

## Supplementary Material for

### Uniformly curated signaling pathways reveal tissue-specific cross-talks and drug target candidates

Tamás Korcsmáros<sup>1,2,\*</sup>, Illés J. Farkas<sup>3,\*</sup>, Máté Szalay<sup>2</sup>, Petra Rovó<sup>1</sup>, Dávid Fazekas<sup>1</sup>, Zoltán Spiró<sup>2</sup>, Csaba Böde<sup>4</sup>, Katalin Lenti<sup>5</sup>, Tibor Vellai<sup>1</sup>, and Péter Csermely<sup>2,+</sup>

<sup>1</sup> Department of Genetics, Eötvös University, Pázmány P. s. 1C, H-1117 Budapest, Hungary

<sup>2</sup> Department of Medical Chemistry, Semmelweis University, PO Box 260, H-1444 Budapest, Hungary

<sup>3</sup> Statistical and Biological Physics Research Group of the Hungarian Academy of Sciences, Pázmány P. s. 1A, H-1117 Budapest, Hungary

<sup>4</sup> Morgan Stanley Hungary Analytics Ltd, Lechner Ö. f. 8, H-1095 Budapest, Hungary

<sup>5</sup> Department of Morphology and Physiology, Semmelweis University, Vas u. 17, H-1088 Budapest, Hungary

\* equal contributions

+ Corresponding author, csermely@eok.sote.hu

See the supporting website for further information: <http://Signalink.org>

## Contents

<b>SUPPLEMENTARY RESULTS AND DATA</b>	<b>2</b>
COMPARISON OF PATHWAYS ACROSS SPECIES	2
<i>Statistical significance tests of cross-talk numbers</i>	3
ANALYSIS OF CROSS-TALKS IN EACH SPECIES	3
MULTI-PATHWAY PROTEINS (PROTEINS CONTAINED BY MORE THAN ONE PATHWAY)	6
<i>Protein numbers and examples</i>	6
<i>Statistical significance tests: Disease-related proteins among multi-pathway proteins</i>	7
CROSS-TALKS IN THE SELECTED HEALTHY TISSUE TYPES AND LIVER CARCINOMAS	7
THE WEBSITE OF SIGNALINK: SEARCH, BROWSE, AND DOWNLOAD OPTIONS	9
<b>SUPPLEMENTARY METHODS</b>	<b>10</b>
DETAILED DESCRIPTION OF THE SIGNALINK DATABASE	10
MANUAL CURATION PROCESS OF SIGNALINK	11
<i>Proteins</i>	11
<i>Interactions</i>	12
<i>Pathways and proteins included into and excluded from Signalink</i>	13
<i>Curation process: Example</i>	14
COMPARING THE APPROACH OF SIGNALINK WITH OTHER APPROACHES	16
COMPARING THE HUMAN SIGNALING PATHWAYS OF KEGG, REACTOME, NETPATH, AND SIGNALINK	17
<i>KEGG and Signalink: A comparison of isoforms and the definition of interacting entities</i>	19
<i>Reactome and Signalink: Complexes and database-specific molecular functions</i>	22
<i>NetPath and Signalink: Interactions, protein identifiers, cited papers, and curation approaches</i>	25
<i>Pairwise comparisons between all four databases: Reactome, NetPath, KEGG, and Signalink</i>	27
<i>Comparing the lists of publications referenced in signaling pathway databases</i>	28
<i>Final conclusions from the database comparisons</i>	30
PROTEIN EXPRESSION IN HEALTHY TISSUE TYPES	31
<i>Selecting the expression data set: eGenetics</i>	31
<i>Selecting the five tissue types</i>	32
<i>Controls</i>	33
STATISTICAL SIGNIFICANCE ANALYSES OF PATHWAY MEMBER PROTEIN FUNCTIONS	34
<i>Gene Ontology term enrichment tests in the curated signaling pathways</i>	34
<i>Gene Ontology term similarity of human multi-pathway proteins and their pathways</i>	37
VALIDATION OF THE HUMAN INTERACTIONS OF SIGNALINK WITH AN EXTERNAL RESOURCE	38
DATABASE UPDATES	39

## Supplementary results and data

### *Comparison of pathways across species*

On the system level one may observe that the worm has more signaling proteins than the fly, which is mostly due to the large number (more than 250) of nuclear hormone receptors (NHR) in the worm. As for single pathways, note that in the worm several proteins orthologous to fly and human JAK/STAT and Hh pathway member proteins are expressed. However, only few interactions of these proteins have been experimentally verified so far (Burglin and Kuwabara, 2006). In other words, experimental evidence cannot support yet that in *C. elegans* these two pathways are functional. It has been suggested that the pathways JAK/STAT and Hh disappeared from the worm during its evolution (Pires-daSilva and Sommer, 2003). A comparison of pathways between the three species shows also that in all three the largest pathway (containing the largest number of proteins) is the EGF/MAPK pathway, while the next three (in the order of protein number) are WNT, TGF, and IGF, respectively. In *C. elegans* and *D. melanogaster* the total weighted number of proteins<sup>1</sup> in these four pathways is approximately the same (154.5 and 117.3 proteins, respectively), while in humans the total is approximately twice this number (295.2). As a contrast, the Hh and Notch pathways have very similar protein numbers in all three species. Note that in humans four (EGF/MAPK, IGF, JAK/STAT, and WNT) of the five largest pathways are kinase-based pathways.

A look at the internal structure of the eight curated signaling pathways reveals further details. See the accompanying website, <http://Signalink.org>, for detailed statistics, a detailed curation manual, and zoomable images of the 8 curated pathways in all 3 organisms. In humans the backbone (core region<sup>2</sup>) of the EGF/MAPK pathway is significantly stronger due to the high number of parallel MAPK cascades. Similarly, the backbone of the WNT pathway is notably larger in humans than in the other two species, which is caused mainly by the high numbers of receptors and ligands. As a contrast, in the IGF, Hh, Notch, and TGF pathways the ratios between the sizes of pathway regions (core and periphery) are almost identical in the three species (in the Notch pathway the core region is small). The JAK/STAT pathway changes strongly across the three investigated species. In *C. elegans* the pathway is not present, in *Drosophila* it has a few proteins (a similar number of core and peripheral proteins) while in humans it is already comparable in size to the largest pathways and its core region is larger than its peripheral region. The large number of human JAK/STAT pathway member proteins and the strong core region of this pathway are mainly due to the high number of ligands (cytokines) and their receptors allowing this pathway to produce a combinatorially large repertoire of immune responses. In other words, the change of the size and structure of the JAK/STAT pathway reflects a major functional change: acquiring a central role in immune responses.

Because of the large number of its experimentally known genetic interactions, the fly has currently many more known interactions among its signaling proteins than the worm. However, experimental approaches have not yet verified a large portion of the genetic interactions in the fly as true physical interactions, and so the evidence provided by genetic interactions is less direct than other evidence types. Thus, before inserting genetic interactions into Signalink, we have carefully filtered them, mostly by searching for other types of

---

<sup>1</sup> A protein participating in  $n$  pathways adds  $1/n$  to the weighted number of proteins in each of these pathways.

<sup>2</sup> For an explanation of the term “core” see below in this document under “Manual curation process of Signalink”.

evidence available for the same interactions. See below under “Manual curation process of SignaLink” for a description for interaction evidence types. As a result in SignaLink the signaling interactions of the fly and the worm have similar levels of confidence and their total numbers are close too.

As the fly is a more complex organism than the worm, one may expect that the number of high-confidence protein-protein interactions in it, including directed signaling interactions, is larger than in the worm. Consequently, we expect that many of the genetic interactions in the fly will be experimentally verified in the near future as true physical protein-protein interactions, many of them as cross-talks (*i.e.*, signaling interactions between pathways).

### Statistical significance tests of cross-talk numbers

We have determined the adjusted Wald Confidence Interval to find the pairs of pathways between which pathway-pathway connections (*i.e.*, cross-talks) are more frequent than the others (Agresti and Coull, 1998). With the adjusted Wald confidence interval we determine the 95% confidence interval, supposing binomial distribution of the cross-talks. If the number of cross-talk lies out of this 95% confidence interval, we may say that the particular pathway connections are more frequent. Supplementary Table 1 shows the relative frequencies of the interaction numbers between pairs of human signaling pathways in SignaLink. Note that this table shows only the 10 highest scoring directed pairs of pathways. Based on the full list of 56 ( $= 8 * 7$ ) directed pairs of pathways the average of the relative frequency is 1.79% the upper border of the adjusted Wald confidence interval lies at 7.27%.

Directed cross-talk between pathways		Number of interactions	Relative frequency	Statistically significantly overrepresented
IGF	EGF/MAPK	36	3.10%	
EGF/MAPK	IGF	50	4.31%	
EGF/MAPK	WNT	53	4.57%	
TGF	WNT	57	4.91%	
WNT	EGF/MAPK	65	5.60%	
JAK/STAT	EGF/MAPK	83	7.16%	
EGF/MAPK	JAK/STAT	90	7.76%	
WNT	TGF	90	7.76%	
TGF	EGF/MAPK	104	8.97%	
EGF/MAPK	TGF	160	13.79%	

**Supplementary Table 1.** Wald statistical scores of the interaction numbers between the 8 human signaling pathways in the SignaLink database. Only the 10 directed pathway pairs with the largest scores are shown here. Relative frequencies above 7.25% are statistically significant (see text above for details).

### Analysis of cross-talks in each species

In this section we first define *cross-talks*. Then we describe and discuss active pathways in the three organisms. Last, we show examples for cross-talking pairs of pathways.

The *classical definition* of cross-talk assumes that each protein is contained by exactly one signaling pathway. According to this definition, if proteins A and B belong to different

signaling pathways, then the signaling *interaction* between proteins A and B is a *cross-talk between their signaling pathways*. SignalLink – and several other databases too – provide a more precise mapping of signaling pathways and allow that a signaling protein belongs to more than one pathway. Therefore, the classical definition of cross-talk has to be extended. A signaling interaction between two proteins, A (contained by pathways  $a_1, a_2, \dots$ ) and B ( $b_1, b_2, \dots$ ), is at the same time a cross-talk between each pair of different a-b pathways, *i.e.*, between any pathway pair  $a_i$ - $b_j$ , supposed that these two pathways are different. Of course, the signaling interaction between proteins A and B is a signaling interaction within each such  $a_i$  pathway that is listed among the  $b_j$  pathways as well.

As an example consider a signaling network containing a total of four signaling proteins – A, B, C, and D – and two directed signaling interactions:  $A \rightarrow B$  and  $C \rightarrow D$ . Assume that in this example network protein A is contained by the pathways EGF/MAPK and Hh, protein B is contained by the pathways EGF/MAPK and WNT, protein C is contained only by the IGF pathway, and protein D is contained only by the Notch pathway. The directed signaling interaction between proteins A and B is at the same time a (directed) cross-talk between the pathways EGF/MAPK and WNT, a (directed) cross-talk between the pathways Hh and EGF/MAPK, a (directed) cross-talk between the Hh and WNT pathways, and it is also a (directed) signaling interaction inside the EGF/MAPK pathway. The signaling interaction between proteins C and D is a (directed) cross-talk between the IGF and Notch pathways.

Note that if a unit weight is assigned to each of these five (directed) pathway-pathway interactions (one cross-talk between each of the three directed pathway pairs  $EGF/MAPK \rightarrow WNT$ ,  $Hh \rightarrow EGF/MAPK$ , and  $Hh \rightarrow WNT$ , and one directed interaction inside the EGF/MAPK pathway due to the interacting (directed) protein pair  $A \rightarrow B$ ; one cross-talk between the directed pathway pair  $IGF \rightarrow Notch$  due to the interacting (directed) protein pair  $C \rightarrow D$ ), then the total weight of the signaling interaction of the (directed) protein pair  $A \rightarrow B$  would be 4 compared to the total weight of the signaling interaction of the (directed) protein pair  $C \rightarrow D$ , which would be only 1. This would incorrectly overestimate the biological significance of the first pair ( $A \rightarrow B$ ) compared to that of the second pair ( $C \rightarrow D$ ). For most interacting pairs of signaling proteins there is no or little *quantitative* information about the total (system-level) effect of their signaling interaction. Thus – due to the absence of such information – we propose that each signaling protein-protein interaction should add the same total weight, 1, to pathway-pathway interactions. Note that this total weight includes both cross-talks ( $a_i$  and  $b_j$ , the selected pathways of the two interacting proteins, are different pathways) and interactions within pathways ( $a_i$  and  $b_j$  are the same pathway).

If in a signaling interaction at least one of the two interacting proteins belongs to more than one pathway, then usually there is no or little *quantitative* information about which one of the connected pathway pairs is linked more strongly by this pair of interacting proteins. Therefore, we propose that if a protein belongs to  $n$  pathways and another protein that it interacts with through a directed signaling interaction belongs to  $m$  pathways, then the directed signaling interaction between these two proteins should add the same weight,  $1 / (n * m)$ , to the cross-talk weight between any two of the directed pathway pairs connected by this directed signaling interaction. In the example network (defined above) the directed signaling interaction of proteins A and B adds a cross-talk weight of  $1/4$  to the three (directed) cross-talks  $EGF/MAPK \rightarrow WNT$ ,  $Hh \rightarrow EGF/MAPK$ ,  $WNT \rightarrow TGF$ , and the interaction of the EGF/MAPK pathway with itself, while the directed signaling interaction of proteins C and D adds a cross-talk weight 1 to the cross-talking (directed)  $IGF \rightarrow Notch$  pathway pair.

Note that since protein-protein interactions are directed, two oppositely directed interactions between the same proteins, *i.e.*,  $A \rightarrow B$  and  $B \rightarrow A$ , are counted separately and contribute separately to the interaction weights (cross-talks and within-pathway) between the directed pairs of their pathways.

## Active pathways and cross-talking pathways

With an increasing complexity of the investigated organisms the number of active pathways increases and the number of cross-talking pathways also increases. In other words, the cross-talk map of the organism becomes denser.

In *C. elegans* only six of the eight curated pathways are active. In addition, the network of pathways is sparse: only 5 of the 30 possible cross-talk types (connecting the 6 active pathways) are present (see the supporting website for detailed statistics). Interestingly, the number of cross-talks in the two possible directions between the same pair of pathways is often different. For example, in *C. elegans* – according to Signalink – the weighted number of directed signaling interactions pointing from the EGF/MAPK pathway to the IGF pathway is 1.5, while in the opposite direction (IGF  $\rightarrow$  EGF/MAPK) there is no signaling interaction. Two further examples are (i) the cross-talks between the Hh and WNT pathways in *D. melanogaster* (for Hh  $\rightarrow$  WNT the weighted number of directed interactions is 6.9 and for WNT  $\rightarrow$  Hh it is 3.1) and (ii) the human EGF/MAPK and TGF pathways where the weighted number of directed signaling interactions pointing towards the EGF/MAPK pathway is much higher as in the opposite direction (47.2 vs. 28.0).

In *Drosophila* all eight curated pathways of Signalink are active (see Fig. 2b of the main text), however, the NHR and JAK/STAT pathways are isolated. Without these two pathways the cross-talk network is already relatively dense: 16 of the total 30 possible cross-talk types are present.

In humans all eight curated signaling pathways are active and all 28 ( $= 8 * 7 / 2$ ) pathway pairs are connected by cross-talks. As signaling interactions are directed, the 28 pairs of pathways allow 56 ( $= 8 * 7$ ) types of directed cross-talk. Of these 56 only 6 are inactive. The highest (weighted) number of cross-talks is between EGF/MAPK and the other four large pathways: WNT, TGF, IGF, and JAK/STAT. In contrast to the other two species, the NHR pathway is active and cross-talks with all other pathways, which is mainly due to the signaling interactions between Nuclear Hormone Receptor proteins and the transcription factors of the other pathways.

