmad3:ppSmad3, ppSmad2:ppSmad3:ppSmad4, and ppSmad2:ppSmad3:ppSmad3:Smad4 complexes in Hepa1-6 cells is sufficient to link complex formation to gene expression.

Modulation of TGF-β-Induced Gene Expression by Changing the Abundance of the Smad Molecules

The analysis of transcription factor binding sites in the promoter region of the selected Smad target genes revealed that most of the target genes contained experimentally validated (Figure 4A, dashed lines) or expert-curated (Figure 4A, solid lines) transcription factor binding sites for at least one of the considered Smad proteins. For Dusp5 and Pdk4, no transcription factor binding site for Smad2, Smad3, or Smad4 was reported, indicating indirect regulation.

To elucidate in Hepa1-6 cells the impact of the identified trimeric Smad complexes on the dynamics of TGF-β-induced gene expression, we perturbed the system with target-specific small interfering RNA (siRNA) against Smad3 and Smad4 and measured the knockdown efficiency for Smad3 and Smad4 by quantitative IB (Figures 4B and S3A). Possibly due to the high expression level of Smad2 in Hepa1-6 cells, only a marginal knockdown effect could be achieved for Smad2. To assess in silico the potential impact of the knockdown of Smad proteins on the TGF-β-induced formation of the three Smad complexes, we adjusted in our mathematical model the initial amounts of Smad3 and Smad4 according to the experimentally measured knockdown efficiency (Figures 4B and S3A) and performed model simulations to predict the time courses of the formation of the three Smad complexes (Figures 4C and S3B). The model predictions suggested that knockdown of Smad3 had a strong negative impact on the formation of the ppSmad2:ppSmad3:ppSmad3 and of the ppSmad2:ppSmad3:Smad4 complex in a dose-dependent manner of the siRNA, but only a minor effect on the formation of the ppSmad2:ppSmad4:Smad4 complex. A similar effect was observed in response to the knockdown of Smad4, which resulted in a negative effect on the ppSmad2:ppSmad3:Smad4 and the ppSmad2:ppSmad4:Smad4 complexes, and only minor effects were observed on the formation of the ppSmad2:ppSmad3:ppSmad3 complex.

The experimental results shown in Figure 4D revealed that Smad3 knockdown altered the expression of all target genes. Positively regulated Smad target genes were suppressed (e.g., Ski, Skil, Klf10, and Sox4), whereas the expression of target genes that are repressed by TGF-β stimulation were up-regulated upon Smad3 knockdown (e.g., Cxc15, Dusp5, and Bmp4) (Figure 4D). Comparable effects were observed upon Smad4 knockdown (Figure S3C), confirming the dependency of the expression of the 12 selected target genes on Smad signaling.

To examine the specific impact on the identified trimeric Smad complexes, we established Hepa1-6 cells overexpressing Flag-tagged Smad2, Smad3, or Smad4, and measured the total amounts of Smad proteins by quantitative IB (Figure 4E). A 2-fold increase of Smad2, a 14-fold increase of Smad3, and a 3-fold increase in the total amount of Smad4 compared with wild-type Hepa1-6 cells were obtained. We observed that overexpression of one of the Smad proteins had no major impact on the expression levels of the other two Smad proteins.

We performed model simulations by adjusting our mathematical model to the measured overexpression levels to predict the time courses of the formation of the three Smad complexes (Figure 4F). The model predictions suggested that overexpression of Smad3 affects the formation of the ppSmad2:ppSmad3:ppSmad3 and of the ppSmad2:ppSmad3:Smad4 complexes, while it negatively affects the dynamics of the ppSmad2:Smad4:Smad4 complex. The model predicted a positive influence of Smad4 overexpression on the formation of the ppSmad2:Smad4:Smad4 complex and, to a lesser extent, on the dynamics of the ppSmad2:ppSmad3:Smad4 complex. On the contrary, the model predicted that Smad2 overexpression had little impact on Smad complex formation. These insights suggest that alterations in the total amount of Smad proteins, in particular of
Smad4 and Smad3, change the extent to which the three relevant complexes are formed.

The time course experiments (Figure 4G, data points) showed that Smad2 overexpression positively influenced the TGF-β-mediated induction of Ski, Skil, Dnmt3a, and Sox4. Smad3 overexpression increased the expression of Klf10 and Dnmt3a at earlier time points, whereas the expression of Sox4 and Jun was affected at later time points and the expression of Ski, Skil, Smad7, and Bmp4 was altered during the entire observation period. Conversely, the TGF-β-induced downregulation of Cxcl15, Dusp5, Tgfa, and Pdk4 expression was augmented by Smad3 overexpression. Smad4 overexpression resulted in an increased upregulation of Klf10 expression and an augmented repression of Cxcl15 and Dusp5. Our mathematical model adjusted to the observed overexpression levels of the Smad2, Smad3, and Smad4 was able to quantitatively describe the majority of the observed gene expression profiles (Figure 4G, solid lines). For Dnmt3a, Jun, Klf10, and Cxcl15, the model trajectories were not able to describe the experimental data, which might be due to the absence of interactions between the identified Smad complexes or with co-factors in our mathematical model.

In sum, the mathematical model was capable of correctly quantifying the connection between Smad2, Smad3, and Smad4 levels, the formation of Smad complexes, and the expression dynamics of the majority of the representative TGF-β target genes. Since upon overexpression or knockdown of Smad proteins the dynamics of the majority of genes remained in the same cluster (Figure S3D), we concluded that, while the TGF-β-induced expression dynamics is a property of each gene, the dynamics can be modulated by a change in the abundance of Smad pathway components.

Figure 5. Mathematical Model-Based Determination of the Impact of Single Smad Complexes on Gene Expression
Prediction of the influence of the single Smad complexes on TGF-β-induced gene expression. Left panel: regulation relative to gene expression at time point 0 min. Green, positive regulation; red, negative regulation; black, marginal impact. Uncertainties of the predictions are visualized by the thickness of the lines. Right panel: normalized area under the curve of the trajectories (log2), with median and SD. The color code of both panels was normalized for each gene individually.

Model-Based Analysis of the Contribution of the Individual Smad Complexes to TGF-β-Induced Gene Expression
To quantify the influence of each Smad complex on TGF-β-induced gene expression, we used our mathematical model to predict the expression profiles of the TGF-β-induced genes in the presence of only one of the three Smad complexes (Figure 5, left panel). We utilized the area under the curve of the model-predicted gene expression profiles as a quantitative measure for the extent of gene activation represented by a heatmap (Figure 5, right panel). In our mathematical model, the relative influence of each Smad complex on the extent of gene expression is determined by the activation and inhibition parameters multiplied by the complex concentrations and further modulated by the turnover rates (Figure S4A). To analyze which activation and inhibition parameter contributed most to the expression dynamics of the respective gene, we performed a sensitivity analysis (Figure S4B). These model-based studies suggested that the ppSmad2:ppSmad3:ppSmad3 complex primarily has a positive influence on the expression of Jun, and a weakly positive impact on Smad7 which is due to a negligible inhibition parameter. The activation parameters of ppSmad2:ppSmad3:ppSmad3 for Ski, Dnmt3a, Sox4, Klf10, Cxcl15, Dusp5, Tgfa, and Pdk4 are estimated as zero. On the contrary, the model predicted that the ppSmad2:Smad4:Smad4 complex induces the expression of Ski, Skil, Sox4, Jun, and Smad7, and represses Cxcl15, Dusp5, Tgfa, and Pdk4. For Smad7, the
inhibitory parameter of the ppSmad2:Smad4:Smad4 complex
was in agreement with zero. Our mathematical model indicated
that the ppSmad2:ppSmad3:Smad4 complex has the strongest
activation parameter for 7 out of the 12 genes and also the
largest inhibitory parameter for Dusp5 and Pdk4.

In summary, the integrative mathematical model enabled
us to dissect the individual contribution of the three
Smad complexes to TGF-β-induced target gene expression.
While expression of most target genes was positively
influenced by the ppSmad2:ppSmad3:Smad4 complex, Dusp5 and Pdk4 were transcriptionally repressed by this
complex. Possibly the ppSmad2:ppSmad3:Smad4 complex is an
activating transcription factor that exerts its downregulating
function by inducing a transcriptional repressor.

**Protein Abundance-Dependent Model Predictions and
Validation in Different Liver Cell Types**

To evaluate Smad complex formation in other liver cells
that potentially harbor different amounts of Smad proteins, we
examined primary mouse hepatocytes, the human hepatoma cell line
HepG2, and primary human hepatocytes.

The abundance of Smad2, Smad3, and Smad4 in these cells
was determined by quantitative IB and quantitative mass spectrometry (Figures S5A–S5D). Compared with Hepa1-6 cells (Figures 1A and 1B), the abundance of Smad2, Smad3, and Smad4 (Table S1) was substantially different in primary mouse hepatocytes, HepG2 cells, and primary human hepatocytes (Figure 6A).

Compared with Hepa1-6 cells (Figures 1A and 1B), in primary mouse hepatocytes the expression of Smad2 was 8-fold lower,
Smad3 was 3-fold lower, and Smad4 was 5-fold lower. Overall the abundance of Smad proteins in HepG2 cells was similar to that in primary mouse hepatocytes, but showing a similar approximately 10:1 ratio between Smad2 and Smad3 proteins as in Hep1-6 cells. Primary human hepatocytes harbored comparable amounts of Smad2, Smad3, and Smad4, and the Smad2 concentration was lower by one order of magnitude compared with HepG2 cells.

