TECHNICAL BRIEF

Pathway Palette: A rich internet application for peptide-, protein- and network-oriented analysis of MS data

Manor Askenazi[1,2,3,4]*, *Shaojuan Li*[1,2]*, *Saurav Singh*[1,2] *and Jarrod A. Marto*[1,2,3]

[1] Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[2] Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, MA, USA
[3] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA
[4] Department of Biological Chemistry, The Hebrew University of Jerusalem, Jerusalem, Israel

Recent improvements in proteomic technologies have collectively yielded data sets that far exceed the capabilities of typical low-throughput interpretation strategies. Unfortunately, tools designed to leverage the ''peptide-centric'' content of MS-based proteomics lag the current rate of data production. Here, we describe Pathway Palette (http://blais-pathways.dfci.harvard.edu), a freely accessible internet application that enables researchers to easily transition from peptides to biological pathways, while simultaneously retaining the qualitative and quantitative aspects of the underlying MS data.

Received: October 27, 2009
Revised: December 22, 2009
Accepted: January 5, 2010

Keywords:
Bioinformatics / Mass Spectrometry proteomics / Network analysis / Protein–protein interaction / Quantitative proteomics / Rich internet application

Mass spectrometry is now well established as the technique of choice for characterization of proteins derived from a plethora of model systems in biomedical research. Moreover, collective advances in sample preparation, enrichment, and fractionation, among other methods [1] now support proteomic-based experiments that are designed to monitor changes in protein expression and post-translational modification state as a function of biological perturbation. In fact, MS-based proteomics is now well integrated with hypothesis-driven research within many laboratories. In addition to sheer quantity, the diversity of relevant information which needs to be integrated in order to glean biological insight from the data poses an enduring challenge to practitioners [2]. A typical example of this problem, which has garnered much attention in recent years, is MS-based analysis of protein–protein interactions (PPIs) derived from both targeted and large-scale affinity purification experiments. This approach provides an ideal means to characterize multi-component protein complexes formed within a physiologically relevant context, *e.g. in vivo* or *in vitro*; nevertheless, reduction of these data to biological pathways and networks remains tedious and challenging. Despite these difficulties, a ''network diagram'' representation of PPI data has become so ubiquitous that it is a *de facto* standard display for results from such large-scale studies [3–5]. As a result, proteomic researchers naturally gravitate toward tools that allow extension of protein lists to meaningful biological annotation. Unfortunately, software tools that currently offer a network perspective tend to focus at the gene level and typically accept either a gene list or one of the few accession lists commonly used in microarray and high-throughput sequencing platforms. This approach is problematic across a wide range of MS-based proteomic experiments. For example, various *ad hoc* and community-derived standards [6] for data reporting suggest that protein identifications based on a single peptide sequence require higher stringency as compared with identifications based on multiple peptides. Unfortunately, simple metrics such as the number of peptides identified *per* protein is lost in gene-centric tools. As a more complex example, consider the case in which protein identification and quantification are

Correspondence: Dr. Jarrod A. Marto, Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Smith 1158A, Boston, MA 02115-6084, USA
E-mail: jarrod_marto@dfci.harvard.edu
Fax: +1-617-582-7737

Abbreviation: PPI, protein–protein interaction

*These authors have contributed equally to this work.

augmented with data for post-translational modification. In this case, quantification data may vary across peptides otherwise assigned to the same protein. Here again, pathway tools that rely solely on gene ID fail to capture the full information content of MS-based proteomic data. As a result of these and other limitations, practitioners typically use proteomic-specific toolsets at the early phases of their data analyses and then switch to more general-purpose software in an effort to place their primary data and observations into a biologically meaningful context. Although some tools [7] have tried to bridge the gap, their functionality is typically limited to a physical–chemical analysis of the inferred proteins, which, while being potentially useful, *e.g.* for optimization of the experimental protocols, is far less relevant for interpretation of the data in the context of biological pathways. With these considerations in mind, we set out to develop Pathway Palette (Figs. 1 and 2), a freely accessible, rich internet application that provides a canvas on which researchers can iteratively explore biological networks, while retaining all peptide-centric annotation of the source proteomic data. Below, we describe the design philosophy and software architecture, along with the logistics of data input and navigation of the Pathway Palette canvas.

