The NIH National Center for Integrative Biomedical Informatics (NCIBI)

Brian D Athey,[1] James D Cavalcoli,[1] H V Jagadish,[2] Gilbert S Omenn,[1] Barbara Mirel,[3] Matthias Kretzler,[4] Charles Burant,[5] Raphael D Isokpehi,[6] Charles DeLisi,[7] the NCIBI faculty, trainees, and staff[8]

[1]Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan, USA
[2]Computer Science and Engineering Division, College of Engineering, University of Michigan, Ann Arbor, Michigan, USA
[3]School of Education, University of Michigan, Ann Arbor, Michigan, USA
[4]Nephology Division, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA
[5]Metabolism and Obesity Center, University of Michigan Medical School, Ann Arbor, Michigan, USA
[6]Center for Bioinformatics & Computational Biology, Jackson State University, Jackson, Mississippi, USA
[7]Biomedical Engineering, Boston University, Boston, Massachusetts, USA
[8]National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, Michigan, USA

Correspondence to
Dr Brian D Athey, 100 Washtenaw Ave, Ann Arbor, MI 48109-2218, USA; bleu@umich.edu

BDA and JDC are co-lead authors.

Received 18 August 2011
Accepted 23 August 2011
Published Online First 19 November 2011

ABSTRACT

The National Center for Integrative and Biomedical Informatics (NCIBI) is one of the eight NCBCs. NCIBI supports information access and data analysis for biomedical researchers, enabling them to build computational and knowledge models of biological systems to address the Driving Biological Problems (DBPs). The NCIBI DBPs have included prostate cancer progression, organ-specific complications of type 1 and 2 diabetes, bipolar disorder, and metabolic analysis of obesity syndrome. Collaborating with these and other partners, NCIBI has developed a series of software tools for exploratory analysis, concept visualization, and literature searches, as well as core database and web services resources. Many of our training and outreach initiatives have been in collaboration with the Research Centers at Minority Institutions (RCMI), integrating NCIBI and RCMI faculty and students, culminating each year in an annual workshop. Our future directions include focusing on the TranSMART data sharing and analysis initiative.

MISSION

The mission of the National Center for Integrative Biomedical Informatics (NCIBI) is to facilitate scientific exploration of complex disease processes on a much larger scale than is currently feasible. The Center develops interactively integrated analytical and modeling technologies to acquire or create context-appropriate molecular biology information from emerging experimental data, international genomic databases, and the published literature. The NCIBI supports information access for and the data analysis workflows of collaborating biomedical researchers, enabling them to build computational and knowledge models of biological systems validated through focused work on specific diseases. Computational advances in NCIBI are focused on developing tools and methods which are directly applied to the complex disease areas which are the focus of our Driving Biological Problems (DBPs), discussed below, and which can be broadly disseminated. Integrated tools and data include the Michigan Molecular Interactions database (MiMI) for protein interactions, information extraction from NLP software algorithms, visualization and modeling data using Cytoscape plugins to the MiMI database and for metabolic data (Metscape), and network mapping tools like SAGA for approximate sub-graph matching. The tools and applications are detailed in table 1 and at our web portal (http://portal.ncibi.org/gateway/tryourtools.html). The Center also focuses on outreach, training, and educational programs to broaden the use of applicable methods and analyses. NCIBI supports graduate students with faculty mentoring them in projects relating to the overall mission of NCIBI (figure 1).

DBPS: WHAT WORKED WELL?

We have had several different DBPs over the first 6 years of NCIBI. The original DBPs were: (1) prostate cancer progression; (2) organ-specific complications of type 1 and 2 diabetes; and (3) genetic susceptibility and phenotypic sub-classification of bipolar depressive disease.[13] We also had several partners as sub-contractors including Lee Hood, Biaoyang Lin (now at the Swedish Medical Center, Seattle), Qiang Tian (oncology and proteomics; Institute for Systems Biology, Seattle), and Rich Watanabe (type 2 diabetes; University of Southern California, Los Angeles). Projects within these categories which were most successful as a result of their engagement with NCIBI were those which were already generating biological data, and had keenly focused hypothetical questions. The value of the NCIBI resources for these scientists lay in the integration of data and tools, allowing them to visually explore and browse through their data, especially with the addition of parsed, tagged, and searchable data from the literature, thus speeding up analysis and leading to new hypotheses.