To assess the impact of these differences in abundance of Smad proteins on Smad complex formation, we utilized the experimentally established cell-type-specific concentration of Smad2, Smad3, and Smad4 as starting values in our mathematical model, and predicted the dynamics of the TGF-β-induced formation of the three Smad complexes for each of the studied cell types. The model predicted that changes in the total amounts of Smad2, Smad3, or Smad4 altered the dynamics of the formation of ppSmad2:ppSmad3:ppSmad4 complexes (Figures S5E–S5G). To experimentally test if these differences propagate to the expression of Smad target genes, we treated primary mouse hepatocytes, HepG2 cells, and primary human hepatocytes with 1 ng/mL TGF-β for up to 10 hr and determined the dynamics of the expression of the representative Smad target genes by qRT-PCR analysis (Figure 6B, filled circles, and Figure S6, open circles). To analyze the obtained data with our mathematical model, we adjusted the model to the cell type-specific abundance of Smad proteins. Since we observed differences in the dynamics of gene activation in the different cell types, the model parameters of TGF-β receptor activation and Smad complex formation were newly estimated, while the model parameters linking the respective Smad complexes to gene expression were retained. We omitted the mouse-specific gene Cxcl15 from this analysis. The resulting model simulations correctly described for each cell type the specific dynamics of the TGF-β-induced expression of SKI, DNMT3A, SOX4, SMAD7, KLF10, BMP4, and TGFA (Figure 6B, continuous lines), DUSP5 was not detectable in primary human hepatocytes and HepG2 cells, and PDK4 was not detectable in HepG2 cells (Figure S6). While a transcriptional response of SKIL and JUN to TGF-β was observed, the cell type-specific expression dynamics was not in line with the predicted model trajectories (Figure S6, continuous lines). The qualitative differences between the model trajectories and the experimental data for these four Smad target genes might be explained by cell-type-specific epigenetic modifications or co-factors that are currently not considered in our mathematical model.

Our result showed that adjusting our mathematical model to the measured cell-type-specific abundance of Smad2, Smad3, and Smad4 was sufficient to predict the dynamics of TGF-β-induced target gene expression. This supports our hypothesis that the link between the identified Smad complexes and the regulation of the expression of the majority of TGF-β target genes is conserved among the four liver cell types studied.

**Gene Expression-Based Prediction and Experimental Validation of Dysregulation of the Abundance of Smad Proteins in Hepatocellular Carcinoma**

An important role in progression of hepatocellular carcinoma (HCC) has been attributed to Smad signal transduction (Dzieran et al., 2013). Currently, primarily genome-wide expression studies are available for patients with HCC. Therefore, we tested whether a reverse approach can be used to predict the abundance of Smad proteins present in patient samples on the basis of gene expression data.

We analyzed tumor-free and tumor tissue samples from 30 patients with HCC (Figure 7A, cohort A). By qRT-PCR the expression of the 12 TGF-β target genes, except for the mouse-specific gene Cxcl15 and TGFA, which was not detectable in human liver tissue, was analyzed. The remaining ten selected TGF-β target genes showed major alterations in their expression levels (Figure 7B), with SKI, SKIL, DNMT3A, and SOX4 being significantly upregulated in tumor samples. Cirrhotic or non-cirrhotic origin of the tumor samples had no impact on gene expression, except for BMP4, DNMT3A, and DUSP5, which were upregulated in cirrhotic compared with non-cirrhotic tissue.

For each patient, we incorporated the differences in the expression level of the selected Smad target genes between the tumor and the tumor-free samples in our mathematical model and predicted the corresponding alterations of the three Smad complexes, as well as TGF-β receptor activation required to achieve the observed expression pattern. Our analysis indicated that no major difference in the formation of the ppSmad2:Smad4:Smad4 complex occurred in the tumor compared with the tumor-free samples. However, the model predicted a significantly higher abundance of the ppSmad2:ppSmad3:ppSmad3 and ppSmad2:ppSmad3:Smad4 complexes in the tumor samples (Figure 7C).

Because of the model-predicted increase of the Smad complexes in the tumor context, we calculated the total abundance of Smad2, Smad3, and Smad4 proteins in the tumor-free and tumor samples on the basis of the predicted amounts of the three Smad complexes. The mathematical model predicted a significant increase in the mean of the sum of all Smad proteins in the tumor samples (Figures 7D and S7A). By quantitative IB we determined the abundance of Smad2, Smad3, and Smad4 proteins in the patient samples (Figure S7B). An upregulation of Smad2, Smad3, and Smad4 (Figure S7C), as well as a significant increase in the sum of all Smad proteins, was observed in the tumor samples compared with the tumor-free samples (Figure 7E). We confirmed that in the tissue setting, the TGF-β-induced activation of the Smad signaling pathway and of Smad target gene expression are far from saturation (Figure S7D).

The mathematical model predicted that activation of the TGF-β receptor is significantly elevated in the tumor compared with the tumor-free samples (Figure 7F). Since it is currently technically not possible to directly quantify the activation status of the TGF-β receptor, we used Smad2 phosphorylation as a proxy to analyze the pathway activation. We collected a new cohort of fresh patient material (cohort B), and determined Smad2 phosphorylation in these samples by quantitative IB (Figure S7E). In line with the model prediction, a significant increase in Smad2 phosphorylation was observed in the HCC samples (Figure 7G) and Smad2 protein levels were significantly increased in the tumor tissue samples (Figure S7F).

In sum, these observations underscore that our reverse modeling approach is capable of inferring quantitative information on the abundance of Smad proteins from gene expression...
data, and that elevated phosphorylation of Smad2 as well as elevated expression of Smad2, Smad3, and Smad4 proteins is characteristic for HCC.

**DISCUSSION**

In this study, we predicted and provided evidence for the three most relevant Smad complexes formed in response to TGF-β stimulation, and utilized a broadly applicable mathematical modeling approach to dissect the impact of these complexes on target gene expression.

Upon TGF-β stimulation only doubly phosphorylated Smad2 and Smad3 engage in complex formation. This notion is in agreement with the crystal structure of the doubly phosphorylated Smad2 homodimer, which revealed an essential role of both phosphoserine residues in stabilizing the complex (Wu et al., 2018).
Our results showed that only a minor fraction of the Smad2 and Smad3 molecules present in a cell contribute to complex formation. Likewise, it was shown that only minor amounts of Smad3 associate with Smad2 (Wu et al., 2001b).

Out of 84 possible combinations of complexes of non-phosphorylated and phosphorylated Smad2, Smad3, and Smad4, we predicted, by combining quantitative experimental data with the development of a mathematical model, the three most relevant trimeric Smad complexes in primary hepatocytes and hepatoma cells: ppSmad2:ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4, and ppSmad2:ppSmad3:Smad4. These results extend previous insights obtained by crystallographic studies showing that the C-terminal domain of Smad4 and the C-terminal domain of phosphorylated Smad2 have the capacity to form homotrimers (Wu et al., 2001a, 2001b). Sedimentation studies examining phosphorylation-induced Smad complex formation with a pseudo-phosphorylated Smad3 showed that Smad3 heterotrimer formation is favored over homotrimer formation (Chacko et al., 2001). In line with the ppSmad2:ppSmad3:Smad4 complex that we detected in our study, an in situ proximity ligation assay showed that TGF-β stimulation induced the formation of complexes consisting of Smad2, Smad3, and Smad4 (Zieba et al., 2012).

Despite chromatin IP and microarray studies (Qin et al., 2009; Zhang et al., 2011), it was not possible to define the specific contribution of individual trimeric Smad complexes to gene expression. By overexpression and knockdown of individual Smad proteins, and a mathematical model that links the TGF-β-induced activation of signal transduction to target gene expression, we dissected the contribution of the identified Smad complexes to target gene expression and revealed their positive or negative regulatory effect.

Our analysis indicated that the ppSmad2:ppSmad3:Smad4 complex is the most important complex involved in TGF-β-induced gene expression with a positive influence on most of the genes, and that ppSmad2:ppSmad3:Smad4 and ppSmad2:Smad4:Smad4 complexes repress the expression of DUSP5 and PDK4. Our transcription factor and simulation analyses suggested an indirect regulation, possibly mediated by cell-type-specific co-factors, such as STAT3, GATA4, and C/EBPβ (Qin et al., 2009). A similar approach was used to investigate erythropoietin-induced heterodimer formation between Stat5a and Stat5b (Boehm et al., 2014). In general, our technology combining IP, quantitative mass spectrometry, and mathematical modeling with L₁ regularization, can be used to quantitatively investigate protein-protein interactions and thereby add quantitative information to protein-protein interactions that were mapped on a genome-wide scale (Li et al., 2017).

Our systems biology approach enabled us to quantitatively link the total amount of Smad proteins and Smad-regulated gene expression in hepatoma cells and in primary hepatocytes. We provided evidence that alterations in the abundance of Smad proteins do not change the type but the amount of the three Smad complexes formed. Our studies show that the link between the dynamics of Smad complex formation and the regulation of gene expression is mostly conserved in primary hepatocytes and hepatoma cells.

