# Bioinformatics Tools for Proteomics Data Interpretation

Karla Grisel Calderón-González, Jesús Hernández-Monge, María Esther Herrera-Aguirre, and Juan Pedro Luna-Arias

#### Abstract

Biological systems function via intricate cellular processes and networks in which RNAs, metabolites, proteins and other cellular compounds have a precise role and are exquisitely regulated (Kumar and Mann, FEBS Lett 583(11):1703–1712, 2009). The development of high-throughput technologies, such as the Next Generation DNA Sequencing (NGS) and DNA microarrays for sequencing genomes or metagenomes, have triggered a dramatic increase in the last few years in the amount of information stored in the GenBank and UniProt Knowledgebase (UniProtKB). GenBank release 210, reported in October 2015, contains 202,237,081,559 nucleotides corresponding to 188,372,017 sequences, whilst there are only 1,222,635,267,498 nucleotides corresponding to 309,198,943 sequences from Whole Genome Shotgun (WGS) projects. In the case of UniProKB/Swiss-Prot, release 2015\_12 (December 9, 2015) contains 196,219,159 amino acids that correspond to 550,116 entries. Meanwhile, UniProtKB/TrEMBL (release 2015\_12 of December 9 2015) contains 1,838,851,8871 amino acids corresponding to 555,270,679 entries. Proteomics has also improved our knowledge of proteins that are being expressed in cells at a certain time of the cell cycle. It has also allowed the identification of molecules forming part of multiprotein complexes and an increasing number of posttranslational modifications (PTMs) that are present in proteins, as well as the variants of proteins expressed.

Departamento de Biología Celular, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav-IPN), Av. Instituto Politécnico Nacional 2508, Col. San Pedro Zacatenco, Gustavo A. Madero, C.P. 07360 Ciudad de México, Mexico

J. Hernández-Monge

H. Mirzaei and M. Carrasco (eds.), *Modern Proteomics – Sample Preparation, Analysis and Practical Applications*, Advances in Experimental Medicine and Biology 919, DOI 10.1007/978-3-319-41448-5\_16

16

K.G. Calderón-González • M.E. Herrera-Aguirre J.P. Luna-Arias (⊠)

e-mail: jpluna@cell.cinvestav.mx; jpluna@cinvestav.mx; jpluna2003@gmail.com

Instituto de Física, Universidad Autónoma de San Luis Potosí, Av. Manuel Nava 6, Zona Universitaria, C.P. 78290 San Luis Potosí, S.L.P., Mexico

<sup>©</sup> Springer International Publishing Switzerland 2016

#### Keywords

Proteomics data interpretation • Interactome mapping • Gene Ontology • STRING • MINT • IntAct • HPRD • BioGRID • PIPs • MPIDB • TAIR • PANTHER • DAVID • KEGG • IPA

Biological systems function via intricate cellular processes and networks in which RNAs, proteins other cellular metabolites, and compounds have a precise role and are exquisitely regulated [1]. The development of highthroughput technologies, such as the Next Generation DNA Sequencing (NGS) and DNA for microarrays sequencing genomes or metagenomes, have triggered a dramatic increase in the last few years in the amount of information GenBank and stored in the UniProt Knowledgebase (UniProtKB). GenBank release 210, reported in October 2015, contains 202,237,081,559 nucleotides corresponding to 188,372,017 sequences, whilst there are only 1,222,635,267,498 nucleotides corresponding to 309,198,943 sequences from Whole Genome Shotgun (WGS) projects. In the case of UniProKB/Swiss-Prot, release 2015 12 (December 9, 2015) contains 196,219,159 amino acids that correspond to 550,116 entries. Meanwhile, UniProtKB/TrEMBL (release 2015\_12 of December 9 2015) contains 1,838,851,8871 amino acids corresponding to 555,270,679 entries. Proteomics has also improved our knowledge of proteins that are being expressed in cells at a certain time of the cell cycle. It has also allowed the identification of molecules forming part of multiprotein complexes and an increasing number of posttranslational modifications (PTMs) that are present in proteins, as well as the variants of proteins expressed.

Considering that human cells contain between 20,000 and 30,000 protein-encoding genes and possibility that there could be approximately four alternative splice variants for each gene [2], the total number of proteins that could be expressed at a certain time would range between 80,000 and

120,000. Moreover, guessing four PTMs in each protein, then, the total number of proteins in a cell would range between 320,000 and 480,000. However, when we consider the more than 400 different PTMs that have been found [3] the number of proteins in a cell would easily grow to more than one million.

Proteins do not function alone; they usually carry their function by interacting with one or more partners. The main goal of the proteinprotein interaction map is to catalogue interactions and to define the interactome. These interactions are currently determined using a vast array of technologies, including yeast two hybrid systems, tag-fusion proteins for the identification of interacting proteins, co-immunoprecipitation, chemical crosslinking, phage display, FRET (Fluorescence Resonance Energy Transfer), SPR (Surface Plasmon Resonance), tandem affinity purification, protein microarrays, protein domains, etc. Many of these techniques, if not all, use mass spectrometry and non-redundant gene and protein databases as the main tools for the identification of peptides and proteins. Many of the cellular protein-protein interaction networks have been catalogued and a number of interactome databases have been established. There are several protein-protein interaction databases freely available via World Wide Web that can be used to determine the putative functions of a protein based on its direct or indirect interactions. Protein-protein interaction maps in these databases are, in general, based on the information published, mostly in PubMed. In this section, we describe some of the most important databases available, including STRING, MINT, IntAct, HPRD, BioGRID, PIPs, MPIDB and TAIR. Furthermore, additional tools such as Gene Ontology, PANTHER, DAVID, KEGG, and IPA, among others, have been developed to facilitate data mapping into these databases. We are certain that these tools will be useful in understanding the intricate interactions and functions of proteins in cells.

## 16.1 Gene Ontology

Many proteins are conserved through evolution and consequently share the same functions. However, the systems of nomenclature for genes and proteins stay divergent despite repeated evaluation of gene similarities by experts [4]. In order to tackle this challenge, the Gene Ontology (GO) consortium was created. The aim of the GO project is to provide a structured vocabulary to define specific biological domains that describe gene products in different organisms [5]. GO project began in 1998 as a collaborative effort between three organism databases: FlyBase (Drosophila), the Mouse Genome Informatics (MIG) project and the Saccharomyces Genome Database (SGD). The GO Consortium has been continuously growing due to the deposition of several animal, microbial and plant genome databases [6], as well as the recent addition of ontology areas, such as cell cycle and cilia-related terms, as well as multicellular organism processes [7]. By using these ontologies, it is possible to graph structures that comprise cellular components, molecular functions, biological and the processes, relationships between them in a speciesindependent manner [7]. In other words, GO is divided in two modules, the ontologies, called GO ontology, which includes defined terms and their relationships, and the GO annotations, which covers gene products and defined terms [8]. The GO annotation is generated either by a curator or automatically through predictive methods (95 % by this method).

The gene ontology relationships are developed like a tree, depicting a hierarchy from more general terms to more specific ones. Terms are linked by three possible relationships: "is\_a", "part\_of", and "positively regulates/negatively regulates". The "is\_a" is a simple relationship between a class and a subclass. The "part\_of" relationship is more complex than the former. C is part of D means that whenever C is present, it always belongs to D; for instance, an organelle (C) is always part of a cell (D), but not all cells have the same organelles. In the GO website (http://geneontology.org), a variety of browsers provide visualization and query capabilities for GO. For example, the AMIGO browser provides a web interface for searching and displaying ontologies, term definitions and associated annotated gene products for diverse organism databases [6]. The GO Online SQL (Structured Query Language) Environment (GOOSE) for AmiGO 2, allows users to freely enter SQL queries in the GO database. On the other hand, the PANTHER Classification System, that is further described next, provides enrichment analysis tools for GO.

## 16.2 PANTHER

PANTHER (Protein ANalysis Through Evolutionary Relationships) is a classification system that combines ontology, gene function, pathways and statistical tools. This classification system can analyze sequencing, gene expression, and proteomics data [9]. PANTHER is a large database of gene families developed as a resource for family and subfamily classification of proteins [10]. PANTHER has two main components: PANTHER library (PANTHER/ LIB) and PANTHER index (PANTHER/X). PANTHER library is a collection of protein families and subfamilies represented as phylogenetic trees assembled using Hidden Markov statistical models (HMMs) and a multiple sequence alignment algorithm (MSA) (Fig. 16.1a) [9–12]. PANTHER index is a set of ontological abbreviated terms that describe the function of proteins in biological processes or molecular functions [10–12]. In addition,



Fig. 16.1 PANTHER data overview. PANTHER has two main modules: (a) PANTHER Library which is a collection of families and subfamilies of proteins. This library is constructed from a selection of sequences built into clusters. These clusters are then used to generate multiple sequence alignments (MSA), phylogenetic trees, and statistical HMMs. (b) PANTHER Pathways are built using literature databases related to pathway

PANTHER has a Pathway module, in which the represented as pathways are a diagram generated with CellDesigner software (Fig. 16.1b) [13]. This module uses a defined vocabulary to describe pathways and their components, including pathway class and components, molecular class, reaction class, reaction relationships, cell type, and cellular components [14, 15]. PANTHER pathways are related to protein sequences in the PANTHER/ LIB and, therefore, are also connected with

components or a particular molecular class. Then, pathways are drawn and curated by expert curators using the CellDesigner software. Pathways are built based on molecular class or pathway component, reaction class and relationships, and cell type or cellular components. The pathway component is a link between various PANTHER modules

families/subfamilies and HMM analysis (Fig. 16.1) [9, 10, 12]. Pathways are created and annotated by expert curators, according to evidence found in the literature. Moreover, pathways can be curated with the Pathway curation software (http://curation.pantherdb. org/) [14, 15]. Some of the pathways included in the PANTHER database are Cell cycle, DNA replication, General transcription regulation, Glycolysis, Tricarboxylic acid cycle, among others (http://www.pantherdb.org/pathway/ pathwayList.jsp). The PANTHER database contains the following information:

- 1. Genes (104 genomes; 1,424,953 total genes; 1,026,421 genes in PANTHER families with phylogenetic trees, MSA and HMMs)
- 2. Families (11,928 families and 83,190 subfamilies)
- 3. Pathways (177 pathways, 3092 pathway components, 2447 sequences related to pathways, and 2447 references captured for the pathways)
- Ontologies (550 terms in PANTHER GO slim, 257 terms corresponding to biological process, 70 cellular components, and

223 molecular functions; 243 terms of protein class; 41,603 terms used in GO database annotations, including 9942 molecular functions, 27,852 biological processes, and 3809 cellular component terms (http://www.pantherdb.org/data).

The main window in PANTHER is composed of two main toolbars. The first one contains different links to individual topics (Fig. 16.2, items 1-5), as well as an option for registration, login and contact (Fig. 16.2, items 6-8). The second toolbar contains different options for data analysis, including gene list analysis, browse, sequence search, cSNP scoring, and keyword



**Fig. 16.2** PANTHER Classification System website. The main window in PANTHER contains two main toolbars. The first toolbar on top has links to different options inclduing: (1) PANTHER data, (2) PANTHER tools, (3) workspace, (4) downloads and (5) help/tutorial, and a section for (6) registration, (7) login, and (8) contact. The second toolbar, right under the first one, is for

data analysis: (9) Gene list analysis, (10) browse, (11) sequence search, (12) cSNP scoping, and (13) keyword search. PANTHER also includes: (14) Quick keyword search, (15) whole genome function views, (16) genome statistics, (17) publications, and (18) recent publications describing PANTHER [16]

search (Fig. 16.2, items 9–13). In addition, PAN-THER has a panel for keyword search and quick links (Fig. 16.2, items 14-18) [16]. In the analysis of list of genes or proteins, different functional classification views can be obtained, including gene list, bar or pie charts. Also, genes or proteins can be statistically analyzed through an enrichment test or a statistical overrepresentation test [17]. The PANTHER Ontology Browser also called PANTHER Prowler, browses and retrieves results (e.g. molecular functions, biological process, cellular component, protein class, pathway, and species) for input data related to ontology terms, such as genes and families [11, 17]. The PANTHER HMM sequence-scoring (sequence search) tool, can be used to search and compare protein sequences with the HMMs of PANTHER library.

The top hit HMM can be observed in the results page, which also contains a statistical value for significance [17]. The Evolutionary Analysis of Coding SNPS (cSNP scoring) tool estimates the probability of a specific amino-acid change [17]. The keyword search tool can be used to obtain a variety of information, such as genes, families, pathways, and ontology terms for the protein of interest. However, we will focus on the generation of graphs for proteins classified in different categories.

## 16.3 PANTHER Gene List Analysis

To perform a gene list analysis using the PAN-THER website (http://pantherdb.org), go to the toolbar gene list analysis (Fig. 16.3) and enter the



**Fig. 16.3** Procedure to perform gene list analysis in PANTHER. The *red* section denotes the three primordial steps: (1) Enter the IDs of proteins to be analyzed, (2) select the organism, and (3) select the type of analysis to be performed

IDs of the genes or proteins in your list (Ensembl, Ensembl\_PRO, Ensembl\_TRS, Gene ID, Gene symbol, GI, HGNC, IPI, UniGene, UniProtKB ID) into the window, separating IDs by a space or comma. IDs can also be uploaded as a txt file. Then select the list type for query data (i.e. ID List, Previously exported gene list, Workspace list or PANTHER Generic Mapping File) and the organism of interest for analysis. In our example, we selected "ID list" and "Homo sapiens". Afterward, choose the type of analysis you like to perform. For example, we selected the "functional classification" viewed as a pie chart. Finally, click on the submit key (Fig. 16.3). In the results webpage, genes can be classified according to Molecular Function, Biological Process, Cellular Component, Protein Class, and Pathway (Fig. 16.4a). The chart obtained for a certain process can change for other processes. In addition, pie charts can be changed to bar charts and vice versa (Fig. 16.4b). The list of genes obtained in each ontological classification can be exported as a txt file. Classification categories may also contain different subcategories. When the cursor is located over a category in a chart, a message containing the following information will be displayed: Category name and its corresponding identifier, number of genes included from your list, the corresponding percentage of gene hits against the total number of identified genes, and the percentage of gene hits against the total number function hits (Fig. 16.4a). When a subcategory is selected, the corresponding gene list will be displayed (Fig. 16.5). As an example, we classified a list of overexpressed proteins in common between Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cells lines, which were recently described by Calderón-González et al. [18]. These proteins were categorized into Molecular functions and Cellular components (Fig. 16.4). In the first category, the most representative processes were: Binding and Catalytic activity with 25 and 21 genes, respectively (Figs. 16.4a and 16.5a). For Cellular component classification, categories with the higher number of genes were: Cell part (14 genes) and Macromolecular complex (10 genes) (Fig. 16.4b).

#### 16.4 DAVID

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) was developed in 2003 to address the emerging challenges posed by the post-genomic era [19]. DAVID, as well as other tools for the analysis of large gene lists, is based on the principle of gene enrichment that are functionally related to an altered gene/protein (generated by high throughput technologies). These enriched genes might potentially cooperate within a determined group and/or biological process [20]. DAVID is composed of the DAVID knowledgebase and five annotation tools:

- 1. DAVID Functional Annotation
- 2. DAVID Gene Functional Classification
- 3. DAVID Gene ID Conversion
- 4. DAVID Gene Name Viewer
- 5. NIAID Pathogen Annotation Browser.

The DAVID Knowledgebase is constructed around the "DAVID Gene Concept", which include tens of millions of gene/protein identifiers from several major public databases. This data concentration eliminates annotation redundancy among different resources and allows the organization of gene identifiers into more than 40 functional classification categories, e.g. Ontology (more than 40 million records), Protein-protein interactions (more than four millions), Disease gene associations (9000), Pathways (above 50,000), Functional categories (more than 6.9 millions), etc. [21].

DAVID Gene Functional Classification: This tool is useful for the exploration of large lists of genes into more feasible modules ordered according to their functional relationships. These functionally organized modules are very useful in processing large amounts of information, switching from a gene centric analysis to a module-centric analysis [21].

DAVID Functional Annotation Tool Suite: The Functional Annotation Tool Suite displays three ways for combining results: Functional Annotation Clustering, Functional Annotation Chart and Functional Annotation Table. The



**Fig. 16.4** Functional classification of proteins up-regulated in both Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cells lines.

The proteins were classified into (**a**) Biological Processes and (**b**) Cellular Components. Figure shows the change of pie chart to a bar graphic as well

Functional Annotation Clustering tool allows the user to group genes depending on the degree of their functional association. It is performed with a novel algorithm that measures relationships among annotation terms. This process is useful to eliminate the redundant relationships that exist in many-genes-to-many-terms cases (i.e. when one gene is associated with many different redundant terms and one term is associated with many genes) [21]. Additional features of this



**Fig. 16.5** Classification of Biological Processes for proteins up-regulated in both Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cells lines (**a**) Biological processes pie chart displaying

different categories of processes, e.g. Metabolic Processes. (b) List of genes involved in the selected Metabolic Processes

clustering tool is the ability to rank the importance of annotation groups with an enrichment score (EASE scores) that uses the geometric mean of all the enrichment p-values of each annotation term in the group; the annotation clustering tool provides a link to a 2-D viewer for related gene-term relationships, allowing a fast way to focus on the genes that have common annotation terms [22]. On the other hand, The Functional Annotation Chart tool can be used to get the typical gene-GO term enrichment analysis (similar to other tools) to identify the most relevant (overrepresented) biological terms associated with a given gene list. However, DAVID offers extended annotation coverage in comparison to other enrichment analysis tools. The enhanced annotation coverage includes not only the GO terms but more than 40 annotation categories, such as protein-protein interactions, protein functional domains, disease associations, bio-pathways, sequence features, gene tissue expression, etc. This tool is helpful to identify enriched annotation terms associated with the gene list of interest in a linear tabular text format. Similar to the Annotation Clustering Tool, the Functional Annotation Chart also provides links to further explore the list of interacting proteins, link gene-disease associations and visualize genes on BioCarta and KEGG pathway maps [21]. Finally, the Functional Annotation Table tool is a query engine for DAVID Knowledgebase without statistical probes. It delivers annotation information in a table format for every gene from the users' gene list. This is a particularly useful tool when users want to have a closer look of some specific interesting genes and explore its annotation information.

DAVID's Gene ID Conversion tool allows conversion of user's input gene or gene product identifiers from any type to another in a more comprehensive and high throughput manner with a uniquely enhanced ID-ID mapping database leveraging heterogeneous annotations [23].

DAVID's Gene Name Viewer is another tool useful to quickly attach meaning to a list of gene IDs, translating them into their corresponding gene names. Thus, before proceeding to an in-depth analysis, researchers can quickly have an overview of gene names to gain insight into their biological system and have *a priori* general idea of interesting processes that might be involved.

DAVID's NIAID Pathogen Browser: The National Institute of Allergy and Infectious Diseases (NIAID) has defined three categories of priority pathogens, A, B and C. These pathogens are important for biodefense purposes and have become attractive study subjects because of the increasing research funding available to study them. The DAVID NIAID Pathogen Browser is provided as a support tool for researchers that would like to explore the biology of the priority pathogens types. For example, one may choose the word "anthrax" and type the key word "toxin", the result is a list of genes from Bacillus anthracis that matches to the typed key word. This tool may assist researchers in understanding the biology of a priority pathogen if the gene list retrieved from the DAVID NIAID Pathogen Browser is further analyzed by one of DAVID's Bioinformatics Resources [21].

Analysis of gene lists: To carry out an optimal gene list analysis, the list should; (1) have enough number of genes/proteins ranging from hundreds to thousands (e.g. 100–2000), (2) only include genes with statistical significance that show a notable up or down regulation, (3) show reproducibility between experimental replicas [22].

DAVID bioinformatics resources website is organized in two main toolbars (Fig. 16.6). There are different links, like Start Analysis, Shortcut to DAVID Tools, Technical Center, among others on top. On the left side, there are other shortcuts to DAVID Tools that also offers a brief explanation for each tool. Recently added DAVID NIAID Pathogen Annotation Browser tool can be found on the top menu in shortcut to DAVID Tools.

It is straightforward to upload a gene list for DAVID bioinformatics analysis (Fig. 16.7a). Firstly, go to https://david.ncifcrf.gov/gene2 gene.jsp and select Start analysis. On the left side choose upload in the list manager, then: (1) Copy/paste the gene lists to be analyzed into box A; a text file or a gene IDs list can also be



**Fig. 16.6** DAVID Bioinformatic Resources Website. This website has two main toolbars. The toolbar on the top has links to: (1) Start Analysis, (2) Shortcut to DAVID Tools, (3) Technical Center, (4) Downloads and APIs, (5) Terms of Service, (6) Why David, and (7) About

Us. And the toolbar on the left side (8) has links to Tools that offer a brief explanation for each of DAVID's tool. Additionally, in (2) we can find the recently added tool NIAID Pathogen Annotation Browser (9)

uploaded in box B, (2) Choose the corresponding gene identifier type for your input gene IDs; alternatively use the ID conversion tool to seek (or convert) the correct gene identifier, (3) Select the type of list you are submitting, either gene list or gene background. The general guideline is to set up a pool of genes as population background. This usually includes all the genes that could be possibly detected (e.g. all the probes included in a particular DNA microarray). Since most of the studies are done in a genome-wide scale, there is no need to set a background (default background is the entire genome), (4) Submit the List. The analysis different suites are displayed (Fig. 16.7b) that will be applied to the submitted gene list shown on the left (highlighted in the

Gene List Manager) (Fig. 16.7b). By clicking Start Analysis, users can go back at any time to upload another gene list or to access any analytical tool suite of interest.

