

# SPECTRA: an integrated knowledge base for comparing tissue and tumor-specific PPI networks in human

Giovanni Micale<sup>1</sup>, Alfredo Ferro<sup>2</sup>, Alfredo Pulvirenti<sup>2\*†</sup> and Rosalba Giugno<sup>2\*†</sup>

<sup>1</sup> Department of Computer Science, University of Pisa, Pisa, Italy, <sup>2</sup> Department of Clinical and Molecular Biomedicine, University of Catania, Catania, Italy

## OPEN ACCESS

### Edited by:

Marco Pellegrini,  
Consiglio Nazionale delle Ricerche,  
Italy

### Reviewed by:

Arsen Arakelyan,  
Institute of Molecular Biology, Armenia  
Mohammed El-Kebir,  
Brown University, USA

### \*Correspondence:

Alfredo Pulvirenti and  
Rosalba Giugno,  
Department of Clinical and Molecular  
Biomedicine, University of Catania, Via  
Andrea Doria 6, Catania 95037, Italy  
apulvirenti@dmi.unict.it;  
giugno@dmi.unict.it

<sup>†</sup> Alfredo Pulvirenti and Rosalba  
Giugno have contributed equally to  
this work.

### Specialty section:

This article was submitted to  
Bioinformatics and Computational  
Biology, a section of the journal  
Frontiers in Bioengineering and  
Biotechnology

Received: 31 October 2014

Accepted: 17 April 2015

Published: 08 May 2015

### Citation:

Micale G, Ferro A, Pulvirenti A and  
Giugno R (2015) SPECTRA: an  
integrated knowledge base for  
comparing tissue and tumor-specific  
PPI networks in human.  
Front. Bioeng. Biotechnol. 3:58.  
doi: 10.3389/fbioe.2015.00058

Protein–protein interaction (PPI) networks available in public repositories usually represent relationships between proteins within the cell. They ignore the specific set of tissues or tumors where the interactions take place. Indeed, proteins can form tissue-selective complexes, while they remain inactive in other tissues. For these reasons, a great attention has been recently paid to tissue-specific PPI networks, in which nodes are proteins of the global PPI network whose corresponding genes are preferentially expressed in specific tissues. In this paper, we present SPECTRA, a knowledge base to build and compare tissue or tumor-specific PPI networks. SPECTRA integrates gene expression and protein interaction data from the most authoritative online repositories. We also provide tools for visualizing and comparing such networks, in order to identify the expression and interaction changes of proteins across tissues, or between the normal and pathological states of the same tissue. SPECTRA is available as a web server at <http://alpha.dmi.unict.it/spectra>.

**Keywords:** tissue, tumor, database, proteins, interactions

## 1. Introduction

In the last 10 years, there has been a rapid growth of available protein–protein interaction (PPI) data. They represent all known physical interactions between proteins within a cell. The collection of PPI data yields a network depicting a global overview of the relationships between proteins.

Nowadays, PPIs of species and associated data are stored in many databases, which usually are weekly or monthly updated. Primary sources of PPI data include BioGRID (Stark et al., 2006), DIP (Xenarios et al., 2000), HPRD (Peri et al., 2004), IntAct teorchard2013mintact, and MINT (Licata et al., 2012).

DIP (Xenarios et al., 2000) was the first database, which combined information from multiple observations and experimental techniques into networks of interacting proteins for different species. HPRD (Peri et al., 2004) contains manually curated proteomic information regarding human proteins, which are annotated and linked to OMIM database (Hamosh et al., 2005). BioGRID (Stark et al., 2006) collects protein–protein and genetic interactions for all major model organisms trying to remove redundancy and create a single mapping of interactions. The IntAct database (Orchard et al., 2013) provides tools for both textual and graphical representations of protein interactions. Interacting proteins can be annotated with GO terms for functional analysis. MINT (Licata et al., 2012), which is based on the IntAct database

infrastructure, collects experimentally verified PPIs by extracting experimental evidences from the scientific literature.

Some databases integrate PPIs data of human and other organisms from primary sources, by removing redundancies and assigning a unique reliability score. These include STRING (Franceschini et al., 2013), IRefIndex (Razick et al., 2008), ConsensusPathDB (Kamburov et al., 2013), and HitPredict (Patil et al., 2011).

STRING (Franceschini et al., 2013) combines physical interaction data and curated pathways of different organisms with predicted interactions from text mining, genomic features and interactions transferred from model organisms based on orthology. IRefIndex (Razick et al., 2008) is a set of tools to index and retrieve proteins and interactions from major public databases. Indexes are built according to protein sequences and taxonomy identifiers and mapping scores evaluate the quality of the mapping. ConsensusPathDB (Kamburov et al., 2013) integrates human protein–protein interactions, biochemical pathways, gene regulatory, and drug–target interactions into a global network, containing genes, proteins, and metabolites, which can be visualized, analyzed, and annotated. HitPredict (Patil et al., 2011) combines PPI data from IntAct (Orchard et al., 2013), BIOGRID (Stark et al., 2006), and HPRD (Peri et al., 2004), by assigning a confidence score based on sequence, structure, and functional annotations of the interacting proteins. The reliability score is calculated using the Bayesian networks.

The analysis of PPI networks has provided novel biological insights on the function of many previously uncharacterized proteins in *Human* through module identification (Bader and Hogue, 2003; Adamcsek et al., 2006; Mete et al., 2008; Rhrissorakrai and Gunsalus, 2011), network querying (Ferro et al., 2007; Banks et al., 2008; Bruckner et al., 2010), and network alignment (Flannick et al., 2006; Kalaev et al., 2009; Liao et al., 2009; Sahraeian and Yoon, 2013; Micale et al., 2014a) algorithms. Furthermore, the annotation of PPI networks with external data (i.e., diseases, expression data, phenotypes) has helped to classify genes according to the expression profiles (Dao et al., 2011), predict new gene–disease associations (Huang et al., 2012; Zhao et al., 2012), and discover new drugs (Huang et al., 2012; Alaimo et al., 2013; Csermerly et al., 2013).

These tasks have been accomplished thanks to the availability of authoritative repositories of gene expression data in normal/cancer tissues and at different diseases stages (Uhlen et al., 2010; Barrett et al., 2013; Rustici et al., 2013). For example, ArrayExpress (Rustici et al., 2013) and GEO (Barrett et al., 2013) include gene expression data from microarray and high-throughput sequencing experiments, which can be easily queried or downloaded. Users can also submit data directly by using the standard MIAME format. More recently, new projects have started with the aim of cataloging tissue or tumor sequencing data. The Cancer Genome Atlas (TCGA)<sup>1</sup> collects complete high-throughput genome data (clinical information, expressions data, methylations, mutations) for specific cancer tissues, with the purpose of helping the diagnosis and the treatment of cancers. The Human Protein Atlas (Uhlen et al., 2010) is a database with

histological images showing the spatial distribution of proteins in normal and cancer tissues. Protein Atlas contains also transcription expression levels, protein expression profiles, and subcellular localization data.

All above PPI networks data are constructed by ignoring the role of proteins in human tissues. On the other hand, human diseases often occur in specific tissues (Lage et al., 2008). Some genes can be predominantly expressed in one or few tissues and can control the formation of protein complexes (Emig and Albrecht, 2011). Furthermore, genes can use alternative splicing as a powerful mechanism to enlarge the number of their interactors and perform distinct functions in different tissues (Emig and Albrecht, 2011). Therefore, the integration of PPI networks with tissue-specific gene expression data can help to highlight the role of some genes in specific disease or tumors. The result of such integration gives the so called Tissue-Specific PPI (TS-PPI) network (Bossi and Lehner, 2009), which is a subgraph of a PPI network where the genes corresponding to both interacting proteins are expressed in one or more selected tissues.

Some studies focus on the analysis of global and local properties of TS-PPI networks. In Bossi and Lehner (2009), authors prove that most housekeeping proteins form highly tissue-specific protein interactions, suggesting a key role of those proteins in tissue-specific biological processes. Emig and Albrecht (2011) show that the number of tissue-specific proteins is very low and the receptor-activated signaling processes and the transcriptional regulation are two key factors for tissue specificity. In Souiai et al. (2011), a gradient model is used to describe the structure of TS-PPI networks, containing interactions of regulatory and developmental functions at the core of the TS-PPI network and physiological functions at the periphery.

Several recent works highlight the advantages of using TS-PPI networks. In Lopes et al. (2011), a set of proteins related to the response of viral infection in a TS-PPI network lead to a more reliable functional enrichment. Magger et al. (2012) use TS-PPI networks to improve the prioritization of candidate disease-causing genes with respect to a generic PPI network. In Chen and Wang (2012), authors identify functional modules in TS-PPI networks using CFinder (Adamcsek et al., 2006) and show that they exhibit more biological meaning than modules in a PPI network. Xiao et al. (2014) propose a new method for the identification of multi-tissue gene co-expression networks associated with specific functional processes relevant for phenotype variation and disease in humans. Barshir et al. (2014) show that genes causing hereditary diseases tend to have higher transcript levels and more interacting partners in the TS-PPI network of disease tissues than in the TS-PPI network of unaffected tissues.

To the best of our knowledge, few tools are available for querying and analyzing TS-PPI networks (Barshir et al., 2013; Nersisyan et al., 2014). CyKeggParser (Nersisyan et al., 2014) is a Cytoscape app for generating and analyzing tissue-specific KEGG pathways. Pathways can be checked for inconsistencies and modified based on gene expression data from normal and cancer tissues. TissueNet (Barshir et al., 2013) is a dataset of TS-PPIs in humans, which integrates a collection of four PPI networks (BioGRID, DIP, IntAct, and MINT) with three expression datasets (GEO, Human Protein Atlas, and Illumina Body Map 2.0). The database provides

<sup>1</sup><http://cancergenome.nih.gov>

a web interface for retrieving tissue-specific interactions of a query protein. However, it handles only 16 normal tissues and does not provide any tool for the analyses of TS-PPI networks.