Microarray studies of gene expression in HCC are well established (Hao et al., 2011). We demonstrated that, with the aid of our mathematical model, it is possible to predict the expression of TGF-β-induced genes based on protein data and to use gene expression data to predict total levels of Smad proteins.

In the presented study, predictions of the mathematical model indicated that the abundance of Smad2, Smad3, and Smad4, as well as Smad2 phosphorylation, are increased in HCC tissue, which was experimentally confirmed in liver samples from HCC patients. Congruently, a previous study showed that mutations in the Smad2 and Smad4 genes might contribute to the development of HCC (Yakicier et al., 1999).

As TGF-β can exert multiple functions depending on the cellular context, it is tempting to speculate that cell-type-specific abundance of Smad proteins might be key to explain pleiotropic effects of TGF-β signaling. Therefore, HCC-specific alterations in the abundance of Smad proteins might affect the extent to which the three complexes are formed, and increase or decrease the expression of target genes contributing to tumor progression.

Our approach represents a generally applicable framework that establishes a quantitative link between complex formation of transcription factors, signaling dynamics, and gene expression. With this approach it is not only possible to use protein information to predict the dynamics of gene expression as commonly practiced, but conversely enables to predict upstream protein abundance based on gene expression dynamics.
QUANTIFICATION AND STATISTICAL ANALYSIS
DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and two tables and can be found with this article online at https://doi.org/10.1016/j.cels.2017.11.010.

ACKNOWLEDGMENTS

We thank Luigi Terracciano from the Institute of Pathology in Basel for histological diagnosis of the human liver tissue samples (cohort B). This work was supported by the NIH (1R01DK090347-01), by the DFG-funded SFB/Transregio 179, and by the German Ministry of Education and Research (EMBF) within the eBio collaborative research projects “Systems Biology of Erythropoietin” (SBEpo, 0311628B) and “Multi-Scale Modeling of Drug Induced Liver Injury” (MS_DILI, 0311074A, 0311074B) and the junior group EAsys (0311080) within the Virtual Liver Network (0315745, 0315741), within the Liver Systems Medicine network (LiSYSM, 0311042, 0311048), within the CancerSys network “LungSys II” (0316042A), and within the German Center for Lung Research (DZL, 82DZL00404).

AUTHOR CONTRIBUTIONS

P.L., M.S., and U.K. designed the project. M.S., U.K., and J.T. supervised the project. Cell culture, quantitative IB, and experiments for mass spectrometry measurements were performed by P.L. Mass spectrometry analysis was performed by M.E.B. and supervised by W.D.L. Microarray measurements were performed by M.M. and N.G., and data analysis was performed by C.K. N.I. performed the Genomatix Pathway System analysis. B.S. contributed in developing the model selection methodology. A.V., P.L., D.D., G.D., and C.K. N.I. performed the Genomatix Pathway System analysis. B.S. contributed in developing the model selection methodology. A.V., P.L., D.D., G.D., and D.S. prepared primary human hepatocytes and the human liver samples of cohort A. M.S.M., M.H., and K.H. prepared the human liver samples of cohort B. P.L., A.V., M.S., and U.K. wrote and all authors approved the manuscript.

Received: November 30, 2016
Revised: June 23, 2017
Accepted: November 14, 2017
Published: December 13, 2017

REFERENCES


# STAR METHODS

## KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antibodies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-Smad2/3</td>
<td>BD Biosciences</td>
<td>Cat#610843; RRID:AB_398161</td>
</tr>
<tr>
<td>anti-Smad2</td>
<td>Cell Signaling Technology</td>
<td>Cat#5339; RRID:AB_10626777</td>
</tr>
<tr>
<td>anti-Smad3</td>
<td>Cell Signaling Technology</td>
<td>Cat#9523; RRID:AB_2193182</td>
</tr>
<tr>
<td>anti-Smad4</td>
<td>Cell Signaling Technology</td>
<td>Cat#9515; RRID:AB_2193344</td>
</tr>
<tr>
<td>anti-Flag</td>
<td>Rockland</td>
<td>Cat#600-401-383; RRID:AB_219374</td>
</tr>
<tr>
<td>anti-pSmad2</td>
<td>Cell Signaling Technology</td>
<td>Cat#3108; RRID:AB_490941</td>
</tr>
<tr>
<td>anti-pSmad3</td>
<td>Cell Signaling Technology</td>
<td>Cat#9520; RRID:AB_2193207</td>
</tr>
<tr>
<td>anti-Smad4</td>
<td>Santa Cruz Biotechnology</td>
<td>Cat#sc-7966; RRID:AB_627905</td>
</tr>
<tr>
<td><strong>Chemicals, Peptides, and Recombinant Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFβ1</td>
<td>R&amp;D Systems</td>
<td>Cat#240-B-010</td>
</tr>
<tr>
<td>SBP-Smad2 calibrator protein</td>
<td>This paper; N/A</td>
<td></td>
</tr>
<tr>
<td>GST-Smad3 calibrator protein</td>
<td>This paper; N/A</td>
<td></td>
</tr>
<tr>
<td>SBP-Smad4 calibrator protein</td>
<td>This paper; N/A</td>
<td></td>
</tr>
<tr>
<td>Actinomycin D</td>
<td>BioChemica</td>
<td>A1489</td>
</tr>
<tr>
<td>SB-431542</td>
<td>Sigma Aldrich</td>
<td>S4317</td>
</tr>
<tr>
<td>Smad3 blocking peptides</td>
<td>Cell Signaling; #1933S</td>
<td></td>
</tr>
<tr>
<td><strong>Critical Commercial Assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNeasy Mini Plus Kit</td>
<td>Qiagen</td>
<td>Cat#74136</td>
</tr>
<tr>
<td>Affymetrix GeneChip Mouse Genome 430 2.0 Array</td>
<td>Thermo</td>
<td>Cat#900495</td>
</tr>
<tr>
<td>Applied Biosystems High-Capacity cDNA Reverse Transcription Kit</td>
<td>Thermo</td>
<td>Cat#4368813</td>
</tr>
<tr>
<td>NucleoSpin RNA Kit</td>
<td>Macherey-Nagel</td>
<td>Cat#740955</td>
</tr>
<tr>
<td><strong>Deposited Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microarray data set: Effect of TGFβ treatment (1 ng/ml) on gene expression in Hepa1-6 cells</td>
<td>This paper</td>
<td><a href="http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90954">http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90954</a></td>
</tr>
<tr>
<td><strong>Experimental Models: Cell Lines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepa1-6</td>
<td>ATCC</td>
<td>Cat#CRL-1830; RRID:CVCL_0327</td>
</tr>
<tr>
<td>HepG2</td>
<td>ATCC</td>
<td>Cat#HB-8065; RRID:CVCL_0027</td>
</tr>
<tr>
<td><strong>Experimental Models: Organisms/Strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C57BL/6N mice</td>
<td>Charles River</td>
<td>Cat#027</td>
</tr>
<tr>
<td><strong>Oligonucleotides</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMARTpool: ON-TARGETplus Smad2 siRNA</td>
<td>Dharmacon</td>
<td>L-040707-00-0005</td>
</tr>
<tr>
<td>SMARTpool: ON-TARGETplus Smad3 siRNA</td>
<td>Dharmacon</td>
<td>L-040706-00-0005</td>
</tr>
<tr>
<td>SMARTpool: ON-TARGETplus Smad4 siRNA</td>
<td>Dharmacon</td>
<td>L-040687-00-0005</td>
</tr>
<tr>
<td>ON-TARGETplus Non-targeting pool</td>
<td>Dharmacon</td>
<td>D-001810-10-20</td>
</tr>
<tr>
<td><strong>Recombinant DNA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pMOWS-Flag-Smad2</td>
<td>This paper; N/A</td>
<td></td>
</tr>
<tr>
<td>pMOWS-Flag-Smad3</td>
<td>This paper; N/A</td>
<td></td>
</tr>
<tr>
<td>pMOWS-Flag-Smad4</td>
<td>This paper; N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Software and Algorithms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xcalibur Version 3.0.63.</td>
<td>Thermo</td>
<td>RRID:SCR_014593</td>
</tr>
<tr>
<td>MaxQuant</td>
<td><a href="http://www.biochem.mpg.de/5111795/maxquant">http://www.biochem.mpg.de/5111795/maxquant</a></td>
<td>version 1.5.0.12</td>
</tr>
</tbody>
</table>

(Continued on next page)
CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Ursula Klingmüller (u.klingmueller@dkfz.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Culture of Cell Lines and Primary Cells

