BMC Genomics

DATABASE Open Access

IntegromeDB: an integrated system and biological search engine

Michael Baitaluk[1*], Sergey Kozhenkov[1], Yulia Dubinina[1] and Julia Ponomarenko[1,2]

Abstract

Background: With the growth of biological data in volume and heterogeneity, web search engines become key tools for researchers. However, general-purpose search engines are not specialized for the search of biological data.

Description: Here, we present an approach at developing a biological web search engine based on the Semantic Web technologies and demonstrate its implementation for retrieving gene- and protein-centered knowledge. The engine is available at http://www.integromedb.org.

Conclusions: The IntegromeDB search engine allows scanning data on gene regulation, gene expression, protein-protein interactions, pathways, metagenomics, mutations, diseases, and other gene- and protein-related data that are automatically retrieved from publicly available databases and web pages using biological ontologies. To perfect the resource design and usability, we welcome and encourage community feedback.

Keywords: data integration, search engine, biological ontologies

Background

Diverse web resources and databases, such as UniProt [1], GenBank [2], GeneCards [3], RGD [4], MGD [5], and many others, catalog various properties of genes, ranging from their mapped chromosomal coordinates to the enzymatic function of the proteins they encode. Scientists have to visit many of these databases or web sites for each gene in their candidate gene list; they must learn and remember how to navigate various web sites, each of which accepts different sets of gene identifiers (Entrez Gene, Ensembl, Refseq, UniGene, and other), thus making the navigation difficult and time-consuming. Although these resources are highly informative individually, the collection of available content has more power, if provided in an integrated, unified, centralized context indexed in a robust manner. Although currently there is no single resource that completely describes everything that a researcher might want to learn about a specific gene, a few integrative approaches towards this goal have been developed and include BioGPS [6], Ondex [7], NIF [8], and several databases for cataloging web resources, such as PathGuide [9] and MouseBook [10]. The integrated and comprehensive use of biological information is hampered by the large number of available databases and their fragmentation. ID/naming problems and conflicts in biological data pose additional obstacles towards comprehensive biological data integration. Because of the diverse history of these databases and resources, integration with commonly used molecular database resources, such as NCBI's Entrez [11], is done on a case-by-case basis. At the same time, new online resources are continually being developed and staying abreast of these tools and evaluating their utility is a time-consuming and recurring task.

Web search engines, such as Google, excel at finding a needle in a haystack: a single fact, a single definitive web page. Often, however, the user's objective is not to find a single fact, but to explore all available data concerning a specific genes or proteins. For example, a person diagnosed with diabetes might want to learn all known scientific data about genes associated with this disease. Or, for a researcher who wants to learn 'all about p53 gene/protein', it is much easier to get the information centralized in one place, instead of searching the whole Internet for this information. Even if the researcher knows exactly what to look for, for example, specific post-translational modifications in p53 protein that result in different types of cancer, he/she still might need to browse many different web resources in search of this information.

* Correspondence: baitaluk@sdsc.edu
[1]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA
Full list of author information is available at the end of the article

Here, we propose a solution to the problem of searching biological data, specifically gene/protein-centered data, via semantic integration of data from a broad collection of molecular biology resources and presentation of these data on one web page for each searchable gene/protein. Data have been automatically integrated from more than a thousand databases listed in the Nucleic Acid Research database list [12], as well as from PubMed, Wikipedia, and millions of public web sites that are referenced from these databases. The presented search engine, http://www.integromedb.org, has been designed for browsing available data on gene regulation, gene expression, protein-protein interactions, pathways, metagenomics, mutations, diseases, and many other gene- and protein-related data. IntegromeDB web page is linked to BiologicalNetworks http://www.biologicalnetworks.org [13], which is a Java web-start application designed for in-depth analysis of biological data, building and visualization of gene regulatory and protein-protein interaction networks; the presented web resource and BiologicalNetworks both rely on the same integrated database [14]. Application Programming Interface (API) available at http://integromedb.org/api.jsp provides access to the most features available through the web application and can be used by third-party users and resources. The batch search and retrieval options are also provided at the IntegromeDB web site, http://www.integromedb.org.

Construction and content

Data Integration

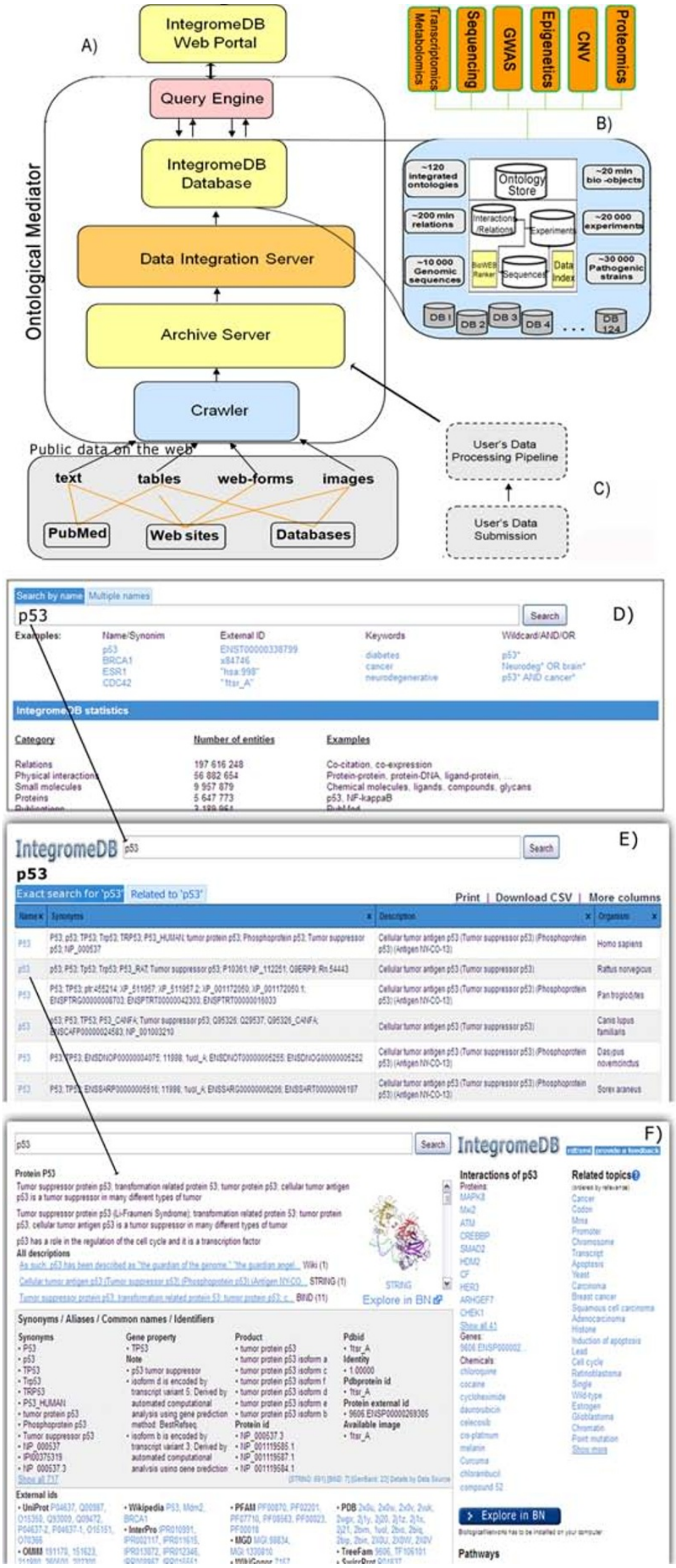
The basic anatomy of the IntegromeDB data integration system that is behind the presented search engine is described in detail in [14]. Briefly, IntegromeDB is a semantic, graph based, 'deep web' data integration system that automatically captures, integrates, and manages publicly available data related to functioning of genes and proteins (Figure 1; see also Figure S1 in Additional file 1). The system's internal database is structured as a metagraph database that contains the features attributable to a given bio-entity, such as gene, protein, pathway, operon, experiment, COG, and other. The system dynamically incorporates new sets of objects and their relations and integrates the following four complex orthogonal data types: Graphs, Tables/Histograms, Trees, and Sequences [14,15]. The problems of data integration are addressed via ontology-driven data mapping, multiple data annotation and heterogeneous data querying, also enabling integration of the user's data. The internal database schema is RDF-compatible [16]; that is, it stores biological data in the RDF-compatible format, the standard format of the Semantic Web [17]. The IntegromeDB web site supports RDF format by providing export of the Search Results web page in RDF format (at the top of the page; Figure 2) and also providing download of some parts of the database that describe organisms, pathways, synonyms, and interactions as the RDF formatted files http://integromedb.org/download_rdf.jsp.

