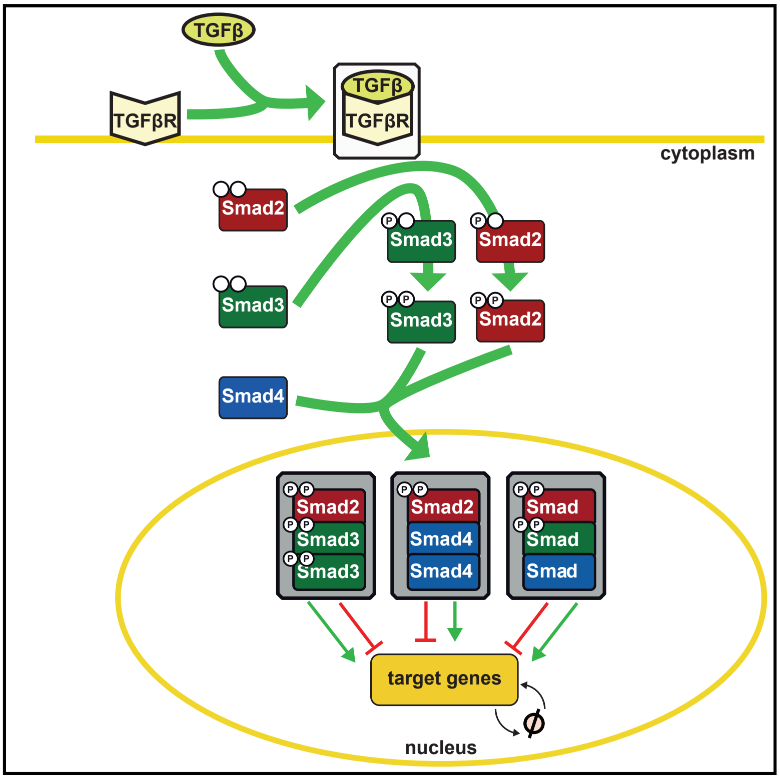
Cell Systems

Article

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

Graphical Abstract



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

Authors

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

Correspondence

u.klingmueller@dkfz.de

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.

Lucarelli et al., 2018, Cell Systems *6*, 1–15
January 24, 2018 © 2017 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2017.11.010

Cell Systems
Article

OPEN ACCESS
CellPress

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,[2,3,13] Artyom Vlasov,[1] Martin E. Boehm,[1,4] Nao Iwamoto,[1] Bernhard Steiert,[2,3] Susen Lattermann,[1] Marvin Wäsch,[1] Markus Stepath,[1] Matthias S. Matter,[5] Mathias Heikenwälder,[6] Katrin Hoffmann,[7] Daniela Deharde,[8] Georg Damm,[8,9] Daniel Seehofer,[8,9] Maria Muciek,[10] Norbert Gretz,[10] Wolf D. Lehmann,[4] Jens Timmer,[2,11] and Ursula Klingmüller[1,12,14,*]

[1]Division Systems Biology of Signal Transduction, German Cancer Research Center (DKFZ), INF 280, 69120 Heidelberg, Germany
[2]Institute of Physics, University of Freiburg, 79104 Freiburg, Germany
[3]Center for Biological Systems Analysis, University of Freiburg, 79104 Freiburg, Germany
[4]Molecular Structure Analysis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[5]Institute of Pathology, University of Basel, 4003 Basel, Switzerland
[6]Division Chronic Inflammation and Cancer, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[7]Department of General and Transplantation Surgery, Ruprecht Karls University Heidelberg, 69120 Heidelberg, Germany
[8]Department of General-, Visceral- and Transplantation Surgery, Charité University Medicine Berlin, 13353 Berlin, Germany
[9]Department of Hepatobiliary Surgery and Visceral Transplantation, University of Leipzig, 04103 Leipzig, Germany
[10]Medical Research Center, Medical Faculty Mannheim, University of Heidelberg, 68167 Mannheim, Germany
[11]BIOSS Centre for Biological Signalling Studies, University of Freiburg, 79104 Freiburg, Germany
[12]Translational Lung Research Center (TLRC), Member of the German Center for Lung Research (DZL), 69120 Heidelberg, Germany
[13]These authors contributed equally
[14]Lead 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 and regulate transcription of hundreds of genes (Moustakas and Heldin, 2002). Smad2 and Smad3 are regulated by phosphorylation, but not Smad4. Based on the three Smad proteins with three phosphorylation states for Smad2 and Smad3 ($n = 7$) and the trimeric complexes ($k = 3$), the theoretical number of different Smad complexes can be calculated according to the formula for unordered sampling with replacement:

$$\binom{n+k-1}{k} = 84 \qquad \text{(Equation 1)}$$

Thus, in principle, 84 different trimeric Smad complexes can form, but it has not been determined which Smad complexes indeed occur in particular cell types such as hepatocytes.

Depending on the cell type, only distinct sets of target genes are induced by TGF-β. This could be mediated by the interaction of Smad complexes with 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

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

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 (Schmierer 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 L_1 regularization (Merkle et al., 2016; Steiert et al., 2016). The L_1 regularization approach can be employed to investigate which model reactions are required to describe experimental data. Therefore the L_1 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 (coIP) 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,

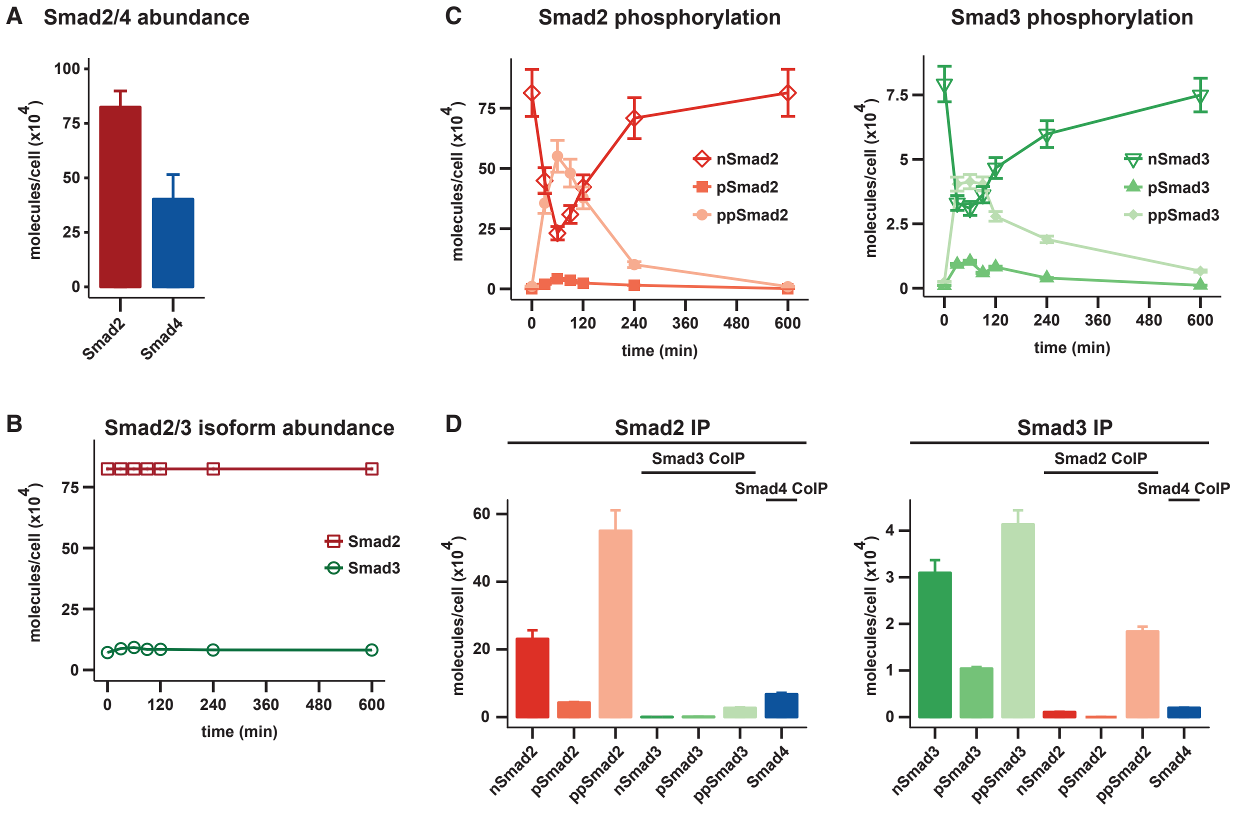


Figure 1. Abundance and Dynamics of TGF-β-Induced Smad Pathway Components

(A) Abundance of Smad2 and Smad4 proteins in Hepa1-6 cells determined by quantitative immunoblotting (IB). Error bars represent SEM (n = 18).

(B) Hepa1-6 cells were stimulated with 1 ng/mL TGF-β and relative protein abundance of Smad2 to Smad3 was determined by mass spectrometry (n = 2).

(C) Whole-cell lysates of Hepa1-6 cells stimulated with 1 ng/mL TGF-β were subjected to IP with anti-Smad2/3 antibodies (n = 3 for Smad2 and n = 2 for Smad3) and analyzed by mass spectrometry for absolute phosphorylation levels of Smad2 and Smad3. n, non-phosphorylated; p, singly phosphorylated; pp, doubly phosphorylated. Error bars represent 5% error from mass spectrometry measurement.

(D) Whole-cell lysates of Hepa1-6 cells stimulated with 1 ng/mL TGF-β for 60 min were used for IP with anti-Smad2 or anti-Smad3 antibodies. Amounts of coIP nSmad2/3, pSmad2/3, ppSmad2/3, and Smad4 are shown. Smad2, Smad3, and Smad4 protein abundance was determined by mass spectrometry (n = 3 for Smad2 and n = 2 for Smad3 and Smad4). Error bars represent 5% error from mass spectrometry measurement.

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 (k_{on}). 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 L_1 regularization term to the k_{on} parameters favoring a minimal number of distinct complexes in the mathematical model. For statistical assessment, we calculated the profile likelihoods for the ten k_{on} 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 k_{on} parameters (Figure 2B), the best parameter estimation value (red asterisk) was compatible with 10^{-14}, which is equivalent to zero. For the other three k_{on} 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