In this section, a couple of examples are presented to showcase a few of the tools from David's toolbox that are most widely used using gene lists corresponding to proteins down regulated in both Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cell lines studied by Calderón-González et al. [18]. Selecting Functional Annotation Tool (Fig. 16.7b), results in Annotation Summary Results, which displays the number and percentage of genes (from the submitted gene list) involved in different GO categories **Fig. 16.7** Uploading data into David's gene list manager. (**a**) On the left side; (*1*) Upload a gene list, (*2*) Choose the corresponding gene identifier, (*3*) Select the type of list, either gene list or gene background, (*4*) Submit the gene list. (**b**) Once the user has submitted the gene list, the Analysis Wizard shows the shortcuts for the different DAVID Analysis tools



(Fig. 16.8). In each category, users can click on Chart to obtain an individual chart report for the selected category. Users can choose a number of categories for further analysis in the Combined Annotation Tools (Fig. 16.8). A table divided in several annotation clusters will be obtained by clicking on Annotation Clustering Tool. Every annotation cluster is formed by a group of terms from functionally related genes. Taken all together, the chance to identify a biological significance increases (Fig. 16.9). The degree of similarity between annotations is measured by Kappa statistics. This tool also provides a link to generate a 2D-view map that allows a fast way to associate genes that have common annotation terms.

From this very specific gene list, we observed an enriched group of genes involved in mitochondrial function. Noteworthy, the high correlation of this result in comparison with other tools previously explored. Since the submitted gene list corresponds to down-regulated genes in a proteomic approach, this result suggests that MCF7, T47D and MDA-MB231 breast cancer



**Fig. 16.8** Functional Annotation Tool Suite. (1) Gene List Manager showing the list that is being analyzed. (2) Annotation Summary results displaying different categories: (3) the number and (4) percentage of genes involved. (5) Clicking on this box will generate a chart

report of functional categories. (6) The user can choose the number of categories to be considered for further analysis in the Combined Annotation Tools (7) by checking the check boxes next to each category

cell lines have an impaired mitochondrial function in comparison to the MCF10A control cell line.

For instance, NADH-coenzyme Q reductase, 3,2 trans-enoyl-Coenzyme A isomerase, cytochrome c oxidase, and malate dehydrogenase are some of the encoding genes that had a high EASE SCORE and are involved in the mitochondrial inner membrane function.

#### 16.5 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource designed for understanding and interpreting biological systems using high-throughput data [24–26]. KEGG is composed of 17 databases organized into four categories:

4	3 The overall scores of ea	encrichmer ich term me	nt score for mbers. The	the group based on higher the more enr	the EASE riched.
Clustering	options and stringency		(2)		
	Functional Annotation Cluster	ing Tarma i	n the ennet	ation alustan	$\sim$
	Current Gene List: DWN_REG LIST Current Background: Homo saplens 49 DAVED (Ds. 8) Options Classification Stringency Medium • Reun using options Cleater Sublist 24 Cluster(s) Caste Sublist 24 Cluster(s) Coste Sublist GotEB4_CC_IAT mitachenidia matrix GotEB4_CC_IAT mitachenidia SP_PR_XENVGROS and UP_S40_IFA1U8E biologister/Juberne UP_S40_IFA1U8E biologister/Juberne	Rela Rela	7 ted term set	ation cluster Genes involved in arch Cont PWar Beams 22 128-16 156-14 24 36-14 46-12 14 26-13 138-13 17 184-1 148-3 17 208-1 148-3 17 208-1 148-3 10 208-1 148-3 10 208-1 10 208-1 10 208-1 10 208-1 10 208-1 10 208-10	5 individual term
	UP_SEQ_FEATURE nucleotide phosphate-bi UP_SEQ_FEATURE binding site.NAD INTERPRO MAC(F)-binding domain	nding region:NAD ALI RT RT		5 3.9E-5 4.4E-3 4 2.7E-4 2.0E-2 5 8.3E-4 1.1E-1	6
Annotatio	n clusters of terms		E	ASE Score (a modifie	ed Fisher Exact
which sha	re functional similarities		p-	value. The smaller, t	he more enriched
Gene Repor	t st: UP_REG LEST1	8	20 View Do	are unlike special. The specialization of special sectors are able of	$\frown$
Current Backgro 49 DAVID IDs 31 record(s)	und: Homo sapiens	<u> </u>		States indiag press 2 secting Mills. (eg. States (eg. 9 (southful)) "Hillings (	9
Q14454	SSON CERE NAME 2.4-diaming Cole reductions 1, mitterhoudering	Helated Genes		And a set of the set o	
F01937	3. Androsciash-metals. Astrofesantases Cold consistent of contribution (A barrakes (F. contr)	85		100 school School genetics N research (March 100 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1	
096000	www.seconder.compare.compare.com	20		Intel artist print 1 - Advention Alternation	
075-489	1909 debutcomare (ubicasare) ferSicrites 3, 300a (3408) contrars 0 (relation)	50		Anticipant or relation to discussion range & design (to constance	
406702	5100 céclum bindine protein At	89		Annual Contract of Contract of State Contraction, Mar. 1999.	
000154	acri-Cal, thiostheraps 2	10		In tractions coupling here, any herein	
F11310	ac-A Constante A Edited resonante, C-4 to C-12 straight chain	89		increase and the second second second	
075390	chier suchase	2	and the second second	and a starting of the starting	an i mismar
075208	psepume 09 hamolog (5. ownisae) oderhoune s. executio	BG AG	-	and a second second second a	
Q12959	data, lerat horselin 1 (Drosachin)	80	1.0	And appendix the proof theory is perpendicul service, and a con-	
P42125 800505	dodecencel-Coencisme A delta Isonerase (3.2 crans-encel-Coencisme A somerase) distante encelore del transporte a 2. entrethendrad (assuttate encelorementerase 2)	85		Addressed Assessed & Maria Internet (21.2 years and Assessed & the waterships assessed & Senting 10, until	
dex2/ds	pictoffions 5-transferrers impas 1	89		And the second s	
P19307 016836	Insultance 1 hydroxyachi-Geerzyme A dehydrosenase	50			
P26440	iscusient Conzume A defudrosenzae	85	11111111111		
Q\$4764	Challer, shell arcatelo	15			
			34116 41 11		

1. Systems information: KEGG PATHWAY 2. Genomic information: KEGG ORTHOLOGY (pathway maps), KEGG BRITE (functional (orthology (KO) groups), KEGG GENOME hierarchies and table files) and KEGG MOD- (complete genomes), KEGG GENES (gene An examinated complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional Annotations) KEGG GENES (gene file (framway, straumula complex, functional complex, functional annotation about genes) (framway, straumula complex, functional genomes) (framway, straumula complex, functional genomes) (framway, straumula complex, functional for (framway, straumula complex, functional complex, functional for (framway, straumula complex, functi

obtained from different databases, such as RefSeq (prokaryotes, eukaryotes, plasmids and viruses), GenBank (prokaryotes), and PubMed (addendum: collection of manually created protein sequences entry)

- Chemical information, also called KEGG LIGAND: KEGG COMPOUND (metabolites and other small molecules), KEGG GLYCAN (glycans), KEGG REACTION (biochemical reactions), KEGG RPAIR (reactant pairs), KEGG RCLASS (reaction class), and KEGG ENZYME (enzyme nomenclature)
- Health information commonly called KEGG MEDICUS: KEGG DISEASE (human diseases), KEGG DRUG (drugs), KEGG DGROUP (drug groups), KEGG ENVIRON (crude drugs and health related substances), JAPIC (drug labels in Japan) and DailyMed (links to drug labels in USA) [26].

The annotation system in KEGG is based on the correlation between functional information and orthologous groups (KEGG Orthology or KO) through the assignment of KO identifiers (K number). This information is stored in the KO database and is independent of the KEGG GENE database that contains completely sequenced genomes [26]. The KO system is essential for connecting the genomic information with systemic functional information resulting in the conversion of genes to K numbers, leading to KEGG reconstruction an automatic of PATHWAYS and other networks [26, 27]. Currently, KEGG has more than 4000 complete genomes annotated with the KO system [26].

KEGG has several analysis tools:

- KEGG Mapper which is the interface used for KEGG Mapping. This is composed of KEGG BRITE, MODULE, and PATHWAY mapping tools, which map genes, proteins, small molecules, etc. (also called objects) into all brite functional hierarchies, modules and pathways maps, respectively [28]
- KEGG Atlas is a graphical interface to navigate the global integrated maps in KEGG. Maps available are Metabolism (Biosynthesis

of amino acids, Biosynthesis of secondary metabolites, Carbon metabolism, Degradation of aromatic compounds, Fatty acid metabolism, Microbial metabolism in diverse environments, and 2-Oxocarboxylic acid metabolism) and Cancer pathway [29]

- BlastKOALA: KOALA is defined as KEGG Orthology And Links Annotation. BlastKOALA is used for the annotation of completely sequenced genomes. This tool utilizes the Pangenomes database
- GhostKOALA: this tool is designed by the metagenome annotation and it uses the Pangenomes and Viruses databases [26, 27], (5) BLAST/FASTA performs searches of similar sequences
- 5. SIMCOMP searches for similar chemical structures

Pathway Maps Analysis To map proteins of interest into Pathways, go to the KEGG website (http://www.genome.jp/kegg/) and on the Dataoriented entry points, click on the KEGG PATH-WAY key (Fig. 16.10). In the Pathway Mapping menu, select the mapping tool of interest: Search Pathway, Search&Color Pathway or Color Pathway. As an example, the up and down-regulated proteins found common between Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cells lines from Calderón-González et al. were analyzed with Search&Color Pathway tool [18]. the Up-regulated proteins were colored in red, whilst down-regulated polypeptides were presented in green (Fig. 16.11). To perform this analysis, an organism must be selected first by clicking on the org key, after which a new window is displayed to find the three to four KEGG organism code. Type the desired organism in the window and then click on select. In this example, H. sapiens has the hsa code. The next step is to introduce IDs in UniProtKB format, followed by the word red or green as mentioned before. Other compatible ID formats are KEGG-Identifiers, NCBI-GeneID and NCBI-ProteinID. Alternatively, a file containing IDs can be uploaded. To perform



**Fig. 16.10** KEGG website. This image shows the different links provided in KEGG's website, including KEGG Home, KEGG Database, KEGG Objects, KEGG Software, among others. The website also provides several

the search, the following options were selected; (1) to include aliases and (2) to display objects not found in the search (Fig. 16.12a). The result window shows a list of pathways where proteins were mapped, as well as a list of protein IDs that were not found (Fig. 16.12a). A list of proteins found in each pathway, including their UniProtKB IDs and KEGG *H. sapiens* database codes is also displayed (Fig. 16.12b). Clicking a

tools for the data analysis including KEGG Mapper, KEGG Atlas, BlastKOALA, Ghost KOALA, BLAST/ FASTA, SIMCOMP. KEGG Pathway modules are highlighted in a *red box* 

particular UniProtKB ID will display the information for the selected ID (Fig. 16.13a). On the other hand, if the code of the *H. sapiens* organism in KEGG is selected, a new window containing KEGG information about that protein, including Gene name, Disease, KEGG Orthology, Structure, Motifs in the protein, and Pathways, among other information will be displayed (Fig. 16.13b). Finally, when a certain pathway is selected, an



**Fig. 16.11** KEGG pathway mapping tool. This image shows the general procedure for mapping proteins in Search & Color Pathway module. The format of IDs as

well as the organism need to be selected. Protein accession numbers are followed with the word *red* or *green* to highlight up- or downregulated proteins, respectively

image is generated where up- or down-regulated proteins are highlighted in red or green respectively (Fig. 16.14). In the case of the breast cancer cell line, most quantified proteins mapped to metabolic processes, with 22 polypeptides [5 up-regulated ( $\uparrow$ ) and 17 down-regulated ( $\downarrow$ )]:  $\downarrow$ 3HIDH,  $\uparrow$  SAHH3,  $\downarrow$  IVD (Amino acid metabolism),  $\uparrow$  CMBL (Hydrolase),  $\downarrow$  CISY (Carbon metabolism, 2-Oxocarboxylic acid metabolism, biosynthesis of amino acids, carbohydrate metabolism),  $\downarrow$  AL1A3 (Carbohydrate metabolism, amino acid metabolism, metabolism of other amino acids, xenobiotics biodegradation and metabolism, chemical carcinogenesis),



**Fig. 16.12** Search & Color Pathway result. (a) A list of proteins that were not found are shown at the *top*. The list of different pathways is also displayed with the

↓ AATM (Carbon metabolism, 2-Oxocarboxylic acid metabolism, biosynthesis of amino acids, amino acid metabolism, fat digestion and absorption), ↓ HCDH (Fatty acid metabolism, carbohydrate metabolism, lipid metabolism, amino acid metabolism), ↓ HXK1 (Carbon metabolism, carbohydrate metabolism, biosynthesis of other secondary metabolites, HIF-1 signaling pathway, insulin signaling pathway, carbohydrate digestion and absorption, central carbon metabolism

number of proteins involved. (b) Two examples of proteins involved in RNA transport and DNA replication processes

in cancer, endocrine and metabolic diseases), ↓ ACADM (Carbon metabolism, fatty acid metabolism, carbohydrate metabolism, lipid metabolism, amino acid metabolism, metabolism of other amino acids, PPAR signaling pathway), ↑ METK2 (Biosynthesis of amino acids, amino acid metabolism), ↓ MDHM (Carbon metabolism, carbohydrate metabolism, amino acid metabolism), ↓ NDUBA, ↓ NDUS3 (Energy metabolism, neurodegenerative diseases,



**Fig. 16.13** Additional information for proteins in KEGG Database. The proteins displayed in each pathway have a link to additional information: (a) UniProtKB website and (b) KEGG database

endocrine and metabolic diseases),  $\downarrow$  DHB12 (Fatty acid metabolism, lipid metabolism),  $\downarrow$  ODPB (Carbon metabolism, carbohydrate metabolism, HIF-1 signaling pathway, glucagon signaling pathway, central carbon metabolism in cancer),  $\uparrow$  PGAM1 (Carbon metabolism, biosynthesis of amino acids, carbohydrate metabolism,

amino acid metabolism, glucagon signaling pathway, central carbon metabolism in cancer), ↓ CYC (Energy metabolism, cellular processes, pathways in cancer, neurodegenerative diseases, cardiovascular diseases, endocrine and metabolic diseases, infectious diseases), ↓ RPN1 (Glycan biosynthesis and metabolism, folding, sorting



**Fig. 16.14** Proteins mapped into KEGG PATHWAYS. Polypeptides found up- or down-regulated in both Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cell lines were submitted to KEGG

mapping. Some of the processes found to be affected are, (a) RNA transport process, and (b) DNA replication process. Up-regulated proteins are colored in *red* and down-regulated proteins are in *green* 

and degradation),  $\downarrow$  NLTP (Lipid metabolism, cellular processes, PPAR signaling pathway),  $\downarrow$  SPEE (Amino acid metabolism, metabolism of other amino acids),  $\uparrow$  PYR1(Nucleotide metabolism, amino acid metabolism). Others

mapped pathways were: RNA transport with 5 proteins  $\uparrow$  IMB1,  $\uparrow$  RAN,  $\uparrow$  EIF3B,  $\uparrow$  EIF3F,  $\uparrow$  EIF3I) (Fig. 16.14a) and DNA replication with 4 polypeptides involved ( $\uparrow$ MCM3,  $\uparrow$  MCM4,  $\uparrow$  MCM6,  $\uparrow$  PCNA) (Fig. 16.14b).

## 16.6 Ingenuity Pathway Analysis (IPA)

Ingenuity Pathway Analysis (IPA, QIAGENs Redwood City, www.qiagen.com/ingenuity) is a software application platform developed for analysis, understanding, integration and interpretation of biological data [30]. Ingenuity can analyze data acquired using platforms such as microarrays, proteomics, metabolomics, etc. IPA uses the QIAGEN's Ingenuity Knowledge Base in which contents extracted from articles, biomedical literature, reviews, internally curated knowledge, and other sources are structured into Ontology terms. The information in this platform are categorized into several knowledgebases:

- Ingenuity expert information, including Ingenuity expert findings and Ingenuity expert assist findings
- Ingenuity supported third party information including MicroRNA-mRNA interactions (miRecords, TarBase, TargetScan)

- Protein-Protein Interactions including BIND, cognia, DIP, Interactome studies, MINT, and MIPS
- 4. Additional sources: An open access database of genome-wide association results, BIOGRID, Breast cancer information core (BIC), Catalogue of somatic mutations in cancer (COSMIC), Chemical Carcinogenesis Research Information System (CCRIS), ClinicalTrials.gov, ClinVar, DrugBank, GO, GVK Biosciences, Hazardous Substances Data Bank (HSDB), HumanCyc, IntAct, miRBase, Mouse Genome Database (MGD), Obesity Gene Map Database, and Online Mendelian Inheritance in Man (OMIM).

The principal components of IPA suite are

- 1. Core Analyze
- 2. IPA-Tox
- 3. IPA-Biomarker
- 4. IPA-Metabolomics (Fig. 16.15)



**Fig. 16.15** The main page of Ingenuity Pathway Analysis suit. All functions are listed via in two main tabs, Learning IPA, and shortcuts. The shortcut tab contains

the dataset- and pathway options, as well as different analysis options, including Core, IPA-Tox, IPA-Biomarker and IPA-Metabolomics

Core Analyze consists of classified data sets mapped into biological processes, networks and pathways. IPA-Tox module includes data classified in the context of toxicological processes. In this tool the toxicity and safety of compounds is evaluated. IPA-Tox keeps track of the biological processes that are related to compound toxicity at various biochemical and molecular levels. IPA-Biomarker tool is used to identify and prioritize potential biomarker candidates. The selection of these putative biomarkers is based on their biological characteristics. Finally, the fourth application IPA-Metabolomics, is able to analyze metabolomics data, which are then contextualized into biological insights (metabolism and cell physiology).

IPA supports several types of identifiers including Affymetrix, Affymetrix SNP ID, Agilent, CAS registry number, CodeLink, dbSNP, Ensembl, GenBank, Entrez gene, Gene Symbol-mouse, Gene Symbol-rat and Gene symbol—Human (Hugo/HGNC), GenPept, GI number, Human Metabolome Database (HMDB), Illumina, Ingenuity, International Protein Index, KEGG, Life Technologies Biosystems), miRBase (Applied (mature), miRBase (stemloop), PubChem CID, RefSeq, UCSC hg18 and 19, UniGene and UniProtKB/ Swiss-Prot accession number. The confidence reported by IPA are either experimentally determined or theoretically predicted. Some tissues and cell lines covered by IPA include tissue and primary cells from nervous and other organ systems and cell lines from breast cancer, cervical, central nervous system (CNS), colon, hepakidney, toma, immune, leukemia, lung, lymphoma, macrophage, melanoma, myeloma, neuroblastoma, osteosarcoma, ovarian, pancreprostate and teratocarcinoma model atic, systems. Mutations covered include functional effect, inheritance mode, translation impact, unclassified mutation, zygosity and wild type.