We designed Pathway Palette to (i) provide seamless integration of experimental results with existing, publicly available data resources, (ii) enable iterative personalized knowledge refinement without a requirement for user registration or tracking of specific user history, and (iii) efficiently divide computational tasks between client- and

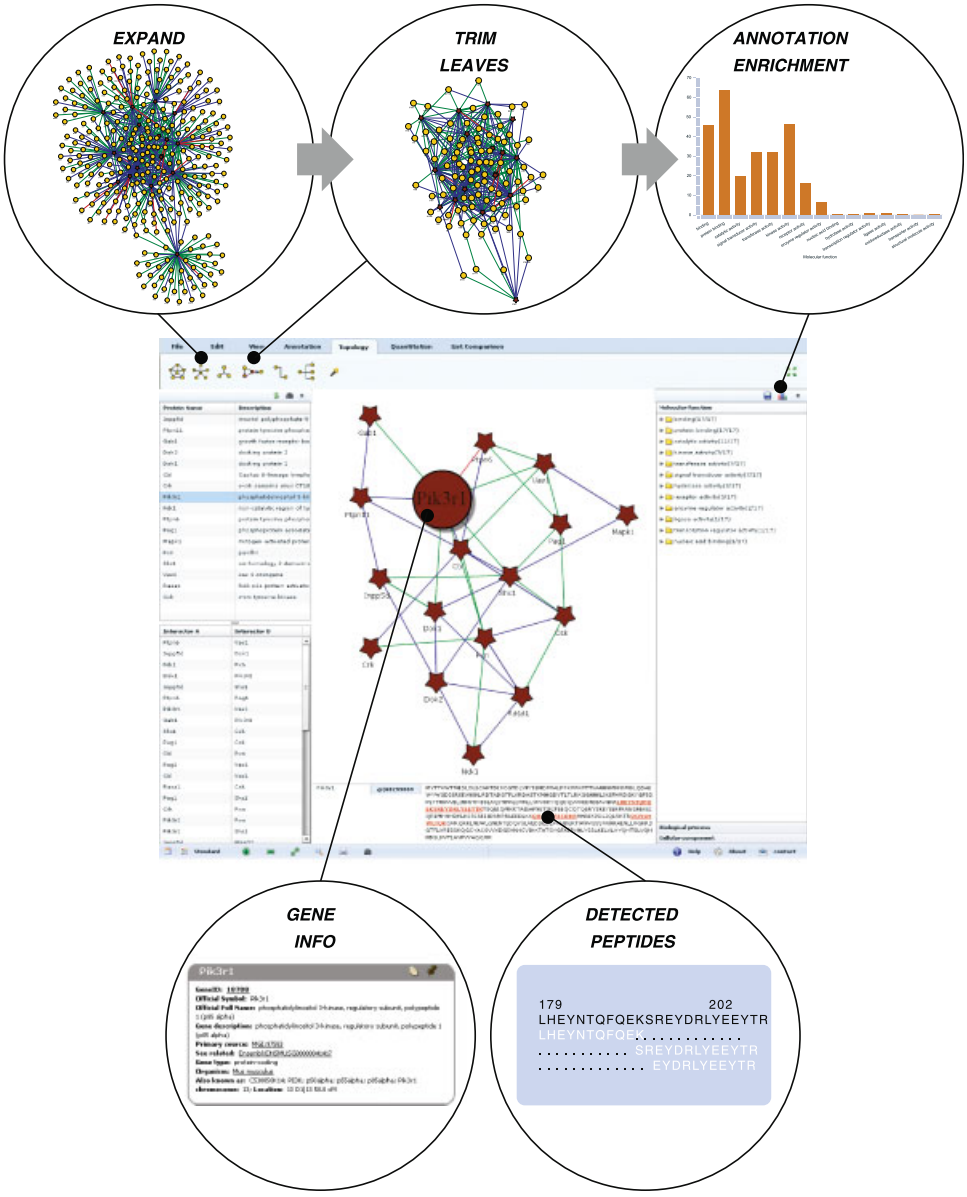

Figure 1. Upon data upload, Pathway Palette creates a protein-interaction network as the primary view (center panel). Users can click on specific nodes to obtain information rich ''gene cards'' (bottom left), and explore detected peptides and protein sequence coverage (bottom right). In addition, researchers can modify the network topology through, for example, addition of edges based on known PPIs (top left) and elimination of leaf nodes from the expanded graph (top center). At any point, users can query the current network view for enriched annotations, including generation of a significance bar plot (top right).