In this paper, we propose SPECTRA (SPECific Tissue/Tumor Related PPI networks Analyzer), a framework to build and analyze TS-PPI networks. SPECTRA integrates tissue and tumor-specific gene expression data from the most authoritative online repositories such as Protein Atlas, ArrayExpress, GEO, and TCGA. Expression data are then integrated with high-quality protein-protein interactions, taken from HPRD, BioGRID, MIPS, IntAct, and the work of Havugimana et al. (2012). We provide a web interface for constructing, visualizing, and comparing TS-PPI networks, with the aim of identifying differential interaction/expression patterns in TS-PPI networks (i.e., distinct tissues, or normal and pathological states of the same tissue). The TS-PPI networks together with the results of differential analysis can be easily visualized by using Cytoscape facilities (Shannon et al., 2003) and downloaded as text files for further investigations. SPECTRA is free for all users and available at <http://alpha.dmi.unict.it/spectra>.

## 2. Materials and Methods

SPECTRA combines protein-protein interactions in human with gene expressions, by integrating 13 authoritative resources. The final integrated SPECTRA database contains 16,435 protein coding genes and 175,841 gene interactions (GIs), 1,350,637 tissue-specific gene expression data entries covering 107 normal tissues, and 2,171,808 tumor-specific expression data entries covering 160 different tumors.

### 2.1. Interaction Datasets

Human protein interaction data were taken from BioGRID<sup>2</sup>, DIP<sup>3</sup>, a recent work by Havugimana et al. (2012), HPRD<sup>4</sup>, IntAct<sup>5</sup>, and MINT<sup>6</sup>.

**Table 1** describes the features of the PPI networks integrated in SPECTRA. Networks taken from the work of Havugimana et al. (2012), IntAct and MINT are weighted with edge weights ranging in [0,1], while the other PPI networks are unweighted. Proteins of the considered PPI networks, including splicing isoforms, were first mapped to the corresponding gene. Next, a global GI network was built, by collecting all interactions reported

<sup>2</sup><http://thebiogrid.org>

<sup>3</sup><http://dip.doe-mbi.ucla.edu/dip>

<sup>4</sup><http://www.hprd.org>

<sup>5</sup><http://www.ebi.ac.uk/intact>

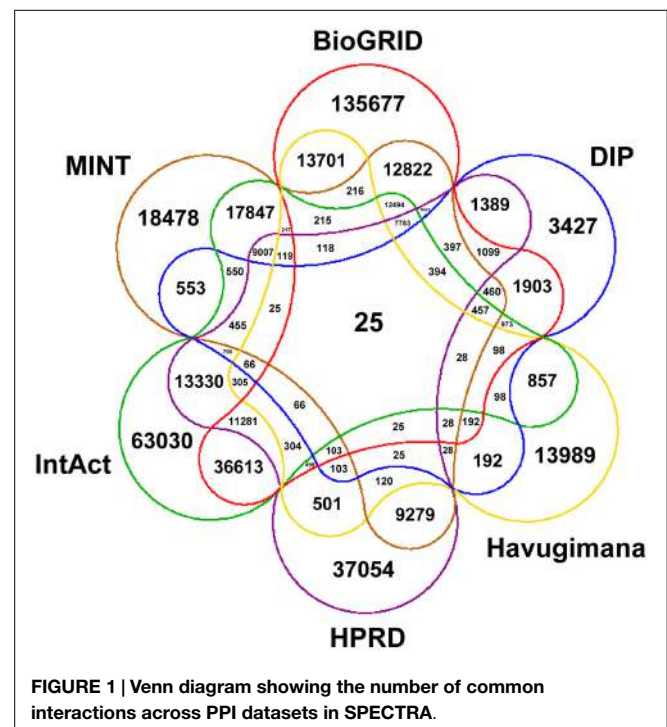
<sup>6</sup><http://mint.bio.uniroma2.it/mint>

**TABLE 1 | Features of PPI networks integrated in SPECTRA.**

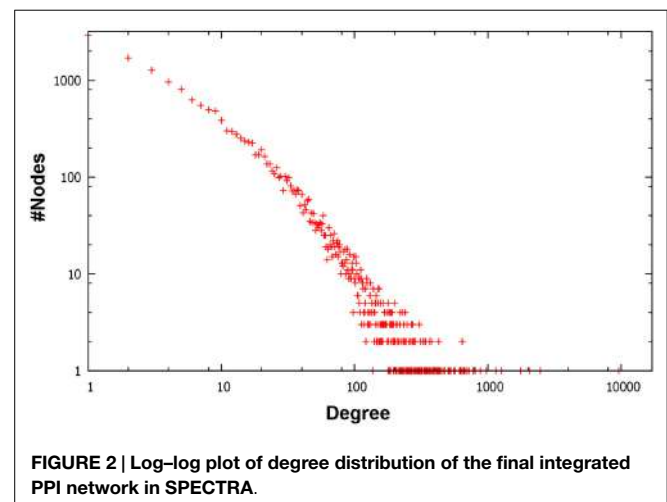
Network	Nodes	Edges	Type
BioGRID	15,290	135,677	Unweighted
DIP	2,338	3,427	Unweighted
Havugimana et al. (2012)	3,003	13,989	Weighted
HPRD	9,506	37,054	Unweighted
IntAct	11,637	63,030	Weighted
MINT	6,551	18,478	Weighted

in at least one dataset. We assigned to each edge a pair consisting of the average value of weights across the datasets that report that interaction and the percentage of datasets giving the interaction (dataset coverage). Average edge weights range from 0.131 to 1.

**Figure 1** depicts a Venn diagram of common gene interactions between PPI datasets. Interaction databases generally show low overlap, with only 25 interactions shared by all datasets and only 7,783 interactions in common between MINT, BioGRID, IntAct, and HPRD, which are the biggest ones. The final integrated network has 16,435 nodes, 175,841 edges and 17 connected components, with a high average diameter (9) and low clustering coefficient (0.289). The average degree is 21.398 and the degree distribution follows a power law (**Figure 2**).



**FIGURE 1 | Venn diagram showing the number of common interactions across PPI datasets in SPECTRA.**



**FIGURE 2 | Log-log plot of degree distribution of the final integrated PPI network in SPECTRA.**

**TABLE 2 | Features of expression datasets integrated in SPECTRA.**

Dataset	Platform	Tissues	Tumors
E-MTAB-62 (Lukk et al., 2010)	GPL96	46	110
GDS181 (Su et al., 2002)	GPL91	29	6
GDS596 (Su et al., 2004)	GPL96	57	5
GDS1096 (Ge et al., 2005)	GPL96	36	0
GDS3113 (Dezso et al., 2008)	GPL2986	32	0
ProteinAtlas	GPL11154	28	33
TCGA	Agilent G4502A-07-3	0	27

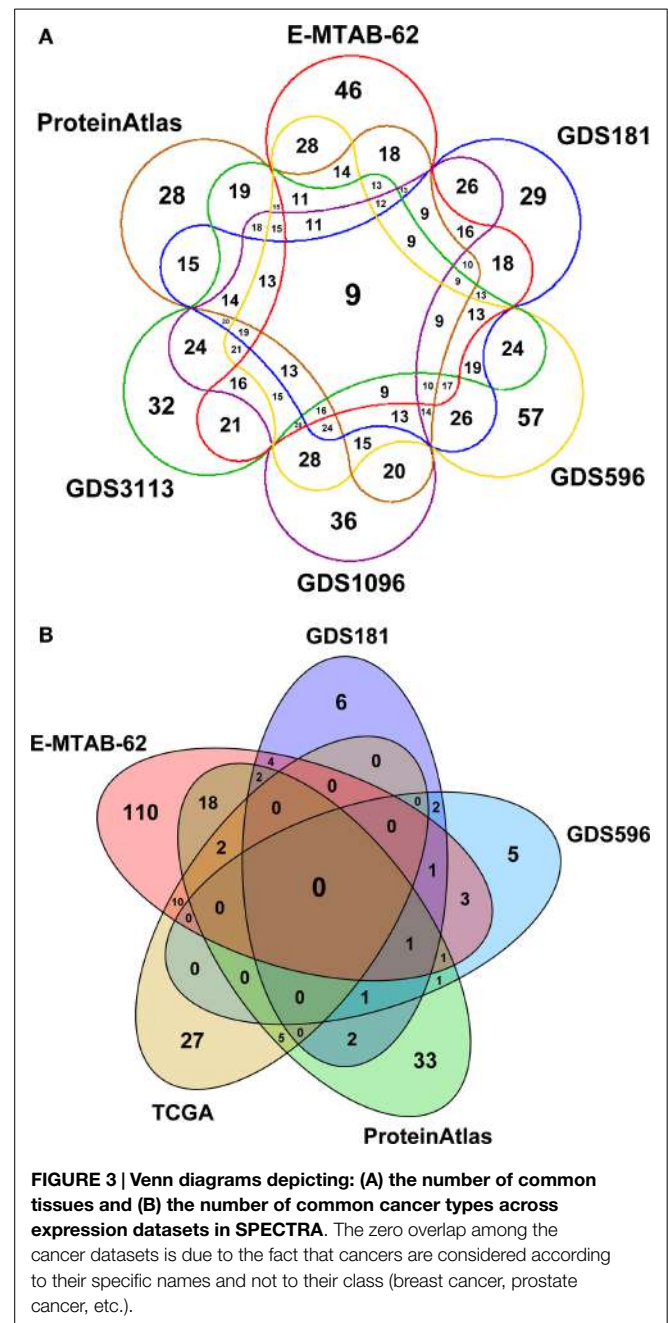
## 2.2. Expression Datasets

Gene expression data for various tissues and tumors were downloaded from ArrayExpress<sup>7</sup>, GEO<sup>8</sup>, ProteinAtlas<sup>9</sup>, and TCGA (see text footnote 1). **Table 2** lists the gene expression datasets integrated within SPECTRA, the platform used to detect the expressions and the number of covered tissues and tumors.