The major architectural blocks of the system are a web-crawler, archive server, data integration server, internal database, and ontological mediator (Figures 1 and 1S1). The crawler automatically searches public databases, PubMed, and web resources for biological data and extracts them from texts and tables. The system accepts the user's data that can be integrated together with public data. The Data Integration Server accepts external data, semantically mapping them to the schema of the Internal Database, and injects the external data into the database. This allows IntegromeDB to tap into the Deep Web, the portion of the web that is concealed behind web forms. Some estimates have pegged the size of the Deep Web at up to 500 times larger than the Surface Web [18]. Briefly, the procedure to process the data behind web forms works as follows. Starting from an initial prediction of candidate keywords by simple ontological analysis of the text on the page (usually it is ether gene, protein, drug, pathway names), the form is tested and if it returns valid results, the common URL pattern of meaningful pages (for example 'http://www.ncbi.nlm.nih.gov/gene/?term=') is analyzed and further keywords will be extracted from the resulting pages. The iterative process is continued until either no new candidate keywords can be extracted or no more new web pages can be retrieved for a given resource.

The IntegromeDB content is available for more than 100,000 genes and proteins from 1252 organisms, more than 2 million experiments, 1000 pathways, 900,000 gene and protein records, and thousands of PubChem [19] records. Since many of the integrated resources are frequently updated, the IntegromeDB database is re-loaded on a quarterly basis. To ensure that the connections between external databases within IntegromeDB are as accurate as possible, an automated process checks for updates of outside databases, downloads files, and populates database tables.

Querying IntegromeDB

Search on the IntegromeDB web page can be executed via keyword, wild card, or multiple word queries. Genes/proteins can be queried using either the gene name (the official HGNC [20] name, Entrez Gene [11], Ensembl [21]), Affymetrix identifier, current UniGene/UniProt [1] cluster identifier, or GenBank [2] accession number of a sequence (Refseq ID) associated with the gene through UniGene/UniProt (Figure 1D). We have assembled the largest collection of gene aliases available on the web by combining synonymous data from more than a thousand databases. IntegromeDB can therefore accept practically any known ID as input. To display data for a specific



Search by name | Multiple names

p53 | Search

Examples:	Name/Synonim	External ID	Keywords	Wildcard/AND/OR
	p53 BRCA1 ESR1 CDC42	ENST00000338799 x84746 "hsa:998" "1tsr_A"	diabetes cancer neurodegenerative	p53* Neurodeg* OR brain* p53* AND cancer*

IntegromeDB statistics

Category	Number of entities	Examples
Relations	197 616 248	Co-citation, co-expression
Physical interactions	56 882 654	Protein-protein, protein-DNA, ligand-protein, ...
Small molecules	9 957 879	Chemical molecules, ligands, compounds, glycans
Proteins	5 647 773	p53, NF-kappaB
Publications	3 189 964	PubMed

IntegromeDB p53 | Search

p53

Exact search for 'p53' | Related to 'p53' — Print | Download CSV | More columns

Name	Synonyms	Description	Organism
P53	P53; p53; TP53; Trp53; TRP53; P53_HUMAN; tumor protein p53; Phosphoprotein p53; Tumor suppressor p53; NP_000537	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13)	Homo sapiens
p53	p53; P53; Tp53; Trp53; P53_RAT; Tumor suppressor p53; P10361; NP_112251; Q9ERP9; Rn.54443	Cellular tumor antigen p53 (Tumor suppressor p53)	Rattus norvegicus
P53	P53; TP53; ptr:455214; XP_511957; XP_511957.2; XP_001172060; XP_001172060.1; ENSPTRG00000008703; ENSPTRT00000042303; ENSPTRT00000016033	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13)	Pan troglodytes
p53	p53; P53; TP53; P53_CANFA; Tumor suppressor p53; Q95326; Q29537; Q95326_CANFA; ENSCAFP00000024583; NP_001003210	Cellular tumor antigen p53 (Tumor suppressor p53)	Canis lupus familiaris
P53	P53; TP53; ENSDNOP00000004075; 11998, 1uol_A; ENSDNOT00000005255; ENSDNOG00000005252	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13)	Dasypus novemcinctus
P53	P53; TP53; ENSSARP00000005516; 11998, 1uol_A; ENSSARG00000006206; ENSSART00000006187	Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13)	Sorex araneus

p53 | Search — IntegromeDB

Protein P53

Tumor suppressor protein p53; transformation related protein 53; tumor protein p53; cellular tumor antigen p53 is a tumor suppressor in many different types of tumor

Tumor suppressor protein p53 (Li-Fraumeni Syndrome); transformation related protein 53; tumor protein p53; cellular tumor antigen p53 is a tumor suppressor in many different types of tumor

p53 has a role in the regulation of the cell cycle and it is a transcription factor

All descriptions

- As such, p53 has been described as "the guardian of the genome," "the guardian angel... Wiki (1)
- Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO... STRING (1)
- Tumor suppressor protein p53; transformation related protein 53; tumor protein p53; c... BIND (11)

STRING — Explore in BN

Synonyms / Aliases / Common names / Identifiers

Synonyms
- P53
- p53
- TP53
- Trp53
- TRP53
- P53_HUMAN
- tumor protein p53
- Phosphoprotein p53
- Tumor suppressor p53
- NP_000537
- IPI00375319
- NP_000537.3

Show all 717

Gene property
- TP53

Note
- p53 tumor suppressor
- isoform d is encoded by transcript variant 5. Derived by automated computational analysis using gene prediction method: BestRefseq
- isoform b is encoded by transcript variant 3. Derived by automated computational analysis using gene prediction

Product
- tumor protein p53
- tumor protein p53 isoform a
- tumor protein p53 isoform c
- tumor protein p53 isoform f
- tumor protein p53 isoform d
- tumor protein p53 isoform e
- tumor protein p53 isoform b

Protein id
- NP_000537.3
- NP_001119585.1
- NP_001119587.1
- NP_001119584.1

Pdbid
- 1tsr_A

Identity
- 1.00000

Pdbprotein id
- 1tsr_A

Protein external id
- 9606.ENSP00000269305

Available image
- 1tsr_A

[STRING: 691] [BIND: 7] [GenBank: 22] Details by Data Source

External ids
- UniProt P04637, Q00987, O15350, Q93009, Q09472, P04637-2, P04637-1, O15151, O70366
- OMIM 191170, 151623,
- Wikipedia P53, Mdm2, BRCA1
- InterPro IPR010991, IPR002117, IPR011615, IPR013872, IPR012346,
- PFAM PF00870, PF02201, PF07710, PF08563, PF00023, PF00018
- MGD MGI:98834, MGI:1330810
- PDB 2x0u, 2x0w, 2x0v, 2wxk, 2wgx, 2j1y, 2j20, 2j1z, 2j1x, 2j21, 2bim, 1uol, 2bio, 2biq, 2bip, 2bin, 2X0U, 2X0W, 2X0V
- TreeFam 9606, TF106101

Interactions of p53

Proteins: MAPK8, Mdm2, ATM, CREBBP, SMAD2, HDM2, CF, HER3, ARHGEF7, CHEK1, Show all 41

Genes: 9606.ENSP000002...