IPA analysis core protocol: To use IPA, a license needs to be purchased but one can use a trial version for a limited period of time. To perform an analysis in IPA, first an analysis dataset need to be created (Fig. 16.16). To create an analysis dataset, go to Annotate datasets

			Genes and Cher	nicals Diseases and	d Functions	Pathways and Tox Lists
set Uplo	ad - IPA Analysis 181215.x	dsx				
elect File	e Format:	Flexible For	rmat	<ul> <li>More In</li> </ul>	fo	
Contains	Column Header:	• Yes O	No			
alast Ida	atifies Town	UniDent/Co	in Dest Assessing		the idea till a b	me found in the detroit
elect lde	entimer type:	UNIPROU SW	iss-Prot Accession	- ppecity	the identifier ty	ype round in the dataset.
Array plat	tform used for experiments	Not specif	ied/applicable	▼ Select n	elevant array pl	atform as a reference set for data analysis.
DIT OBSE	ERVATION NAMES INFER	OBSERVATION	More Info			
	D	OBSERVATION	More Info Observation 2	Observation 3 👻	Ignore	-
	D  Cobservation names D Cobservation names D Cobservation Exp R	oeservation	More Info Observation 2 Exp Ratio	Observation 3   Exp Ratio	Ignore	
	D Obser D III4:113	OBSERVATION	More Info Observation 2 Exp Ratio	Observation 3 Exp Ratio 115:113	Ignore	-
1 Pro 2 Q9	D Obser Exp R oteins 114:113 UG63 1.48788	observation 1	5 More Info Observation 2 ▼ Exp Ratio ▼ 117:113 1.71181595325469999	Observation 3 Exp Ratio 115:113 1.75942599773407	Ignore	•
1 Pro 2 Q9 3 Q9	ERVATION NAMES PHER D Obser Exp R oteins 114:113 30UG63 1.48788 38WF3 1.97668	OBSERVATION vation 1 • atio • 797855377 30205917401	S         More Info           Observation 2         ▼           Exp Ratio         ▼           117:113         ▼           1.63702499866486         1.63702499866486	Observation 3 Exp Ratio 115:113 1.75942599773407 1.3063050306499101	Ignore	•
1 Pro 2 Q9 3 Q9 4 Q9	ERVATION NAMES INTER D Obser Exp R oteins 114:113 JUG63 1.48788 JBWF3 1.97668 JP147 2.38947	OBSERVATION vation 1 • atio • 797855377 30205917401 29614257799	5 More Info Cbservation 2 ▼ Exp Ratio ▼ 117:113 1.7118159532546999 1.63702499866486 2.0250289440154998	Observation 3  Exp Ratio I15:113 1.75942599773407 1.3063050508499101 1.395080086898799	Ignore International International	-
1 Pro 2 Q9 3 Q9 4 Q9 5 Q9	D         Obser           D         Closer           114:113         114:113           JUG63         1.487:83           18WF3         1.97668           199747         2.38947           06HN2         2.86102	OBSERVATION vation 1  vation 1 vation	More Info           Observation 2           Exp Ratio           117:113           1.7118159532546999           1.63702499806486           2.0250289440154998           2.82182097434998	Observation 3  Exp Ratio  115:113 1.75942599773407 1.3063050508499101 1.395038006698799 1.7519400119781501	Ignore 	
1 Pro 2 Q9 3 Q9 4 Q9 5 Q9 6 Q9	D         Obser           D         Cobser           Exp R.         Cobser           BWP3         114:113           JUG63         1.48788           BWP3         1.97668           J9747         2.38947           S6H02         2.86102           S6D06         3.34422	OBSERVATION vation 1 v atio v 797855377 30205917401 29614257799 60486602801 707557678	More Info           Observation 2           Exp Ratio           117:113           1.77118159532546999           1.637024993066485           2.0250289440154998           2.221623097434998           4.2122168540954599	Observation 3  Exp Ratio  International Inte	Ignore	•
1 Pro 2 99 3 99 4 99 5 99 6 99 7 99	D         Obser           D         Cobser           Exp R           oteins         114:113           80G63         1.46788           98WF3         1.97668           99747         2.38947           96H02         2.86102           96D66         3.34422           92545         1.4115	vation 1 v atio v 797855377 30205917401 29614257799 60486602801 707557678 40031433101	More Info           Observation 2           Exp Ratio           117:113           1.71181595325466999           1.63702499866486           2.0250289440154998           2.82182097434998           4.2122168540954599           1.3763279914855599	Observation 3  Exp Ratio I15:113 1.75942599773407 1.3063050508499101 1.35038006889799 1.7519400119781501 1.7594920396804801 1.5192824983851517	Ignore 	
1 Pro 2 Q9 3 Q9 5 Q9 6 Q9 7 Q9 8 Q8	D         Obset           Exp R.         Obset           Exp R.         Obset           BWW3         1.4713           SWW3         1.4788           SWW3         1.9766           SP147         2.38947           S6H02         2.86102           S6H02         3.34422           22545         1.44115           SWWMW7         1.42413	OBSERVATION vation 1 v atic v 797855377 30205917401 29614257799 60486602801 707557678 40031433101 60425949099	More Info           Observation 2         ▼           Exp Ratio         ▼           117:113         ▼           1.718:1595325466999         1.63702499866486           2.0250289-440154998         2.22182597434998           4.2122168540954599         1.3765279914855999           1.617674946784097         1.617674946784097	Observation 3  Exp Ratio T15:113 1.75942599773407 1.3063050508499101 1.3950380086898799 1.75944920396804801 1.91282498836510 1.91282498836510 1.91282498836510	Ignore	
1 Pro 2 99 3 99 6 99 7 99 8 9 96	D         ✓         Obser           D         ✓         Obser           Exp R.         Exp R.         Exp R.           000000000000000000000000000000000000	OBSERVATION vation 1 v atio v 797855377 30205917401 29614257799 60486602801 707557678 40031433101 60425949099 202991486	More Info           Observation 2           Exp Ratio           177.113           1.7718159532546999           1.63702499806486           2.82182097434998           4.2122168540954599           1.6763279914855999           1.6767494678497           1.4387738986774301	Observation 3  Exp Ratio Exp Ratio T15:113 1.75942599773407 1.3063050508499101 1.91282498836517 1.91282498836517 1.4782110452652 1.2218159437175601	<b>Ignore</b>	
1 Pro 2 99 3 99 4 99 5 99 6 99 7 98 8 9 6 10 02	D         Obset           Exp R.         Obset           Exp R.         Obset           SUG63         1.48788           SWW3         1.97666           SHW72         2.86102           SG6402         2.86102           SG066         3.34422           S2945         1.44115           SP206         1.39203           SP209         1.39203           SP209         1.39203	OBSERVATION vation 1 • atio • 797855377 30205917401 29614257799 60485602801 707557678 40031433101 60425949099 202991486 69718933101	More Info           Observation 2         ▼           Exp Ratio         ▼           17:113         ▼           1.7:10539325466999         1.63702499806486           2.0202894400154998         2.22182097434998           4.2:1221685409545999         1.61767494678497           1.3763279514855599         1.61767494678497           1.38772010165771         1.38177010165771	Observation 3  Exp Ratio Exp Ratio 115:113 1.75942599773407 1.30603050506499101 1.395038008689799 1.7519400119781501 1.759420398604801 1.759420198604801 1.75942048048051 1.759420480517 1.78210452652 1.221819437179601 1.379564048057 1.379564048057 1.379564048057 1.379564048057 1.379564048057 1.3795640480 1.379564048057 1.3795640480 1.3795640480 1.3795640480 1.3795640480 1.3795640480 1.3795640480 1.3795640480 1.3795640480 1.3795640480 1.379564 1.3795640 1.379564 1.379564 1.379564 1.379564 1.379564 1.379564 1.379564 1.379564 1.379564 1.379564 1.3795 1.379 1.3795 1.379 1.3795 1.379	Ignore	
1 Pro 2 99 3 99 4 99 5 99 6 99 7 99 8 8 86 9 90 10 02 111 01	D         ✓         Obser           D         ✓         Obser           Exp R.         Exp R.           BW#3         1.4:113           JUG63         1.4:8788           BW#7         2.36947           J6HN2         2.86102           J6D606         3.3:4422           J29245         1.4:4115           SVPQ9         1.39923           SVPQ9         1.39923           Z7/81         1.2:4144	OBSERVATION vation 1 v atio v 797855377 30205917401 20514257799 60486602801 707557678 40031433101 60425949099 202991486 69718933101 70353317301	More Info           Observation 2         ▼           Exp Ratio         ▼           117.113         1.7118159532546999           1.63702499866486         2.2020289440154998           2.2020289440154998         2.42182097434998           1.376279914855999         1.3763279914855999           1.3767494678497         1.381770101165771           1.3805696566504         1.3905649566504	Observation 3         ▼           Exp Ratio         ▼           115:113         ▼           1.35942599773407         ▼           1.30500006499101         1.36938064801           1.9242498836517         1.718110452652           1.221819437179601         1.3795640465574601	Ignore	
1 Pro 2 99 3 99 6 99 7 99 8 08 9 96 10 02 11 02	D         Obser           D         Obser           Exp R.         Obser           BUG63         1.48788           80WF3         1.9766           9747         2.38947           96H82         2.86102           96G6         3.4422           8245         1.44113           9747         2.38947           95H82         2.86102           96D66         3.4422           82929         1.39923           97299         1.39923           9729         1.39592           16531         1.65377	OBSERVATION vation 1 • 797855377 30205917401 29614257799 60486602801 60485602801 60425949099 202991485 69718933101 70353317301 7035282699	More Info           Observation 2           Exp Ratio           I7:113           1.7:118159532546999           1.63702495806486           2.22168540954599           1.3763279914855999           1.3763279914855999           1.3763279914855999           1.3763279914855997           1.36170101165771           1.30649566504           1.370040950122099	Observation 3  Exp Ratio T15:113 T15:113 T1.594259973407 T1.3063050506499101 T.759490019781501 T.75949020386828979 T.7519400119781501 T.79482038682517 T.4782110452652 T.2218159437179601 T.3795640468597401 T.3809180259704601 T.37956101139	Ignore  Ignore	
1 Pro 2 Q9 3 Q9 4 Q9 5 Q9 6 Q9 7 Q9 6 Q9 7 Q9 8 Q8 9 Q6 10 Q2 11 Q1 12 Q1 13 Q1	D         Obser           D         Cobser           Exp R.         Obser           BWW3         114:113           SUG63         1.48788           BWW7         2.38947           S0606         3.34422           S0606         3.34422           S0606         3.34422           S245         1.44115           SP209         1.39923           S72181         1.24144           16576         1.95682           6531         1.65377           16270         2.67023	OBSERVATION vation 1  vation 1 vation 1 vation 1 vation 1 vation 2 vation	More Info           Observation 2         ▼           Exp Ratio         ▼           117.113         1.7118159532546999           1.63702495866486         2.0250289440154998           2.0250289440154998         2.4212097434998           1.3763279674807         1.3767494578497           1.3617010165771         1.3656774307           1.305649566504         1.200568566504           1.02084956650122099         4.0298419341931197	Observation 3         ▼           Exp Ratio         ▼           115:113         ▼           1.3063050506499101         1.3063050506499101           1.35038008886799         1.7514400119781501           1.75242038604801         1.91282498836517           1.4782110452652         1.227152652           1.23795640468597401         1.379564048597401           1.2772680521011399         2.68053007125854	Ignore  Ignore	

Fig. 16.16 Creation of a dataset with the IPA software. *Red* rectangles spotlight the basic steps to perform an analysis for a dataset

option in the IPA window (Fig. 16.15), select the file you wish to analyze and save the file. For illustration purposes, we analyzed proteins differentially expressed in common in Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cell lines from Calderón-González et al. [18]. It is necessary to specify the following information for the data that you wish to analyze:

- 1. File format: Flexible format
- 2. Column header: Yes
- 3. Identifier type: UniProt/Swiss-Prot accession
- 4. Array platform: In this case, it does not apply

Then the observation names must be edited, specifying the ID of proteins; in our case, the observation option 1 was selected (114:113. MCF7/MCF 10A), 2 (117:113. T47D/MCF 10A), 3 (115:113 MDA-MB-231/MCF10A), according to data number. Finally, the quantitative data format must be specified, which in our case we chose Exp Ratio (Fig. 16.16).

To carry out IPA Core analyses, we first uploaded the dataset previously created and then specified the parameters according to the goals of our study. The IPA platform gives different options to filter the data. We filtered the parameters for breast cancer disease as follows:

- General settings: Ingenuity knowledge base (genes only). Considering direct and indirect relationships
- Networks: 25 interaction networks with 35 molecules per interactome. Include endogenous chemicals (default parameters)
- 3. Data sources: All
- 4. Confidence: All
- 5. Species: Human with stringent filter
- Tissues and cell lines: Mammary gland as organ and all breast cancer cell lines of database
- 7. Mutations: All.

At the end of the page, cutoff values are selected. We focused on up- and down-regulated proteins (Fig. 16.17). The statistical significance was determined by Fisher's Exact Test, for which the p-value cutoff was set at 0.05. As a result of this analysis, we obtained three summary results, one for each observation. Then, we performed a Core Comparison Analysis. This analysis was performed using the following option (Core: Compare analysis). The procedure also requires

e Core Analysis - [analysis : IPA Analysi	s 181215]	
General Settings	Population of genes to consider for p-value calculations:	
Networks Interaction	Reference Set Ingenuity Knowledge Base (Genes Only)	Ŧ
Data Sources All	Relationships to consider:	Optional Analyses:
Confidence Experimentally Ob	Affects networks and upstream regulator analysis	🗃 🗹 My Project
Species All	Direct and Indirect Relationships     Direct Relationships	V My Pathways
Tissues & Cell Lines	Contraction of the second s	
Mutation All		
ADVANCED SAVE AS REFAULTS		
Apression value type Cutott Ra	nge Pocus on	
Exp Fold Change -4.	5131 to 0.433 Both Up/Downregulated T	eady molecules across observations
eview Dataset IPA Analysis 181215	Observation: Observation 1 (98) 💌	

Fig. 16.17 Core parameters needed for IPA analysis. Figure shows the different parameters that need to be set to perform and delimit a Core Analysis. In this case the analysis was focus on breast cancer disease

selecting files for comparison. The summary results for all observation are reported in a single file. The Core Analysis result window shows different tool bars:

- 1. Canonical Pathways (Chart and HeatMap)
- 2. Upstream Analysis (Table and HeatMap)
- 3. Diseases & Functions (Chart and HeatMap)
- 4. Regulator effects (Table)

- Networks (Networks for each observation or overlapping networks)
- 6. Molecules (Tables).

We focused our analysis on canonical pathway result obtained as a chart (Fig. 16.18a) or a HeatMap (Fig. 16.18b). In both cases, the number of up- and down-regulated proteins and their statistical probability were reported. Some of the



**Fig. 16.18** Classification of proteins found up- or down-regulated in both Luminal A and Claudin-Low breast cancer cell lines into canonical pathways with IPA software. The result can be displayed as (**a**) Bar chart or (**b**) Heatmap

processes affected were: Fatty acid oxidation I  $(\downarrow ACADM, \downarrow ECI1, \downarrow HADH, \downarrow IVD, \downarrow SCP2,$  $\downarrow$ SLC27A4 with a p-value 3.57  $\times$  10<sup>-8</sup>), aspartate degradation II ( $\downarrow$ GOT2 and  $\downarrow$ MDH2, p-value of  $3.78 \times 10^{-4}$ ), cell cycle control of chromosomal replication ( $\uparrow$ MCM3,  $\uparrow$ MCM4 and  $\uparrow$ MCM6, p-value 1.01  $\times$  10<sup>-3</sup>), telomere extension by telomerase ( $\uparrow$ XRCC5 and  $\uparrow$ XRCC6, p-value  $5.44 \times 10^{-3}$ ), and protein and ubiquitination pathway (HSP90AB1, ↑PSMA3, ↑PSMC1, ↑PSMD2, ↓PSMD3, and ↑PSMD7, pvalue  $8.65 \times 10^{-3}$ ).

Diseases functions are divided into two categories, Diseases and Bio Functions and Tox Functions. We only obtained the first category. We found the affected processes to be:

- 1. Cell-to-cell signaling and interaction: Formation of focal adhesions ( $\downarrow$ CTNND1 and  $\uparrow$ STMN1, p-value 1.30  $\times$  10<sup>-3</sup>)
- 2. Cellular assembly and organization: Formation of focal adhesions ( $\downarrow$ CTNND1 and  $\uparrow$ STMN1, p- value 2.39 × 10<sup>-2</sup>) and polymerization of microtubules ( $\uparrow$ STMN1, p-value 2.39 × 10<sup>-2</sup>)
- 3. Cellular function and maintenance: Formation of focal adhesions ( $\downarrow$ CTNND1 and  $\uparrow$ STMN1, p-value 1.30 × 10<sup>-3</sup>) and polymerization of microtubules ( $\uparrow$ STMN1, p-value 2.39 × 10<sup>-2</sup>)
- 4. Cell death and survival: Anoikis ( $\downarrow$ CTNND1 and  $\uparrow$ ILK, p-value 3.99  $\times$  10<sup>-3</sup>) and cytotoxicity of breast cancer cell lines ( $\downarrow$ RELA, p-value 3.17  $\times$  10<sup>-2</sup>)
- 5. Drug metabolism: Synthesis and oxidation of tretinoin ( $\downarrow$ ALDH1A3, p-value 8.02 × 10<sup>-3</sup>)
- 6. Cellular development: Epithelial-mesenchymal transition of breast cancer cell lines ( $\uparrow$ ILK and  $\uparrow$ STMN1, p-value 4.45 × 10<sup>-2</sup>) among other processes

The interactome data obtained in three separate experiments were processed resulting in identification of two principal networks related to: (1) Cellular development, cellular growth and proliferation, cellular movement, cell death and survival, and cancer, with a score of 19 and 14 molecules involved (\LALDH1A3, \CTSD,

 $\downarrow$ DLG1,  $\downarrow$ EZR,  $\uparrow$ FUS,  $\uparrow$ ILK,  $\uparrow$ KPNB1,  $\downarrow$ MVP,  $\downarrow$ RELA,  $\downarrow$ S100A8,  $\uparrow$ SET,  $\downarrow$ SLC25A5,  $\uparrow$ XRCC5 and  $\uparrow$ XRCC6) (Fig. 16.19a). (2) Cell death and survival, cellular development, DNA replication, recombination and repair, cancer and hereditary disorder obtained 12 proteins (↑ABCF2, ↑CAD, CTNND1. ↓CYCS, ↑HSP90AB1,  $\downarrow$ LGALS3BP,  $\uparrow$ MAT2A,  $\uparrow$ MCM6,  $\uparrow$ MSH6,  $\uparrow$ NUMA1,  $\uparrow$ PCNA,  $\uparrow$ SNRPG) with a score of 15 (Fig. 16.19b). Proteins in red and green represent the up- and down- regulated proteins, respectively. Small molecules are shown in gray color to highlight their relationship with our proteins. Created Networks can be exported to IPA pathway for subcellular localization and decoration of network with organelles and backgrounds.

## 16.7 Biomarkers Module

To perform biomarker filtration, we used the Biomarkers module. As a first step in using the Biomarker module, we selected the analysis dataset function and choose a dataset created previously. Next we chose the following parameters:

- 1. Species: Human
- 2. Tissues and cell lines: mammary gland as organ and breast cancer cell lines
- 3. Molecules: All
- 4. Diseases: Cancer
- 5. Biofluids: All
- 6. Biomarkers: All biomarkers application (diagnosis, disease progression, efficacy, not applicable, prognosis, response to therapy, safety and unspecified application) and breast disease (breast cancer, breast carcinoma, ductal carcinoma, ductal carcinoma in situ, ductal infiltrating breast carcinoma, infiltrating lobular breast carcinoma, invasive ductal breast cancer, lobular breast cancer, mammary neoplasm, metastasic breast cancer) (Fig. 16.20a).

We then ran the analysis, saved the results, and performed a comparative analysis on our Fig. 16.19 IPA Networks of proteins found up- or down-regulated in both Luminal A and Claudin-Low breast cancer cell lines. The up- and downregulated proteins are represented by molecules in red and green color, respectively. (a) Interactome related to cellular development, cellular growth and proliferation, cellular movement, cell death and survival, and cancer. (b) Interactome involved in cell death and survival, cellular development, DNA replication, recombination and repair, cancer and hereditary disorder



datasets. In this analysis, we had three datasets to compare (Fig. 16.20b) and only considered proteins found in all three datasets. We found four candidate biomarkers common between the luminal A and Claudin-low cells falling into different biomarker application categories: unspecified application ( $\uparrow$ KHSRP protein found in nucleus and  $\downarrow$ S100A8 with cytoplasmic localization), diagnosis, efficacy ( $\downarrow$ RELA localized in nucleus and  $\uparrow$ STMN1 found in cytoplasm) RELA was also found related to the drug NF-kappa B decoy (Fig. 16.21). All proteins were found in blood and all are related to cancer; however, they are not unique to this disease, as they are found in other diseases.

## 16.8 Protein-Protein Interactions Databases

## 16.8.1 STRING

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a database of



**Fig. 16.20** Filter parameters for biomarker analysis in IPA software. (a) Creating a filter for putative biomarkers. (b) Comparison analysis between all observations (MCF7, T47D and MDA-MB-231)

known and predicted protein interactions [31]. This database was developed by the Center for Protein Research (CPR), The European Molecular Biology Laboratory (EMBL), The Swiss Institute of Bioinformatics (SIB), The University of Copenhagen (KU), The Technische Universität Dresden (TUD), and The Universität Zürich (UZH). STRING version 10.0 has 9,643,763 proteins from 2031 organisms. The

main objective of this database is to integrate, predict and unify several protein-protein interactions [31, 32]. Associations between proteins can be physical (direct) or functional (indirect). The functional associations are defined as the interaction between two proteins that participate or contribute in the same cellular process or metabolic pathway, as well as other functional processes [32–34].

Co	mparicon bio	markers										ಕರ್
	000000	Unique Bi Click a Bi O Group O Group	iomarkers omarker Filte p 1 p 2 p 3	r Result name to view the	: potential biomarke	rs unique to it versus the others.	Cick Cick comp	ton Biomarkers the link below to view po ared. iew Common Biomarker;	tential biomarkers cor	nmon	across all Biomarker Filter Result	is that were
Th	ommon bion to to MV PA ere are 4 gene	narkers \ Compariso	n Details \ IIST CUSTO									
	/ Symbol	Entrez Gene Name	Location	Family	Drug(s)	UniProt/Swiss-Prot Accession(A1)	Exp Fold Change	A1) Exp Fold Change(A2)	Exp Fold Change(A3)	Bloo	d Bronchoalveolar Lavage Fluid	Cerebral Spinal Flu
	KHSRP	KH-type splicing regulatory protein	Nucleus	enzyme		Q92945	1.441	1.376	1.913	x		
	RELA	v-rel avian reticuloendothelios s viral oncogene homolog A	Nucleus	transcription regulator	NF-kappaB decoy	Q04206	-2.293	-1.591	-1.337	×		
	\$100A8	S100 calcium binding protein A8	Cytoplasm	other		P05109	-2.124	-2.348	-2.417	×	x	
-	STMN1	stathmin 1	Cytoplasm	other		P16949	2.572	2.178	2.736	×	×	

Fig. 16.21 Result of biomarker filter. Figure shows the four common biomarkers between. Luminal A and Claudinlow breast cancer cell lines

STRING database uses the following type of information to predict possible interaction:

- 1. Genomic data
- 2. High throughput experiments
- 3. Co-expression
- 4. Data extracted from literature

STRING import knowledge about proteinprotein interactions from other databases such as IntAct, MINT, BioGRID, Reactome, KEGG, BIND, HPRD, DIP, NCI-Nature Pathway Interaction, GO, and EcoCyc [33]. In addition, STRING has a large collection of predicted interactions that are produced de novo using prediction algorithms [33, 35]. De novo predictions are made using genomic context such as conserved genomic neighborhood, gene fusion events, and co-occurrence of genes across the genome [34]. STRING also performs searches for genes with similar transcriptional response through a variety of conditions (co-expression) [33]. Information extracted from literature is another source used to extract protein association information from. In this case, STRING obtains information from all abstracts in PubMed database directly [36]. Finally, STRING assigns a probabilistic confidence score to all associations obtained through comparison of the association predictions against a reference database. STRING uses the KEGG database because this is manually curated [32, 37].