**Figure 3** depicts a Venn diagram of common tissues and tumors across expression datasets. While tissue names are generally shared, tumor names are much differentiated, resulting in a poor overlap between datasets. In particular, TCGA contains data for very specific tumors and partially overlap only with E-MTAB-62 dataset, which is the richest one. Note that the numbers reported in **Figure 3** only refer to specific tumors and not to tumor classes. So, for instance, “breast carcinoma” and “breast adenocarcinoma” are considered distinct tumors, even though they belong to the same class of tumors, “breast cancer.”

As regards the integration of expression data, we followed the work of Guo et al. (2013), where authors show that there is a positive correlation between normalized Affymetrix and RNA-Seq data. We performed RMA normalization (McCall et al., 2010) for datasets based on Affymetrix platforms (GDS181, GDS596, and GDS1096), using the corresponding R Bioconductor package. For GDS3113, ProteinAtlas, and TCGA, we first computed the  $\log_2$  of the number of fragments and then we normalized values using the quantile normalization method of Bolstad et al. (2003). For GDS181, GDS596, and GDS1096 datasets, normalized values were computed from raw data. Then, probes, which were present in a particular microarray dataset, were mapped to the corresponding genes. Finally, the expression of a gene for a specific tissue was computed as the average expression value of probes mapping to that gene in the tissue. For GDS3113, ProteinAtlas, and TCGA, instead, we directly normalized gene expression values for different tissues. We did not normalize EMTAB-62 data, since its source values were already normalized with RMA.

Finally, we assigned to each pair gene–tissue a unique positive expression score, given by the average normalized expression value of the gene in that tissue, according to the different datasets. Expression scores in SPECTRA range from 3.566 to 17.366 for tissues and from 0.01 to 17.343 for tumors.



## 2.3. Data Schema

SPECTRA database is structured as a MySQL relational database with six tables: *Genes*, *Tissues*, *Tumors*, *Interactions*, *Expr\_normal*, and *Expr\_tumor*.

The *Genes* table contains the list of all expressed and interacting genes. Each entry is identified by the gene symbol and contains associated data, including a description string, aliases, and cross references to Entrez Gene (if available).

The *Tissues* and *Tumors* tables have the same structure. Tissues and tumors are associated to different classes, depending on the organism part they refer to. Each entry is identified by a unique number and contains a description and the corresponding class.

<sup>7</sup><http://www.ebi.ac.uk/arrayexpress>

<sup>8</sup><http://www.ncbi.nlm.nih.gov/geo>

<sup>9</sup><http://www.proteinatlas.org>

SPECTRA contains 26 distinct classes of tissues and 32 distinct classes of tumors.

The *Interactions* table lists all the PPIs integrated in SPECTRA. Interactions are identified by a couple of gene symbols and the edge weight for each integrated dataset (when available) is stored, together with the average interaction weight across dataset reporting that interaction and the dataset coverage.

*Expr\_normal* and *Expr\_tumor* contain all the gene expressions in normal and cancer tissues. The unique identifier of *Expr\_normal* is a couple gene–tissue, while entries in *Expr\_tumor* are uniquely identified by the couple gene–tumor. In both tables, the normalized expression value for each integrated dataset (where available) and the average expression score are included as associated data.

## 2.4. An Algorithm for Differential Local Alignment of TS-PPI Networks

TS-PPI networks are compared in SPECTRA for identifying patterns of differential gene expressions between multiple TS-PPI networks.

Our goal is to find conserved sub-regions in the TS-PPI networks, which maximize the difference of expression values of aligned genes. The problem is related to that of finding maximal-scoring connected subgraphs, which is NP-hard, even in a common simpler setting where the aligning TS-PPI networks have the same set of nodes and edges (e.g., TS-PPI networks built starting from different expression data and the same interaction datasets) (Ideker et al., 2002).

In the case of two TS-PPI networks with the same set of nodes and edges (representing for instance case and control expression data), heuristic (Ideker et al., 2002; Sohler et al., 2004; Cabusora et al., 2005; Rajagopalan and Agarwal, 2005; Guo et al., 2007) and exact (Dittrich et al., 2008) solutions have been proposed. However, as far as we are concerned, no solutions are known for the multiple case. Here, we propose an approximate solution to the multiple differential alignment problem based on a modified version of the GASOLINE algorithm (Micale et al., 2014a). For simplicity, we consider TS-PPI networks with no multiple edges between two nodes.

### 2.4.1. The GASOLINE Algorithm

GASOLINE (Micale et al., 2014a) is a greedy and stochastic algorithm for multiple local alignment of protein–protein interaction networks. Flowchart in Figure 4 provides a general description of GASOLINE.

Given  $N$  weighted PPI networks of different species, where edge weights are probabilities of interaction between proteins, local alignment aims at finding a set of connected subnetworks, one from each network, that are conserved in their sequence and interaction pattern. Such subnetworks could represent evolutionary conserved complexes or pathways across different organisms.

Such a problem is related to subgraph isomorphism, which is known to be NP-complete (Cook, 1971). GASOLINE proposes an approximate solution through a stochastic-greedy strategy consisting of two phases.

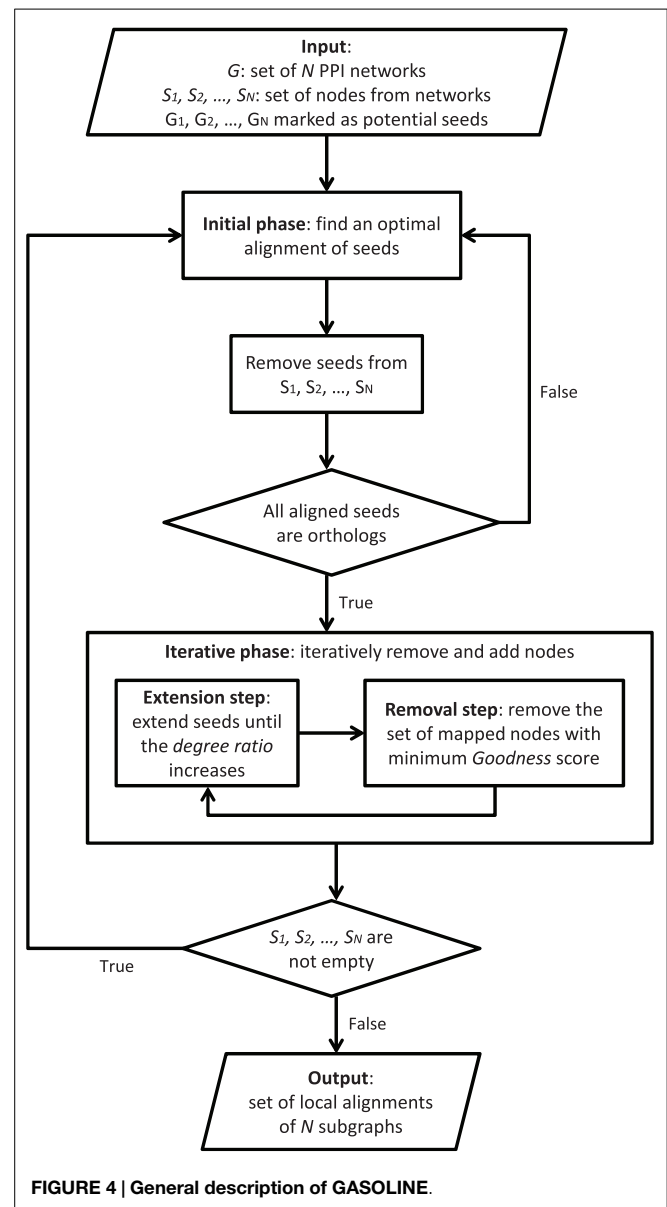


FIGURE 4 | General description of GASOLINE.

In the first step, called bootstrap phase, we look for orthologous proteins across the networks and build a set of seeds. The set of seeds initially consists of proteins, one from each network, and includes all the starting nodes of the suboptimal local network alignment.

The second step, called iterative phase, repeatedly adds (extension step) or removes (removal step) nodes in the network alignment, trying to maximize the final alignment score. Each extension step adds, in each network, a single node to the corresponding seed. During the extension step, the seeds grow up producing a set of subgraphs, one from each network. The extension process is regulated by a properly defined degree ratio measuring the average density of the aligned subgraphs with respect to their neighbors in the networks. The extension is performed until the degree ratio increases.

Each removal step replaces from the current alignment the set of proteins (one from each network) with minimum topology similarity score.

The bootstrap phase and each extension step are performed through Gibbs sampling (Geman and Geman, 1984). In both cases, the Gibbs sampling builds a chain, where each state represents a combination (i.e., alignment) of single proteins, one from each network. First, a random initial state is selected. Then, the sampling method iteratively performs a transition from a state to another, by replacing a randomly chosen protein of the current alignment with a protein of the same network, according to a properly defined transition probability distribution.

Due to its non-deterministic nature, different iterations of GASOLINE may produce different local alignments. The above steps are iterated to produce a set of local networks alignments, which are then ranked according to an Index of Structural Conservation (ISC) score. ISC score measures the percentage of conserved interactions in the final alignment. The higher is ISC, the better is the alignment.

GASOLINE implements preprocessing and post-processing steps. During preprocessing, the search space for potential seeds is reduced. This is obtained by marking only proteins having orthologs in all aligning networks and with a significant interaction degree in each network. All marked nodes in each network  $G_i (1 \leq i \leq N)$  are added to a set called  $S_i$ . These sets will be used in the initial phase and will be updated at each iteration. Finally, a post-processing filters the final set of local alignments returned by GASOLINE by removing highly overlapping complexes.

GASOLINE does not allow many-to-many mapping between aligned nodes. However, experimental results show that the algorithm can produce more reliable results than methods implementing many-to-many mapping. Moreover, GASOLINE is clearly faster than the state-of-art algorithms (Micale et al., 2014a).

### 2.4.2. The Adapted GASOLINE

We implemented a customized version of GASOLINE to compare two or more Tissue-Specific PPI (TS-PPI) networks for local differential alignment problem. GASOLINE algorithm was extended to deal with gene expressions as weights to the nodes.