Prostate cancer progression: from androgen-regulated signaling pathways to causal gene fusions to mediation of metastatic phenotypes

The discovery of androgen-responsive TMPRSS2/ETS family fusion genes in a high percentage of prostate cancers by NCIBI investigator Arul Chinnaiyan and colleagues has stimulated a whole new thrust of bioinformatics-driven research focused on multi-dimensional characterization of gene fusions in solid tumors.[14] The Oncomine database and the hypothesis of heterogeneity among similarly diagnosed patients were essential to this discovery, which has seen broad use in the National Cancer Institute (NCI) community, and has led to the creation of the spin-off Compendia Biosciences (Ann Arbor, Michigan, USA).[5] Commercial biomarker development is underway. We have used the full NCIBI suite of tools and advanced experimental validation methods to generate an innovative iterative approach to reveal driver gene fusions from paired-end sequencing data

Table 1 The National Center for Integrative Biomedical Informatics suite of tools and data and partial usage statistics

Database/tool name	Function	Reference/usage
Michigan Molecular Interactions (MiMI); MiMI Cytoscape plug-in	Data merged and integrated from numerous protein interaction databases (GRID, intAct, BIND, HPRD, MINT, DIP, MDC, MIPS, Reactome, and KEGG) along with gene and pathway information; maintains provenance	Jayapandian *et al*,[1] Tarcea *et al*[2] (MIMi website 9733 visits (4133 unique), 445 plug-in downloads)
NLP Parsed PubMed and PMCOA	Literature parsed and tagged for sentence parts, as well as gene names, gene name synonyms, and metabolites	http://nlp.ncibi.org/about.html
Nephromine	Compendium of publicly available renal gene expression profiles, a sophisticated analysis engine, and a powerful web application designed for data mining and visualization of gene expression data	Martini *et al*[3]
Cell line ontology	A cell line ontology for tagging cell line names in biomedical text	Sarntivijai *et al*[4]
Oncomine research edition (ORE)	Comprehensive resource for cancer gene expression datasets, with embedded statistical and informatics tools; used to collect, standardize, analyze, and deliver data to the research community	Rhodes *et al*[5]
Tools for exploratory analysis		
SAGA/TALE	SAGA (Substructure Index-based Approximate Graph Alignment) is a tool for querying a biological graph database to retrieve matches between sub-graphs of molecular interactions that scientists select and biological networks. TALE is used for graphs with larger numbers of nodes and edges (up to 200).	Tian *et al*[6]
ConceptGen	A web-based tool to identify biological gene sets (called concepts) enriched with differentially expressed genes (or any other user-identified gene list), and to explore networks of relationships among biological concepts	Sartor *et al*[7] (7289 visits, 783 unique)
Metscape 1.01, v2.11	This Cytoscape plug-in is used to visualize and analyze metabolomic data. It uses data from the Edinburgh Human Metabolic Network reconstruction.[8]	Gao *et al*[9] (1173 visits, 336 downloads)
LRPath	A logistic regression method for identifying enriched biological groups in gene expression data	Sartor *et al*[10]
Literature search analysis tools		
Gene2MeSH	An annotation tool that associates Medical Subject Heading (MeSH) terms with genes using PubMed. Associations are ranked by (Fisher's exact test) significance score between genes and MeSH terms.	http://gene2mesh.ncibi.org/
Metab2MeSH	Similar to Gene2MeSH; annotated for metabolites rather than genes	http://metab2mesh.ncibi.org/
PubAnatomy	Provides new ways to explore relationships in anatomical structures, pathophysiological processes, gene expression levels, and protein—protein interactions in Medline literature and experimental data.	Xuan *et al*[11]
PubOnto	Provides multiple ontologies from the Open Biomedical Ontology to help researchers explore literature from different perspectives. Quickly locates and filters articles by concepts (eg, Gene Ontology (GO)).	Xuan *et al*[12]

in cancer.[15] Xiaosong Wang and Maureen Sartor (Core 1) developed a concept signature score and an integrative pipeline that assimilates multi-dimensional data and distinguishes fusions and mutations. Using this approach, we identified a novel R3HDM2-NFE2 gene fusion in aggressive lung cancers and are validating additional important gene fusions.

