Bioinformatics, 32(4), 2016, 571–578
doi: 10.1093/bioinformatics/btv620
Advance Access Publication Date: 27 October 2015
Original Paper

Systems biology

PPIXpress: construction of condition-specific protein interaction networks based on transcript expression

Thorsten Will[1,2] and Volkhard Helms[1,*]

[1]Center for Bioinformatics and [2]Graduate School of Computer Science, Saarland University, Saarbrücken, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 5, 2015; revised on September 28, 2015; accepted on October 20, 2015

Abstract

Summary: Protein–protein interaction networks are an important component of modern systems biology. Yet, comparatively few efforts have been made to tailor their topology to the actual cellular condition being studied. Here, we present a network construction method that exploits expression data at the transcript-level and thus reveals alterations in protein connectivity not only caused by differential gene expression but also by alternative splicing. We achieved this by establishing a direct correspondence between individual protein interactions and underlying domain interactions in a complete but condition-unspecific protein interaction network. This knowledge was then used to infer the condition-specific presence of interactions from the dominant protein isoforms. When we compared contextualized interaction networks of matched normal and tumor samples in breast cancer, our transcript-based construction identified more significant alterations that affected proteins associated with cancerogenesis than a method that only uses gene expression data. The approach is provided as the user-friendly tool PPIXpress.

Availability and implementation: PPIXpress is available at https://sourceforge.net/projects/ppixpress/.

Contact: volkhard.helms@bioinformatik.uni-saarland.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein–protein interaction networks (PPIN) are an important pillar of data integration in computational biology and have been used in a large number of studies and approaches. Generally, such networks are collections of physical interactions between pairs of proteins compiled from different experiments (Vidal *et al.*, 2011).

Full PPINs provide a convenient overview of the interactome of an organism. Yet, they do not reflect the true wiring exhibited by the cell in a specific state, because an interaction can only be realized if both partners are available. Pruning the full network to the set of proteins whose genes are expressed in the same condition has proven to be a straightforward solution for this. This allowed investigating the interaction landscape across tissues (Bossi and Lehner, 2009; Lopes *et al.*, 2011; Sinha and Nagarajaram, 2014) as well as the origin of tissue-specific diseases (Barshir *et al.*, 2014). Furthermore, it improved the prediction of disease genes (Magger *et al.*, 2012).

An estimated 95% of human multi-exon genes undergo alternative splicing (AS) (Pan *et al.*, 2008) and the specific isoform of a protein was shown to have a considerable impact on its ability to bind interaction partners (Buljan *et al.*, 2012; Ellis *et al.*, 2012; Miederer *et al.*, 2015). Thanks to the ability of quantifying individual transcripts nowadays, it thus appears worthwhile to also increase the granularity of condition-specific networks to this resolution.

Domain–domain interaction networks (DDIN) depict interactions between individual protein domains and provide a convenient framework to relate interaction sites with sequence information. In contrast to models based on atomistic structural data, DDINs allow for universal applicability (Ozawa *et al.*, 2010; Ma *et al.*, 2012; Will and Helms, 2014). So far, the only methodical effort regarding the effect of AS on interaction networks is found in the Cytoscape 2.x plugin DomainGraph. When linked to the AS analysis tool AltAnalyze, DomainGraph can highlight protein domains in DDINs that are affected by differential exon usage (Emig *et al.*, 2010). However, this tool is intended for visual exploration. While the user can manually estimate the implications of respective changes as PPIN and DDIN are visualized together, the tool does not allow to automatically infer conclusions for the PPIN on a whole-proteome scale.

With PPIXpress, we aim here at providing a simple standalone solution for the automatic construction of condition-specific protein interaction networks based on domain information and transcript expression. In addition to this core functionality, the tool is able to retrieve current protein interaction data, to add functional association scores from the STRING database (Szklarczyk *et al.*, 2015) to unweighted networks, and it can output the underlying condition-specific DDIN for each sample. It allows the usage of compressed input files in the gzip-format and is well-suited for batch-processing.

2 Materials and methods

2.1 PPIXpress

The input data for PPIXpress consists of a reference PPIN with condition-unspecific interactions and at least one sample of transcript- or gene-level expression data. From that, the tool constructs the condition-specific subnetworks for each transcriptome. Thus each network only comprises those interactions from the reference that are considered active in the sample.

Networks can be provided in the simple input format (either interacting UniProt, HGNC or Ensembl gene pairs line-by-line, optionally with a weight) or can alternatively be retrieved from IntAct (Orchard *et al.*, 2014) for a certain organism. The current version supports expression data in the following formats: Cufflinks FPKM files (Trapnell *et al.*, 2010), GENCODE or comparable Ensembl-annotated GTF files (Harrow *et al.*, 2006), TCGA RNASeq data or textfiles with expression-levels per line as commonly exported by popular R-based tools. All other data sources that are used internally are automatically retrieved in their most current versions. Furthermore, the user may optionally change the expression threshold (absolute or percentile-based) that is applied, limit the analysis to the gene-level, or inquire specific versions of retrieved data.

The basic principle of PPIXpress is outlined in Figure 1 and will be explained in the following paragraphs. Details regarding the annotation with domain data and datasets are covered in Section 2.2.