Let  $A$  and  $B$  two genes and  $Expr(A)$  and  $Expr(B)$  their expression values, with  $Expr(A) \geq Expr(B)$ . In order to evaluate the expression difference between  $A$  and  $B$ , we compute the *log fold change*, defined as follows:

$$\text{LogFold}(A, B) = \log_2 \left( \frac{Expr(A)}{Expr(B)} \right) \quad (1)$$

Given  $N$  TS-PPI networks and a set of aligned genes  $G = \{G_1, G_2, \dots, G_N\}$ , one for each TS-PPI network,  $\text{MaxLogFold}$  is the maximum value of  $\text{LogFold}$  function among all pairs of genes in  $G$ :

$$\text{MaxLogFold}(G) = \max\{\text{LogFold}(G_i, G_j) \forall 1 \leq i < j \leq n\} \quad (2)$$

We applied the following changes to original GASOLINE algorithm:

- We included the  $\text{LogFold}$  function in the Gibbs sampling procedure of bootstrap and iterative phases, by multiplying it by the topology and homology scores in the computation of node similarities;

- The number of iterations of Gibbs sampling both in the bootstrap and in the extension phase is governed by a new parameter, is  $\alpha$ , which is a probability threshold related to  $N$ , the number of networks, according to the following formula:

$$k = \max \left\{ k' : \left( \frac{N-1}{N} \right)^{k'} > \alpha \right\} \quad (3)$$

where  $P = \left( \frac{N-1}{N} \right)^{k'}$  is the probability that a gene is never selected in  $k'$  consecutive iterations of Gibbs sampling. The idea is to stop Gibbs sampling when an alignment does not change for  $k$  consecutive iterations. The lower is  $\alpha$ , the higher is  $k$ , so the more precise and slower will be the sampling procedure:

- We introduced a new threshold, *MaxLogFoldThreshold*, for the value of  $\text{MaxLogFold}$  function, and we used it to tune the extension process in place of the degree ratio: in particular, we extend the current alignment until the average value of *MaxLogFold* between the sets of aligned nodes is above such a threshold;
- In the remove phase, the set of aligning nodes with minimum value of  $\text{MaxLogFold}$  is deleted from the current local alignment;
- Given a local alignment  $A = \{A_1, A_2, \dots, A_w\}$ , where  $w$  is the size of the alignment and  $A_1, \dots, A_w$  are the set of aligned genes, an average value of  $\text{MaxLogFold}(A_i)$  is computed together with the *ISC* score to evaluate the quality of the alignment.

## 3. Results

SPECTRA is a framework for retrieving and analyzing protein-protein interaction data specific for a given set of normal or cancer tissues. The underlying graph model in SPECTRA is the Tissues-Specific PPI network (or TS-PPI network), in which the genes of corresponding interacting proteins are both expressed in one or more tissues. The architecture of SPECTRA is composed by (i) the *searching tool*, which allows to build TS-PPIs; (ii) the *comparison tool* to look for shared differential expressions patterns between genes of two or more TS-PPI networks. Results can be graphically visualized by using Cytoscape.js or downloaded as text files.

### 3.1. SPECTRA Search Tool: Building TS-PPI Networks in SPECTRA

SPECTRA builds TS-PPI networks starting from a user-defined set of genes, tissues, expression data, and interaction data. **Figure 5** depicts the search interface of SPECTRA.

In the “Gene data” section (**Figure 5A**), the user can look for all genes expressed in a set of tissues or restrict the search to a specific list of genes. Genes can be provided with their official names or Aliases (e.g., Ensembl Gene, Entrez Gene, Affy).

In the “Expression data” section (**Figure 5B**), the user limits the search to a set of tissues/tumors and to a set of expression datasets or uploads a text file with custom expression data. Note that the two options are mutually exclusive, that is, all the settings concerning datasets and tissue/tumors will be ignored if the user

Home
Search
Compare
Documentation
Contacts

**A Gene data**

Search for all genes in SPECTRA ?
     
  Search for selected genes ?

---

**B Expression data**

Select parameters for expression data ?
     
  Upload expression data ?

Select one or more tissues
  Select one or more tumors

Class	Subclass		Input list
adipose tissue	adrenal cortex		adrenal gland
adrenal gland	adrenal gland	>>	
blood		<<	
bone			
bone marrow			
brain			
breast			

Gene expressions must be reported AT LEAST by: ?

- EMTAB62
- GDS181
- GDS596
- GDS1096
- GDS3113
- ProteinAtlas

Minimum expression value for genes (between 0 and 16):
 
-
+
?

---

**C Interaction data**

Interactions must be reported AT LEAST by: ?

- BioGrid
- DIP
- Havugimana
- HPRD
- IntAct
- MINT

Minimum average weight for gene interactions (between 0 and 1):
 
-
+
?

Minimum dataset coverage for gene interactions (0-100%):
 
-
+
?

Search

**FIGURE 5 | SPECTRA search tabbed panel.** Red boxes highlight the three sections: **(A)** “Gene data,” **(B)** “Expression data,” and **(C)** “Interaction data.” In this case, the parameters have been set to indicate that we want to retrieve all the interactions that are present at least in Havugimana and HPRD, involving

genes that are expressed in adrenal gland” tissue according at least to GDS3113 and ProteinAtlas. In this example, we neither restrict our search to a predefined set of genes nor provide a threshold for interaction weights, dataset coverage, and expression scores.

provides a custom text file. Available tissues and tumors in SPECTRA are listed in a table and can be easily included in the input query list with a double click in each entry. When no data are provided, all the tissues and tumors in SPECTRA are considered. Tissues and tumors are also mutually exclusive, meaning that a TS-PPI network built-in SPECTRA cannot contain interactions defined on both normal and tumor tissues. However, two TS-PPI networks defined upon a specific set of tissues and tumors, respectively, can be always compared for differential analysis with the adapted GASOLINE. The user can also select one or more datasets from which the expression have to be reported. When the expression is in other datasets it will be also given. When no dataset is selected, all expression data in SPECTRA are considered.

Finally, a further filter on genes can be applied by indicating a threshold for the minimum normalized value of gene expressions to be considered.

The “Interaction data” section (Figure 5C) contains the parameters for filtering interaction data. As above, the user can select one or more datasets where protein interactions have to be reported. If no interaction dataset is selected, all PPIs in SPECTRA are considered. A threshold can be provided to select interaction weights above a given value and a minimum dataset coverage.

When all input parameters have been specified, the user clicks on the “Search” button. At the end of the process, all the TS-PPIs found are listed in a result table (Figure 6). For each TS-PPI, we show the interacting genes, the tissues where they are expressed, the expression values of genes in tissues, the average interaction weights and dataset coverages of corresponding proteins. Results are ordered by dataset coverage and average interaction weight. Expression values and interaction weights are depicted with colored progress bar, where colors range from cyan (low values) to red (high values).

By selecting a specific TS-PPI in the result table, additional data about the interaction and the interacting genes are shown (Figures 6 and 7). A list of datasets reporting the interaction and the corresponding interaction weight is reported on the right

of the result table (Figure 6). Below the result table, two panels with details about the interacting genes are shown (Figure 7). For each gene, description and aliases are provided, together with the lists of tissues and tumors where the gene is expressed, according to the different expression datasets, ordered by expression score.

### 3.2. SPECTRA Comparison Part: Compare TS-PPI Subnetworks

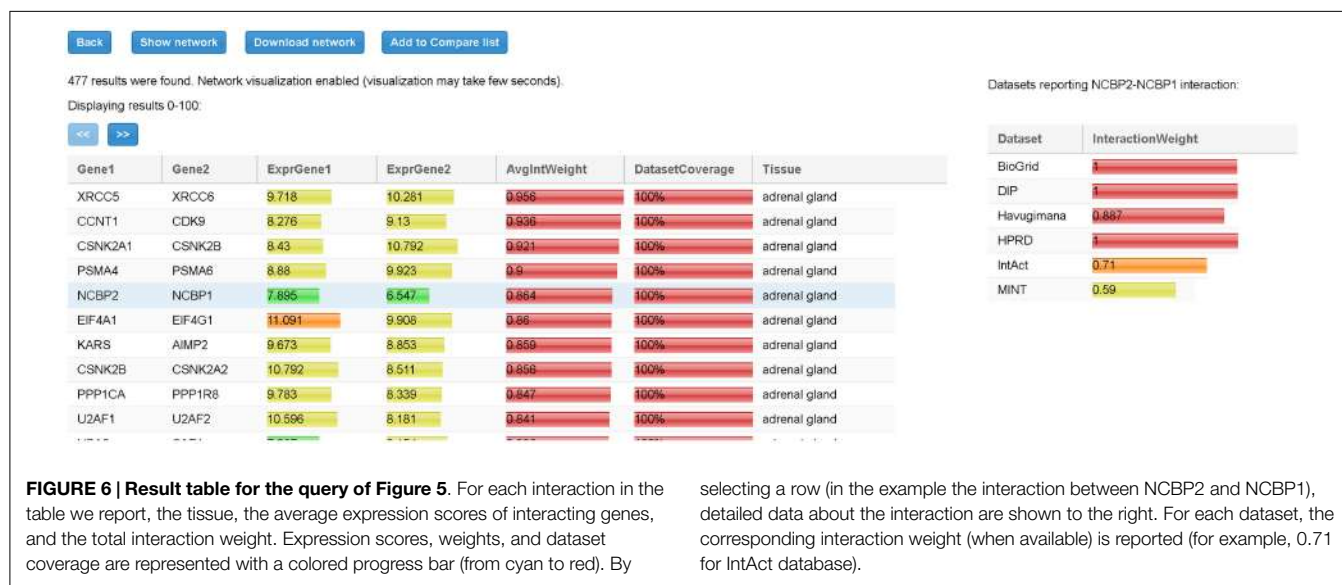
TS-PPI networks can be compared in SPECTRA for identifying patterns of differential gene expressions between multiple TS-PPI networks. The goal is to find conserved sub-regions in the TS-PPI networks, which maximize the difference of expression values of aligned genes.