## Examples: Cross-talks in healthy human tissues

To underline the importance of signaling cross-talk in the three investigated species, we list here first a few general examples about the role of cross-talk in organ development. Then we provide examples for cross-talk between two core proteins, a core and a peripheral protein, and two peripheral proteins.

In *C. elegans* EGF/MAPK and WNT signaling can cooperate to establish embryonic polarity and to specify cell fate. The decision between reproductive and dauer development is jointly controlled by TGF and IGF signaling, while vulval development is determined by the cross-talks between EGF/MAPK, WNT, and Notch signaling (Hanna-Rose and Han, 2000). The

patterning of the *Drosophila* retina is cooperatively controlled by multiple pathways: Hh, Decapentaplegic (Dpp) from the TGF-beta family, Notch, and EGF/MAPK. Similarly, the development of the wing is controlled jointly by Hh, Dpp, WNT, and EGF/MAPK (Baker, 2007). Recent studies show that the development of human hematopoietic stem cells is jointly controlled by the Hh, JAK/STAT, Notch, TGF, and WNT pathways (Blank et al., 2008).

Interestingly, a cross-talking pathway may interact through not just one, but many of its pathway sections (receptor, mediator, *etc.*) and both pathway regions (core, periphery). In humans ALK-6 and ROR2 are core proteins of the TGF and WNT pathways, respectively, and they are both receptors (Sammar et al., 2004). Similarly, EGF/MAPK and IGF cross-talk through their core proteins GRB2 and IRS1 (Claeys et al., 2002), two receptor adaptor proteins, while EGF/MAPK, Hh, and WNT cross-talk through their core proteins RSK1 and GSK3 (De Mesquita et al., 2001), which are mediators (kinases). A cross-talk involving a core and a peripheral protein is the interaction between PIAS4 (JAK/STAT) and SMAD4 (TGF) (Long et al., 2003), and a cross-talk between two non-core proteins is the interaction of CTBP1 (Notch) and HIPK2 (TGF) (Zhang et al., 2003).

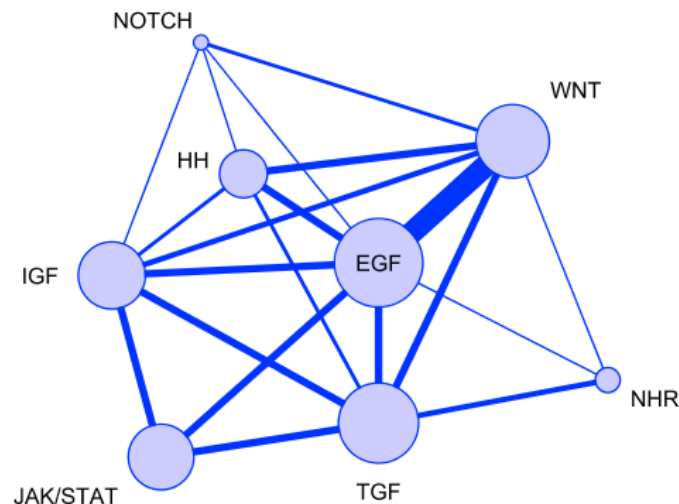
### ***Multi-pathway proteins (proteins contained by more than one pathway)***

#### **Protein numbers and examples**

During the curation process we found that several proteins are contained by more than one signaling pathway. We call these proteins here multi-pathway proteins. We note that multi-pathway proteins and their possible functions were reported in (Komarova et al, 2005). In the worm, the fly, and humans we found a total of 6, 12, and 62 such proteins, respectively. In humans the eight curated pathways in Signalink have the following numbers of multi-pathway proteins (*i.e.*, proteins that are contained by at least one other pathway too): EGF/MAPK (41), Hh (9), IGF (17), JAK/STAT (16), NHR (3), Notch (2), TGF (29), and WNT (22). In summary, within one human signaling pathway the ratio of multi-pathway proteins varies from 5% (Notch) to 46% (IGF).

Supplementary Figure 1 shows the eight human signaling pathways curated in Signalink and the numbers of multi-pathway proteins. Observe that EGF/MAPK – the pathway containing the largest number of proteins – is the only pathway that shares proteins with all other pathways. On the other end of the spectrum are the Notch, JAK/STAT, and NHR pathways; they share proteins only with 3 or 4 other pathways.

We list now a few examples for human multi-pathway proteins. First, the scaffold protein Axin and the kinase GSK3 are both central components, *i.e.*, core proteins, in more than one signaling pathway (Luo and Lin, 2004; Frame and Cohen, 2001). Second, the WNT and Hh pathways both contain the same essential ubiquitin ligase complex (Maniatis, 1999). Moreover, ubiquitin ligases highly similar to this complex tune the TGF and Notch pathways (Itoh and ten Dijke, 2007; Itoh et al., 2003).



**Supplementary Figure 1.** In the 8 human signaling pathways curated in SignaLink there are 62 multi-pathway proteins. In this figure each node represents one human signaling pathway. The size of a node is proportional to the number of proteins contained by the pathway (represented by the node) and the width of a link between two nodes is proportional to the number of proteins contained by both pathways. See the text for details.

### Statistical significance tests: Disease-related proteins among multi-pathway proteins

Khi-square tests showed the significance of known disease-related proteins (Amberger et al., 2009; Becker et al., 2004; Dennis, Jr. et al., 2003) among multi-pathway proteins in different sets of proteins: The set of 62 human multi-pathway proteins is enriched with disease-related proteins: 45% (28) of them are known to be disease-related, while in the 8 human signaling pathways only 25.5% (165 of 646):  $\chi^2 = 10.9817$ ;  $p = 0.0009$ . The ratio of multi-pathway proteins among all human proteins listed by Ensembl is only 20% (3,929 of 19,534):  $\chi^2 = 24.0607$ ;  $p = 0.9 \cdot 10^{-6}$ .

### *Cross-talks in the selected healthy tissue types and liver carcinomas*

According to eGenetics (Kelso et al., 2003) in the investigated healthy human tissues 40% (muscle) to 52% (colorectal) of the signaling proteins listed by SignaLink are expressed. We obtained protein expression in two screens of liver carcinomas (Wurmbach et al., 2007; Chen et al., 2002) from Oncomine (Rhodes et al., 2007). In these two liver carcinomas we considered a protein differentially expressed if the p value of its expression in at least one of the two screens – compared to healthy liver tissues and computed by a t-test of Oncomine – was below 0.05. We found that 46% of all SignaLink proteins have a modified expression compared to healthy liver tissue. Despite the similar ratios of active signaling proteins and the possibility of cross-talk between any two pathways, cross-talks are active in a tissue-specific manner. This flexibility allows the development and maintenance of different phenotypes, even in extreme cases such as cancer.

We continue with two examples for the tissue-specific expression of cross-talk types. First, observe in Fig. 3a of the main text that (in the selected tissue types) between the following

pairs of large pathways the ratio of active signaling cross-talks is lower than the average: EGF/MAPK-IGF, IGF-JAK/STAT, EGF/MAPK-JAK/STAT, IGF-TGF, and IGF-WNT. Taking the pathway pair EGF/MAPK-IGF as an example, we found only two primary publications reporting signaling cross-talk between them (Roudabush et al., 2000; Vardy et al., 1995). The second sexample, standing in contrast to the first, is the following. For 3 larger (and several smaller) pairs of pathways cross-talk is more frequent in the investigated tissue types than one would assume from the sizes of the two cross-talking pathways: EGF/MAPK-NHR, NHR-TGF, and NHR-WNT. These fields are colored red in the chart on Fig. 3a. One reason for the higher than average ratio of active cross-talks between these four pathways is that SMAD3, a likely key protein in fibrogenesis (Flanders, 2004; Roberts et al., 2001), participates in all four of them and it is expressed more often in the selected tissue types than the average signaling protein.

Next, we discuss the figure on liver carcinomas (Figure 3b) of the main text. In general liver carcinomas are known to involve genetic mutations in the WNT and JAK/STAT pathways (Takigawa and Brown, 2008; Niwa et al., 2005), an increased activity of the EGF/MAPK and IGF pathways and their cross-talk (Desbois-Mouthon et al., 2006), and also an increased activity of the JAK/STAT pathway (Calvisi et al., 2006). A less known result is that several proteins of the Hh pathway are activated in human HCCs (hepatocellular carcinomas) (Huang et al., 2006). In addition to these Figure 3b of the main text shows the amount by which each cross-talk type is related to the investigated liver carcinomas. The cross-talks of EGF/MAPK with most other pathways are frequently changed (turned on or off), similarly to the interactions within the group EGF/MAPK-WNT-TGF-IGF. On the other hand, the cross-talk activity of NHR and Notch changes only weakly in the analyzed human liver carcinomas. (The low number of cross-talks of the Notch pathway may distort the results for this pathway.)

Finally, we concluded that the pathways EGF/MAPK, JAK/STAT, and Notch are clear examples for three distinct types of signaling behavior: (i) high expression in normal tissue types and strong changes in cancer, (ii) high expression, but small changes, and (iii) low expression with small changes. Below, we show literature support for these 3 major claims.

The high expression of the EGF/MAPK pathway in all selected tissue types corresponds to its central role in the evolution and normal function of complex signaling networks (Manning et al., 2002). Moreover, deregulation of its pathway members is the major cause of several cancer types (Amit et al., 2007). The JAK/STAT pathway is the main pathway utilized by immunologically important receptors, *e.g.*, cytokine receptors. These receptors are present in many tissue types, and similar to other highly expressed JAK/STAT pathway members they receive and transmit signals from immune cells to the given cell's nucleus (Haan et al., 2006). The Notch pathway is an important pathway for cell differentiation but less important in mature cell types, therefore most of its pathway members – many that are responsible only for the regulation of the pathway – are not expressed in the selected tissue types (Bolos et al., 2007). However, the Notch pathway can be crucial in the development of liver cancer: a recent study showed that in normal cells it could inhibit other oncogenic pathways (*e.g.*, WNT) but in liver cancer major components of the Notch pathway is downregulated (Wang et al., 2009b).



## ***The website of Signalink: Search, Browse, and Download options***

At <http://signalink.org> the user can **search** for signaling proteins and interactions of *C. elegans*, *D. melanogaster*, and humans by protein name, UniProt ID (or accession), and species-specific IDs (WormBase ID, FlyBase ID, and Ensembl protein ID). Result pages contain the following information (all hyperlinked): protein IDs (protein name, UniProt ID, UniProt accession, and species-specific IDs), the roles of proteins (pathway sections: ligand, receptor, mediator, co-factor, transcription factor, and other; and pathway regions: core and peripheral), orthologous proteins, and interaction types (incoming, outgoing, within pathways, and between pathways, *i.e.*, cross-talk).

The user can **browse** the pathways in each species and view zoomable high-resolution maps (networks) of each pathway colored by (i) pathway sections or (ii) by pathway regions or (iii) by ortholog information. For pathway sections and regions networks were exported from Cytoscape 2.5.2 using the Cytoscape plugins Golorize (Release date: Feb. 12, 2007) and BiNoM (1.0). Node colors are listed on the website. A pie chart (with Golorize) indicates that the given protein belongs to more than one pathway or more than one pathway section. We used the following parameters for the force-directed layout algorithm producing the networks. Default spring coefficient and length: both -10, node mass: 6, number of iterations: 1000, selected integration algorithm: Runge-Kutta (minimum and maximum edge weights: 0 and a very large number, respectively).

With the **ortholog network viewer** the user can explore all 8 curated signaling pathways separately or together in each of the 3 species. This viewer shows the predicted pathway member proteins too. In all visualizations direct and indirect interactions are shown as normal and dashed lines, respectively. Arrows and blunted arrows stand for stimulation and inhibition, respectively.

You can **download** from <http://signalink.org> the curated signaling pathways of each species separately, all pathways of one selected species, and all pathways of all three species. Available formats are the following. (i) *All interactions* in a simple Excel file; (ii) *signaling proteins and interactions* in these formats: MS Excel, CSV (comma separated values), SQL dump, SBML (Systems Biology Markup Language), CYS and SVG files exported from Cytoscape, and PNG formatted images; (iii) *multi-pathway proteins* listed in a plain text file with TAB-separated fields; (iv) *predicted pathway member proteins* in an MS Excel file; (v) the *list of review papers* used for compiling Signalink, grouped by pathways, in RefMan, MEDLINE, and BibTeX formats; (vi) the list of most likely *drug target candidates*, sorted by the number of analyzed properties that make them possible drug targets. Note that in the downloadable Cytoscape CYS files proteins functioning in more than one signaling pathway are colored pink, because currently the parameters of Golorize cannot yet be fully exported into CYS files.

## Supplementary methods

### *Detailed description of the SignaLink database*

SignaLink uses 8 biochemically defined pathways important in metazoans: EGF/MAPK, Insulin and Insulin-like Growth Factor (IGF), TGF- $\beta$  (TGF), Wingless (WNT), Hedgehog (Hh), JAK/STAT, Notch, and Nuclear hormone receptors (NHR). Instead of focusing on one particular function or phenotype, SignaLink defines each pathway by its unique combination of biochemical characteristics. For example, the JAK/STAT pathway has cytokine receptors, it is dominated by non-receptor tyrosine kinases, and its proteins frequently dimerise. Proteins are assigned to pathways (i) based on reviews of pathways and protein families and (ii) using biochemical interactions between proteins which are manually identified from the original publications reporting experimental evidence.

For each protein SignaLink stores four identifiers: (i) the UniProt ID of the protein (*e.g.*, P92172), (ii) the name of the gene (*e.g.*, *daf-7*), (iii) a species-specific gene ID (for *C. elegans* and *D. melanogaster*) or the protein's Ensembl ID (for humans), and (iv) an internal database ID (a positive integer). In addition to these each protein in SignaLink is manually tagged with (i) the pathway(s) it participates in (one or more of the 8 pathways curated in SignaLink), (ii) its pathway region(s) (core, peripheral) in each of its pathways separately and (iii) its pathway section(s) (one or two of: ligand, receptor, mediator, co-factor, transcription factor, and other).

In SignaLink we allowed a maximum of two pathway sections for each protein. A few specific cases and consequences are the following. First, we note that, strictly speaking, in the Notch pathway most proteins have more than two roles: receptor, mediator, and transcription factor. However, we labelled Notch proteins only with the pathway sections receptor and transcription factor, and not with mediator, because when Notch proteins are mediating they have almost no signaling interactions. Second, in the Nuclear hormone receptor pathway we listed only the actual NHR proteins (each both as receptor and transcription factor) and considered each NHR signaling protein to belong to the core. No ligands were included, because in the NHR pathway ligands are not proteins, but lipids. We also skipped the co-factors of the NHR pathway, because their function is nuclear regulation and not the adjustment of the signaling flow. We believe that the co-factors of nuclear hormone receptors and other nuclear factors, *e.g.*, chromatin remodelling factors, should be examined in the transcription regulatory network, not in the network of signaling interactions, thus, they belong to another type of cellular database.

The pathway region of a protein indicates whether it is essential in the signaling flow of the given pathway. Based on experimental evidence we allowed during the curation process explicitly that a protein belongs to more than one pathway and even (this happened rarely) that it is both a core and a peripheral component within the same pathway. As an extreme case the human protein SMAD3 belongs to four pathways and five regions within those four pathways: EGF/MAPK (core), WNT (both core and peripheral), TGF (core), and NHR.