Systems biology of diabetic nephropathy

Molecular signatures and mechanisms could define progression of complications in individual patients with diabetes. Matthias Kretzler and colleagues have created an international Renal BioBank Network with kidney biopsies from 2600 patients with eight causal categories of glomerular nephropathy. Specimens were micro-dissected to generate glomerular and tubulo-interstitial compartments, then analyzed with the NCIBI-MIT GenePattern pipeline for microarray data processing and NCIBI data analysis tools. In analogy with Oncomine for cancers, we created Nephromine, a database resource for the international community.[3] A major new drive in this DBP is the integration of transcriptomic data sets with comprehensive metabolite networks for identification of putative diagnostic markers and novel pathways using Metscape. NCIBI has provided critical support for the University of Michigan George M. O'Brien Kidney Research Core Center (P30) to successfully compete for the NIDDK RO1 Molecular Predictors of Progressive Renal Failure in the Chronic Renal Insufficiency Cohort grant, and the NIDDK R24 Integrated Systems Biology Approach to Diabetic Microvascular Complications grant. In addition, it allowed Dr Kretzler to effectively compete for an NIH Office of Rare Disease U54 funding mechanism for the Nephrotic Syndrome Study Network (NEPTUNE). NEPTUNE currently has 15 participating academic health centers across North America.

Metabolism and obesity studies

An important area of collaboration is the development of Metscape, a Cytoscape plugin that is being utilized to provide new insights into biology by integrating and visually displaying metabolomic and gene expression data.[9] Several types of informatics tools have been developed to represent, visualize, and analyze biological networks and experimental data. Metscape represents the next generation of programs that provide dynamic networks, rather than just static images, and enables direct integration and comparisons of different types of experimental data. In recent studies, new insights into metabolic pathways associated with exercise tolerance in preclinical rat studies have been uncovered using Metscape.

NCIBI OUTPUTS AND DELIVERABLES

NCIBI tools and data which drove these successes

Over the years, NCIBI has had many partners as sub-contractors including Robert Murphy (Carnegie Mellon University), Jill Mesirov and Michael Reich (GenePattern; Broad Institute), Mark

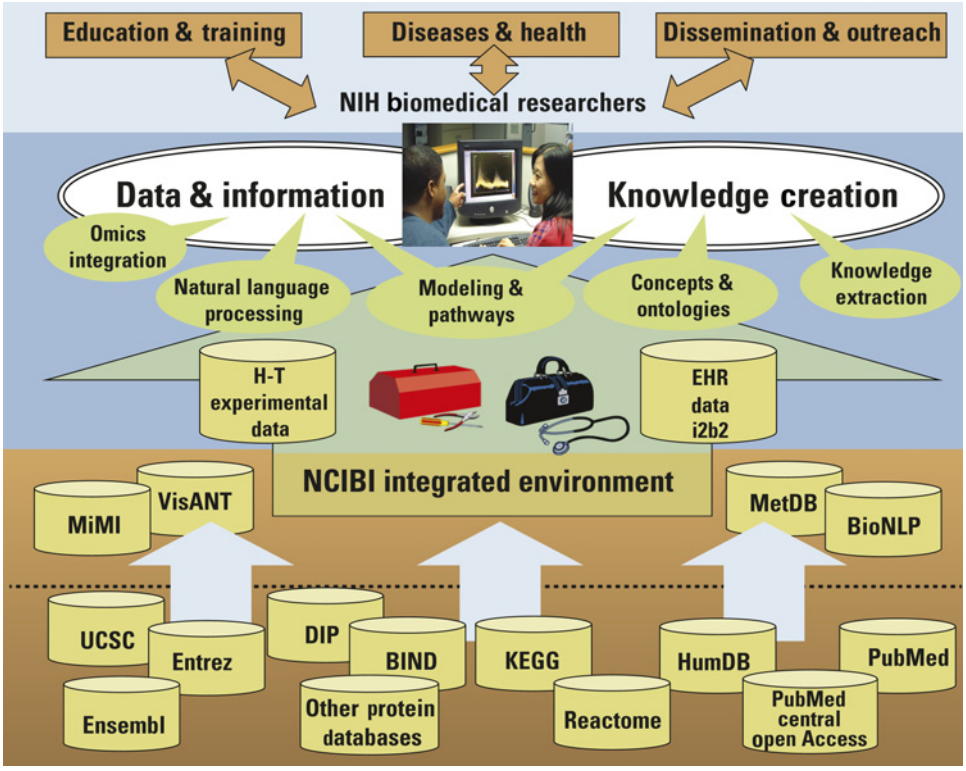

Figure 1 The overall vision of the National Center for Integrative Biomedical Informatics is to apply novel methods of deep multiscale data integration, linked with analytic tools, to enhance biomedical research. Driving Biological Problems (DBPs) use the integrated environment of data and information to accelerate the creation of new knowledge.