The mouse hepatoma cell line Hepa1-6 (ATCC CRL-1830, female) and the human hepatoma cell line HepG2 (ATCC HB-8065, male) were cultivated in Dulbecco’s modified Eagle’s Medium (DMEM, Gibco) supplemented with 10% (v/v) fetal bovine serum (FBS, Life Technologies), 1% 100 U penicillin/streptomycin (Gibco) and 1% 200 mM glutamine (Gibco). Primary human hepatocytes were isolated from macroscopically healthy tissue that remained from resected human liver of three patients with primary or secondary liver tumors or benign local liver diseases by a two-step EDTA/collagenase perfusion technique (Kegel et al., 2016). Informed consent of the patients for the use of tissue for research purposes was obtained according to the ethical guidelines of the Charité University Medicine Berlin. Detailed donor anamnesis of the three patients providing primary human hepatocytes is stated in the table below. Primary mouse hepatocytes were isolated as previously described (Mueller et al., 2015) from 8- to 12-week-old male C57BL/6N mice (Charles River) housed at the DKFZ animal facility under a constant light/dark cycle, maintained on a standard mouse diet, and allowed ad libitum access to food and water. All animal experiments were approved by the governmental review committee on animal care of the state Baden-Württemberg, Germany (reference number A-24/10). Primary mouse hepatocytes and primary human hepatocytes were cultivated in phenol red-free Williams E medium (Biochrom) supplemented with 10% (v/v) fetal bovine serum (FBS, Life Technologies), 0.1 µM dexamethasone, 10 µg/ml insulin (Sigma-Aldrich), 2 mM L-glutamine (Gibco) and 1% (v/v) penicillin/streptomycin 100× (Gibco) using collagen I-coated cell dishes (BD Biosciences). 24 hours before the experiment, 1.2×10⁶ Hepa1-6 cells and 2×10⁶ HepG2 cells, primary mouse and primary human hepatocytes (for IB and qRT-PCR experiments) and 7.5×10⁶ cells (for mass spectrometry experiments) were seeded. 4 hours before the experiment, the different cell types were washed three times with PBS and kept in growth factor depleted medium supplemented with 1% penicillin/streptomycin (Gibco) and 1% glucose (Gibco).

Donor anamnesis of the three patients providing primary human hepatocytes

<table>
<thead>
<tr>
<th>Donor</th>
<th>Age (years)</th>
<th>Sex</th>
<th>BMI</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58</td>
<td>female</td>
<td>24.7</td>
<td>Liver metastases</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>female</td>
<td>22.6</td>
<td>Liver metastases</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>male</td>
<td>27</td>
<td>Liver metastases</td>
</tr>
</tbody>
</table>

Hepatocellular Carcinoma and Tumor-free Tissue Samples

A first cohort (Cohort A) of samples from liver tumor patients and corresponding tumor-free liver tissue were provided by the Charité University Medicine Berlin. Informed consent of the patients for the use of tissue for research purposes was obtained corresponding to the ethical guidelines of Charité University Medicine Berlin. Detailed donor anamnesis is stated in the table below.

Donor anamnesis of HCC and tumor-free tissue samples (Cohort A)

<table>
<thead>
<tr>
<th>Donor</th>
<th>Cirrhosis</th>
<th>Age (years)</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>No</td>
<td>71</td>
<td>M</td>
</tr>
<tr>
<td>A2</td>
<td>No</td>
<td>74</td>
<td>M</td>
</tr>
</tbody>
</table>

(Continued on next page)
A second cohort (Cohort B) of freshly frozen samples from liver tumor patients and corresponding tumor-free liver tissue were provided by the University Hospital Heidelberg and University of Basel to measure phosphorylation of Smad proteins. Informed consent of the patients for the use of tissue for research purposes was obtained corresponding to the ethical guidelines of University Hospital Heidelberg and University of Basel. Detailed donor anamnesis is stated in the table below.

**Donor anamnesis of HCC and tumor-free tissue samples (Cohort A)**

<table>
<thead>
<tr>
<th>Donor</th>
<th>Cirrhosis</th>
<th>Age (years)</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td>No</td>
<td>58</td>
<td>F</td>
</tr>
<tr>
<td>A4</td>
<td>No</td>
<td>78</td>
<td>F</td>
</tr>
<tr>
<td>A5</td>
<td>No</td>
<td>58</td>
<td>F</td>
</tr>
<tr>
<td>A6</td>
<td>No</td>
<td>68</td>
<td>M</td>
</tr>
<tr>
<td>A7</td>
<td>No</td>
<td>66</td>
<td>M</td>
</tr>
<tr>
<td>A8</td>
<td>No</td>
<td>87</td>
<td>M</td>
</tr>
<tr>
<td>A9</td>
<td>No</td>
<td>76</td>
<td>M</td>
</tr>
<tr>
<td>A10</td>
<td>No</td>
<td>76</td>
<td>M</td>
</tr>
<tr>
<td>A11</td>
<td>No</td>
<td>71</td>
<td>M</td>
</tr>
<tr>
<td>A12</td>
<td>No</td>
<td>61</td>
<td>F</td>
</tr>
<tr>
<td>A13</td>
<td>No</td>
<td>61</td>
<td>M</td>
</tr>
<tr>
<td>A14</td>
<td>No</td>
<td>73</td>
<td>M</td>
</tr>
<tr>
<td>A15</td>
<td>No</td>
<td>65</td>
<td>M</td>
</tr>
<tr>
<td>A16</td>
<td>Yes</td>
<td>66</td>
<td>M</td>
</tr>
<tr>
<td>A17</td>
<td>Yes</td>
<td>79</td>
<td>M</td>
</tr>
<tr>
<td>A18</td>
<td>Yes</td>
<td>56</td>
<td>F</td>
</tr>
<tr>
<td>A19</td>
<td>Yes</td>
<td>76</td>
<td>M</td>
</tr>
<tr>
<td>A20</td>
<td>Yes</td>
<td>70</td>
<td>F</td>
</tr>
<tr>
<td>A21</td>
<td>Yes</td>
<td>67</td>
<td>M</td>
</tr>
<tr>
<td>A22</td>
<td>Yes</td>
<td>69</td>
<td>M</td>
</tr>
<tr>
<td>A23</td>
<td>Yes</td>
<td>59</td>
<td>M</td>
</tr>
<tr>
<td>A24</td>
<td>Yes</td>
<td>76</td>
<td>M</td>
</tr>
<tr>
<td>A25</td>
<td>Yes</td>
<td>71</td>
<td>F</td>
</tr>
<tr>
<td>A26</td>
<td>Yes</td>
<td>57</td>
<td>F</td>
</tr>
<tr>
<td>A27</td>
<td>Yes</td>
<td>71</td>
<td>M</td>
</tr>
<tr>
<td>A28</td>
<td>Yes</td>
<td>71</td>
<td>M</td>
</tr>
<tr>
<td>A29</td>
<td>Yes</td>
<td>67</td>
<td>M</td>
</tr>
<tr>
<td>A30</td>
<td>Yes</td>
<td>63</td>
<td>M</td>
</tr>
</tbody>
</table>

**Donor anamnesis of HCC and tumor-free tissue samples (Cohort B)**

<table>
<thead>
<tr>
<th>Donor</th>
<th>Cirrhosis</th>
<th>Age (years)</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Yes</td>
<td>66</td>
<td>M</td>
</tr>
<tr>
<td>B2</td>
<td>Yes</td>
<td>72</td>
<td>M</td>
</tr>
<tr>
<td>B3</td>
<td>No</td>
<td>68</td>
<td>M</td>
</tr>
<tr>
<td>B4</td>
<td>No</td>
<td>75</td>
<td>M</td>
</tr>
<tr>
<td>B5</td>
<td>No</td>
<td>80</td>
<td>M</td>
</tr>
<tr>
<td>B6</td>
<td>Yes</td>
<td>55</td>
<td>W</td>
</tr>
<tr>
<td>B7</td>
<td>No</td>
<td>73</td>
<td>M</td>
</tr>
<tr>
<td>B8</td>
<td>Yes</td>
<td>64</td>
<td>M</td>
</tr>
<tr>
<td>B9</td>
<td>Yes</td>
<td>57</td>
<td>M</td>
</tr>
<tr>
<td>B10</td>
<td>No</td>
<td>58</td>
<td>M</td>
</tr>
<tr>
<td>B11</td>
<td>No</td>
<td>59</td>
<td>M</td>
</tr>
<tr>
<td>B12</td>
<td>No</td>
<td>82</td>
<td>W</td>
</tr>
</tbody>
</table>
METHOD DETAILS

Stimulation, Lysis and SDS-PAGE

Cells were stimulated with 1 ng/ml of TGFβ1 (R&D Systems, Cat #240-B-010) for up to 10 hours. For IP, Hepa1-6 cells (2 × 10⁶) were lysed in total cell lysis buffer (1% NP40, 150 mM NaCl, 20 mM Tris-HCl pH 7.4, 10 mM NaF, 1 mM EDTA pH 8.0, 2 mM ZnCl₂ pH 4.0, 1 mM MgCl₂, 2 mM Na₂VO₃, 20% glycerol, 2 μg/ml aprotinin and 200 μg/ml AEBSF). Lysates were rotated for 30 minutes at 4°C, sonicated and centrifuged for 10 minutes at 20,800 × g and 4°C. The supernatant was subjected to IPs with anti-Smad2/3, anti-Smad2, anti-Smad3, anti-Smad4 or anti-Flag antibodies (BD-610843, Cell Signaling #5339; #9523; #9515, Rockland 600-401-383 respectively, dilution 1:100), supplemented with Protein A sepharose (GE Healthcare) and recombinant calibrator proteins. The IPs were rotated overnight at 4°C. Immunoprecipitated proteins were separated by SDS-PAGE. Gels used for mass spectrometry were washed three times with water for 5 minutes and Coomassie stained for 1 hour according to the SimplyBlue SafeStain (Invitrogen) instructions. For IB, proteins were transferred to nitrocellulose membranes. IB was performed with anti-pSmad2 (Cell Signaling, #3108), anti-pSmad3 (Cell Signaling, #9520), anti-Smad4 (Santa Cruz, sc-7966), anti-Smad2/3 antibodies (BD-610843) and anti-Flag (Rockland 600-401-383) antibodies. Horseradish peroxidase (HRP) conjugated anti-mouse IgG HRP (Dianova 115-035-146), anti-rabbit IgG HRP (Dianova 111-035-144) and anti-Protein A HRP (GE Healthcare NA9120) secondary antibodies were used for chemiluminescence detection employing ECL substrate (GE Healthcare). Chemiluminescence was measured with an ImageQuant LAS 4000 device (GE Healthcare) utilizing a CCD-camera allowing the detection in a broad linear range. Band intensities were quantified using the ImageQuantTL Software (GE Healthcare).