In SignaLink each interaction is manually tagged with the PubMed ID of the original publication reporting experimental evidence for the interaction. This is necessary for selecting interactions with experimental evidence from the large set of 'common knowledge' (but often not sufficiently verified) interactions used in many reviews (Cusick et al., 2009). All interactions in SignaLink are directed. For most of the interacting protein pairs in SignaLink

the unidirectional (only  $A \rightarrow B$ ) and for a few pairs it is bidirectional (both  $A \rightarrow B$  and  $B \rightarrow A$ ). The interaction type descriptors ‘direct’ and ‘indirect’ indicate whether there is biochemical (direct) or other type of evidence (indirect), respectively. For details see the next section.

## *Manual curation process of Signalink*

All pathways of the three species were compiled (*i.e.*, manually curated) separately. Signalink contains no interologs (Cusick et al., 2009; Yu et al., 2004). For each pathway, we performed the following three main steps. First, we searched for pathway-specific review articles and databases. Next, we added to each pathway the signaling proteins listed as the members of that pathway in the review papers. Finally, we searched again for additional signaling interactions of the listed pathway proteins.

## **Proteins**

Data on proteins were collected as follows.

- ORFs and Ensembl protein IDs from species-specific databases: WormBase (version 191) (Rogers et al., 2008), FlyBase (version 2008.6) (Drysdale, 2008), and Ensembl (version 49) (Flicek et al., 2008).
- UniProt IDs from UniProt (version 87) (Boutet et al., 2007). If more than one UniProt ID was available for the same protein, then the ID(s) of the protein(s) with the longest amino acid sequence was (were) used.
- Protein name synonyms were listed from review papers, *e.g.*, Ref. (Kyriakis and Avruch, 2001), and the “synonym” field of the iHOP database (May 2008 version) (Fernandez et al., 2007).
- For protein ID conversions we used also PICR, the Protein Identifier Cross-Reference Service (Cote et al., 2007) and the Synergizer (Berriz and Roth, 2008).
- For each protein we listed its orthologs in the other two species with the help of the ortholog clusters of the InParanoid database (version 6.1) (Berglund et al., 2008). Consider an InParanoid cluster containing protein  $P_A$  from species A and proteins  $P_{B,1}$  and  $P_{B,2}$  from species B. From the  $P_B$  proteins in the cluster we included into Signalink only those with an InParalog score equal to or above 0.3.
- When inserting a protein into Signalink we assigned it to one pathway and – within that pathway – to one pathway region (*e.g.*, EGF/MAPK core, TGF non-core, or NHR core). Later further pathways and pathway regions were added for this protein, if necessary.
  - We marked a protein a ‘core’ component of a pathway, if it is essential for transmitting the signal of its pathway and has at least one of the pathway’s biochemical characteristics, *e.g.*, ‘Ser/Tyr-kinase activity’.
  - ‘Non-core’ (same as ‘peripheral’) indicates that the protein does not belong to the core of the pathway, but it modulates the pathway’s core proteins.
- We determined the pathway section(s) of each protein separately, *e.g.*, ligand or receptor. A maximum of two sections per protein were allowed.
  - The pathway position **ligand** indicates that the protein starts the signal of its pathway.
  - A **receptor** is the immediate receiver of this signal.
  - A **mediator** is a member of the pathway mediating the signal from the receptors towards transcription factors.

- A **co-factor** modulates the function of any other protein. Note that co-factors are often in the peripheral (non-core) region of their pathways.
- **Transcription factor**: (i) transmits the signal received from its pathway towards another transcription factor (TF) protein, or (ii) forms a complex with other TF proteins, or (iii) binds to a specific promoter region on the DNA.
- **Other**: non-signaling proteins with the roles of cellular motion, transport, and membrane anchoring. In the worm, the fly, and humans SignaLink contains 14, 10, and 13 proteins, respectively annotated with the pathway position “other”. Out of these proteins 3, 1, and 5, respectively are co-factors at the same time too, and the remaining 11, 9, and 8 proteins have no additional pathway position, only “other”. In summary, in SignaLink the pathway position “other” is used rarely, and it appears in combination with co-factor only.
- **Unknown**: the position of this signaling protein in its pathway is not known.
- After compiling the lists of proteins and signaling interactions we performed two further steps:
  - We went through the list of proteins again and for each protein, P, we listed all those interacting proteins, R, for which SignaLink contains a  $P \rightarrow R$  directed interaction. We added to the list of pathway memberships of P all pathways of each of these (R) proteins.
  - For each signaling protein in one species (A) we listed its orthologs from the other two species (B, C). For each of these orthologs we checked whether it has been found to interact with signaling proteins already listed in SignaLink (for species B or C). If yes, then we inserted the ortholog into SignaLink too.

## Interactions

We used the following rules for collecting data on interactions.

- We collected the signaling interactions of a protein from primary research articles listed in
  - review papers,
  - species-specific databases (FlyBase, WormBase) and UniProt,
  - iHOP, ChiliBot (Chen and Sharp, 2004), and PubMed search results.
- Details:
  - Each interaction inserted into SignaLink had to be directed.
  - We made an effort to find the earliest possible experimental article describing the interaction.
  - For all review papers and approximately 70% of all primary experimental articles, we examined not only the abstract, but also the main text.
  - In SignaLink the experimental *evidence* for a signaling interaction between the two involved proteins is *either direct or indirect*:
    - *Direct experimental evidence*. There is published biochemical evidence for the signaling interaction between the two proteins.
    - *Indirect experimental evidence*. There is no direct biochemical evidence for an interaction, however, published experimental results suggest that the interaction is very likely possible. Evidence types accepted here are (1) changes in mRNA/protein levels, enzyme activities, concentrations of the products of catalyzed reactions, and (2) docking domain structures.
  - The *effect of an interaction* can be activating or inhibitory. For interactions with indirect evidence, ++ and -- mean activating, while +- and -+ mean inhibitory interaction. A unidirectional interaction (A and B interact as either

A→B or B→A) has only one type of effect, but for the few bidirectional interactions (A→B and B→A are both present) more than one type of effect is possible between the two proteins.

- Two signaling interactions between the same two proteins in opposing directions are listed separately in SignalLink.

## **Pathways and proteins included into and excluded from SignalLink**

Signal transduction – briefly summarized – is the transmission of extracellular signals towards the nucleus. Based on this definition some protein families, pathways, and cellular systems were fully or partially excluded from SignalLink.

Based on the summary of Gerhart (Gerhart, 1999) Pires da Silva and Sommer selected (Pires-daSilva and Sommer, 2003) seven signaling pathways that are continuously active during development, have an essential role after the developmental processes, and are conserved in metazoans. We selected the same seven pathways and separated the RTK pathway group to EGF/MAPK and Ins/IGF.

Several further signaling pathways used by Gerhart (for example, the integrin pathway or the apoptosis pathway), are active only occasionally (not continuously) and contribute mainly to adult physiology. These were excluded by Pires da Silva and Sommer. We excluded these pathways from SignalLink too. In addition to their late activity, the excluded pathways are often controlled by input signals that are not extracellular. Examples are the TOR and apoptosis pathways, some immune pathways (*e.g.*, the NF-κB pathway), and cell cycle-related pathways such as the p53 system.

For an additional protein to be included into SignalLink during the manual curation process, we required that (1) it has an experimentally verified role in one of the eight curated pathways or (2) it has a signaling interaction with an already listed SignalLink protein.

The following cellular systems perform tasks other than signal transduction, thus only those of their proteins are present in SignalLink that have a direct connection to signaling proteins:

- Clathrin- and non-clathrin-mediated endocytosis
- Nuclear import/export system
- Chromatin remodelling
- The ubiquitin system

Interactions between transcription factors (TFs) are mostly well-separated from signal transduction (ST) due to their different time scale, mechanisms, and cellular location (Papin et al., 2005). Thus, most TF-TF interactions have been omitted from SignalLink. Note that signal transduction pathways end on TFs, thus, many proteins are contained by both the ST system and the transcription regulation system. We stopped at the first TF of a pathway and did not insert its neighbors into SignalLink. If we found that several TFs directly interact with non-TF signaling proteins and these TFs interact themselves, then we inserted these TF-TF interactions into SignalLink too (*e.g.*, between STAT3 and SMAD3). However, the co-factors of the NHR pathway were excluded from SignalLink, because they are members of the TF network and not the ST system.

From the following protein families only those members have been included into Signalink for which we could identify an experimentally verified role in at least one of the eight selected signaling pathways or an interaction with a Signalink protein:

- Heterotrimeric G proteins,
- Non-receptor tyrosine kinases,
- Receptor tyrosine kinases (RTK).

From the 58 currently known RTK receptors Signalink contains only 4 EGF and 4 IGF receptors (4 EGFR, 2 IGFR, IRR, and INSR) for the following reasons.

- From the RTK pathway group, EGF/MAPK and IGF/INS are two pathways with sufficiently large sizes for comparisons.
- Inserting all 58 RTK receptors and associated proteins (approximately 3-4 ligands and mediators per receptor) would merge the currently largest pathway, EGF/MAPK, with Ins/IGF, resulting in a single pathway significantly larger than all other pathways (Hh, JAK/STAT, NHR, Notch, TGF, and WNT) combined. Thus, inserting all RTK receptors and associated proteins would strongly reduce the usability of pathway comparisons.
- In experimental work EGF and IGF receptors are often discussed separately. They are considered to be the two most important and representative receptor groups of the RTK family. Within the RTK pathway system EGF and IGF receptors are two of the most frequently researched receptor groups and they have clearly different functional roles.

Note that despite their clearly different roles, the downstream effects of the two receptor groups (EGF and IGF) do have similarities: IGF receptors can also forward signals to the MAPK section of the EGF/MAPK pathway (Robinson et al., 2000).

### **Curation process: Example**

As an example we discuss here the human Notch pathway and one of its components, the human NOTCH1 protein.

- According to pathway-based reviews (Baron, 2003), in humans there are 4 Notch family member proteins: NOTCH1, NOTCH2, NOTCH3, and NOTCH4.
- Notch proteins have a specific role in transmitting signals (Weinmaster, 1997).
- The ligands and transcription factors between which Notch proteins transmit signals as well as several further proteins tuning the functions of Notch proteins are known (Bray, 2006).

### **Protein data**

- Most databases and review papers do not differentiate between splice variants. For the human NOTCH1 protein Ensembl (Flicek et al., 2008) contains two splice variants: ENSP00000277541 and ENSP00000360765. From these two the InParanoid database (Berglund et al., 2008) contains only the first, ENSP00000277541, therefore, we inserted into Signalink only this splice variant. For proteins that have more than one splice variant, but none of them is present in the InParanoid database, we inserted into Signalink the splice variant that has a primary UniProt accession (AC), as listed by Ensembl version 49.
- From Ensembl we included into Signalink the UniProt accession(s) of the protein and from UniProt we used the following data fields: the protein's description, the protein reference – if it contained interaction data, – and the protein's cellular component. In

addition, data from the protein's description and interaction fields were manually tested in the primary publications for further information.

- According to Ref. (Ilagan and Kopan, 2007), NOTCH1 is a core protein of its pathway and it functions as a receptor, mediator, or transcription factor. However, when in the Notch pathway, NOTCH1 functions either a receptor or a transcription factor, thus, we have included into Signalink only these two pathway sections for NOTCH1.
- To make Signalink as complete as possible, we searched for orthologs of the human NOTCH1 protein. Orthologs without known signaling interactions became predicted pathway proteins in Signalink.
- From the InParanoid database we identified the *C. elegans* and *D. melanogaster* orthologs of the human NOTCH1 protein (ENSP00000277541). (In several cases we searched by both the UniProt and Ensembl protein IDs in InParanoid to find the protein.) Interestingly, the human NOTCH1 has two orthologs in worm (LIN-12 and GLP-1), but only one ortholog in fly (the protein N). We inserted all three orthologs into Signalink. We listed the species-specific protein IDs and the UniProt ACs of the orthologs from WormBase and FlyBase.
- For ligands and transcription factors interacting with NOTCH1 we followed the same steps.

### Interactions

- We listed articles describing signaling interactions between NOTCH1 and other proteins by browsing through the references of the above mentioned review papers and by using the search engines iHOP (Fernandez et al., 2007) and ChiliBot (Chen and Sharp, 2004). The search engine iHOP allows the user to search for all abstracts with interactions containing NOTCH1. With ChiliBot the interactions between two selected proteins can be directly searched for.
- Consider, for example, the interaction between NOTCH1 and TACE/ADAM17. This interaction is described in the experimental article (Brou et al., 2000). After reading the article we found that it describes (i) a putative cleavage site for TACE on NOTCH1 and (ii) a correlation between the in vitro enzymatic activity of TACE and the activity of NOTCH1. Thus, this article provides evidence for the activation of NOTCH1 by TACE.
- In addition, we searched directly for interactions between the orthologs of TACE and NOTCH1 in the other two species.

## Comparing the approach of SignaLink with other approaches

Signaling pathways are the building blocks of signaling networks, however, their definition is not yet standardised. Thus, currently used definitions of signaling pathways are based on a variety of functional, structural, and cell specific classifications (Bader et al., 2006). As a consequence of this, more than one definition is in use for the system-level examination of signaling pathways. For example, the definition of cross-talks has not yet been standardised.

Most of these approaches have a specific purpose, *e.g.*, the characterisation of cell fate signals (Tseng and Dustin, 2002), the analysis of selected cell types (Grunwald and Klein, 2002) or selected pathways (Fuerer et al., 2008). Supplementary Table 2 lists several approaches to cross-talk analysis in the increasing order of the complexity of the examined system. The table starts with general cross-talk reviews of one selected pathway to intercellular cross-talk studies. While previous studies compared and discussed the cooperation of pathway pairs at most (Wang et al., 2009a; Nusse, 2003), SignaLink allows the user to perform both pathway-specific and system-level analyses for all eight pathways, with an emphasis on cross-talk.

Complexity ↓	Focus of the analysis	Reference
	<b>Cross-talks of</b>	(Guo and Wang, 2009)
	in healthy cell types	(Hurlbut et al., 2007)
	one pathway in stem cells	(Katoh and Katoh, 2007)
	two pathways	(Borisov et al., 2009; Wang et al., 2009a)
	selected pathways in one tissue type	(Natarajan et al., 2006)
	many pathways in one tissue type	(Sternberg, 2005)
	many pathways in stem cells	(Blank et al., 2008)
	many pathways in one organ	(Yan et al., 2004)
	<b>Intercellular cross-talk</b>	(Boswell et al., 2008)
	between two pathways in one organ	

**Supplementary Table 2.** Approaches to signaling cross-talk analysis in metazoans.

Databases usually agree on the ‘cores’ (backbones) of signaling pathways, but peripheries are often captured very differently. To circumvent the ambiguity of pathway boundaries, Gerstein and colleagues have embedded pathway ‘cores’ into the large-scale network of all known PPIs (Lu et al., 2007). Lu *et. al.* also point out that the individual signaling databases of pathways are often assembled by several professionals applying different collection criteria and thus, it can be misleading to merge such pathway collections into a single database and perform cross-talk analyzes on them. The features setting SignaLink clearly apart from other pathway databases are shown in Supplementary Table 3.



## ***Comparing the human signaling pathways of KEGG, Reactome, NetPath, and Signalink***

Manual curation projects are (i) inherently error-prone (Cusick et al., 2009) and (ii) require specific goals that define their curation criteria. Comparisons between both the curation protocols (Rzhetsky et al., 2009) and the resulting databases are essential. In the current paper we publish a database and some of its possible applications. Therefore, not the curation process, but the database itself – Signalink – should be compared to others.