Musen (NCBO; Stanford University), Ben Keller (Eastern Michigan University), and Kirstie Bellman (Aerospace Corp). These outstanding partners have contributed to the development of the overall NCIBI tool suite. The NCIBI tools and data are accessible via the NCIBI tools page (http://portal.ncibi.org/gateway/tryourtools.html) which includes links to all the tools, tutorials, and demonstrations of tools usage. The NCIBI computational algorithms and bioinformatics development cores have produced a suite of tools and data that can be organized in three major categories (see table 1):

1. *Core databases:* These databases serve as the building blocks for all the other development tools and as a source for annotation. Some are mirrors of external data (eg, NCBI EntrezGene and NLM PubMed), some are adapted from external repositories and developed by NCIBI (eg, NLP Parsed PubMed, MiMI), and others were developed for specific DBP purposes (eg, Oncomine, Nephromine).
2. *Exploratory analysis tools:* These tools for exploring experimental data sets link to the core databases, provide analysis and annotation frameworks for research data, and facilitate integration across the NCIBI suite of tools, making it easier for scientists to move between applications.
3. *Conceptual literature search tools:* Using the biomedical literature as a source of data and annotation is a key feature of many of the NCIBI tools. The tools in this category are used to browse literature and for annotation based on concepts of keywords rather than with existing datasets. This is especially useful when beginning or broadening a research project and looking for novel connections and relationships outside the initial research domain.

Even among these categories, there are clearly three areas where our tools have driven success, and been utilized broadly by the scientific community. First, programmatic interfaces which allow other informatics tools to rapidly query our backend databases such as NLP Parsed PubMed, MiMI, Conceptgen and other databases. This allows other bioinformatics users to link to our data without having to utilize web frontends. Second, natural language processing of biomedical literature has provided a rich resource on tagged entities such as proteins, genes, and metabolites, and also has led NCIBI to make good progress in developing systems for classifying the interaction types found in biomedical text. Third, metabolomic tools and data are clearly lacking and represent a key niche for NCIBI. The Metscape plugin for Cytoscape provides a bioinformatics framework for the visualization and interpretation of metabolomic and expression profiling data in the context of human metabolism (figure 2). It allows users to build and analyze networks of genes and compounds, identify enriched pathways from expression profiling data, and visualize changes in metabolite data. Gene expression and/or compound concentration data can be loaded from file(s) (in CSV, TSV, or Excel formats) or the user can directly enter individual compounds/genes (using KEGG compound IDs, or Entrez Gene IDs) to build metabolic networks without loading a file. Metscape uses an internal relational database stored at NCIBI that integrates data from KEGG and EHMN.

As of May, 2011, the NCIBI tools and resources had been utilized as follows. NCIBI had 12 123 unique web portal visitors from 100 countries; 51% were new visitors and 49% were returning visitors. Users viewed an average of 2.8 pages per visit, and 9733 (4133 unique) visits and 560 558 queries (520 unique) were logged. The NCIBI web portal http://portal.ncibi.org/gateway/tryourtools.html had a total of 12 132 unique visitors from over 110 different countries. The NCIBI web services are also widely used around the world.

Education and outreach

NCIBI supports five to six graduate students annually as part of its investment in new development, training, and outreach. Over the past 6 years, we have supported 17 different students who have worked closely with professional developers and learnt 'hands-on' techniques for developing robust software. It is