Quantitative Mass Spectrometry

After Smad protein enrichment by IP, the following sample preparation steps were performed: purification per 1D SDS-PAGE, staining with Coomassie, gel band extraction, destaining, reduction with dithiothreitol (Sigma) and alkalylation with iodoacetamide (Sigma). To analyze the Smad2 and Smad3 degree of phosphorylation and relative protein abundance, both proteins were cut out together and digested with LysC (Roche Diagnostics). To analyze the relative protein abundance of Smad4, extracts of these gel bands were additionally subjected to tryptic digestion (Trypsin Sequencing Grade from bovine pancreas, Roche Diagnostics). The digestion buffer was 100 mM NH₄HCO₃ in 5% acetonitrile. Following overnight incubation, peptide extraction was performed by transferring the supernatant to an extra vial and performing three further extraction steps with acetonitrile, 5% formic acid and again acetonitrile. Samples were concentrated in a Speedvac (Eppendorf) and desalted with C18 ZipTips (Millipore) applying a protocol based on water, acetonitrile and trifluoroacetic acid. To equalize the recovery of peptides and corresponding phosphopeptides from the LC system, we added citrate to a final concentration of 20 mM to LysC digested samples. Samples were measured by nanoUPLC (nanoAcquity UPLC, Waters) coupled to an LTQ-Orbitrap XL mass spectrometer (Thermo Scientific). We applied a precolumn setup and acetonitrile based gradients (0-40% in < 1 hour). Smad2 and Smad3 protein ratios as well as degrees of phosphorylation were analyzed by manual peak integration using Thermo Xcalibur Version 3.0.63. Relative protein abundances of Smad2, Smad3 and Smad4 were analyzed using peptide raw intensities generated by MaxQuant (1.5.0.12).

For pairwise relative Smad2/Smad3 isoform quantification by mass spectrometry, bands from 1D-PAGE were excised, ensuring that both Smad2 and Smad3 are quantitatively present in one band. Because both isoforms exhibit a high degree of sequence similarity, digestion of Smad2 and Smad3 using LysC leads to three categories of peptides. One category consists of identical peptides for Smad2 and Smad3, the next comprises highly similar peptides that differ only in one or a few amino acids and the third category is formed by completely different peptides for both isoforms. The isoform abundance for Smad2 and Smad3 can be determined by comparing the signal intensities within pairs of highly similar peptides. For accurate relative quantification we analyzed two such pairs. Each pair (Smad2 vs. Smad3) differed in just one amino acid: acSSILPF-pT-PPVK vs. acSSILPF-pT-PPVK and (K)TGRLEDEK vs. (K)TGQDELEK. All signals detected from these peptides, such as different charge states, threonine phosphorylation (first pair) and deamidation for the second pair as well as versions with and without N-terminal lysine (second pair) were considered. By applying this strategy, the quantification of the isoform abundances were highly precise (SD < 2%, n = 13, biological replicates). Isoform ratios calculated from both peptide pairs showed good agreement within 5%. The Smad2 and Smad3 ratio used for mathematical modeling is the mean of both pairs.

For bulk-based relative Smad4 quantification by mass spectrometry, a lower molecular weight region was additionally excised from the 1D-PAGE gel. Regarding the amino acid sequence, Smad2 and Smad3 are much more similar to each other than to Smad4. Therefore, following digestion no highly similar peptides are formed that would be suitable for a pairwise relative quantification between Smad4 and Smad2/3. For this reason, the selected quantification strategy relied on the accumulated intensities of all peptides detectable for each Smad protein. To maximize the number of peptides, tryptic digestion was performed. For relative protein quantification among all three Smad proteins, the sum of all Smad4 peptide intensities (up to 27 peptides) was compared to the sum of all Smad2 and Smad3 peptides (up to 16 unique Smad2 peptides, 12 unique Smad3 peptides and 11 common Smad2 and Smad3 peptides). After that, the highly accurate Smad2 to Smad3 ratio from a corresponding LysC-digested and pairwise relatively quantified aliquot was used to adjust the accurate ratio among all three Smad proteins.

Relative quantification of Smad2 and Smad3 phosphorylation occupancies was performed by analyzing non-phosphorylated, singly phosphorylated and doubly phosphorylated C-terminal LysC peptides. The amino acid sequences of the peptides from both isoforms differ only in exchange of two amino acids (ppSmad2: VLTQMGSPSVR-camC-S-pS-M-pS, ppSmad3: VLTQMGSPSIR-camC-S-pS-V-pS). Among all detected Smad2 and Smad3 peptides, these C-terminal LysC-fragments showed...
the most abundant signals. To equalize recovery of the different phospho-forms from the LC, samples were injected in 20 mM citrate. For the standard-free quantification method applied, signal intensities of all detectable charge states were taken into account. Cysteine residues were present as carbamidomethylated (cam) modified residue, because samples were treated with dithiothreitol and iodoacetamide during the workflow. An additional frequent modification within the analyte peptide is methionine oxidation. This modification turned out to be phosphorylation-independent. A different methionine oxidation status leads to a retention time shift (C18 column) of up to several minutes. For Smad2 and Smad3 standard-free phosphorylation status determination of the three different phospho-forms (non-oxidized, singly oxidized for Smad2 and Smad3 and additionally a doubly oxidized version for Smad2) showed no significant difference in degrees of phosphorylation, confirming a correct quantification with low random and systematical errors.

**Validation of Complex Formation by Sequential Immunoprecipitations**

7.5x10^6 Hepa1-6 cells were stimulated with 1 ng/ml of TGFβ for 60 minutes or were left untreated. Cells were lysed and processed as described above. For Smad3 or Smad4 depletion, three sequential IPs were performed with anti-Smad3 (Cell Signaling #9523) or anti-Smad4 antibodies (Cell Signaling #38454), respectively. The lysates and the depleted supernatants were subjected to an anti-Smad4 or anti-Smad3 IP, respectively, and IB was performed with an anti-Smad2 (Cell Signaling #5339) antibody, followed by an anti-Smad4 (Santa Cruz sc-7966) antibody or an anti-Smad3 (Cell Signaling #9523) antibody.

For the validation of the heterotrimeric complex, the lysates were first subjected to an IP with ant-Smad3 (Cell Signaling #9523). After overnight incubation, bead-bound proteins were dissociated by the addition of a Smad3 blocking peptide (Cell Signaling #19335) with a 5-fold excess by weight compared to the antibody. The obtained supernatants were subjected to an IP with an anti-Smad4 antibody (Cell Signaling #38454). Quantitative IB was performed with an anti-Smad2 antibody (Cell Signaling #5339). For the detection of the immunoblots chemiluminescence in combination with a CCD camera based device, ImageQuant, was used.

**Microarray Analysis of Gene Expression Data**

2x10^6 Hepa1-6 cells were stimulated with 1 ng/ml of TGFβ for 0, 1, 3, 6 and 10 hours in biological duplicates. As controls, the same time points were evaluated in duplicates without TGFβ treatment. RNA was extracted using the RNeasy Mini Plus Kit (Qiagen) according to the manufacturer’s instructions. High-throughput quantification of the gene expression induced by TGFβ was performed using Affymetrix Mouse genome 430 2.0 microarrays according to the manufacturer’s instructions. For preprocessing the R statistical computing environment was used and the robust multiarray average (RMA) algorithm was applied as implemented in the simpleaffy package. The expression data were deposited in the Gene Expression Omnibus (GEO) database under the accession number GEO: GSE90954.

The selection criterion for the TGFβ target genes was a significant (p<0.01) and more than 1.5-fold induction compared to untreated controls. For this analysis, a linear model accounting for time and treatment effects was applied and p-values were calculated based on the t-statistic. Genes that were not constant over time, i.e. showing a maximal regulation of more than 1.5-fold at one point in time in the untreated controls were discarded from further analyses. The duplicates were averaged and standard errors of the means were calculated. Since it is not feasible to reliably calculate standard errors from duplicates, the median over all standard errors was used as uncertainty for further analyses. The dynamics of the induced gene expression as shown in Figure 3D was then estimated by fitting a five parameter transient function

\[ f(t) = A_{\text{trans}} \left(1 - e^{-\frac{t}{\tau_1}}\right) + A_{\text{trans}} \left(1 - e^{-\frac{t}{\tau_2}}\right) e^{-\frac{t}{\tau_3}} + p_0 \]

(Equation 2)

to the time courses for each gene and both treatment conditions.

