

Facilitating the use of large-scale biological data and tools in the era of translational bioinformatics

Irene Kouskoumvekaki, Nour Shublaq and Søren Brunak

Submitted: 7th May 2013; Received (in revised form): 9th July 2013

Abstract

As both the amount of generated biological data and the processing compute power increase, computational experimentation is no longer the exclusivity of bioinformaticians, but it is moving across all biomedical domains. For bioinformatics to realize its translational potential, domain experts need access to user-friendly solutions to navigate, integrate and extract information out of biological databases, as well as to combine tools and data resources in bioinformatics workflows. In this review, we present services that assist biomedical scientists in incorporating bioinformatics tools into their research. We review recent applications of Cytoscape, BioGPS and DAVID for data visualization, integration and functional enrichment. Moreover, we illustrate the use of Taverna, Kepler, GenePattern, and Galaxy as open-access workbenches for bioinformatics workflows. Finally, we mention services that facilitate the integration of biomedical ontologies and bioinformatics tools in computational workflows.

Keywords: translational bioinformatics; data integration; biological networks; scientific workflows; biomedical ontologies

INTRODUCTION

Recent technological advances, such as Next Generation Sequencing, have contributed to an explosion of high-throughput biological data [1], which has been accompanied by an equal outbreak in databases and bioinformatics tools for life sciences research [2–4]. Supported by an equal increase of processing computer power, new opportunities have arisen in relation to the data analysis and formation of new biological hypotheses. Computational experimentation is gradually embraced by the biomedical community, and, likewise, bioinformatics is moving from a provider of discoveries confined to basic biological research, to a provider of solutions that can directly affect clinical care, as captured by the term ‘translational bioinformatics’ [5].

For bioinformatics to realize its translational potential, domain experts with limited computing background and scripting experience should have access to user-friendly solutions to navigate, integrate and extract information out of biological databases. Moreover, the use of bioinformatics tools for data analyses and combination of tools and data resources in scientific workflows should be facilitated for researchers that are not accustomed or always pre-trained to handling different computational tools and working environments.

In this review, we present services that assist biomedical scientists in incorporating bioinformatics tools into their research. The services span from aggregates of tools that visualize, enrich and integrate biological information to services that act as

Corresponding author. Irene Kouskoumvekaki, Department of Systems Biology, Center for Biological Sequence Analysis (CBS), Technical University of Denmark (DTU), DK-2800 Kgs. Lyngby, Denmark. Tel.: +45-4525-6162; Fax: +45-4593-1585; E-mail: irene@cbs.dtu.dk

Irene Kouskoumvekaki is Associate Professor at the Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark (DTU) with expertise in chemoinformatics and systems chemical biology.

Nour Shublaq is Senior Research Associate at the Centre for Computational Science of University College London (UCL), and the Strategic Coordinator of the UCL Computational Life and Medical Sciences Network. Her research interests and expertise span computational biomedicine, biomedical engineering and biomedical informatics.

Søren Brunak is the founding Director of CBS at DTU and Professor of Disease Systems Biology at the Novo Nordisk Foundation for Protein Research. His current research interests include text mining for biomedical research and heterogeneous data integration in systems biology.

computational workbenches for running bioinformatics workflows. We focus on services that are open access, have a user-friendly Graphical User Interface and are the most embraced by the biomedical research community. For each service, we mention the type of data and tools it supports, and we give examples from the current literature of its application in medical research. As the key for the translation of bioinformatics to the clinic lies in the integration of phenotype information that resides in the clinical world (i.e. disease diagnoses, therapeutic effects of drugs, drug side effects, patient medical history) with high-throughput experimentation and data analysis, we have included in the review biomedical ontologies that are incorporated in the aforementioned services and assist in the integration of phenotype and genotype data.

Services for visualization, integration and enrichment of biological information

With the plethora of biological data being currently generated, the reductionist approach of looking at a single gene at a time has become outdated. We now have the opportunity to get a holistic picture of a biological system (be it an organism, a tissue, a cell or a disease pathway) and study its emergent properties as a result of the interactions and dependencies among its components. This requires the ability to structure the data in a way that would make it possible to visualize them, merge them with reference or other types of data or get all information currently available (e.g. in publicly available databases) about the studied system. Various computational tools have been developed for the visualization, integration, functional annotation and enrichment of biological networks such as protein–protein interaction (PPI), enzyme–substrate interaction (metabolic) and DNA–protein interaction (gene regulatory) networks [6]. Later in the text, we focus on three open-access services, Cytoscape, BioGPS and Database for Annotation, Visualization and Integrated Discovery (DAVID) that we consider as the most prominent in the network biology arena, which aggregate computational tools for bioinformatics analyses (Figure 1). Table 1 summarizes recent examples of the use of these services for biological data visualization, integration and information enrichment by the biomedical research community. As shown in the table, biomedical scientists use Cytoscape, BioGPS and DAVID for the elucidation of the macroscopic