2.1.1 Relating protein and domain interactions

In the initial mapping stage, a one-to-at-least-one relationship between interactions in the given PPIN and the corresponding DDIN is established such that all PPIs found in the reference PPIN should be supported by at least one underlying DDI. In this step, PPIXpress considers the longest isoform of each protein as its representative in the DDIN, because large-scale experimental analyses and most databases usually declare it as the principal variant (Talavera *et al.*, 2013; Rodriguez *et al.*, 2013). Hence, the annotated domain compositions of the longest isoforms are used to construct a network on the domain-level that is then related to the reference PPIN. According to a dataset of feasible interactions between domain types (see Section 2.2), edges between interacting domains of distinct proteins are established in the DDIN if the protein pair is also connected in the PPIN. Thereby, it is noted which DDI or which DDIs support each individual interaction between proteins in the network because different DDIs may ratify the same PPI on this level.

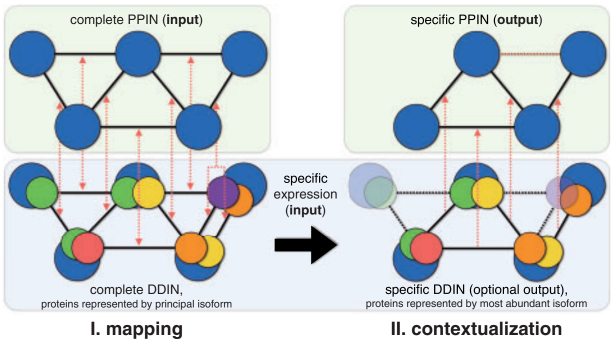
complete PPIN (**input**)
specific PPIN (**output**)
specific
expression
(**input**)
complete DDIN,
proteins represented by principal isoform
specific DDIN (optional output),
proteins represented by most abundant isoform
I. mapping
II. contextualization

Fig. 1. The PPIXpress approach can be divided into two stages. Initially, complete PPIN and DDIN are related to each other, whereby artificial domains (here shown in green) may be introduced to ensure a complete connectivity on the domain-level. This correspondence is then used to filter sample-specific domain–domain interactions (DDI) derived from transcript expression data and to map these back to the supported protein–protein interactions (PPI). The details are covered in the main text. Proteins without any expressed transcripts, as well as domains that are not found in the most abundant transcript are shown translucent. In this example, a method that only uses gene expression data would miss the disappearance of the protein interaction shown as a red dashed horizontal line in the top right picture

If a protein interaction cannot be assigned to any domain interaction at this stage, we add artificial domains to the affected proteins. Those domains are utilized to also establish links in the DDIN between those interaction partners in the reference PPIN whose binding cannot be explained otherwise by available domain interaction data. Introduction of such artificial domains allows our approach to sustain a complete correspondence between the two network-layers. Adding fictitious protein domains to overcome the sparsity of domain-level data was introduced before to improve the performance of protein complex prediction approaches that make use of such data (Ma *et al.*, 2012; Will and Helms, 2014). While Ma *et al.* (2012) introduced this idea in a non-deterministic way, PPIXpress uses a deterministic approach as described in Will and Helms (2014). These non-physical domains are thought to be present in every transcript coding for the protein and thus implement the behavior of gene-based methods where domain-level annotation fails to explain the protein-level outcome. This way, the methodology guarantees a seemless and safe transition to the performance of the gene-level approach whenever available data coverage on DDIs cannot explain the macroscopic observation.

2.1.2 Condition-specific construction

After the sample-independent mapping step, the expression data is incorporated to contextualize the network. Initially, all transcripts above a user-defined threshold are determined for the specific expression sample. From those, only the most abundant transcript of each protein is chosen to build a sample-specific DDIN, whereas all others are neglected. Based on this specific DDIN derived from the expression data and the previously determined mapping of DDIs to PPIs, a specific PPIN is constructed that only contains interactions that are supported by domain-level evidence. Here, it is not important if an individual PPI is backed by one or several DDIs; the existence of a single support is sufficient.

Viewed differently, PPIXpress used with transcript data first prunes the reference PPIN in a node-specific manner such as the established methods that are based on gene expression, but additionally trims the network in an edge-specific way guided by the domain data. The resulting network is therefore always a subnetwork of one obtained from a construction method based one gene expression. Figure 1 shows an example for this (red dashed interaction). These additionally considered 'edgetic' changes, as they are called in recent literature, are increasingly thought to be of crucial importance for phenotypic traits (Zhong *et al.*, 2009; Sahni *et al.*, 2013, 2015). If PPIXpress is switched to the gene-level mode all genes with expressed transcripts (or all above the threshold if only gene expression data is given) are taken into account. The longest coding transcript, the same reference as in the initial mapping, is selected as the representative of the protein. Thus the gene-level behavior is replicated while the specific DDINs are also reported.

Although data and methodology would in principle allow to process the contribution of a weighted ensemble of transcripts at this stage, we decided to introduce the strong assumption to discard all but the most abundant transcripts per protein. On the one hand, there is increasing biological evidence that generally only one dominant transcript per gene acts as the main contributor in a cellular condition (Gonzalez-Porta *et al.*, 2013; Ezkurdia *et al.*, 2015; Mele *et al.*, 2015). On the other hand, quantifying the distribution would require several additional parameters that may render the model unnecessarily complex and consequently the tool less appealing to the user. The discretization thus equally satisfies biological as well as practical considerations.