The first term represents a sustained response with amplitude \(A_{\text{trans}}\) and time constant \(\tau_1\). The second term accounts for transient up- or down-regulation with amplitude \(A_{\text{trans}}\) with the same time constant \(\tau_1\) for induction and a second time scale \(\tau_2\) for relaxation. The last parameter \(p_0\) is the offset which is specified during fitting primarily by the measurement at \(t=0\). To prevent overfitting, only the two mentioned time scales were allowed and the parameters were restricted to reasonable ranges. For both time scales, it was assumed that they are smaller than two times the whole measurement interval, i.e. \(\tau_1, \tau_2 < 2 \tau_{\text{max}} = 20\) hours. As lower bounds for the two time scales, one half of the smallest sampling time interval, i.e. \((t_2 - t_1)/2 = 0.5\) hours was assumed. Smaller time scales, i.e. a faster dynamics could not be resolved by the available experimental data and would therefore lead to overfitting of the data and to large uncertainties of the predicted dynamics. For the amplitudes, the interval \([1 \times 10^{-12}, 2 \Delta y]\) was used as constraint where \(\Delta y\) denotes the observed range of the measurements. For the offset, \(p_0 \in [\min(y) - \Delta y/2, \max(y)]\) was allowed where \(\min(y)\) and \(\max(y)\) denotes the smallest and the largest observation. All five parameters were estimated by maximum likelihood. Optimization was performed on the logarithmic parameter scale. Next, k-means clustering with \(k = 12\) was performed as implemented in MATLAB using Euclidean distances. For this purpose, the dynamics estimated for the treated and untreated conditions were merged. Therefore, genes with a similar dynamics after treatment but with a distinct basal expression could be assigned to different clusters.

For the selection of a representative gene for each cluster, we performed literature mining using the software suites MetaCore (version 6.31 build 68930) and Ingenuity IPA (build 441680M) to identify potential Smad2-, Smad3-, Smad4- and TGFβ-specific target genes.
RNA Extraction and Quantitative Real-time PCR

2×10^6 Hepa1-6 cells, HepG2 cells or primary mouse hepatocytes were stimulated with 1 ng/ml TGFβ for up to 10 hours in biological triplicates. Primary human hepatocytes were seeded on collagen-coated 6-well-plates (BD Biosciences) and cultivated as described for primary mouse hepatocytes (Mueller et al., 2015). Briefly, cells were cultured in full medium at 37°C, 5% CO2 in a humidified incubator for 24 hours. For Hepa1-6 and HepG2 cells, medium was then changed to serum-free medium for additional 24 hours. Prior to TGFβ treatment, cells were cultured for 4 hours in serum and dexamethasone-free medium for equilibration. Primary human hepatocytes were then incubated with 1 ng/ml TGFβ for 30, 60, 120, 240 and 600 minutes. Cells without TGFβ treatment served as negative control. For the experiments with SB-431542 (Sigma Aldrich, S4317), Hepa1-6 cells were pre-treated for 30 minutes with 5 µM SB-431542 prior to TGFβ stimulation. For samples treated with TGFβ alone, as solvent control the same amount of DMSO was applied. RNA was extracted using the RNeasy Mini Plus Kit (Qiagen) according to the manufacturer’s instructions. Complementary DNA was generated with the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) and analyzed using the LightCycler 480 with the hydrolysis-based Universal Probe Library (UPL) platform (Roche Diagnostics). Gene-specific primers and UPL probes are displayed in the table below. Crossing point values were calculated using the second-derivative-maximum method of the LightCycler 480 Basic Software (Roche Applied Science). PCR efficiency correction was performed for each PCR setup individually. mRNA data was normalized against HPRT.

<table>
<thead>
<tr>
<th>Gene-specific mouse (m) and human (h) primers and UPL probes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
</tr>
<tr>
<td>mBmp4</td>
</tr>
<tr>
<td>mCxcl15</td>
</tr>
<tr>
<td>mDnmt3a</td>
</tr>
<tr>
<td>mDusp5</td>
</tr>
<tr>
<td>mHprt</td>
</tr>
<tr>
<td>mJun</td>
</tr>
<tr>
<td>mKlf10</td>
</tr>
<tr>
<td>mPdk4</td>
</tr>
<tr>
<td>mSki</td>
</tr>
<tr>
<td>mSkil</td>
</tr>
<tr>
<td>mSmad7</td>
</tr>
<tr>
<td>mSox4</td>
</tr>
<tr>
<td>mTgfa</td>
</tr>
<tr>
<td>hBMP4</td>
</tr>
<tr>
<td>hDNMT3A</td>
</tr>
<tr>
<td>hDUSP5</td>
</tr>
<tr>
<td>hHPRT</td>
</tr>
<tr>
<td>hJUN</td>
</tr>
<tr>
<td>hKLF10</td>
</tr>
<tr>
<td>hPKD1</td>
</tr>
<tr>
<td>hSki</td>
</tr>
<tr>
<td>hSKIL</td>
</tr>
<tr>
<td>hSMAD7</td>
</tr>
<tr>
<td>hSOX4</td>
</tr>
<tr>
<td>hTGFα</td>
</tr>
</tbody>
</table>

Mathematical Model

Development of the mathematical model based on ordinary differential equations (ODEs) and model simulations were performed using the MATLAB-based modeling environment D2D (www.data2dynamics.org) (Raue et al., 2015). All reactions at the pathway level were implemented as mass-action kinetics and the impact of the Smad complexes was described by Michaelis-Menten kinetics. Parameters that were estimated to be very low were set to zero without changing the fit nor the predicted dynamics. Estimated model parameters for the reduced model extended to gene expression are shown in Table S2. The mathematical model and the data sets are open source and available to the public at www.data2dynamics.org.
Description of the Comprehensive Mathematical Model

TGFβ is binding to the TGFβ receptor and is subsequently leading to its activation. In addition the activated receptor can be down-regulated by degradation (Derynick and Feng, 1997; Itoh and ten Dijke, 2007). The non-phosphorylated Smad2 and Smad3 monomers are susceptible to be double phosphorylated by the active receptor (Massague et al., 2005). The active double phosphorylated Smad monomers can be inactivated by a two-step dephosphorylation into single phosphorylated and afterwards in non-phosphorylated Smad monomers. The dissociation of each trimeric Smad complex is dependent on the dephosphorylation of the double phosphorylated Smad2 and Smad3 in the heterotrimeric Smad complexes or on the dissociation of the homotrimeric Smad4 complex. The double phosphorylated Smad2/Smad3 and Smad4 are able to form different complexes (Wrana, 2002). The active Smad complexes activate or inhibit target gene expression (Levy and Hill, 2006; Qin et al., 2009; Zhang et al., 2011).

Reactions of the Comprehensive Mathematical Model

The reactions of the comprehensive model are as follows:

\[ v_1 = [\text{Rec}] \cdot [\text{Rec}_{\text{act}}] \cdot [\text{TGFb}] \] (Equation 3)

\[ v_2 = [\text{TGFb}_{\text{pRec}}] \cdot [\text{Rec}_{\text{degind}}] \] (Equation 4)

\[ v_3 = k_{\text{on}_{222}} \cdot [\text{ppS2}]^3 \] (Equation 5)

\[ v_4 = 3 \cdot [\text{S}_{\text{dephosphos}}] \cdot [\text{ppS2}_{\text{ppS2}}] \] (Equation 6)

\[ v_5 = k_{\text{on}_{333}} \cdot [\text{ppS3}]^3 \] (Equation 7)

\[ v_6 = 3 \cdot [\text{S}_{\text{dephosphos}}] \cdot [\text{ppS3}_{\text{ppS3}}] \] (Equation 8)

\[ v_7 = [\text{S4}]^3 \cdot k_{\text{on}_{444}} \] (Equation 9)

\[ v_8 = [\text{S4}_{\text{S4}}] \cdot k_{\text{diss}_{\text{SS}}} \] (Equation 10)

\[ v_9 = [\text{S2}] \cdot [\text{S}_{\text{phos}}] \cdot [\text{TGFb}_{\text{pRec}}] \] (Equation 11)

\[ v_{10} = [\text{S}_{\text{dephosphos}}] \cdot [\text{ppS2}] \] (Equation 12)

\[ v_{11} = [\text{S}_{\text{dephosphos}}] \cdot [\text{pS2}] \] (Equation 13)

\[ v_{12} = [\text{S3}] \cdot [\text{S}_{\text{phos}}] \cdot [\text{TGFb}_{\text{pRec}}] \] (Equation 14)

\[ v_{13} = [\text{S}_{\text{dephosphos}}] \cdot [\text{ppS3}] \] (Equation 15)

\[ v_{14} = [\text{S}_{\text{dephosphos}}] \cdot [\text{pS3}] \] (Equation 16)

\[ v_{15} = k_{\text{on}_{223}} \cdot [\text{ppS2}]^2 \cdot [\text{ppS3}] \] (Equation 17)

\[ v_{16} = 2 \cdot [\text{S}_{\text{dephosphos}}] \cdot [\text{ppS2}_{\text{ppS2}}] \] (Equation 18)

\[ v_{17} = [\text{S}_{\text{dephosphos}}] \cdot [\text{ppS2}_{\text{ppS2}}] \] (Equation 19)
\( v_{18} = [S4] \cdot k_{on,224} \cdot [ppS2]^2 \)  
(Equation 20)