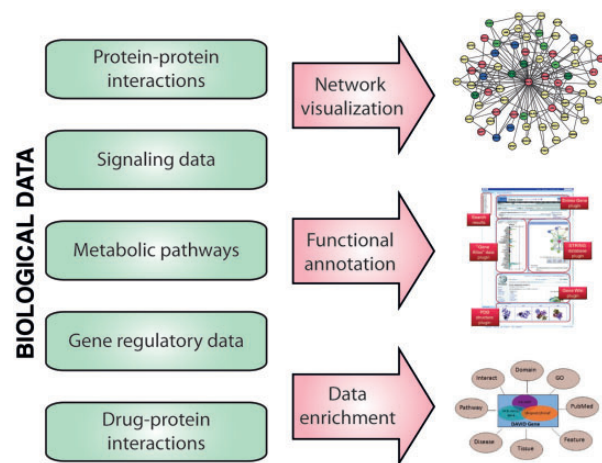


Figure 1: Open-access services such as Cytoscape, BioGPS and DAVID provide tools for the bioinformatics analysis of biological data, such as PPIs, gene regulation and drug–protein interactions. The integration, functional annotation and enrichment analysis of such data offer a bird’s eye view of a biological system (e.g. a disease pathway) and allow the study of its emergent properties as a result of the interactions among its components. The two lower schematic illustrations at the right side of the figure have been adapted from the corresponding publications [7,8].

structures and functions of biological systems via bioinformatics analyses of biological data at the genomic, transcriptomic, proteomic and metabolomic levels. The applications cover a wide spectrum of diseases, from cardiovascular to cancers and genetic diseases.

Cytoscape

Cytoscape, now in its 2× version, is perhaps the most popular bioinformatics package for biological network visualization and integration of complex data sets [30]. Cytoscape creates graphs where the nodes represent the entities of interest (genes, proteins, drugs, etc.) and the edges the relationships (interactions) between them. However, Cytoscape offers much more than complex network visualization. The core Cytoscape framework is extendable through the implementation of a vast variety of Apps (formerly called plugins): software components that add specific capabilities to the core Cytoscape service, the majority of which have been developed to serve the biomedical community. These Apps usually originate from independent software developers and range from literature mining tools (e.g. AgilentLiteratureSearch [31]) to advanced network analysis (e.g. MCODE [32]) and biological data enrichment tools

Table 1: Recent examples of the use of open-access tools for biological data visualization, integration and enrichment

Disease area	Research scope	Tool (Apps/plugins)	References
Cryptorchidism	Genomic, transcriptomic and proteomic data integration for the construction of disease genes atlas	Cytoscape (<i>PiNGO</i> [9]) and DAVID (<i>Functional annotation chart</i>)	Cannistraci et al. [10]
Cardiovascular	Network based on signaling, gene expression and drug-target data for the identification of off-target effects of torcetrapib	Cytoscape (<i>BisoGenet</i> [11], <i>NetworkAnalyzer</i> , <i>ClusterONE</i> [12]) and DAVID (<i>Gene functional classification</i>)	Fan et al. [13]
Pterygium	Pathway analysis of mRNA data to elucidate the mechanism of action of doxycycline	Cytoscape (<i>Reactome FI</i> [14], <i>Genemania</i> [15])	Larráyoiz et al. [16]
Type 2 diabetes	Global pathway analysis for the identification of dysregulated genes	Cytoscape (<i>jActiveModules</i> , <i>NetworkAnalyzer</i>)	Skov et al. [17]
Wound healing defects	Functional analysis of differentially expressed genes to elucidate the involvement of Factor XIII in wound response	Cytoscape (<i>BiNGO</i> [18])	Töröcsik et al. [19]
Low-grade serous carcinoma	Pathway analysis of differentially expressed genes to identify candidates related to carcinogenesis	Cytoscape (<i>IntAct</i>)	May et al. [20]
Atherosclerosis	Disease gene association network and pathway analysis based on transcriptomics data	Cytoscape (<i>ExpressionCorrelation</i> , <i>AgilentLiteratureSearch</i> , <i>NetworkAnalyzer</i> , <i>jActiveModules</i> , <i>BiNGO</i> , <i>ClueGO</i> [21], <i>Golorize</i> [22])	Diez et al. [23]
Ischemic dilated cardiomyopathy	Layered PPI network to reveal novel drug targets	Cytoscape (<i>BionetBuilder</i> , <i>NetworkAnalyzer</i> , <i>Cerebral</i> , <i>GenePro</i>) and DAVID (<i>Functional annotation clustering</i>)	Zhu et al. [24]
Multiple sclerosis	PPI network analysis to identify disease-relevant loci	BioGPS (<i>Data set Library</i>) and DAVID (<i>Functional annotation chart</i>)	Ragnedda et al. [25]
Breast cancer	Comparison of tumor gene expression profiles for biomarker discovery	BioGPS (<i>Data set Library</i>)	Pennings et al. [26]
Trisomies 18 and 21	Analysis of transcriptome data for the identification of biomarkers of abnormal human neurodevelopment	BioGPS (<i>Gene Expression Atlas</i>) and DAVID (<i>Functional annotation clustering tool</i>)	Hui et al. [27]
Pregnancy-related complications	Analysis of transcriptome data to retrieve functional information about fetal cells	BioGPS (<i>Gene Expression Atlas</i>)	Pritchard et al. [28]
Pharmacogenomics	Comparison of mRNA expression differences between resistant and sensitive cell lines to identify SNPs related to paclitaxel response	BioGPS (<i>Data set Library</i>)	Eng et al. [29]