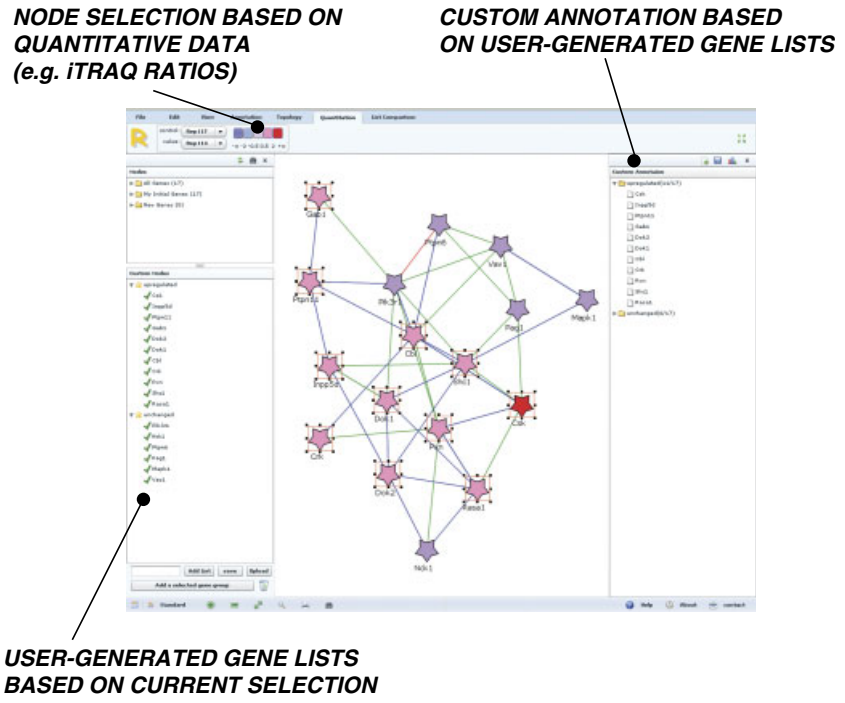

Figure 2. Data for relative quantification, in this case based on iTRAQ, may be uploaded from multiplierz-formatted spreadsheets and visualized as a color gradient. Users can then select nodes based on discrete ratio boundaries expressed as $\log_2$ ratios (top left). Selected nodes, based on ratio, topological features, *etc.*, can be captured as a gene list (bottom left). Individual or sets of user-defined gene lists can then be downloaded as a human-readable text files. These gene lists can in turn be used as custom annotations (top right), which can then be analyzed for enrichment, in a manner analogous to that used with the standard annotation sources (*e.g.* GO, KEGG and OMIM).

server-side resources. As Pathway Palette is built upon a foundation of database repositories that are keyed by a wide range of annotations, including gene, protein, disease, and pathway, there is already a fundamental benefit in simply making this information available through a convenient, unified interface. However, our broader objective is for researchers to use Pathway Palette as a navigational tool to iteratively build new knowledge, typically in the form of new gene lists, annotations, or novel connections between otherwise unrelated molecules, in biological pathways and networks.

An implicit requirement to establish an enthusiastic and significant user base for Pathway Palette is the efficient management of protein lists. Although seemingly a trivial task, we have made two design choices related to list management which, we believe, will encourage broad community participation and ultimately publication of analyses that leverage Pathway Palette functionality. First, use of Pathway Palette does not require login credentials nor is user history tracked within or between sessions. In addition, simple text files can be used as input. As a result, users may anonymously re-visit previous analyses in an iterative manner, with their history and strategy stored client-side within human-readable files. Second, Pathway Palette treats user-created lists as first class annotation sources. This means that users can upload lists, created from primary data, directly into the very same analytic framework used to assess significance of biological annotations such as GO [8], KEGG [9], and OMIM [10]. This enables users to quickly detect over-representation of lists created in the previous experiments, and hence validate that new knowledge has been generated. We believe that Pathway Palette is the first freely and anonymously accessible proteomic tool that provides this form of cyclical knowledge refinement.