STRING website is composed of two components, the first component deals with protein analysis and the second covers the platforms (Fig. 16.22). The window of results displays the networks of protein-protein associations. The resulting interactome is represented by connecting lines. Each one of these lines represents different types of evidence. Networks can be viewed in three forms:

- Evidence view in which connections are color coded as follows, neighborhood (green), gene fusion (red), co-occurrence (blue), co-expression (black), experiments (purple), database (light blue), text mining (yellow), and homology (gray)
- Confidence view in which the thickness of connecting lines correlates with the strength of the associations
- 3. Interaction view in which the type of interactions is color coded as follows; activation (brilliant green), inhibition (red), binding (blue), phenotype (brilliant blue), catalysis (purple), posttranslational modifications (lilac), reaction (black) and expression (olive green)

by name protein sequence names sequences	What it does
protein name: (examples: <u>#1 #2 #3)</u> Q9Y266	STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:
(STRING understands a variety of protein names and accessions: you can also try a random entry)	Genomic High-throughput (Conserved) Previous
organism:	
Homo sapiens V	STRING quantitatively integrates interaction data from these sources for
interactors wanted:	a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 9'643'763
COGs Proteins Reset GO!	proteins from 2'031 organisms.
please enter your protein of interest	
More Info Funding / Support Acknowledgement	ts Use Scenarios
STRING (Search Tool for the Retrieval of Interacting Gen	enes/Proteins) is being developed at CPR, EMBL, SIB, KU, TUD and UZH.
STRING references: <u>Szklarczyk et al. 2015</u> / <u>2013</u> / . Miscellaneous: Access Statistics, Robot Access Guide, S	<u>2011</u> / <u>2009</u> / <u>2007</u> / <u>2005</u> / <u>2003</u> / <u>Snel et al. 2000</u> . Supported Browsers.
Mast's Nam2 This is warsian 10 of CTDINC name save	ering more than 2000 organisms, and with improved prediction algorithms!
what s new? This is version 10 or STRING - now cove	sister and the bally of CTDING data

**Fig. 16.22** STRING window view. The STRING webpage has different options to perform interaction analysis. The search can be done by the name of the protein or

STRING has also an interactive view. In this option the network can by reordered by moving the proteins in the network. In advanced option, the network can be enriched into a GO Biological Processes, GO Molecular functions, GO Cellular components, KEGG Pathways, PFAM domains, INTERPRO domains, and Protein- Protein interactions. In each enrichment category, a new window is displayed containing a list of interactors, which contains different processes, the number of proteins involved as well as a pvalue.

## 16.8.2 Protein-Protein Interaction Networks

To determine the protein-protein interaction of overexpressed NUDC protein exclusively found in Claudin-low breast cancer cell line [18], we a protein sequence. The analysis can be performed for multiple proteins in the same way. In addition, the main page has various tabs with information about this platform

accessed the STRING website http://string-db. org/.

To generate a network of protein interactions, a list (one or more) of protein names, accession number, or sequence, as well as the organism or species they originated from, need to be specified (Fig. 16.22). At the bottom of the result window there is a parameter box. The options in the parameter box are used to select the active prediction algorithm. The confidence score as well as the number of interactors can be adjusted as well (Fig. 16.23). The interactome can be seen according to evidence (Fig. 16.24a), confidence (Fig. 16.24b) and action (Fig. 16.24c). In each network, a score is generated according to each protein's interaction evidence. In addition, a brief description for each protein is also displayed (Fig. 16.24). NUDC protein is associated with PAFAH1B1



Fig. 16.23 STRING results view. A window containing different parameters is shown at the *bottom*. The active prediction methods as well as the confidence of the interactions in the network can be selected in this window

(platelet-activating factor acetylhydrolase 1b), PLK1 (polo-like kinase 1), NDEL1 (nudE nuclear distribution E homolog (A. nidulans)like 1), HSP90AA1 (heat shock protein 90 kDa alpha), BTRC (beta-transducin repeat containing E3 ubiquitin protein ligase), NDE1 (nudE nuclear distribution E homolog 1 (A. nidulans)), ZW10 (ZW10, kinetochore associated, homolog (Drosophila), FBXW11 (F-box and WD repeat domain containing 11), CLIP1 (CAP-GLY domain containing linker protein 1) and ZWILCH (Zwilch, kinetochore associated, homolog (Drosophila)). All interactions have more than 0.90 score. In



**Fig. 16.24** Interaction network of NUDC protein. This polypeptide is overexpressed exclusively in Claudin-low breast cancer cell line. The interactome can be seen in three options. (a) Evidence view, where the color lines represent the diverse evidences of interactions: *Green*, neighborhood; *red*, gene fusion; *blue*, co-occurrence; *black*, co-expression; *purple*, experiments; *light blue*, database; *yellow*, text mining;

gray, homology. (b) Confidence view where thicker lines represent stronger associations. (c) Interaction view, where the different modes of action are represented by different colors. *Brilliant green*, activation; *red*, inhibition; *blue*, binding; *brilliant blue*, phenotype; *purple*, catalysis, lilac, PTMs; *black*, reaction; *olive green*, expression. The three view modes provide a score of the different evidence of interaction addition, the network was enriched into GO Biological Processes. Processes showed Enrichment with statistical significance were:

- 1. Mitotic prometaphase  $(4.940 \times 10^{-13})$
- 2. Mitotic anaphase  $(8.089 \times 10^{-12})$
- 3. Mitotic M phase  $(6.309 \times 10^{-11})$
- 4. M phase  $(6.309 \times 10^{-11})$
- 5. Mitotic cell cycle phase  $(4.300 \times 10^{-10})$
- 6. Cell cycle phase  $(4.300 \times 10^{-10})$

All processes mentioned above have at least eight proteins involved. We selected the cell cycle phase process as an example. The proteins enriched in this process are shown in color red (Fig. 16.25a). We selected the interacting proteins NUDC and ZW10 as examples to extract interaction information. ZW10 was selected because it is an essential component of the mitotic checkpoint that prevents cells from prematurely exiting mitosis. The evidence supporting the functional link between these two proteins are the following:

- 1. Co-expression (putative homologs are co-expressed in other species, score 0.065)
- 2. Association in curated database (score 0.900)
- 3. Co-mentioned in PubMed abstracts (score 0.285)

Also putative homologs are mentioned together in other species (score 0.192). The combined score is 0.938. There is also activity evidence, such as catalysis (score 0.900), binding (score 0.900) and reaction (score of 0.900) that support the interaction between these two proteins (Fig. 16.25b). For proteins selected in a network, STRING displays a window with information about their 3D structure, as well as links to Ensembl, GeneCards, KEGG, Nextprot and UniProt. Also, STRING can show the protein sequence and the sequence of its homologs in organisms stored in STRING. NUDC has three 3D structures obtained from Protein DataBase (PDB) (Fig. 16.25c). As mentioned above, STRING can perform network analysis for multiple proteins as well. We performed an interactome analysis for the up- and down-regulated proteins common in Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cells lines [18]. In this case, we used the highest confidence (0.900) possible to generate our interaction network. The network has several interaction nodes related to:

- 1. Energy metabolism
- 2. Translation
- 3. Proteasome
- 4. Replication and repair
- 5. Transcription

Red and green arrows indicate up- and down-regulated proteins, respectively (Fig. 16.26).

#### 16.8.3 MINT

The Molecular INTeraction database or MINT is an open source protein-protein interaction database developed at the Università degli Studi di Roma Tor Vergata that has been experimentally verified [38, 39]. The webpage can be found at http://mint.bio.uniroma2.it/mint/Welcome.do (Fig. 16.27). The current version of MINT database (November 2015) contains 241,458 interactions, corresponding to 35,553 proteins and 5554 PMIDS (PubMed unique identifiers). Species included are Drosophila melanogaster, Saccharomyces cerevisiae, Caenorhabditis elegans, mammals and viruses, with mammal databases being the main datasets. Evidences for protein-protein interactions include association studies, co-localization, direct interactions, interactions in form of complexes, enzymatic reactions, and high throughput studies. Protein-protein interactions have been identified by a number of methods including coimmunoprecipitation with either anti-bait or anti-tag antibodies, fluorescence microscopy, peptide arrays, protein arrays, pull down experiments, SPR, tandem affinity isolation, two hybrid arrays, two hybrid pooling, and two hybrid systems, etc. Additionally, the MINT database is freely available for academic and commercial users.



**Fig. 16.25** Interaction network of NUDC overexpressed protein found exclusively in Claudin- low breast cancer cell line. STRING platform provides different information for the generated network. (a) Network enrichment for GO Biological Processes. The proteins in *red* which

have a statistical significance (*p*-value) are involved in cell cycle phase. (b) Evidence supporting interaction between NUDC and ZW10. (c) 3D protein structure information

There are three additional databases available via MINT website including HomoMINT, Domino, and VirusMINT. The first one is an inferred network for human; the second is specialized in domain-peptide interactions, and the last is a protein-protein interaction database specialized on viruses.

Protein interaction searches in MINT database (Fig. 16.28a) can be carried out using PubMed ID, D.O.I, or author's name. Alternatively, this



**Fig. 16.26** STRING interaction network of proteins found up- or down-regulated in both Luminal A (MCF7 and T47D) and Claudin-low (MDA-MB-231) breast cancer cell lines. This list has interaction nodes related to: (1) Energy metabolism, (2) Translation, (3) Proteosome degradation, (4) Replication and repair, (5) Transcription. Colored lines represent different evidence of interaction:

*Green*, neighborhood; *red*, gene fusion; *blue*, co-occurrence; *black*, co-expression; *purple*, experiments; *light blue*, database; *yellow*, text-mining; *gray*, homology. *Red arrows* indicate up-regulation and *green arrows* down-regulation. A box with information about some proteins is also shown



Fig. 16.27 Homepage of the Molecular INTeraction database, MINT

database can be searched against protein or gene name, protein accession number (Protein AN) or keywords. Protein accession numbers recognized by MINT search engine are FlyBase, Ensembl, Human Identified Gene Encoded Large Protein Analyzed database (HUGE), Nematode database (WormBase), OMIM, REACTOME pathway database, the *Saccharomyces* Genome Database (SGD), and Universal Protein Resource Knowledgebase (UniProtKB).

To demonstrate how MINT database works, we selected the vesicle-fusing ATPase NSF (P46459) for analysis. This protein is part of a set of proteins that were found overexpressed in several breast cancer cell lines [18]. To follow our analysis, click on the Search tab and type P46459 (Fig. 16.28, arrow 1) and then select the organism (Fig. 16.28, arrow 2) and then press the Search key (Fig. 16.28, arrow 3). Results show certain information for the queried protein including its ID, species, synonyms, domains found in query, a link to its role in diseases, its gene ontology, references covering the target protein, prediction of its modular domain interactions (ADAN), and its orthologs in MINT database (Fig. 16.28). Results also display a window containing a list of molecules interacting with the target according to MINT database, evidence for each interaction and a global score for each interaction (Fig. 16.28).