(e.g. BiNGO [18]). The aforementioned examples constitute the most popular Apps based on the total number of downloads, as estimated by Saito et al. [33] in a recent review of the landscape of Cytoscape Apps. Users of Cytoscape most typically apply the Apps on their own experimental interaction data to visualize them in a user-friendly format that allows them to form or rule out certain biological hypotheses. In addition, one can enrich new experimental data with data and auxiliary information from publicly available databases to get a holistic or more comprehensive picture of the biological system in question. Since early 2013, Cytoscape has

been in the middle of a transition to a whole new version, v3.x, with modular architecture and new user interfaces.

BioGPS

BioGPS is a centralized gene portal with aggregated federated gene annotation resources [7]. Each resource is gene-centric and is wrapped within a plugin. The BioGPS library currently contains >300 active plugins provided by the research community, which are grouped into categories such as genetics, literature, pathway, expression and protein, to name a few of the most popular. Each plugin can

belong to more than one category and can be queried either on its own or combined with others in custom layouts. For example, for a given candidate gene list, the researcher can create custom layouts by adding the plugins of interest (e.g. *Gene Identifiers* plugin, *KEGG (human)* plugin for pathway maps from KEGG [34], *STRING-Proteins and their Interactions* plugin for PPI data from the STRING database [35] and *Gene expression/activity chart* plugin with reference gene expression data sets), which will then aggregate the respective information in a single output. However, owing to the lack of commonly used identifiers among the plugins, it is not possible to combine pieces of data from multiple plugin sources into a single analysis or visualization [36]. In addition to plugins, BioGPS contains a Data Set Library of >2000 reference data sets grouped by disease categories that can be browsed via the Gene expression/activity chart plugin.

DAVID

The DAVID contains pre-packaged modules dedicated to the systematic and comprehensive functional annotation of gene lists [37]. A module in DAVID is a standardized set of bioinformatics tools that have been assembled together to carry out a specific task. DAVID is built on a centralized Knowledgebase [8], which aggregates heterogeneous data from a variety of resources in a centralized location. During the process of data aggregation, all gene identifiers are converted to internal DAVID identifiers, which make it possible to cross-reference information among the different data resources, spot redundant information and make use of all available data for an input gene. For a given gene ID, the user can view the gene name with the Gene name batch viewer, perform functional classification with the corresponding module [38], do enrichment analysis with the use of the Functional annotation chart module or conduct further functional annotation analyses.

Services for bioinformatics workflows

Bioinformatics workflows are assemblies of bioinformatics tools (commonly referred to as modules) and biological data, which when combined in a structured set of steps, result in automated or semi-automated procedures for providing answers to biological questions. A bioinformatics workflow can be repeated on the same or a different project, which facilitates data sharing and research reproducibility. As illustrated in Figure 2, through the use of

workbenches for bioinformatics workflows, computational analysis of different types of biological data [e.g. DNA, exome or ChIP (chromatin immunoprecipitation) sequencing, flow cytometry or small molecule interaction screen data] can be effectively translated to outputs with direct application to the clinic. Later in the text, we discuss some of the most popular open-access workbenches for bioinformatics workflows (Taverna, Kepler, GenePattern and Galaxy), and we illustrate recent examples of their use by the biomedical research community (Tables 2 and 3).

Taverna

Taverna [73] is an open-source workflow management system that integrates tools and data resources in scientific workflows. The newest version includes an Interaction Service plugin, which allows for user interaction as an integrated part of any given scientific workflow. Taverna is a general-purpose system, and as such, it does not come with pre-packaged tools and can work with any data type; it has, however, been particularly embraced by scientists working in the field of systems biology. As of version 2.1.2, Taverna has a built-in support for browsing BioCatalogue (<http://www.biocatalogue.org>) and myExperiment (<http://www.myexperiment.org>) in the form of plugins that can be installed in the Taverna Workbench. BioCatalogue is a public, centralized, curated and monitored registry of Life Science Web services that can be used for building workflows. myExperiment is a scientific networking site designed for sharing, discovering and re-using workflows [74] for computational experiments and currently hosts >1800 workflows designed for Taverna, primarily in the field of bioinformatics [75]. Two of the most popular—in terms of times they have been downloaded—are the workflow on ‘Pathways and Gene annotations for QTL region [76]’ that uses BioMart [77] and KEGG [34] and the ‘Lymphoma type prediction based on microarray data [78]’ that uses caArray [79]. In a recent review article on scientific workflow management in proteomics, De Bruin *et al.* [80] illustrate, in four different case studies, how tools in Taverna can be incorporated in workflows for carrying out complex data analysis tasks in proteomics.