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

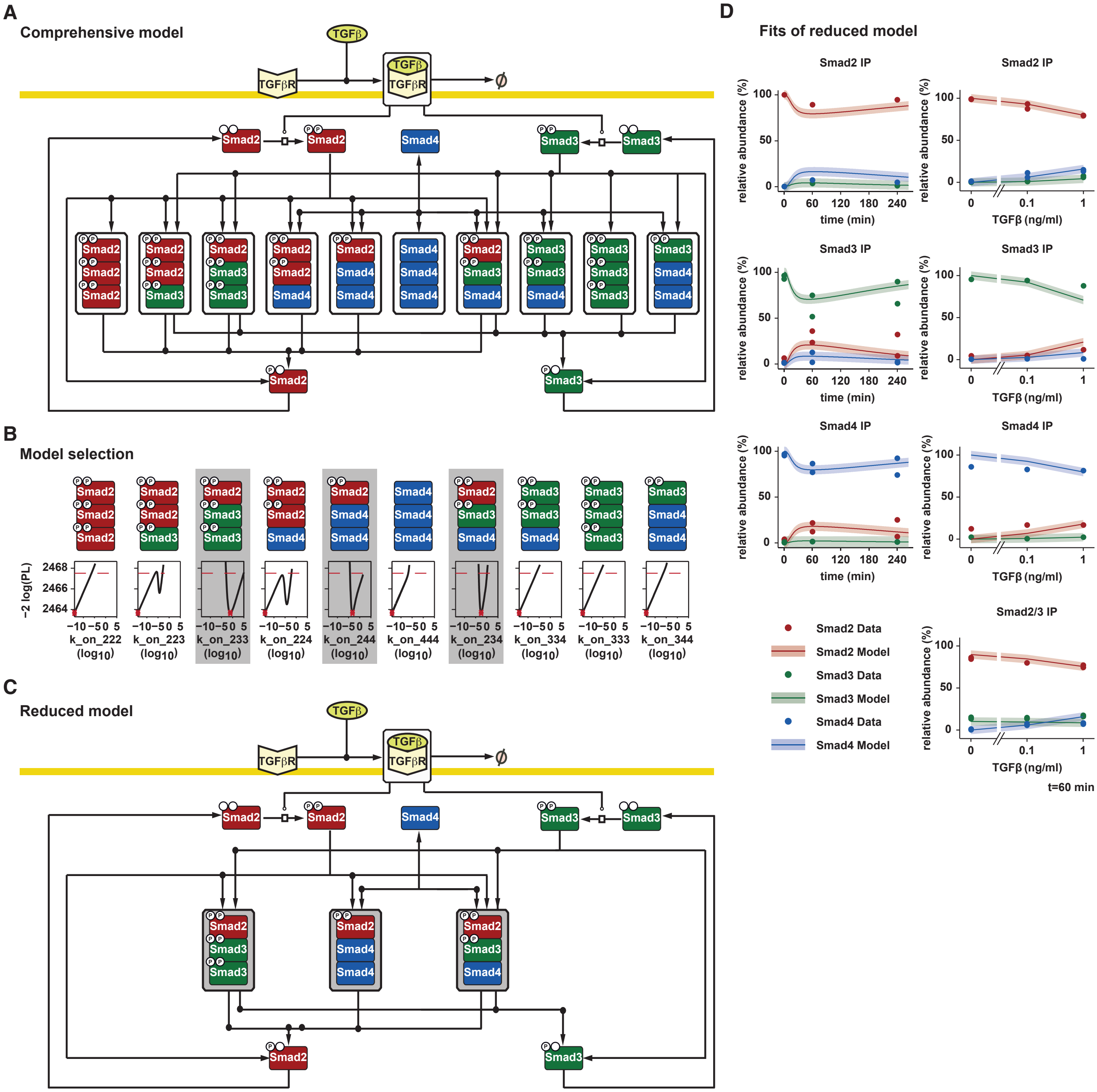


Figure 2. Data-Driven Model Reduction Identifies Relevant Smad Complexes

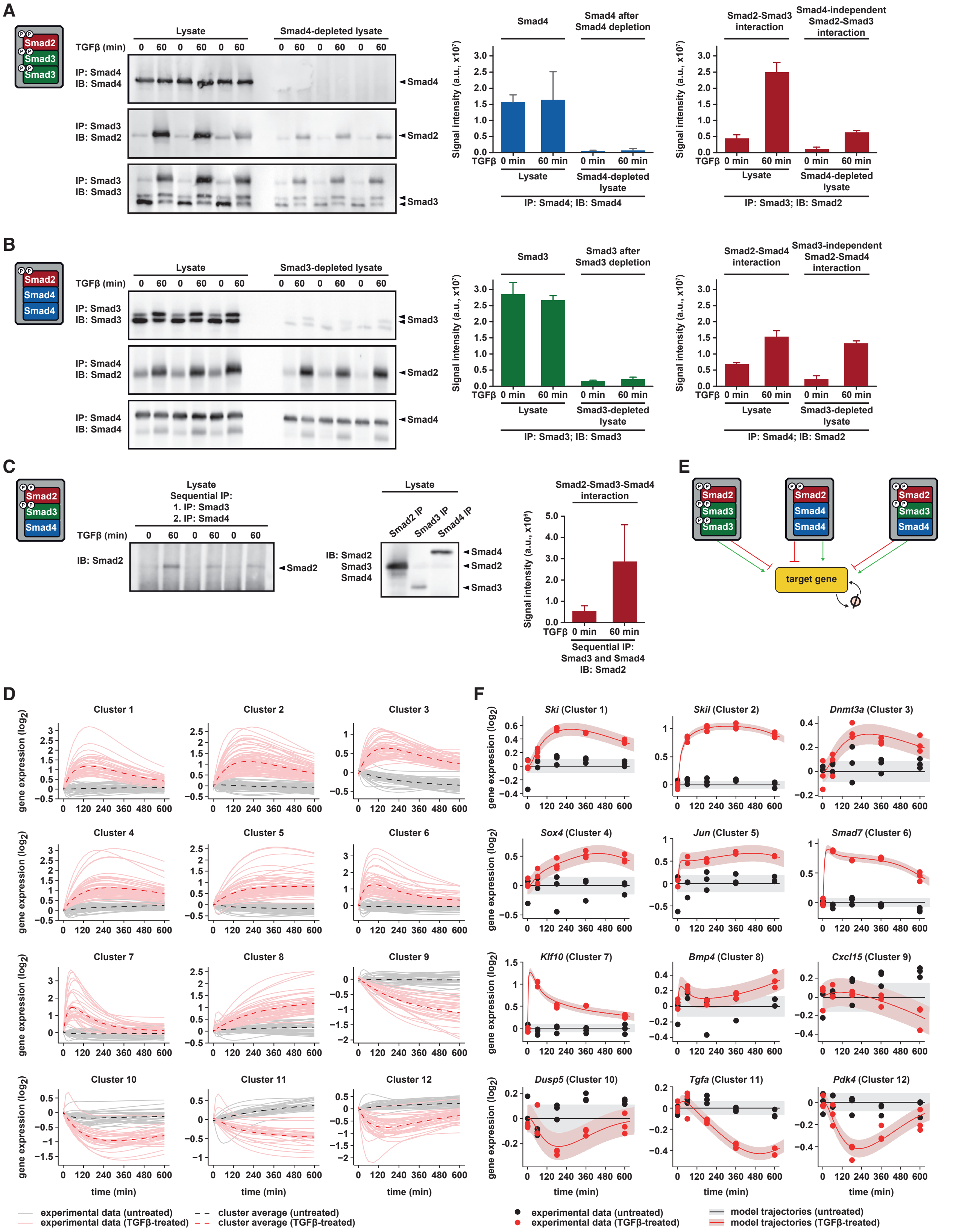
(A) A mathematical model describes the formation of ten different homo- and heterotrimers comprising ppSmad2, ppSmad3, and Smad4.

(B) Employment of model reduction by L_1 regularization to identify the relevant Smad complexes required to explain the experimental data. The black curves indicate the profile likelihood for the association rate (k_{on}) of a specific complex. For three complexes (gray background) the profile likelihood, −2 log(PL), increases above the statistical threshold (red dashed line) if the association rate is deviating from the estimated value (red asterisk).

(C) Structure of the reduced model.

(D) Description of the experimental data by the reduced model. Left panels: Smad2 (n = 1), Smad3 (n = 2), and Smad4 (n = 2). Right panels: Smad2 (n = 2), Smad3 (n = 1), Smad4 (n = 1), and Smad2/3 (n = 2). Dots, experimental mass spectrometric data; continuous line, model trajectories; shading, 5% error.

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 $\chi^2 = \Sigma_i((y_i - f_i)/\sigma_i)^2$, 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.



(legend on next page)

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

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 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:Smad4: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 Smad4 IP, and the detection of Smad2 by IB confirmed the coIP 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, *Cxcl15* with its human ortholog *CXCL8*/IL8 (Ge et al., 2010) for cluster 9, *Dusp5* (Tao et al., 2016) for cluster 10, *Tgfa* (Nozato et al., 2003) for cluster 11, and *Pdk4* (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.

(A) 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.

(B) 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.

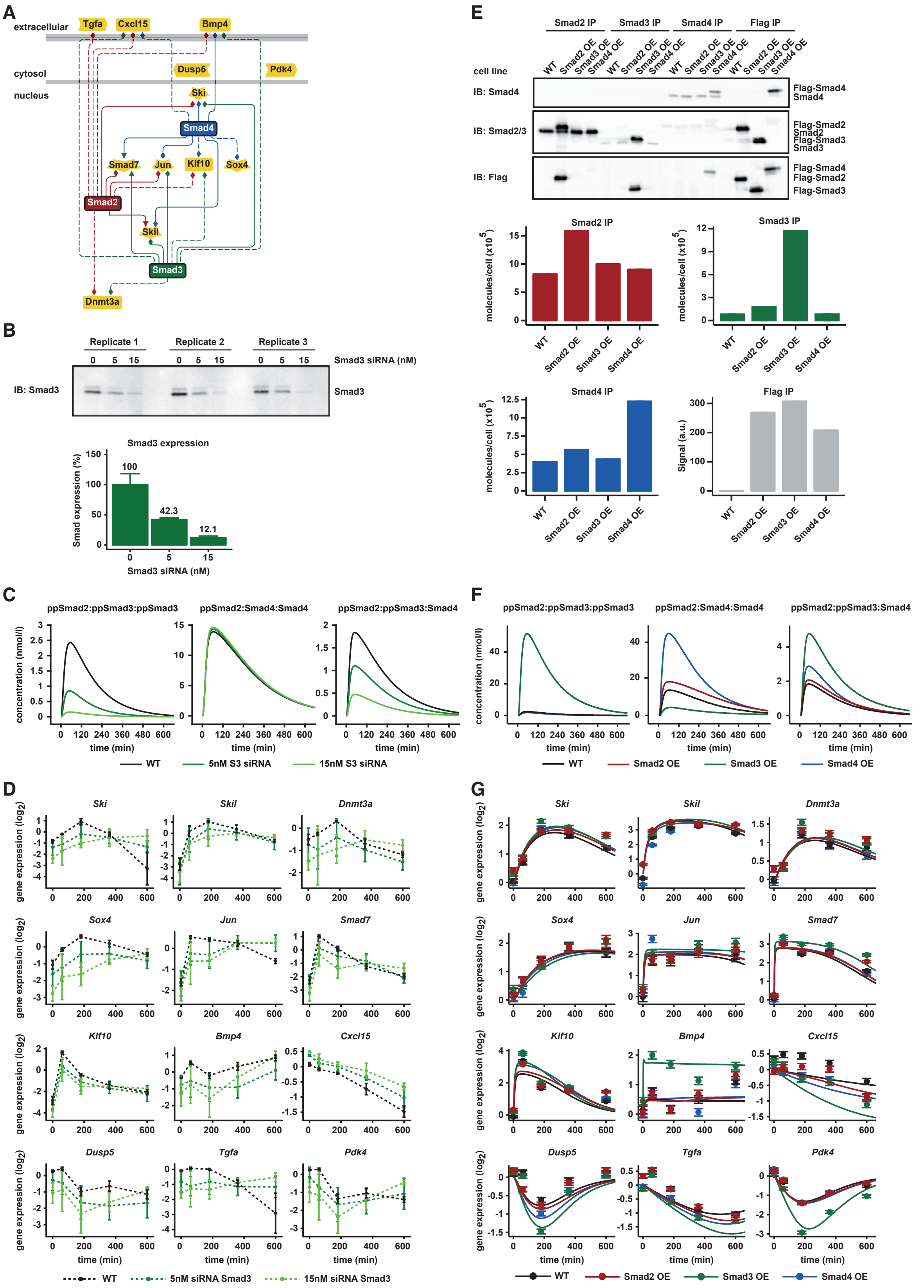
(C) Lysates were used for Smad3 IP and proteins were dissociated from beads. The supernatants were subjected to Smad4 IP and the associated Smad2 was detected by IB. IPs were confirmed by IB. Experiments were performed in biological triplicates and means and SD are shown.

(D) 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.

(E) 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.

(F) 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.

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010



(legend on next page)

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

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 knowledge on the TGF-β-induced dynamics of the ppSmad2:ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4, and ppSmad2: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:Smad4: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:Smad4: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 upregulated upon Smad3 knockdown (e.g., *Cxcl15, 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

Figure 4. Influence of Smad Protein Abundance on the Dynamics of Smad Complexes and TGF-β-Induced Gene Expression

(A) Analysis of the connection between Smad2 (red), Smad3 (green), Smad4 (blue), and the 12 selected TGF-β target genes (yellow) by the Genomatix Pathway System (GePS). Dashed lines, experimentally validated; solid lines, expert-curated connections; rounded rectangles, proteins; right chevrons, kinases; left chevrons, phosphatases; stars, co-factors; arrows, activation; diamonds, Smad binding sites.

(B) Smad3 protein was downregulated using two different concentrations of target siRNA. Lysates of Hepa1-6 were subjected to Smad3 IP and IB. Experiments were performed in biological triplicates and means and SD are shown.

(C) Model simulations of the dynamics of complex abundance after Smad3 knockdown. Continuous lines, model simulations; WT, Hepa1-6 wild-type.

(D) TGF-β-induced gene expression after Smad3 knockdown in Hepa1-6 cells determined by qRT-PCR. Experiments were performed in biological triplicates and means and SD are shown.

(E) Overexpression (OE) of FLAG-tagged Smad2, Smad3, and Smad4 proteins in Hepa1-6 cells analyzed by IB.

(F) Model simulations of the dynamics of complexes upon overexpression of different Smad molecules.

(G) Analysis of TGF-β-induced expression of the selected target genes in Smad-overexpressing Hepa1-6 cells by qRT-PCR (n = 3). The error bars represent the SE resulting from scaling of the experimental data by a mixed effects alignment model.

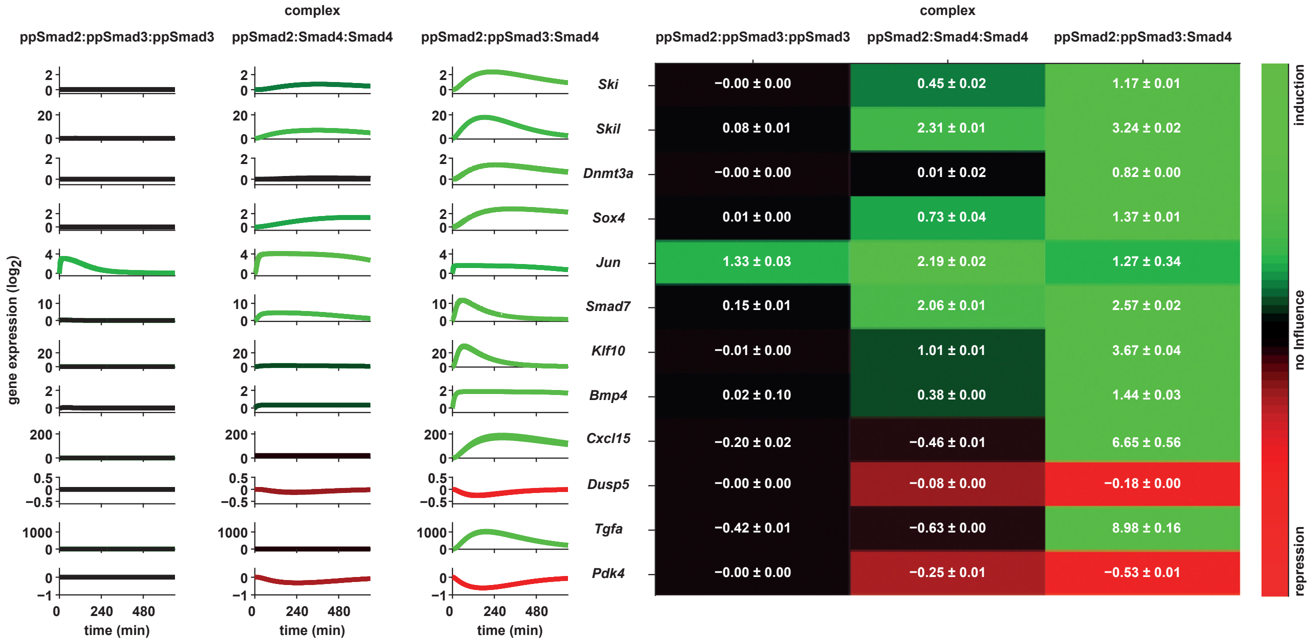


Figure 5. Mathematical Model-Based Determination of the Impact of Single Smad Complexes on Gene Expression

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 (log$_2$), with median and SD. The color code of both panels was normalized for each gene individually.

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.