2.2 PPIXpress: datasets and protocols

2.2.1 Protein interaction networks

Self-interactions between proteins are not considered by PPIXpress because they interfere with classical types of network analysis such as complex prediction or disease gene prioritization. Furthermore, if the input PPIN is annotated with one of the non-UniProt accessions, they are converted using the HGNC webservice or Biomart (Smedley *et al.*, 2015), depending on the identifiers at hand.

2.2.2 Domain annotations

Internally, the tool queries UniProt (Bateman *et al.*, 2015) to infer the organism that is dealt with from the network data. With this knowledge, all required annotation data is retrieved from the appropriate and most recent Ensembl database (Cunningham *et al.*, 2015) by MySQL queries. The data comprise the relations between proteins, transcripts and genes, but also the assignment of resulting protein domains to transcripts. Only transcripts that can be directly associated to Swiss-Prot proteins are considered.

PPIXpress uses domain annotations derived from the manually curated Pfam-A database (Finn *et al.*, 2014b). Pfam domains in Ensembl are detected for each transcript individually using InterProScan (Jones *et al.*, 2014) and are automatically updated with every new release. As Pfam-A domains are non-overlapping and have predetermined family-specific detection thresholds that are used by InterProScan to filter for matches, neither additional parameters nor any postprocessing are needed within PPIXpress for this step (Finn *et al.*, 2010). Moreover, queries and internal data structures are designed to reflect the repeated occurrence of the same domain type within a protein in the optionally returned sample-specific DDINs.

2.2.3 Domain interaction data

To provide a comprehensive knowledgebase of physical interactions between protein domain types with PPIXpress, we precompiled high-confidence domain interaction data from DOMINE (Yellaboina *et al.*, 2011) and IDDI (Kim *et al.*, 2012). Both are integrated databases that assign reliability estimations to their available datasets. In DOMINE (version 2.0) interactions were classified into disjoint categories according to their estimated confidence. In IDDI (release May 2011) numerical confidence values were assigned to each interaction. As those primary resources appear not to be updated anymore, we additionally integrated automatic retrieval of current data from the 3did (Mosca *et al.*, 2014) and iPfam (Finn *et al.*, 2014a) databases whose interaction data is exclusively inferred and automatically updated from the RCSB Protein Data Bank (Berman *et al.*, 2000).

By default, PPIXpress uses a high-confidence subset of DOMINE and IDDI (see definition of PRE_{HC} in Section 2.3.4) expanded by the most recent 3did/iPfam data. Interactions between domains of the same type are taken into account if they are annotated.

2.3 Evaluation: datasets and protocols

2.3.1 Protein interaction networks

Data of experimentally determined physical interactions between proteins in human (*H. sapiens*, taxon 9606), mouse (*M. musculus*, taxon 10090), fruit fly (*D. melanogaster*, taxon 7227) and yeast (*S. cerevisiae S288c*, taxon 559292) were retrieved from IntAct (release 189) (Orchard *et al.*, 2014) using PPIXpress. For human we additionally compiled a second PPIN from physical interactions between human proteins in BioGRID (release 3.3.124) (Chatr-Aryamontri *et al.*, 2015). Here, a conversion to UniProt accessions was carried out using mapping data from HGNC (Gray *et al.*, 2015) that was downloaded on May 5., 2015.

2.3.2 Expression data

For the case study, transcript expression data for breast cancer (BRCA) was retrieved from TCGA (Koboldt *et al.*, 2012) as level 3 Illumina HiSeq-RNASeq V2 data based on RSEM quantification (Li and Dewey, 2011) and filtered to the portion of 112 matched normal/tumor samples (last updated January 14, 2015).

Since it is a common threshold across popular RNA-seq quantification methods (Diez *et al.*, 2014; Barshir *et al.*, 2014; Sinha and Nagarajaram, 2014), by default all transcripts (or genes if only gene expression data is given) with an abundance value above 1.0 are considered as expressed in PPIXpress. For the case study we also used this standard threshold.

2.3.3 Domain annotations

For all conducted analyses we used data from Ensembl release 79.

2.3.4 Domain interaction data

To evaluate the potential influence of the DDI dataset on the results, we compiled different subsets of data from the aforementioned sources (see Section 2.2.3): PRE_{HC} only contained those interactions from DOMINE that were inferred from structure or within the category of highest-confidence predictions and those interactions from IDDI whose confidence values exceed a threshold associated with an accuracy of 90% in the benchmarks of their original publication. PRE_{VHC} is a subset of PRE_{HC} that was restricted to the experimentally known interactions in DOMINE and the portion of IDDI that achieved the highest accuracy of 98% in Kim *et al.* (2012). 3did/iPfam contains the retrieved data from these two structure-based databases and ALL-DDI denotes the merged dataset $\text{PRE}_{\text{HC}} \cup$ 3did/iPfam. All conducted analyses were based on data from iPfam version 1.0 and 3did version 2015_02. Table 1 outlines the respective sizes of the four DDI datasets used.

Table 1. Amount of domain–domain interaction data in the different datasets used

	ALL-DDI	PRE_{HC}	3did/iPfam	PRE_{VHC}
Domain types	7449	6193	5920	4354
Domain interactions	30 551	26 377	10 953	6285

2.3.5 Whole-genome rewiring of protein interaction networks

For all 112 cases in TCGA with matched BRCA data from both normal and tumor tissue from the same patient, we constructed condition-specific protein interaction networks for both states and counted the changes in every comparison across all matched samples. To keep track of the changes within the interactome we marked each interaction that was only observable in the network of the disease sample as positive count and those that appeared only in the network of the healthy sample as negative. Figure 2 illustrates the overall approach for the network construction and comparison. For the evaluation the networks were constructed with PPIXpress using different methodologies and data. All steps were assessed for all settings individually. To obtain comparable abundance thresholds for protein precursors in all methods, a gene was considered to be abundant if at least one of its transcripts was abundant in a given dataset.