GenePattern

GenePattern [81] is a workbench for bioinformatics research that follows the concept of Reproducible Research Systems [82]. GenePattern provides access

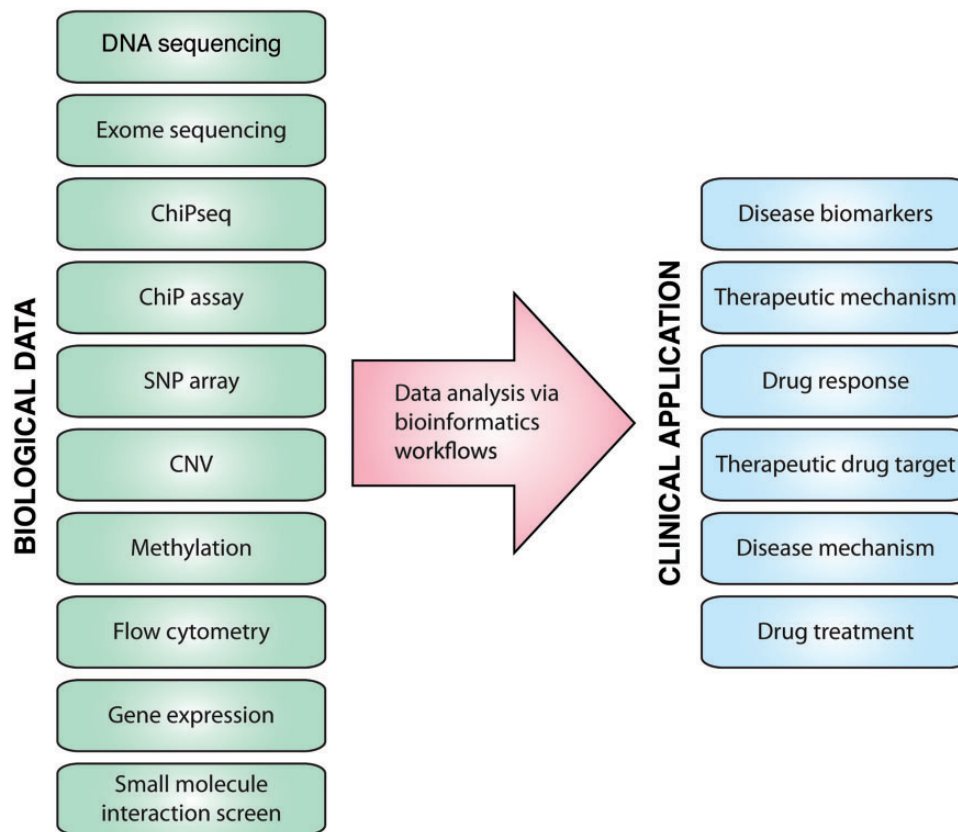


Figure 2: Workbenches for bioinformatics workflows provide the framework for the computational analysis of large-scale genetic data, and–omics data in general. Via assemblies of bioinformatics tools and reference databases combined in a structured set of steps, one can develop automated or semi-automated workflows for answering questions of direct clinical relevance.

to >230 tools for gene expression analysis, proteomics, Single-nucleotide polymorphism (SNP) analysis, flow cytometry, RNA-seq analysis and common data processing tasks. In addition, it provides access to reference genome data for a number of species, via the TopHat, Bowtie and Burrows–Wheeler Aligner (BWA) modules. A web-based interface ensures easy access to these tools and allows the creation of multi-step analysis workflows that enable reproducible computational research. GenePattern has been used extensively in biomedical research and has contributed to high impact projects on a variety of disease areas, as illustrated in the examples of Table 2. The work of Weigert *et al.* is an excellent example of using tools from GenePattern with the aim to identify mutations that confer resistance to JAK–2 inhibitors [52]. The authors combined the Robust Multi-array Average, the Variation Filter and the Comparative Marker Selection tools for differential analysis of the gene expression data from leukemia cell lines providing evidence that three mutations within the JAK2 kinase domain confer resistance across a panel of JAK inhibitors.