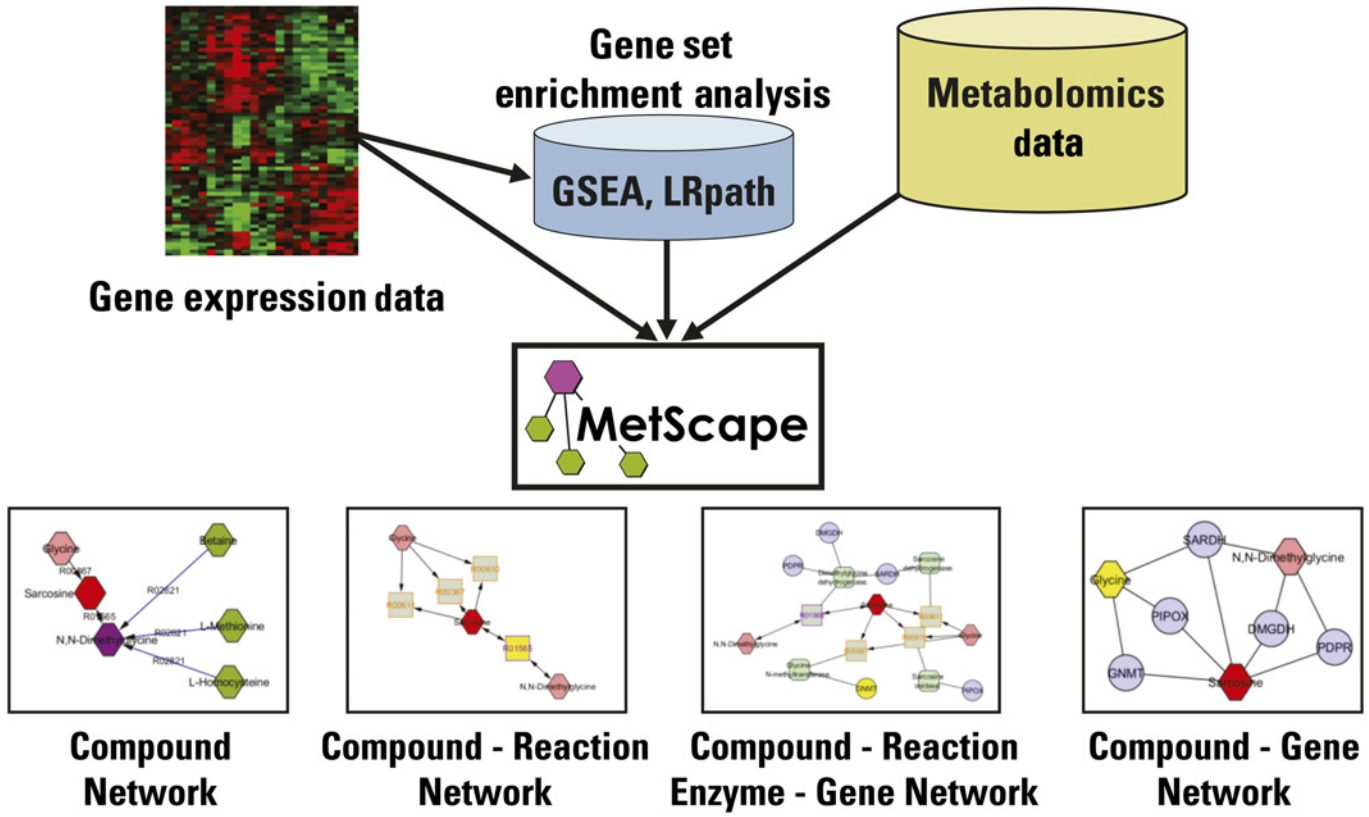

Figure 2 Metscape (a Cytoscape plugin) used to visualize high and low capacity runners' metabolites after 10 min of exercise (http://metscape.ncibi.org/).

also important to recognize that NCIBI's goal of usability and user testing of its software tools is part of the education process for graduate students. Graduate students learn these skills outside of the classroom, and in accordance with the overall development plans of NCIBI. Graduates of the program include: Gunnes Erkan (Google), Carlos Santos (CEO of a pharmaceutical start-up company), Adrianne Chapman (Mitre, Virginia), Magesh Jayapandian (IBM Silicon Valley Lab) and Yuanyuan Tian (IBM Almaden Research Center). In addition, there were several post-doctoral trainees including Xiaosong Wang (Assistant Professor at Baylor), who worked on our cancer DBP with Gil Omenn and Arul Chinnaiyan, and Rich McEachin (Research Investigator, UM), who was building bridges with Scott and Nancy Saccone (Washington University, St. Louis).

Our outreach and training program includes regular web-based 'Tools and Technology' seminars which are accessible by live web interface, and are archived on our website. Ongoing outreach (via local, distance, and remote mechanisms) in conjunction with NCI and our Health Sciences Library provides training on individual tools such as Cytoscape, MiMI, Conceptgen and Metscape. In addition, beginning in our third year, we have held three annual NCIBI/Research Centers in Minority Institutions (RCMI) summer workshops to help provide experience and access to the NCIBI tools and data. We have ongoing collaborations with Jackson State University, the RCMI Translational Research Network, and other RCMI locations.