Pathway Palette is delivered to the user in the form of a rich internet application deployed through an architecture that combines a Flex/Flash front-end and a backend built on PHP augmented with Python scripts, as well as a dedicated graph layout server implemented as a java servlet (http://www.yworks.com) from an Apache/Tomcat/Zend server (http://www.apache.org/, http://www.zend.com) (Fig. 3). Applet-based solutions which compute graph layouts on the client-side were not considered since protein-interaction graphs can be prohibitively large and are not necessarily amenable to layout on a modern netbook or legacy desktop. In addition, it is well recognized that Flash has a much higher installation base in the modern internet browser ecosystem (http://riastats.com/, http://www.statowl.com/) as compared with Sun's Java JDK/JRE (http://www.java.com).

Pathway Palette is intended to provide primary support for proteomic research and therefore enables direct upload of data elements that are unique to MS. In principle, these input data could be based on community-defined standards such as mzIdentML (http://www.psidev.info/index.php?q = node/319), however, since the associated definitions have not yet stabilized (http://www.psidev.info/index.php?q = node/408), we opted in the first deployment of Pathway Palette to use human-readable text and spreadsheets for data input. Two primary modes of input are available to the user:

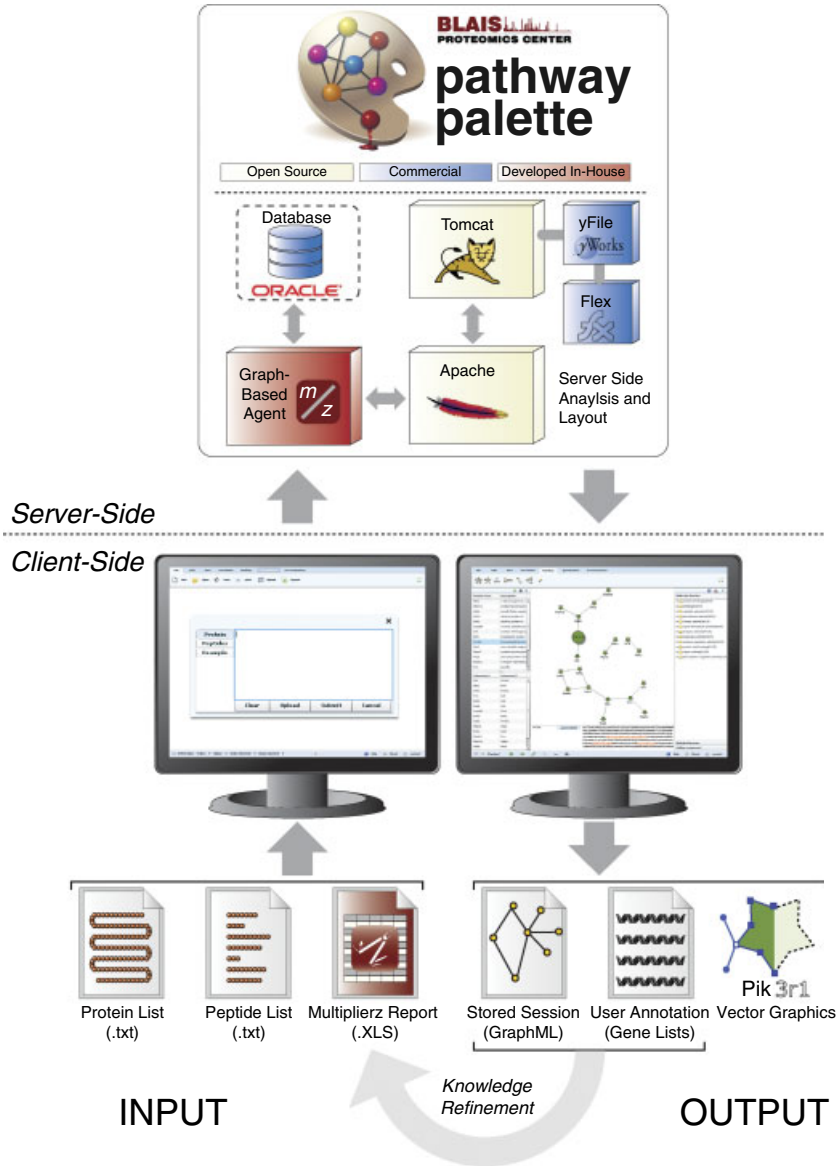

Figure 3. Pathway Palette features a robust web architecture that is driven server-side (top) by an Oracle database and powerful graph agent for execution of computationally intensive operations that are delivered to the client (bottom) through a rich Adobe Flex front-end. Users can upload data in a variety of formats (bottom left), including peptide sequences, protein accession lists, Entrez Gene Symbols and quantitative proteomic data from multiplierz-formatted spreadsheets. Users can also upload custom gene lists defined in the previous Pathway Palette sessions (bottom right), thus enabling an intuitive, and anonymous, knowledge refinement cycle. Finally, Pathway Pallet generates vector-compatible output to facilitate production of publication quality protein network graphics.