Galaxy

A more recent development in high-throughput analysis of biological data with weight put on research reproducibility is Galaxy [83]. The public Galaxy server has a pre-packaged toolbox for a variety of sequence analyses, which is easily extensible to host additional tools. Galaxy’s main page is divided into three sections: tools (left side), analysis (middle) and history (right side). Users can either import data sets from reference data warehouses or upload their own. When Galaxy is installed locally, users have the possibility to add new tools or work with data sets that are either confidential or too large to be processed on the public server. Research reproducibility is ensured by the Share Data service, where the users can store and make their own data libraries and workflows publicly available or have access to those that others have made public. A detailed comparison of the functionalities of Galaxy in relation to other systems is given in Goecks *et al.* [83]. Recent examples of workflows developed for Galaxy and Applications of Galaxy in biomedical research are summarized in Table 3. Galaxy workflows

Table 2: Selected examples of the use of GenePattern in biomedical research

Disease area	Research scope	Tools	References
T-ALL	High-resolution comparative genomic hybridization array for the study of molecular pathogenesis of T-ALL	Preprocess Data set, Comparative Marker Selection [39]	Gutierrez et al. [40]
Diabetic neuropathy	Gene expression profiles associated with disease progression	Robust Multi-array Average	Hur et al. [41]
Multiple sclerosis	Large scale immunophenotyping for disease biomarker identification	Comparative Marker Selection, Non-negative Matrix Factorization, SVM	De Jager et al. [42]
Leukemia	Epigenetic profiling for therapeutic target identification	Not stated	Bernt et al. [43]
Pancreatic cancer	DNA microarray and sequencing-based analysis to identify mechanism of tumor suppression	GISTIC [44]	Shain et al. [45]
Influenza virus	Genome-wide transcriptional profiling of lung ILCs for the study of lung tissue homeostasis following virus infection	ClassNeighbors	Monticelli et al. [46]
Lung adenocarcinoma	Characterization of copy number alterations in primary lung adenocarcinoma	PM-MM Difference Model, dChipSNP	Weir et al. [47]
Wilms tumor	Comparison of genome-wide chromatin profiles for the identification of transcriptional and epigenetic mechanisms as disease drivers	Not stated	Aiden et al. [48]
Gastric cancer	Gene expression analysis of cancer cell lines to identify subtypes associated with survival and drug response	Nearest Template Prediction [49]	Tan et al. [50]
Tumors	Genetic screen to identify genes that promote target cell resistance to human natural killer cell-mediated killing	Comparative Marker Selection	Bellucci et al. [51]
JAK-2 driven cancers	Gene expression profiling for the identification of mutations that confer resistance to JAK-2 inhibitors	Robust Multi-array Average, Variation Filter, Comparative Marker Selection	Weigert et al. [52]
Hepatocellular carcinoma	Genome wide expression profiling of fixed cancer tissues to discover a gene-expression signature associated with survival	IlluminaDASL, Nearest Template Prediction, Survival Analysis	Hoshida et al. [53]
T-ALL	Gene expression profiling for the elucidation of TLX1 oncogenes role in disease pathogenesis	Comparative Marker Selection	De Keersmaecker et al. [54]
Metabolic syndrome	Study of the anti-inflammatory properties of regulatory T-cells based on gene expression data	Robust Multi-array Average	Feuerer et al. [55]
Glioblastomas	Study of the interplay between genomic and epigenomic changes in cancers and identification of mutation associated with glioma hypermethylator phenotype	Not stated	Turcan et al. [56]
Myelodysplastic syndrome	Gene expression profiling to identify signatures associated with drug response	Comparative Marker Selection, Revlimid	Ebert et al. [57]
Type-I diabetes	Coupling interaction screen data with gene expression analysis for associating human genetic variation to small molecule sensitivity	Comparative Marker Selection, Gene Set Enrichment Analysis	Shaw et al. [58]
Head and neck squamous cell carcinoma	Whole-genome sequencing data analysis to unravel the mutational landscape of disease	Firehose pipeline	Stransky et al. [59]

T-ALL, T-cell acute lymphoblastic leukemia; ILC, innate lymphoid cell; JAK-2, Janus kinase 2; kNN, k-nearest neighbors; SVM Support Vector Machine, GISTIC: Genomic Identification of Significant Targets in Cancer.