Figure 8 shows the “Compare” tabbed panel in SPECTRA. Before running the adapted GASOLINE, the user has to upload at least two TS-PPI networks. For each network, the number of nodes and edges are reported. Networks can also be renamed by double clicking on the corresponding cell. Note that uploaded TS-PPI networks with multi-edges between nodes will be always treated as simple networks, where multi-edges are replaced by a single edge with weight equals to the average weight of multi-edges and label given by the concatenation of the multi-edge labels.

Once the networks have been uploaded, the user can click on “Run GASOLINE” button to set the input parameters for the adapted GASOLINE (Figure 8).

We briefly describe their meaning (default values are reported in brackets):

- “Sigma”: the minimum degree of candidate nodes for the initial alignment of seeds (1);
- “Alpha”: a value between 0 and 1, which regulates the number of iterations of Gibbs sampling in the bootstrap and extend phases (default 0.05);
- “Overlap threshold”: a maximum average overlap threshold between local alignments, which is used to remove highly



selecting a row (in the example the interaction between NCBP2 and NCBP1), detailed data about the interaction are shown to the right. For each dataset, the corresponding interaction weight (when available) is reported (for example, 0.71 for IntAct database).



**Details for gene NCBP2**

Gene symbol: NCBP2  
 Description: Nuclear cap binding protein subunit 2, 20kDa  
 Entrez id: [22916](#)

Aliases:  
 NCBP2, 32789\_at, P52298, 32790\_at, 201521\_s\_at, 201517\_at, ENSG00000114503, NP\_031388

Tissue	EMTAB62	GDS181	GDS596	GDS1096	GDS3113	ProteinAtlas	AvgScore ↓
mammary gland	Not reported	Not reported	Not reported	Not reported	10.472	Not reported	10.472
fetal thymus	Not reported	Not reported	Not reported	Not reported	9.9	Not reported	9.9
whole body	7.941	Not reported	Not reported	Not reported	11.292	Not reported	9.617
retina	Not reported	Not reported	Not reported	Not reported	9.466	Not reported	9.466
gallbladder	Not reported	Not reported	Not reported	Not reported	Not reported	9.342	9.342
fetal kidney	7.965	Not reported	Not reported	Not reported	9.298	9.837	9.033
duodenum	Not reported	Not reported	Not reported	Not reported	Not reported	8.954	8.954
colon	Not reported	Not reported	Not reported	7.351	9.665	9.499	8.838
stomach	Not reported	Not reported	Not reported	7.064	Not reported	9.187	8.126
esophagus	7.064	Not reported	Not reported	Not reported	Not reported	9.185	8.125

Tumor	EMTAB62	GDS181	GDS596	ProteinAtlas	TCGA	AvgScore ↓
uterine carcinosarcoma	Not reported	Not reported	Not reported	Not reported	11.538	11.538
lower grade glioma	Not reported	Not reported	Not reported	Not reported	11.326	11.326
rectum adenocarcinoma	Not reported	Not reported	Not reported	Not reported	11.231	11.231
glioblastoma multiforme	Not reported	Not reported	Not reported	Not reported	11.222	11.222
uterine corpus endometrioid carcinoma	Not reported	Not reported	Not reported	Not reported	11.207	11.207
ovarian serous cystadenocarcinoma	Not reported	Not reported	Not reported	11.409	10.855	11.132
bladder urothelial carcinoma	Not reported	Not reported	Not reported	Not reported	11.091	11.091
lymphoid neoplasm diffuse large B-cell lym...	Not reported	Not reported	Not reported	Not reported	10.999	10.999
monocytic lymphoma	Not reported	Not reported	Not reported	10.763	Not reported	10.763
thyroid carcinoma	Not reported	Not reported	Not reported	Not reported	10.735	10.735

**FIGURE 7 | The panel with detailed information of a gene.** When an interaction is selected from the result table (Figure 6), two panels with additional data, one for each interacting gene, are shown. This example refers to the detailed panel for gene NCBP2, which appears when the row table of Figure 6 is selected. In the detailed panel, the gene symbol, the

description, the corresponding ID in Entrez Gene database (when available), and aliases (including references in other databases) are reported. Finally, two tables with the set of tissues and tumors where the gene is expressed are shown. These are shown in decreasing order with respect to the average expression scores.

overlapping alignments. It takes values between 0 and 1 (default 0.5, which means 50%);

- “Refine iterations”: the number of iterations of the iterative phase, i.e., extend steps followed by a removal step (default 10);
- “Minimum alignment size”: the minimum size of a local alignment. Local alignments with size lower this minimum size are not reported in final list (default 3);
- “Minimum gene expression log fold change threshold”: value for *MaxLogFoldThreshold*, which controls the extension process (default 0.6).

According to the experiments reported in Micale et al. (2014a,b), we assigned to each parameter default values, which guarantee a good tradeoff between speed and accuracy of GASOLINE.

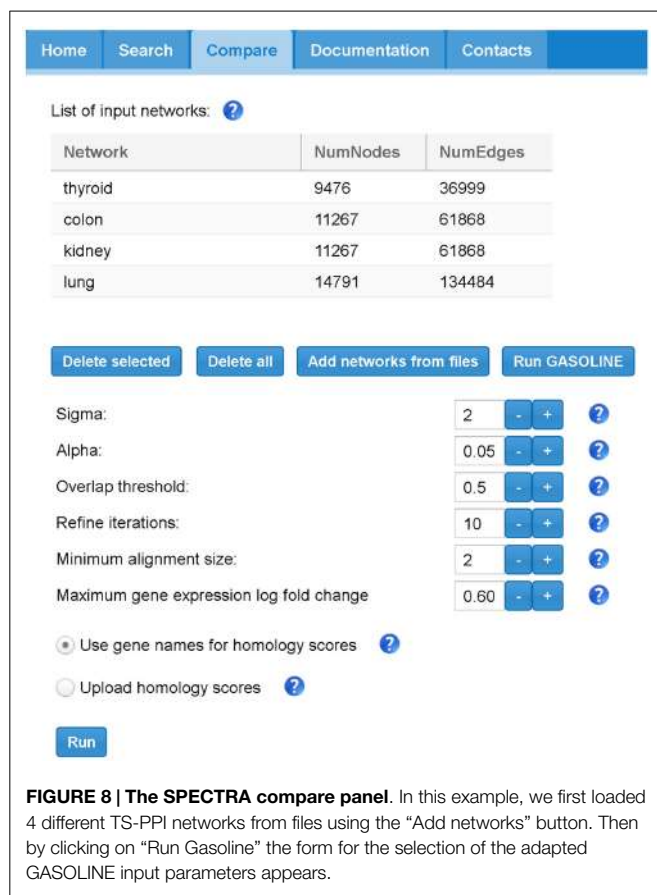
“Alpha” and “Refine iterations” parameters are strictly related to the stochastic nature of the algorithm. Lower values for “Alpha” and higher values for “Iter Refine” can be assigned to improve accuracy; however, the suggested default values are enough to yield good alignment results. Higher values of “Sigma” can be used to restrict the search to alignments starting from central genes in the input networks and to speedup the algorithm. Lower values of

“Overlap threshold” and higher values of “Minimum alignment size” allow to prune the final set of local alignments.

*MaxLogFoldThreshold* is the most critical parameter for GASOLINE. By increasing this threshold, the number and the size of final local alignments can be highly decreased and the algorithm could become much faster. Notice that there is no constant ideal value for *MaxLogFoldThreshold*, because it is highly dependent on the properties of input expression data. For log-transformed gene expression data, like the one which are present in SPECTRA database, low values of *MaxLogFoldThreshold* (0.2–1) are recommended.

Before running the adapted GASOLINE by clicking on “Run GASOLINE” button, the user has to indicate an homology scoring scheme between proteins of different aligning TS-PPI networks (Figure 8). The default naive solution is to use gene names for computing similarities: if two nodes have the same label, then they are considered homologs. Otherwise, user can upload an homology score file.

When the adapted GASOLINE ends, it gives as output a list of local alignments (if any, see Figure 9). For each alignment, the size, the average value of *MaxLogFold*, and the ISC score are reported.



The screenshot shows the SPECTRA compare panel. At the top, there are navigation tabs: Home, Search, Compare (selected), Documentation, and Contacts. Below the tabs, there is a section titled "List of input networks:" with a help icon. A table lists four networks: thyroid, colon, kidney, and lung, with their respective number of nodes and edges. Below the table are buttons for "Delete selected", "Delete all", "Add networks from files", and "Run GASOLINE". Underneath, there are input fields for "Sigma", "Alpha", "Overlap threshold", "Refine iterations", "Minimum alignment size", and "Maximum gene expression log fold change", each with a numeric input and +/- buttons. At the bottom, there are radio buttons for "Use gene names for homology scores" (selected) and "Upload homology scores", and a "Run" button.

Network	NumNodes	NumEdges
thyroid	9476	36999
colon	11267	61868
kidney	11267	61868
lung	14791	134484

**FIGURE 8 | The SPECTRA compare panel.** In this example, we first loaded 4 different TS-PPI networks from files using the “Add networks” button. Then by clicking on “Run Gasoline” the form for the selection of the adapted GASOLINE input parameters appears.

By selecting an alignment, its details are reported on the right (Figure 9). Alignment data include the set of nodes and edges attributes. The final mapping of aligned nodes is represented as a matrix in which columns contain nodes of the same network and rows represent the mapped genes.

### 3.3. Alternative Input for SPECTRA

User can upload text files in SPECTRA for building and comparing network. Expression data can be provided as text files in the “Expression data” section (Figure 5B) by selecting the “Upload expression data” option. Expression data files should have a matrix format with a row header representing tissues, a column header representing genes, and matrix elements indicating the gene expression value in a tissue.

There are two ways to provide input TS-PPI networks for comparison. User can either upload a text file or create the TS-PPI network with the SPECTRA searching tool and pass it to the comparison page. In the first case, network files are uploaded by clicking on “Add networks from files” in the “Compare” tabbed panel (Figure 8).