Besides noting the changes across all samples, a rewiring probability P_{rew} per interaction was individually computed for each matched sample pair and then averaged over all samples. P_{rew} was approximated as the number of rewiring events (interactions added + interactions removed between samples) divided by the number of interactions in normal or tumor, whichever was smaller. Since such differential networks summarized over all matched samples will inevitably contain many changes that occur only in few patients, we added a filtering step. On the basis of P_{rew} a one-tailed binomial test was applied to check how likely it was to observe a certain number of rewiring events of an individual interaction over all samples by chance. P-values were adjusted using Benjamini–Hochberg (Benjamini and Hochberg, 1995) and only the interactions with adjusted p-values below 0.05 were retained.

2.3.6 Randomized implementation of PPIXpress

To assess the assumption that only the most abundant transcript of each protein contributes to the specific DDIN (see Section 2.1.2), we modified PPIXpress to randomly select any of the transcripts above the expression threshold for each protein instead. For this randomized implementation we repeated the evaluation of the case study 100 times. As the construction method was applied to 112 × 2 samples (all matched pairs) per iteration during that process and the variance among the results was quite low, we think 100 iterations were sufficient for this comparison. Since the ALL-DDI dataset was used with the randomized method it is referred to as RANDOM(ALL-DDI) in the following tables.

2.3.7 'Hallmarks of cancer' data and analysis

We associated proteins with 10 currently established hallmarks of cancer on the basis of a handcrafted list of relevant GO terms by Suzuki *et al.* (2014). We retrieved all proteins in human with such an annotation using QuickGO (Binns *et al.*, 2009) on May 5, 2015. Associations inferred from automatic annotation (GO evidence IEA) were discarded as those are often inferred from protein interactions. A protein interaction was associated with a hallmark term if at least one of its involved proteins was part of the corresponding set of hallmark proteins.

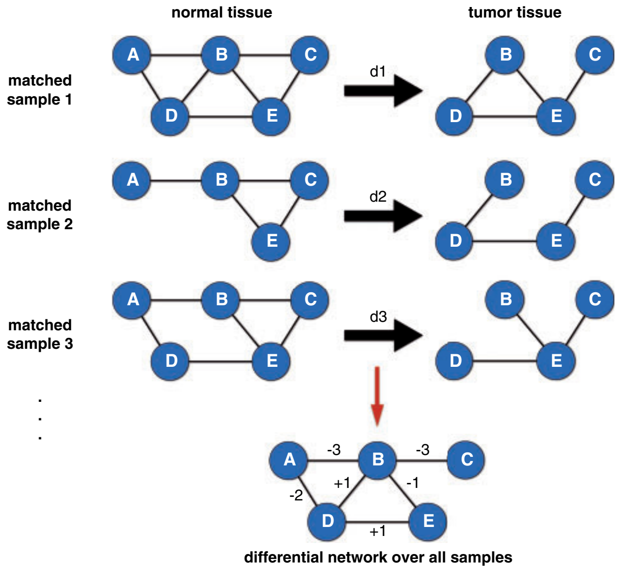

Fig. 2. For all matched BRCA samples from TCGA we built protein interaction networks using different methodologies. d1 to d112 denote changes in topology between normal and tumor interaction network in each matched pair. These differences were determined in every single patient and summed up in a differential network shown at the bottom. Here, the interaction between proteins A and B, for example, disappeared for all three shown matched sample pairs. Thus the edge between A and B is annotated with −3 in the differential network shown at the bottom

2.3.8 Enrichment analysis

Enrichment analysis was performed using DAVID 6.7 (Huang *et al.*, 2009). We specifically checked for enriched KEGG pathways (Kanehisa *et al.*, 2014) and GO biological processes (Blake *et al.*, 2015), set the proteins included in the respective input network as the background and kept the default settings of DAVID otherwise.

3 Results and discussion

3.1 Coverage of DDI datasets in practice

We first examined how many protein interactions are typically supported by at least one non-artificial domain interaction in the mapping stage of PPIXpress (interaction coverage) and how many proteins have domain annotations that contribute to that (protein coverage). We did this across various reference PPINs of several organisms and on the basis of different high-confidence DDI datasets as described in Section 2.3.4.