Table 3: Recent examples of the use of Galaxy in biomedical research

Disease area	Research scope	Tool	References
Congenital heart disease	Generation of genome-wide CNV data to study the contribution of global rare CNVs to disease risk	Genomic Intervals	Soemedi <i>et al.</i> [60]
Myoclonus epilepsy	Sequence enrichment array for the identification of mutations associated with disease	Not stated	Corbett <i>et al.</i> [61]
Synovial sarcoma	Genome-wide analysis to determine the role of oncogene in tumorigenesis	Coverage, Intersect	Garcia <i>et al.</i> [62]
Breast cancer	Elucidation of transcription factor involvement in estrogen receptor-mediated gene expression in breast cancer cells	Not stated	Ross-Innes <i>et al.</i> [63]
Asthma	GWAS of lymphocyte count and analysis of gene expression data to identify asthma susceptibility candidate genes	Not stated	Cusanovich <i>et al.</i> [64]
Atypical chronic myeloid leukemia	Analysis of exome sequencing data to identify new oncogene	BWA alignment [65]	Piazza <i>et al.</i> [66]
Cellular immunity	Development of method for high resolution HLA genotyping and allele discovery in large patient cohorts	BLAT [67] and custom tools and scripts	Lank <i>et al.</i> [68]
Cancer	Development of curated database of cancer stem cell experiments	GeneSigDB [69], MSigDB [70], WikiPathways [71]	Ho Sui <i>et al.</i> [72]

CNV, copy number variants; GWAS, genome-wide association study; BWA, Burrows-Wheeler Aligner; HLA, human leukocyte antigen; BLAT, the BLAST-like alignment tool; GeneSigDB, Gene Signature Database; MSigDB, Molecular Signatures Database.

have been used on a wide range of disease areas, from cancers to epilepsy and asthma, and have assisted in the analysis of biological data ranging from copy number variation (CNV) and gene expression data to exome and DNA sequencing. In the work of Garcia *et al.*, for example, the authors used the Coverage and Intersect tools to determine overlapping regions between ChIP-seq and MeDIP (methylated DNA immunoprecipitation) -ChIP data sets from the Gene Expression Omnibus repository and the SYT-SSX2 oncogene. SYT-SSX2 is associated with initiation and development of synovial sarcoma, and the overall aim of the study was to analyze the genome-wide binding of SYT-SSX2 and to elucidate its mechanism of control of gene expression networks [62].

A Taverna-Galaxy integration has been recently implemented (<https://trac.nbic.nl/elabfactory/wiki/eGalaxy>), which bridges the two by allowing Taverna workflows (i.e. all workflows available in myExperiment repository) to be converted to Galaxy tools and thus operate directly via the Galaxy interface [84].

Kepler

Last but not least, Kepler [85] is a well-established system for scientific workflow management with its

emphasis so far, however, on ecology and geology. Within life sciences research, it has been recently used by Stropp *et al.* as a workbench for the development of a series of workflows for the analysis of Nimblegen ChIP-chip microarray data and data from Affymetrix gene expression microarrays, as well as a workflow for the design of PCR primers for microarray validation experiments. Stropp *et al.* [86] provide a comprehensive comparison of Kepler, Taverna and Galaxy in relation to bioinformatics analyses in general, and the analysis of microarray data in particular.

Workflows in Kepler are depicted by graphs where the nodes (i.e. the computational components) are called ‘actors’ and the edges (i.e. the data flow) are called ‘channels’. The recent release of Kepler v-2.4 library currently contains over 350 ‘actors’ that can be customized and connected with one another. In addition, the Kepler webpage (<https://kepler-project.org>) provides a list of add-on modules (standardized suites of tools) that, when incorporated into Kepler, increase its functionality. For example, the Provenance add-on allows the recording of the execution history of a workflow, which facilitates the sharing and reproducibility of a computational experiment. BioKepler [87], a

module that is planned to be distributed on top of the core Kepler scientific workflow system, will serve the analysis of large scale biological data, specifically data generated by next-generation sequencing. BioKepler aims to offer workflow components for bioinformatics analyses with distributed and parallel processing capabilities.

Integration of biomedical ontologies in computational workflows

Bioinformatics has traditionally dealt with the computational study of biological entities, such as DNA, genes and proteins. From early on, the need to standardize entity names across species and databases became apparent, and it has led in 2000 to the development of the Gene Ontology [88], which currently consists of >30 000 terms [89], and is probably the most well-known example of a biomedical ontology. Ontologies provide a set of concepts that standardize the representation of a domain. Researchers can then use ontology concepts to store their data in a structured form, disseminate information in a computer-readable format, identify relevant information on other resources and express relationships among sets of concepts in a semantically interoperable way. Although substantial progress has been achieved in relation to the standardization of omics data [90], phenotype information that resides in the clinical world such as diseases, therapeutic interventions, side effects, patient information and laboratory measurements is still, to a large extent, disorganized. When it comes to adverse drug reactions, for example, there exists no generally accepted comprehensive dictionary and standardized ontology.

Unlike stand-alone software, where the user needs to manually modify the output of one tool to be accepted as an input of another, workflows greatly depend on the unhindered flow of data from one computational step to the next. Thus, the success and wider adoption of workbenches for bioinformatics workflows heavily rely on their ability to carry out data exchange and data integration across computational tools in a comprehensive and seamless manner. Efforts to create services that facilitate the integration of biomedical ontologies and bioinformatics tools in computational workflows have recently started to materialize.