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

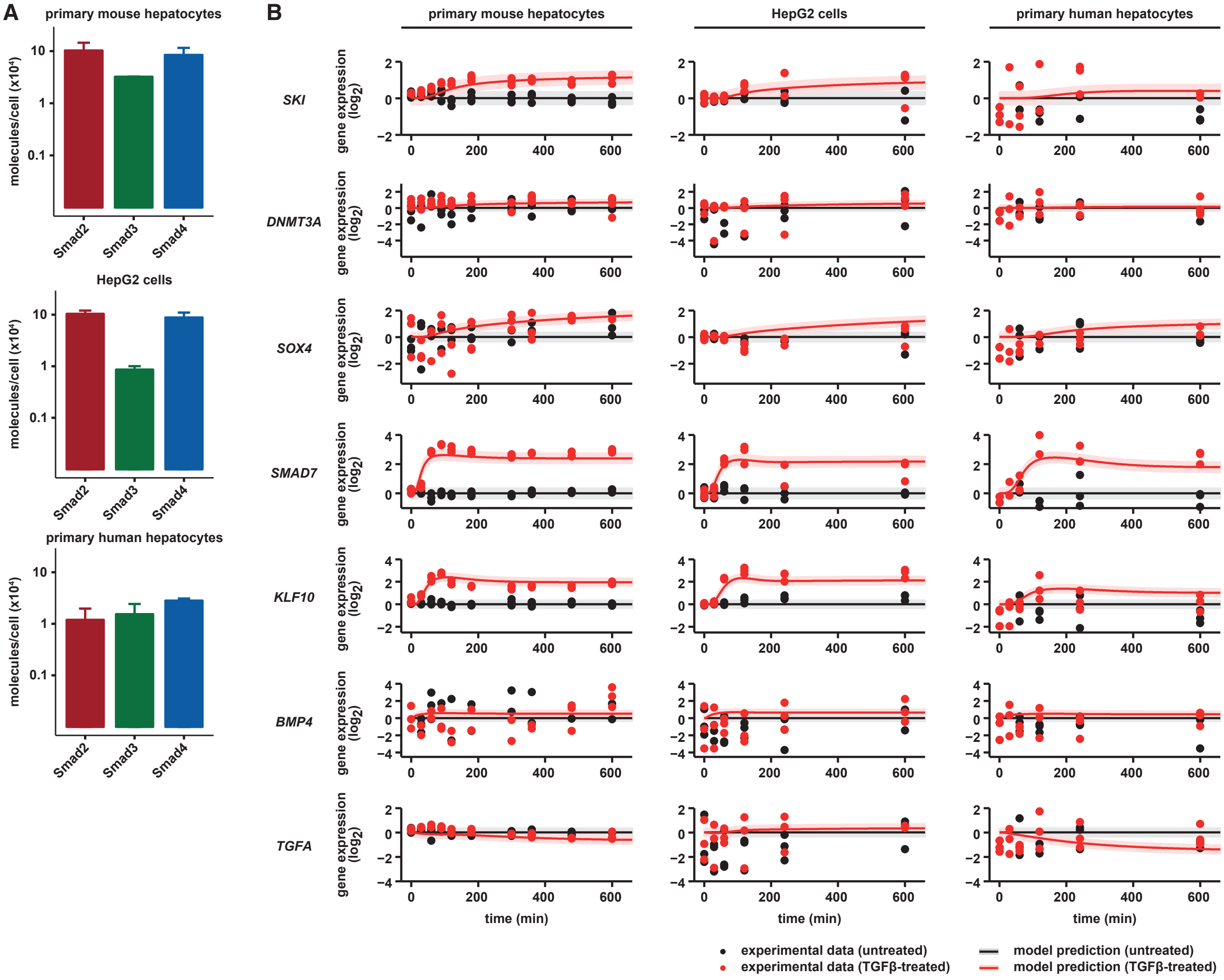


Figure 6. Conservation of Connection of Smad Complexes and TGF-β-Induced Target Genes in Primary Hepatocytes and Hepatoma Cells

(A) Smad protein abundance in primary mouse hepatocytes, HepG2 cells, and primary human hepatocytes determined by mass spectrometry and IB. Error bars represent SEM (n = 4).

(B) Model simulations of TGF-β-induced gene expression in primary mouse hepatocytes, HepG2 cells, and primary human hepatocytes. Smad2, Smad3, and Smad4 protein abundance was incorporated in the model. Conserved model parameters linking the three Smad complexes to gene expression were assumed and cell type-specific TGF-β signaling dynamics were estimated based on the qRT-PCR data (dots, n = 3). Continuous lines, model simulations; shaded area, estimated error.

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 Hepa1-6 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:ppSmad3, ppSmad2:Smad4:Smad4, and ppSmad2:ppSmad3:Smad4 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

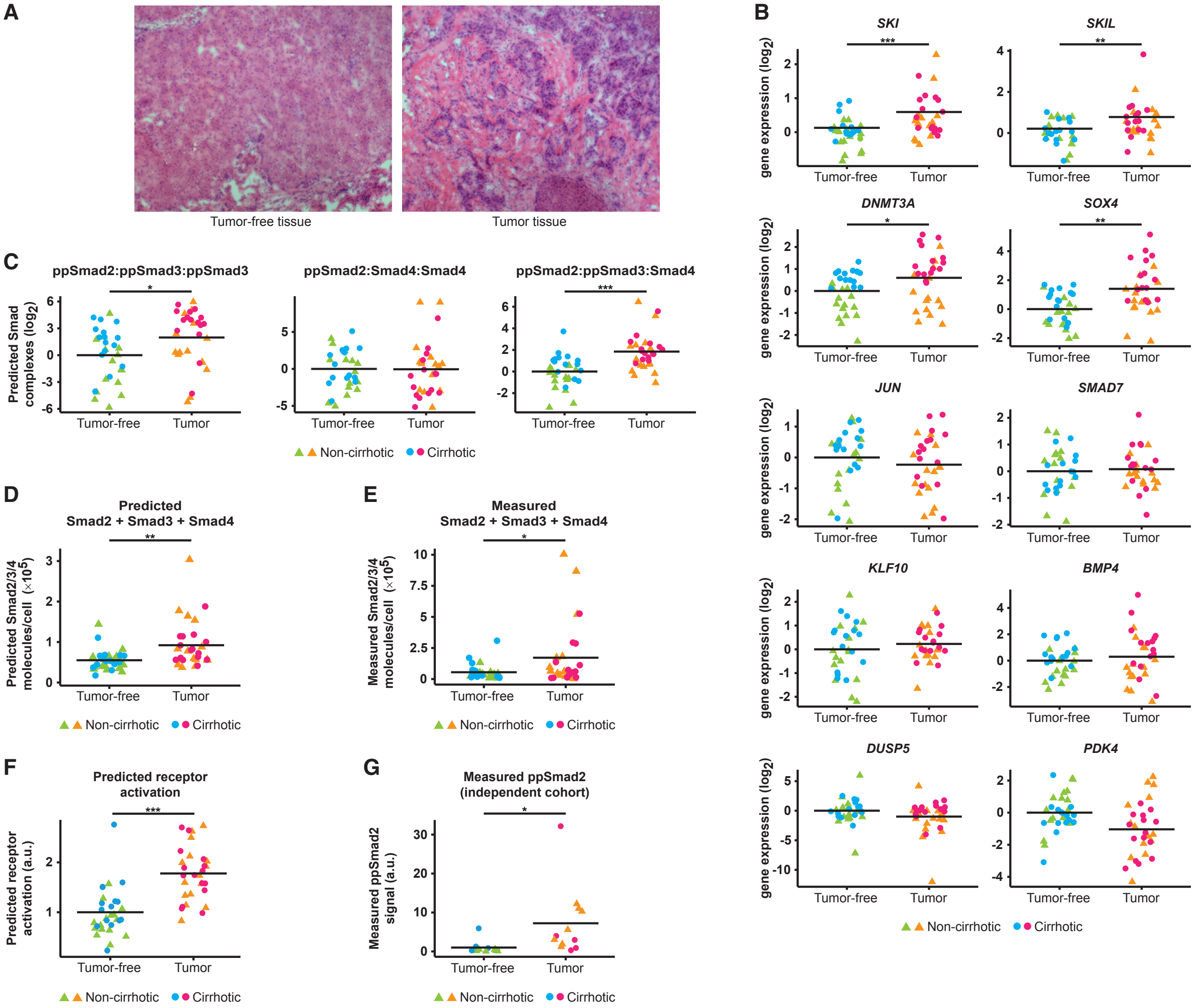


Figure 7. Prediction of Changes in Smad Abundance Based on Gene Expression in HCC Tissue Samples

(A) Histological images (magnification, ×100) of tumor-free (left) and tumor tissue (right) from a patient with cirrhosis who developed HCC.

(B) Expression of TGF-β target genes in tumor-free and tumor tissue samples determined by qRT-PCR ($n = 30$).

(C) Model-based prediction of the relative amounts of the Smad complexes based on target gene expression shown in (B) ($n = 30$).

(D) Prediction of amounts of Smad2, Smad3, and Smad4 based on the abundance of the Smad complexes shown in (C.) ($n = 30$).

(E) Determination of the amount of Smad2, Smad3, and Smad4 for each patient by quantitative IB. Mean signal of tumor-free samples was adjusted to molecules per cell measurements in primary human hepatocytes ($n = 29$).

(F) Prediction of the amount of TGF-β receptor phosphorylation by the mathematical model based on the Smad complexes shown in (C) ($n = 30$).

(G) Determination of the amount of phosphorylated Smad2 by IB for each patient in an independent cohort B (Figure S7E). Different blots were scaled to each other relative to the amount of phosphorylated Smad2 in TGF-β-treated HepG2 cells (Figure S7E) ($n = 12$).

All data are shown relative to the mean of the tumor-free samples. Horizontal lines indicate mean values. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; paired t tests.

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., 2001b). 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 L_1 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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Culture of Cell Lines and Primary Cells
  - ○ Hepatocellular Carcinoma and Tumor-free Tissue Samples
- METHOD DETAILS
  - ○ Stimulation, Lysis and SDS-PAGE
  - ○ Quantitative Mass Spectrometry
  - ○ Validation of Complex Formation by Sequential Immunoprecipitations
  - ○ Microarray Analysis of Gene Expression Data
  - ○ RNA Extraction and Quantitative Real-time PCR
  - ○ Mathematical Model
  - ○ Description of the Comprehensive Mathematical Model
  - ○ Reactions of the Comprehensive Mathematical Model
  - ○ ODE System of the Comprehensive Mathematical Model
  - ○ Identification of the Occurring Smad Complex Using L_1 Regularization
  - ○ Model Prediction and Experimental Validation of mRNA Half-lives
  - ○ Gradual Knock-down of Smad Proteins
  - ○ Overexpression of Smad Proteins
  - ○ Transcriptional Activity of Smad Complexes
  - ○ Analysis of Hepatocellular Carcinoma Samples
  - ○ Prediction of Complexes and Total Amounts

- 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 (BMBF) within the e:Bio collaborative research projects “Systems Biology of Erythropoietin” (SBEpo, 0316182B) and “Multi-Scale Modeling of Drug Induced Liver Injury” (MS_DILI, 031L0074A, 031L0074B) and the junior group EAsys (031L0080) within the Virtual Liver Network (0315745, 0315741), within the Liver Systems Medicine network (LiSyM, 031L0042, 031L0048), 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. Stable cell lines and measurements of protein and mRNA levels by qRT-PCR were conducted by P.L., S.L., M.W., and A.V. Data analysis, quantitative dynamic modeling, and further model analyses were performed by P.L., M.S., 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. A.V., 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

Ali, N.A., and Molloy, M.P. (2011). Quantitative phosphoproteomics of transforming growth factor-beta signaling in colon cancer cells. Proteomics *11*, 3390–3401.

Boehm, M.E., Adlung, L., Schilling, M., Roth, S., Klingmuller, U., and Lehmann, W.D. (2014). Identification of isoform-specific dynamics in phosphorylation-dependent STAT5 dimerization by quantitative mass spectrometry and mathematical modeling. J. Proteome Res. *13*, 5685–5694.

Chacko, B.M., Qin, B., Correia, J.J., Lam, S.S., de Caestecker, M.P., and Lin, K. (2001). The L3 loop and C-terminal phosphorylation jointly define Smad protein trimerization. Nat. Struct. Biol. *8*, 248–253.

D’Souza, R.C., Knittle, A.M., Nagaraj, N., van Dinther, M., Choudhary, C., ten Dijke, P., Mann, M., and Sharma, K. (2014). Time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF-beta. Sci. Signal. *7*, rs5.

Derynck, R., and Feng, X.H. (1997). TGF-beta receptor signaling. Biochim. Biophys. Acta *1333*, F105–F150.

Domingo-Gonzalez, R., Wilke, C.A., Huang, S.K., Laouar, Y., Brown, J.P., Freeman, C.M., Curtis, J.L., Yanik, G.A., and Moore, B.B. (2015). Transforming growth factor-beta induces microRNA-29b to promote murine alveolar macrophage dysfunction after bone marrow transplantation. Am. J. Physiol. Lung Cell. Mol. Physiol. *308*, L86–L95.

Dosen-Dahl, G., Munthe, E., Nygren, M.K., Stubberud, H., Hystad, M.E., and Rian, E. (2008). Bone marrow stroma cells regulate TIEG1 expression in acute lymphoblastic leukemia cells: role of TGFbeta/BMP-6 and TIEG1 in chemotherapy escape. Int. J. Cancer *123*, 2759–2766.

Dzieran, J., Fabian, J., Feng, T., Coulouarn, C., Ilkavets, I., Kyselova, A., Breuhahn, K., Dooley, S., and Meindl-Beinker, N.M. (2013). Comparative analysis of TGF-beta/Smad signaling dependent cytostasis in human hepatocellular carcinoma cell lines. PLoS One *8*, e72252.

Feng, X.H., and Derynck, R. (2005). Specificity and versatility in TGF-beta signaling through Smads. Annu. Rev. Cell Dev. Biol. *21*, 659–693.

Ge, Q., Moir, L.M., Black, J.L., Oliver, B.G., and Burgess, J.K. (2010). TGFbeta1 induces IL-6 and inhibits IL-8 release in human bronchial epithelial cells: the role of Smad2/3. J. Cell Physiol. *225*, 846–854.

Greber, B., Lehrach, H., and Adjaye, J. (2007). Fibroblast growth factor 2 modulates transforming growth factor beta signaling in mouse embryonic fibroblasts and human ESCs (hESCs) to support hESC self-renewal. Stem Cells *25*, 455–464.

Hao, K., Lamb, J., Zhang, C., Xie, T., Wang, K., Zhang, B., Chudin, E., Lee, N.P., Mao, M., Zhong, H., et al. (2011). Clinicopathologic and gene expression parameters predict liver cancer prognosis. BMC Cancer *11*, 481.

Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., and Woodward, C.S. (2005). SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. ACM Trans. Math. Softw. *31*, 363–396.

Inman, G.J., Nicolas, F.J., Callahan, J.F., Harling, J.D., Gaster, L.M., Reith, A.D., Laping, N.J., and Hill, C.S. (2002). SB-431542 is a potent and specific inhibitor of transforming growth factor-beta superfamily type I activin receptor-like kinase (ALK) receptors ALK4, ALK5, and ALK7. Mol. Pharmacol. *62*, 65–74.

Itoh, S., and ten Dijke, P. (2007). Negative regulation of TGF-beta receptor/Smad signal transduction. Curr. Opin. Cell Biol. *19*, 176–184.

Kegel, V., Deharde, D., Pfeiffer, E., Zeilinger, K., Seehofer, D., and Damm, G. (2016). Protocol for isolation of primary human hepatocytes and corresponding major populations of non-parenchymal liver cells. J. Vis. Exp. e53069.