The results are shown in Table 2. In all cases, a larger dataset allowed to relate a larger part of the reference protein interactions to known domain interactions. $PRE_{HC,}$ for example, contained around 2.4 times as many DDIs as 3did/iPfam (compare respective columns in Table 1) and could relate 2.3–4.8 times more PPIs to DDIs depending on the reference network examined (compare PRE_{HC} and 3did/iPfam in Table 2). The addition of recent structural data from 3did/iPfam to the precompiled integrated dataset only led to a small improvement in interaction coverage (for all networks consistently below 1%, compare ALL-DDI and PRE_{HC} in Table 2). The human interactomes had the best coverage of interactions for all DDI data examined. However, even in the best case, still only about half of the proteins and roughly a fourth of the interactions could be associated with supporting domain information at all. Since the density of the PPINs was very heterogeneous (avg. degrees ranged from 2.2 to 12.7, see Table 2), the ratio of proteome and

Table 2. The annotation coverage of DDINs for different reference PPINs and different DDI datasets

		Size of network		Fraction of contributing proteins/fraction of matched PPIs			
Organism	Data source	Proteins/interactions	Avg. degree	ALL-DDI	$\mathrm{PRE}_{\mathrm{HC}}$	3did/iPfam	$\mathrm{PRE}_{\mathrm{VHC}}$
Human	BioGRID	15 086/156 271	10.4	0.517/0.264	0.506/0.256	0.364/0.099	0.334/0.093
Human	IntAct	13 665/81 460	6.0	0.437/0.246	0.428/0.241	0.287/0.103	0.259/0.097
Mouse	IntAct	7149/15 742	2.2	0.264/0.173	0.259/0.169	0.144/0.068	0.135/0.068
Fruit fly	IntAct	10 178/38 592	3.8	0.102/0.039	0.099/0.037	0.051/0.014	0.041/0.012
Yeast	IntAct	5993/76 003	12.7	0.530/0.186	0.513/0.181	0.272/0.038	0.230/0.036

Contributing proteins are all proteins that had at least one annotated domain that was used to span the DDIN in the mapping stage of PPIXpress. Matched PPIs are all PPIs that could be related to at least one DDI at this stage. The degree of a protein is the number of interaction partners in the respective network.

interactome coverage is not meaningful across networks. Interestingly, the coverage of proteins associated with hallmarks of cancer was higher than those of non-hallmark proteins (Supplementary Tables S1/S2).

Whereas this analysis suggests that the partial coverage of domain annotation may reduce the value of the proposed approach, it emphasizes the importance of a flexible approach that is able to integrate both well and poorly annotated proteins seamlessly. What are the reasons for that tenuous coverage? While the vast majority of proteins in human and yeast are annotated with at least one Pfam domain (Mistry *et al.*, 2013), even in our largest dataset only 7449 of the 14 831 domain types in Pfam 27.0 (Finn *et al.*, 2014b) have known domain interactions. Aside from the fact that some domain types may not be meant to facilitate protein interactions anyhow, experimental coverage of domain interactions is still sparse and not expected to near completion in the near future (Yellaboina *et al.*, 2011; Goncearenco *et al.*, 2014). Even if data on actual interactions was more comprehensive, respective binding interfaces have to be related to conserved protein building blocks of any kind to make the data universally applicable. However, interactions can also be mediated by disordered regions between domains (Dunker *et al.*, 2005) which are difficult to account for by a general annotation scheme as Pfam and are underrepresented there (Mistry *et al.*, 2013). This is, for example, the case for the pluripotency transcription factor Oct4 (Esch *et al.*, 2013).

3.2 Case study: rewiring of protein interactions in breast cancer

Since deregulation of splicing factors and accompanying alterations in protein products are known to contribute to tumorigenesis (David and Manley, 2010; Venables *et al.*, 2009; Danan-Gotthold *et al.*, 2015), transcript-based network construction may benefit an analysis in that context. Thus we present as a case study a comparison of the changes in the interactome between matched healthy and tumor samples from 112 breast cancer patients as explained in Section 2.3.5 that we conducted with different network construction approaches.

In cancer, changes in the interaction network can be expected to include proteins that are associated with certain hallmarks of cancer and that are frequently found in biological processes and pathways related to cancerogenesis. Based on this assumption we assessed whether the transcript-based methodology of PPIXpress was advantageous to the established gene-based network adjustment and to what extent selecting particular DDI datasets influenced the results.

Moreover, we examined the effect of our decision to exclusively rely on the domain annotation of the most abundant transcript above the threshold for each protein (see Section 2.1.2). To evaluate this, we randomized the transcript selection in PPIXpress as explained in Section 2.3.6. On average only 1.55 ± 0.038 transcripts were expressed per protein of the BioGRID network in each sample and 1.58 ± 0.040 in IntAct. Unsurprisingly, the number of discrete domain assemblies among those expressed transcripts was even smaller (Supplementary Tables S3/S4) and they mostly resembled the domain composition of the longest-coding transcript (Supplementary Tables S5/S6), since the principal domain composition often remains consistent among different isoforms (Ezkurdia *et al.*, 2012).

An overview of the network sizes during the construction phase and some statistics are provided in Supplementary Tables S7–S9. Using either gene- or transcript-based filtering of the input PPIN, the normal to tumor conversion was accompanied with a net loss of around 130-150 proteins and 900–1200 interactions, depending on the reference PPIN (compare respective rows in Supplementary Tables S7/S8). In line with expectations, the networks constructed by the gene-based approach were always the largest ones and had on average around 20 proteins and 800 interactions more than networks built from transcript data using the ALL-DDI dataset, for example. Furthermore, the sizes of networks built using transcript-based approaches slightly decreased with the amount of DDI data involved (compare columns in Supplementary Tables S7/S8 and see Supplementary Table S9 for a statistical evaluation). Networks constructed with transcript resolution and the largest dataset ALL-DDI possessed on average around 10 proteins and 450 interactions less than those built using the smallest dataset $\mathrm{PRE}_{\mathrm{VHC.}}$ Both observations can be accounted for by the subnetwork property of our transcript-based construction that we explained in Section 2.1.2. Since the sizes of the constructed networks were similar to results by Barshir *et al.* (2014) who considered 12 669 protein-coding genes to be expressed in healthy breast tissue, we deem our expression threshold to be suitable. Independently of DDI data considered, all transcript-based methods detected more rewiring events across the individual matched sample pairs and overall a higher number of significant changes in interactions compared to the gene-based approach (see first two rows per network in Table 3 and respective rows in Supplementary Tables S7/S8). In the case of the ALL-DDI dataset, including transcript data into PPIXpress allowed to detect 357 additional significant rewiring events in BioGRID and 120 additional events in the IntAct network (compare respective columns in Table 3).