\( v_{19} = 2 \cdot S_{dephosphos} \cdot [ppS2 \cdot ppS2 \cdot S4] \)  
(Equation 21)

\( v_{20} = k_{on,233} \cdot [ppS2] \cdot [ppS3]^2 \)  
(Equation 22)

\( v_{21} = S_{dephosphos} \cdot [ppS2 \cdot ppS3 \cdot ppS3] \)  
(Equation 23)

\( v_{22} = 2 \cdot S_{dephosphos} \cdot [ppS2 \cdot ppS3 \cdot ppS3] \)  
(Equation 24)

\( v_{23} = [S4] \cdot k_{on,334} \cdot [ppS3]^2 \)  
(Equation 25)

\( v_{24} = 2 \cdot S_{dephosphos} \cdot [ppS3 \cdot ppS3 \cdot S4] \)  
(Equation 26)

\( v_{25} = [S4]^2 \cdot k_{on,244} \cdot [ppS2] \)  
(Equation 27)

\( v_{26} = S_{dephosphos} \cdot [ppS2 \cdot S4 \cdot S4] \)  
(Equation 28)

\( v_{27} = [S4]^2 \cdot k_{on,344} \cdot [ppS3] \)  
(Equation 29)

\( v_{28} = S_{dephosphos} \cdot [ppS3 \cdot S4 \cdot S4] \)  
(Equation 30)

\( v_{29} = [S4] \cdot k_{on,234} \cdot [ppS2] \cdot [ppS3] \)  
(Equation 31)

\( v_{30} = S_{dephosphos} \cdot [ppS2 \cdot ppS3 \cdot S4] \)  
(Equation 32)

\( v_{31} = S_{dephosphos} \cdot [ppS2 \cdot ppS3 \cdot S4] \)  
(Equation 33)

For the model extension, the genes are linked to the complexes with the following reactions:

\( v_{32} = \frac{\text{gene\_turn} + \text{gene\_act1} \cdot [ppS2 \cdot ppS3 \cdot ppS3] + \text{gene\_act2} \cdot [ppS2 \cdot S4 \cdot S4] + \text{gene\_act3} \cdot [ppS2 \cdot ppS3 \cdot S4]}{\text{gene\_inh1} \cdot [ppS2 \cdot ppS3 \cdot ppS3] + \text{gene\_inh2} \cdot [ppS2 \cdot S4 \cdot S4] + \text{gene\_inh3} \cdot [ppS2 \cdot ppS3 \cdot S4] + 1} \)  
(Equation 34)

\( v_{33} = [\text{gene}] \cdot \text{gene\_turn} \)  
(Equation 35)

with “gene” representing SKI, SKIL, DNMT3A, SOX4, JUN, SMAD7, KLF10, BMP4, CXCL15, DUSP5, TGFA and PDK4 and “gene\_turn” describing the gene-specific and TGF\( \beta \)-independent gene turnover.

**ODE System of the Comprehensive Mathematical Model**

The ODE system of the comprehensive model determining the time evolution of the dynamical variables is given by:

\[ d[TGFb]/dt = -v_1 \]  
(Equation 36)

\[ d[Rec]/dt = -v_1 \]  
(Equation 37)

\[ d[TGFb\cdot pRec]/dt = +v_1 - v_2 \]  
(Equation 38)
system was solved by a parallelized implementation of the CVODES algorithm (Hindmarsh et al., 2005). It also supplies the parameter sensitivities utilized for parameter estimation.

Identification of the Occurring Smad Complex Using L1 Regularization

In the context of ODE models of signaling pathways, L1 regularization was applied to determine the cell-type specific parameters (Merkle et al., 2016) and was described in detail in the setting of ODE models (Steiert et al., 2016).

with “gene” representing $S_{KI}$, $SKIL$, $DNMT3A$, $SOX4$, $JUN$, $SMAD7$, $KLF10$, $BMP4$, $CXCL15$, $DUSP5$, $TGFA$ and $PDK4$. The ODE system was solved by a parallelized implementation of the CVODES algorithm (Hindmarsh et al., 2005). It also supplies the parameter sensitivities utilized for parameter estimation.

Identification of the Occurring Smad Complex Using L1 Regularization

$L_1$ regularization is a general methodology to establish models with a minimal number of parameters, i.e. to reduce the complexity of models down to a level which is required to explain the data. In the context of ODE models of signaling pathways, $L_1$ regularization was applied to determine the cell-type specific parameters (Merkle et al., 2016) and was described in detail in the setting of ODE models (Steiert et al., 2016).

\[
\frac{d[S2]}{dt} = -v_9 + v_{11} \quad \text{(Equation 39)}
\]

\[
\frac{d[S3]}{dt} = -v_{12} + v_{14} \quad \text{(Equation 40)}
\]

\[
\frac{d[S4]}{dt} = -3 \cdot v_7 + v_9 + v_{10} - 2 \cdot v_{15} + v_{16} + 2 \cdot v_{21} + v_{22} - 2 \cdot v_{23} + v_{24} - 2 \cdot v_{27} - v_{28} - v_{30} + v_{31} \quad \text{(Equation 41)}
\]

\[
\frac{d[ppS2\_ppS2\_ppS2]}{dt} = +v_3 - v_4 \quad \text{(Equation 42)}
\]

\[
\frac{d[ppS3\_ppS3\_ppS3]}{dt} = +v_5 - v_6 \quad \text{(Equation 43)}
\]

\[
\frac{d[S4\_S4\_S4]}{dt} = +v_7 - v_8 \quad \text{(Equation 44)}
\]

\[
\frac{d[pS2]}{dt} = +v_4 + v_{10} - v_{11} + v_{16} + v_{19} + v_{21} + v_{26} + v_{30} \quad \text{(Equation 45)}
\]

\[
\frac{d[pS3]}{dt} = +v_6 + v_{13} - v_{14} + v_{17} + v_{22} + v_{24} + v_{28} + v_{31} \quad \text{(Equation 46)}
\]

\[
\frac{d[ppS2\_ppS2\_ppS2]}{dt} = 3 \cdot v_7 + v_2 + v_9 - v_{10} - 2 \cdot v_{15} + v_{16} + 2 \cdot v_{17} + v_{18} + v_{20} - v_{22} - v_{25} - v_{29} + v_{31} \quad \text{(Equation 47)}
\]

\[
\frac{d[ppS3]\_d}{dt} = 3 \cdot v_5 + 2 \cdot v_6 + v_{12} - v_{13} - v_{15} + v_{16} - 2 \cdot v_{20} + 2 \cdot v_{21} + v_{22} - 2 \cdot v_{23} + v_{24} - v_{27} - v_{29} + v_{30} \quad \text{(Equation 48)}
\]

\[
\frac{d[ppS2\_ppS2\_S4]}{dt} = +v_8 - v_{19} \quad \text{(Equation 49)}
\]

\[
\frac{d[ppS2\_ppS2\_ppS3]}{dt} = +v_5 - v_{16} - v_{17} \quad \text{(Equation 50)}
\]

\[
\frac{d[ppS2\_ppS3\_ppS3]}{dt} = +v_{20} - v_{21} - v_{22} \quad \text{(Equation 51)}
\]

\[
\frac{d[ppS3\_ppS3\_S4]}{dt} = +v_{23} - v_{24} \quad \text{(Equation 52)}
\]

\[
\frac{d[ppS2\_ppS3\_S4]}{dt} = +v_{29} - v_{30} - v_{31} \quad \text{(Equation 53)}
\]

\[
\frac{d[ppS3\_S4\_S4]}{dt} = +v_{27} - v_{28} \quad \text{(Equation 54)}
\]

\[
\frac{d[ppS2\_S4\_S4]}{dt} = +v_{25} - v_{26} \quad \text{(Equation 55)}
\]

\[
\frac{d[\text{gene}]}{dt} = +v_{32} - v_{33} \quad \text{(Equation 56)}
\]
The idea of L1 regularization is to minimize an objective function $\chi^2_{\text{pen}} = \chi^2_{\text{data}} + \lambda \sum_j |\theta_j|^2$ assessing agreement of the data $y_i$, $i=1,…,n_{\text{data}}$, with the model $g_i$ and a second term $\chi^2_{\text{nl}} = \lambda \Sigma_j |\theta_j|$, penalizing parameters $\theta_j$, $j=1,…,n_{\text{param}}$, which are different from zero. Since the penalty terms in $\chi^2_{\text{nl}}$ have a non-vanishing gradient ±1 for all values unequal to zero, parameters which improve $\chi^2_{\text{data}}$ less than 1/\lambda are estimated equals to zero.

In our context, L1 regularization is applied to identify the complexes which are required to describe the coIP data. For these complexes, $\chi^2_{\text{pen}}$ is optimal for non-vanishing association rates. For complexes which are not required, the penalty $\chi^2_{\text{nl}}$ dominates the data contribution $\chi^2_{\text{data}}$ and therefore the association rates are estimated to zero.