The Open Biomedical Ontologies (OBO; <http://www.obofoundry.org>) Foundry [91] provides a web-based service for scientists to locate and use a variety of phenotype ontologies. Ontologies such as

the Foundational Model of Anatomy [92], The Mammalian Phenotype Ontology [93] and the Human Phenotype Ontology [94] are included. New ontologies undergo a peer review process before getting accepted in OBO Foundry, to evaluate whether they comply with the shared principles and common controlled vocabularies that will allow for their interoperability. The ONTO-ToolKit supports the analysis and handling of OBO-formatted ontologies via the Galaxy interface [95]. The Disease Ontology of the Disease and Gene Annotation plugin of BioGPS makes use of OBO for the controlled vocabulary of disease terms.

The National Center for Biomedical Ontology (NCBO) BioPortal [96] is a repository that provides access via web browsers and web services to different biomedical ontologies. In addition, it includes the NCBO Annotator (formerly referred to as the Open Biomedical Annotator [97]), which is an automated text tagger based on biomedical concepts from the Unified Medical Language Systems and NCBO ontologies. NCBO-Galaxy [98] provides BioPortal functionality via Galaxy.

CONCLUSION

The use of open access services that facilitate the computational analysis of biological data has been gaining ground in the field of biomedical research. It has brought bioinformatics tools closer to the end user and has enabled their adoption by researchers that are not experts in scripting. It has also enabled the analysis of vast amounts of genomics data that are being nowadays generated. Services for data integration, visualization and analysis of biological networks enable domain experts to visualize and extract knowledge out of large amounts of data. Additionally, one can use newly generated data together with data from publicly available reference databases. Access to large reference data allows scientific findings to be reflected in a context that incorporates the current knowledge in the field. Last but not least, services that facilitate the development of bioinformatics workflows can promote data sharing and research reproducibility.

There is increasing awareness that, for translational bioinformatics to fully materialize its potential in biomedical research, there is a strong need to develop biomedical ontologies that systematize biological data and create a common language of communication across different bioinformatics tools. Linking heterogeneous biological data is central in many biomedical

applications ranging from disease gene finding and disease susceptibility to biomarker discovery and drug response. Recent efforts towards this end include, for example, the integration of biomedical ontologies in workbenches for bioinformatics workflows, and ensuring consistent terminology across data and tools. Establishing biological data based on comprehensive dictionaries, ontologies and standards may sound like a costly, tedious and bureaucratic task, but it is essential for making the leap to the translation of bioinformatics to biomedical research.

Key Points

- The use of open access services that facilitate the computational analysis of biological data has been gaining ground in the field of biomedical research.
- Experts in different biomedical domains need access to user-friendly solutions to navigate, integrate and extract information from large-scale biological data, as well as to combine bioinformatics tools and data resources.
- Cytoscape, BioGPS and DAVID can be used for integration, network analysis and information enrichment of biological data at the genomic, transcriptomic, proteomic and metabolomic levels. Their applications cover a wide spectrum of diseases, from cardiovascular to cancers and genetic diseases.
- Some of the most popular open-access workbenches for bioinformatics workflows that provide the framework for the computational analysis of large-scale-omics data include Taverna, Kepler, GenePattern and Galaxy.
- For bioinformatics to fully realize its translational potential in biomedical research, the use and further development of biomedical ontologies should be facilitated for the data integration and exchange across tools in a coherent and seamless manner.

Acknowledgements

The corresponding author would like to thank Assoc. Professor Gianni Panagiotou for critically reviewing and providing feedback during the preparation of the manuscript and Melani Khodaie for her assistance on the figures.

FUNDING

INBIOMEDvision Consortium. INBIOMEDvision (Project No. 270107) is a Coordination Support Action funded by the European Commission under the Seventh Framework Programme (FP7/2007–2013).