The number of databases collecting signaling interactions is constantly growing (Bader et al., 2006). However, many of these databases are proprietary, list fewer than 200 molecules, or only selected types of pathway components (*e.g.*, protein kinases). Even among the few databases passing these criteria – free for academic use, more than 200 molecules, contains all pathway component types – there are currently no gold standards of signal transduction pathways compiled with similar purposes and methods as Signalink. Thus, Signalink as a whole cannot be validated at this moment.

Due to the absence of appropriate gold standards we compared the human signaling pathways of Signalink with those from 3 widely used pathway databases: KEGG, Reactome, and NetPath. For reference we compared the human signaling pathways of these three databases to each other too. Except where explicitly stated, for the comparisons we mapped all protein IDs to UniProt primary accessions (*e.g.*, P42224). For the tables and text below we converted UniProt primary accessions to HGNC symbols (*e.g.*, we mapped P42224 to STAT1).

As discussed below, there are several important differences between these databases:

- Availability of a detailed description of the curation method
- Data sources (none, research articles, or only reviews)
- Entities connected by the interactions differ: pathways, proteins, ortholog groups, protein families, *etc.*
- Usage of protein isoform names
- Protein complexes allowed or not as a source of binary interactions
- Additional attributes of the interactions (*e.g.*, directed/undirected)
- Type of protein identifier used (UniProt AC, RefSeq ID, *etc.*)
- Number of cross-talks
- Database-specific (overrepresented) molecular functions

We believe that a comprehensive analysis of these differences is essential when deciding about which database should be used for a particular application. Below we compare the curation methods and contents (proteins and interactions) of the four databases. First through pairwise comparisons between the databases, *e.g.*, KEGG-Signalink or Reactome-Signalink, and then in a summary encompassing all four databases.

	KEGG	BiOCARTA	pSTING	REACTOME	Science Signaling – STKE	NCI Nature – Pathway Interaction Database	HPRD-NetPath	Signalink
<b>Number of species</b>	Many but most networks are based on humans	Humans and mouse	Humans, rat, mouse, fly, worm, and yeast	Many	Many	Only humans	Only humans	3 ( <i>C. elegans</i> , <i>Drosophila</i> , Humans)
<b>Source of pathway components</b>	Subjective manual curation from the literature	Subjective manual curation from the literature	Subjective manual curation from the literature	Reaction based manual curation	Subjective manual curation from the literature	Subjective manual curation from the literature	Subjective manual curation from HPRD	Many pathway reviews, further manual curation
<b>Source of interactions</b>	Only reviews	Mostly reviews	PubMed ID listed for each interaction	PubMed ID listed for each interaction	PubMed ID listed for some interactions (not all)	PubMed ID listed for each interaction	PubMed ID and cited reaction details listed for each interaction	PubMed ID listed for each interaction
<b>Type of interactions</b>	Directed	Directed	Directed	Directed	Directed	Directed	Undirected	Directed
<b>Number of signaling pathways (only differences from Signalink are shown)</b>	10 (no NHR)	Many	12 (no WNT, TGF, Notch, IGF, Hh, NHR )	10 but not for every species (no Hh, NHR, JAK/STAT)	49 canonical, 32 specific (no NHR)	82 (no NHR, JAK/STAT)	20 (no WNT, IGF, NHR, JAK/STAT)	8 (EGF/MAPK, WNT, TGF, Notch, IGF, Hh, NHR, JAK/STAT)
<b>Definition of pathways</b>	Not defined	Not defined	Immune pathways	Not defined; uniform curation rules	Not defined	Not defined	Cancer or immune pathways	Biochemically based, important in development; uniform curation rules
<b>Pathways with (i) uniform curation criteria and (ii) readily available for cross-talk analysis</b>	Not possible	Not possible	Possible	Possible, but no global pathway view or common platform	Not possible	Not possible	Not possible	Possible
<b>Availability (cost, data structure, website)</b>	Free, web based, <a href="http://www.genome.jp/kegg">www.genome.jp/kegg</a>	Free, but only diagram image and protein list, <a href="http://www.biocarta.com">www.biocarta.com</a>	Free, web based, <a href="http://pstiing.licr.org">pstiing.licr.org</a>	Free, SQL dump, <a href="http://www.reactome.org">www.reactome.org</a>	Free after registration, Web based, SBML upon request, <a href="http://stke.sciencemag.org">stke.sciencemag.org</a>	Free, PID XML and BioPAX, <a href="http://pid.nci.nih.gov">pid.nci.nih.gov</a>	Free, in many network formats <a href="http://www.netpath.org">www.netpath.org</a>	Free, SQL dump, SBML <a href="http://signalink.org">signalink.org</a>

**Supplementary Table 3.** Comparison of Signalink with other manually curated signaling databases.

## KEGG and SignaLink: A comparison of isoforms and the definition of interacting entities

### Listing signaling interactions from KEGG pathway files

The Kyoto Encyclopedia of Genes and Genomes (Ogata et al., 1999) contains a large number of pathways, among them signal transduction and endocrine pathways. Out of the 8 pathways curated in SignaLink 7 are listed in KEGG: MAPK, WNT, Notch, Hedgehog, TGF-beta, Jak-STAT, and Insulin.

The pathways of KEGG contain seven types of entries: compound, group, map, ortholog, gene, enzyme, and other. From these we used only orthologs and their interactions. In KEGG each 'ortholog' entry is a group of proteins, which may contain more than one human protein. In KEGG an interaction connects two entries, e.g., two ortholog groups, and – to the best of our knowledge – an interaction between two ortholog groups does not contain information about which proteins from the two ortholog groups are the actual interactors.

For each of the above mentioned seven pathways we first downloaded the corresponding PSI-MI formatted xml pathway file from <ftp://ftp.genome.jp/pub/kegg/pathway/organisms/hsa> in April 2009. These seven KEGG pathways contain four types of entries (usage in brackets): ortholog (373), map (40), group (15), and compound (11). We discarded 'map' (a linked pathway map, which denotes an entire pathway) and 'compound' (a chemical compound including, for example, glycan) entries, and retained 'ortholog' and 'group' (a complex of gene products, mostly a protein complex) entries.

In the seven downloaded xml formatted pathway files the name of an 'ortholog' type entry contained one or more 'KEGG ortholog set (ko)'. For example, in the file 04330\_Notch.xml the type of entry number 3 is 'ortholog' and its name is a space-separated list containing three KEGG ortholog sets: "ko:K04505 ko:K04522 ko:K06060". Each of these three KEGG ortholog sets contains one human protein. We used the file hsa\_ko.list to map these KEGG ortholog set identifiers to KEGG's *Homo sapiens* protein numbers. In the example we mapped ko:K04505 to hsa:5663, ko:04522 to hsa:5664, and we could not map ko:06060 (this ID was discarded). Next, we mapped KEGG's *Homo sapiens* protein numbers to UniProt accessions with the file hsa\_uniprot.list, in the example for hsa:5663 two UniProt accessions were listed (B2R6D3 and P49768), we mapped hsa:5663 to P49768 and similarly, we mapped hsa:5664 to the second of the two listed UniProt accessions (B1AP21, P49810). Note that in both cases, the first UniProt AC is from TrEMBL, while the second is from Swiss-Prot. By checking a few examples in the file hsa\_uniprot.list we concluded that if for a KEGG *Homo sapiens* protein number more than one UniProt accession (AC) is listed and at least one of these ACs is from Swiss-Prot, then the last AC is from Swiss-Prot. Thus, to use a more reliable protein AC – one from Swiss-Prot, if possible – we used always the last AC.

The 'group' entries in the downloaded xml files of KEGG pathways contain components which are themselves entries from the same xml file. We replaced each 'group' entry with the list of its 'ortholog' components and we replaced each of these ortholog components with the list of the UniProt accessions found as explained above. In summary each pathway entry that we kept for the analysis contained one or more protein name (in UniProt AC format).

KEGG pathways list directed interactions between entries. Each of the entries we used contained one or more human protein. We used only those interactions where the two connected entries were both among the ortholog or group entries selected above. For the seven pathways this selected list of directed interactions gave 285 nodes (protein groups) and 210 directed links between these groups. The total number of proteins in these groups was 429.

### **Constructing the network of signaling interactions from KEGG pathways**

In the previous subsection we explained how we listed the directed signaling interactions connecting groups of human proteins from the KEGG pathways MAPK (we compare it to the EGF/MAPK pathway of SignaLink), WNT, Notch, Hedgehog, TGF-beta, JAK/STAT, and Insulin. We used two alternatives to construct a network of directed signaling interactions from these data. We used the ‘matrix’ and ‘spoke’ methods, which are frequently applied for constructing binary protein-protein interaction lists from protein complex data (Orchard et al., 2007).

We called the two groups of proteins connected in KEGG by a directed signaling interaction a *source group* and a *target group*. For the first KEGG signaling network (‘KEGG matrix’) we connected with a directed signaling interaction each protein of the source group with each protein of the target group. For the second KEGG signaling network (‘KEGG spoke’) we selected one protein from the source group and one protein from the target group and added a directed signaling interaction only between these two proteins. Both protein groups contained UniProt primary accessions (allowing both Swiss-Prot and TrEMBL proteins). We picked the selected protein from a group by (i) putting accessions into alphabetical order and then (ii) picking either the first Swiss-Prot accession from the list or – if only accessions from TrEMBL were listed – the last accession.

### **Comparing the human signaling pathways of KEGG and SignaLink**

To compare KEGG with SignaLink, we considered the ‘KEGG matrix’ and ‘KEGG spoke’ networks (see the previous section) constructed from the KEGG pathways MAPK, WNT, Notch, Hedgehog, TGF-beta, JAK/STAT, and Insulin. Note in Supplementary Table 4 that the number of ‘SignaLink only’ proteins is well above the number of KEGG spoke ‘KEGG only’ proteins. Only in the ‘KEGG matrix’ network – which very likely contains many false positive links – is the number of signaling interactions comparable to that of SignaLink.

For the comparisons in Supplementary Table 4 we mapped first each protein name (both in KEGG and SignaLink) to a UniProt primary accession (AC). The comparisons labelled “all isoforms names separately” were performed on these lists. Next we mapped each UniProt primary AC to the HGNC gene name listed in UniProt or – if there was no HGNC name – the truncated ID (the ID minus the “\_HUMAN” tag). Finally, in these name lists we removed from the name of each protein any of the following patterns, if they followed a number: (i) a single letter (for example, WNT10A was replaced with WNT10), (ii) a letter and a number (e.g., CSNK2A1 was replaced with CSNK2), and (iii) a letter, a number and a letter (e.g., CSNK1A1L was replaced with CSNK1). In addition to these rules MAPK protein names (*i.e.*, names starting with “MAPK”) were not changed (protein names starting with “MAP” followed by zero or more numbers, a “K” and then optional further characters were

considered to be MAPK protein names). We performed the comparisons labeled “merged isoform names” (M) on these final name lists.

Pathway	Proteins and interactions				
	M: isoform names merged, S: all isoforms listed separately				
	only in KEGG		intersection: both in KEGG 'matrix' and SignalLink	only in SignalLink	
	KEGG 'matrix'	KEGG 'spoke'			
	M (S)	M (S)	M (S)	M (S)	
Proteins	EGF/MAPK	85 (101)	57 (59)	63 (63)	98 (99)
	Hh	16 (20)	4 (5)	14 (13)	25 (26)
	IGF	65 (65)	51 (51)	9 (9)	26 (28)
	JAK/STAT	24 (24)	14 (14)	54 (55)	55 (66)
	Notch	10 (10)	8 (8)	14 (14)	25 (25)
	TGF	26 (26)	19 (19)	19 (19)	64 (67)
	WNT	52 (62)	40 (43)	32 (26)	41 (49)
	Total				
	all proteins	196 (224)	137 (145)	210 (205)	252 (278)
multi-pathway proteins	27 (34)	19 (19)	8 (7)	36 (36)	
Inter-Actions	Total	1332 (1411)	301 (307)	199 (191)	775 (799)
	Cross-talks (weighted number)	214.1 (199.8)	74.5 (71.5)	11.3 (10.3)	216.8 (214.9)

**Supplementary Table 4.** Protein and interaction numbers in the seven signaling pathways curated by both KEGG and SignalLink. The label “only in KEGG” indicates proteins (or interactions) listed only in KEGG, “intersection” indicates proteins (or interactions) listed in the given pathway(s) of both databases, and “only in SignalLink” indicates names listed only in SignalLink. In KEGG each directed signaling interaction connects two groups of proteins (a source group and a target group). In the KEGG ‘matrix’ network for each directed interaction we added a directed link between each protein of the source group and each protein of the target group. In the KEGG ‘spoke’ network only one protein from the source group was connected with one protein from the target group. See the section “Analysis of cross-talk in each species” in “Supplementary Results” above for the definition of weighted link numbers and a detailed description. Pathways in which the removal of isoform names significantly reduces the number of proteins are highlighted with gray background. See text for the discussion of this table.

Removing isoform names (*i.e.*, merging name variants) has a significant effect on the lists of proteins in several pathways. We found that KEGG contains a significant number of isoforms in both the MAPK and WNT pathways, and SignalLink contains some in the WNT pathway. In addition, both databases contain many – but not the same – isoforms in the JAK/STAT pathway. Observe also in the ‘M’ columns (where isoform names are merged) of Supplementary Table 4 that the KEGG ‘matrix’ data set contains significantly more proteins than SignalLink in the IGF pathway (74 vs. 35) and notably more proteins (84 vs. 73) in the WNT pathway, while all other pathways are more populous in SignalLink. In total SignalLink and the KEGG matrix data set have similar numbers of proteins and almost the same (weighted) cross-talk numbers. However, the KEGG matrix data set – as all networks prepared with the matrix method – is very likely to contain many false positive interactions. Thus, in our opinion SignalLink contains significantly more biologically meaningful signaling interactions, including cross-talks, than KEGG.

## Conclusions for the comparison between KEGG and Signalink

Considering the seven signaling pathways curated in both databases, we found the following differences between KEGG and Signalink:

- With the most permissive approach – KEGG ‘matrix’ with isoforms – the selected 7 pathways of KEGG have 20% fewer proteins than the same 7 pathways in Signalink (429 vs. 483 proteins). However, with the same approach KEGG has 50% more (undirected) interactions (1 502 vs. 990). This large number of links probably includes many that were introduced only by the ‘matrix’ method, and do not exist in reality.
- Although in Signalink the weighted number of cross-talks is higher, KEGG has significantly more (50%) interactions between fewer (20%) proteins. This is mainly due to the higher number of interactions within the pathways of KEGG.
- Several pathways contain large numbers of isoform names. After merging isoform names, the total number of proteins contained exclusively by either of the two databases decreases, while the total number of proteins contained by both databases increases.