MINT Home	Search	Curation								
earch pubmed id/D.O.I/author: search			Statistic	s Download		C	ontacts/Lin	ks/Linking		-
search				Search MINT-ID/IMEx ID:		10	Viewer sea	arch		
earch proteins in MINT by:				Organism:						
Protein or gene name:				all Mammalia V	iruses		* uniprot	db, ensembl	, flybas	е,
<ul> <li>Protein accession number *:</li> </ul>	1	1	1 -	Homo sapiens			reactome	)	n, nage	,
la contra de la Ordena de la		*	2	Saccharomyces cerevi	siae		Coordh			
<ul> <li>keywords: (min 6 characters)</li> </ul>	•)			Drosophila melanogast	er	27	<b>Search</b>	ų.		
				Caenorhabditis elegans	5	3				
CONNECT proteins	Enter a list of produced	oteins (e.g. pro	teins in a	Blast proteins in MIN	r: Blast					-
interactions connecti	ing them (use cros	s references: u	uniprotkb,	(paste sequence in FAS	STA format)					
CONNECT	o, wormbase, om	m, nuge) 🔹 Ir	iciude	11						
connecting proteins i	not in the list O	only consider p	roteins in this	11						
				11						
				11						
				11						
				8						
										0
HomoMINT: an inferred human netwo	ork D	omino: a domain	peptide interactio	rs database	VirusMINT: a v	virus proteir	n interactions	database		•
: HomoMINT: an inferred human netwo MTATT Home Ser	ork D arch Ci	omino: a domain	peptide interactio Statistics	ns database Download	VirusMINT: a v	virus proteir	n interactions o	database /Linking		(
: HomoMINT: an inferred human netwo א לאוד Home Ser sicle-fusing ATPase	ork D arch Ca Binary Intera	omino: a domain iration	peptide interaction Statiatica	ts database Download export partners sequences	VirusMINT: a v	virus proteir Cr Fasta	n interactions ( ontacts/Links MINT viewer	database A.inkina	3 Vien	w ir
: HomoMINT: an inferred human netwo געדאוד אסייר אסייר אסייר icle-fusing ATPase	ork D arch Ci Binary Inter MINT vie	omino: a domain iration	peptide interaction Statiatics	ns database Download export partners sequences i	VirusMINT: a v	virus proteil Co Fasta	n interactions ontacts/Links MINT viewer	database Alinkino	3 View hom	w ir
HomoMINT: an inferred human netwo MTATT Home Ser Icle-fusing ATPase	ork D arch Ca Binary Inter MINT vie	omino: a domain rration uctions wer	peptide interaction Statiatica	Is database Download export partners sequences MSF: Vesicle-tusing ATPart 5 partnerfol found in MINT	VirusMINT: a v n Fasta format ie (P46459)	virus protein Cr Fasta	n interactions o ontacts/Links MINT viewer	database Alinking	3 View	w ii
HomoMINT: an inferred human netwo ATAIT Home Ser Icle-fusing ATPase ProtKB AC P46459, Q9UK22, A8K22	ork D arch Cri Binary Inter MINT vier D9, QBN607,	omino: a domain tration	poptide interaction Statistics	Is database Download export partners sequences i NSF: Vesicie-fusing ATPart 15 partner(s) found in NIVT. Your query also matches 130	VirusMINT: a v n Fasta format ee (P46459) 49 interaction	rirus proteir Ca Fasta	n Interactions ( ontacts/Links MINT viewer (s) from other	database A.Inking	Vien horr	
: HomoMINT: an inferred human netwi געדאוד Home Sea sicle-fusing ATPase ProtKB AC P46459, Q9UK22, ABK22 Anism Homo sapiens (9606) NER Nethymaleinstice.	ork D arch Ct Binary Inter MINT vie D9, Q8N6D7, antilitye fusion protein	omino: a domain rration uctions wer	peptide interaction Statiation	Is database Download export partners sequences in Spartner(s) found in MINT. Your query also matches 14 protein	VirusMINT: a v n Fasta format ie (P46459) 49 Interaction	rirus proteil Ca Fasta evidence	n interactions ontacts/Links MINT viewer (s) from other	database A Inkine • databases ir ss coloc coo	View hor	
HomoMINT: an inferred human netw MTNT Home Sar Icle-fusing ATPase ProtKB AC P48459, C9UK22, A8K21 anism Homo sapiens (9606) NSF, N-ethylmalemide-s providese S16 (PBP011	ork D arch Cu Binary Interr MINT vie D9, Q8N6D7, ensitive fusion protein	omino: a domain irration interions wer	peptide Interaction Statistics	Ins database Download export partners sequences SPSF: Vesicle-tusing ATPart 15 partner(s) found in MINT. Your query also matches 143 protein X © GABBR2 Home sapiens ( X © GABBR2 Home sapiens (	VirusMINT: a v in Fasta format e (P46459) 49 Interaction (2V 075899) (20UBS5)	rirus protein Ca Fasta evidence (dences s 4 0 3 0	n interactions ontacte/Links MINT viewer (s) from other corror circuit a 0.74 1 0.74 1	database ALinking databases ir databases ir ss colog en 2 1 1	Vien hor	
HomoMINT: an inferred human netwo A/T A1T' Home Sev icle-fusing ATPase ProtKB AC P46459, Q9UK22, ABK21 anise Homo saplens (9606) enames and NSF, N-ethy/maleimide- protase, S16 (IPR0013 (PF0003308), AAA, sub (I	ork D arch Ca Binary Interr MINT vier D9, QBN6D7, aensitive fusion proteil 384), Cdo48_2 (IPR00 1PR003860), AAA_AT	omino: a domain iration ictions wer v. Vesicular-fusion 14201), ATPase Zontr (IPRC	peptide Interaction Statietica Protein NSF, NT_N 003959),	export partners sequences i export partners sequences i sequences	VirusMINT: a v n Fasta format e (P46459) 49 Interaction 090055) 2262)	rirus protein Ca Fasta evidence (dences s 4 0 3 0 2 0	n interactions ontacte/Links MINT viewer (s) from other corror circol a 0.74 1 0.59 1 0.59 1	database A Linkino databases in ss color for 2 1 1 1	Vien horr	
HomoMINT: an inferred human netwi MTN1T Home Sear licle-fusing ATPase ProtKB AC P46459, Q9UK22, A8K21 anism Homo sapiens (9606) enames and norms Peptidase. 316 (IPR00338), MAA, sub ( AAA, ATPase (IPR00388), MAA, sub ( AAA, ATPASE (IPR0048	ork D arch Co Binary Inter MINT vier D9, Q8N6D7, ensitive fusion protein 549), Cdc43. 2 (IPRO 19R003960), AAA, AT 3), Asp_decarb_fold	omino: a domain aration actions wer a, Vesicular-Iusion 4201), ATPase Cent (UPR009010),	peptide interaction Statistics	ts database Download export partners sequences MSF: Vesicle-tusing ATPart 15 partner(s) found in MINT. Your query also matches 143 Protein X © GABBRI Homo sapiens (X X © GABBRI Homo sapiens (X X © GABBRI Homo sapiens (X X © GABBRI Homo sapiens (X) X © GABBRI Homo sapiens (X) X © GABBRI Homo sapiens (X) X © GABBRI Homo sapiens (X)	VirusMINT: a v n Fasta format e (P46459) 49 Interaction (275899) (29055) 2282(2) 378) 21	evidencei	n interactions of contacts/Links MINT viewer (s) from other 0.74 1 0.59 1 0.55 1 0.55 1 0.55 1 0.55 1	database A Inkina databases ir color op 2 1 1 1 1 1	Vien hom	
HomoMINT: an inferred human netwi A T h 1 T Home Ser icle-fusing ATPase ProtKB AC P46459, O9UKZ2, A8K22 anism Homo saplens (9606) anames and NSF, N-ethylmaleimide-s nyms Peptidase. S16 (IPR00359 sains (PR003359, AAA, sub ( AAA, ATPase (IPR00359 asses OMIM: (601633), GC>0005829	ork D Binary Inter MINT vier D9, Q8N607, ensitive fusion proteil (Ph03960), AAA, AT (Ph03960), AAA, AT (Stage of the stage	omino: a domain irration inctions wer h. Vesicular-fusion H4201), ATPaseV/ Pase_centr (IPR Pase_centr (IPR	peptide interaction Statiatics	s database Download export partners sequences NSF: Vesicle-tusing ATPR 15 partner(s) found in MINT, Your query also matches 13 Prove Homo sapiens (P4 © GABBRT Homo sapiens (P4 © GRIA2 Homo sapiens (P1650 © FES Homo sapiens (P1650 © FES Homo sapiens (P1650	VirusMINT: a v n Fasta format ee (P46459) 49 interaction (075899) (09U855) 2262) 378) 1) 2)	rirus proteil Cr Fasta Iconces S 4 ( 2 ( 2 ( 1 ( 1 (	n interactions of a contractive links MINT viewer (s) from other core circle a 0.55 1 0.55 1 0 0.55 1 0 0.55 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	database A inking databases in ss color (m 2 1 1 1 1 1 1 1 1 1 1 1	View horr	
HomoMINT: an inferred human netwo At 7 h17 Home Ser Icle-fusing ATPase Icle-fusing ATPase TrotKB AC P46459, Q9UK22, A8K22 Infism Homo sapiens (9606) anames and nyms Peptidase, S16 (IPR0015 (IPR00338), AAA, sub (IPR0035 alns AAA, ATPase (IPR0035 ases OMIM: (601633), G0-2005629 G0-2005629 G0-2005631	ork D arch Ci Binary Inter MINT vie D9, Q8N6D7, ensitive fusion protein PR03960), AAA, AT 3), Asp_decarb_fold i	omino: a domain rration uctions wer n, Vesicular-fusion v4201), ATPaseV/ Pase_centr (IPRR IPR009010),	poptide interaction Statiation	Is database Download export partners sequences in Spartner(s) found in NIXT. Your query also matches 143 CABBR2 Homo sapiens (PA CABBR2 Homo sapiens (PA CABBR2 Homo sapiens (PA FTPN Homo sapiens (PA FTPN Homo sapiens (PA FTPS HOMO sapiens (PA	VirusMINT: a v i i i i i i i i i i i i i i i i i i i	rirus proteila Fasta Idonces S 4 0 2 0 1 0 1 0	n interactions ontects/Links MINT viewer (c) from other corps (inced a 0.55 1 0.55 1 0.55 1 0.43 1 0.43 1 0.43 1 0.28 1	database A inking databases in sa color en 2 1 1 1 1 1 1 1 1	View hom	
HomoMINT: an inferred human network           λ/Tλ17         Home         Sar           icle-fusing ATPase         Icle-fusing ATPase           rotKB AC         P46459, O9UK22, ABK21           mism         Home sapiens (9606)           mames and myms         NSF, N=ethylmaleimide-sam           ains         (PPI00338), AAA_aub ( AAA_ATPase (IPP00354)           see         OMMIs (60 1633), GO-3005629           gO-3005524         GO-3005531 GO-3005531           sontology         GO-3004776	ork D arch Ca Binary Interr MINT vie D9, Q8N6D7, ensitive fusion protein (PR003960), AAA, A1 (3), Asp_decarb_fold in	omino: a domain irration actions wer 1, Vesicular-fusion (4201), ATPase V Pase, centr (IPRO (IPR009010),	peptide Interaction Statistics	Is database Download export partners sequences Second Second S	VirusMINT: a v n Fasta format ie (P46459) 49 interaction (200855) 2262) 2378) 1) 2) U(RFS) 459)	evidences dences	n interactions antiacte/Links MINT viewer (s) from other 0000 Circol ( 0.74 1 0.55 1 0.55 1 0.55 1 0.55 1 0.55 1 0.43 1 0.43 1 0.43 1 0.43 1 0.28 1 0.28 1 0.28 1 0.29 1	database Alinkino databases ir databases i	View hom PSICQ compl	win
HomoMINT: an inferred human netwo Ad TA1T Home Sea Icle-fusing ATPase IrotKB AC P46459, Q9UK22, ABK21 Initiam Homo sapiens (9606) NSF, N-ethylmaleimide-s names and NSF, N-ethylmaleimide-s Peptidase_S16 (IPP0033 Initians (IPP00338), AAA_sub ( AAA_ATPase (IPP00338), AAA_sub ( AAA_ATPase (IPP00338), aAA_sub ( GO:0005829 GO:0005524 GO:0005524 GO:0005524 GO:0005524 GO:0005524	ork D arch Ca Binary Inter MINT vie D9, QBN607, aensitive fusion protein 384), Cdc48_2 (PR00 BP003600, AAA, A1 3), Asp_decarb_fold i	omino: a domain ration ctions wer h, Vesicular-fusion 4201), ATPaseVPR Pase_cent (VPR Pase_cent (VPR Pase_cent (VPR)),	peptide interaction Statistics	ts database Download export partners sequences is partners) found in MINT. Your query also matches 143 Consequences Conseq	VirusMINT: a v n Fasta format e (P46459) 49 Interaction (075899) (04UIS5) (2282) (04UIS5) (2282) (04UIS5) (04UI	evidence 6 2 1 1 1 1 1 1 1 1 1 1 1 1 1	n interactions of contacts/Links MINT viewer (s) from other (s) fr	database A Inkino	View hon	W
HomoMINT: an inferred human netwith         \u03c6 ATPase         Icle-fusing ATPase         icle-fusing ATPase         ProtKB AC       P46459, O9UKZ2, A8K22         nism       Homo sapiens (9606)         names and nymes       Peptdase. S16 (IPP00139)         alms       (PP003339, AAA, sub (IP000359)         ac0:0005829       GC:0005829         GC:0005829       GC:0005824         GC:0004725       GC:0004252         GC:0004252       GC:0004252         GC:0004252       GC:0004252         GC:0004252       GC:0004252         GC:000508       B	ork D arch Cri Binary Inter MINT vier D9, G8N6D7, ensitive fusion proteil 884), Cdo48_2 (IPR0 (IPR003690), AAA, AT 3), Asp_decarb_fold i	omino: a domain irration inctions wer h. Vesicular-fusion H201), ATPaseV/ Pase_centr (IPR (IPRD09010),	poptide interaction Statiatics	Is database Download export partners sequences is NSF: Vesicle-fusing ATPart T5 partner(s) found in NIKT, Your query also matches 143 protein X & GABBRT Homo sapiens (PA B GABBRT Homo sapiens (P1659 FTEN Homo sapiens (P1659) FTEN Homo sapiens (P1659 FTEN Homo sapiens (P1659) FTEN Hom	ViruaMINT: a v in Fasta format in Fasta format 49 interaction (75899) (240055) 2282) 378) 1) 2) UKR5) 42) UKR5) 459) 459) 459) 292240)	rirus protein Fasta evidences 100nces 2 0 2 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1	n interactions of ontacts/Links MINT viewer (s) from other core circles 0.55 1 0.55 1 0.55 1 0.55 1 0.55 1 0.55 2 0.43 0 0.28 1 0.28 1	database //Linking //databases in (s) coloci (n) 2 1 1 1 1 1 1 1 1 1	Vien hom PSICQ	w i i
HomoMINT: an inferred human netw At TA1T Home Sar Icle-fusing ATPase ProtKB AC P48459, C9UK22, A8K21 Infism Homo sapiens (9606) NSF, N-ethylmalemide-s payms Peptidase. S16 (PPR015 AA, ATPase (IPPR0359 ases OMIM: (601633), G0.2005629 G0.	ork D arch Ct Binary Inter MINT vie D9, Q8N6D7, ensitive fusion protein 984), Cdo48, 2 (IPR0 19R003960), AAA, AT 3), Asp_decarb_fold 1 efface: NP_006169, 2 bio8163,1 (RP-006169, 2 bio8163,1 (RP-0061	omino: a domain rration actions wer h, Vesicular-fusion 14201), ATPaseV/ Pase_centr (IPRr Pase_centr (IPR))	peptide interactio Statiation	Is database Download export partners sequences i SNSF: Vesicle-tusing ATPart 15 partner(s) found in NIXT. Your query also matches 143 CabBR2 Homo sapiens (PA CabBR2 Homo sapiens (PA CabBR2 Homo sapiens (PA FES Homo sapiens (PA CabBR2 Homo sapiens (PA FES Homo sapiens (PA CabCsL Homo	VirusMINT: a v in Fasta format e (P46459) 49 interaction 075899) (240855) 2262) 2482) 2482) 2482) 2482) 292 292) 292 292) 292 292 292 292 292	rirus proteil Fasta evidences ddnccs	n interactions antacta/Links MINT viewer (a) from other corp circcl a 0.74 1 0.55 1 0.55 1 0.55 1 0.43 1 0.28 1 0.	database A Linkino databases ir databases ir datababases ir databases ir databases ir databases ir databases	View hon	winol
HomoMINT: an inferred human netwo A/T A1T' Home Sev icle-fusing ATPase ProtKB AC P46459, O9UK22, ABK21 anise Homo saplens (9606) enames and NSF, N-ethyfmaleimide-s protKB AC P46459, O9UK22, ABK21 anise (P400338), AAA_aub ( AAA_ATPase (IP40038), AAA_AUB ( AAA_ATPASE (IP40048), AAA_AUB ( AAA_ATPASE (IP40048), AAA_AUB ( AAA_ATPASE (IP40048), AAA_AUB ( AAA_ATPASE (IP40048), AAA_AUB (	ork D arch Ca Binary Interr MINT vie D9, Q8N6D7, ensitive fusion protei (PR003960), AAA, A1 (PR003960), AAA, A1 (S), Asp_decarb_fold ( afseq: NP_006169.2 H30613.1 (B003061) afseq: MP_006169.2	omino: a domain iration ictions wer (1) Vesicular-fusion (4201), ATPase V Pase_centr (IPRO (IPRO06010), (IPR006000), (IPR0	peptide Interaction Statiatica protein NSF, NT_N 203959), 2T_13885, 51 ensembl: 886927 enfect:	Is database Download export partners sequences i seport partners sequences Sequences i Sequences i Sequ	VirusMINT: a v n Fasta format e (P46459) 49 Interaction 075899) 0700 075899) 0700 075899 0700 075899 0700 075899 0700 0700 0700 0700 0700 0700 0700 0	rirus proteil Cri Fasta (dence) (denc)	a interactions antactiv/Linka MINT viewer (a) from other (b) from other (c) from	database A Inkino	View horr	will
HomoMINT: an inferred human netwo ATATT Home Sea icle-fusing ATPase ProtKB AC P46459, O9UK22, ABK21 anism Homo sapiens (9606) names and norms Peptidase_S16 (IPR003 nains (PR00338), AAA, sub ( AAA, ATPase (IPR0038) nains (OC005829 GO:0005829 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005824 GO:0005825 GO:0005825 GO:0005825 GO:0005825 GO:0005825 GO:0005825 GO:0005825 GO:0005828 GO:000588 GO:	ork D arch Ca Binary Interr MINT vier D9, QBN6D7, aensitive fusion proteil 384), Cdo48_2 (IPR00 197003960), AAA, AT 3(3), Asp_decarb_fold I efface: NP_006169.2 H30613.1 (BC030615) t; MINT-1669916 MIN	omino: a domain iration ictions wer 4201), ATPaseVA 4201), ATPaseVA Pase_cent (IPR (IPR009010), reactome: REA( ), jet: IPR0000644 Tr-212200 MINT-2	peptide interaction Statistics	ts database Download export partners sequences i Spartner(s) found in MINT. Your query also matches 143 protein X © GABBRT Homo sapiens (PA CABBRT Homo sapiens (P1650 FER Homo sapiens (P1650 SNVPH Homo sapiens (P1650) SNVPH Homo sapiens (P1650 SNVPH Homo sapiens (P1650) SNVPH Homo sap	VirusMINT: a v n Fasta format e (P46459) 49 interaction (77599) (2005) 2282 (2015) 2282 (2015) 2392H0) 2392H0) 2392H0) 2392H0) 2392H0) 2392H0) 2386) 2395 2305	rirus proteili Fasta evidencei donccei 4 0 2 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1	n interactions of contacts/Links MINT viewer (s) from other (s) fr	database A Inkino databases ir so color or 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	View hor	w i uu
HomoMINT: an inferred human netwi A T A 17 Home Sear icle-fusing ATPase ProtKB AC P46459, O9UKZ2, A8K22 anism Homo sapiens (8606) enames and onyms. Peptdase. S16 (IPR00139) mains (PPR00339), AAA, sub ( AAA, ATPase (IPR00359) sases OMIM: (601633), GC 0005824 GC 0005824 GC 0005824 GC 0004252 GC 00015031 GC 0004252 GC 0004252 Socolder 2 GC 0004252 Socolder 2 HNT-5005109 Prediction of protein-prote In Metry Metades Julian	ork D arch Cri Binary Inter: MINT vier D9, 08N607, ensitive fusion protein 084), Cd-48_2 (IPRoC IPROC3660), AAA, A1 03), Asp_decarb_fold I ensitive fusion protein 031, Asp_decarb_fold I ensitive fusion protein ensitive fusion protein	omino: a domain rration intions wer A. Vesicular-fusion 4201), ATPaseV/ Pase_centr (IPR (IPR009010), reactome: REAC (), [pi: IP1000064 Tr212200 MINTs Dular domaster (P)	poptide interaction Statiatics	Is database Download export partners sequences in the partners sequences in the partners because a sequences in the partners in the partners protein the GABBRT Home sequences in GABBRT Home sequences in the GABBRT Home sequences (CB EDES Home sequences (CB EDES L HOME se	VirusMINT: a v in Fasta format in Fasta format (P46459) 49 interaction (75899) (240055) 2282) 378) 1) 2) UKR5) 429 2282) 378) 1) 2) 20 20 386) 573)	Arros protein           Fasta           evidences           Idences           2           2           2           1           0           1           0           1           0           1           0           1           0           1           0           1           0           1           0           1           0           1           0           1           0	n interactions of ontacts/Links MINT viewer core circest a 0.55 1 0.55 1 0.55 1 0.55 1 0.43 0 0.28 0 00000000000000000000000000000000000	database // Linking databases ir science of color of 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	View home of the second	
HomoMINT: an inferred human network       At TA1T     Home     Sar       Icle-fusing ATPase       ProtKB AC     P48459, O9UK22, A8K21       Inlism     Home saplens (8606)       names and snyms     NSF, N-ethylmalemide-s       Peptidase.S16 (IPR0015 (PF000339), AAA, sub of (PF00339), AAA, sub of (PF00339), AAA, sub of (PF00339), AAA, at Pase (IPR00159)       aases     OMIM: (601633), G0.2005524 (G0.2006472 G0.2004252 G0.2004252 G0.2004252       rs Xrefs:     Patient NINT-5004254, p (Bab/emb/genabark: AA ENSG0000073669 min NSF2. Vesicle-fusing nat-1: Vesicle-fusing nat-1: Vesicle-fusing	ork D arch Ci Binary Inter MINT vie D9, Q8N6D7, ensitive fusion protein 849, Cdc48, 2 (IPRO 1PR03960), AAA, AT 3), Asp_decarb_fold I efseq: NP_006169.2 H30618.1 (BC030615 ± MINT-1669916 MIN ein InterAction of mol ATPase Caenonhabc	omino: a domain rration actions wer h, Vesicular-fusion 4201), ATPase V Pase_centr (IPR( IPR009010), reactome: REA4 h, jejt: IPR000044 IT-212260 MINT-2 bular domAiNs melanogaster (PH tile elegano (C944	peptide interaction Statiation Statiation protein NSF, NT_N 2003959), 27_13685, 51 ensembl: 886692 refseq: 54351) 9922 e 1)	Is database Download export partners sequences in the partner(s) found in MIXT. Your query alow matches 143 protein x © GABBR1 Homo sapiens (PA x © GABC Homo sapiens (PA x © CLOSL Homo sapiens (PA x WIX Homo sapiens (PA x Proteins linked to a dieses.	VirusMINT: a v in Fasta format 49 Interaction 075999) 029UB35) 029	Arrus protein           Fasta           evidence           fdences           2           2           1           2           1           2           1	n interactions ontacts/Links MINT viewer (a) from other corp circle () 0.74 1 0.55 1 0.55 1 0.43 1 0.45 1 0.28 1 0	database ALinking databases in ssi coloci en 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	View PSICQ COTTO 1 1 1 1	win

**Fig. 16.28** MINT search webpage. (a) Search in MINT can be performed using: (1) Gene or protein name, Protein ID or keywords and the species of interest or the whole database, (2) Protein sequence in FASTA format,

(3) a list of proteins. (**b**, **c**) Result of a query for vesiclefusing ATPase NSF from Homo sapiens (UniProtKB/ Swiss-Prot ID P46459). (**c**) List of NSF interactors are shown

Clicking on the MINT viewer will generate a list of interactions that are displayed as a function of score threshold. For each partner, a number showing evidence for interaction is shown (Fig. 16.29). As an example, we clicked on number 4 and a new window appeared showing the partner name, ID, and techniques used to determine the interaction, as well as a PubMed identifier containing this information (Fig. 16.29).

qo





1 evidence by imaging technique

```
details
```

Fig. 16.29 Binary interactions of the N-ethylmaleimidesensitive fusion protein NSF viewed in MINT database. (a) Basic information queried for NSF. (b) Binary interaction map of NSF with 15 interactors found in MINT database. (c) Selecting number 4 in (b), a new window is displayed showing the name of the corresponding interactor (GABBR2, Gamma-aminobutyric acid type B receptor subunit 2) and the experimental methods used to determine this interaction, as well as the PMID ID for the publication describing it

## 16.8.4 IntAct

IntAct is a database of protein-protein interactions, as well as a suite of analytical tools at The European Bioinformatics Institute (EBI), which is part of the European Molecular Biology Laboratory (EMBL) [40, 41]. All information has been curated by experts at the IntAct team.

This freely available database can be accessed through its webpage http://www.ebi. ac.uk/intact/.

As of November 26th, 2015 this database had registered 355,819 interactions, which included 89,340 interactors (proteins) described in 36,864 experiments, 13,892 PMIDs, and 564,831 binary interactions. Methods used for the determination of protein-protein interactions include tandem affinity purification, anti-tag co-immunoprecipitation, two hybrid systems, pull down experiments, two hybrid arrays, anti-bait co-immunoprecipitation, two hybrid pooling approach, and co-sedimentation, among others. The source of information mainly comes from human (42.5 %), various S. cerevisiae strains (22.8 %), Mus musculus (11.3 %), and D. melanogaster (8.1 %). Other species included are Escherichia coli, C. elegans, A. thaliana, Campylobacter jejuni, etc. MINT and IntAct databases have recently joined their individual efforts to optimize resources as the MIntAct of project. thus avoiding duplication activities [42].

IntAct model has three main components, interactions, interactors, and experiments used to determine interactions. Protein interactions are inferred using scientific publications, including binary interactions or complexes. An interactor can be defined as a biological molecule (mainly a protein) involved in a specific interaction. An interaction is not circumscribed to binary interactions only; it also includes interactions with more partners identified in the experiment performed, e.g. precipitation of multi-protein complexes. Search in IntAct database can be performed in different ways, including name of gene, protein, RNA or chemical compound, or UniProtKB, ChEBI (Chemical Entities of Biological Interest), RNA Central, PMID or IMEx (International Molecular Exchange) IDs. The principal page of IntAct (Fig. 16.30) contains links to other websites the might be of interest. These sites include MINT, UniProtKB, The Swiss Institute of Bioinformatics (SIB), The Interologous Interaction Database (I2D), The Innate Immune Response Database (Innate Database), Molecular Connections, The Extracellular Matrix Interactions Database (MatrixDB), The Modular Approach to Cellular Functions Resource (MB Info), a curated resource for functional analysis of agricultural plant and animal gene products (AgBase), and The cardiovascular Gene Annotation database at the London's Global University (UCL).

As an example of the function of IntAct, we selected the protein XRCC6 (X-ray repair crosscomplementing protein 6, UniProtKB ID P12956), which was found overexpressed in both Luminal A and MDA-MB-231 breast cancer cell lines [18]. This protein is a singlestranded DNA-dependent and ATP-dependent 3'-5' DNA helicase involved in DNA non-homologous end joining (NHEJ) required for double-strand break repair and V(D)J recombination. To reproduce our analysis, in the search window (Fig. 16.30) type XRCC6 or P12956 ID and push the search key. A new window will appear on screen with the results for your query (Fig. 16.31). There are 324 binary interaction found for XRCC6 protein up to date. These interactions are displayed as a table, where molecule A is your query or bait, and B molecules are proteins interacting with your query. For each interaction, a list of interaction methods used for the determination of such interactions is shown, their corresponding IDs, and the source database as well. When you click on the interactors tab, a new page will be shown containing a list of all interactors, showing the type of interactor, the number of interactions described, a link to access the description in UniProtKB, and a description of the interaction (Fig. 16.32). More information, including interactions described, the

EMBL-EBI 🍈				Services	Research Training About us
Int Act					
Home Advanced Search About Resources Download	đ				• Feedback
IntAct Molecular Interaction	on Database			il Data Con	itent
IntAct provides a freely available, open source database curation or direct user submissions and are freely availa	e system and analysis tools for molecular interaction data able. The IntAct Team also produce the Complex Portal $\mathscr D$	a. All interactions are derived from li	terature	<ul> <li>Publications: 1:</li> <li>Interactions: 5:</li> <li>Interactors: 89</li> </ul>	3892 64831 430
Search in IntAct	& Examples	Submission		Vews	W Follow
Enter search term(s) Gearch @ Search Tips	<ul> <li>Gene, Protein, RNA or Chemical name: BRCA2, <u>Staurosponine</u></li> <li>Unihrotik or ChEB1AC: Q06609, CHEB1:15996</li> <li>Unihrotik ID: <u>LCK_HUMAN</u></li> <li>RNMCentral ID: UNSQ006(C954_S59292)</li> <li>PMID: <u>25146956</u></li> <li>IMEX_ID: <u>IN-23318</u></li> </ul>	Submit your data to IntAct to visibility and usability! III Training Online & upcoming courses	o increase its	Helio GCa beene We're teac Coytosci ow:j//Ubi ta Retwee Expand	H Training @Elitraining 18 Nov minidge, Uni atudjetat Today zhing network knatykis using per and @PSICQUIC: Biu Med by InfAct at EBI
Dataset of the month: November A human protein interactome in three quantitati Hein et al. Interactor access Go to Archive Contributors Manually curated content is added to IntAct by curated MINT UniProc. MBinfo AgBase	we dimensions.  Citing IntAct  The MIntAct project-Inf  The MIntAct project-Inf  The MIntAct Project-Inf  The MIntAct Project-Inf  The MintAct Project Inf  The MintAct Pr	tAct as a common curation plat tabases. 4451]∉ [Full Text]∉ IntAct is a member of the IMEx∉ Consortium.	form for 11	EMBL-EE Porta Si porta Si port	I Training d EBmaining 3 Nov mac L project Gemolobi and L project Gemolobi Med Dy IndXA at EBI herein Al Binding dEBmaining 2 Nov al Binding project tells ut a growth completee e.g. profer completee e.g. profer completee e.g. profer completee e.g. brit moto I Training dEBmaining 2 Nov chard Stracks at EBI herein mount de Stracks at CBI herein commission de Stracks at CBI herein project
IntAct View version: 4.2.1	Besarch	Training	Industry		About us
ews.es by topic Brochures By name (A-2) Context us Help & Support Intranet	Verview Publications Research groups Postdocs & PhDs	Train at ESI Train autide ESI Train and ESI Train and ESI Contact organisers	Overview Members Area Workshops <u>SME</u> Forum Contact Industry	programme	Voerview Leadership Funding Background Collaboration Jobs People & groups News Events Visit us

Fig. 16.30 Homepage of the IntAct Molecular Interaction Database

chromosome location in Ensembl webpage, the mRNA expression for interactor in the Expression Atlas webpage, and pathways is displayed when interactors are searched separately. The map of interactions for your query can be displayed in three layouts, force directed (Fig. 16.33), radial (Fig. 16.34) or circle (Fig. 16.35). In all cases, you can zoom in the graph with the tool window at the bottom.

Search can also be performed for a list of identifiers. The result will be more complex as all interactions for each member of your list will be shown. As an example, we only show the graph for ten proteins overexpressed in Luminal A and MDA-MB-231 breast cancer cell lines [18], where a total of 1101 binary interactions were found in database (Figs. 16.36, 16.37 and 16.38).