Koinuma, D., Tsutsumi, S., Kamimura, N., Taniguchi, H., Miyazawa, K., Sunamura, M., Imamura, T., Miyazono, K., and Aburatani, H. (2009). Chromatin immunoprecipitation on microarray analysis of Smad2/3 binding sites reveals roles of ETS1 and TFAP2A in transforming growth factor beta signaling. Mol. Cell. Biol. *29*, 172–186.

Lebrun, J.J., Takabe, K., Chen, Y., and Vale, W. (1999). Roles of pathway-specific and inhibitory Smads in activin receptor signaling. Mol. Endocrinol. *13*, 15–23.

Levy, L., and Hill, C.S. (2006). Alterations in components of the TGF-beta superfamily signaling pathways in human cancer. Cytokine Growth Factor Rev. *17*, 41–58.

Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Staerfeldt, H.H., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods *14*, 61–64.

Luo, K., Stroschein, S.L., Wang, W., Chen, D., Martens, E., Zhou, S., and Zhou, Q. (1999). The Ski oncoprotein interacts with the Smad proteins to repress TGFbeta signaling. Genes Dev. *13*, 2196–2206.

Massague, J., Seoane, J., and Wotton, D. (2005). Smad transcription factors. Genes Dev. *19*, 2783–2810.

Merkle, R., Steiert, B., Salopiata, F., Depner, S., Raue, A., Iwamoto, N., Schelker, M., Hass, H., Wasch, M., Bohm, M.E., et al. (2016). Identification of cell type-specific differences in erythropoietin receptor signaling in primary erythroid and lung cancer cells. PLoS Comput. Biol. *12*, e1005049.

Moustakas, A., and Heldin, C.H. (2002). From mono- to oligo-Smads: the heart of the matter in TGF-beta signal transduction. Genes Dev. *16*, 1867–1871.

Mueller, S., Huard, J., Waldow, K., Huang, X., D’Alessandro, L.A., Bohl, S., Borner, K., Grimm, D., Klamt, S., Klingmuller, U., et al. (2015). T160-phosphorylated CDK2 defines threshold for HGF dependent proliferation in primary hepatocytes. Mol. Syst. Biol. *11*, 795.

Nozato, E., Shiraishi, M., and Nishimaki, T. (2003). Up-regulation of hepatocyte growth factor caused by an over-expression of transforming growth factor beta, in the rat model of fulminant hepatic failure. J. Surg. Res. *115*, 226–234.

Qin, H., Chan, M.W., Liyanarachchi, S., Balch, C., Potter, D., Souriraj, I.J., Cheng, A.S., Agosto-Perez, F.J., Nikonova, E.V., Yan, P.S., et al. (2009). An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. BMC Syst. Biol. *3*, 73.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics *25*, 1923–1929.

Raue, A., Steiert, B., Schelker, M., Kreutz, C., Maiwald, T., Hass, H., Vanlier, J., Tonsing, C., Adlung, L., Engesser, R., et al. (2015). Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. Bioinformatics *31*, 3558–3560.

Schmierer, B., Tournier, A.L., Bates, P.A., and Hill, C.S. (2008). Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. Proc. Natl. Acad. Sci. USA *105*, 6608–6613.

Steiert, B., Timmer, J., and Kreutz, C. (2016). L1 regularization facilitates detection of cell type-specific parameters in dynamical systems. Bioinformatics *32*, i718–i726.

Stockert, J., Adhikary, T., Kaddatz, K., Finkernagel, F., Meissner, W., Muller-Brusselbach, S., and Muller, R. (2011). Reverse crosstalk of TGFbeta and PPARbeta/delta signaling identified by transcriptional profiling. Nucleic Acids Res. *39*, 119–131.

Stroschein, S.L., Wang, W., Zhou, S., Zhou, Q., and Luo, K. (1999). Negative feedback regulation of TGF-beta signaling by the SnoN oncoprotein. Science *286*, 771–774.

Tao, H., Yang, J.J., Hu, W., Shi, K.H., Deng, Z.Y., and Li, J. (2016). MeCP2 regulation of cardiac fibroblast proliferation and fibrosis by down-regulation of DUSP5. Int. J. Biol. Macromol. *82*, 68–75.

Wrana, J.L. (2002). Phosphoserine-dependent regulation of protein-protein interactions in the Smad pathway. Structure *10*, 5–7.

Wu, J.W., Fairman, R., Penry, J., and Shi, Y. (2001a). Formation of a stable heterodimer between Smad2 and Smad4. J. Biol. Chem. *276*, 20688–20694.

Wu, J.W., Hu, M., Chai, J., Seoane, J., Huse, M., Li, C., Rigotti, D.J., Kyin, S., Muir, T.W., Fairman, R., et al. (2001b). Crystal structure of a phosphorylated Smad2. Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signaling. Mol. Cell *8*, 1277–1289.

Yakicier, M.C., Irmak, M.B., Romano, A., Kew, M., and Ozturk, M. (1999). Smad2 and Smad4 gene mutations in hepatocellular carcinoma. Oncogene *18*, 4879–4883.

Zhang, Y., Handley, D., Kaplan, T., Yu, H., Bais, A.S., Richards, T., Pandit, K.V., Zeng, Q., Benos, P.V., Friedman, N., et al. (2011). High throughput determination of TGFbeta1/SMAD3 targets in A549 lung epithelial cells. PLoS One *6*, e20319.

Zi, Z., Feng, Z., Chapnick, D.A., Dahl, M., Deng, D., Klipp, E., Moustakas, A., and Liu, X. (2011). Quantitative analysis of transient and sustained transforming growth factor-beta signaling dynamics. Mol. Syst. Biol. *7*, 492.

Zieba, A., Pardali, K., Soderberg, O., Lindbom, L., Nystrom, E., Moustakas, A., Heldin, C.H., and Landegren, U. (2012). Intercellular variation in signaling through the TGF-beta pathway and its relation to cell density and cell cycle phase. Mol. Cell Proteomics *11*, M111.013482.

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

STAR★METHODS

KEY RESOURCES TABLE

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

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R Project for Statistical Computing	https://www.r-project.org/	RRID:SCR_001905
Data2Dynamics modeling environment	Raue et al., 2015 Bioinformatics 31(21):3558-60. https://doi.org/10.1093/bioinformatics/btv405	http://data2dynamics.org
MetaCore	Thomson Reuters	version 6.31 build 68930
Ingenuity IPA	Qiagen	build 441680M
Matlab	The Mathworks	R2016b

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× 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 were used. 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, 1.2×10^6 Hepa1-6 cells and 2×10^6 HepG2 cells, primary mouse and primary human hepatocytes (for IB and qRT-PCR experiments) and 7.5×10^6 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% glutamine (Gibco).

Donor anamnesis of the three patients providing primary human hepatocytes

Donor	Age (years)	Sex	BMI	Diagnosis
1	58	female	24.7	Liver metastases
2	65	female	22.6	Liver metastases
3	75	male	27	Liver metastases

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)

Donor	Cirrhosis	Age (years)	Sex
A1	No	71	M
A2	No	74	M

(*Continued on next page*)

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

Continued

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

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

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 B)

Donor	Cirrhosis	Age (years)	Sex
B1	Yes	66	M
B2	Yes	72	M
B3	No	68	M
B4	No	75	M
B5	No	80	M
B6	Yes	55	W
B7	No	73	M
B8	Yes	64	M
B9	Yes	57	M
B10	No	58	M
B11	No	59	M
B12	No	82	W

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 (2×10^6) 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 $ZnCl_2$ pH 4.0, 1 mM $MgCl_2$, 2 mM Na_3VO_4, 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 alkylation 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 NH_4HCO_3 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 20mM 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-PPVVK vs. acSSILPF-pT-PPIVK and (K)TGRLDELEK vs. (K)TGQLDELEK. 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

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

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

7.5×10^6 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 anit-Smad3 (Cell Signaling #9523). After overnight incubation, bead-bound proteins were dissociated by the addition of a Smad3 blocking peptide (Cell Signaling #1933S) 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

2×10^6 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

$$f(t) = A_{\mathrm{sus}}\left(1 - e^{-\frac{t}{\tau_1}}\right) + A_{\mathrm{trans}}\left(1 - e^{-\frac{t}{\tau_1}}\right)e^{-\frac{t}{\tau_2}} + p_0 \qquad \text{(Equation 2)}$$

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

The first term represents a sustained response with amplitude A_{sus} and time constant τ_1. The second term accounts for transient up- or down-regulation with amplitude A_{trans} with the same time constant τ_1 for induction and a second time scale τ_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. τ_1, $\tau_2 < 2\ t_{\mathrm{max}} = 20$ 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 $[1\times10^{-10}, 2\ \Delta y]$ was used as constraint where Δy 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% CO_2 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.

Gene-specific mouse (m) and human (h) primers and UPL probes

Gene	Forward primer	Reverse primer	UPL
mBmp4	gaggagtttccatcacgaaga	gctctgccgaggagatca	89
mCxcl15	tgctcaaggctggtccat	gacatcgtagctcttgagtgtca	18
mDnmt3a	aaacggaaacgggatgagt	actgcaattaccttggctttct	75
mDusp5	gatcgaaggcgagagaagc	ggaagggaaggatttcaacc	102
mHprt	cctcctcagaccgcttttt	aacctggttcatcatcgctaa	95
mJun	tttgattcaaa	agggacccatggaag	12
mKlf10	agccaaccatgctcaacttc	ggcttttcagaaattagttccatt	67
mPdk4	cgcttagtgaacactccttcg	cttctgggctcttctcatgg	22
mSki	gagaaagagacgtccccaca	tcaaagctcttgtaggagtagaagc	33
mSkil	gacagggaggccgagtatg	ccgctcctgtctgagttcat	96
mSmad7	acccccatcaccttagtcg	gaaaatccattgggtatctgga	63
mSox4	ctcgctctcctcgtcctct	cgtcttcgaactcgtcgtc	63
mTgfa	cctggtggtggtctccatt	cagtgtttgcggagctga	81
hBMP4	ctgcaaccgttcagaggtc	tgctcgggatggcactac	17
hDNMT3A	cctgaagcctcaagagcagt	tggtctccttctgttctttgc	46
hDUSP5	caaatggatccctgtggaa	cccttttccctgacacagtc	5
hHPRT	tgaccttgatttattttgcatacc	cgagcaagacgttcagtcct	73
hJUN	ccaaaggatagtgcgatgttt	ctgtccctctccactgcaac	19
hKLF10	tctgaaggcccacacgag	acctcctttcacaacctttcc	2
hPDK4	cagtgcaattggttaaaagctg	ggtcatctgggcttttctca	31
hSKI	gaagcaggaggagaagctcag	ccacgcgtaggaactcca	22
hSKIL	gaggctgaatatgcaggacag	cttgcctatcggcctcag	13
hSMAD7	acccgatggattttctcaaa	aggggccagataattcgttc	69
hSOX4	caacgccaactccagctc	accgaccttgtctcccttc	25
hTGFA	ttgctgccactcagaaacag	atctgccacagtccacctg	63

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 downregulated by degradation (Derynck 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 phosphosphorylated 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 comprehensive model contains the following reactions:

$$v_1 = [\mathrm{Rec}] \cdot \mathrm{Rec}_{\mathrm{act}} \cdot [\mathrm{TGFb}] \tag{Equation 3}$$

$$v_2 = [\mathrm{TGFb\_pRect}] \cdot \mathrm{Rec\_degind} \tag{Equation 4}$$

$$v_3 = \mathrm{k\_on\_222} \cdot [\mathrm{ppS2}]^3 \tag{Equation 5}$$

$$v_4 = 3 \cdot \mathrm{S\_dephosphos} \cdot [\mathrm{ppS2\_ppS2\_ppS2}] \tag{Equation 6}$$

$$v_5 = \mathrm{k\_on\_333} \cdot [\mathrm{ppS3}]^3 \tag{Equation 7}$$

$$v_6 = 3 \cdot \mathrm{S\_dephosphos} \cdot [\mathrm{ppS3\_ppS3\_ppS3}] \tag{Equation 8}$$

$$v_7 = [\mathrm{S4}]^3 \cdot \mathrm{k\_on\_444} \tag{Equation 9}$$

$$v_8 = [\mathrm{S4\_S4\_S4}] \cdot \mathrm{kdiss\_SS} \tag{Equation 10}$$

$$v_9 = [\mathrm{S2}] \cdot \mathrm{S\_phos} \cdot [\mathrm{TGFb\_pRec}] \tag{Equation 11}$$

$$v_{10} = \mathrm{S\_dephosphos} \cdot [\mathrm{ppS2}] \tag{Equation 12}$$

$$v_{11} = \mathrm{S\_dephos} \cdot [\mathrm{pS2}] \tag{Equation 13}$$

$$v_{12} = [\mathrm{S3}] \cdot \mathrm{S\_phos} \cdot [\mathrm{TGFb\_pRec}] \tag{Equation 14}$$

$$v_{13} = \mathrm{S\_dephosphos} \cdot [\mathrm{ppS3}] \tag{Equation 15}$$

$$v_{14} = \mathrm{S\_dephos} \cdot [\mathrm{pS3}] \tag{Equation 16}$$

$$v_{15} = \mathrm{k\_on\_223} \cdot [\mathrm{ppS2}]^2 \cdot [\mathrm{ppS3}] \tag{Equation 17}$$

$$v_{16} = 2 \cdot \mathrm{S\_dephosphos} \cdot [\mathrm{ppS2\_ppS2\_ppS3}] \tag{Equation 18}$$

$$v_{17} = \mathrm{S\_dephosphos} \cdot [\mathrm{ppS2\_ppS2\_ppS3}] \tag{Equation 19}$$

$$v_{18}=[\text{S4}]\cdot \text{k\_on\_224}\cdot[\text{ppS2}]^2 \qquad \text{(Equation 20)}$$

$$v_{19}=2\cdot \text{S\_dephosphos}\cdot[\text{ppS2\_ppS2\_S4}] \qquad \text{(Equation 21)}$$

$$v_{20}=\text{k\_on\_233}\cdot[\text{ppS2}]\cdot[\text{ppS3}]^2 \qquad \text{(Equation 22)}$$

$$v_{21}=\text{S\_dephosphos}\cdot[\text{ppS2\_ppS3\_ppS3}] \qquad \text{(Equation 23)}$$

$$v_{22}=2\cdot \text{S\_dephosphos}\cdot[\text{ppS2\_ppS3\_ppS3}] \qquad \text{(Equation 24)}$$

$$v_{23}=[\text{S4}]\cdot \text{k\_on\_334}\cdot[\text{ppS3}]^2 \qquad \text{(Equation 25)}$$

$$v_{24}=2\cdot \text{S\_dephosphos}\cdot[\text{ppS3\_ppS3\_S4}] \qquad \text{(Equation 26)}$$

$$v_{25}=[\text{S4}]^2\cdot \text{kon\_244}\cdot[\text{ppS2}] \qquad \text{(Equation 27)}$$

$$v_{26}=\text{S\_dephosphos}\cdot[\text{ppS2\_S4\_S4}] \qquad \text{(Equation 28)}$$

$$v_{27}=[\text{S4}]^2\cdot \text{kon\_344}\cdot[\text{ppS3}] \qquad \text{(Equation 29)}$$

$$v_{28}=\text{S\_dephosphos}\cdot[\text{ppS3\_S4\_S4}] \qquad \text{(Equation 30)}$$

$$v_{29}=[\text{S4}]\cdot \text{k\_on\_234}\cdot[\text{ppS2}]\ \cdot[\text{ppS3}] \qquad \text{(Equation 31)}$$

$$v_{30}=\text{S\_dephosphos}\cdot[\text{ppS2\_ppS3\_S4}] \qquad \text{(Equation 32)}$$

$$v_{31}=\text{S\_dephosphos}\cdot[\text{ppS2\_ppS3\_S4}] \qquad \text{(Equation 33)}$$