(i) The most unstructured form of input is a list. The list can either be a peptide list, a protein list, or a gene list. This mode also enables users to cut and paste directly from their usual data sources into the Pathway Palette browser input fields.

(ii) Users may also upload structured spreadsheets generated by multiplierz, our recently described, mzAPI-based framework for MS data analysis [11] (http://blais.dfci.harvard.edu/multiplierz). These spreadsheets can also be used in mode one above, through copy and paste of peptide or protein data into the website.

Direct upload of the latter provides convenient and comprehensive support for quantification data, meaning that in subsequent network analyses (see below) users can readily inspect all peptide identification and quantification evidence for each protein node. For either input mode described above, Pathway Palette will accept peptide sequences, protein identifiers from RefSeq [12] (*e.g.* gi|62362414) or SGD [13] (*e.g.* YHR023W), and gene symbols based on the NCBI's Entrez Gene [14] (currently, the system supports the Human, Mouse and Yeast proteomes). In cases where the input consists of peptide sequences, users must specify whether the system should generate the largest set of proteins consistent with the input list, or use a greedy algorithm to calculate a minimal set. These options constitute the two extreme interpretations of the provided peptide list: either every protein containing any of the given peptides is listed, or the system iteratively selects a protein which accounts for the largest number of peptides, reports that protein and then eliminates the constituent peptides from subsequent iterations.

The Pathway Palette interactive canvas facilitates functional analysis of MS data within the context of a protein network. Upon data upload, Pathway Palette generates a PPI network based on information curated by BioGRID [15]. Furthermore, we reduced the 25 interaction types in the BioGRID ontology into four major subsets: (i) Low-throughput techniques (green), (ii) HTP/complex detection, *i.e.* pull-downs (blue), (iii) HTP/pairwise, *i.e.* yeast-two-hybrid (red), and (iv) Genetic interactions (not selected by default). Other data sources that are made available include: (v) NCBI Entrez database (PPI entries from: HPRD [16], BioGRID and BIND [17]), (vi) interologs [18] (putative interactions through homology), as well as (vii) the top 10% of the STRING [19] predicted PPI database. The system also supports input of user-defined PPI data as a simple tab-delimited list of Entrez Gene ID pairs. Pathway Palette enables the user to filter, prioritize, and color-code edges based on the underlying data source (rather than displaying all data sources for a given edge, Pathway Palette colors edges based on the highest priority data source containing the edge in question). In the case of peptide sequence input, the resulting nodes are colored (i) green if they represent a gene uniquely identified by peptide evidence, (ii) red if they can be removed from the plot without loss of peptide evidence, or (iii) blue if they are not uniquely supported by the peptide evidence.

The user can extend the network further by interactively selecting nodes and incorporating known interactors not present in the original data set. Pathway Palette provides graph–theoretic operations which help manage exploration of the network, including: (i) extension of the network to include new nodes based on various PPI data sources, (ii) removal of leaf nodes, (iii) retrieval of missing edges, (iv) shortest path calculations, (v) Steiner Tree generation [20], among others. The resulting network is user-editable, from the manual addition and removal of nodes, through graph layout algorithms, to manipulation of individual node shape, color, outline, *etc*. The resulting graphs can be stored as GraphML files [21] on the user's desktop (for later upload) or printed in publication quality (vector graphics compatible) output. We recognize that much of the functionality of Pathway Palette is available in Cytoscape [22], a platform renowned for its advanced features and plugin architecture. While locally hosted software environments are suitable for many tasks and research environments, we chose to deploy Pathway Palette from a central server to alleviate administrative overhead for casual or other users who lack the necessary skills or infrastructure to support local instances of Cytoscape or similar platforms.

Although the PPI network view constitutes the central data view in Pathway Palette, two additional representations are available which rely directly on MS as the primary data source:

(i) *Peptide view*: allows the user to inspect the exact peptides upon which a given protein (and hence gene) identification is based. The protein sequence is shown with the areas covered by peptide reads highlighted and underlined. By clicking on a highlighted region, all relevant peptides covering the region are shown (''coverage drill-down'').