3.2.1 Hallmarks of cancer

We first checked how many of the significantly rewired interactions per construction method affected proteins that can be related to hallmarks of cancer (see Section 2.3.7 for definition). Table 3 shows aggregated results for this analysis. Details for the individual hallmark terms are given in Supplementary Tables S10/S11. A statistical

Table 3. Results for the rewiring analysis of the breast cancer versus normal interaction networks in terms of rewired interactions that affect proteins associated with hallmarks of cancer as defined by Suzuki *et al.* (2014)

	GENE	ALL-DDI	PRE_{VHC}	RANDOM(ALL-DDI)
BioGRID				
P_{rew}	0.067 ± 0.016	0.069 ± 0.017	0.068 ± 0.017	0.078 ± 0.017
Sign. rewired interactions	9754	10 111	10 022	8661 ± 55
Part. in any hallmark term	7028	7343	7273	6265 ± 42
Fraction in any hallmark term	0.721	0.726	0.726	0.723 ± 0.002
Avg. part. per hallmark term	1749	1841	1820	1529 ± 15
Avg. fraction per hallmark term	0.179	0.182	0.182	0.177 ± 0.001
IntAct				
P_{rew}	0.077 ± 0.019	0.079 ± 0.020	0.078 ± 0.019	0.086 ± 0.018
Sign. rewired interactions	5184	5304	5280	4783 ± 25
Part. in any hallmark term	3484	3587	3571	3168 ± 25
Fraction in any hallmark term	0.672	0.676	0.676	0.662 ± 0.002
Avg. part. per hallmark term	808	835	834	704 ± 9
Avg. fraction per hallmark term	0.156	0.157	0.158	0.147 ± 0.001

The rewiring of an interaction was defined as significant according to the statistical protocol described in Section 2.3.5. An interaction was said to participate in a hallmark term if one of its associated proteins belonged to the corresponding set of hallmark proteins. The rows labelled 'fraction' depict the relative proportion of hallmark-associated interactions among all detected interactions. Comprehensive results for the individual hallmark terms and all DDI datasets are provided in Supplementary Table S10 for the BioGRID network and in Supplementary Table S11 for IntAct, respectively.

assessment of the differences between the methods on the basis of Wilcoxon signed-rank test is made in Supplementary Tables S12/S13. Overall, a construction based on transcript expression was, independently of the reference PPIN and DDI dataset used, able to find a significantly larger number of differential interactions that could be associated with hallmarks of cancer than the gene-level approach (P <0.001 in all cases, see first column in Supplementary Table S12). For example, of the statistically relevant rewiring events revealed by the transcript-based construction on the basis of the ALL-DDI dataset 315 more interactions indeed affected at least one protein associated with any hallmark term compared to the interactions detected by gene expression and 103 more in IntAct, respectively (compare third row per network in Table 3 and Supplementary Tables S10/S11). The fraction of such interactions among all differential interactions was not significantly higher (see first column in Supplementary Table S13). Thus only the amount but not the density of relevant information was higher compared to a gene-based construction. With few exceptions this also held for individual hallmark terms (see fifth row per network in Table 3 for mean values and Supplementary Tables S10/S11 for details). Regarding the absolute number of interactions affecting certain protein sets, only in 'Genome Instability and Mutation' and 'Avoiding Immune Destruction' in IntAct some of the transcript-based runs performed slightly worse than the gene-based method (see respective rows in Supplementary Table S11). The runs based on the highest-confidence DDI dataset PRE_{VHC} never reported fewer interactions in any term category, though.

When the transcript per protein was randomized as explained in Section 2.3.6, the transcript-based analysis gave worse results than the gene-based and all non-randomized transcript-based approaches (Supplementary Tables S10–S13). In particular, significantly less hallmark-relevant interactions were detected ($P < 0.001$ in all cases, last row in Supplementary Table S12). The difference in relevant fractions per hallmark term was not significant, though (see last row Supplementary Table S13). Interestingly, it still found more interactions related to 'Enabling Replicative Immortality' in both reference networks than the gene-based methodology (see respective row in Supplementary Tables S10/S11). As this was the only example in all analyses where the randomized method was superior to the gene-based one, we examined in what regard the proteins in that set differed from all others, and how the gene-based method could lose that much predictive power there. There was no noteworthy difference in the coverage of the interactions among this subset of proteins but an increase in protein coverage compared to all other hallmark sets with 82% in BioGRID (avg. hallmark proteins: 68%, details in Supplementary Table S1) and 76% in IntAct (avg. hallmark proteins: 63%, details in Supplementary Table S2). Thus comparatively many proteins had at least one annotated domain that contributed to the DDI/PPI mapping, but the majority of the networks was still covered by artificial domains. Furthermore, the proteins associated with 'Enabling Replicative Immortality' had the most variable domain compositions among the expressed transcripts per protein in both BioGRID (8% more than any other protein subset, see Supplementary Table S3) and IntAct (6% more than any other protein subset, see Supplementary Table S4). Intriguingly, they also had the smallest fraction of expressed transcripts per protein that had the same domain composition as the principal protein isoform (2.5-3.4% smaller than any other protein subset, see Supplementary Tables S5/S6). Consequently, the proteins in 'Enabling Replicative Immortality' had the largest divergence from the principal domain composition among all protein subsets that we examined and thus behaved most different compared to the gene-based construction.