Chemicals: chloroquine, cocaine, cycloheximide, daunorubicin, celecoxib, cis-platinum, melanin, Curcuma, chlorambucil, compound 52

Related topics (ordered by relevance)

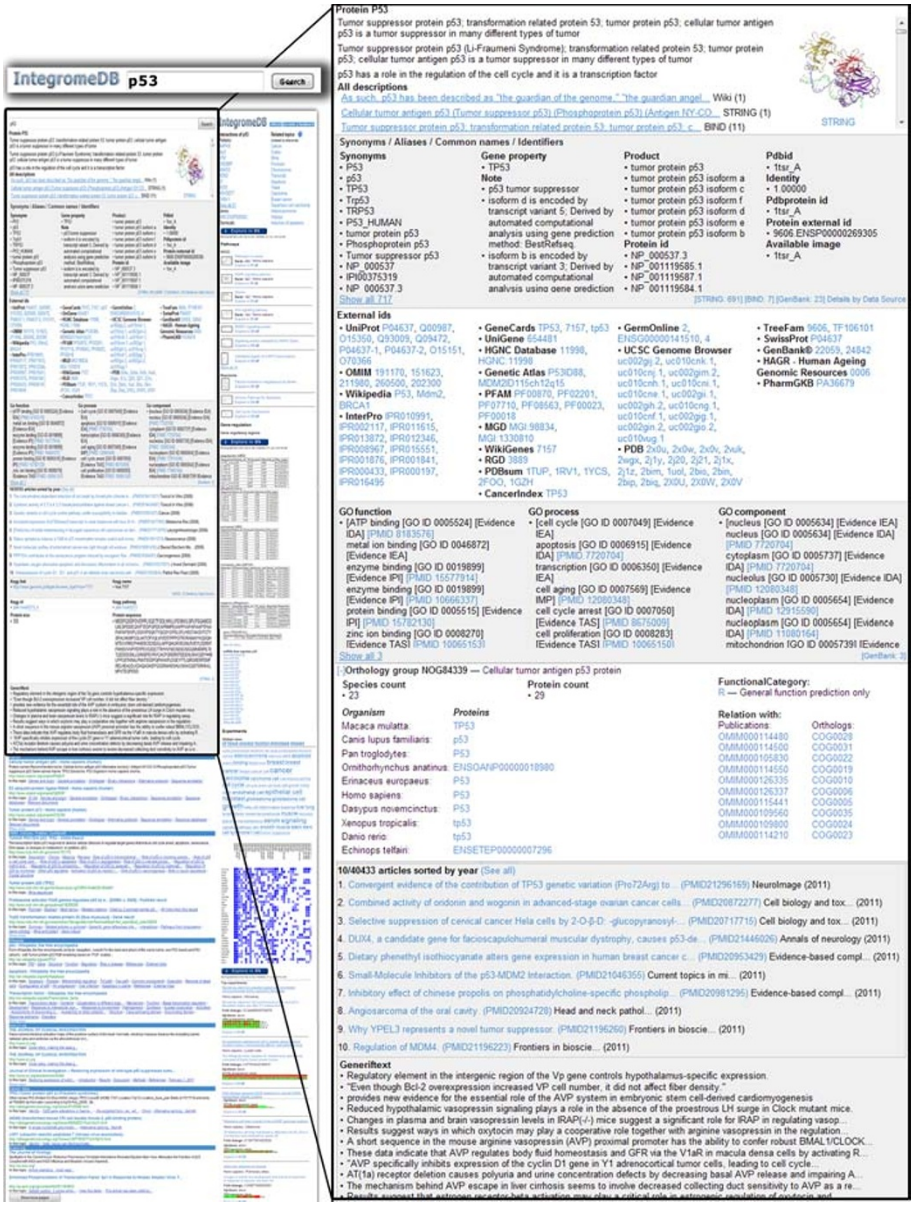
Cancer, Codon, Mma, Promoter, Chromosome, Transcript, Apoptosis, Yeast, Carcinoma, Breast cancer, Squamous cell carcinoma, Adenocarcinoma, Histone, Induction of apoptosis, Lead, Cell cycle, Retinoblastoma, Single, Wild-type, Estrogen, Glioblastoma, Chromatin, Point mutation, Show more

Explore in BN

BiologicalNetworks has to be installed on your computer

Pathways

Figure 1 IntegromeDB architecture. A) Major architectural blocks of the IntegromeDB system. **B)** Internal database of the IntegromeDB; described in detail in [14]. **C)** The pipeline processing the user's data. **D)** IntegromeDB web page with search examples and search options. For a batch search, input can be uploaded as a text file or by pasting the gene/protein list into a text box. **E)** Search results sorted by relevance score. Clicking the gene/protein name takes the user to the respective gene/protein report page **(F)** (see Figure 2).



Protein P53

Tumor suppressor protein p53; transformation related protein 53; tumor protein p53; cellular tumor antigen p53 is a tumor suppressor in many different types of tumor

Tumor suppressor protein p53 (Li-Fraumeni Syndrome); transformation related protein 53; tumor protein p53; cellular tumor antigen p53 is a tumor suppressor in many different types of tumor

p53 has a role in the regulation of the cell cycle and it is a transcription factor

All descriptions

As such, p53 has been described as "the guardian of the genome," "the guardian angel... Wiki (1)
Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO... STRING (1)
Tumor suppressor protein p53; transformation related protein 53; tumor protein p53; c... BIND (11)

Synonyms / Aliases / Common names / Identifiers

Synonyms
- P53
- p53
- TP53
- Trp53
- TRP53
- P53_HUMAN
- tumor protein p53
- Phosphoprotein p53
- Tumor suppressor p53
- NP_000537
- IPI00375319
- NP_000537.3

Show all 717

Gene property
- TP53

Note
- p53 tumor suppressor
- isoform d is encoded by transcript variant 5; Derived by automated computational analysis using gene prediction method: BestRefseq.
- isoform b is encoded by transcript variant 3; Derived by automated computational analysis using gene prediction

Product
- tumor protein p53
- tumor protein p53 isoform a
- tumor protein p53 isoform c
- tumor protein p53 isoform f
- tumor protein p53 isoform d
- tumor protein p53 isoform e
- tumor protein p53 isoform b