Support for National Centers for Biomedical Computing overall initiatives

NCIBI has provided hosting and support for the overall National Centers for Biomedical Computing (NCBC) website (http://www.ncbcs.org), and continues to maintain and update it. This site provides links to all the other NCBC programs, as well as links to archives of All Hands meetings, and wiki pages summarizing the overall NCBC efforts. One such effort has been Biositemaps. Biositemaps was developed with support from several of the NCBCs including NCIBI, to develop technologies to address locating, querying, and mining biomedical resources (tools and data). This technology allows for groups to add and curate their own resources using a defined editor, which generates the resource information in a defined RDF schema which conforms to the Biomedical Resource Ontology. This technology has been actively taken up by the Clinical and Translational Science Awards (CTSA) community and NCIBI continues to support its development.

Current/future goals

Currently NCIBI is working to further integrate the tools and data sources, and continues to update these and release new versions on a regular basis. In addition, further outreach and training on our existing tools is being carried out, and refinement of tool integration. Future plans involve tighter integration with the i2b2 platform and insertion of NCIBI tools and data (table 1) into the i2b2 Hive for broad dissemination into NCBC i2b2 performance sites. This capability is also being added to the tranSMART platform in collaboration with the Johnson and Johnson Corporation.[16]

Funding This study was supported by NIH grant number U54-DA021519.

Competing interests None.

Ethics approval The University of Michigan Medical School Institutional Review Board (IRBMED) approved this study.

Provenance and peer review Commissioned; internally peer reviewed.

Data sharing statement NCIBI tools and resources are available free of charge to the research community for non-commercial use. Please refer to http://portal.ncibi.org/gateway/pdf/Terms%20of%20use-web.pdf for the general terms of use.

REFERENCES

1. **Jayapandian M,** Chapman A, Tarcea VG, *et al*. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Res* 2007;**35**: D566—71.
2. **Tarcea VG,** Weymouth T, Ade A, *et al*. Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res* 2009;**37**:D642—6.
3. **Martini S,** Eichinger F, Nair V, *et al*. Defining human diabetic nephropathy on the molecular level: integration of transcriptomic profiles with biological knowledge. *Rev Endocr Metab Disord* 2008;**9**:267—74.
4. **Sarntivijai S,** Ade AS, Athey BD, *et al*. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics* 2008;**24**:2760—6.
5. **Rhodes DR,** Kalyana-Sundaram S, Mahavisno V, *et al*. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 2007;**9**:166—80.
6. **Tian Y,** McEachin RC, Santos C, *et al*. SAGA: a subgraph matching tool for biological graphs. *Bioinformatics* 2007;**23**:232—9.
7. **Sartor MA,** Mahavisno V, Keshamouni VG, *et al*. ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* 2009;**26**:456—63.
8. **Ma H,** Sorokin A, Mazein A, *et al*. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* 2007;**3**:135.
9. **Gao J,** Tarcea VG, Karnovsky A, *et al*. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* 2010;**26**:971—3.
10. **Sartor MA,** Mahavisno V, Keshamouni VG, *et al*. ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* 2010;**26**:456—63.
11. **Xuan W,** Dai M, Buckner J, *et al*. Cross-domain neurobiology data integration and exploration. *BMC Genomics* 2010;**11**:S6.
12. **Xuan W,** Dai M, Mirel B, *et al*. Open Biomedical Ontology-based Medline exploration. *BMC Bioinformatics* 2009;**10**(Suppl 5):S6.
13. **McEachin RC,** Saccone NL, Saccone SF, *et al*. Modeling complex genetic and environmental influences on comorbid bipolar disorder with tobacco use disorder. *BMC Med Genet* 2010;**11**:14—30.
14. **Tomlins SA,** Rhodes DR, Perner S, *et al*. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005;**310**:644—8.
15. **Wang XS,** Prensner JR, Chen G, *et al*. An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat Biotechnol* 2009;**27**:1005—11.
16. **Perakslis ED,** Van Dam J, Szalma S. How informatics can potentiate precompetitive open-source collaboration to jump-start drug discovery and development. *Clinical Pharmacology & Therapeutics* 2010;**87**:614—16.