#### 16.8.5 HPRD

The Human Protein Reference Database (HPRD) is a free web resource containing information of human proteins, including an information summary for each protein, their PTMs, protein-protein interactions, expression levels in tissues, mRNA and protein sequences, non-protein interactions, alternate names, participation in diseases, and domains found in proteins. All the information stored in this database is curated by a group of expert biologists from the Pandey Lab at Johns Hopkins University and the Institute of Bioinformatics in Bangalore, India [43]. The current version of HPRD is 9. It contains information for 30,047 proteins, 41,327 protein-protein interactions, 93,710 PTMs, 112,158

Act :	> In	Act Search R	esults	for and for		B10050		+ Show more data from EMBL-
24	b	inary int	eractions	found for	rsearch	term P12956		
nte	racti	ons (324)	Interactors	Interaction Detail	s Graph			
0	Filt	er out the spo	oke expanded co-	complexes 🕘 (2	06)	Your query also matches <u>1</u> bio Your query also matches <u>2,210</u> in databases. Your query also matches <u>4</u> interact databases.	logical complexes in IntAct. tteraction evidences from 11 other ttion evidences from 1 other IMEx	What is this view?
	Cus	tomize view	Select format	to Download 🗘	Download			
				(	1 of 17)	1 2 3 4 5 6 7 8 9 10	20 0	
	Dts	Molecule 'A'	Links 'A'	Molecule 'B'	Links 'B'	Interaction Detection Method	Interaction AC	Source Database
0	9	XRCC6	P12956 EBI-353208	XRCC5	P13010 EBI-357997	tandem affinity purification	EBI-4370616 imex : IM-16919-1	Molecular Connections
D	9					anti tag coimmunoprecipitation	EBI-11057566 imex : IM-24272-440	IntAct
D	9					anti tag coimmunoprecipitation	EBI-11057764 imex: IM-24272-441	IntAct
D	9					x-ray crystallography	EBI-516722 1JEQ reactome : REACT_3482.1	MINT
•	2					anti bait coimmunoprecipitation	EBI-707473	MINT
•	9					anti bait coimmunoprecipitation	EBI-1563979	IntAct
•	2					anti bait coimmunoprecipitation	EBI-1563986	IntAct
•	2					anti bait coimmunoprecipitation	EBI-1563993	IntAct
•	8					anti bait coimmunoprecipitation	EBI-1563999	IntAct
>	9					electron microscopy	EBI-7162081 MINT-4051838 imex : IM-11282-1	MINT
•	9.					blochemical	EBI-7162105 MINT-4051792 imex : IM-11282-2	MINT
•	9.					anti bait coimmunoprecipitation	EBI-8505439 MINT-8052786 imex : IM-15672-8	MINT
•	9					anti bait coimmunoprecipitation	EBI-1201176 imex: IM-19911-4	IntAct
•	2	XRCC6	P12956 EBI-353208	PRKDC	P78527 EBI-352053	anti bait colmmunoprecipitation	EBI-3956213 imex : IM-16532-16	IntAct
•	8					anti tag coimmunoprecipitation	EBI-11057764 imex: IM-24272-441	IntAct
•	8					anti bait coimmunoprecipitation	EBI-1563993	IntAct
D	9					proximity ligation assay	EBI-3388690 imex: IM-15308-4	12D
D	9					protein kinase assay	EBI-2307851 imex: IM-12076-6	IntAct
D	8					protein kinase assay	EBI-2307862 imex : IM-12076-8	IntAct
D	9	XRCC6	P12956 EBI-353208	WRN	Q14191 EBI-368417	anti tag coimmunoprecipitation	EBI-11057764 imex : IM-24272-441	IntAct

Fig. 16.31 List of binary interactions found for XRCC6 (the X-ray repair cross-complementing protein 6 from *Homo sapiens*, UniProtKB/Swiss-Prot ID P12956) in

sites of protein expression, 22,490 sites of intracellular localization, 470 domains, and 453,521 PMIDs. In addition, two other applications have been recently added, the PhosphoMotif Finder and NetPath resources, which allow the identification of phosphorylation motifs for known kinases/phosphatases and binding motifs for phospho serine/threonine or phospho tyrosine in a compendium of signaling pathways in humans [43]. IntAct database. A total of 324 interactions were found for this protein

To perform a search, click on the Query key, type your query and push the Search button on the upper left part on screen (Fig. 16.39, arrow). There are several options for a query, including Protein Name, Accession Number (RefSeq, GenBank, OMIM, UniProtKB and Entrez Gene Name), HPRD identifier, Gene Symbol, Chromosome locus, Molecular Class (e.g. Nuclease, Serine Proteinase, Translation Regulatory protein, Glycosylase, etc.), PTMs (e.g. ADP

h	inc	ny interact	tione f	ound for	search to	P1204	56	
					Searchite	111 1 1230		
ract	ions	(324) Interact	ors Int	eraction Details	Graph			
rote	eins (	(150) Compour	nds (3)	Nucleic Acids	(26) Genes (4	)		
Acti	on fo	r selection: Sear	ch Interac	tions   Chrom	osome Location	mRNA Expression	Pathways	What is this view
						1 2 3 4 5	678 **	
		Names	Туре	Interactions	Links	Species	Accession	Description
	1	tonsl_human	protein	3	EBI-1052467	human (9606)	EBI-1052467	Tonsoku-like protein
	2	usf1_human	protein	2	EBI-1054489	human (9606)	EBI-1054489	Upstream stimulatory factor 1
	3	cebpa_human	protein	2	EBI-1172054	human (9606)	EBI-1172054	CCAAT/enhancer-binding protein alpha
	4	cebpa_rat	protein	3	EBI-1172084	rat ( <u>10116</u> )	EBI-1172084	CCAAT/enhancer-binding protein alpha
	5	hxb7_human	protein	8	EBI-1248457	human (9606)	EBI-1248457	Homeobox protein Hox-B7
	6	aplf_human	protein	4	EBI-1256044	human (9606)	EBI-1256044	Aprataxin and PNK-like factor
	7	sir1_human	protein	2	EBI-1802965	human (9606)	EBI-1802965	NAD-dependent protein deacetylase sirtuin-1
	8	prkdc_human	protein	6	EBI-352053	human (9606)	EBI-352053	DNA-dependent protein kinase catalytic subunit
	9	xrcc6_human	protein	324	EBI-353208	human (9606)	EBI-353208	X-ray repair cross-complementing protein 6
	10	parp1_human	protein	<u>6</u>	EBI-355676	human (9606)	EBI-355676	Poly [ADP-ribose] polymerase 1
	11	xrcc5_human	protein	13	EBI-357997	human (9606)	EBI-357997	X-ray repair cross-complementing protein 5
	12	wrn_human	protein	8	EBI-368417	human (9606)	EBI-368417	Werner syndrome ATP-dependent helicase
	13	b2y833_human	protein	5	EBI-3952893	human (9606)	EBI-3952893	
	14	myc_human	protein	3	EBI-447544	human ( <u>9606</u> )	EBI-447544	Myc proto-oncogene protein
	15	hd_human	protein	3	EBI-466029	human ( <u>9606</u> )	EBI-466029	Huntingtin
	16	vcam1_human	protein	3	EBI-6189824	human (9606)	EBI-6189824	Vascular cell adhesion protein 1
	17	ppid_human	protein	4	EBI-716596	human (9606)	EBI-716596	Peptidyl-prolyl cis-trans isomerase D
	18	te2ip_human	protein	3	EBI-750109	human (9606)	EBI-750109	Telomeric repeat-binding factor 2-interacting protein 1
	19	tf7l2_human	protein	9	EBI-924724	human (9606)	EBI-924724	Transcription factor 7-like 2
	20	coil_human	protein	3	EBI-945751	human (9606)	EBI-945751	Collin

**Fig. 16.32** List of binary interactions found for XRCC6 (the X-ray repair cross-complementing protein 6 from Homo sapiens, UniProtKB/Swiss-Prot ID P12956) in IntAct database. There are 150 proteins, three chemical compounds (XAV939, 15-deoxy-Delta(12,14)-

prostaglandin J2 and Midostaurin), 26 nucleic acid molecules, and four genes (Klk3, kallikrein-related peptidase 3 encoding gene; Tmps2, Transmembrame protease serine 2). here only a list of 20 protein interactors is shown

Ribosylation, Glycation, Nitration, Sumoylation. Ubiquitination), Cellular Component, Domain Name, Motif, Expression Site, Length of Protein sequence, Molecular Mass, and Diseases (Fig. 16.40). To present an example, we searched NUMA1. Results are shown in Fig. 16.41. Information retrieved includes the name of protein (NUMA1 corresponds to the Nuclear mitotic apparatus protein 1, isoform 1), Molecular Class (Structural protein), Molecular Function (Structural molecule activity), and Biological Process (Cell growth and/or maintenance). Seven additional tabs are provided, which are Summary, Sequence, Interactions, External Links, Alternate Names, Diseases, PTMs, and Substrates. The General tab contains the



Fig. 16.33 Force-directed layout of the interaction map found for XRCC6 in IntAct database. XRCC6 protein is at the center of the map



Fig. 16.34 Radial layout of the interaction map found for XRCC6 in IntAct database. XRCC6 protein query is at the center of the map



Fig. 16.35 Circle layout of the interaction map found for XRCC6 in IntAct database. XRCC6 protein query is located at the *top* of the map

corresponding HPRD ID 01236, Gene symbol NUMA1, Molecular Weight 238259 Da, Chromosome location 11q13, intracellular localization, domains and motifs, and sites of tissue gene expression (Fig. 16.41). The sequence of NUMA1 and its corresponding mRNA are obtained by clicking on Sequence tab (Fig. 16.42). A list of proteins that interact with NUMA1, and types of experiment and interactions (direct or in a complex) are shown in Fig. 16.43.

Alternatively, it is possible to search HPRD by browsing Molecule Class, Domains, Motifs, PTMs, and Localization by pushing the Browse key on the right of the main webpage (Fig. 16.39). Furthermore, access to Human Proteinpedia, Pathways, PhosphoMotif Finder, or downloading the complete HPRD are possible using the main menu.

#### 16.8.6 BioGRID

The Biological General Repository for Interaction Datasets (BioGRID, http://thebiogrid.org), as many other protein-protein interactions databases, has as main goals to curate, organize and make it freely available. The funding partners of this important database are the National Institutes of Health (NIH), the



Fig. 16.36 Interaction map found for PSA3, SYWC, MCM4, SMAP, DDB1, EIF3, PYR1, MCM3, SSRP1 and METK2 proteins in IntAct database. Force directed

layout of the network showing many more interactions that are contained in the IntAct database

Canadian Institutes of Health Research (CIHR), the Genome Canada, and GenomeQuébec. Many other institutions have joined efforts to BioGRID, including the Université de Montréal, Princeton University, Mount Sinai Hospital, University of Edinburgh, SGD, FlyBase, GeneDB, NCBI, WormBase, MaizeGDB, MINT, IntAct, String, MatrixDB, SIB, GO, UniProt, Reactome, Cytoscape, and many others that can be found in the BioGRID webpage. The current version of BioGRID database (3.4.131, December 2015) has information for several model organisms, including *A. thaliana*, *C. elegans, Candida albicans, Danio rerio, Dictyostellium discoideum, D. melanogaster, H. sapiens, Mus musculus, Neurospora crassa, Plasmodium falciparum, S. cerevisiae, Schizosaccharomyces pombe, Xenopus laevis,* 



Fig. 16.37 Radial layout of the network found for PSA3, SYWC, MCM4, SMAP, DDB1, EIF3, PYR1, MCM3, SSRP1 and METK2 proteins in IntAct database

among other eukaryotic organisms. Furthermore, it has information of prokaryotic cells, such as *B. subtilis, E. coli, Mycobacterium tuberculosis, and Streptococcus pneumoniae*. Some viruses are included as well, e.g. Hepatitis C virus, Human Herpesvirus, Human Immunodeficiency virus, and Human Papillomavirus type 16 [44–46]. In its current version, the BioGRID database contains 749,213 non- redundant interactions, corresponding to 63,026 gene products and 45,623 unique publications. BioGRID database also includes 11,329 non-redundant interactions between 4851 unique chemical compounds and 2464 gene products accumulated from 8875 scientific publications. BioGRID also contains PTMs information. A total of 19,981 PTMs corresponding to 18,578 unassigned sites, 3165 unique proteins, 14,999 genes retrieved from 4317 publications are stored in this database.



**Fig. 16.38** Circle layout of the interaction map found for PSA3, SYWC, MCM4, SMAP, DDB1, EIF3, PYR1, MCM3, SSRP1 and METK2 proteins in IntAct database

To perform a search in BioGRID database, type your query (gene name, identifier or keywords) in the gene search window and select the species (Fig. 16.44). It is important to note that only one protein at a time can be searched. Alternatively, searches can be done by PubMed publication. However, searching of Multiple Genes or Publications will be available soon. As an example of a search, we selected the MCM6 protein, which was found overexpressed in both Luminal A and MDA-MB-231 breast cancer cell lines [18]. Results indicates that MCM6, the Minichromosome maintenance complex component 6, is involved in four GO Biological Processes:

- 1. DNA replication
- 2. DNA strand elongation involved in DNA replication
- 3. G1/S transition of mitotic cell cycle
- 4. Mitotic cell cycle

Reference Da	Protein labase	Py Octor	
Corry Course Cou	Nexter       Thuman Proteingedia enables data sharing of human proteins" in February 2008 issue of Nature Biotechnology         Wetchnology       Phospholtotil Finder, published in February 2007 issue of Nature Biotechnology         Wetchnology       Comparison of Protein-Protein Interaction Databases, published in BMC Bioinformatice	Highlights Phospholtotil Einder Allows you to check if your protein contai phosphorylation motif described in the lit Pathways A set of 36 curated signaling pathways ar of a new pathway resource that we have WetPath HPRD Release 9 New The latest Release 9 is available for down Search by PubMed New	ns any ersture e available as part feveloped called load. <u>Click here</u>
Phospholider		Statistics	
Walacula Anthony		Protein Entries	30,047
		Protein-Protein Interactions	41,327
		PTMs	93,710
		Protein Expression	112,158
		Subcellular Localization	22,490
		Domains	470
		PubMed Links	453,521
	About HPRD COMMERCIAL ENTITIES MAY NOT USE THIS SITE WITHOUT PROR LICENSING AUTHORIZATION. PLASE SEND AN <u>EXAM</u> , FOR FURTHER The Human Potein Reference Database represents a certrained pationm to visually depict and integrate information pertaining to domain anythotic add protein the human proteins. Althe Information in PRPD has been manaaly estanded from the Berlance by expert biologists who real, rive patibase in Zope, an open source web application terver; that provides versality in query functions and allows data to be displayed dynamically. Preses of the biologing reference that database. Pasad, T.S. K, et al. (2009) Human Protein Reference Database - 2009 Update. Nucleic Acids Research. 37, DTE-72. (PubMed)	INFORMATION ABOUT LICENSING. Is, post-transitional modifications, interaction networks and deserved and analyze the published data. HPRD has been created usin	ise association for g an object oriented
Carriel	Please send any questions or comments about the Human Protein Reference Database to beil     Councipt & Johns Hopkins University and the Institute of Bioritomatics.	2	
	This is a joint project between:		

Fig. 16.39 Homepage of the Human Protein Reference Database HPRD

Tou are at: (HPRD >> Query			
Query			
The default behavior if more	than one term is e	ntered within th	in a field is 'AND.' e.g. entering 'SH2 SH3' in 'Domain' search field will search for all the proteins that have both SH2 and SH3 domains. Similarly, if mo han one field is filled in, it will be treated as an 'AND' query. For more information go to the FAQ.
Protein Name	NUMA1		
Accession Number	OMIM	0	
HPRD Identifier			
Gene Symbol			
Chromosome Locus			
Molecular Class			See List
PTMs			See List
Cellular Component			See List
Domain Name			See List
Motif			See List
Expression			See List
Length of Protein Sequence	From :	to :	in amino acida
Molecular Weight	From :	to :	in KDa
Diseases			
			Search Clear
	Ple	ase send an	iny questions or comments about the Human Protein Reference Database to belo
		Cop	pyright @ Johns Hopkins University and the Institute of Bioinformatics.
			This is a joint project between:

Fig. 16.40 Query webpage of the Human Protein Reference Database HPRD

Nuclear mitotic apparatus protein 1	
	Wensie Ceen Structural protein
	Structural molecule activity
Isoform 1	terrigite research cell growth and/or maintenance
TT TT TT P(	
	60
ALTERNATE DISEASES PIMETPATER	
SUMMARY   SEQUENCE   INTERACTIONS EXTERNAL LINK	G
L Canada	
A STATE OF	
HPRD ID: 01236	Molecular Weight (Da): 238259
Gene Sympol: NUMAL	Gene map Locus. 11(15
Discologica 1	
Localization	
Primary Nucleus Alternate Nucleo	olus 🕶 Cytosol 🕶 Microtubule 🕶 Mitochondrion 🕶 Cytoplasm 🕶
2 ·	
Demains and Mettle	Fernandan
Domains and Motifs	Expression
Domains and Motifs Domains Mic	otfs Site of Expression
Domains and Motifs Domains CC 1474 - 1699 CC 213 - 637	oths Brain Site of Expression Brain Evolution
Domains         Mc           Domains         CC         1474 - 1699         Mc           CC         213 - 937         CC         927 - 1444	otfis Brain Site of Expression Brain Ecolomia Interfect Parts
Domains and Motifs         Mo           Domains         GC 1472 - 1899           GC 213 - 937         GC 962 - 1444	otife Expression Site of Expression Expressi Expression Expression Expression Expression
Domains and Motifs Bomains Me CC 1474-1099 CC 213-937 CC 962-1444	etifis Expression Site of Expression Ecidophies Ecidoph
Domains         Me           Domains         CC 1474 - 1999           CC 213 - 537         CC 962 - 1444	etifs Brain Site of Expression Brain Epidomia Epidomia Epidomia Exelection Exolocite Leutocht Luno Marmierz pland
Domains and Motifs         Mo           Domains         CC 1474-1099         Mo           CC 273-807         CC 962-1444         Mo	stife Expression Eratio Eratio Expression Ex
Domains         Me           Domains         CC 1474-1099           CC 213-837         CC 962-1444	otife  Expression  Expression  Ecidio/mis  Interior  Levicode  Lun  MammaxyJand Placenta Prostate Salver.gland
Domains         Me           Domains         CC 1474 - 1899           CC 213 - 537         CC 962 - 1444	etifs Brain Site of Expression Existentia Existentia Existentia Existentia Existentia Existentia Elicentia Elicentia Esistentia Salviori, gind Salviori, gind Salviori, gind
Domains and Motifs Domains CC 1474-1099 CC 213-937 CC 962-1444	Expression Expression Expression Exiting Exiti
Domains         Me           Domains         CC 1474 - 1699           CC 213 - 937         CC 962 - 1444	etifs Brain Site of Expression Brain Eckloymis Interface Levicoyte Lunio Mammary Jand Postare Salitance Salita Salitance Salit
Domains and Motifs Bomains Mo CC 1474-1099 CC 213-937 CC 962-1444	etits Expression Site of Expression Eadity Site of Expression Editors Interime Leditors Interime Leditors Interime Leditors Interime Editors Interimated Interime Editors Interime Editors Interime Editors Interi
Domains         Me           CC         1474 - 1099         CC         213 - 137           CC         962 - 1444         CC         962 - 1444	otify  Expression  Expression  Brain Ecidio/mile Intention Levidoode Lun Ammary.gland Placenta Prostate Salivar.gland Sketell muscle Skin Scien Stomath Tastis Throuid gland
Domains and Motifs         Motifs           Domains         CC 1474-1099         Mo           CC 213-807         CC 962-1444         Mo	Expression Expression Entity Entity Entity Entity Expression Expre
Domains         Me           CC         1474 - 1099         CC         213 - 137         Me           CC         902 - 1444         CC         902 - 1444         CC         1444	etife  Expression  Expression  Eadio/mis  Eadio/mis Eadio/mis Eadio/mis Eadio/mis Eadio/mis Eadio/mis Eadio/mis Eadio/mis
Domains and Motifs         Motifs           Domains         CC 1474-1099           CC 273-807         CC 962-1444	etifs Site of Expression  Expression  Expression  Exidentia  Exide
Domains         Mo           CC         1474 - 1099         CC         213 - 637         Mo           CC         002 - 1444         CC         902 - 1444         CC         100 <td>etite  Expression  Expression</td>	etite  Expression

**Fig. 16.41** HPRD query result for the Nuclear Mitotic Apparatus Protein 1, NUMA1. This screenshot shows a putative PTM map as well as a summary for NUMA1

indicating the chromosome localization, subcellular localization, domains, and tissues where the protein is expressed

This protein is also involved in four GO Functions:

- 1. ATP binding
- 2. ATP-dependent DNA helicase activity
- 3. Identical protein binding
- 4. Protein binding

MCM6 is also part of three GO Components:

- 1. MCM complex
- 2. Nucleoplasm
- 3. Nucleus (Fig. 16.45, arrows 1–3)

In order of significance according to the number of physical interactions, MCM6 has 82 interactors which are MCM2, MCM4, MCM7, MCM10, MCMBP, MCM3, CDT1, TONSL, MCM5, HIST1H4A, SSRP1, ASF1B, CDKN2A, ASF1A, MMS22L, and ING5 (Fig. 16.45). When the interactions option is selected, a list of 142 interactions are displayed on screen, indicating the name of interactor, its role in the interaction, name of the species, code for the experimental evidence, source of the dataset, whether interaction is from high or low high throughput screening experiments, a