Protein id
- NP_000537.3
- NP_001119585.1
- NP_001119587.1
- NP_001119584.1

Pdbid
- 1tsr_A

Identity
- 1.00000

Pdbprotein id
- 1tsr_A

Protein external id
- 9606.ENSP00000269305

Available image
- 1tsr_A

[STRING: 691] [BIND: 7] [GenBank: 23] Details by Data Source

External ids
- **UniProt** P04637, Q00987, O15350, Q93009, Q09472, P04637-1, P04637-2, O15151, O70366
- **OMIM** 191170, 151623, 211980, 260500, 202300
- **Wikipedia** P53, Mdm2, BRCA1
- **InterPro** IPR010991, IPR002117, IPR011615, IPR013872, IPR012346, IPR008967, IPR015551, IPR001876, IPR001841, IPR000433, IPR000197, IPR016495
- **GeneCards** TP53, 7157, tp53
- **UniGene** 654481
- **HGNC Database** 11998, HGNC:11998
- **Genetic Atlas** P53ID88, MDM2ID115ch12q15
- **PFAM** PF00870, PF02201, PF07710, PF08563, PF00023, PF00018
- **MGD** MGI:98834, MGI:1330810
- **WikiGenes** 7157
- **RGD** 3889
- **PDBsum** 1TUP, 1RV1, 1YCS, 2FOO, 1GZH
- **CancerIndex** TP53
- **GermOnline** 2, ENSG00000141510, 4
- **UCSC Genome Browser** uc002gij.2, uc010cnk.1, uc010cnj.1, uc002gim.2, uc010cnh.1, uc010cni.1, uc010cne.1, uc002gii.1, uc002gih.2, uc010cng.1, uc010cnf.1, uc002gig.1, uc002gin.2, uc002gio.2, uc010vug.1
- **PDB** 2x0u, 2x0w, 2x0v, 2vuk, 2wgx, 2j1y, 2j20, 2j21, 2j1x, 2j1z, 2bim, 1uol, 2bio, 2bin, 2bip, 2biq, 2X0U, 2X0W, 2X0V
- **TreeFam** 9606, TF106101
- **SwissProt** P04637
- **GenBank®** 22059, 24842
- **HAGR - Human Ageing Genomic Resources** 0006
- **PharmGKB** PA36679

GO function
- [ATP binding [GO ID 0005524] [Evidence IDA] [PMID 8183576]
  metal ion binding [GO ID 0046872] [Evidence IEA]
  enzyme binding [GO ID 0019899] [Evidence IPI] [PMID 15577914]
  enzyme binding [GO ID 0019899] [Evidence IPI] [PMID 10666337]
  protein binding [GO ID 0005515] [Evidence IPI] [PMID 15782130]
  zinc ion binding [GO ID 0008270] [Evidence TAS] [PMID 10065153]

Show all 3

GO process
- [cell cycle [GO ID 0007049] [Evidence IEA]
  apoptosis [GO ID 0006915] [Evidence IDA] [PMID 7720704]
  transcription [GO ID 0006350] [Evidence IEA]
  cell aging [GO ID 0007569] [Evidence IMP] [PMID 12080348]
  cell cycle arrest [GO ID 0007050] [Evidence TAS] [PMID 8675009]
  cell proliferation [GO ID 0008283] [Evidence TAS] [PMID 10065150]

GO component
- [nucleus [GO ID 0005634] [Evidence IEA]
  nucleus [GO ID 0005634] [Evidence IDA] [PMID 7720704]
  cytoplasm [GO ID 0005737] [Evidence IDA] [PMID 7720704]
  nucleolus [GO ID 0005730] [Evidence IDA] [PMID 12080348]
  nucleoplasm [GO ID 0005654] [Evidence IDA] [PMID 12915590]
  nucleoplasm [GO ID 0005654] [Evidence IDA] [PMID 11080164]
  mitochondrion [GO ID 0005739] [Evidence

[GenBank: 3]

[-]Orthology group NOG84339 — Cellular tumor antigen p53 protein

Species count
- 23

Protein count
- 29

FunctionalCategory:
R — General function prediction only

Organism	*Proteins*
Macaca mulatta:	TP53
Canis lupus familiaris:	p53
Pan troglodytes:	P53
Ornithorhynchus anatinus:	ENSOANP00000018980
Erinaceus europaeus:	P53
Homo sapiens:	P53
Dasypus novemcinctus:	P53
Xenopus tropicalis:	tp53
Danio rerio:	tp53
Echinops telfairi:	ENSETEP00000007296

Relation with:

Publications:	Orthologs:
OMIM000114480	COG0028
OMIM000114500	COG0031
OMIM000105830	COG0022
OMIM000114550	COG0019
OMIM000126335	COG0010
OMIM000126337	COG0006
OMIM000115441	COG0005
OMIM000109560	COG0035
OMIM000109800	COG0024
OMIM000114210	COG0023

10/40433 articles sorted by year (See all)

1. Convergent evidence of the contribution of TP53 genetic variation (Pro72Arg) to... (PMID21296169) NeuroImage (2011)
2. Combined activity of oridonin and wogonin in advanced-stage ovarian cancer cells... (PMID20872277) Cell biology and tox... (2011)
3. Selective suppression of cervical cancer Hela cells by 2-O-β-D: -glucopyranosyl-... (PMID20717715) Cell biology and tox... (2011)
4. DUX4, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-de... (PMID21446026) Annals of neurology (2011)
5. Dietary phenethyl isothiocyanate alters gene expression in human breast cancer c... (PMID20953429) Evidence-based compl... (2011)
6. Small-Molecule Inhibitors of the p53-MDM2 Interaction. (PMID21046355) Current topics in mi... (2011)
7. Inhibitory effect of chinese propolis on phosphatidylcholine-specific phospholip... (PMID20981295) Evidence-based compl... (2011)
8. Angiosarcoma of the oral cavity. (PMID20924728) Head and neck pathol... (2011)
9. Why YPEL3 represents a novel tumor suppressor. (PMID21196260) Frontiers in bioscie... (2011)
10. Regulation of MDM4. (PMID21196223) Frontiers in bioscie... (2011)

Generiftext
- Regulatory element in the intergenic region of the Vp gene controls hypothalamus-specific expression.
- "Even though Bcl-2 overexpression increased VP cell number, it did not affect fiber density."
- provides new evidence for the essential role of the AVP system in embryonic stem cell-derived cardiomyogenesis
- Reduced hypothalamic vasopressin signaling plays a role in the absence of the proestrous LH surge in Clock mutant mice.
- Changes in plasma and brain vasopressin levels in IRAP(-/-) mice suggest a significant role for IRAP in regulating vasop...
- Results suggest ways in which oxytocin may play a cooperative role together with arginine vasopressin in the regulation...
- A short sequence in the mouse arginine vasopressin (AVP) proximal promoter has the ability to confer robust BMAL1/CLOCK...
- These data indicate that AVP regulates body fluid homeostasis and GFR via the V1aR in macula densa cells by activating R...
- "AVP specifically inhibits expression of the cyclin D1 gene in Y1 adrenocortical tumor cells, leading to cell cycle...
- AT(1a) receptor deletion causes polyuria and urine concentration defects by decreasing basal AVP release and impairing A...
- The mechanism behind AVP escape in liver cirrhosis seems to involve decreased collecting duct sensitivity to AVP as a re...
- Results suggest that estrogen receptor-beta activation may play a critical role in estrogenic regulation of oxytocin and

Figure 2 IntegromeDB gene/protein report page. Gene/protein attributes (descriptions, synonyms, external ids, GO annotations, orthology, publications, and others) from a variety of data sources are ordered by relevance and clustered by data source.

gene (Figures 1F, 2), the user must first select a gene from the search result list (Figure 1E). The Search results are divided into two groups: Exact Search "p53" and Relevant to "p53". The first group contains genes which names contain the query term (p53); the second group contains genes which were found in the same publications or the same database records with the query term.

One of the important features of IntegromeDB is the ability to simultaneously extract data for multiple genes in a batch, thus eliminating the need for laborious cross-referencing of data from external databases. The batch search is particularly useful for functional genomics studies, where it is necessary to regularly update annotations associated with genes/proteins being examined. For example, researchers interested in the mapped position or subcellular localization of a list of genes can extract these attributes and perform further analyses, assessing the enrichment of transcription factor binding sites or a certain functional attributes within clusters of genes. List of genes can be input as a text file via uploading on the server or by pasting it into a search box. The batch search is currently limited to 1000 genes.