**Model Prediction and Experimental Validation of mRNA Half-lives**

mRNA half-lives were calculated based on the gene-specific turnover parameters, with half-life = ln(2)/turnover. The mRNA turnover of each gene was classified as fast (half-life < 10 minutes), intermediate (half-life between 10 and 100 minutes) or slow (half-life > 100 minutes). The confidence interval of each gene-specific turnover parameter was determined by the profile likelihood method (Raue et al., 2009). To experimentally determine mRNA stability, Hepa1-6 cells were cultivated and growth factor depleted as described above and were stimulated with 1 ng/ml TGFβ for 2 hours followed by treatment with 1 μg/ml Actinomycin D to inhibit transcription. Total RNA was extracted at specific time points and was analyzed using qRT-PCR. The mRNA half-life was estimated by fitting the mRNA expression to a 3-parameter exponential decay function: $y = y_0 + a \exp(-b \cdot x)$. mRNA half-lives were calculated as: half-life = ln(2)/b. Confidence intervals were calculated based on the standard errors of the estimates.

**Gradual Knock-down of Smad Proteins**

0.5x10^6 Hepa1-6 cells were cultivated in DMEM (Gibco) supplemented with 10% (v/v) FBS (Life Technologies) and 1% 200 mM glutamine (Gibco) for 24 hours and siRNA transfection was performed according to the Lipofectamine RNAiMax protocol (Invitrogen, Cat. #13778-150). SMARTpool siRNA against Smad2 (L-040707-00-0005), Smad3 (L-040706-00-0005) and Smad4 (L-040687-00-0005) was obtained from Dharmacon. ON-TARGETplus Non-targeting pool (D-001810-10-20) served as siRNA control. After 24 hours, medium was exchanged with fresh medium containing 1% 100x penicillin/streptomycin (Gibco) for another 24 hours. Growth factor depletion was performed as described above and were stimulated with 1 ng/ml TGFβ for 2 hours followed by treatment with 1 μg/ml Actinomycin D to inhibit transcription. Total RNA was extracted at specific time points and was analyzed using qRT-PCR. The mRNA half-life was estimated by fitting the mRNA expression to a 3-parameter exponential decay function: $y = y_0 + a \exp(-b \cdot x)$. mRNA half-lives were calculated as: half-life = ln(2)/b. Confidence intervals were calculated based on the standard errors of the estimates.

**Overexpression of Smad Proteins**

Mouse Smad2, -Smad3 and -Smad4 inserts were obtained from total RNA of primary mouse hepatocytes. The RNA was reverse-transcribed into cDNA. The insert was re-cloned into the retroviral expression vector pMOWS-Flag-MCS using PacI and EcoRI restriction sites. Transfection of Phoenix eco packaging cell line was performed by using calcium phosphate precipitation. Transducing supernatants were generated 24 hours after transfection by passing through a 0.45 μm filter, supplemented with 8 μg/ml polybrene (Sigma). Stably transduced Hepa1-6 cells were selected in the presence of 1 μg/ml puromycin (Sigma) 24 hours after transduction.

**Transcriptional Activity of Smad Complexes**

The mathematical model was utilized to assess the gene regulatory impacts of the individual Smad complexes ppSmad2: ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4 and ppSmad2:ppSmad3:Smad4. For this purpose, the transcriptional activities of all complexes except a single complex of interest c were virtually prohibited in the mathematical model by setting the respective activation- and inhibition parameters to zero. For quantitative assessment, the areas under the curves

$$AUC_{cg}(p) = \frac{1}{t_{\text{max}}} \int_{t=0}^{t_{\text{max}}} \log_2(x_c(t,p)) / x_g(0,p) \, dt$$

(Equation 57)

of the log2-ratios of the concentrations $x_c(t,p)$ after TGFβ stimulation relative to the steady state expression at time point t=0 were calculated for all genes $g$, each individual complex, and a given parameter vector $p$. Negative $AUC_{cg}$ indicates negative regulation of the respective Smad complex c on gene g, a positive $AUC_{cg}$ is obtained for positive regulators. To translate uncertainties in the estimated parameters $p$ to $AUC_{cg}(p)$, the analysis was repeated for all statistically valid parameters obtained for the profile likelihood.

**Analysis of Hepatocellular Carcinoma Samples**

Each liver tissue piece (Cohort A) was cut into 20-30 mg pieces and homogenized in total cell lysis buffer for protein analysis or in lysis buffer RA1 (Macheroy & Nagel) for RNA isolation in a Precellys 24 homogenizer (VWR Life Science). Protein lysates were rotated for 30 minutes at 4°C, sonicated and centrifuged for 10 minutes at 20 800 x g and 4°C. Supernatants were subjected to IPs with anti-Smad2/3 (BD-610843) or anti-Smad3 (Cell Signaling #9523) antibodies, respectively. Lysates were rotated overnight at 4°C and analyzed by quantitative IB. Protein signals were determined as described before and normalized against the respective calibrator signal to...
compare samples from different IBs. Subsequently, the mean of the tumor-free samples was set to the absolute value determined for the molecules per cell of Smad2, Smad3 and Smad4 in primary human hepatocytes. Accordingly, the relative signal intensities were translated into molecules per cell values for each sample. RNA was isolated from the homogenized samples according to the manufacturers protocol (Macheroy & Nagel). RNA integrity number (RIN) was measured to assess the quality of the isolated RNA revealing a RIN value of 7-9 for all samples. Samples were further subjected to qRT-PCR analysis as described above. mRNA data was normalized against the geometric mean of TBP and UBE2R2.

Phosphorylation of Smad2 was detected in human tissue samples from Cohort B. Samples were homogenized with a plastic pestle and were processed as described before. Lysates of tissue samples (1000 μg protein) as well as of reference samples from unstimulated or stimulated (1 ng/ml TGFβ) for 60 minutes) HepG2 cells (100 μg protein) were subjected to IP experiments with an anti-Smad2 (Cell Signaling #5339) antibody and were supplemented with Protein A sepharose (GE healthcare 17-0963-03). IPs were rotated overnight at 4°C and were analyzed by quantitative IB. Protein signals were determined as described above and normalized against the TGFβ-stimulated reference sample from the HepG2 cell line.

**Prediction of Complexes and Total Amounts**

The integrative dynamic model for TGFβ-induced Smad complex formation and the subsequent effect on target genes was used to predict the regulation at the level of the Smad complexes from observed gene expression level in patients suffering from hepatocellular carcinoma. As a first step in this analysis, the steady state concentrations of the genes and the complexes were calculated for the parameters fitted for Hepa1-6. For this purpose, a receptor activity of 10% relative to the estimated maximal activity after the treatment of Hepa1-6 with 1 ng/ml TGFβ was assumed. Then, the observed fold-changes at the gene expression level for an individual patient relative to the average over all tumor-free samples were added to the gene expression level in the mathematical model. Next, three parameters for concentration fold-changes of ppSmad2:ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4 and ppSmad2:ppSmad3:Smad4 were introduced and estimated for a single patient while keeping all other parameters fixed. For this step, only the part of the model linking the complexes to the genes was required. Structural identifiability was checked using the profile likelihood approach (Raue et al., 2009). In addition, a weak prior log10 fold ~ N(μ,σ²) with μ=0 and σ²=4 was used to decrease the variability of the estimates in the case of weakly informative expression data. In analogy to the analysis for the Smad complexes, the observed regulation at the gene expression level was also used to predict regulation at the level of the total Smad concentrations. Again, the steady states for the gene expression were calculated for the model fitted for Hepa1-6 and by assuming 10% of the maximal experimentally observed receptor activity for the cell line. Three fold-parameters S2fold, S3fold, S4fold were introduced representing the alteration in Smad2, Smad3, and Smad4 abundances. Then, the measured fold-changes for the gene expression in tumor-free and tumor tissue in individual patients relative to the average over all tumor samples were added to the steady state levels of the model. For these changes at the gene expression level, the corresponding four parameters were then estimated to predict altered activation levels of TGFβ receptors and fold-changes of Smad2, Smad3, and Smad4.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

Microarray expression data was considered significantly regulated if p<0.01 as tested by a two-factorial linear model and if they showed an at least 1.5-fold increase compared to untreated controls. The dynamics of gene expression was estimated based on a mathematical function approximating the trajectories for all genes. Transcripts that were not constant over time (higher fold-change than 1.5) in the untreated control were discarded and transcripts were considered as significantly regulated if p<0.01 (two-factorial linear model) and if they showed an at least 1.5-fold increase compared to untreated controls.

qRT-PCR data and predicted Smad complexes from paired tumor-free and tumor tissue (n=29) were expressed as log2 values with the average of the tumor-free value set to zero. Here, significance was tested by paired two-sided t-tests, with significances defined as *, p<0.05; **, p<0.01; ***, p<0.001. The predicted and measured sum of Smad2, Smad3 and Smad4 from paired tumor-free and tumor tissue (n=29) were expressed as molecules/cell with the average of the tumor-free value set to the measured values in primary human hepatocytes. Significance was tested by paired two-sided t-tests, with significances defined as *, p<0.05; **, p<0.01. Significance of IB data of Smad2 phosphorylation amounts (n=12) was tested by paired two-sided t-tests, with significances defined as *, p<0.05; **, p<0.01; ***, p<0.001.

**DATA AND SOFTWARE AVAILABILITY**

The modeling framework, the mathematical model and the data sets are open source and are available here: http://www.data2dynamics.org.

The microarray expression data were deposited in the Gene Expression Omnibus (GEO) database under the accession number GEO: GSE90954: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90954.