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

$$v_{32}=\frac{\text{gene\_turn}+\text{gene\_act1}\cdot[\text{ppS2\_ppS3\_ppS3}]+\text{gene\_act2}\cdot[\text{ppS2\_S4\_S4}]+\text{gene\_act3}\cdot[\text{ppS2\_ppS3\_S4}]}{\text{gene\_inh1}\cdot[\text{ppS2\_ppS3\_ppS3}]+\text{gene\_inh2}\cdot[\text{ppS2\_S4\_S4}]+\text{gene\_inh3}\cdot[\text{ppS2\_ppS3\_S4}]+1} \qquad \text{(Equation 34)}$$

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

with “gene” representing *SKI*, *SKIL*, *DNMT3A*, *SOX4*, *JUN*, *SMAD7*, *KLF10*, *BMP4*, *CXCL15*, *DUSP5*, *TGFA* and *PDK4* and “gene_turn” describing the gene-specific and TGFβ-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:

$$d[\text{TGFb}]/dt=-v_1 \qquad \text{(Equation 36)}$$

$$d[\text{Rec}]/dt=-v_1 \qquad \text{(Equation 37)}$$

$$d[\text{TGFb\_pRec}]/dt=+v_1-v_2 \qquad \text{(Equation 38)}$$

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

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

$$d[\mathrm{S4}]/dt = -3 \cdot v_7 + 3 \cdot v_8 - v_{18} + v_{19} - v_{23} + v_{24} - 2 \cdot v_{25} + 2 \cdot v_{26} - 2 \cdot v_{27} + 2 \cdot v_{28} - v_{29} + v_{30} + v_{31} \quad \text{(Equation 41)}$$

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

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

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

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

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

$$d[\mathrm{ppS2}]/dt = -3 \cdot v_3 + 2 \cdot v_4 + v_9 - v_{10} - 2 \cdot v_{15} + v_{16} + 2 \cdot v_{17} - 2 \cdot v_{18} + v_{19} - v_{20} + v_{22} - v_{25} - v_{29} + v_{31} \quad \text{(Equation 47)}$$

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

$$d[\mathrm{ppS2\_ppS2\_S4}]/dt = +v_{18} - v_{19} \quad \text{(Equation 49)}$$

$$d[\mathrm{ppS2\_ppS2\_ppS3}]/dt = +v_{15} - v_{16} - v_{17} \quad \text{(Equation 50)}$$

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

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

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

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

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

$$d[\mathrm{gene}]/dt = +v_{32} - v_{33} \quad \text{(Equation 56)}$$

with "gene" representing *SKI*, *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 L_1 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).

The idea of L_1 regularization is to minimize an objective function $\chi^2_{pen} = \chi^2_{data} + \chi^2_{l1}$ which is a sum of a term $\chi^2_{data} = \Sigma_i (y_i - g_i)^2/\sigma^2$ assessing agreement of the data y_i, $i=1,\ldots,n_{data}$ with the model g_i and a second term $\chi^2_{l1} = \lambda\ \Sigma_j \mid \theta_j \mid$ penalizing parameters θ_j, $j=1,\ldots,n_{para}$, which are different from zero. Since the penalty terms in χ^2_{l1} have a non-vanishing gradient $\pm\lambda$ for all values unequal to zero, parameters which improve χ^2_{data} less than $\lambda\theta$ are estimated equals to zero.

In our context, L_1 regularization is applied to identify the complexes which are required to describe the coIP data. For these complexes, χ^2_{pen} is optimal for non-vanishing association rates. For complexes which are not required, the penalty χ^2_{l1} dominates the data contribution χ^2_{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 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

0.5×10^6 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 before. Hepa1-6 cells were treated with 1 ng/ml TGFβ and RNA was harvested at the indicated time points as described above. Additionally, unstimulated cells were lysed as described above to access the knock-down efficiency. IP was performed with specific antibodies against Smad2 (Cell Signaling, #5339), Smad3 (Cell Signaling #9523) and Smad4 (Cell Signaling #38454). Quantitative IB was performed with anti-Smad2 (Cell Signaling #5339), anti-Smad3 (Cell Signaling #9523) and anti-Smad4 (Cell Signaling #38454) antibodies, respectively. Knock-down efficiency was assessed by immunblotting relatively to the impact of control siRNA.

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

$$AUC_{c,g}(p) = \frac{1}{t_{\max}} \int_{t=0}^{t_{\max}} log_2\left(x_g(t,p)\right)/x_g(0,p)\, dt \qquad \text{(Equation 57)}$$

of the $\log_2$-ratios of the concentrations $x_g(t,p)$ 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 $AUC_{c,g}$ indicates negative regulation of the respective Smad complex c on gene g, a positive $AUC_{c,g}$ is obtained for positive regulators. To translate uncertainties in the estimated parameters p to $AUC_{c,g}(p)$, 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 × *g* and 4°C. Supernatants were subjected to IPs with anti-Smad2/3 (BD-610843) or anti-Smad4 (Cell Signaling #9515) antibodies and supplemented with Protein A sepharose (GE healthcare 17-0963-03) and 1 ng GST-Smad3 or 1 ng SBP-Smad4 calibrators, 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

Please cite this article in press as: Lucarelli et al., Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression, Cell Systems (2017), https://doi.org/10.1016/j.cels.2017.11.010

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 $\log_{10}$ fold $\sim N(\mu,\sigma^2)$ with μ=0 and σ^2=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 $S2_{\mathrm{fold}}$, $S3_{\mathrm{fold}}$, $\mathrm{S4}_{\mathrm{fold}}$ 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 $\log_2$ 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.

Cell Systems, Volume *6*

Supplemental Information

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

Philippe Lucarelli, Marcel Schilling, Clemens Kreutz, Artyom Vlasov, Martin E. Boehm, Nao Iwamoto, Bernhard Steiert, Susen Lattermann, Marvin Wäsch, Markus Stepath, Matthias S. Matter, Mathias Heikenwälder, Katrin Hoffmann, Daniela Deharde, Georg Damm, Daniel Seehofer, Maria Muciek, Norbert Gretz, Wolf D. Lehmann, Jens Timmer, and Ursula Klingmüller

Contents

1 Supplemental Figures

1.1 Figure S1

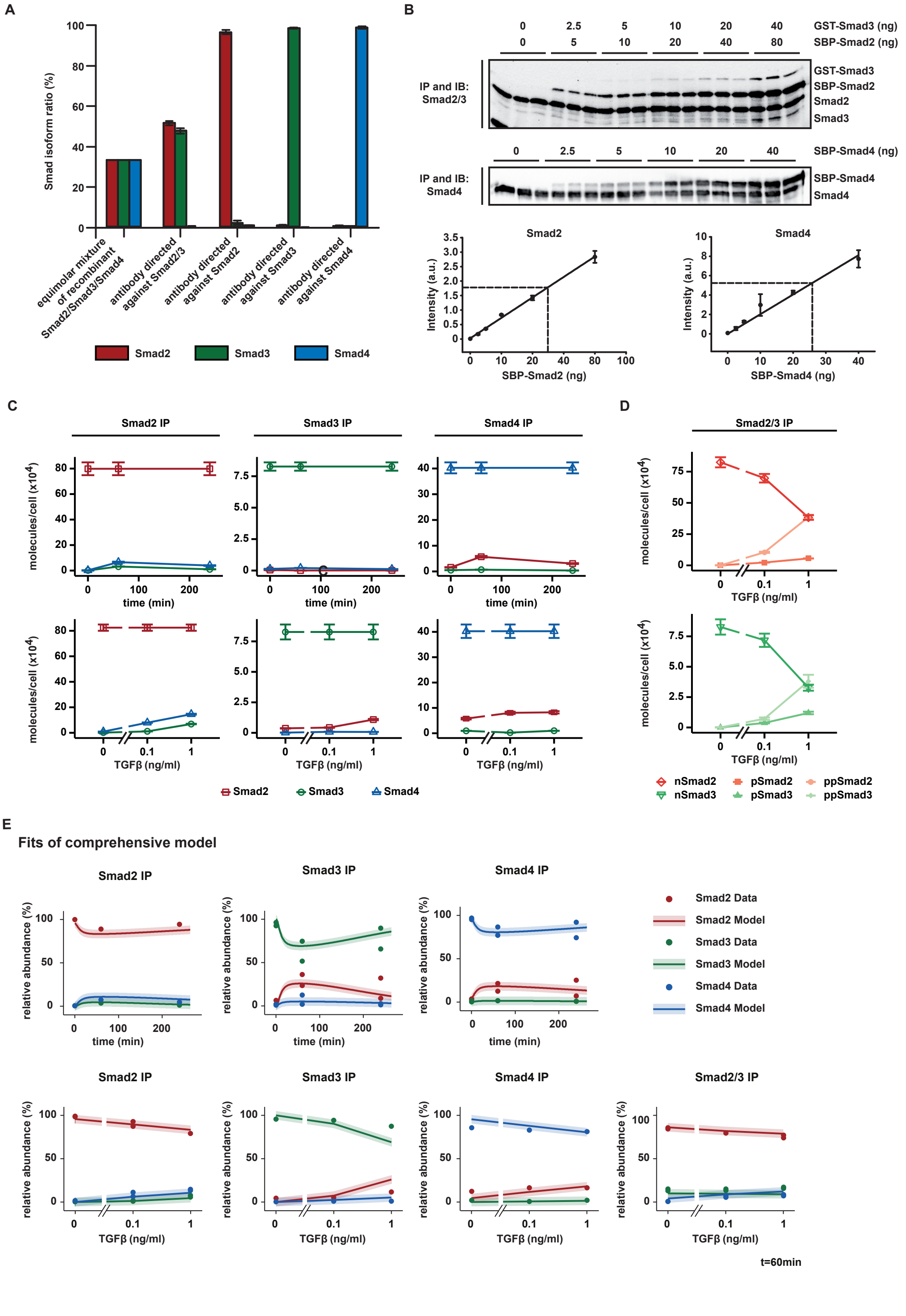


Figure S1 (related to Figure 1 and Figure 2): Comparison of Smad antibody specificities, determination of the total amount of Smad molecules and complex formation and phosphorylation of Smad molecules in Hepa1-6 cells and fits of the comprehensive mathematical model

Figure S1 (Continued) **A.** Equimolar amounts of recombinantly produced GST- or SBP-tagged Smad2, Smad3 and Smad4 proteins were mixed and subjected to immunoprecipitations with antibodies directed against Smad2, Smad3, Smad4 and Smad2/3. Protein ratios of precipitated proteins were determined by quantitative mass spectrometry. **B.** Whole cell lysates of 2×10^6 Hepa1-6 cells were analyzed for endogenous levels of Smad2 and Smad4 by means of dilution series of added recombinant calibrator proteins SBP-Smad2 and SBP-Smad4, respectively. Protein levels were determined by immunoprecipitation (IP) and subsequent quantitative immunoblotting (IB). Data of GST-Smad3 was not used due to the resulting high background of the immunoblot. Linear regression analysis (solid lines) was performed for recombinant calibrators SBP-Smad2 and SBP-Smad4, respectively. Endogenous protein levels of unstimulated cells (dashed lines) were calculated by means of the respective regression functions. a.u.: arbitrary units. **C.** Hepa1-6 cells were treated with 1 ng/ml TGFβ in a time-resolved manner (upper panel) or stimulated for 60 minutes with different concentrations of TGFβ (lower panel). Whole cell lysates were subjected to immunoprecipitation with antibodies directed against Smad2, Smad3 and Smad4 and analyzed by quantitative mass spectrometry for the relative protein abundance. Error bars represent 5% error for mass spectrometry measurements (n=1). **D.** Hepa1-6 cells were treated with TGFβ for 60 minutes with different concentrations of TGFβ. Whole cell lysates were subjected to immunoprecipitation with antibodies directed against Smad2/3 and analyzed by quantitative mass spectrometry. Data was converted to molecules/cell based on the results depicted in Figure 1. n: non-phosphorylated; p: single phosphorylated; pp: double phosphorylated. **E.** Time-resolved mass spectrometry analysis of the relative protein abundance for Smad proteins and complexes. Whole cell lysates of Hepa1-6 cells were stimulated with TGFβ for 240 minutes, lysed at indicated time points, subjected to IP with anti-Smad2 (n=1), anti-Smad3 (n=2) and anti-Smad4 (n=2) antibodies and analyzed by quantitative mass spectrometry (dots: experimental data, continuous line: model trajectories; shading: corresponds to 5% error, which is typical for mass spectrometry experiments). Dose response experiment after 60 minutes stimulation with 0, 0.1 and 1 ng/ml of TGFβ (Smad2 IP n=2, Smad3 IP n=1, Smad4 IP n=1, Smad2/3 IP n=2).