The cornerstone of the IntegromeDB search engine is the IntegromeDB ontology and Categorization Engine

(ICE) (Figure S1 in Additional file 1). The IntegromeDB ontology, which is described in detail in [13,14], integrates over a hundred OBO ontologies and consists of millions of terms organized as a directed acyclic graph (DAG), reflecting 'is-a', 'has-a', and 'part-of' relationships; for example, 'p53 is-a transcription factor'. Given a user query, ICE determines the nodes in the ontology that are most closely connected with a query. For example, for the 'p53' query, ICE first determines that p53 is a tumor suppressor, which is related to carcinogenesis and cancer pathogenesis. It also determines that p53 is a family of transcription factors related to cell signaling and playing role in the immune system and cell proliferation.

For each ontological term A and for each web page, or a document, X, integrated into the system, the relevance RL of the document X to the term A, is calculated as follows:

$$RL(X, A) = K1 \times PO(A) + K2 \times PP(X) + K3 \times PT(X) + K4 \times PR(X),$$

where $K1$-$K4$ - empirical coefficients (K1 = 0.2; K2 = 0.5; K3 = 0.4; K4 = 0.8); PO - frequency of a term A in the ontology which the term A belongs to; PP - frequency of a term A in the document X, this value corresponds to the *Lucene* keyword search score [22]; PT - frequency of the term A in different HTML tag fields of the web page (document) X; PR - PageRank [23] of the document X. In the process of executing the query containing the term A, ICE retrieves the top one thousand ranked documents.

Documents, or web pages, that are found to be relevant to the query gene/protein are further grouped together by data sources (Figure 3) and by similarity among documents. The similarity between two documents is defined as follows. Assume that a document is seen as a sequence of 'words', where a 'word' can be any combination of letters and digitals, representing biological information. It can be a simple word, like *'diabetes'*, or an abbreviation, like *'Leu'*, a *sequence* word 'ATGCTGGG' or 'MEPQSDVL', a number, or a combination of letters and numbers, like 'Hs.654481'. Further assume that for each word i a threshold frequency to be in the document is set at $b(i)$. Now, for each document, we can compute a vector in which the i-th component is set to 1, if the value of relative frequency of the word i in the document is greater than $b(i)$, or 0, otherwise. This binary vector is regarded as a fuzzy digital signature of the document and is built on the *descriptive set of words*. Vectors with 3 or less words above threshold are removed from further consideration. Low complexity words, such as 'ATGC', 'MEPQSDVL', or '0123456789', are considered as unique sets of symbols and are not considered in the vectors. The descriptive set of N words, which is defined for the whole database once it is updated, consists of the 'quality' words that cover nearly the maximum amount of documents thus that N is nearly minimal, where N is determined empirically. The 'quality' of a word is determined as relative stability of the corresponding component of the vector in similar documents. The threshold $b(i)$ is chosen thus that the most of the documents have non-zero vectors. To estimate the similarity between two documents, the normalized Levenshtein distance [24] is used. Two documents are deemed similar if the difference between them does not exceed 10%. This rough estimate was obtained via assertion of the similarity between randomly chosen 300 documents by six researchers. Figures 4 and 5 show the documents (web pages) grouped by synonyms, sequences, pathways, regulatory and other types of data for the query 'p53'.

Utility

IntegromeDB web interface

IntegromeDB report page uses a two-dimensional newspaper-like layout (Figure 2) rather than as a search results page as in standard search engines (*e.g.*, Google). Data are grouped by data sources and similarity and sorted in each group by the relevance to the query gene/protein as is shown in Figures 2, 3, 4 and 5. The report page also provides a list of topics related to the query gene/protein (Figure 4). The links to the corresponding data sources are provided. Among data displayed for each gene/protein on the left panel of the report page (Figure 3) are aliases and external IDs, ontological terms, chromosome localization with links to NCBI and UCSC genome browsers, orthologs/homologs from multiple organisms, protein and genomic sequences, and related publications. On the right panel of the report page (Figures 4, 5) are shown interaction data and pathways (from KEGG [25], NCI [26], REACTOME [27], BioCARTA [28]), experiments (e.g., expression, metabolomics and proteomics data from GEO [29] and ArrayExpress [30]), miRNA data (from miRBase [31] and microRNA.org [32]), images (*e.g.*, protein structures from PDB [33] and Wikipedia [34]), and relative mRNA expression frequencies derived from various cell and tissue types (from descriptions/metadata of experiments in GEO and ArrayExpress).

For example, for the query 'p53', the report page (Figures 2, 3, 4 and 5) describes that p53 is involved in apoptosis, cancer, prostoglandin metabolism pathway, and MAPK, Wnt, cell cycle and other canonical pathways. Detailed information for each pathway, including genes, proteins, and small molecules involved in it is provided. For example, for prostoglandin metabolism pathway, the user can learn that it is involved in metabolizing lipids into prostaglandins and plays an important role in pain and inflammation; that the protein encoded by human PTGS1 gene is involved in the conversion of prostaglandin