## Reactome and Signalink: Complexes and database-specific molecular functions

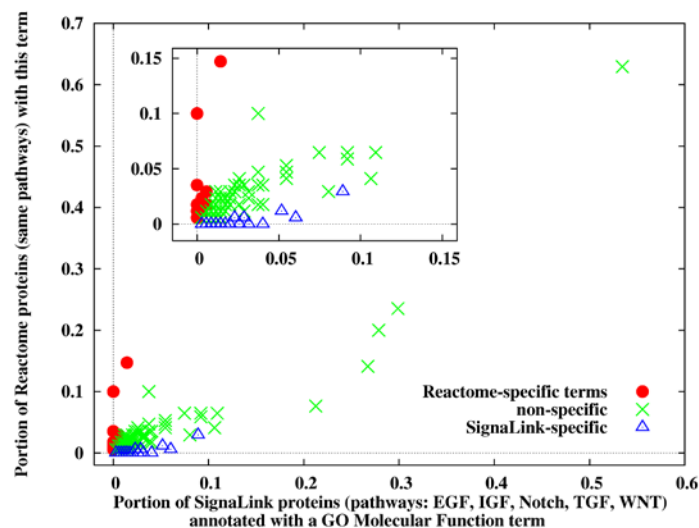
We downloaded the files `uniprot_2_pathways.stid.txt` and `homo_sapiens.interactions.txt` in April 2009 from the web directory <http://www.reactome.org/download/current> of Reactome (Joshi-Tope et al., 2005). There are five signaling pathways that have been curated both in Reactome and in Signalink: EGF (EGFR in Reactome and EGF/MAPK in Signalink), IGF/Ins, Notch, TGF, and WNT.

From the first of the two data files (`uniprot_2_pathways.stid.txt`) we extracted the protein lists of the EGFR pathway (“Signaling by EGFR”, to be compared to the EGF/MAPK pathway of Signalink), the IGF/Ins pathway (“Signaling by Insulin receptor”), the Notch pathway (“Signaling by Notch”), the TGF pathway (“Signaling by TGF beta” merged with “Signaling by BMP”), and the WNT pathway (“Signaling by Wnt”). Next we converted protein identifiers to UniProt primary accessions. From the second of the two data files (`homo_sapiens.interactions.txt`) we listed all interactions between the proteins of the five pathways. Reactome lists binary interactions classified into four interaction types: ‘neighbouring\_reaction’ (528 of them connect proteins of the five selected pathways), ‘reaction’ (535), ‘indirect\_complex’ (349), and ‘direct\_complex’ (950).

Signalink contains binary protein-protein interactions, but no binary associations between two proteins inferred from the membership of the two proteins in the same protein complex. Reactome contains both. In Reactome ‘direct\_complex’ and ‘indirect\_complex’ are the labels indicating binary associations between two proteins inferred from the membership of the two proteins in the same protein complex. For a mathematically and biologically meaningful comparison of Reactome and Signalink only true binary protein-protein interactions should be used. In Supplementary Table 5 this comparison is marked ‘N’ (Binary interactions of Reactome inferred from membership in the same protein complex are excluded). As Reactome interaction data are often used in the literature without a distinction between Reactome data types, we have also performed the comparison by including ‘complex’ associations: in Supplementary Table 5 ‘C’ indicates that in this comparison the Reactome data set includes binary protein-protein associations derived from the membership of the two proteins in the same protein complex.

	Pathway	Proteins and interactions		
		N: Reactome without complexes C: Reactome with complexes		
		only in Reactome  N (C)	both in Reactome and SignaLink  N (C)	only in SignaLink
Proteins	EGF/MAPK	28 (28)	16 (16)	146
	IGF	32 (32)	7 (7)	30
	Notch	2 (2)	14 (14)	25
	TGF	10 (11)	19 (19)	67
	WNT	5 (53)	4 (5)	70
	Total			
	all proteins	56 (105)	64 (65)	283
	multi-pathway proteins	6 (15)	0 (2)	39
Inter- actions	Total			
	all interactions	622 (1454)	67 (73)	615
	cross-talks (weighted sum)	39.3 (39.3)	0 (0)	216.5

**Supplementary Table 5.** Statistics of the five signaling pathways curated in both Reactome and Signalink. Observe that (i) protein lists significantly differ and (ii) in each pathway Signalink contains a larger number of proteins. Note that the removal of ‘complex’ interactions from Reactome reduces from 58 to 9 the number of its WNT pathway proteins that have at least one interaction with another WNT pathway protein. On the other hand the removal of ‘complex’ interactions leaves the list of cross-talks in Reactome’s selected five pathways unchanged. Thus, ‘complex’ interactions in Reactome connect pairs of proteins within pathways, mostly within the WNT pathway. Note also that the the number of proteins contained by both databases is low.



**Supplementary Figure 2.** The relative frequency of the usage of each GO Molecular Function term annotating the proteins of the five signaling pathways curated in both Reactome and Signalink. The terms called “Reactome-specific” and “Signalink-specific” are those terms that annotate at least three times as many proteins (compared to the total number of proteins) in one data set as in the other.

GO Molecular Function term	Number of proteins annotated with this term	
	in Reactome	in Signalink

**Reactome-specific terms** (at least 3 times more frequently used in the Reactome subset)

GO:0008233 peptidase activity	25	5
GO:0004298 threonine-type endopeptidase activity	17	-
GO:0016887 ATPase activity	6	-
GO:0043560 insulin receptor substrate binding	5	2
GO:0008601 protein phosphatase type 2A regulator activity	5	2

*etc.*

**Signalink-specific terms** (at least 3 times more frequently used in the Reactome subset)

GO:0008083 growth factor activity	5	31
GO:0004713 protein tyrosine kinase activity	2	18
GO:0043565 sequence-specific DNA binding	1	21
GO:0004722 protein serine/threonine phosphatase activity	1	8
GO:0004708 MAP kinase kinase activity	1	10
GO:0004707 MAP kinase activity	1	10

*etc.*

**Non-specific terms** (most frequently used non-specific terms)

GO:0005515 protein binding	107	186
GO:0000166 nucleotide binding	40	104
GO:0005524 ATP binding	34	97
GO:0016740 transferase activity	24	93
GO:0005509 calcium ion binding	17	13
GO:0004674 protein serine/threonine kinase activity	13	74

*etc.*

**Supplementary Table 6.** The usage of GO Molecular Function terms annotating the proteins of the five signaling pathways curated in both Reactome and Signalink. The total number of proteins is 170 in the Reactome subset and 348 in the Signalink subset.

In the five compared pathways both the protein and interaction lists of Reactome and Signalink significantly differ (Supplementary Table 5). To quantify the possible reasons for this difference we first listed all Gene Ontology (GO) Molecular Function terms (Harris et al., 2004) that annotate at least one protein in Reactome or Signalink. For each of these terms we counted the number of proteins it annotates in Reactome and the number of proteins it annotates in Signalink. Supplementary Figure 2 and Supplementary Table 6 show these two numbers for each GO Molecular Function term and the two sets of terms typical for Reactome and Signalink.

## Conclusions for the comparison between Reactome and Signalink

In summary, we found the following main differences between the five signaling pathways curated in both Reactome and Signalink:

- The selected five pathways of Signalink contain more than twice as many proteins (348) as the same pathways in Reactome (170). After removing interactions labelled 'direct\_complex' and 'indirect\_complex' from Reactome, this difference becomes even larger: the five selected pathways of Reactome will contain only 56 such proteins that have at least one signaling interaction with another signaling protein in these five pathways.
- Without complexes Reactome has almost the same number of interactions in the five pathways as Signalink (689 vs. 682). However, Signalink has almost three times as



many cross-talks (226 vs. 79). After adding Reactome's complex interactions ('direct\_complex' and 'indirect\_complex') the number of its within-pathway interactions increases more than twofold (to 1527), however, the number of cross-talks is unchanged. Thus, the 'complex' interactions of Reactome (*i.e.*, the binary signaling interactions inferred from the membership of two proteins in the same complex) are mostly within pathways.

- In Reactome the WNT pathway is dominated by interactions inferred from the membership of two proteins in the same complex.
- Several binding functions frequently occur in both databases. However when comparing the two databases to each other we found that proteolytic functions are more frequent in Reactome and kinases are more frequent in Signalink.

### **NetPath and Signalink: Interactions, protein identifiers, cited papers, and curation approaches**

NetPath was originally published as a part of the Human Protein Reference Database (Keshava Prasad et al., 2009) and was recently published in a separate database article (Kandasamy et al., 2010). It contains human proteins and interactions grouped into 20 pathways. In NetPath a protein is allowed to belong to more than one signaling pathway. There are 5 signaling pathways that are curated in both NetPath and Signalink: EGFR1, Hedgehog, Notch, TGF beta Receptor, and WNT. The other 15 pathways in NetPath are specific for immune cells or cancers.

For each interaction NetPath provides a list of interactors. To convert these interactions to a list of interacting pairs, we used the 'matrix' model (Bader and Hogue, 2002): each pair of human proteins appearing together in at least one interaction was listed as an interacting pair. We discarded self-interactions, *i.e.*, interactions containing a protein interacting with itself. We note here that the number of human proteins in a NetPath interaction is typically two or three, thus, for the matrix model is a reasonably precise way to map the list of interactions (protein sets) to a list of interacting pairs.

For each of the five pathways we have

1. Downloaded pathway data in a single file from [http://www.netpath.org/data/psimi\\_2](http://www.netpath.org/data/psimi_2),
2. Selected the human proteins (by the taxonomy ID: 9606) in each interaction,
3. Listed all human proteins of the pathway,
4. Listed pairs of interacting proteins with the 'matrix' model, and
5. Listed NetPath proteins with their RefSeq protein IDs, and mapped these later to UniProt primary accession (AC) identifiers.

For the comparisons shown in Supplementary Tables 7 and 8 we considered the proteins of the five selected pathways of NetPath and all interactions between these proteins. We did the same in Signalink.

In Supplementary Table 8 we compare the lists of articles cited by the two databases for the five selected pathways. NetPath lists PubMed IDs (PMIDs) in each of its pathway files, *e.g.*, at [http://www.netpath.org/data/psimi\\_2/NetPath\\_10\\_PSIMI.xml](http://www.netpath.org/data/psimi_2/NetPath_10_PSIMI.xml) for the EGFR1 pathway. From Signalink we listed for each pathway (or the five pathways together) the PMIDs of those articles that provided experimental evidence for the interactions connecting two proteins that are both contained by the selected pathway (or pathway group).

Pathway	Number of proteins or interactions listed for this pathway		
	only in NetPath	both in NetPath and SignaLink	only in SignaLink
<b>EGF/MAPK</b>			
<b>All proteins</b>	<b>115</b>	<b>19</b>	<b>143</b>
Listed in UniProt/Swiss-Prot	72	19	143
Listed in UniProt/TrEMBL	42	0	0
Not found in UniProt	1	0	0
<b>All interactions (undirected)</b>	<b>141</b>	<b>0</b>	<b>351</b>
<b>Hh</b>			
<b>All proteins</b>	<b>5</b>	<b>12</b>	<b>27</b>
Listed in UniProt/Swiss-Prot	2	12	27
Listed in UniProt/TrEMBL	2	0	0
Not found in UniProt	1	0	0
<b>All interactions (undirected)</b>	<b>10</b>	<b>4</b>	<b>53</b>
<b>Notch</b>			
<b>All proteins</b>	<b>42</b>	<b>11</b>	<b>28</b>
Listed in UniProt/Swiss-Prot	25	11	27
Listed in UniProt/TrEMBL	16	0	0
Not found in UniProt	1	0	1
<b>All interactions (undirected)</b>	<b>38</b>	<b>2</b>	<b>58</b>
<b>TGF</b>			
<b>All proteins</b>	<b>103</b>	<b>17</b>	<b>69</b>
Listed in UniProt/Swiss-Prot	63	17	68
Listed in UniProt/TrEMBL	36	0	1
Not found in UniProt	4	0	0
<b>All interactions (undirected)</b>	<b>186</b>	<b>0</b>	<b>188</b>
<b>WNT</b>			
<b>All proteins</b>	<b>61</b>	<b>20</b>	<b>55</b>
Listed in UniProt/Swiss-Prot	36	20	55
Listed in UniProt/TrEMBL	20	0	0
Not found in UniProt	5	0	0
<b>All interactions (undirected)</b>	<b>84</b>	<b>4</b>	<b>94</b>
<b>Five pathways together</b>			
<b>All proteins</b>	<b>281</b>	<b>81</b>	<b>274</b>
Listed in UniProt/Swiss-Prot	171	81	273
Listed in UniProt/TrEMBL	100	0	1
Not found in UniProt	10	0	0
Multi-pathway proteins	15	2	39
<b>All interactions (undirected)</b>	<b>446</b>	<b>11</b>	<b>690</b>
<b>Cross-talks (weighted sum)</b>	<b>58.3</b>	<b>1.5</b>	<b>169.6</b>

**Supplementary Table 7.** Statistics of the five signaling pathways curated in both NetPath and SignaLink. For the comparison all protein names – RefSeq IDs from NetPath and UniProt accessions (AC) from SignaLink – have been converted to UniProt primary accessions. For the five signaling pathways NetPath and SignaLink list roughly the same number of proteins: 362 and 355, respectively. Altogether the two databases list 636 proteins for the five pathways, but only 81 protein names are listed in both cases. This large difference is in part due to the significant number of proteins in NetPath that have either not yet been verified fully with experiments and are listed in UniProt/TrEMBL (100 such proteins) or are not listed in UniProt at all (10 proteins). The number of cross-talks is much larger in SignaLink (weighted number: 171.1 vs. 59.8) and only a few of these cross-talks are present in both databases (weighted number: 1.5). Observe that in SignaLink the amount of proteins that belong to more than one pathway is more than twice as many as in NetPath. Note also that the low number of proteins contained by both NetPath and SignaLink makes the number of interactions contained by both databases also very small.

Pathway	Number of papers cited for the interactions of this pathway		
	only in NetPath	both in NetPath and Signalink	only in Signalink
<b>EGF/MAPK</b>	132	3	206
<b>Hh</b>	9	5	23
<b>Notch</b>	47	10	24
<b>TGF</b>	95	18	74
<b>WNT</b>	53	17	61
<b>All 5 pathways</b> all papers	327	61	337

**Supplementary Table 8.** The number of articles cited for the interactions of the five signaling pathways (EGF/MAPK, Hedgehog, Notch, TGF, and WNT) that are curated in both NetPath and Signalink.

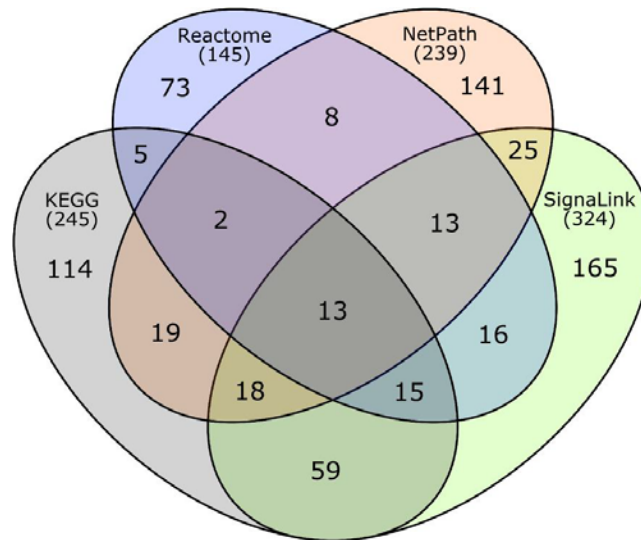
## Conclusions for the comparison between NetPath and Signalink

We conclude that the major differences between the five signaling pathways curated in both NetPath and Signalink are the following:

- Signalink contains more than two times as many multi-pathway proteins and nearly three times as many cross-talks as NetPath.
- NetPath – compared to Signalink – contains a much higher number of proteins listed only in UniProt/TrEMBL, but not in UniProt/Swiss-Prot. Note that proteins listed in UniProt/TrEMBL have not yet been fully experimentally verified as known proteins.
- The two databases (NetPath and Signalink) were compiled with very different concepts in mind. We assume that NetPath was compiled using protein properties listed in UniProt (both Swiss-Prot and TrEMBL) and interactions in HPRD. As a contrast, Signalink is review-based and its concept is to identify the functions of known proteins within their signaling pathways. This conceptual difference between NetPath and Signalink may partly explain why the number of papers used for the compilation of both NetPath and Signalink is low (see Supplementary Figure 8).
- Signalink contains directed binary signaling interactions, while NetPath lists groups of interacting signaling proteins.
- Signalink is available for three metazoan species, while NetPath is available only for humans.