(ii) *Quant-Network*: At any point in the analysis, by navigating to the quantitation tab, the user can color proteins in the PPI network by fold change based on the uploaded spreadsheet data. Ratio values can be used to select proteins and can therefore serve as a basis for user-defined protein-lists and custom annotations (Fig. 2).

In addition to network visualization and exploration, Pathway Palette supports functional annotation of protein networks. Enrichment analysis (represented graphically as a bar plot of negative log *p*-values) can be generated for KEGG, OMIM, and GO annotations (slim version [23]). In addition to these standard annotation sources, users can upload previously curated protein lists which can effectively be used as a private form of annotation, accessible through the same tools as the standard annotation sources. Finally, Pathway Palette includes additional functions for comparison of protein lists as well as other miscellaneous operations that are beyond the scope of this brief report. Detailed descriptions of the functionality provided by the system as well as detailed tutorials are all provided on the accompanying web-page: http://blaispathways.dfci.harvard.edu/tutorial.

Pathway Palette is a rich, web-based resource which provides researchers with a convenient means to place mass spectrometry-based proteomic data within the context of biological pathways and networks. Support for anonymous user access along with server-side deployment of computationally intensive operations will lower barriers for community participation and ensure that the scale of graph analysis and network visualization is not limited by client-side computational resources. We will continue to engineer improvements to the website, driven by our own research needs in addition to user suggestions for future functionality.

The authors thank Eric Smith for preparation of figures, and also members of the Marto Lab for valuable discussions and testing of Pathway Palette. This work was supported by the Dana-Farber Cancer Institute and the National Human Genome Research Institute (P50HG004233).

The authors have declared no conflict of interest.

References

[1] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198–207.

[2] Patterson, S. D., Data analysis-the Achilles heel of proteomics. *Nat. Biotechnol.* 2003, *21*, 221.

[3] Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R. *et al.*, Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001, *292*, 929–934.

[4] Gavin, A., Bosche, M., Krause, R., Grandi, P. *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, *415*, 141–147.

[5] Yan, W., Lee, H., Yi, E., Reiss, D. *et al.*, System-based proteomic analysis of the interferon response in human liver cells. *Genome Biol.* 2004, *5*, R54.

[6] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[7] Park, D., Kim, B., Cho, S., Park, S. *et al.*, MassNet: a functional annotation service for protein mass spectrometry data. *Nucleic Acids Res.* 2008, *36*, W491–W495.

[8] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000, *25*, 25–29.

[9] Kanehisa, M., Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000, *28*, 27–30.

[10] Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C. *et al.*, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2002, *30*, 52–55.

[11] Parikh, J., Askenazi, M., Ficarro, S., Cashorali, T. *et al.*, Multiplierz: an extensible API based desktop environment for proteomics data analysis. *Biomed. Chromatogr. Bioinformatics* 2009, *10*, 364.

[12] Pruitt, K. D., Tatusova, T., Maglott, D. R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007, *35*, D61–D65.

[13] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A. *et al.*, SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* 1998, *26*, 73–79.

[14] Maglott, D., Ostell, J., Pruitt, K. D., Tatusova, T., Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2007, *35*, D26–D31.

[15] Stark, C., Breitkreutz, B., Reguly, T., Boucher, L. *et al.*, BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006, *34*, D535–D539.

[16] Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N. *et al.*, Human protein reference database – 2006 update. *Nucleic Acids Res.* 2006, *34*, D411–D414.

[17] Bader, G. D., Betel, D., Hogue, C. W. V., BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 2003, *31*, 248–250.

[18] Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X. *et al.*, Annotation transfer between genomes: protein–protein interologs and protein-DNA regulogs. *Genome Res.* 2004, *14*, 1107–1118.

[19] von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S. *et al.*, STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 2003, *31*, 258–261.

[20] Scott, M. S., Perkins, T., Bunnell, S., Pepin, F. *et al.*, Identifying regulatory subnetworks for a set of genes. *Mol. Cell. Proteomics* 2005, *4*, 683–692.

[21] Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M., *Graph Drawing* 2002, 109–112.

[22] Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A. *et al.*, Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* 2007, *2*, 2366–2382.

[23] Barrell, D., Dimmer, E., Huntley, R. P., Binns, D. *et al.*, The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* 2009, *37*, D396–D403.