1.2 Figure S2

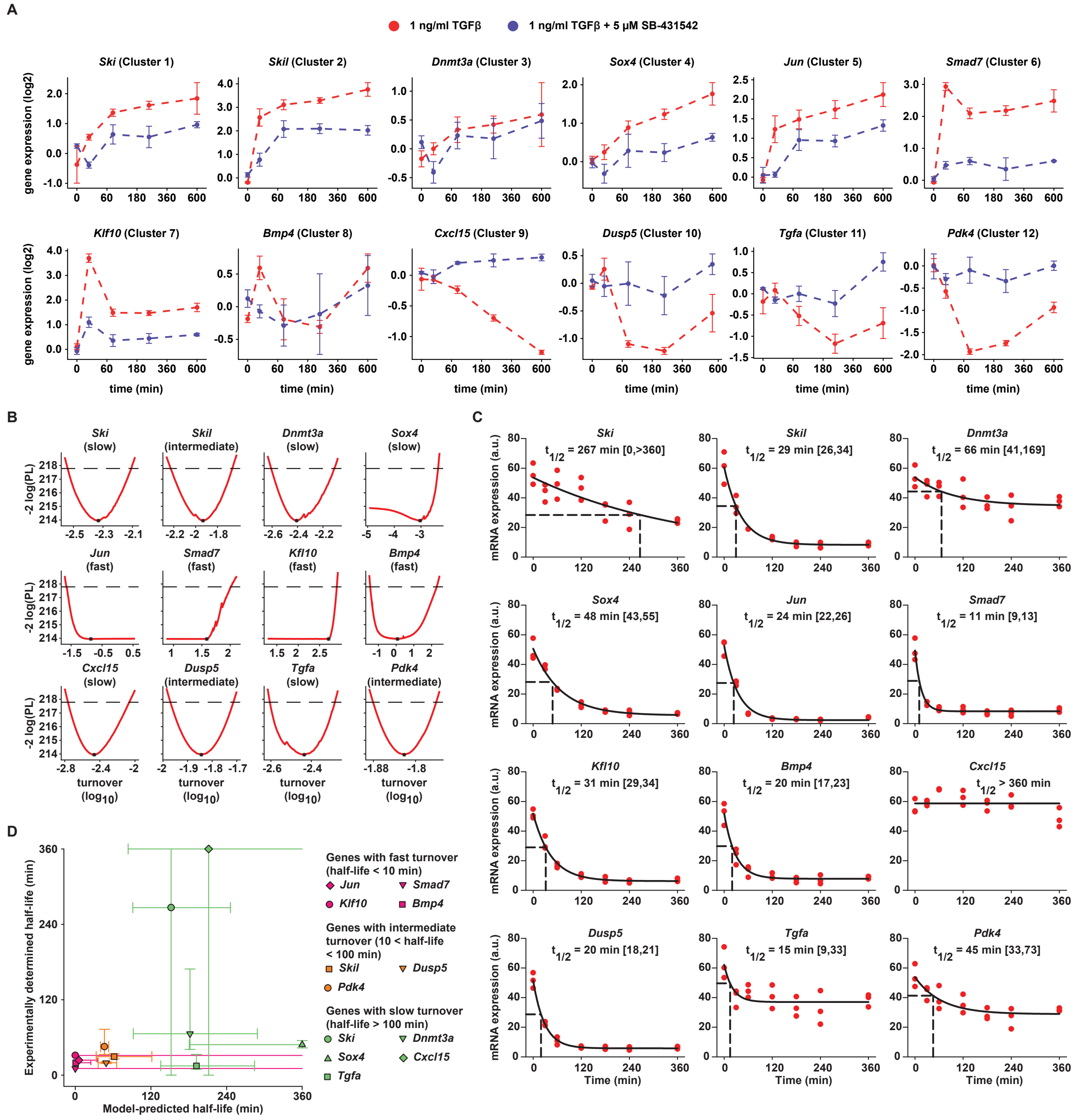


Figure S2 (related to Figure 3): TGFβ target gene validation and comparison of model-predicted to experimentally measured gene stability

Figure S2 (Continued) **A.** Hepa1-6 cells were pre-treated with 5 μM SB-431542 or DMSO as control for 30 minutes prior to stimulation with 1 ng/ml of TGFβ. Total RNA was extracted and analyzed by qRT-PCR for the twelve representative Smad target genes. Experiments were performed in biological triplicates and mean and standard deviations are shown. **B.** The uncertainty of each gene-specific turnover parameters was analyzed by the profile likelihood method. X-axes show the parameter value of the analyzed turnover parameter. Y-axes describe the difference in likelihood. Black stars denote the value for the best fit. Red dashed lines define the point-wise confidence interval in likelihood difference, defining the confidence interval of the model-predicted half-lives. Based on the best fit, the turnover parameters were classified as fast (half-life $<$ 10 minutes), intermediate (half-life between 10 and 100 minutes) or slow (half-life $>$ 100 minutes), with half-life $= ln(2)/$turnover. **C.** Hepa1-6 cells were pre-stimulated with 1 ng/ml of TGFβ for 120 minutes and subsequently treated with 1 μg/ml of Actinomycin D to inhibit transcription. mRNA stability was measured for up to 360 minutes. Total RNA was extracted and analyzed by qRT-PCR for the twelve representative Smad target genes. mRNA half-lives (black dashed lines) were estimated by exponential decay functions (black solid lines). Confidence intervals are given in squared brackets. **D.** Model-predicted half-lives determined in (A) are compared to experimentally measured half-lives as indicated in (B). Error bars represent respective confidence intervals.

1.3 Figure S3

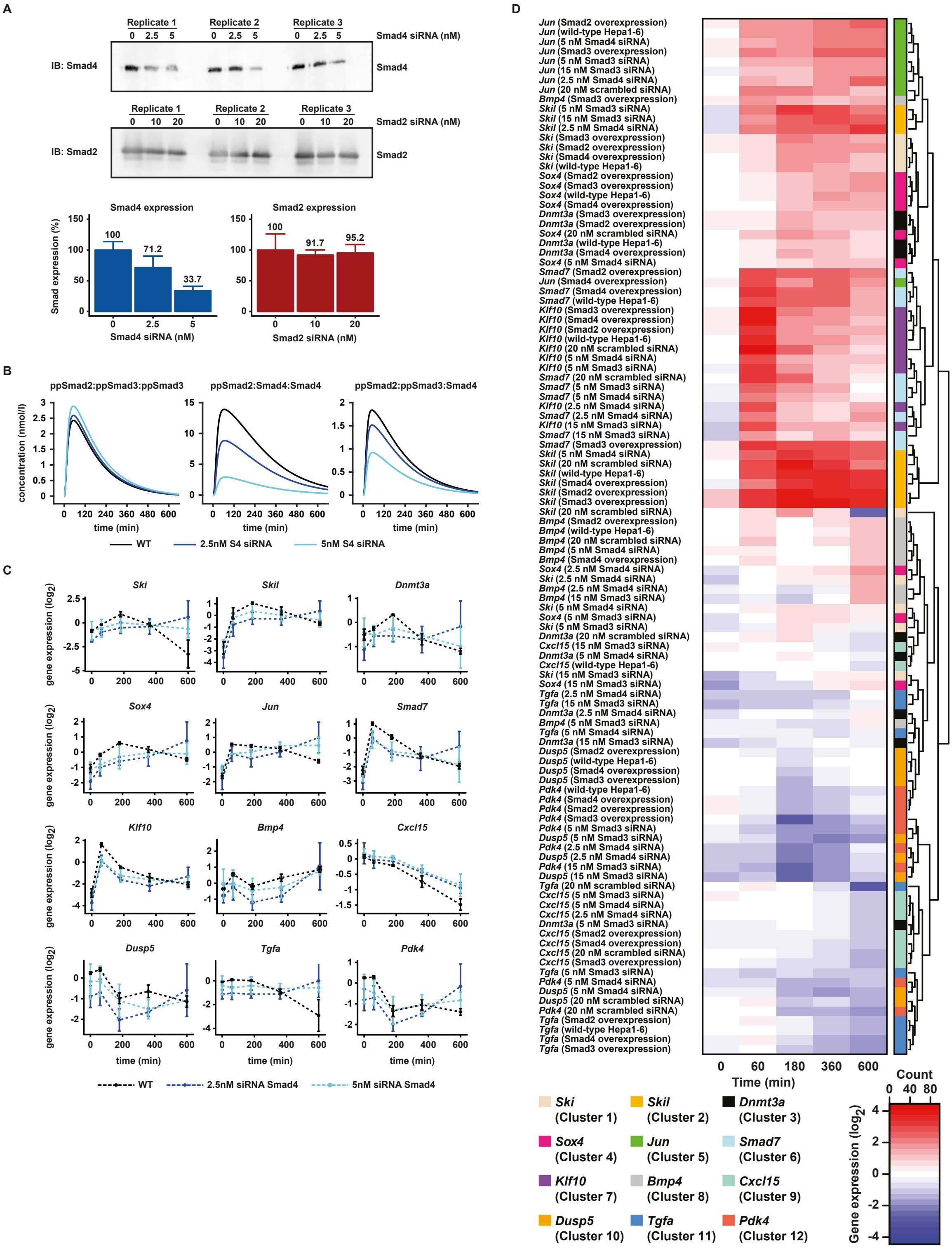


Figure S3 (related to Figure 4): TGFβ-induced gene expression in Hepa1-6 cells upon knockdown of Smad4 and cluster stability of TGFβ target genes

Figure S3 (Continued) **A.** Smad4 and Smad2 proteins were down-regulated in a gradual way by using two different concentration of target siRNA. Whole cell lysates of Hepa1-6 were subjected to immunoprecipitations with anti-Smad4 and anti-Smad2 antibodies, respectively. Signals were quantified by an ImageQuant device utilizing a CCD-camera and Smad expression was accessed relative to the scrambled siRNA control. Experiments were performed in biological triplicates and mean and standard deviations are shown. **B.** Simulations with the integrative mathematical model for the time-resolved complex abundance after a gradual knock-down of Smad4 in presence of 1 ng/ml of TGFβ were performed (continuous lines: predicted model trajectories). WT: Hepa1-6 wild type. **C.** TGFβ-induced gene expression after Smad4 knock-down in Hepa1-6 cells was determined for the selected twelve TGFβ target genes by qRT-PCR. Experiments were performed in biological triplicates and mean and standard deviations are shown. WT: Hepa1-6 wild type. **D.** Hierarchical clustering was performed for averaged, $\log_2$ transformed gene expression data of 1 ng/ml of TGFβ-stimulated Hepa1-6 cells. Gene expression was measured under the following nine conditions: Hepa1-6 wild-type cells, Hepa1-6 cells overexpressing Smad2, Smad3 or Smad4, Hepa1-6 cells treated with 5 nM or 15 nM Smad3 siRNA, with 2.5 nM or 5 nM Smad4 siRNA or with 20 nM scrambled siRNA.

1.4 Figure S4

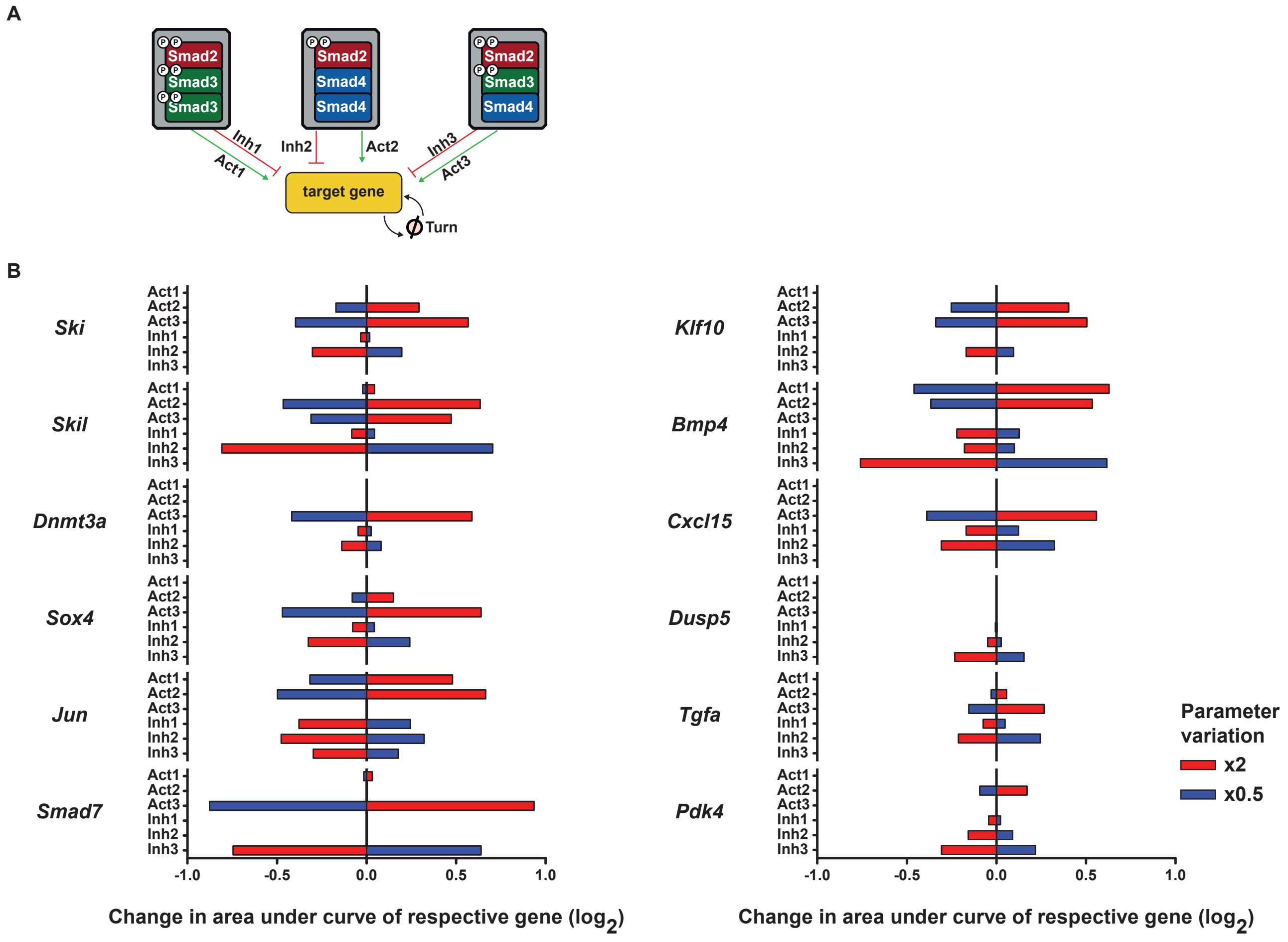


Figure S4 (related to Figure 5): Sensitivity analysis of activation and deactivation parameters contributing to target gene expression
A. TGFβ target genes were linked to the reduced pathway model, establishing an integrative mathematical model. The induction of target gene expression by the respective Smad complex was modeled by an activating (Act1, Act2 Act3; green) and by an inhibitory (Inh1, Inh2, Inh3; red) parameter. Turn: Gene-specific turnover rate. **B.** Sensitivity analysis of the activating and inhibitory parameters on target gene expression. Parameter values were increased (red) or decreased (blue) by two-fold compared to the parameter value estimated for the best fit of the reduced model shown in Figure 2C, and the relative change of the area under the curve of the respective gene was calculated as a measure of the impact of parameter variation on target gene expression.