3.2.2 Enrichment of exclusively found interactions

Next we examined for all transcript-based variants one-by-one if significant changes were missed compared to a gene-based construction and which alterations were found in addition. To quantify the relevance of rewiring events exclusively reported by either method, enrichment analysis was performed on the affected proteins (see Section 2.3.8). An outline of the results is presented in Table 4, details are listed in Supplementary Tables S14/S15.

Generally, the results originating from gene- and transcript-based construction methods diverged more strongly the more DDI data was incorporated (see first rows in Supplementary Tables S14/S15). While a larger DDI dataset enabled transcript-based approaches to detect more interactions that were not considered by the gene-based adjustment, also more interactions that were

Table 4. Shown are significantly rewired interactions that were exclusively found either by the gene- or transcript-based methods and the proteins that are affected by them

	GENE/ALL-DDI	GENE/PRE$_{VHC}$	GENE/RANDOM(ALL-DDI)
BioGRID			
Common rew. interactions	9665	9716	8308 ± 7
Exclusive rew. interactions	89/446	38/306	$1445 \pm 7/352 \pm 54$
Affected proteins	117/424	58/326	$1401 \pm 5/344 \pm 47$
KEGG PWC	$1.0/1 \times 10^{-17}$	$1.0/5 \times 10^{-15}$	$(2 \pm 1) \times 10^{-10}/(1 \pm 8) \times 10^{-7}$
Enriched terms (KEGG)	2/34	1/19	$16 \pm 1/18 \pm 5$
Highest enrichment (KEGG)	$0.0013/2 \times 10^{-17}$	$0.0037/9 \times 10^{-16}$	$(2 \pm 0.1) \times 10^{-9}/(1 \pm 8) \times 10^{-10}$
Enriched terms (GO BP)	6/116	0/87	$108 \pm 4/97 \pm 15$
Highest enrichment (GO BP)	$7 \times 10^{-5}/4 \times 10^{-16}$	$1.0/6 \times 10^{-16}$	$(3 \pm 1) \times 10^{-12}/(5 \pm 0.4) \times 10^{-12}$
IntAct			
Common rew. interactions	5118	5155	4653 ± 16
Exclusive rew. interactions	66/186	29/125	$531 \pm 16/130 \pm 17$
Affected proteins	87/213	46/145	$571 \pm 8/150 \pm 17$
KEGG PWC	$1.0/9 \times 10^{-11}$	$1.0/4 \times 10^{-6}$	$(9 \pm 9) \times 10^{-6}/0.12 \pm 0.32$
Enriched terms (KEGG)	0/15	0/11	$21 \pm 1/7 \pm 4$
Highest enrichment (KEGG)	$1.0/4 \times 10^{-13}$	$1.0/4 \times 10^{-8}$	$(4 \pm 5) \times 10^{-13}/(3 \pm 0.1) \times 10^{-6}$
Enriched terms (GO BP)	0/35	1/20	$26 \pm 2/13 \pm 8$
Highest enrichment (GO BP)	$1.0/3 \times 10^{-6}$	$0.0367/3 \times 10^{-5}$	$(6 \pm 0.1) \times 10^{-12}/(4 \pm 8) \times 10^{-4}$

In all but the first lines per network, left values denote the outcome regarding interactions and proteins exclusively found in the gene-based approach and right values the same for the transcript-based construction. Affected proteins are proteins linked to significantly rewired interactions. Enrichment was determined according to Section 2.3.8 and defined as $P < 0.05$ (Bonferroni-adjusted). KEGG PWC abbreviates the enrichment of KEGG pathway 'hsa05200:Pathways in cancer' and GO BP abbreviates the GO category 'biological process'. Comprehensive results for all DDI datasets are provided in Supplementary Table S14 for the BioGRID network and in Supplementary Table S15 for IntAct, respectively.

detected by the gene-based approach were not detected (see second row per network in Table 4 and respective rows in Supplementary Tables S14/S15). In all cases, the exclusively found significant changes revealed by network construction based on transcripts featured many more enriched pathways and GO processes and also a much higher enrichment of individual terms compared to the portion of interactions that were only found by the gene-based approach. KEGG term 'hsa05200:Pathways in cancer', for example, was not enriched in the exclusive results of the gene-based approach but strongly in those of the transcript-based method, independent of the DDI dataset used (adjusted $P < 10^{-5}$ in all cases, see fourth row per network in Table 4 and respective rows in Supplementary Tables S14/S15). It is worth pointing out that the identified enriched terms are closely linked to carcinogenetic processes suggesting that the rewired interactions are not simply random alterations overall (Supplementary Tables S16/S17). The most prevalent changes exclusively found by the transcript-based method using the largest DDI dataset, for example, were found across 66 matched samples in both networks. Four of the five exclusively found rewiring events across both networks that occurred 66 times were related to the loss of an interaction of FLT1 (P17948, 'Vascular endothelial growth factor receptor 1'), a tyrosine-protein kinase that acts as a cell-surface receptor for several cancer-relevant signaling cascades.