SUN	MARY SEQU	IENCE INTER	and the second sec												
(P)		SERIEL BRIEFE	ACTIONS L	NKS											
	rotein Sequer	100 2115AA	NP_006176.	2											
	HTLHATRGAA LOKAPVPSTC DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND DLATEKSOND ALIGATIPAK	LLSWYNSLEV SSTIPPELSP RKINOLSEEN KOULSSLITD KEKAAKLEIL RAGRAGLEAR ATSKYVARLE GEDOELAKLR AOVARGROEA LPAKHLCOOL KYEGAKVKVL OAGLKYKEAR TTQIINITHT PQETLSRASH GKAKH	ADPVEAVLQL PSHQAXHEIR GDLSFKLREF LOSSISNLSO OOGLOVANEA LOOIGEANDA TLVNKAGEOO GLEAAOIKEL ERNNGLISSL OAEDAADKK EERORPORER OTCHRITAOV KKLDVEEPDS QPIQIAEGTG	QDCSIFIKII FLELQKVASS ASHLOOLODA AKTELLOASO RDSA075VTO ETTASRELVE FELROTVKOL IETXSELVE HELROTVKOL IETXSILSEO OKLTAOVEOL ANSSFIETRS ITTRQQRKKV	DRIHGTEEGQ SSGNNTLAGG LINELTERISK ANGARLATAOY AOREKALLSR EAMMAORTAE FARAGORGHE KEGLAKKEKE KEGLAKKEKE AAGGLARLLA KYZOREOTEQ DOGLARLARF APASQASLRA SLEPEQGFGT	QILKOPVSER PASPMODILO ATOEMLEKOA ASLISELATL KVEELOACVE SECEOLVKEV MLEDQOROF HASGSGAOSE LXRLVMASSE RAORELOELI VEELSKKLAD QVATDALKSR TSSTOLKSR TSSTOLKSR	LDFVCSFLQK TPOTOMRALK OLEKELSANL NATIOOODE TAROEOHEAO AANRENYEDS CSTOAALOAM AAGRTEPTGP KSOKLEELB SDOASKYOOO EPOANPOLD PROKYADOE SPOASKYOOO EPOANPOLD PRPHTPRDRH	NRKHPSSPEC KOLADERSKR ODKECLEENS LAGLROOANE AGVARLELOL OGEENJGAM EREAROMISKE KLEALRAFYS LLOATLASS KTAOOLRAFK KLEAVOAOGG SIDSLDLAGE SIDSLDLAGE EGRKQSTTEA	LVSAQKVLEG DELELELAEN EILOGKLSOL KOAOLAOTLO RSEOOKATEK FOEDIMITKE LEBLRAALME KLEOOCOFOO ARAALRSSAL ESOOEAOKLO EGTELSITSK RSSARRSONG QKKAAPASTK	SELELANNTH RELLITENDAD ERHLGOLON ODEDASOGLE ERVADEROL SCEKAROELO SOCOOCEERG EDADSLERSL KEIVOSLERSL MEIVOSLERSL MOLNILOADL LPREOPDOTS VSSGAPPGRN QADRROSMAP	LLLYHSTMSS IANGOGRIDG PPORKGEVIG HOVFOLSSSL OCOLOALKES EAXERVAGIE OCEREVARIT EAXEAVAGIE OCENTATION SCIEDARSEN SPYMOTODE SILNTPEKLG	KSPRDWEQFE LALLNEKOAA BVLOLETLKO KOKEOOLKKY LKVTKOSLEE SHSELOISRO DESGRAADAL DESLEVIED KLOMEKAKTH QRLPFKVEL KLOMEKAKTH NELLRRGASK	YKIQAELAVI SPLEPZELEE EAATLAANNT AEEOEATROD IXRRAADALE ONSTLAELSIAN ALEKADARAEL LEEKADELGE RAEELGOELK QAREKYVOEL SLIFTPIPA AELQORNEVC KALSKASPNT	LKFVLDHEDG LRENNESLTM GLORKVENLE HAOGLATAAE EOORCISELK LARALOOVOL SUSALASAOR AHOIKTFOKE AVRADAETE LOEOLRSLEO RSOAPLESCL RSOAPLESCL RSOTRRSPRI	LNLNEDLEN RLHETLEOC TERGOGEAN BELASLÆR AETRSLÆR AETRSLÆR ORVERADLA ELAAFVRADLA LAAFVORAD LOKENKELE DSLGOVFLD SSRPSLSLG ATTTASAAT
D	NA Sequence	Open Rea	ding Frame: 1	73 to 6520	NM_006185.2										
	GCGCTGGCGC	GAGCCCACGA	AGAGGTACGA	TTCCGGAGAA	TCGCGAGGCA	GAGCGGGAGC	GCGCAGCCAG	GTGGAAACTA	ATTCTAAGCC	AGACTGCTGG	AGATCACCCT	GTTCTAGTGT	GTGGAGOCTT	CCACCAGGAG	GCGCATTGG
	GTGACTGTCT	GOCATCACCA	AGATGACACT	COACGCCACC	COOCCOCCTC	CACTOCICTC	TTOOUTGAAC	AGTCTACACO	TOGCTGACCC	TGTGGAGOCT	GTGCTGCAGC	TCCAGGACTG	CAGCATCTTC	ATCAAGATCA	CARCAGAA
	OGAACTOOCG	AAGATGACCA	TOCTOCTOTT	ATACCACTCT	ACCATGAOCT	CCAAAAGTCC	CAGGGACTOG	GAACAGTTTG	ANTATAAAAT	TCAGOCTOAG	TTOOCTUTCA	TTCTTAAATT	TOTOCTOGAC	CATGAGGACG	GOCTAAACC
	TAATGAGGAC	CTAGAGAACT	TCCTACAGAA	AGCTCCTGTG	CCTTCTACCT	GTTCTAGCAC	ATTCCCTGAA	GAGCTCTCCC	CACCTAGCCA	CCAGGCCAAG	AGGGAGATTC	GCTTCCTAGA	GCTACAGAAG	GTTGCCTCCT	CTTCCAGTG
	GAACAACTTT	CTCTCAGGTT	CTCCAGCTTC	TCCCATGGGT	GATATCCTGC	AGACCCCACA	GTTCCAGATG	AGACOGCTGA	AGAAGCAGCT	TOCTGATGAG	AGAAGTAATA	GGGATGAGCT	GGAGCTOGAG	CTAGCTGAGA	ACCGCARGO
	CCTCACCGAG	AAGGATGCAC	AGATAGCCAT	GATOCAGCAG	CGCATTGACC	GCCTAGCCCT	GCTGAATGAG	ANGCAGGCGG	CCAGCCCACT	GGAGCCCAAG	GAGCTTGAGG	AGCTGCGTGA	CAAGAATGAG	AGCCTTACCA	TGCGGCTGC
	TURARCOCTO	ANGCAUTOCC	AGGACCTGAA	GACAGAGAGA	AGCCAGATOG	ATCUCARAAT	CAACCAUCTT	TCOGROGAGOOG	TOGROGACET	TICCITIANS	CTOCOGGADT	TTOCCASTCA	TCTOCAGCAG	CTACAGGATG	COUTCAATG
	CTTGTCCCAG	CTGCAGGATA	ACCCACCCCA	GGAGAAGGOC	GAOGTOCTOG	GTGATGTCTT	GCAGCTOGAA	ACCTTGAAGC	ANGAGGCAGC	CACTCTTOCT	GCAAACAACA	CACAGCTCCA	ACCAGOGTA	GAGATOCTOG	AGACTGAGC
	AGOCCAGCAG	GAAGCCAAGC	TOCTTOCTGA	GCGGGGGCCAC	TTCGANGANG	AAAAOCAOCA	OCTOTOTAGC	CTGATCACTO	ACCTOCAGAG	CTCCATCTCC	AACCTCAGCC	AGOCCAAGGA	AGAGCTOGAG	CAGGCCTCCC	AGOCTCATO
	GOCCCGOTTO	ACTOCCCAGE	TESCCTCTCT	GACCTCTGAG	CTCACCACAC	TCAATGCCAC	CATCCAGCAA	CAGGATCANO	AACTOGCTOG	CCTGAAGCAG	CAGOCCANAG	AGAAGCAGGC	CCAGCTAGCA	CAGACCOTCC	ANCAGCANO
	ACAGGCCTCC	CAGGGCCTCC	GCCACCAGGT	GGAGCAGCTA	AGCAGTAGCC	TGAAGCAGAA	GGNGCNGCNG	TTGANGGNOG	TAGCOGAGAA	GCAGGAGGCA	ACTAGGCAGG	ACCATOCCCA	GCAACTOOCC	ACTOCTOCAG	AGGAOCGAG
	GOCCTCCTTA	AGGGAOCGOG	ATOCOCCTCT	CANGCAGCTG	GAOGCACTOG	AGAAOGAGAA	GOCTOCCANG	CTOGAGATTC	TOCAGCAGCA	ACTTCAGGTG	GCTANTGAAG	CCCOGGACAG	TOCCCAGACC	TCAGTGACAC	AGOCCCAGO
	GOCCCAGGAG	AAGGACCAGC	TCCAGGAGCA	GCTCCAGGCC	CTCAAMGAGT	CCTTGAAGGT	CACCANGOOC	ACCTTGAAG	ACGAGAGACCC	CADGOCTOCA	GATOCCCTOG	ANGAGCAGCA	GCGTTGTATC	TCTGAGCTGA	AGCAGAGA
	CCGAAGCCTG	OTGGAGCAGC	ATAAGCGGGA	ACGAAAGGAG	CTOGANGANG	AGAGOGCTOG	OCOCANGOOG	CTGGAGGCTC	GATTACAGCA	GCTTGGGGAG	GCCCATCAGG	CTGAGACTGA	AGTCCTGCGG	COCGAOCTOG	CAGAGGCCA
	GOCTOCCCAG	CACACAGCTG	AGAGTGAGTG	TGAGCAGCTC	OTCAAAGAAG	TAGCTGCCTG	GCOTGAGCOG	TATGAGGATA	GCCAGCAAGA	GENEGCACAG	TATOGCOCCA	TOTTCCAOGA	ACAGCTGATO	ACTITICAAGG	AGGAATGTG
	GAAGGCCCGC	CAGGAGCTOC	AGGAGOCAAA	GGAGAAGOTG	GCAGGCATAG	ANTCOCACAG	CGAGCTCCAG	ATAMOCCOOC	MCAGAACGA	ACTAGCTGAG	CTCCATGCCA	ACCTGGCCAG	AGCACTCCAG	CAGGTCCAAG	MGANOGANO
	CAGGGCCCAG	AAGCTTOCAG	ATGACCTOTC	CACTCTGCAG	GAAAAGATGG	CTGCCACCAG	CAAAGAGGTG	GCCCGCTTGG	AGACCTTGGT	GCGCAAGGCA	GOTGAGCAGC	AGGAAACAGC	CTCCCGGGAG	TTAGTCAAGG	AGCOTOCGA
	GCAGCAGCAG	GAGGAOCOTG	GGCAGCAGGA	ANDOGNOGTO	GCGCGGCTGA	CCCAGGAGCG	GEGECETOCC	CAGGCTGACC	TTGCCCTGGA	GAAGOCGGCC	MGAOCAGAGC	TTGAGATOCG	GCTGCAGAAC	GCCCTCAACG	AGCAGCGTG
	GGAGTTCOCT	ACCCTOCANG	AGOCACTOOC	TCATGCCCTG	ACCOMANAGE	ANGOCANOGA	CCAGGAGTTG	GCCAMGCTTC	<b>GTGGTCTGGA</b>	GOCAGCCCAG	ATAAAAAAAC	TOGAGGAACT	TCOOCANACC	GTGAAGCAAC	TGAAGGAAC
	GCTGGCTANG	AAAGAAAAGG	AGCACGCATC	TOGCTCAGGA	GCCCANTOTS	AGGCTGCTGG	CAGGACAGAG	CCAACAGGCC	CCAAGCTGGA	GCACTGCGG	GCAGAGGTGA	OCAMOCTOGA	ACAGCAATGC	CAGAAGCAGC	AGGAGCAGG
	TUACAGOCTO	BAACGCAGCC	TCGAGGCTGA	CARGOCCICC	COOCTGACC	ACCOUNTED ACCOUNTED	TETGGAGACT	CIOCAGGGCC	CTTAGAOGA	GAAGGCCCAG	UNGCTAGOOC	ACAOTCAGAG	TUCCTTNOCC .	TUBBCCCAAC	COGAGTTOG
	GANGGAOOGG	GAGAGCAAOG	AGTTGAAGCO	GCTGGTGATG	GCCGAGTCAG	AGAAGAGCCA	GAAGCTOGAG	GAGAOGCTOC	OCCTOCTOCA	OGCAGAGACA	GCCAGCAACA	GTOCCAGAGC	TOCAGAACGC	AGCTCTOCTC	TGCGGGAGG
	GOTGCAGAGC	CTCCGOGAGG	AGOCTGAGAA	ACAOCOOGTO	GCTTCAGAGA	ACCTOCOGCA	GGAGCTGACC	TCACAGGETO	ACCOTOCOGA	GENECTOOOC	CAAGAATTGA	AGOCGTOOCA	GGAGAAOTTC	TTCCAGAAAG	AGCAGOCCC
	CTOCACCCTG	CAGCTOGAOC	ACACCAGCAC	ACAGGCCCTG	GTGAOTGAOC	TOCTOCCAOC	TANGCACCTC	TOCCAGCAGC	TOCAOGCCGA	GCAGGCCGCT	GCCGAGAAAC	GCCACCGTGA	GGAGCTOGAG	CAGAGCAAGC	AGOCCOCTG
	GGGACTOCGG	GCAGAGCTGC	TOCOOCCCA	GCGGGAGCTT	GGGGAOCTGA	TICCTCTGCG	GCAGAAGGTG	GCAGAGCAGO	AGCGAACAGC	TCAGCAGCTG	COGOCAGAGA	AGGCCAGCTA	TOCAGAGCAG	CTGAGCATGC	TGAAGAAGG
	GOTOCAGOGA	GANGCACAGA	GCACTGCCCG	GENOCTOCAG	OTGATGACTG	CCAACTATCA	OGOTOCCANG	GTCANGOTCC	TOCAGGAGAG	OCAGOGGTTC	CAGGAAGAGA	OCCAGABAACT	CACTOCCCAG	OTCOACCACC	TAGACOTAT
	TCAGAGAGAG	CAAACTAAGC	AGOTOGAAGA	ACTGAGTANG	AAACTOGCTO	ACTOTGACCA	AGCCAGCANG	GTOCAGCAGC	AGAAOCTGAA	OCTOTOCAG	GCTCAGOGAG	OCGAGAGCCA	GCAGGAGOCC	CAGCOCCTCC	AGOCCCAGO
	GAATGAACTG	CANGCOCAGE	TGAGCCAGAA	GGAOCAOOCA	GCTGAOCACT	ATAAOCTOCA	GATGGAGAAA	OCCANANCAC	ATTATGATOC	CANGANGCAG	CAGAACCAAG	AGCTGCAGGA	GCAGCTOCOG	AGCCTOGAGC	AGCTOCAGA
	GGAAAACAAA	GAGCTOCGAG	CTGAAGCTGA	ACCOCTOCOC	CATGACCTAC	NOCADOCTOC	GCTGAAGACC	ARGANGECTO	AACAGACCTG	CCGCCACCTT	ACTOCCCAOG	TGCOCAGCCT	GGNOGCACAG	GTTGCCCATG	CAGACCAGC
	TACCCACCAC	CHOCOCAAAT	CONTROCTOR	AGAACCATGCT	TIAAAGAGCC	OTGAGCCCCA	GOCCOCCANG	CAGCTOGACT	TUAGTATIGA	CROCCTOGAT	CTGAOCTOCG	COCCACCOCAC	COCACTCAGT	ATCACCAGCA GACAGCAGCA	ROCKCETC
	DOGAGACOTC	TTCCTOGACT	COOGTOGTAN	GACCOGCTCC	GCTCGTCGGC	GCACCACGCA	GATCATCAAC	ATCACCATGA	CCAAGAAGCT	AGATOTOGAA	GAGCCAGACA	GCGCCAACTC	ATCOTTCTAC	ACCACOCOCT	CTOCTOCTO
	TTCCCAOOCT	AGCCTGCGAG	CCACCTCCTC	TACTCAGTCT	CTAGCTCOCC	TOGGTTCTCC	CGATTATOGC	AACTCAGCCC	TOCTCAGCTT	GCCTOGCTAC	COCCCCACCA	CTCGCAGTTC	TOCTCOTCOT	TCCCAGGCCG	OGOTOTOCA
	TOOGGCCCCT	CCAGGAAGGA	ACAGCTTCTA	CATOGGCACT	TOCCAOGATO	MOCCTGAGCA	OCTOGATOAC	TOGAACCOCA	TTOCMONOCT	GCAGCAGCGC	AATCGAGTGT	OCCCCCCACA	TOTGANGACO	TECTATOCCC	TOGAGTCCA
	GCCTTCCCTG	AGCCTOOGCA	CCATCACAGA	TGAGGAGATG	AAAACTGGAG	ACCCCCAAGA	GACCETGOGC	CGAGCCAGCA	TOCAGCCAAT	CCAGATAGCC	GAGOGCACTG	GCATCACCAC	CCGGCAGCAG	COCANACOGO	TCTCCCTAG
	COORCEACCAG	TOGATOGOAN	TENGENTER	CANCACAGOCC	ANCANCIGTT	TOCCACGCCC GGAACAGCCC	TOTOCOCO	CONGACCOAC	ATGAAGGGCG	CRAACAGAGC	ACTACTGAGG	CTOCCAGAAGAA	ABCAGCTCCA	DOTTOTACTA	TROCCACCO
	CACAGOCAGO	GCCGCCACTG	CTGCCGCCAT	TOGTOCCACC	COTOGAGOCA	AGGGCAAGGC	AAAGCACTAA	AGGGCCAGTA	CCAGTGAGTG	GCCCCACCTG	TGTCCCCGAT	GCTGACCTCA.	CCTGGTCCTC	COCCTACTOT	CCCTCTCAG
	GCCTTCTCTC	AGCTCCCAGG	CCAACAGTAG	CCANACCOCT	AGAGACAGTG	ATGCCTGCCC	GCACCCTGGC	CTGGTCCCTG	GTCCTTCACT	GGCGCCTTCT	COGAGCTOGC	CCAGGGGGGCC	TOGAGCATGO	ACAGTGTGGG	CGCTCTCCC
	ACCTTGCCTC	CTITITCTT	AAAGCAAAGT	CACTTCTCCA	TCACAACCAG	ATTTGAGGCT	GGTTTTGATG	OCTGOGTCCT	TEGECCTECC	CAGTCTTCCT	CTTAGCCTCT	GGATCTAGAA	GGGACCATAA	GAGGAGTAGG	CCCTGGTTC
	TGCTGTCCTG	GTGGCTGGGC	CCAGCAGGGG	CCCTCACTCT	TGAAGTCCAG	GACTOGOTCT	GACCTOSTOS	GAGCACCTGC	CAGAGGATGC	TCTTTCCCAG	GACOGATGOG	CCCTATGTCT	CAGGAGTOGG	GTTGGGGGAC	AGCCTTCAG

Fig. 16.42 Protein and DNA sequences for NUMA1 in HPRD

score for each interaction, the name of the person who curated the information, and additional notes (Fig. 16.46). When the Network tab is selected, three different layouts can be obtained: Concentric circles (Fig. 16.47), Single circle (Fig. 16.48), and Grid (Fig. 16.49). If the number of minimum evidence is changed to five for example, the number of interactions will drop (Fig. 16.50), thus reducing the complexity of the interaction map. When the PTM sites tab is selected, the amino acid sequence of the query is displayed and those residues with an identified PTM are highlighted in blue. Additional information such as the type of modification indicated as well as the source of information are also provided if PTM option is selected (Fig. 16.51). In the case of MCM6, there are 35 Lysine residues marked as ubiquitinated and two additional non-assigned PTMs (neddylation and sumoylation) (Fig. 16.52).

#### 16.8.7 PIPs

The Human Protein-Protein Interaction Prediction (PIPs) is a specialized database containing a

EXTERNAL		
SUMMARY SEQUENCE INTERACTIONS LINKS		
Protein Interactions		
PROTEIN INTERACTORS		
Name of Interactor	Experiment Type	Туре
PMI PMI SMC1 Glona anothed sequence 41 Band 4.1 like protein 1 Tankyrase.1	In Jorden J. Harden, Jonethan Arganization In 1990 - Jin Mahoo In 1990 - Jin Mahoo In 1990 - Jin Mahoo J. Sheeti Z. Hybrid In 1990 - Jin Mahoo J. Sheeti Z. Hybrid In 1990 - Jin Mahoo J. Sheeti Z. Hybrid In 1990 - Jin Mahoo J. Sheeti Z. Hybrid	Direct Direct Direct Direct Direct
Action mellated problem J.A. Nuclear revelopment coactivator. 6 RAD21 LGN protein Entythrocyte membrane protein band. 4.1-like. 2 G. protein. Jointa inhibiting. 1	In JVino In JVino In JVino In JVino In JVino In JVino In JVino	Direct Direct Direct Direct Direct Direct
SMC1 Stromal antigen 1 Stromal antigen 2 Sector 2 SOC112 contein HSP008	in Vivo : in Vito	Complex
T.Comdex.1	in.Veo	Complex
Addim alibita.1 Ubiquita activatios enzyme.1 Ubiquita activatios enzyme.1 Proteascome 2005 subular.0 pro 5 Mitochondrial lacialización ISNA a yothetaane Eukanyoliz anteliation exinopation factor 2 El 253 Laminin receptor.1 Ribonomal archier.Jamp.PD Bibonomal archier.Jamp.PD Bibonomal archier.Jamp.PD Bibonomal archier.JSB Bibonomal archier.JSB Bibonomal archier.JSB Bibonomal archier.JSB Bibonomal archier.JSB Bibonomal archier.JSB Bibonomal archier.JSB		

Fig. 16.43 List of protein interactors of NUMA1 queried in HPRD

catalogue of predicted human protein-protein interactions that have been probabilistically determined using a Bayesian model, which takes into account several modules: Expression, Orthology, Localization, Domain co-occurrence, PTMs co-occurrence, Disorder, and Transitive. Expression considers information from a number of gene expression profiles. Orthology uses the interactions that have been determined for orthologues from fly, human, worm and yeast. Localization is determined by using a human subcellular localization predictor (PSLT) in different subcellular compartments. Domain co-occurrence uses the information stored in InterPro (Protein sequence analysis and classification, http://www.ebi.ac.uk/interpro) and Pfam (Protein families, http://pfam.xfam.org) protein domain databases. PTM co-occurrence uses the information contained in HPRD and UniProtKB. Disorder refers to the prediction of intrinsic disorder of protein found in VLS2 prediction. Finally, Transitive is a module which involves the local topology of networks, considering all modules described above [47].