1.5 Figure S5

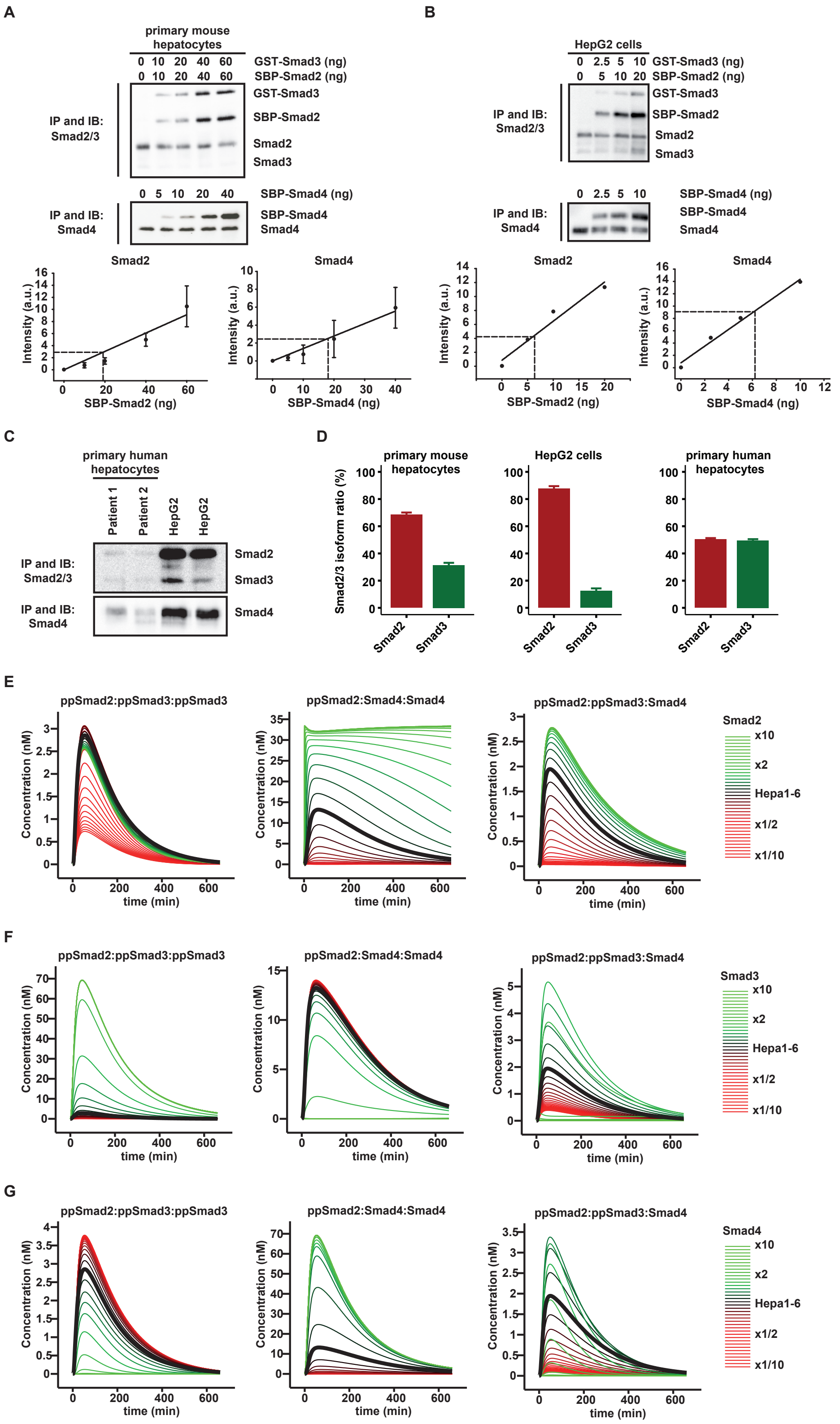


Figure S5 (related to Figure 6): Determination of the total amount of Smad molecules in primary mouse hepatocytes, the HepG2 cell line and primary human hepatocytes and model predictions for the impact of Smad amounts on the dynamics of the complexes

Figure S5 (Continued) Whole cell lysates of primary mouse hepatocytes **(A.)** and HepG2 cells **(B.)** were analyzed for endogenous levels of Smad2 and Smad4 by means of dilution series of added recombinant calibrator proteins SBP-Smad2 and SBP-Smad4, respectively. Protein levels were determined by immunoprecipitation (IP) and subsequent quantitative immunoblotting (IB). Data of GST-Smad3 was not used due to the resulting high background of the immunoblot. Linear regression analysis (solid lines) was performed for the recombinant calibrators SBP-Smad2 and SBP-Smad4. Endogenous protein levels of unstimulated cells (dashed lines) were calculated by means of the respective regression functions. a.u.: arbitrary units. **C.** Whole cell lysates of unstimulated primary human hepatocytes from two patients were analyzed for endogenous levels of Smad2 and Smad4 by immunoprecipitation (IP) and subsequent quantitative immunoblotting (IB). Lysates of HepG2 cells were used as comparison to calculate the total amount of Smad molecules. **D.** Immunoprecipitations of whole cell lysates with antibodies directed against Smad2/3 were performed and Smad isoform ratios were analyzed by quantitative mass spectrometry (assuming 5% measurement error) to determine Smad3 protein abundance (n=3). Error bars represent standard error of the mean. Model simulations indicating the effect of increasing (green) and decreasing (red) Smad2 **(E.)**, Smad3 **(F.)** and Smad4 **(G.)** protein levels relative to the amount in Hepa1-6 cells (black) on the complexes formation of ppSmad:ppSmad3:ppSmad3, ppSmad2:Smad4:Smad4 and ppSmad2:ppSmad3:Smad4. Simulations have been performed for 1 ng/ml TGFβ stimulation. **E.** Increasing amounts of Smad2 enhance the abundance of ppSmad2:Smad4:Smad4 and ppSmad2:ppSmad3:Smad4 and renders the formation of ppSmad2:Smad4:Smad4 from transient to sustained. **F.** Increasing amounts of Smad3 enhance the production of the ppSmad2:ppSmad3:ppSmad3 complex but at the same time reduce the formation of the ppSmad2:Smad4:Smad4 and ppSmad2:ppSmad3:Smad4 complexes. **G.** Increasing amounts of Smad4 enhance formation of ppSmad2:Smad4:Smad4; but this increase at the same time reduces the production of ppSmad2:ppSmad3:ppSmad3 and ppSmad2:ppSmad3:Smad4.

1.6 Figure S6

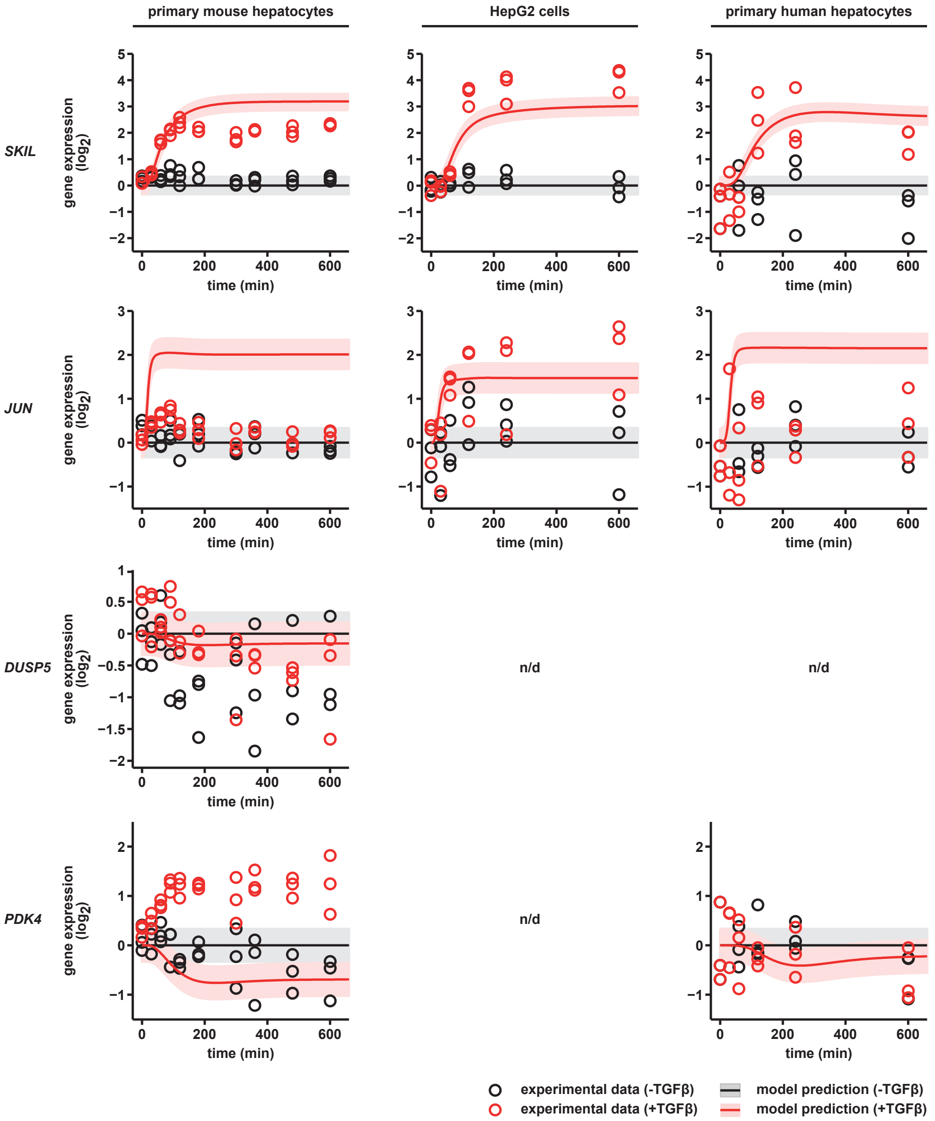


Figure S6 (related to Figure 6): Model predictions for gene expression revealed indirect TGFβ target genes in primary mouse hepatocytes, HepG2 cells and primary human hepatocytes.
Continuous lines represent the trajectories and shaded area the estimated error of model simulations based on the Smad protein abundance (Figure 6A) and the qRT-PCR data (Figure 6B) in primary mouse hepatocytes, HepG2 cells and primary human hepatocytes. Empty circles represent experimental data points (n=3) that were not used for parameter estimation. See STAR Methods for donor anamnesis of primary human hepatocytes. n/d: not detected.

1.7 Figure S7

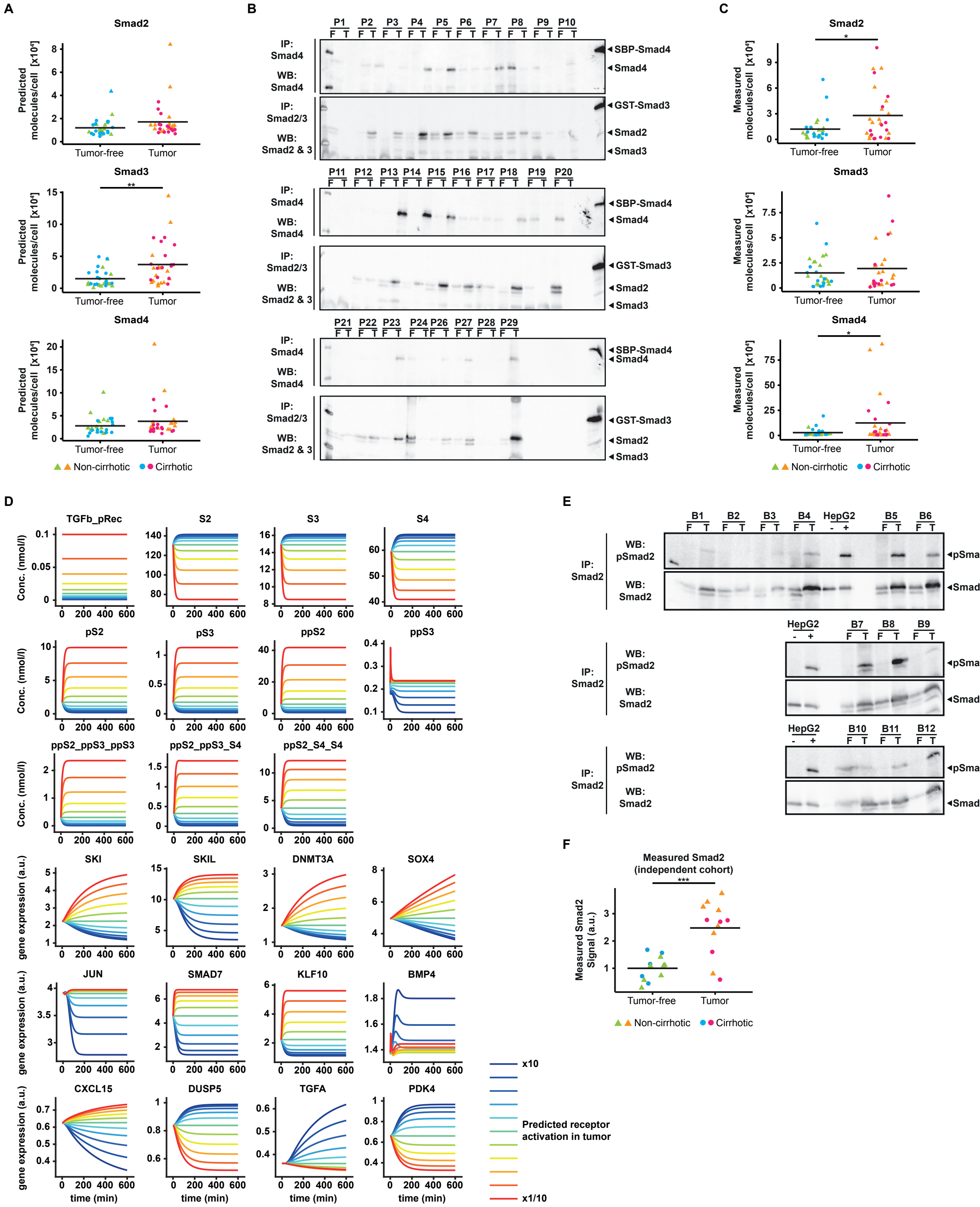


Figure S7 (related to Figure 7): Smad protein abundance and phosphorylation in hepatocellular carcinoma and corresponding non-tumor tissue samples