TS-PPI network files for comparison follows the same format of the result table in SPECTRA (Figure 6), except for the dataset coverage, with fields separated by tab characters. In the second case, one or more TS-PPI networks for specific tissues are passed to the comparison tool, by clicking on the “Add to compare list” button. The network is then added as input to the comparison list (Figure 8). By default, networks are added with the name

of the corresponding tissue, optionally followed by a progressive number whenever two or more TS-PPI networks for the same tissue are already present in the table. Anyway, networks can be later renamed by the user from the comparison table, before running GASOLINE.

In the homology file, needed to run the adapted GASOLINE algorithm, each row contains a pair of nodes of different TS-PPI networks, followed by a positive score value.

### 3.4. SPECTRA Output

TS-PPI networks (or subnetworks of them) are downloadable from the result panel, by clicking on “Download network” button (Figure 6). The user can filter the set of tissues upon which the TS-PPI network is defined. TS-PPI networks will be saved into different text files, one for each selected tissue or tumor. The file format is the same of the result table (Figure 6), with fields separated by tab characters.

The set of differential alignments returned by the adapted GASOLINE can be saved as .zip archive. The archive will contain a text file for each alignment. Each file contains the same alignment information reported in Figure 9.

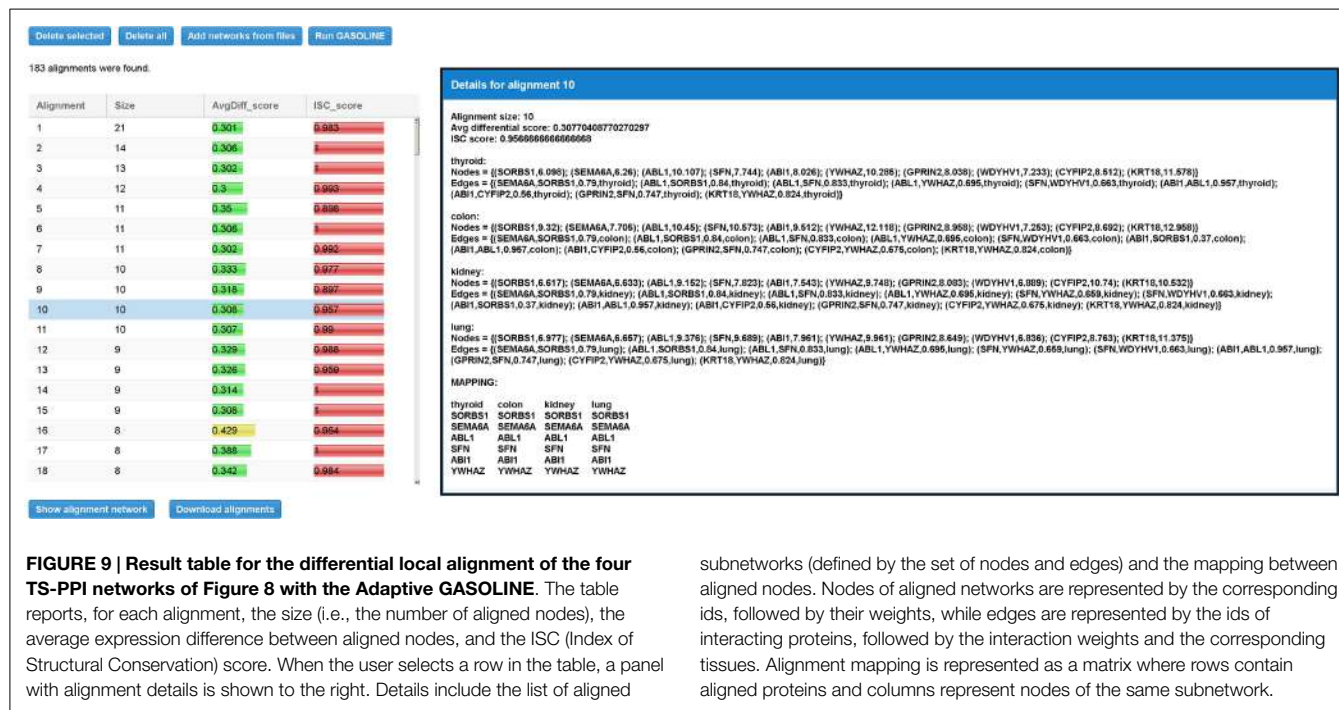
Results can also be visualized by using Cytoscape.js<sup>10</sup>, a JavaScript library for the analysis and visualization of networks. In the 2D visualization, TS-PPI networks can be navigated and zoomed. A TS-PPI network can be visualized from the result panel (Figure 6). Figure 10 shows two different examples of visualizations of TS-PPI networks within SPECTRA, with one (Figure 10A) or more (Figure 10B) tissues. Nodes and edges are differently colored according to the tissues of the TS-PPI network. Nodes are represented as pies with multiple colored slices. The diameter of the pie is proportional to the total expression score of the gene (considering all tissues of the TS-PPI network) and the size of each pie slice is proportional to the expression score of the gene in the corresponding tissue. Edge line widths are proportional to the interaction weights.

The alignments can be visualized in 2D (Figure 11), by selecting them from the list of local alignments and clicking on the “Show alignment network” button (Figure 9). Aligned nodes are colored according to the network they belong to and their sizes are proportional to the genes expressions. Edges are divided into two categories: intra-edges and inter-edges. Intra-edges connect nodes of the same subnetwork and are represented with solid lines with variable width, depending on the interaction weights. Inter-edges connect aligned nodes of different networks and are drawn with dashed black lines. In both cases, we used the Constraint-Based Layout (COLA) for network visualization.

### 3.5. Case Study

In this section, we show a practical usage of SPECTRA through a case study. We compared a set of four TS-PPI networks, built from genes expression data in normal and well differentiated, moderately differentiated, and poorly differentiated breast cancer tissues. The aim is to identify subnetworks of differentially expressed genes across the normal breast and the three different grades of breast tumors.

<sup>10</sup><http://js.cytoscape.org>



### 3.5.1. Data Preprocessing

We downloaded four breast cancer expression datasets for which information about the stage of breast tumors were available: GSE2361 (Ge et al., 2005), GSE2990 (Sotiriou et al., 2006), GSE4922 (Ivshina et al., 2006), and GSE7390 (Desmedt et al., 2007). We normalized data using RMA in R Bioconductor package (McCall et al., 2010).

The four expression datasets were then combined using COMBAT (Johnson and Li, 2007) into the R InSilicoDbMerging package. Finally, we grouped samples of the integrated dataset into four categories according to the grade of breast tumor (0 for normal tissue, 1 for well-differentiated tumor cells, 2 for moderately differentiated cells, and 3 for poorly differentiated cells). For each category, we computed the average expression value of each gene among samples. Results are stored into four different files (one per category).

### 3.5.2. Uploading Data in SPECTRA and Building Breast TS-PPI Networks

We loaded the expression files in the “Expression data” panel in SPECTRA (Figure 5B) and we selected BioGRID and IntAct as PPI datasets in the “Interaction data” panel (Figure 5C). SPECTRA builds four TS-PPI networks, each of them has 7,472 nodes and 29,765 edges. We added each network to the comparison list of GASOLINE (Figure 8), by clicking on *Add to compare list* from the Result panel (Figure 6).

### 3.5.3. Results of GASOLINE on TS-PPI Networks

Networks have been aligned by clicking on *Run GASOLINE* with the following parameters:

- Sigma = 1;
- Alpha = 0.05;
- Overlap threshold = 0.5;

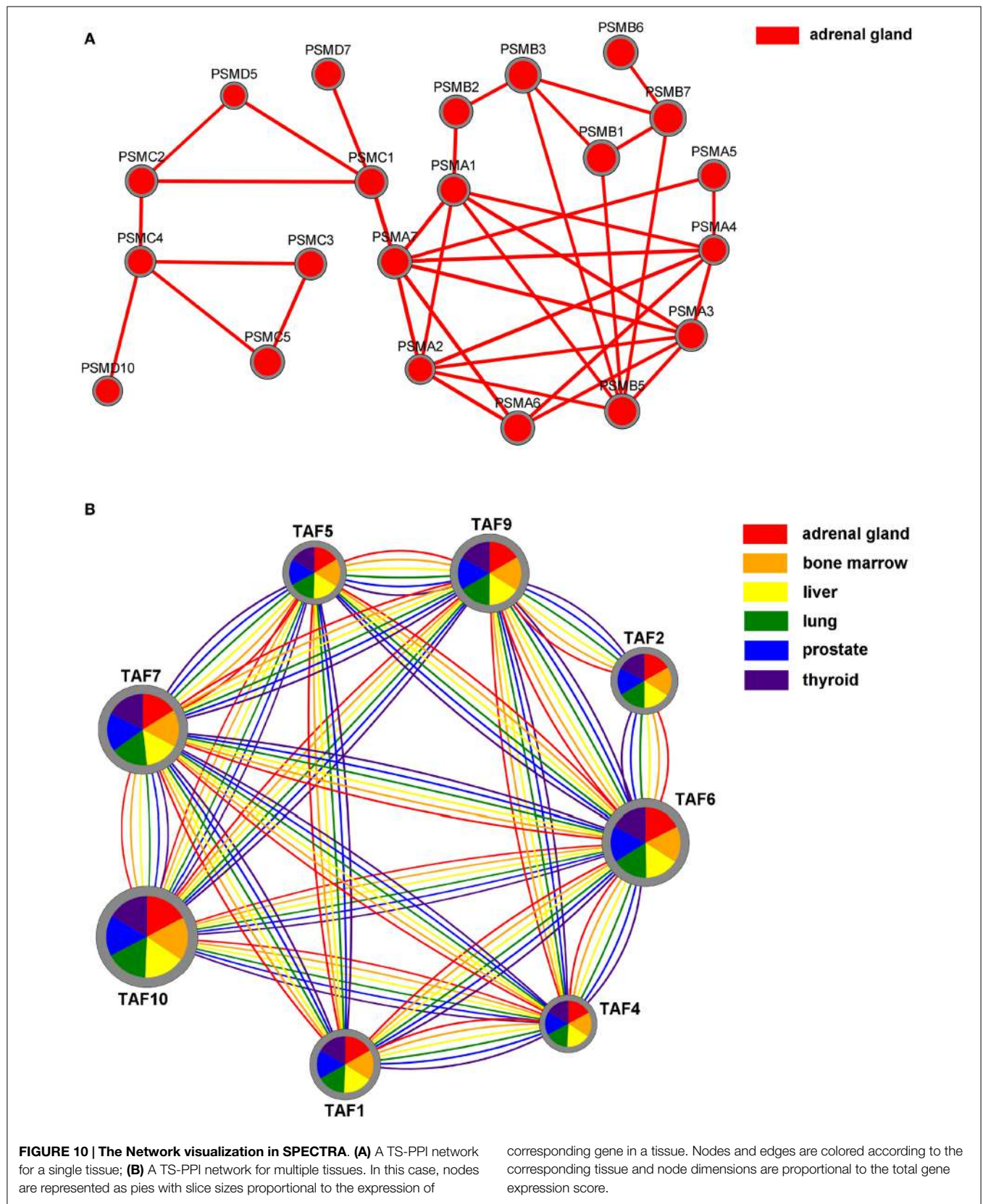
subnetworks (defined by the set of nodes and edges) and the mapping between aligned nodes. Nodes of aligned networks are represented by the corresponding ids, followed by their weights, while edges are represented by the ids of interacting proteins, followed by the interaction weights and the corresponding tissues. Alignment mapping is represented as a matrix where rows contain aligned proteins and columns represent nodes of the same subnetwork.