## Pairwise comparisons between all four databases: Reactome, NetPath, KEGG, and Signalink

As shown in the previous sections the four compared databases list significantly different numbers of proteins for several signaling pathways. The following signaling pathways have been curated in all four databases: EGF/MAPK (MAPK, EGFR1), Notch, TGF, and WNT. In Supplementary Figure 3 we compare all four databases on one diagram using these four pathways.



**Supplementary Figure 3.** Venn diagram comparing the four signaling pathways (EGF/MAPK, Notch, TGF, and WNT) curated in all four databases: Reactome, NetPath, KEGG, and SignalLink. Each section of the diagram shows the number of proteins in that set. The total number of proteins in the four pathways of each database is shown below the name of the corresponding database. Only proteins listed in UniProt/Swiss-Prot are included (proteins listed in UniProt/TrEMBL have been excluded from this comparison). Diagram layout from [http://en.wikipedia.org/wiki/Venn\\_diagram](http://en.wikipedia.org/wiki/Venn_diagram).

### Comparing the lists of publications referenced in signaling pathway databases

Database name	Number of all pathways (signaling and other)	Number of human pathways that have an equivalent in SignalLink	Number of equivalent human pathways in SignalLink
BioCarta	351 <sup>*</sup>	34	8
KEGG	270 <sup>**</sup>	7	7
NCI	107 <sup>*</sup>	20	5
NetPath	20 <sup>*</sup>	5	5
pSTIING	38 <sup>*</sup>	10	3
Reactome	997 <sup>*</sup>	5	5
SignalLink	8 <sup>**</sup>	8	8
STKE	84 <sup>**</sup>	35	8

**Supplementary Table 9a.** Number of human pathways in cellular signaling databases. <sup>\*</sup>Only human pathways.

<sup>\*\*</sup>A pathway available in more than one organism is counted as one pathway.

Database name	Total number of cited papers (signaling and other pathways)	Number of cited papers for the signaling pathways available in Signalink	
		Number of cited papers	Average number per pathway
BioCarta*	1174	86	2
KEGG	2388	73	10.4
NCI*	1287	200	10
NetPath*	1109	351	70.2
pSTIING <sup>†,*</sup>	n.a.	n.a.	n.a.
Reactome	4491	166	33.2
Signalink <sup>‡</sup>	941	941	42.8
STKE <sup>†</sup>	n.a.	n.a.	n.a.

**Supplementary Table 9b.** Statistics of the publications cited in cellular signaling databases. Included are those publications that have been used for manual curation in the given pathway database and are available on its website either as a downloadable file or for viewing with a browser. Note that while Supplementary Table 8 counts the publications cited for the interactions of a given pathway in Signalink, in this table the numbers in the last column indicate the publications cited for pathway membership in Signalink.\*Only human pathways. <sup>†</sup>No online source of the list of publications used for curation could be identified (n.a.: not available). <sup>‡</sup>To curate pathway membership and interactions Signalink uses a total of 941 review and research articles while for curating pathway membership only review papers have been used.

In Supplementary Table 9b we compare the numbers of publications cited by BioCarta, KEGG, NCI (NCI-Nature Pathway Interaction Database), NetPath, pSTIING, Reactome, and STKE (Signal Transduction Knowledge Environment, Database of Cell Signaling) with the list of publications cited by Signalink. We found that in Signalink the number of PubMed citations used per pathway is above the average of the listed databases.

### Listing cited papers

For **BioCarta** we started from the URL <http://www.biocarta.com/genes/allPathways.asp> and downloaded the papers cited under “References” on the page of each pathway. From **NetPath** we downloaded and unzipped the file <http://www.netpath.org/download/zipped/PSI-MI.zip> and then listed pathways from this file. For **KEGG** we listed cited papers from the file <ftp://ftp.genome.jp/pub/kegg/pathway/pathway>. For the **NCI** database we started from the file [http://pid.nci.nih.gov/browse\\_pathways.shtml](http://pid.nci.nih.gov/browse_pathways.shtml) and downloaded the references of all NCI-curated pathways (subnetworks excluded). For **pSTIING** we started from the URL [http://pstiing.licr.org/search/a\\_start\\_pathway.jsp](http://pstiing.licr.org/search/a_start_pathway.jsp), but could not locate the PubMed identifiers of the articles used for manual curation. For **Reactome**, we first downloaded the MySQL dump of the full database from <http://www.reactome.org/download/index.html>, but could not find the PubMed IDs in it. However, it was possible to download the list of PubMed IDs from the BioMart server at <http://www.reactome.org/cgi-bin/mart>. In the case of **Science STKE** (Signal Transduction Knowledge Environment) we started at the following URL: <http://stke.sciencemag.org/cgi/collection>, but could not find the list of PubMed IDs used for manual curation of the pathways.

## Final conclusions from the database comparisons

It is widely assumed that so far in each investigated organism only a small portion of all protein-protein interactions – this includes signaling interactions – have been identified and most interactions remain to be mapped (Venkatesan et al, 2009). To minimise the number of false positives entered into Signalink we have strictly adhered to the rules described above in the section titled “Manual curation process of Signalink” during the curation process. Since the number of missing interactions is large in each signaling database and Signalink was optimised towards minimizing the number of false positive interactions, we think that the high number of pathway proteins and the high number of interactions is one of the advantages of Signalink.

Signalink was compiled based on reviews and primary research articles, therefore, the high numbers of proteins, cross-talks, and multi-pathway proteins indicate a high coverage (true positives). It is increasingly recognized that signaling pathways are not isolated entities, but form a single pathway system, thus, cross-talks and multi-pathway proteins are especially informative, and a high number of these entities (true positives) is also an advantage. As of yet only rough estimates can be made for the number of signaling interactions in the three investigated organisms, but the detailed description of the curation process of Signalink provides strong support for its higher coverage.

We note that Cusick et. al. (Cusick et al, 2009) have pointed out several drawbacks of manually curating “low-throughput” (LTP) data compared to collecting data with high-throughput (HTP) methods. For the specific case of intracellular signaling we add the following comments to the results of Cusick et. al. In intracellular signaling (i) an exceptionally large portion of interactions is localised to, *e.g.*, membranes, and cannot be efficiently detected with current HTP techniques, and (ii) the manual curation process of Signalink eliminates some of the common disadvantageous features of manually curated databases (*e.g.*, absence of proper documentation for the interactions) that have been highlighted by Cusick *et.al.* (2009). Most importantly, the manual curation process of Signalink is based on a combination of experimental (research) and review articles. In the first set of sources (original research articles) experimental evidence is published, while in the second set of sources (review articles) specialists analyze available evidence about the relevance and quality of experimental data.

Based on the comparisons between signaling databases, we found that compared to Reactome, NetPath, and KEGG the advantages of Signalink are the following:

- Precisely defined and documented curation protocol
- Highest numbers of proteins and interactions for the curated signaling pathways
- Highest numbers of cross-talks and multi-pathway proteins
- Largest overlap with the other three databases
- Number of publications per pathway higher than average
- Minimal usage of isoforms
- No binary interactions inferred from membership in the same protein complex
- Low number of proteins for which there is not yet enough experimental evidence (these are proteins listed in UniProt/TrEMBL, not in UniProt/Swiss-Prot).

## *Protein expression in healthy tissue types*

### **Selecting the expression data set: eGenetics**

Two of the currently available major, up-to-date sources of human tissue- (and organ-) specific protein expression data are the GNF Atlas (Su et al., 2004) and the eGenetics data set (Kelso et al., 2003). Both are integrated into Ensembl. The GNF Atlas was produced in a single series of experiments and it contains protein expression for 46 human tissues, organs, and cell lines. The eGenetics data set is a collection of small-scale experiments (7016 cDNA libraries represented in dbEST and 104 SAGE libraries) with protein expression data in a total of 50 different human tissue and organ types. In their top 19 tissue types the GNF Atlas and eGenetics contain protein expression for an average of 251.6 and 242.4 Signalink proteins, respectively. In other words, the two expression data sets provide approximately the same coverage of Signalink. We decided to use eGenetics because its data were more easily accessible.

The eGenetics data set lists for each protein the tissue/organ types where that protein is expressed. This data set contains a total of 524 proteins from Signalink. Out of these 323 are expressed in at least one tissue type (in the eGenetics ontology “anatomical system” means all tissue types). In Supplementary Table 10 we list the tissue and organ types where – according to eGenetics – the highest numbers of Signalink proteins are expressed. This is the full list of tissue/organ types used in eGenetics (in alphabetical order): adrenal gland, alimentary, artery, auditory apparatus, basal nuclei, brain, brain stem, breast, cardiovascular, central nervous system, cerebellum, cerebral cortex, cerebrum, colorectal, dermal, diencephalon, endocrine, endocrine pancreas, epithalamus, female genitals, ganglion, haematological, heart, intestine, joint, liver and biliary system, lung, lymph node, lymphoreticular, male genitals, meninges, mesoderm, midbrain, muscle, musculoskeletal, omentum, oral cavity, osseous labyrinth, penis, peripheral nervous system, pharynx, respiratory, retina, skin, small intestine, subthalamus, urinary, urogenital, uterus, visual apparatus.

eGenetics ontology	Number of expressed Signalink proteins	Percent of human Signalink proteins with pathway information present in eGenetics
anatomical system (all tissue/organ types)	323	61.8%
respiratory	303	57.9%
male genitals	294	56.2%
female genitals	293	56.0%
central nervous system	288	55.1%
Alimentary	264	50.5%
<b>Colorectal</b>	<b>261</b>	<b>49.9%</b>
Endocrine	249	47.6%
Urinary	247	47.2%
<b>Dermal</b>	<b>246</b>	<b>47.0%</b>
visual apparatus	238	45.5%
Musculoskeletal	237	45.3%
<b>liver and biliary system</b>	<b>236</b>	<b>45.1%</b>
Lymphoreticular	235	44.9%
<b>Cardiovascular</b>	<b>230</b>	<b>44.0%</b>
Uterus	220	42.1%
Haematological	212	40.5%
peripheral nervous system	190	36.3%
<b>Muscle</b>	<b>188</b>	<b>36.0%</b>
Breast	174	33.2%

**Supplementary Table 10.** The number of expressed signaling proteins (from Signalink) in each human tissue/organ type according to eGenetics. The five tissue types selected for the analyzes are marked with dark background color.

## Selecting the five tissue types

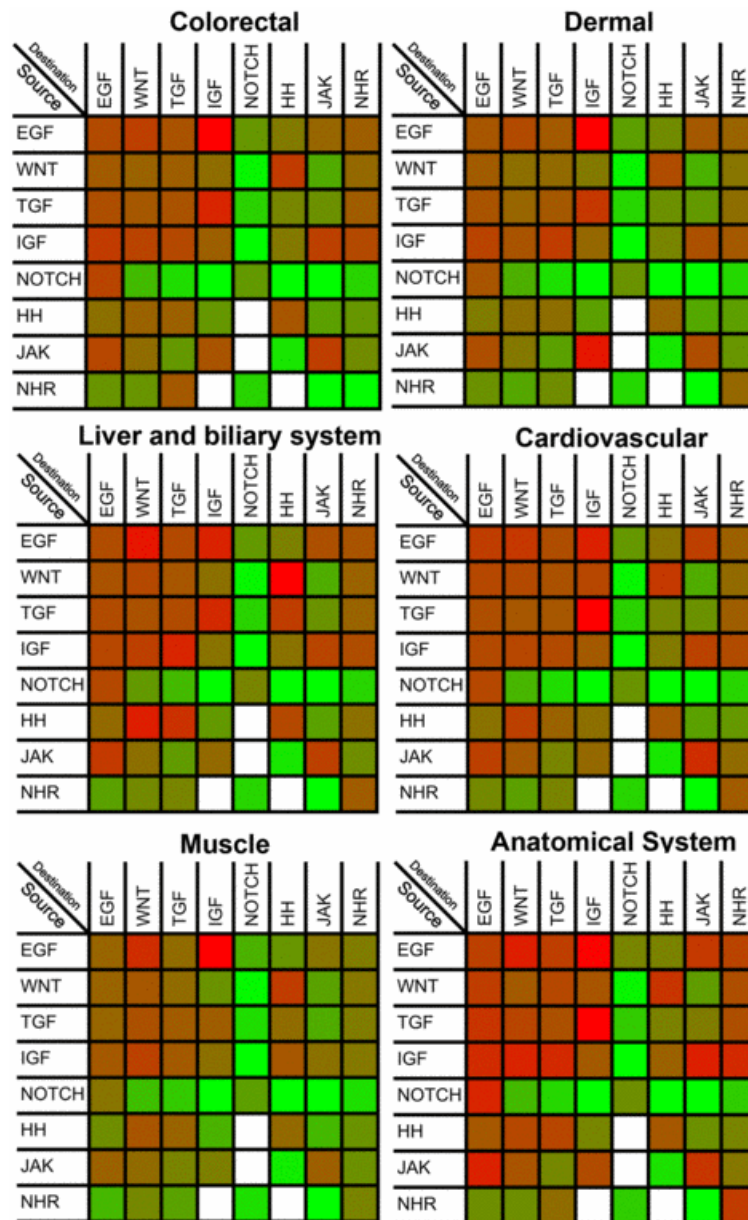
Our goal was to check whether human cross-talks are selectively active, *i.e.*, whether cross-talks connecting different pathway pairs are active in different numbers of tissue types. To obtain a statistically reliable statement, we decided to analyze several tissue types together. On the other hand, using expression from all tissue types together would blur the cross-talk differences we would like to detect. As an optimum we selected five tissue types. We listed from the full eGenetics expression data set only protein expression in these five tissues and merged this reduced expression data set with SignaLink. The criteria for selecting the five tissue types were the following: (i) eGenetics should contain expression data for a high number of SignaLink proteins in these five tissues, *i.e.*, in eGenetics these five tissues should be as specific for SignaLink proteins as possible; (ii) each selected tissue type should be biologically homogeneous, and (iii) they should be of different origin (ecto-, meso-, entodermal) and function (*e.g.*, skin, liver).

To satisfy (i) we first listed the tissue types where the highest numbers of SignaLink proteins were expressed. Observe in Supplementary Table 10 that the first few tissue types on the list are rather heterogeneous: each of them is distributed among several organs, *e.g.*, male and female genitals or the alimentary and endocrine systems. To satisfy conditions (ii) and (iii), we looked on this list for those five tissue types that were the most homogeneous and of sufficiently different origin. Thus, from the top of the list we selected cardiovascular, colorectal, and dermal tissues, the liver/biliary system, and muscle. Supplementary Table 11 shows for each curated signaling pathway the portion of its proteins expressed in the five selected tissue types.

Pathway	Average expression of the pathway's proteins in the five selected tissue types
Notch	48.6%
IGF	45.9%
TGF	44.2%
EGF/MAPK	42.3%
NHR	38.8%
Hh	34.2%
WNT	33.3%
JAK/STAT	25.8%

**Supplementary Table 11.** Average percentage of the proteins of each human signaling pathway expressed in the five selected tissues types. For a given pathway we first computed the percent of its proteins expressed in each of the five selected tissue types and then we averaged this value over the five tissues.