Figure S7 (Continued) **A.** The relative amount of Smad2, Smad3, Smad4 was predicted by the integrative mathematical model based on the Smad complex formation. Mean signal (crossbar) of tumor-free samples was adjusted to molecules per cell measurements in primary human hepatocytes (Fig. 6) (n=30). $^{*}, p < 0.05$; $^{**}, p < 0.01$; paired t-test. **B.** Proteins were extracted from hepatocellular carcinoma (T) and corresponding tumor-free (F) tissue samples (cohort A). Immunoprecipitation (IP) and quantitative immunblotting (IB) was performed. Calibrator for the Smad proteins were used to normalize the signals, thus making them comparable between different blots. Patient samples (P) 1 to 15 correspond to non-cirrhotic livers, patient samples 16 to 30 to cirrhotic livers. See STAR Methods for donor anamnesis of tissue samples (n=29). **C.** The amount of Smad2, Smad3 and Smad4 was measured by quantitative immunoblotting for each patient. Mean signal (crossbar) of tumor-free samples was adjusted to molecules per cell measurements in primary human hepatocytes (Fig. 6A) (n=29). $^{*}, p < 0.05$; paired t-test. **D.** For the Smad pathway activity in liver tissue, a constant level of active TGFβ receptors at a level of 10% compared to the stimulation setting for Hepa1-6 was assumed. Saturation in the pathway and at the promoter levels was checked in this setup by ten-fold decreasing and increasing the number of active receptors. **E.** Proteins were extracted from hepatocellular carcinoma (T) and corresponding tumor-free (F) tissue samples (cohort B). Immunoprecipitation (IP) and quantitative immunblotting (IB) was performed. Data was normalized to a sample of stimula**F.** The relative amount of Smad2 was measured by quantitative immunoblotting for each patient. Values are expressed relative to the mean signal (crossbar) of the tumor-free samples (n=12). $^{***}, p < 0.001$; paired t-test. a.u.: arbitrary units.

2 Supplemental Tables

2.1 Table S1

Table S1 (related to Figure 6): Smad abundances in different cell types

cell type	Smad2	Smad3	Smad4
	(molecules/cell)	(molecules/cell)	(molecules/cell)
Hepa1-6 cells	825×10^3	83×10^3	402×10^3
primary mouse hepatocytes	100×10^3	30×10^3	85×10^3
HepG2 cells	100×10^3	12.5×10^3	90×10^3
primary human hepatocytes	12×10^3	15×10^3	28×10^3

2.2 Table S2

Table S2 (related to STAR methods): Estimated model parameters for the reduced model extended to gene expression

For the reduced mathematical model, the seven k$_{on}$ parameters of the irrelevant complex parameter were set to 0. In total 108 parameters are estimated from the experimental data. The best fit yields a value of the objective function $-2\log(L) = 213.832$ for a total of 1755 data points. The model parameters were estimated by maximum likelihood estimation. The parameter name prefix sd_ indicates the estimated error of a observable. $\hat{\theta}$ indicates the estimated value of the parameters. θ_{min} and θ_{max} indicate the upper and lower bounds for the parameters. The log-column indicates if the value of a parameter was log-transformed. The fitted-column indicates if the parameter value was estimated (1) or was fixed (0). Parameters highlighted in grey color indicate parameters that were not estimated, while red color indicate parameter values close to their boundaries. 24 parameters representing activation or inhibition strengths were estimated to the presumed lower bound (1×10^{-5}) indicating vanishing effects of the respective complexes on certain genes. Since these parameters did not couple to others, they could be set to zero without changing the fit nor the predicted dynamics.

	name	θ_{min}	$\hat{\theta}$	θ_{max}	log	non-log $\hat{\theta}$	estimated
1	Rec_act	-5	-2.2220	+3	1	$+6.00 \cdot 10^{-03}$	1
2	S2tot	-5	+2.1547	+3	1	$+1.43 \cdot 10^{+02}$	1
3	S3tot	-5	+1.2111	+3	1	$+1.63 \cdot 10^{+01}$	1
4	S4tot	-5	+1.8264	+3	1	$+6.71 \cdot 10^{+01}$	1
5	S_dephos	-5	-0.5387	+3	1	$+2.89 \cdot 10^{-01}$	1
6	S_dephosphos	-5	-1.3055	+3	1	$+4.95 \cdot 10^{-02}$	1
7	S_phos	-5	-0.4202	+3	1	$+3.80 \cdot 10^{-01}$	1
8	Ski_act3	-5	-1.8502	+3	1	$+1.41 \cdot 10^{-02}$	1
9	Ski_act2	-5	-3.0936	+3	1	$+8.06 \cdot 10^{-04}$	1
10	Ski_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
11	Ski_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
12	Ski_inh2	-5	-1.3314	+3	1	$+4.66 \cdot 10^{-02}$	1
13	Ski_inh1	-5	-1.5721	+3	1	$+2.68 \cdot 10^{-02}$	1
14	Ski_turn	-5	-2.3412	+3	1	$+4.56 \cdot 10^{-03}$	1
15	Skil_act3	-5	-0.3347	+3	1	$+4.63 \cdot 10^{-01}$	1
16	Skil_act2	-5	-1.1277	+3	1	$+7.45 \cdot 10^{-02}$	1
17	Skil_act1	-5	-1.5129	+3	1	$+3.07 \cdot 10^{-02}$	1

	name	θ_{min}	$\hat{\theta}$	θ_{max}	log	non-log $\hat{\theta}$	estimated
18	Skil_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
19	Skil_inh2	-5	-0.1233	+3	1	$+7.53 \cdot 10^{-01}$	1
20	Skil_inh1	-5	-0.3871	+3	1	$+4.10 \cdot 10^{-01}$	1
21	Skil_turn	-5	-1.9489	+3	1	$+1.12 \cdot 10^{-02}$	1
22	Dnmt3a_act3	-5	-2.1455	+3	1	$+7.15 \cdot 10^{-03}$	1
23	Dnmt3a_act2	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
24	Dnmt3a_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
25	Dnmt3a_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
26	Dnmt3a_inh2	-5	-1.7140	+3	1	$+1.93 \cdot 10^{-02}$	1
27	Dnmt3a_inh1	-5	-1.4416	+3	1	$+3.62 \cdot 10^{-02}$	1
28	Dnmt3a_turn	-5	-2.4191	+3	1	$+3.81 \cdot 10^{-03}$	1
29	Sox4_act3	-5	-1.9521	+3	1	$+1.12 \cdot 10^{-02}$	1
30	Sox4_act2	-5	-3.5660	+3	1	$+2.72 \cdot 10^{-04}$	1
31	Sox4_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
32	Sox4_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
33	Sox4_inh2	-5	-1.0590	+3	1	$+8.73 \cdot 10^{-02}$	1
34	Sox4_inh1	-5	-0.9873	+3	1	$+1.03 \cdot 10^{-01}$	1
35	Sox4_turn	-5	-3.0958	+3	1	$+8.02 \cdot 10^{-04}$	1
36	Jun_act3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
37	Jun_act2	-5	+0.0083	+3	1	$+1.02 \cdot 10^{+00}$	1
38	Jun_act1	-5	+0.7776	+3	1	$+5.99 \cdot 10^{+00}$	1
39	Jun_inh3	-5	+0.9139	+3	1	$+8.20 \cdot 10^{+00}$	1
40	Jun_inh2	-5	+0.1635	+3	1	$+1.46 \cdot 10^{+00}$	1
41	Jun_inh1	-5	+0.9893	+3	1	$+9.76 \cdot 10^{+00}$	1
42	Jun_turn	-5	-0.9166	+3	1	$+1.21 \cdot 10^{-01}$	1
43	Smad7_act3	-5	+2.9992	+3	1	$+9.98 \cdot 10^{+02}$	1
44	Smad7_act2	-5	-0.6740	+3	1	$+2.12 \cdot 10^{-01}$	1
45	Smad7_act1	-5	+1.3198	+3	1	$+2.09 \cdot 10^{+01}$	1
46	Smad7_inh3	-5	+0.5644	+3	1	$+3.67 \cdot 10^{+00}$	1
47	Smad7_inh2	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
48	Smad7_inh1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
49	Smad7_turn	-5	+1.5612	+3	1	$+3.64 \cdot 10^{+01}$	1
50	Klf10_act3	-5	+2.9998	+3	1	$+1.00 \cdot 10^{+03}$	1
51	Klf10_act2	-5	+1.9569	+3	1	$+9.05 \cdot 10^{+01}$	1
52	Klf10_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
53	Klf10_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
54	Klf10_inh2	-5	-1.7483	+3	1	$+1.79 \cdot 10^{-02}$	1
55	Klf10_inh1	-5	-3.1199	+3	1	$+7.59 \cdot 10^{-04}$	1
56	Klf10_turn	-5	+2.6754	+3	1	$+4.74 \cdot 10^{+02}$	1
57	Bmp4_act3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
58	Bmp4_act2	-5	+1.9563	+3	1	$+9.04 \cdot 10^{+01}$	1
59	Bmp4_act1	-5	+3.0000	+3	1	$+1.00 \cdot 10^{+03}$	1
60	Bmp4_inh3	-5	+3.0000	+3	1	$+1.00 \cdot 10^{+03}$	1
61	Bmp4_inh2	-5	+1.2875	+3	1	$+1.94 \cdot 10^{+01}$	1
62	Bmp4_inh1	-5	+2.3394	+3	1	$+2.18 \cdot 10^{+02}$	1
63	Bmp4_turn	-5	-0.0037	+3	1	$+9.91 \cdot 10^{-01}$	1
64	Cxcl15_act3	-5	+0.9658	+3	1	$+9.24 \cdot 10^{+00}$	1
65	Cxcl15_act2	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
66	Cxcl15_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
67	Cxcl15_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
68	Cxcl15_inh2	-5	+2.5045	+3	1	$+3.20 \cdot 10^{+02}$	1
69	Cxcl15_inh1	-5	+3.0000	+3	1	$+1.00 \cdot 10^{+03}$	1
70	Cxcl15_turn	-5	-2.4845	+3	1	$+3.28 \cdot 10^{-03}$	1

	name	θ_{min}	$\hat{\theta}$	θ_{max}	log	non-log $\hat{\theta}$	estimated
71	Dusp5_act3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
72	Dusp5_act2	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
73	Dusp5_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
74	Dusp5_inh3	-5	-0.3265	+3	1	$+4.72 \cdot 10^{-01}$	1
75	Dusp5_inh2	-5	-1.9383	+3	1	$+1.15 \cdot 10^{-02}$	1
76	Dusp5_inh1	-5	-1.9620	+3	1	$+1.09 \cdot 10^{-02}$	1
77	Dusp5_turn	-5	-1.8497	+3	1	$+1.41 \cdot 10^{-02}$	1
78	Tgfa_act3	-5	-1.8941	+3	1	$+1.28 \cdot 10^{-02}$	1
79	Tgfa_act2	-5	-3.5337	+3	1	$+2.93 \cdot 10^{-04}$	1
80	Tgfa_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
81	Tgfa_inh3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
82	Tgfa_inh2	-5	+0.1420	+3	1	$+1.39 \cdot 10^{+00}$	1
83	Tgfa_inh1	-5	+0.3965	+3	1	$+2.49 \cdot 10^{+00}$	1
84	Tgfa_turn	-5	-2.4431	+3	1	$+3.60 \cdot 10^{-03}$	1
85	Pdk4_act3	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
86	Pdk4_act2	-5	-3.3134	+3	1	$+4.86 \cdot 10^{-04}$	1
87	Pdk4_act1	-5	-5.0000	+3	1	$+1.00 \cdot 10^{-05}$	1
88	Pdk4_inh3	-5	+0.1104	+3	1	$+1.29 \cdot 10^{+00}$	1
89	Pdk4_inh2	-5	-1.1915	+3	1	$+6.43 \cdot 10^{-02}$	1
90	Pdk4_inh1	-5	-0.8540	+3	1	$+1.40 \cdot 10^{-01}$	1
91	Pdk4_turn	-5	-1.8241	+3	1	$+1.50 \cdot 10^{-02}$	1
92	init_Rec	-5	+0.2639	+3	1	$+1.84 \cdot 10^{+00}$	1
93	k_on_223	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
94	k_on_224	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
95	k_on_233	-8	-0.8326	+3	1	$+1.47 \cdot 10^{-01}$	1
96	k_on_234	-8	-3.3944	+3	1	$+4.03 \cdot 10^{-04}$	1
97	k_on_244	-8	-5.0641	+3	1	$+8.63 \cdot 10^{-06}$	1
98	k_on_334	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
99	k_on_344	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
100	k_on_222	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
101	k_on_333	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
102	k_on_444	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
103	kdiss_SS	+0	+0.0000	+3	0	$+0.00 \cdot 10^{+00}$	0
104	pRec_degind	-5	-1.3968	+3	1	$+4.01 \cdot 10^{-02}$	1
105	sd_Bmp4	-5	-0.8506	+3	1	$+1.41 \cdot 10^{-01}$	1
106	sd_Cxcl15	-5	-0.8712	+3	1	$+1.35 \cdot 10^{-01}$	1
107	sd_Dnmt3a	-5	-1.0139	+3	1	$+9.68 \cdot 10^{-02}$	1
108	sd_Dusp5	-5	-0.9758	+3	1	$+1.06 \cdot 10^{-01}$	1
109	sd_Jun	-5	-0.7831	+3	1	$+1.65 \cdot 10^{-01}$	1
110	sd_Klf10	-5	-0.9290	+3	1	$+1.18 \cdot 10^{-01}$	1
111	sd_Pdk4	-5	-1.1309	+3	1	$+7.40 \cdot 10^{-02}$	1
112	sd_Ski	-5	-0.9903	+3	1	$+1.02 \cdot 10^{-01}$	1
113	sd_Skil	-5	-0.7570	+3	1	$+1.75 \cdot 10^{-01}$	1
114	sd_Smad7	-5	-1.1523	+3	1	$+7.04 \cdot 10^{-02}$	1
115	sd_Sox4	-5	-0.8985	+3	1	$+1.26 \cdot 10^{-01}$	1
116	sd_Tgfa	-5	-1.1122	+3	1	$+7.72 \cdot 10^{-02}$	1