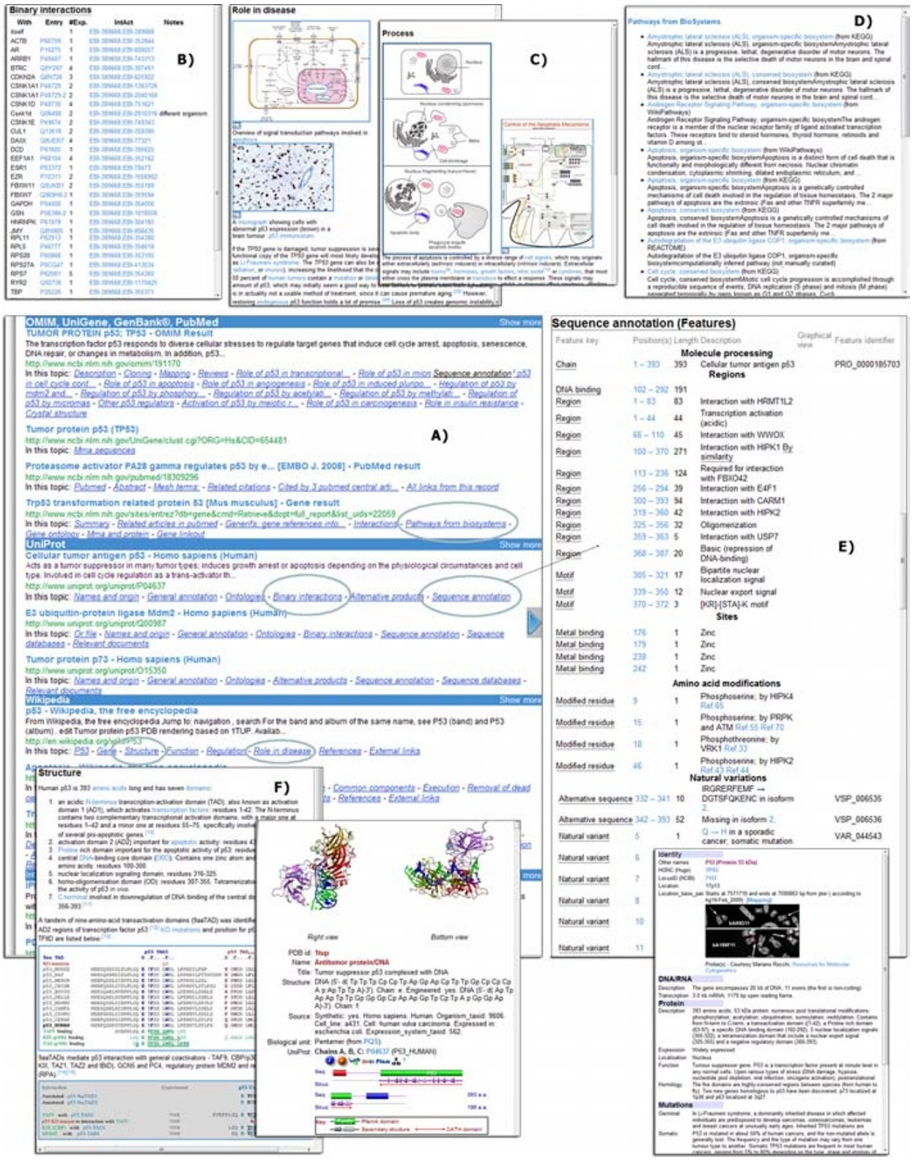

Binary interactions
B)
Role in disease
Process
C)
D)
Pathways from BioSystems
OMIM, UniGene, GenBank®, PubMed
TUMOR PROTEIN p53; TP53 - OMIM Result
Tumor protein p53 (TP53)
Proteasome activator PA28 gamma regulates p53 by e... [EMBO J. 2008] - PubMed result
Trp53 transformation related protein 53 [Mus musculus] - Gene result
A)
UniProt
Cellular tumor antigen p53 - Homo sapiens (Human)
E3 ubiquitin-protein ligase Mdm2 - Homo sapiens (Human)
Tumor protein p73 - Homo sapiens (Human)
Wikipedia
p53 - Wikipedia, the free encyclopedia
Show more
Sequence annotation (Features)
Molecule processing
Regions
Sites
Amino acid modifications
Natural variations
E)
Structure
F)
Right view
Bottom view

Figure 3 Detailed view of the web data on the IntegromeDB report page. A) Search results are grouped by data sources (UniProt, GenBank, PDB, *etc*.), and groups are separated by blue bands. In every group, data are presented according to their relevance to the search criteria. Rolling over the 'In this topic' item (in circles) shows an Instant Preview of the respective item, for example: **B)** Binary Interactions, **C)** Role in disease and Process, **D)** Pathways, **E)** Sequence Annotation, **F)** Structure.

PGG2 into inflammation-causing prostaglandin PGH2; and aspirin has been shown to bind to the PTGS1 gene product (prostaglandin-endoperoxide synthase 1), blocking the ability of this enzyme to produce PGH2 and thereby reducing pain and inflammation.

Web Data

For each gene/protein, information retrieved from web pages is clustered by data sources (Figure 3). The user can scan the content of the web page by rolling over the respective term (e.g. 'Binary Interactions', 'Sequence Annotation', 'Structure') at the bottom of the snippet under *'In this topic'*. For example, for the 'p53' query, among the most relevant (valuable) resources were OMIM, UniProt, UniGene, GenBank, PubMed, Wikipedia, GeneCards, InterPro, p53.free.fr (mutation database), SYSTERS (protein families database), tp53.org. They were followed by (accessible via clicking the *'Show More Pages'* button) EMBL/EBI, MGD, STRING, WikiGenes, Genetic Atlas, CancerIndex, ProteinAtlas, KEGG, UCSC genome Browser, HAGR (Human Ageing Genomic Resource), SwissProt, PharmGKB and others.

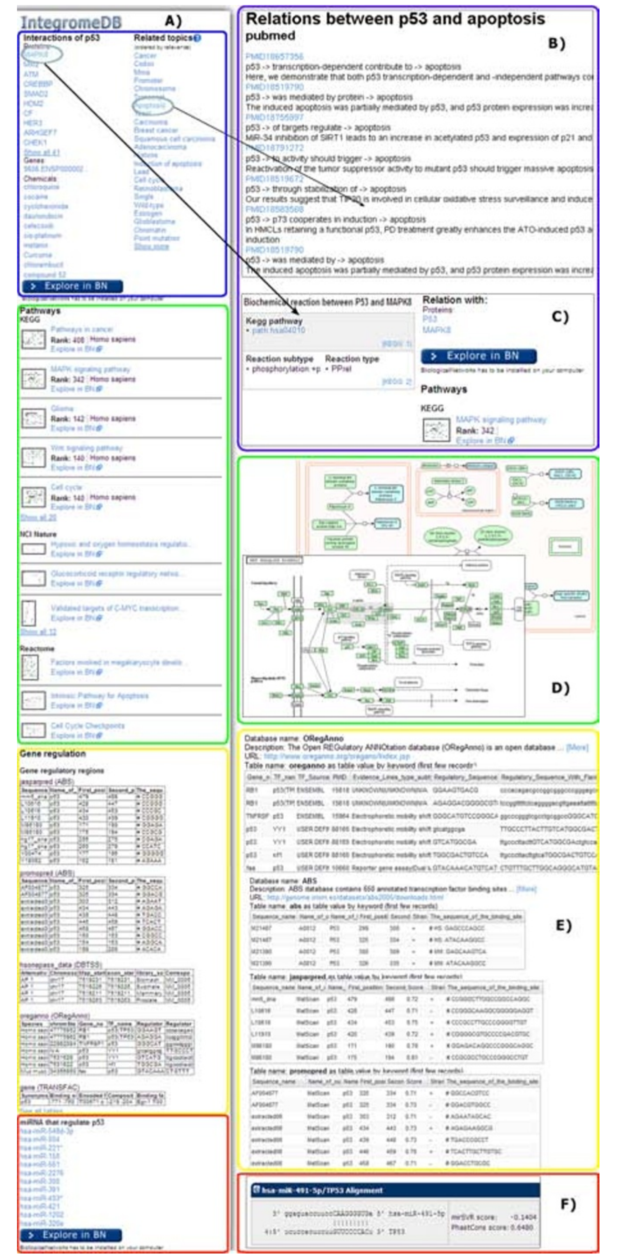


Figure 4 Representation of table and relational data on the IntegromeDB report page. A) Overview of the right panel of the report page. **B)** Relational data presented as topics, for example, for relations between p53 and apoptosis. **C)** Data on interactions and pathways, for example, the interaction between P53 and MAPK8 with detailed view of associated canonical pathways **(D). E)** Tables contain data on transcription factors and gene regulatory regions. **F)** Data on non-coding RNA.

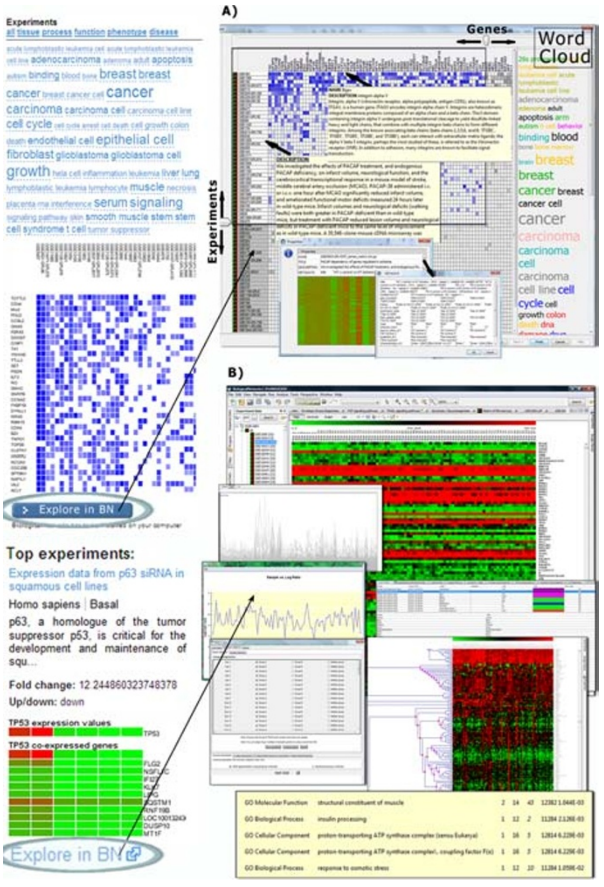

A)
Experiments
Genes
Word Cloud
Experiments
Explore in BN
B)
Top experiments:
Expression data from p63 siRNA in squamous cell lines
Homo sapiens | Basal
p63, a homologue of the tumor suppressor p53, is critical for the development and maintenance of squ...
Fold change: 12.244860323748378
Up/down: down
TP53 expression values
TP53 co-expressed genes
Explore in BN

Figure 5 Representation of experimental time-value data on the IntegromeDB report page. A) Microarray experiments are clustered and represented as the multi-experiment matrix of the most often co-expressed genes with a query gene over a variety of tissues, cell types and diseases, which are represented as a cloud of words. **B)** The matrix and individual experiments can be further explored in BiologicalNetworks http://www.bioloigcalnetworks.org application by clicking on 'Explore in BN.'