**Supplementary Figure 4.** Control for Figure 3a of the main text. See text for details.

## Controls

For Figure 3a of the main text we applied two types of controls shown in Supplementary Figure 4. First, we performed separately for each of the five selected tissue types the same analysis as in Fig. 3a. Second, we selected the category “anatomical system” (*i.e.*, all tissue types pooled). The main characteristics – discussed in the main text and in the supplementary material above – are preserved.

## Statistical significance analyses of pathway member protein functions

### Gene Ontology term enrichment tests in the curated signaling pathways

For a rigorous statistical analysis of the functions of proteins in the curated pathways and their overlaps we performed statistical enrichment tests of Gene Ontology Biological Process terms using the hypergeometric distribution and a (Bonferroni) multiple hypothesis correction. We performed this test in each of the signaling pathways and their overlaps with the 'GO Termfinder' toolbox, implemented by (Boyle et al., 2004). Note that GO TermFinder uses a list of GO terms not only as a plain list, but it makes of the Directed Acyclic Graph (DAG) of GO terms to quantify the significance of term enrichment in a group of proteins.

For each of the 22 signaling pathways available in SignaLink Supplementary Table 12 lists the 5 (for the EGF/MAPK pathway 6 or 7) most significant GO Biological Process terms. This analysis has pointed out that each pathway does have its own specific terms, many of which are explicitly are the GO terms named after the pathway, *e.g.*, MAPKKK cascade. In each of the 3 species the listed (most significant) GO terms of different pathways are different terms, in other words: the listed pathways are functionally distinct entities.

### *Caenorhabditis elegans*

Pathway in SignaLink	Most significant GO Biological Process terms of the proteins of the pathway (with corrected P-values using a Bonferroni correction)		
EGF/MAPK	GO:0040025	vulval development	$7.3 * 10^{-31}$
	GO:0007165	signal transduction	$2.1 * 10^{-28}$
	GO:0007154	cell communication	$4.5 * 10^{-28}$
	GO:0040028	regulation of vulval development	$1.2 * 10^{-23}$
	GO:0048580	regulation of post-embryonic development	$1.2 * 10^{-23}$
	...		
	GO:0000165	MAPKKK cascade	$6.7 * 10^{-16}$
	...		
IGF	GO:0040024	dauer larval development	$4.7 * 10^{-16}$
	GO:0042221	response to chemical stimulus	$2.6 * 10^{-14}$
	GO:0008286	insulin receptor signaling pathway	$9.6 * 10^{-14}$
	GO:0032868	response to insulin stimulus	$9.6 * 10^{-14}$
	GO:0032869	cellular response to insulin stimulus	$9.6 * 10^{-14}$
	...		
NHR	GO:0006350	transcription	0 (for all five GO terms)  The SignaLink NHR pathway contains all 246 <i>C. elegans</i> proteins in these 5 GO groups.
	GO:0006351	transcription, DNA-dependent	
	GO:0006355	regulation of transcription, DNA-dependent	
	GO:0009889	regulation of biosynthetic process	
	GO:0010468	regulation of gene expression	
	...		
NOTCH	GO:0007219	Notch signaling pathway	$1.1 * 10^{-19}$
	GO:0008593	regulation of Notch signaling pathway	$2.4 * 10^{-12}$
	GO:0001708	cell fate specification	$2.2 * 10^{-11}$
	GO:0045165	cell fate commitment	$1.6 * 10^{-10}$
	GO:0045747	positive regulation of Notch signaling pathway	$2.0 * 10^{-8}$
	...		
TGF	GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	$6.4 * 10^{-15}$
	GO:0007179	transforming growth factor beta receptor signaling pathway	$3.3 * 10^{-12}$
	GO:0007167	enzyme linked receptor protein signaling pathway	$7.6 * 10^{-12}$

	GO:0040024	dauer larval development	$2.0 * 10^{-11}$
	GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	$1.9 * 10^{-10}$
	...		
WNT	GO:0045165	cell fate commitment	$3.2 * 10^{-37}$
	GO:0030154	cell differentiation	$1.2 * 10^{-31}$
	GO:0048869	cellular developmental process	$7.7 * 10^{-31}$
	GO:0001708	cell fate specification	$2.5 * 10^{-28}$
	GO:0050794	regulation of cellular process	$5.8 * 10^{-25}$
	...		

**Supplementary Table 12a.** The statistically most significant GO Biological Process terms for each signaling pathway of *C. elegans* curated in SignaLink. See the text above for details.

### *Drosophila melanogaster*

Pathway in SignaLink	Most significant GO Biological Process terms of the proteins of the pathway (with corrected P-values using a Bonferroni correction)		
EGF/MAPK	GO:0007165	signal transduction	$3.9 * 10^{-39}$
	GO:0007154	cell communication	$1.1 * 10^{-35}$
	GO:0050794	regulation of cellular process	$8.7 * 10^{-34}$
	GO:0050789	regulation of biological process	$9.2 * 10^{-33}$
	GO:0007167	enzyme linked receptor protein signaling pathway	$5.1 * 10^{-32}$
	...		
	GO:0000165	MAPKKK cascade	$4.9 * 10^{-24}$
HH	GO:0007173	epidermal growth factor receptor signaling pathway	$6.5 * 10^{-24}$
	...		
	GO:0007224	smoothened signaling pathway	$1.2 * 10^{-47}$
	GO:0008589	regulation of smoothened signaling pathway	$1.2 * 10^{-29}$
	GO:0007166	cell surface receptor linked signal transduction	$5.6 * 10^{-27}$
	GO:0007165	signal transduction	$9.0 * 10^{-23}$
IGF	GO:0007154	cell communication	$1.1 * 10^{-20}$
	...		
	GO:0008286	insulin receptor signaling pathway	$4.5 * 10^{-36}$
	GO:0032868	response to insulin stimulus	$4.5 * 10^{-36}$
	GO:0032869	cellular response to insulin stimulus	$4.5 * 10^{-36}$
	GO:0043434	response to peptide hormone stimulus	$4.5 * 10^{-36}$
JAK-STAT	GO:0032870	cellular response to hormone stimulus	$8.3 * 10^{-34}$
	...		
	GO:0007259	JAK-STAT cascade	$8.1 * 10^{-26}$
	GO:0007243	protein kinase cascade	$2.2 * 10^{-18}$
	GO:0046425	regulation of JAK-STAT cascade	$1.2 * 10^{-16}$
	GO:0007242	intracellular signaling cascade	$2.4 * 10^{-12}$
NHR	GO:0010627	regulation of protein kinase cascade	$3.3 * 10^{-9}$
	...		
	GO:0006355	regulation of transcription, DNA-dependent	$2.9 * 10^{-20}$
	GO:0051252	regulation of RNA metabolic process	$1.5 * 10^{-19}$
	GO:0006351	transcription, DNA-dependent	$2.2 * 10^{-19}$
	GO:0032774	RNA biosynthetic process	$2.3 * 10^{-19}$
NOTCH	GO:0045449	regulation of transcription	$6.9 * 10^{-19}$
	...		
	GO:0007219	Notch signaling pathway	$2.5 * 10^{-43}$
	GO:0007166	cell surface receptor linked signal transduction	$1.8 * 10^{-26}$
	GO:0007154	cell communication	$3.6 * 10^{-25}$
	GO:0007165	signal transduction	$2.2 * 10^{-24}$

TGF	GO:0008593	regulation of Notch signaling pathway	$2.2 * 10^{-19}$
	...		
	GO:0007179	transforming growth factor beta receptor signaling pathway	$1.5 * 10^{-30}$
	GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	$2.6 * 10^{-28}$
	GO:0007167	enzyme linked receptor protein signaling pathway	$1.1 * 10^{-19}$
	GO:0007165	signal transduction	$2.4 * 10^{-18}$
	GO:0007166	cell surface receptor linked signal transduction	$2.8 * 10^{-17}$
WNT	...		
	GO:0016055	Wnt receptor signaling pathway	$8.8 * 10^{-54}$
	GO:0007166	cell surface receptor linked signal transduction	$1.7 * 10^{-34}$
	GO:0007165	signal transduction	$4.6 * 10^{-27}$
	GO:0030111	regulation of Wnt receptor signaling pathway	$5.6 * 10^{-25}$
	GO:0007154	cell communication	$1.2 * 10^{-24}$
	...		

**Supplementary Table 12b.** The statistically most significant GO Biological Process terms for each signaling pathway of *D. melanogaster* curated in SignaLink. See the text above for details.

### *Homo sapiens*

Pathway in SignaLink	Most significant GO Biological Process terms of the proteins of the pathway (with corrected P-values using a Bonferroni correction)		
EGF/MAPK	GO:0007243	protein kinase cascade	$5.8 * 10^{-88}$
	GO:0006793	phosphorus metabolic process	$1.8 * 10^{-80}$
	GO:0006796	phosphate metabolic process	$1.8 * 10^{-87}$
	GO:0007242	intracellular signaling cascade	$1.0 * 10^{-85}$
	GO:0016310	phosphorylation	$1.8 * 10^{-83}$
	...		
	GO:0000165	MAPKKK cascade	$4.8 * 10^{-76}$
HH	...		
	GO:0007224	smoothened signaling pathway	$1.4 * 10^{-15}$
	GO:0009953	dorsal/ventral pattern formation	$2.1 * 10^{-13}$
	GO:0007275	multicellular organismal development	$4.9 * 10^{-13}$
	GO:0032502	developmental process	$1.1 * 10^{-11}$
	GO:0048856	anatomical structure development	$5.5 * 10^{-11}$
IGF	...		
	GO:0032870	cellular response to hormone stimulus	$1.7 * 10^{-26}$
	GO:0008286	insulin receptor signaling pathway	$3.0 * 10^{-25}$
	GO:0032869	cellular response to insulin stimulus	$1.2 * 10^{-24}$
	GO:0032868	response to insulin stimulus	$1.6 * 10^{-24}$
	GO:0043434	response to peptide hormone stimulus	$8.2 * 10^{-24}$
JAK-STAT	...		
	GO:0007259	JAK-STAT cascade	$1.3 * 10^{-60}$
	GO:0018108	peptidyl-tyrosine phosphorylation	$5.7 * 10^{-48}$
	GO:0018212	peptidyl-tyrosine modification	$1.2 * 10^{-47}$
	GO:0007243	protein kinase cascade	$1.2 * 10^{-44}$
	GO:0007260	tyrosine phosphorylation of STAT protein	$2.0 * 10^{-41}$
NHR	...		
	GO:0006355	regulation of transcription, DNA-dependent	$5.3 * 10^{-49}$
	GO:0051252	regulation of RNA metabolic process	$1.7 * 10^{-48}$
	GO:0006351	transcription, DNA-dependent	$1.6 * 10^{-47}$
	GO:0032774	RNA biosynthetic process	$1.7 * 10^{-47}$
	GO:0016070	RNA metabolic process	$2.1 * 10^{-41}$
NOTCH	...		
	GO:0007219	Notch signaling pathway	$3.2 * 10^{-49}$

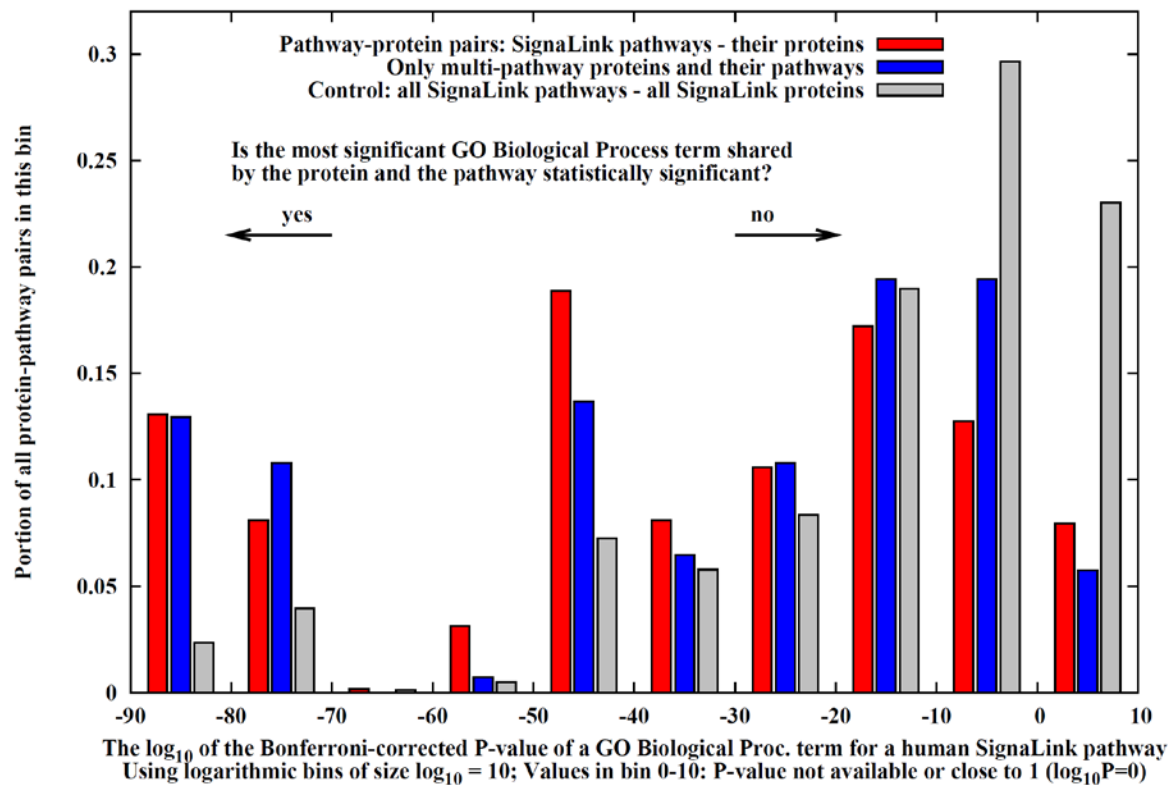
	GO:0007166	cell surface receptor linked signal transduction	$7.5 * 10^{-15}$
	GO:0032502	developmental process	$4.1 * 10^{-12}$
	GO:0031293	membrane protein intracellular domain proteolysis	$4.8 * 10^{-12}$
	GO:0007389	pattern specification process	$5.9 * 10^{-12}$
	...		
TGF	GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	$3.4 * 10^{-54}$
	GO:0007179	transforming growth factor beta receptor signaling pathway	$4.1 * 10^{-41}$
	GO:0007167	enzyme linked receptor protein signaling pathway	$3.2 * 10^{-39}$
	GO:0043687	post-translational protein modification	$1.2 * 10^{-35}$
	GO:0006464	protein modification process	$5.5 * 10^{-33}$
	...		
WNT	GO:0016055	Wnt receptor signaling pathway	$1.8 * 10^{-88}$
	GO:0007166	cell surface receptor linked signal transduction	$6.8 * 10^{-45}$
	GO:0007165	signal transduction	$3.7 * 10^{-33}$
	GO:0007154	cell communication	$1.0 * 10^{-30}$
	GO:0050794	regulation of cellular process	$3.0 * 10^{-25}$
	...		

**Supplementary Table 12c.** The statistically most significant GO Biological Process terms for each human signaling pathway of curated in SignaLink. See the text above for details.