PIPs database is located at the University of Dundee and the current version (December 2015) contains 37,606 interactions with a score > 1.0, indicating a high probability of occurrence. To





perform a search, an ID in IPI, RefSeq or UniProtKB format must be entered in the search window. As an example, when TBP was used to initiate a query, results were displayed in several boxes each containing a number of interactions with a certain score. In this case, there are 65 interactions when a score value  $\geq 1.0$  was

selected. For score values equal or larger than 2.5, 12.5, 25, 250, and 2500, there were 33, 15, 13, 7, and 3 interactions, respectively. When the number of interactions for a score  $\geq 1.0$  is selected, a list of interactors and the scores for each module used will be displayed on the screen.



Fig. 16.45 Result summary for the Minichromosome Maintenance Complex Component 6, MCM6, queried in BioGRID. A total of 82 interactors were found in database

## 16.8.8 MPIDB

The Microbial Protein Interaction Database (MPIDB) at the Craig Venter Institute (http://jcvi.org/mpidb/about.php) is a database whose main goal is to gather information for all known protein interactions from microbial organisms [48]. The current version of MPIDB is 2009-11-

18 and contains 24,295 interactions that have been experimentally determined for 250 species of bacteria. This number of interactions corresponds to 7810 proteins and 24,295 interactors. Like many other databases, MPIDB also imports information from other databases, including IntAct, Database of Interacting Proteins (DIP), The Biomolecular Interaction

MCM6			Hom	o sapiens	Stats & Options	_			
MCG40308, M	is5, P10	5MCM			<b>Current Statis</b>	tics		Public	ations: 5
minichromos	ome ma	aintenance c	omplex component 6		High Throughput	142 Physical	Interactio	Low T	hroughp
					0 (0%)	0 Genetic In	teractions	0	(100%
UBI NEDD	SUMO				Search Filters	Custom	ize how you	r results are o	lisolayed
GO Proc	ess (4)	G	GO Function (4) GO Compone	ent (3)	No Filter: Show	All Association	5		
OMIM	HGNC	EXTERNAL VEGA   Entrez	DATABASE LINKOUTS Gene   RefSeq   UniprotKB   Ensembl   HPR Nished Interactions For This Protein	D					
Switch View:	Interact	ors (82) In	teractions (142) Network PTM Sit	tes (38)					
Displaying 14	2 total	unique inte	ractions						
Interactor	Role	Organism	Experimental Evidence Code	Dataset		Throughput	Score	Curated By	Notes
ACLY	BAIT	H. sapiens	Co-fractionation	Kristens	en AR (2012)	High		BioGRID	۲
ALS2CR11	HIT	H. sapiens	Two-hybrid 🕀	Rual JF (	(2005)	High	-	BioGRID	-
ASF1A	BAIT	H. sapiens	Affinity Capture-MS 🕀	Huttlin E	L (2014/pre-pub)	High	0.9992	BioGRID	Ð
ASF1B	BAIT	H. sapiens	Affinity Capture-MS 🕀	Huttlin E	L (2014/pre-pub)	High	0.9955	BioGRID	
BAG3	BAIT	H. sapiens	Affinity Capture-MS 🕀	Chen Y (	(2013)	High	-	BioGRID	-
BARD1	BAIT	H. sapiens	Affinity Capture-MS 🕀	Woods N	IT (2012)	High	-	BioGRID	-
CAP1	BAIT	H. sapiens	Co-fractionation	Kristens	en AR (2012)	High	-	BioGRID	۲
CCDC8	BAIT	H. sapiens	Affinity Capture-MS 🕀	Hanson	D (2014)	High	-	BioGRID	-
CCNA1	BAIT	H. sapiens	Reconstituted Complex	Ramach	andran N (2004)	High	-	BioGRID	-
CDC45	BAIT	H. sapiens	Two-hybrid	Kneissl 1	4 (2003)	High		BioGRID	
CDC45	HIT	H. sapiens	Two-hybrid 🕀	Kneissl I	4 (2003)	High	-	BioGRID	-
CDK2	BAIT	H. sapiens	Affinity Capture-MS	Neganov	va I (2011)	High	-	BioGRID	
CUL3	BAIT	H. sapiens	Affinity Capture-MS 🕀	Bennett	EJ (2010)	High	-	BioGRID	
DCTN2	BAIT	H. sapiens	Co-fractionation	Havugim	nana PC (2012)	High	0.764	BioGRID	
ECT2	BAIT	H. sapiens	Affinity Capture-MS 🕀	Woods N	IT (2012)	High	-	BioGRID	
FAM161A	HIT	H. sapiens	Two-hybrid 🕀	Rolland	T (2014)	High	-	BioGRID	-
FN1	BAIT	H. sapiens	Affinity Capture-MS 🕀	Humphri	ies JD (2009)	High	-	BioGRID	1
GMNN	BAIT	H. sapiens	Reconstituted Complex	Ramach	andran N (2004)	High	-	BioGRID	
H2AFX	BAIT	H. sapiens	Affinity Capture-MS 🕀	Huttlin B	L (2014/pre-pub)	High	0.9262	BioGRID	
HIST1H2BA	BAIT	H. sapiens	Affinity Capture-MS ®	Huttlin E	L (2015)	High	0.9894	BioGRID	B

Fig. 16.46 List of interactions found for MCM6 in BioGRID

Network Database (BIND) and MINT. Search can be performed using the name of a protein (UniProtKB ID or locus name) or by selecting species name. Results will be displayed as a table containing the UniProtKB ID, name of protein, interactor, loci of query and interactor, species for query and interactor and the number of evidences for such interaction.

#### 16.8.9 TAIR

The Arabidopsis Information Resource (TAIR) at Phoenix Bioinformatics (https://www. arabidopsis.org) is a database of information for plant research model *A. thaliana*.

This database contains the whole *A. thaliana* genome sequence, analysis, structure and



Fig. 16.47 Map of interactions for MCM6 in BioGRID database. Layout of interaction map is shown in concentric circles, where query protein is at the center

annotation of genes, information for all proteins encoded in its genome, data from gene expression experiments, genome maps, pathways, and other information useful to the scientific community [49]. Like other databases, experts from TAIR curate information using published experiments before entering them in this database. Search in TAIR can be performed in several ways: DNA/Clones, Ecotypes, Genes, Gene Ontology, Plant Ontology, Keywords, Locus, Markers, Microarray element, Microarray expression, People/Labs, Polymorphism/Alleles, Protein, Protocols, PMIDS, Seed/Germplasm, and Text. TAIR webpage also contains tools for



**Fig. 16.48** Map of interactions for MCM6 in BioGRID database. Layout of interaction map is shown as a *single circle*, where MCM6 query protein is located at the *top* of the map

analysis of sequences, as well as viewers for maps and sequences. It is recommended to register in TAIR to download the whole genome sequence.

## 16.8.10 GeneCards

The Human Gene Database (GeneCards, http:// www.genecards.org) is another useful database covering the human genome [50–53]. This database was created by scientists at the Weizmann Institute of Science and LifeMap Sciences. Search can be done using keywords, symbols, aliases, or identifiers. Information that can be retrieved from this database include:

- 1. Aliases for query
- Links to HGNC (HUGO Gene Nomenclature Committee, http://www.genenames. org), Entrez Gene at NCBI, Ensembl (genome databases for vertebrates and other



Fig. 16.49 Grid layout of the map of interactions for MCM6 in BioGRID database. MCM6 query protein is located at the *top left corner* of the map

eukaryotic species, http://www.ensembl.org/ index.html), OMIM http://www.omim.org), and UniProtKB

- 3. Summaries of queries retrieved from different sources
- Genomics data for query, including Regulatory Elements, Genomic location, Genomic region view, and RefSeq DNA sequence
- 5. Protein information such as Protein ID, Length in amino acids, Molecular Mass, Quaternary structure, Three dimensional structure from OCA (Brower-database for protein structure/function, http://oca. weizmann.ac.il/oca-docs/oca-home.html), collaborative Proteopedia (The free, D-encyclopedia of proteins & other



Fig. 16.50 Grid layout of the map of interactions for MCM6 in BioGRID database using a minimum value of 5 as evidence

molecules, http://proteopedia.org/wiki/ index.php/Main\_Page), Alternative splice forms, Data of protein expression in Proteomics DB (https://www.proteomicsdb.org/ proteomicsdb/#overview), PaxDB (Protein Abundance Across Organisms, http://paxdb.org/#!home), MOPED (Multi-Omics Profiling Expression Database, https:// www.proteinspire.org/MOPED/mopedviews/ proteinExpressionDatabase.jsf), MaxQB (The MaxQuant DataBase, http://maxqb. biochem.mpg.de/mxdb/), and PTMs, (6) Domains in InterPro (Protein sequence,

analysis and classification, http://www. ebi.ac.uk/interpro), ProtoNet (Automatic Hierarchical Classification of Proteins, http://www.protonet.cs.huji.ac.il/requested/ cluster\_card.php?global=protonetlnol6l611 lifetimel1l2l2&cluster=4023630&releaseid= 6&firstEnterTimeClient=&blast=11053692l 274977&clusteringNum=61)

 Functions retrieved from UniProtKB, Enzyme Number; Gene Ontology; Phenotypes; Animal models for query; links to CRISPR products, miRNAs, siRNAs, shRNAs, clone products, etc.

MCM6				Homo sapiens	Stats & Optio	ns			
MCG40308, Mi minichromos	is5, P105MC	M nance complex co	mponent 6		Current Stat Total Proteins Total PTM Site	istics with PTMs: 2 s: 38	PTM Pub	lications: 1	
UBI NEDD	SUMO			PTM Site Distribution					
GO Proc	cess (4)	GO Function	(4) GO (	component (3)	Sumoylation	1 Site from 1 Public     1 Site from 2 Public	cations		
ONIN	EXT	TERNAL DATABAS	ELINKOUTS	AN L NOOD	Ubiquitination	- 36 Sites from 14	Publications		
	± Downloa	d 135 Published Intera	ctions For This Protei	in in in it in the second s					
witch View: 🛙	Interactors (	82) Interactions	(142) Network	PTM Sites (38)					
isplaying 2 s ort By: [Def	sets with 38 fault] [Alp	total post transl habetical] [PTM	ational modificati count] [Length	ons ] [ <b>Status</b> ]					
NP_00590	6 licensing fact	or MCM6 : 821						35 [details]	
1 MDL 71 FNQ	AAAAEPG QLSTTIQ THPVHPF	AGSQHLEVRD EEFYRVYPYL LVSGTELCLD	EVAEKCQKLF CRALKTFVKD	LDFLEEFQSS RKEIPLAKDF	DGEIKYLQLA YVAFQDLPTR BNPVCANBBB	EELIRPERNT HKIRELTSSR	LVVSFVDLEQ IGLLTRISGQ		
211 TQA 281 IRG	ELPRGSI	PRSLEVILRA DLSYRLVFLA	EAVESAQAGD	KCDFTGTLIV GGKELRDEEQ	VPDVSKLSTP TAESIKNQMT	GARAETNSRV VKEWEKVFEM	SGVDGYETEG SQDKNLYHNL		
351 CTS 421 GKA 491 ATL	LFPTIHG SSAAGLT NARTSIL	NDEVKRGVLL AAVVRDEESH AAANPISGHY	MLFGGVPKTT EFVIEAGALM DRSKSLKONI	GEGTSLRGDI LADNGVCCID NLSAPIMSRF	NVCIVGDPST EFDKMDVRDQ DLFFILVDEC	AKSQFLKHVE VAIHEAMEQQ NEVTDYAJAR	EFSPRAVYTS TISITKAGVK RIVDLHSRIF		
561 ESI 631 MAR	DRVYSLD MHCCDEV	DIRRYLLFAR QP <b>K</b> HV <b>K</b> EAFR	QFKPKISKES LLNKSIIRVE	EDFIVEQY <b>K</b> H TPDVNLDQEE	LRQRDGSGVT EIQMEVDEGA	KSSWRITVRQ GGINGHADSP	LESMIRLSEA APVNGINGYN		
701 EDI 771 RII	NQESAPK E <b>K</b> VIHRL	ASLRLGFSEY THYDHVLIEL	TQAGLKGSTE	GSESYEEDPY	LVVNPNYLLE	D	DSEEELIN <b>KK</b>		
Location		ртм	31	Residue	Sou	ırce(s)	Note	(s)	
28		Ubiquitination		к	Udeshi	ND (2013)	-		
95		Ubiquitination		к	Stes Udeshi	E (2014) ND (2013)			
102		Ubiquitination		к	Sarraf Udeshi	SA (2013) ND (2013)			
108		Ubiquitination		к	Beltra Sarraf	o P (2012) SA (2013)	:		
					Beltra	o P (2012)	-		
173		Ubiquitination		к	Kim Stes Udeshi	W (2011) E (2014) ND (2013)	-		
					Beltra Kim	o P (2012) W (2011)	-		
197		Ubiquitination		ĸ	Sarraf Stes Wagne	SA (2013) E (2014) r SA (2011)	-		

Fig. 16.51 PTMs reported for MCM6 in BioGRID database. There are a few sites shown to carry ubiquitination for MCM6. Reference is also provided

- 7. Localization of genes in chromosomes and subcellular location of proteins
- 8. Pathways
- 9. Drugs for query
- 10. Transcripts: Reference sequence (RefSeq), Enseml, Unigene Clusters
- Expression in tissues: GeneAnalytics (http:// geneanalytics.genecards.org/?utm\_source= genecards&utm\_medium=banner&utm\_

campaign=genecards&utm\_content=banner\_ expression)

- 12. Orthologs
- 13. Paralogs
- 14. Variants
- 15. Disorders in MalaCards (The Humans Disease Database, http://www.malacards.org)
- 16. Publications

733         Ubiguitination         K         Beltrao P (2012)           Kim W (2011)         Sarraf SA (2013)         Sarraf SA (2013)	-
733 Ubiquitination K Kim W (2011) Sarraf SA (2013)	-
Beltrao P (2012)	
Emanuele MJ (2011)	
744 Ubiquitination K Kim W (2011)	-
Sarraf SA (2013)	
Stes E (2014)	·
Beltrao P (2012)	
769 Ubiquitination K Kim W (2011)	-
Povlsen LK (2012)	-
770 Ubiquitination K Beltrao P (2012)	
Povisen LK (2012)	
Beltrao P (2012)	-
Kim W (2011)	
775 Ubiquitination K Povisen LK (2012)	
Udeshi ND (2013)	
Wagner SA (2011)	
Beltrao P (2012)	
796 Ubiguitination K Emanuele MJ (2011)	
Kim W (2011)	
NON-ASSIGNED PTMS	1 1 1
rost translational modifications that are currently unassigned to a specific residue location.	[details
	-
- Redoylation R Jones J (2008)	•
- Sumoylation K Bruderer R (2011)	13
Danielsen IM (2011)	13
Emanuele MJ (2011)	13
Kim W (2011)	13
- Ubiquitination K Meierhofer D (2008)	13
Zhou J (2011)	13
[show 6 more ⊮]	
RELATIONSHIPS	
Protein Relationship Location PTM Residue Identity Source(s)	1
NEDDB Conjugate - Neddylation K PTM Jones J (200	(8)
SUMO2 Conjugate - Sumovilation K PTM Bruderer R (2	011)
wen D (201	4)
SUM03 Conjugate - Sumoylation K PTM Wen D (201	4)
Danielsen JM (	2011)
Emanuele MJ (	1)
UBC Conjugate - Ubiquitination K PTM Kim W (201	2008)
Zhou J (201	1)
[show 6 more	e •]
Earth a martine and a martine an	
	_
	_

Fig. 16.52 PTMs reported for MCM6 in BioGRID database. Other PTMs are also shown in this figure for MCM6, including neddylation, sumoylation, as well as other ubiquitination sites

In addition, there are a lot of links to companies that might have products for the protein of interests, such as antibodies, immunofluorescence, animal models, silencing, etc.

Acknowledgements We thank the Instituto de Ciencia y Tecnología del Distrito Federal (ICyTDF), now renamed Secretaría de Ciencia, Tecnología e Innovación de la Ciudad de México (SECITI), for its support with the project ICyTDF-J.LA (CM-272/12-SECITI/033/2012),

and Consejo Nacional de Ciencia y Tecnología (Conacyt) from Mexico, with the project number SALUD-2009-01-113674, both granted to Dr. Juan Pedro Luna Arias.

#### References

 Kumar C, Mann M (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Lett 583(11):1703–1712

- Su Z, Wang J, Yu J, Huang X, Gu X (2006) Evolution of alternative splicing after gene duplication. Genome Res 16(2):182–189
- Twyman RM (2004) Principles of proteomics. Garland Biosciences/BIOS Scientific Publishers, Hampshire
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29
- Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11(8):1425–1433
- Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C et al (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue):D258–D261
- Gene Ontology C (2015) Gene ontology consortium: going forward. Nucleic Acids Res 43(Database issue): D1049–D1056
- Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9(7):509–515
- Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 8 (8):1551–1566
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13(9):2129–2141
- 11. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky- Lazareva B, Muruganujan A, Rabkin S et al (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res 31 (1):334–341
- 12. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ et al (2005) The PAN-THER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33(Database issue):D284–D288
- Funahashi A, Jouraku A, Matsuoka Y, Morohashi M, Kikuchi N, Kitano H (2008) CellDesigner 3.5: a versatile modeling tool for biochemical networks. Proc IEEE 96(8):1254
- Mi H, Guo N, Kejariwal A, Thomas PD (2007) PAN-THER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res 35(Database issue): D247–D252
- Mi H, Thomas P (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol 563:123–140

- PANTHER User Manual (2015). http://pantherdb.org/ help/PANTHER\_user\_manual.pdf
- 17. Mi H, Muruganujan A, Thomas PD (2013) PAN-THER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41(Database issue):D377–D386
- 18. Calderon-Gonzalez KG, Valero Rustarazo ML, Labra-Barrios ML, Bazan-Mendez CI, Tavera-Tapia-A, Herrera-Aguirre M, Sanchez Del Pino MM, Gallegos-Perez JL, Gonzalez- Marquez H, Hernandez-Hernandez JM et al (2015) Data set of the protein expression profiles of Luminal A, Claudin-low and overexpressing HER2(+) breast cancer cell lines by iTRAQ labelling and tandem mass spectrometry. Data Brief 4:292–301
- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4(5):P3
- 20. da Huang W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13
- 21. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC et al (2007) DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35(Web Server issue): W169–W175
- 22. da Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57
- 23. da Huang W, Sherman BT, Stephens R, Baseler MW, Lane HC, Lempicki RA (2008) DAVID gene ID conversion tool. Bioinformation 2(10):428–430
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32(Database issue): D277–D280
- 25. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34(Database issue):D354–D357
- 26. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2015) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:457
- Kanehisa M, Sato Y, Morishima K (2015) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726
- 28. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(Database issue):D109–D114

- 29. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 36(Web Server issue): W423–W426
- 30. Chaiboonchoe A, Samarasinghe S, Kulasiri D, Salehi-Ashtiani K (2014) Integrated analysis of gene network in childhood leukemia from microarray and pathway databases. BioMed Res Int 2014:278748
- 31. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35 (Database issue):D358–D362
- 32. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33(Database issue):D433–D437
- 33. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M et al (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37(Database issue):D412–D416
- 34. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31(1):258–261
- Harrington ED, Jensen LJ, Bork P (2008) Predicting biological networks from genomic data. FEBS Lett 582(8):1251–1258
- Marcotte EM, Xenarios I, Eisenberg D (2001) Mining literature for protein-protein interactions. Bioinformatics 17(4):359–363
- 37. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39(Database issue):D561–D568
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a molecular INTeraction database. FEBS Lett 513(1):135–140
- 39. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E et al (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40(Database issue):D857–D861
- 40. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A et al (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32(Database issue):D452–D455
- 41. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M,

Friedrichsen A, Huntley R et al (2007) IntAct-open source resource for molecular interaction data. Nucleic Acids Res 35(Database issue):D561–D565

- 42. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N et al (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(Database issue):D358–D363
- 43. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A et al (2009) Human protein reference database--2009 update. Nucleic Acids Res 37(Database issue):D767– D772
- 44. Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the general repository for interaction datasets. Genome Biol 4(3):R23
- 45. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34(Database issue):D535–D539
- 46. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L et al (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43(Database issue):D470–D478
- Scott MS, Barton GJ (2007) Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinf 8:239
- Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P (2008) MPIDB: the microbial protein interaction database. Bioinformatics 24(15):1743–1744
- 49. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40 (Database issue):D1202–D1210
- Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13(4):163
- 51. Safran MC-CV, Shmueli O, Rosen N, Benjamin-Rodrig H, Ophir R, Yanai I, Shmoish M, Lancet D (2003) The GeneCards family of databases: GeneCards, GeneLoc, GeneNote and GeneAnnot. In: Proceedings of the IEEE Computer Science Bioinformatics Conference CSB2003
- 52. Stelzer GHA, Dalah A, Rosen N, Shmoish M, Iny Stein T, Sirota A, Madi A, Safran M, Lancet D (2008) GeneCards: one stop site for human gene research. FISEB (ILANIT)
- Harel A, Inger A, Stelzer G, Strichman-Almashanu L, Dalah I, Safran M, Lancet D (2009) GIFtS: annotation landscape analysis with GeneCards. BMC Bioinf 10:348