Table data

In contrast to table data stored in listed and well-maintained databases, the tables published in the PubMed articles, as well as separate tables distributed across the web, are barely searchable by any search engine. Here, these data are integrated and become searchable using the same approach that is applied to the tables in the databases. Specifically, it concerns tabular data on gene regulatory regions, gene/protein interactions, and gene expression experiments.

To integrate the tables, IntegromeDB looks for web pages containing the HTML 'Table' tag and distinguishes relational tables (that are further parsed) from non-relational (that cannot be integrated), using a combination of in-house and statistical classifiers, e.g., calculating the ontology term in each column of the table; the classifiers have been trained on manually selected examples and use the information already integrated in IntegromeDB. If the columns containing significant percentage (empirically defined value) of the object IDs from IntegromeDB can be found, the table will be further processed; otherwise, filtered out. For example, a two-column table containing gene names and promoter sequences will be identified as a relational table as the percentage of ontologically defined objects in the first (gene names) column is high. Data from the other column will be considered as gene attributes if types of data can be defined, or named; that is, the column contains a data label in the first row. Thus, if in the example table the first cell of the second column contains the word 'promoter' and the other cells, DNA sequences, the postprocessor will understand that data in this column are promoter sequences. It will therefore scan the upstream region of the respective gene from the first column, and in case of finding the matching sequence, remembers its chromosomal localization in the global genomic coordinates and assigns to the gene the sequence in the corresponding cell as the attribute of the type *promoter_sequence*. If the sequence is not found, the postprocessor will assign the sequence as a new attribute with the name as provided in the first cell of the second column, and the type of the attribute will be *sequence*. Figure 4E shows gene regulatory data that was found for the 'p53'query from the databases provided information on gene regulatory regions and transcription factors.

Experimental tables containing, for example, microarray data, are clustered based on the Z-scores obtained using Fisher r-to-Z transformation of Pearson correlation coefficients calculating co-expression of the query gene with each gene in each experiment (data for each gene are averaged over multiple probes). The resulting co-expression matrix (Figure 5) highlights the experiments and genes (the intensity of blue correspond to the increasing value of the Pearson correlation coefficient above 0.75 and up to 1.0) co-expressed with the query gene (p53, in this case), and the word cloud shows a variety of conditions, tissues, and disease states found in the descriptions of these experiments. Visual inspection of the matrix allows detecting patterns of correlation across data sets and spot significantly strong co-expression profiles. Most often correlating genes (with the largest average Z-score over all experiments) and most related experiments (with the largest average Z-score over all genes) are located in the left-top corner of the matrix. For example, p53 was found to be overexpressed in cancers (breast, brain, bone marrow, squamous carcinoma, etc.), fibroblasts, astrocytes, and other cell types.

Integration with other resources

One of the central aims of the IntegromeDB is to maintain cross-connectivity and integration with other public resources in a user-friendly manner. Therefore, we provide the programmatic access to our SQL database that is accessible via the following routes. First, the integrated content of IntegromeDB is available via the IntegromeDB API, which is implemented in Java. Through an XML-RPC service, API provides functions to access programmatically most of the features available in the IntegromeDB web interface, such as retrieving aliases, promoter sequences, or transcriptional regulators for a set of genes. Example code of using API and access to the XML-RPC service are available at http://integromedb.org/api.jsp ('API XML-RPC' tab). Second, if the external user/resource wants to visualize retrieved object(s) (interaction network of a protein, promoter region of a gene, microarray experiments for a set of genes, *etc.*) on his web resource using the BiologicalNetworks integrated research environment [13] he/she should use API access described at the http://integromedb.org/api.jsp 'API BiologicalNetworks' tab. Thus IntegromeDB/BiologicalNetworks maintains mutual cross-referencing with other web resources, that is not limited to simple text-based HTML links, but also enables partner websites to embed visualization of the BiologicalNetworks objects within their own web pages.

Most of the data (pathways, networks, microarray experiments, sequences, etc.) returned by the IntegromeDB web site as a search results can be opened for detailed exploration and analysis in BiologicalNetworks. Figure 6 illustrates several examples of the interactions between IntegromeDB and BiologicalNetworks; in these scenarios, IntegromeDB is used for browsing large volumes of data, while BiologicalNetworks is used for exploring individual datasets in finer detail.

Partner websites or third-party software programs can choose to embed the entire IntegromeDB website into their own software. Thus, an IntegromeDB 'plugin' can be established at the BioGPS [6] portal, which provides 'plug-ins' through which the users can connect any number of external websites into freely configurable screen layouts.

Discussion and Conclusions

Centralized and publicly available resources and search engines for integrated biological data are critical in enabling biologists to efficiently analyze genome-scale data. To meet this demand, we developed IntegromeDB, the general purpose unifying resource and search engine that can take the exploration of molecular biology data to the next level. IntegromeDB uses a hybrid approach to the Deep Web that combines elements of the crawl and federated search and data preprocessing approaches. It follows the trend established by other integrative resources, BioGPS [6] and Entrez [11], for example, that aim at presenting available information in a gene-centered manner. However, the technology behind IntegromeDB differs from that of other resources as it relies on data integration via automatic establishment of Object-Object Property relationships rather than on hyperlinks and manual curation. This approach is scalable in respect of how many data sources can be integrated (the number can be limited only by the hardware availability) and relatively inexpensive to

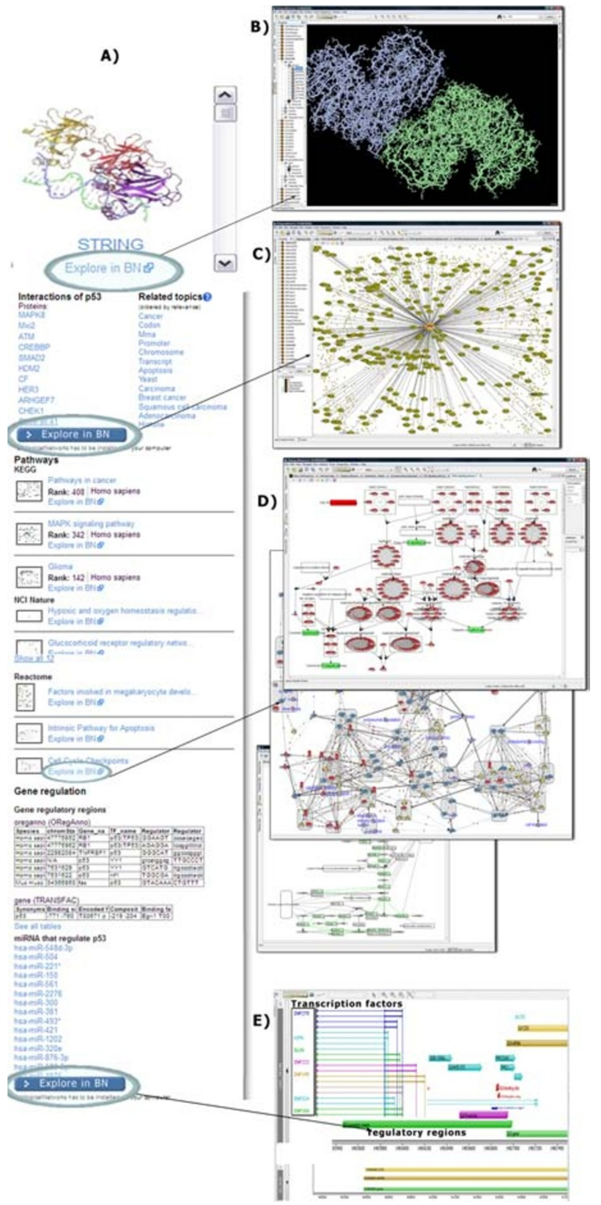


Figure 6 Visualization and exploration of different types of biological information in BiologicalNetworks integrated research environment. A) Overview of the right panel of the report page. **B)** Visualization of 3D protein structure. **C)** Visualization of protein-protein interaction networks. **D)** Visualization of pathways data. **E)** Visualization of data on gene regulatory regions.