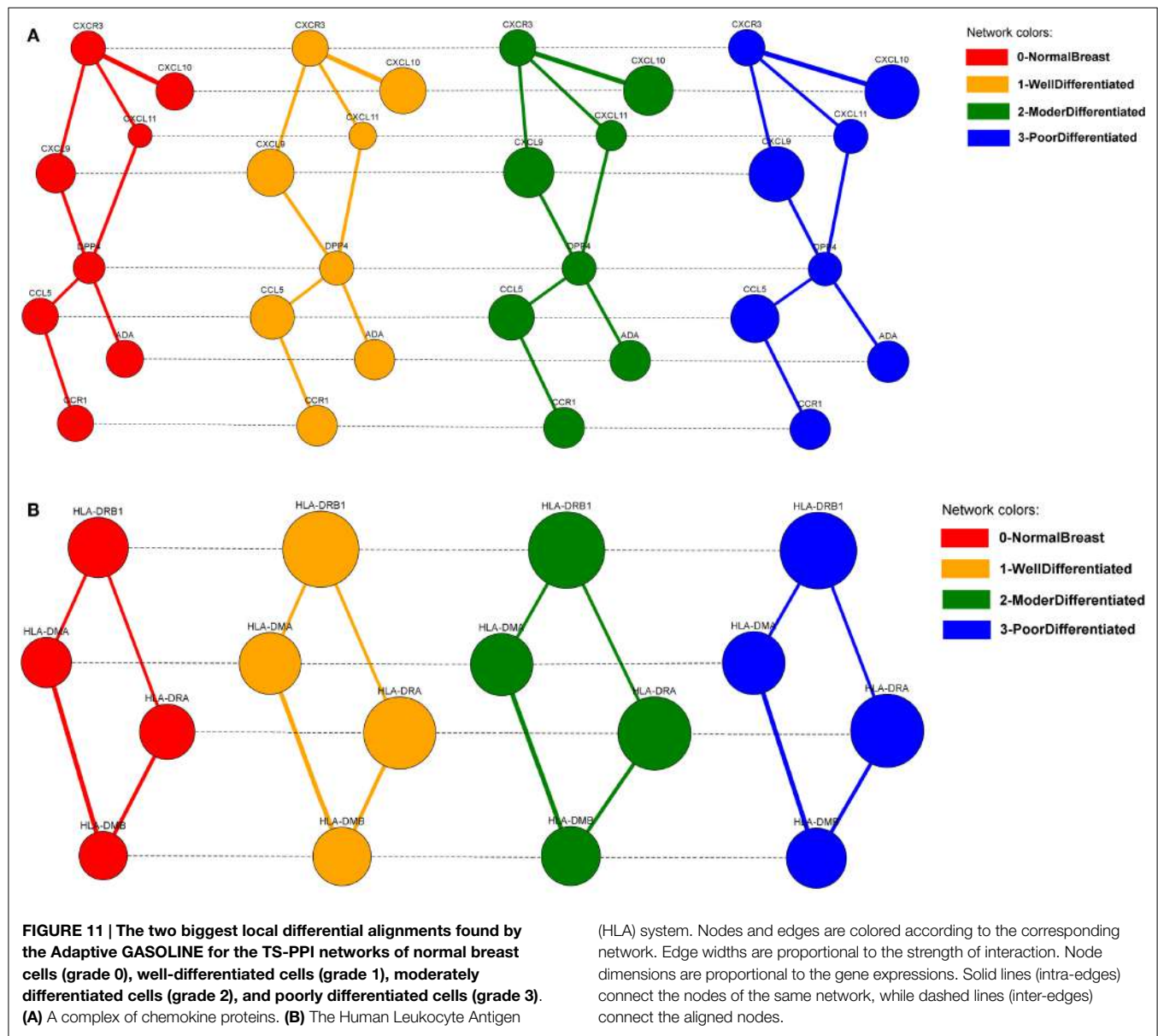
- Refine iterations = 10;
- Minimum complex size = 2;
- Maximum gene expression log fold change threshold = 0.3;
- Use gene names for homology score.

GASOLINE took 27 s to complete the task and returned 20 local alignments. In Figure 11, the two biggest alignments are shown using the SPECTRA visualization tool. Both alignments contain genes that are known to be involved in breast cancer at different stages.

More precisely, the major group of aligned nodes in Figure 11A is formed by the chemokine proteins (CXCL10, CXCL9, CXCL11, CCL5) and the chemokine receptors CXCR3 and CCR1, which are all highly overexpressed across the different grades of breast tumor. Chemokines can be responsible for leukocyte migration during processes of tissue development and formation, or can attract immune cells to a site of inflammation. Chemokines and chemokine receptors are known to have an important role on cancer metastasis, by facilitating tumor dissemination (Muller et al., 2000; Karnoub and Weinberg, 2007). DPP4 gene has a lower expression variation but ensures the communication between CCL5, CCR1, and the other chemokine proteins. This result agrees with the key role of DPP4 in signal transduction and tumor progression (Pro and Dang, 2004).

The alignment of Figure 11B is characterized by the Human Leukocyte Antigen (HLA) system (HLA-DRB1, HLA-DMB, HLA-DMA, HLA-DRA). The HLA system is composed by proteins on cell surface that are responsible for regulation of the immune system. HLA genes exhibit very high differential expression between normal and tumor cells and their overexpression in breast cancers is confirmed by several papers (Bartek et al., 1987; Kaneko et al., 2011; Da Silva et al., 2013).





The above case study highlights the capability of SPECTRA in helping researchers in producing novel biologically sound hypothesis and insight in the study of tissue-specific diseases.

## 4. Discussion

SPECTRA is a knowledge base to build and compare tissue or tumor-specific PPI networks. It overcomes the current PPI network analysis limitations mainly due (i) to the spreading of data in several databases with low overlap; (ii) to be unaware of the role of proteins in human tissues and diseases. SPECTRA integrates 13 databases of both protein–protein interactions and expressions data. Moreover, it provides an algorithm to compare built-in or custom tissue and tumor-specific PPI networks and identify subnetworks of differentially expressed genes. Finally, the results can easily be browsed through

a lightweight web application equipped with a 2D visualization network tool based on Cytoscape.js. Experiments performed on four TS-PPI networks built from gene expression data consisting of normal and breast cancer tissues show that the comparison algorithm can produce biologically significant results. SPECTRA database will go under update twice a year, with a semi-automatic curation of data downloaded from the online repositories. Future developments of SPECTRA aim to provide further network mining algorithms devoted to the analysis of expression data and the validation and annotation with ontologies of results.

## Acknowledgments

Publication of this article has been funded by PON 2007-2013 grant, SIGMA – PON01 00683 – CUP B61H11000380005.