When the transcripts used to construct the specific DDIN were randomized, the positive impact of the transcript-level data vanished in comparison to the established gene-based methodology (see last column in Table 4).

4 Conclusion

PPIXpress exploits domain interaction data to adapt protein interaction networks to specific cellular conditions at transcript-level detail. For the example of protein interactions in breast cancer we showed how this increase in granularity positively affected the performance of the network construction compared to a method that only makes use of gene expression data. A platform-independent and dependency as well as installation-free implementation is provided that only requires little manual effort by the user.

Funding

T.W. was supported by DFG via SFB 1027.

Conflict of Interest: none declared.

References

Barshir,R. *et al.* (2014) Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput. Biol.*, **10**, e1003632.

Bateman,A.A. *et al.* (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Binns,D. *et al.* (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045–3046.

Blake,J.A. *et al.* (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

Bossi,A. and Lehner,B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, **5**, 260.

Buljan,M. *et al.* (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell*, **46**, 871–883.

Chatr-Aryamontri,A. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.

Cunningham,F. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.

Danan-Gotthold,M. *et al.* (2015) Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Res.*, **43**, 5130–5144.

David,C.J. and Manley,J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.*, **24**, 2343–2364.

Diez,D. *et al.* (2014) Systematic identification of transcriptional regulatory modules from protein–protein interaction networks. *Nucleic Acids Res.*, **42**, e6.

Dunker,A.K. *et al.* (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.*, **272**, 5129–5148.

Ellis,J.D. *et al.* (2012) Tissue-specific alternative splicing remodels protein–protein interaction networks. *Mol. Cell*, **46**, 884–892.

Emig,D. *et al.* (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.*, **38**, W755–W762.

Esch,D. *et al.* (2013) A unique Oct4 interface is crucial for reprogramming to pluripotency. *Nat. Cell Biol.*, **15**, 295–301.

Ezkurdia,I. *et al.* (2012) Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol. Biol. Evol.*, **29**, 2265–2283.

Ezkurdia,I. *et al.* (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, **14**, 1880–1887.

Finn,R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Finn,R.D. *et al.* (2014a) iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res.*, **42**, D364–D373.

Finn,R.D. *et al.* (2014b) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Goncearenco,A. *et al.* (2014) Coverage of protein domain families with structural protein–protein interactions: current progress and future trends. *Prog. Biophys. Mol. Biol.*, **116**, 187–193.

Gonzalez-Porta,M. *et al.* (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.*, **14**, R70.

Gray,K.A. *et al.* (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, D1079–D1085.

Harrow,J. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**, 1–9.

Huang,D.A.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Jones,P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, 199–205.

Kim,Y. *et al.* (2012) IDDI: integrated domain–domain interaction and protein interaction analysis system. *Proteome Sci.*, **10**, S9.

Koboldt,D.C. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Lopes,T.J. *et al.* (2011) Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, **27**, 2414–2421.

Ma,W. *et al.* (2012) Protein complex prediction based on maximum matching with domain–domain interaction. *Biochim. Biophys. Acta*, **1824**, 1418–1424.

Magger,O. *et al.* (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.*, **8**, e1002690.

Mele,M. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Miederer,A.M. *et al.* (2015) A STIM2 splice variant negatively regulates store-operated calcium entry. *Nat. Commun.*, **6**, 6899.

Mistry,J. *et al.* (2013) The challenge of increasing Pfam coverage of the human proteome. *Database (Oxford)*, **2013**, bat023.

Mosca,R. *et al.* (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **42**, D374–D379.

Orchard,S. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.

Ozawa,Y. *et al.* (2010) Protein complex prediction via verifying and reconstructing the topology of domain–domain interactions. *BMC Bioinformatics*, **11**, 350.

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.

Rodriguez,J.M. *et al.* (2013) APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.*, **41**, D110–117.

Sahni,N. *et al.* (2013) Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.*, **23**, 649–657.

Sahni,N. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.

Sinha,A. and Nagarajaram,H.A. (2014) Nodes occupying central positions in human tissue specific PPI networks are enriched with many splice variants. *Proteomics*, **14**, 2242–2248.

Smedley,D. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.

Suzuki,A. *et al.* (2014) Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic Acids Res.*, **42**, 13557–13572.

Szklarczyk,D. *et al.* (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

Talavera,D. *et al.* (2013) Alternative splicing and protein interaction data sets. *Nat. Biotechnol.*, **31**, 292–293.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Venables,J.P. *et al.* (2009) Cancer-associated regulation of alternative splicing. *Nat. Struct. Mol. Biol.*, **16**, 670–676.

Vidal,M. *et al.* (2011) Interactome networks and human disease. *Cell*, **144**, 986–998.

Will,T. and Helms,V. (2014) Identifying transcription factor complexes and their roles. *Bioinformatics*, **30**, i415–i421.

Yellaboina,S. *et al.* (2011) DOMINE: a comprehensive collection of known and predicted domain–domain interactions. *Nucleic Acids Res.*, **39**, D730–D735.

Zhong,Q. *et al.* (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.*, **5**, 321.