### Gene Ontology term similarity of human multi-pathway proteins and their pathways

As explained above, we have computed with GO TermFinder the corrected P-values for the list of significantly enriched Gene Ontology terms in each signaling pathway. To estimate the significance with which a protein belongs to a pathway a natural first step is to find GO terms that annotate both the protein and the pathway. Here we quantify the significance with which an arbitrary protein belongs to an arbitrary signaling pathway as the smallest (most significant) P-value of a GO term of the pathway that annotates the selected protein too.

In Supplementary Figure 5 we plot the results of three comparisons. First, we compute the above defined similarity for each human protein – SignaLink pathway pair, and plot the distribution of similarity scores. This is the control set shown in gray color in the figure. Next, we restrict the same comparison to protein-pathway pairs where the protein is a member of that pathway in SignaLink (red in the figure). Third, we restrict this comparison even further and use only such protein-pathway pairs, where the protein is a member of that pathway and it is a member of at least one other pathway, too, *i.e.*, it is a multi-pathway protein (see blue boxes in the figure). We found that the similarity between multi-pathway proteins and their pathways has a peak at low P-values (high significance). The same holds for the comparisons between all pathway member protein and their pathways. On the other hand, in the control case (all proteins compared to all pathways) the peak of the distribution of similarities is at high P-values (low significance).



**Supplementary Figure 5.** Distribution of the similarity scores between the functions (GO Biological Process) of a protein and the functions of a SignaLink pathway. For each of the 8 human signaling pathways curated in SignaLink, GO TermFinder was used to compute the P-values of GO Biological Process terms. Given a set of proteins, their GO terms (in one of three aspects: Biological Process, Cellular component, or Molecular Function) and the DAG (Directed Acyclic Graph) formatted hierarchy of these GO terms, GO TermFinder computes the P-value for the group of proteins as a whole. A summary of this analysis is shown in Supplementary Table 12. See also the text above for details. Here we found that those pathway member proteins that belong to more than one pathway are significantly more similar to their pathways than the proteins in the control case (all proteins compared to all pathways). However, they are less similar to their pathways than those proteins that belong to one pathway only, most likely because they have multiple functions. See text for further details.

### ***Validation of the human interactions of SignaLink with an external resource***

We have evaluated the human interactions listed in SignaLink with the PRINCESS web service (*Protein Interaction Confidence Evaluation System with Multiple Data Sources*) (Li et al., 2008). Although PRINCESS can evaluate only human interactions, the majority of the interactions in SignaLink are human (991 from the total 1461), therefore, this analysis can provide representative information about the percent of true positive interactions in SignaLink. PRINCESS uses 7 different sources of interactions: (1) PPI databases; (2) Known interolog interactions; (3) Possibility for domain interaction; (4) GO\_co-annotation; (5) Genome\_context; (6) Gene\_co-expression; (7) Network\_topology information.

From the 991 human protein-protein interactions listed in SignaLink, the PRINCESS web service recognized 982 interactions and from these 890 interactions received a likelihood score above the cutoff threshold for reliable interactions. Thus, according to this validation test the ratio of high confidence interactions among the human signaling interactions listed in SignaLink is 90.6%. (For detailed statistics see Supplementary Table 13.) The interactions for *C. elegans* and *D. melanogaster* were compiled with the same curation rules as for human, therefore, we expect similar ratios of true positive interactions in these two species too.

	Number of interactions
Interactions tested* (Signalink human)	982
High-confidence interactions**	890
<b>Source of evidence for high-confidence interactions</b>	
PPI_Database	201
Interolog	13
Interacting_Domain	810
GO_Co-annotation	754
Genome_Context	25
Gene_Co-expression	117
Network_Topology	451

**Supplementary Table 13.** Statistics of the evaluation of human interactions listed in Signalink with the PRINCESS web service. \*From the 991 human interactions listed in Signalink, PRINCESS recognized 982 interactions. \*\*Interactions with a likelihood ratio above 2.

### ***Database updates***

The next update of the Signalink database is scheduled for the period January, 2011 – June, 2011.

### **Data sources and methods:**

- PubMed search for review papers of the curated signaling pathways.
- Semi-automatic filtering of the PubMed abstracts returned after a search for the current member proteins of Signalink pathways.
- Publications listed (by PubMed ID) in the updates of major signaling databases.
- Search for recent high-confident high-throughput databases on protein-protein interactions of signaling proteins.

### **Participants:**

- T. Korcsmaros, I. J. Farkas, T. Vellai, and P. Csermely: improving curation guidelines.
- T. Korcsmaros and D. Fazekas: curation.
- I. J. Farkas and D. Fazekas: programming.

## Reference List

- Agresti,A. and Coull,B.A. (1998) Approximate is better than "Exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126.
- Amberger,J. et al. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793-D796.
- Amit,I., Wides,R. and Yarden,Y. (2007) Evolvable signaling networks of receptor tyrosine kinases: relevance of robustness to malignancy and to cancer therapy. *Mol. Syst. Biol.*, **3**, 151.
- Bader,G.D., Cary,M.P. and Sander,C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504-D506.
- Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991-997.
- Baker,N.E. (2007) Patterning signals and proliferation in *Drosophila* imaginal discs. *Curr. Opin. Genet. Dev.*, **17**, 287-293.
- Baron,M. (2003) An overview of the Notch signalling pathway. *Semin. Cell Dev. Biol.*, **14**, 113-119.
- Becker,K.G. et al. (2004) The genetic association database. *Nat. Genet.*, **36**, 431-432.
- Berglund,A.C. et al. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36**, D263-D266.
- Berriz,G.F. and Roth,F.P. (2008) The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics.*, **24**, 2272-2273.
- Blank,U., Karlsson,G. and Karlsson,S. (2008) Signaling pathways governing stem-cell fate. *Blood*, **111**, 492-503.
- Bolos,V., Grego-Bessa,J. and de la Pompa,J.L. (2007) Notch signaling in development and cancer. *Endocr. Rev.*, **28**, 339-363.
- Borisov,N. et al. (2009) Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. *Mol. Syst. Biol.*, **5**, 256.
- Boswell,B.A., Lein,P.J. and Musil,L.S. (2008) Cross-talk between fibroblast growth factor and bone morphogenetic proteins regulates gap junction-mediated intercellular communication in lens cells. *Mol. Biol. Cell*, **19**, 2631-2641.
- Boutet,E. et al. (2007) UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. *Methods Mol. Biol.*, **406**, 89-112.
- Boyle,E.I. et al. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics.*, **20**, 3710-3715.



- Bray,S.J. (2006) Notch signalling: a simple pathway becomes complex. *Nat. Rev. Mol. Cell Biol.*, 7, 678-689.
- Brou,C. et al. (2000) A novel proteolytic cleavage involved in Notch signaling: the role of the disintegrin-metalloprotease TACE. *Mol. Cell*, 5, 207-216.
- Burglin,T.R. and Kuwabara,P.E. (2006) Homologs of the Hh signalling network in *C. elegans*. *WormBook.*, 1-14.
- Calvisi,D.F. et al. (2006) Ubiquitous activation of Ras and Jak/Stat pathways in human HCC. *Gastroenterology*, 130, 1117-1128.
- Chen,H. and Sharp,B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC. Bioinformatics.*, 5, 147.
- Chen,X. et al. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, 13, 1929-1939.
- Claeys,I. et al. (2002) Insulin-related peptides and their conserved signal transduction pathway. *Peptides*, 23, 807-816.
- Cote,R.G. et al. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC. Bioinformatics.*, 8, 401.
- Cusick,M.E. et al. (2009) Literature-curated protein interaction datasets. *Nat. Methods*, 6, 39-46.
- De Mesquita,D.D. et al. (2001) p90-RSK and Akt may promote rapid phosphorylation/inactivation of glycogen synthase kinase 3 in chemoattractant-stimulated neutrophils. *FEBS Lett.*, 502, 84-88.
- Dennis,G., Jr. et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4, 3.
- Desbois-Mouthon,C. et al. (2006) Impact of IGF-1R/EGFR cross-talks on hepatoma cell sensitivity to gefitinib. *Int. J. Cancer*, 119, 2557-2566.
- Drysdale,R. (2008) FlyBase : a database for the *Drosophila* research community. *Methods Mol. Biol.*, 420, 45-59.
- Fernandez,J.M., Hoffmann,R. and Valencia,A. (2007) iHOP web services. *Nucleic Acids Res.*
- Flanders,K.C. (2004) Smad3 as a mediator of the fibrotic response. *Int. J. Exp. Pathol.*, 85, 47-64.
- Flicek,P. et al. (2008) Ensembl 2008. *Nucleic Acids Res.*, 36, D707-D714.
- Frame,S. and Cohen,P. (2001) GSK3 takes centre stage more than 20 years after its discovery. *Biochem. J.*, 359, 1-16.

- Fuerer,C., Nusse,R. and Ten Berge,D. (2008) Wnt signalling in development and disease. Max Delbruck Center for Molecular Medicine meeting on Wnt signaling in Development and Disease. *EMBO Rep.*, 9, 134-138.
- Gerhart,J. (1999) 1998 Warkany lecture: signaling pathways in development. *Teratology*, 60, 226-239.
- Grunwald,I.C. and Klein,R. (2002) Axon guidance: receptor complexes and signaling mechanisms. *Curr. Opin. Neurobiol.*, 12, 250-259.
- Guo,X. and Wang,X.F. (2009) Signaling cross-talk between TGF-beta/BMP and other pathways. *Cell Res.*, 19, 71-88.
- Haan,C. et al. (2006) Jaks and cytokine receptors--an intimate relationship. *Biochem. Pharmacol.*, 72, 1538-1546.
- Hanna-Rose,W. and Han,M. (2000) Getting signals crossed in C. elegans. *Curr. Opin. Genet. Dev.*, 10, 523-528.
- Harris,M.A. et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32, D258-D261.
- Huang,S. et al. (2006) Activation of the hedgehog pathway in human hepatocellular carcinomas. *Carcinogenesis*, 27, 1334-1340.
- Hurlbut,G.D. et al. (2007) Crossing paths with Notch in the hyper-network. *Curr. Opin. Cell Biol.*, 19, 166-175.
- Ilagan,M.X. and Kopan,R. (2007) SnapShot: notch signaling pathway. *Cell*, 128, 1246.
- Itoh,M. et al. (2003) Mind bomb is a ubiquitin ligase that is essential for efficient activation of Notch signaling by Delta. *Dev. Cell*, 4, 67-82.
- Itoh,S. and ten Dijke,P. (2007) Negative regulation of TGF-beta receptor/Smad signal transduction. *Curr. Opin. Cell Biol.*, 19, 176-184.
- Joshi-Tope,G. et al. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33, D428-D432.
- Kandasamy,K. et al. (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biol*, 11, R3.
- Katoh,M. and Katoh,M. (2007) WNT signaling pathway and stem cell signaling network. *Clin. Cancer Res.*, 13, 4042-4045.
- Kelso,J. et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, 13, 1222-1230.
- Keshava Prasad,T.S. et al. (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, 37, D767-D772.

- Kyriakis,J.M. and Avruch,J. (2001) Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation. *Physiol Rev.*, 81, 807-869.
- Li,D. et al. (2008) PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Mol. Cell Proteomics.*, 7, 1043-1052.
- Long,J. et al. (2003) Repression of Smad transcriptional activity by PIASy, an inhibitor of activated STAT. *Proc. Natl. Acad. Sci. U. S. A.*, 100, 9791-9796.
- Lu,L.J. et al. (2007) Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends Biochem. Sci.*, 32, 320-331.
- Luo,W. and Lin,S.C. (2004) Axin: a master scaffold for multiple signaling pathways. *Neurosignals.*, 13, 99-113.
- Maniatis,T. (1999) A ubiquitin ligase complex essential for the NF-kappaB, Wnt/Wingless, and Hedgehog signaling pathways. *Genes Dev.*, 13, 505-510.
- Manning,G. et al. (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.*, 27, 514-520.
- Natarajan,M. et al. (2006) A global analysis of cross-talk in a mammalian cellular signalling network. *Nat. Cell Biol.*, 8, 571-580.
- Niwa,Y. et al. (2005) Methylation silencing of SOCS-3 promotes cell growth and migration by enhancing JAK/STAT and FAK signalings in human hepatocellular carcinoma. *Oncogene*, 24, 6406-6417.
- Nusse,R. (2003) Wnts and Hedgehogs: lipid-modified proteins and similarities in signaling mechanisms at the cell surface. *Development*, 130, 5297-5305.
- Ogata,H. et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27, 29-34.
- Orchard,S. et al. (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, 25, 894-898.
- Papin,J.A. et al. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.*, 6, 99-111.
- Pires-daSilva,A. and Sommer,R.J. (2003) The evolution of signalling pathways in animal development. *Nat. Rev. Genet.*, 4, 39-49.
- Rhodes,D.R. et al. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia.*, 9, 166-180.
- Roberts,A.B. et al. (2001) Is Smad3 a major player in signal transduction pathways leading to fibrogenesis? *Chest*, 120, 43S-47S.
- Robinson,D.R., Wu,Y.M. and Lin,S.F. (2000) The protein tyrosine kinase family of the human genome. *Oncogene*, 19, 5548-5557.

- Rogers,A. et al. (2008) WormBase 2007. *Nucleic Acids Res.*, 36, D612-D617.
- Roudabush,F.L. et al. (2000) Transactivation of the EGF receptor mediates IGF-1-stimulated shc phosphorylation and ERK1/2 activation in COS-7 cells. *J. Biol. Chem.*, 275, 22583-22589.
- Rzhetsky,A., Shatkay,H. and Wilbur,W.J. (2009) How to get the most out of your curation effort. *PLoS. Comput. Biol.*, 5, e1000391.
- Sammar,M. et al. (2004) Modulation of GDF5/BRI-b signalling through interaction with the tyrosine kinase receptor Ror2. *Genes Cells*, 9, 1227-1238.
- Sternberg,P.W. (2005) Vulval development. *WormBook.*, 1-28.
- Su,A.I. et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 6062-6067.
- Takigawa,Y. and Brown,A.M. (2008) Wnt signaling in liver cancer. *Curr. Drug Targets.*, 9, 1013-1024.
- Tseng,S.Y. and Dustin,M.L. (2002) T-cell activation: a multidimensional signaling network. *Curr. Opin. Cell Biol.*, 14, 575-580.
- Vardy,D.A. et al. (1995) Induction of autocrine epidermal growth factor receptor ligands in human keratinocytes by insulin/insulin-like growth factor-1. *J. Cell Physiol*, 163, 257-265.
- Wang,C.C., Cirit,M. and Haugh,J.M. (2009a) PI3K-dependent cross-talk interactions converge with Ras as quantifiable inputs integrated by Erk. *Mol. Syst. Biol.*, 5, 246.
- Wang,M. et al. (2009b) Expression of Notch1, Jagged1 and beta-catenin and their clinicopathological significance in hepatocellular carcinoma. *Neoplasma*, 56, 533-541.
- Weinmaster,G. (1997) The ins and outs of notch signaling. *Mol. Cell Neurosci.*, 9, 91-102.
- Wurmbach,E. et al. (2007) Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology*, 45, 938-947.
- Yan,S.J. et al. (2004) Multiple signaling pathways and a selector protein sequentially regulate Drosophila wing development. *Development*, 131, 285-298.
- Yu,H. et al. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, 14, 1107-1118.
- Zhang,Q. et al. (2003) Homeodomain interacting protein kinase 2 promotes apoptosis by downregulating the transcriptional corepressor CtBP. *Cell*, 115, 177-186.