## References

- Adamcsek, B., Palla, G., Farkas, I., Derenyi, I., and Vicsek, T. (2006). Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023. doi:10.1093/bioinformatics/btl039
- Alaimo, S., Pulvirenti, A., Giugno, R., and Ferro, A. (2013). Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 29, 2004–2008. doi:10.1093/bioinformatics/btt307
- Bader, G., and Hogue, C. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi:10.1186/1471-2105-4-2
- Banks, E., Nabieva, E., Peterson, R., and Singh, M. (2008). Netgrep: fast network schema searches in interactomes. *Genome Biol.* 9, R138. doi:10.1186/gb-2008-9-9-r138
- Barrett, T., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). Ncbi geo: archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193
- Barshir, R., Basha, O., Eluk, A., Smoly, I., Lan, A., and Yeger-Lotem, E. (2013). The tissuenet database of human tissue protein-protein interactions. *Nucleic Acids Res.* 41, D841–D844. doi:10.1093/nar/gks1198
- Barshir, R., Schwartz, O., Smoly, I., and Yeger-Lotem, E. (2014). Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput. Biol.* 10:e1003632. doi:10.1371/journal.pcbi.1003632
- Bartek, J., Petrek, M., Vojtesek, B., Bartkova, J., Kovarik, J., and Rejthar, A. (1987). Hla-dr antigens on differentiating human mammary gland epithelium and breast tumours. *Br. J. Cancer* 56, 727–733. doi:10.1038/bjc.1987.278
- Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi:10.1093/bioinformatics/19.2.185
- Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 5, 260. doi:10.1038/msb.2009.17
- Bruckner, S., Huffner, F., Karp, R., Shamir, R., and Sharan, R. (2010). Topology-free querying of protein interaction networks. *J. Comput. Biol.* 17, 237–252. doi:10.1089/cmb.2009.0170
- Cabusora, L., Sutton, E., Fulmer, A., and Forst, C. (2005). Differential network expression during drug and stress response. *Bioinformatics* 21, 2898–2905. doi:10.1093/bioinformatics/bti440
- Hamosh, A., Scott, A., Amberger, J., Bocchini, C., and McKusick, V. (2005). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi:10.1093/nar/gki033
- Chen, G., and Wang, J. (2012). Identifying functional modules in tissue specific protein interaction network. *IEEE Int. Conf. Bioinform. Biomed. Workshops* 2012, 581–586. doi:10.1109/BIBMW.2012.6470204
- Cook, S. (1971). “The complexity of theorem-proving procedures,” in *Proc. 3rd ACM Symposium on Theory of Computing* (New York: ACM), 151–158. doi:10.1145/800157.805047
- Csermerly, P., Korcsmaros, T., Kiss, H., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* 138, 333–408. doi:10.1016/j.pharmthera.2013.01.016
- Da Silva, G., Silva, T., Duarte, R., Neto, N., Carrara, H., Donadi, E. A., et al. (2013). Expression of the classical and nonclassical hla molecules in breast cancer. *Int. J. Breast Cancer* 2013, 250435. doi:10.1155/2013/250435
- Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., and Sahinalp, S. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* 27, i205–i213. doi:10.1093/bioinformatics/btr245
- Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin. Cancer Res.* 13, 3207–3214. doi:10.1158/1078-0432.CCR-06-2765
- Dezso, Z., Nikolski, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., et al. (2008). A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.* 6:49. doi:10.1186/1741-7007-6-49
- Dittrich, M., Klau, G., Rosenwald, A., Dandekar, T., and Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24, i223–i231. doi:10.1093/bioinformatics/btn161
- Emig, D., and Albrecht, M. (2011). Tissue-specific proteins and functional implications. *J. Proteome Res.* 10, 1893–1903. doi:10.1021/pr101132h
- Ferro, A., Giugno, R., Pigola, G., Pulvirenti, A., Skripin, D., Bader, G., et al. (2007). Netmatch: a cytoscape plugin for searching biological networks. *Bioinformatics* 23, 910–912. doi:10.1093/bioinformatics/btm032
- Flannick, J., Novak, A., Srinivasan, B., McAdams, H., and Batzoglou, S. (2006). Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res.* 16, 1169–1181. doi:10.1101/gr.5235706
- Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). String v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi:10.1093/nar/gks1094
- Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M., et al. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 86, 127–141. doi:10.1016/j.ygeno.2005.04.008
- Geman, S., and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741. doi:10.1109/TPAMI.1984.4767596
- Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D., and Shyr, Y. (2013). Large scale comparison of gene expression levels by microarrays and RNAseq using tcga data. *PLoS ONE* 8:e71462. doi:10.1371/journal.pone.0071462
- Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., et al. (2007). Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics* 23, 2121–2128. doi:10.1093/bioinformatics/btm294
- Havugimana, P., Hart, G., Nepusz, T., Yang, H., Turinsky, A., and Zhihua, A. (2012). A census of human soluble protein complexes. *Cell* 150, 1068–1081. doi:10.1016/j.cell.2012.08.011
- Huang, H., Wu, X., Pandey, R., Li, J., Zhao, G., Ibrahim, S., et al. (2012). C2maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships. *BMC Genomics* 13:S17. doi:10.1186/1471-2164-13-S6-S17
- Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18, S233–S240. doi:10.1093/bioinformatics/18.suppl\_1.S233
- Ivshina, A., Joshy, G., Senko, O., Mow, B., Putti, T. C., Smeds, J., et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* 66, 10292–10301. doi:10.1158/0008-5472.CAN-05-4414
- Johnson, W., and Li, C. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8, 118–127. doi:10.1093/biostatistics/kxj037
- Kalaev, M., Bafna, V., and Sharan, R. (2009). Fast and accurate alignment of multiple protein networks. *J. Comput. Biol.* 16, 989–999. doi:10.1089/cmb.2009.0136
- Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The consensuspathdb interaction database: 2013 update. *Nucleic Acids Res.* 41, D793–D800. doi:10.1093/nar/gks1055
- Kaneko, K., Ishigami, S., Kijima, Y., Funasako, Y., Hirata, M., Okumura, H., et al. (2011). Clinical implication of hla class i expression in breast cancer. *BMC Cancer* 11:454. doi:10.1186/1471-2407-11-454
- Karnoub, A., and Weinberg, R. (2007). Chemokine networks and breast cancer metastasis. *Breast Dis.* 26, 75–85.
- Lage, K., Hansen, N., Karlberg, E., Eklund, A., Roque, F., Donahoe, P. K., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20870–20875. doi:10.1073/pnas.0810772105
- Liao, C., Lu, K., Baym, M., Singh, R., and Berger, B. (2009). Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics* 25, 253–258. doi:10.1093/bioinformatics/btp203
- Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., et al. (2012). Mint, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40, D857–D861. doi:10.1093/nar/gkr930
- Lopes, T., Schaefer, M., Shoemaker, J., Matsuoka, Y., Fontaine, J. F., Neumann, G., et al. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* 27, 2414–2421. doi:10.1093/bioinformatics/btr414

- Lukk, M., Kapushesky, M., Nikkila, J., Parkinson, H., Goncalves, A., Huber, W., et al. (2010). A global map of human gene expression. *Nat. Biotechnol.* 28, 322–324. doi:10.1038/nbt0410-322
- Magger, O., Waldman, Y., Rupp, E., and Sharan, R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.* 8:e1002690. doi:10.1371/journal.pcbi.1002690
- McCall, M., Bolstad, B., and Irizarry, R. (2010). Frozen robust multiarray analysis (frma). *Biostatistics* 11, 242–253. doi:10.1093/biostatistics/kxp059
- Mete, M., Tang, F., Xu, X., and Nurcan, Y. (2008). A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics* 9:S19. doi:10.1186/1471-2105-9-S9-S19
- Micale, G., Pulvirenti, A., Giugno, R., and Ferro, R. (2014a). Gasoline: a greedy and stochastic algorithm for optimal local multiple alignment of interaction networks. *PLoS ONE* 9:e98750. doi:10.1371/journal.pone.0098750
- Micale, G., Pulvirenti, A., Giugno, R., and Ferro, R. (2014b). Proteins comparison through probabilistic optimal structure local alignment. *Front. Genet.* 5:302. doi:10.3389/fgene.2014.00302
- Muller, A., Homey, B., Soto, H., Ge, N., Catron, D., Buchanan, M. E., et al. (2000). Involvement of chemokine receptors in breast cancer metastasis. *Nature* 410, 50–56. doi:10.1038/35065016
- Nersisyan, L., Samsyan, R., and Arakelyan, A. (2014). Cykeggparser: tailoring kegg pathways to fit into systems biology analysis workflows. *F1000Res.* 3, 145. doi:10.12688/f1000research.4410.2
- Orchard, S., Ammari, M., Aranda, B., Brueza, L., Briganti, L., Broackes-Carter, F., et al. (2013). The mintact project-intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi:10.1093/nar/gkt1115
- Patil, A., Nakai, K., and Nakamura, H. (2011). Hitpredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res.* 39, D744–D749. doi:10.1093/nar/gkq897
- Peri, S., Navarro, J., Kristiansen, T., Amanchy, R., Surendranath, V., Muthusamy, B., et al. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32, D497–D501. doi:10.1093/nar/gkh070
- Pro, B., and Dang, N. (2004). Cd26/dipeptidyl peptidase iv and its role in cancer. *Histol. Histopathol.* 19, 1345–1351.
- Rajagopalan, D., and Agarwal, P. (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* 21, 788–793. doi:10.1093/bioinformatics/bti069
- Razick, S., Magklaras, G., and Donaldson, I. (2008). irefindex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9:405. doi:10.1186/1471-2105-9-405
- Rhrissorakrai, K., and Gunsalus, K. (2011). Mine: module identification in networks. *BMC Bioinformatics* 12:192. doi:10.1186/1471-2105-12-192
- Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., et al. (2013). Arrayexpress update-trends in database growth and links to data analysis tools. *Nucleic Acids Res.* 41, D987–D990. doi:10.1093/nar/gks1174
- Sahraeian, S. M. E., and Yoon, B. (2013). Smetana: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PLoS ONE* 8:e67995. doi:10.1371/journal.pone.0067995
- Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303
- Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics* 20, 1517–1521. doi:10.1093/bioinformatics/bth112
- Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* 98, 262–272. doi:10.1093/jnci/djj052
- Souiai, O., Becker, E., Prieto, C., Benkahla, A., De Las Rivas, J., and Brun, C. (2011). Functional integrative levels in the human interactome recapitulate organ organization. *PLoS ONE* 6:e22051. doi:10.1371/journal.pone.0022051
- Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539. doi:10.1093/nar/gkj109
- Su, A., Cooke, M., Ching, K., Hakak, Y., Walker, J., Wiltshire, T., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4465–4470. doi:10.1073/pnas.012025199
- Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 6062–6067. doi:10.1073/pnas.0400782101
- Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* 28, 1248–1250. doi:10.1038/nbt1210-1248
- Xenarios, I., Salwinski, L., Duan, X., Higney, P., Kim, S., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291. doi:10.1093/nar/28.1.289
- Xiao, X., Moreno-Moral, A., Rotival, M., Bottolo, L., and Petretto, E. (2014). Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. *PLoS Genet.* 10:e1004006. doi:10.1371/journal.pgen.1004006
- Zhao, J., Lee, S., Huss, M., and Holme, P. (2012). The network organization of cancer-associated protein complexes in human tissues. *Sci. Rep.* 3, 1583. doi:10.1038/srep01583

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Micale, Ferro, Pulvirenti and Giugno. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.