develop and maintain as we deliberately keep human intervention at a minimum. However, as any automatic approach it has the drawback that unrelated data can be occasionally retrieved. With more usage of the site and more errors reported, the search algorithm will be improved. We encourage the users to report the errors and provide critique using the "Provide Feedback" button at the top right of the search results page (Figure 2).

To further enhance the ability of researchers to extract and manipulate the data, we will continue to provide the access to IntegromeDB through the API and data integration tools available in BiologicalNetworks making the most often used tools implemented as web services, thus eliminating the necessity to download BiologicalNetworks. We will also work on providing execution of SPARQL queries on the IntegromeDB data and create a SPARQL to SQL adaptor, using specifically designed dictionaries (that provide description of triples along with exact mappings onto tables).

We plan to continue extending the resource by adding new data sources and more experimental data and further developing the search engine's usability and speed. To improve the resource design and usability, we will rely on external experts (subject of funds availability) and we welcome and encourage community feedback that can be provided using the "Provide Feedback" button at the top right of the search results page. The speed of the presented search engine depends on both software and hardware; and while we will continue optimizing the search algorithms, we also plan to utilize a new hardware, specifically, a supercomputer Gordon that will become available at the San Diego Supercomputer Center in 2012.

Availability and requirements

- **Project name:** IntegromeDB
- **Project home page:** http://www.integromedb.org
- **Operating system(s):** Platform independent. Tested on the Internet Explorer 8, FireFox 8, Google Chrome web browsers on Windows 2000/XP/Vista/7, Linux/Ubuntu/Redhat, and MacOSX OS.
- **Programming language:** Java
- **Other requirements:** Java 1.6 for running BiologicalNetworks program;
- **License:** GNU GPL version 3
- **Any restrictions to use by non-academics:** none

Additional material

Additional file 1: Components of the IntegromeDB architecture and Search algorithm. Major components of the IntegromeDB search engine.

Acknowledgements
This work was supported by the National Institute of Health grants R01GM084881 to MB and R01GM085325 to JP.

Author details
[1]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA. [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA.

Authors' contributions

SK, YD and MB designed and implemented the system and performed major programming work, JP contributed to the system concept, MB and JP wrote the manuscript. This work was coordinated by MB. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 19 October 2011 Accepted: 19 January 2012
Published: 19 January 2012

References

1. Wu C, Bairoch A, Apweiler R, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic Acids Res* 2006, **34**: D187-191.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2011, **39**:32-7.
3. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D: **GeneCards Version 3: the human gene integrator.** *Database* 2010, doi: 10.1093/database/baq020.
4. Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ, RGD Team: **The Rat Genome Database, update 2007 - easing the path from disease to data and back again.** *Nucleic Acids Res* 2007, **35**:658-62.
5. Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT, the Mouse Genome Database Group: **The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics.** *Nucleic Acids Res* 2011, **39(suppl 1)**:842-848.
6. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, Su AI: **BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.** *Genome Biol* 2009, **10(11)**:R130.
7. Köhler J, *et al*: **Graph-based analysis and visualization of experimental results with Ondex.** *Bioinformatics* 2006, **22(11)**:1383-1390.
8. Gupta A, Bug W, Marenco L, Qian X, Condit C, Rangarajan A, Müller HM, Miller PL, Sanders B, Grethe JS, Astakhov V, Shepherd G, Sternberg PW, Martone ME: **Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF).** *Neuroinformatics* 2008, **6(3)**:205-17.
9. Bader GD, Cary MP, Sander C: **Pathguide: a pathway resource list.** *Nucleic Acids Res* 2006, **34**:D504-6.
10. Blake A, Pickford K, Greenaway S, Thomas S, Pickard A, Williamson CM, Adams NC, Walling A, Beck T, Fray M, Peters J, Weaver T, Brown SD, Hancock JM, Mallon AM: **MouseBook: an integrated portal of mouse resources.** *Nucleic Acids Res* 2010, **38**:D593-9.
11. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucl Acids Res* 2011, **39**:D52-D57.
12. Galperin MY, Cochrane GR: **The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.** *Nucleic Acids Res* 2011, **39**:D1-6.
13. Kozhenkov S, Dubinina Y, Sedova M, Gupta A, Ponomarenko J, Baitaluk M: **BiologicalNetworks 2.0–an integrative view of genome biology data.** *BMC Bioinformatics* 2010, **11**:610.
14. Baitaluk M, Ponomarenko J: **Semantic Integration of Data on Transcriptional Regulation.** *Bioinformatics* 2010, **26(13)**:1651-1661.
15. Kozhenkov S, Sedova M, Dubinina Y, Gupta A, Ray A, Ponomarenko J, Baitaluk M: **BiologicalNetworks–tools enabling the integration of multi-scale data for the host-pathogen studies.** *BMC Syst Biol* 2011, **5**:7.
16. **RDF.** [http://www.w3.org/RDF/].
17. Good BM, Wilkinson MD: **The Life Sciences Semantic Web is full of creeps!** *Brief Bioinform* 2006, **7**:275-286.
18. Wright A: **Searching the Deep Web.** *CACM* 2008, **51(10)**:14-15.
19. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH: **An overview of the PubChemBioAssay resource.** *Nucleic Acids Res* 2010, **38**:D255-66.
20. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA: **genenames.org: the HGNC resources in 2011.** *Nucleic Acids Res* 2011, **39**:D519-9.
21. Hubbard TJP, *et al*: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35**:D610-D617.
22. **Apache Lucene project.** [http://lucene.apache.org].
23. Page L, Brin S, Motwani R, Winograd T: **The PageRank citation ranking: Bringing order to the Web.** 1999.
24. Levenshtein VI: **Binary codes capable of correcting deletions, insertions, and reversals.** *Soviet Physics Doklady* 1966, **10**:707-10.
25. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**:D355-D360.
26. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: The Pathway Interaction Database.** *Nucleic Acids Res* 2009, **37**:D674-9.
27. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P: **Reactome knowledgebase of biological pathways and processes.** *Nucleic Acids Res* 2009, **37**:D619-22.
28. **Biocarta pathways.** [http://www.biocarta.com].
29. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets–10 years on.** *Nucleic Acids Res* 2011, **39**: D1005-10.
30. Parkinson , *et al*: **ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression.** *Nucl Acids Res* 2009, 37:D868-72.
31. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res* 2011, **39**:D152-D157.
32. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression.** *Nucleic Acids Res* 2008, **36**:D149-53.
33. **Protein Data Bank (PDB).** [http://www.rcsb.org].
34. **Wikipedia.** [http://www.wikipedia.org].

doi:10.1186/1471-2164-13-35
Cite this article as: Baitaluk *et al.*: IntegromeDB: an integrated system and biological search engine. *BMC Genomics* 2012 **13**:35.