References

- Shublaq N. A Strategic report on Genotype-Phenotype Resources in the European Union. 2012. (www.inbiomedvision.eu/PDF/Report-GenotypePhenotype-Final.pdf)
- Galperin MY, Fernández-Suárez XM. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res* 2012;**40**:D1–8.
- Fernandez-Suarez XM, Galperin MY. The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res* 2012;**41**:D1–7.
- Benson G. Editorial. *Nucleic Acids Res* 2012;**40**:W1–2.
- Altman RB. Translational bioinformatics: linking the molecular world to the clinical world. *Clin Pharmacol Ther* 2012;**91**:994–1000.
- Gehlenborg N, O'Donoghue SI, Baliga NS, et al. Visualization of omics data for systems biology. *Nat Methods* 2010;**7**:S56–68.
- Wu C, Orozco C, Boyer J, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009;**10**:R130.
- Sherman BT, Huang DW, Tan Q, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* 2007;**8**:426.
- Smoot M, Ono K, Ideker T, et al. PiNGO: a Cytoscape plugin to find candidate genes in biological networks. *Bioinformatics* 2011;**27**:1030–1.
- Cannistraci CV, Ogorevc J, Zorc M, et al. Pivotal role of the muscle-contraction pathway in cryptorchidism and evidence for genomic connections with cardiomyopathy pathways in RASopathies. *BMC Med Genomics* 2013;**6**:5.
- Martin A, Ochagavia ME, Rabasa LC, et al. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics* 2010;**11**:91.
- Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 2012;**9**:471–5.
- Fan S, Geng Q, Pan Z, et al. Clarifying off-target effects for torcetrapib using network pharmacology and reverse docking approach. *BMC Syst Biol* 2012;**6**:152.
- Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* 2010;**11**:R53.
- Montejo J, Zuberi K, Rodriguez H, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* 2010;**26**:2927–8.
- Larráyoz IM, de Luis A, Rúa O, et al. Molecular effects of doxycycline treatment on pterygium as revealed by massive transcriptome sequencing. *PLoS One* 2012;**7**:e39359.
- Skov V, Knudsen S, Olesen M, et al. Global gene expression profiling displays a network of dysregulated genes in non-atherosclerotic arterial tissue from patients with type 2 diabetes. *Cardiovasc Diabetol* 2012;**11**:15.
- Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 2005;**21**:3448–9.
- Töröcsik D, Szeles L, Paragh G, Jr, et al. Factor XIII-A is involved in the regulation of gene expression in alternatively activated human macrophages. *Thromb Haemost* 2010;**104**:709–17.
- May T, Virtanen C, Sharma M, et al. Low malignant potential tumors with micropapillary features are molecularly similar to low-grade serous carcinoma of the ovary. *Gynecol Oncol* 2010;**117**:9–17.

21. Bindea G, Mlecnik B, Hackl H, *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 2009;**25**:1091–3.
22. García O, Saveanu C, Cline M, *et al.* GOLORize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics* 2007;**23**:394–6.
23. Diez D, Wheelock AM, Goto S, *et al.* The use of network analyses for elucidating mechanisms in cardiovascular disease. *Mol Biosyst* 2010;**6**:289–304.
24. Zhu W, Yang L, Du Z. Layered functional network analysis of gene expression in human heart failure. *PLoS One* 2009;**4**:e6288.
25. Ragnedda G, Disanto G, Giovannoni G, *et al.* Protein-protein interaction analysis highlights additional loci of interest for multiple sclerosis. *PLoS One* 2012;**7**:e46730.
26. Pennings JL, Van Dycke KC, van Oostrom CT, *et al.* Biomarker discovery using a comparative omics approach in a mouse model developing heterogeneous mammary cancer subtypes. *Proteomics* 2012;**12**:2149–57.
27. Hui L, Slonim DK, Wick HC, *et al.* Novel neurodevelopmental information revealed in amniotic fluid supernatant transcripts from fetuses with trisomies 18 and 21. *Hum Genet* 2012;**131**:1751–9.
28. Pritchard S, Wick HC, Slonim DK, *et al.* Comprehensive analysis of genes expressed by rare microchimeric fetal cells in the maternal mouse lung. *Biol Reprod* 2012;**87**:42.
29. Eng L, Ibrahim-zada I, Jarjanazi H, *et al.* Bioinformatic analyses identifies novel protein-coding pharmacogenomic markers associated with paclitaxel sensitivity in NCI60 cancer cell lines. *BMC Med Genomics* 2011;**4**:18.
30. Smoot M, Ono K, Ruscheinski J, *et al.* Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011;**27**:431–2.
31. Vailaya A, Bluvus P, Kincaid R, *et al.* An architecture for biological information extraction and representation. *Bioinformatics* 2005;**21**:430–8.
32. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;**4**:2.
33. Saito R, Smoot ME, Ono K, *et al.* A travel guide to Cytoscape plugins. *Nat Methods* 2012;**9**:1069–76.
34. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
35. Franceschini A, Szklarczyk D, Frankild S, *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013;**41**:D808–15.
36. Ringwald M, Wu C, Su AI. BioGPS and GXD: mouse gene expression data—the benefits and challenges of data integration. *Mamm Genome* 2012;**23**:550–8.
37. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**:44–57.
38. Huang DW, Sherman BT, Tan Q, *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 2007;**8**:R183.
39. Gould J, Getz G, Monti S, *et al.* Comparative gene marker selection suite. *Bioinformatics* 2006;**22**:1924–5.
40. Gutierrez A, Sanda T, Ma W, *et al.* Inactivation of LEF1 in T-cell acute lymphoblastic leukemia. *Blood* 2010;**115**:2845–51.
41. Hur J, Sullivan KA, Pande M, *et al.* The identification of gene expression profiles associated with progression of human diabetic neuropathy. *Brain* 2011;**134**:3222–35.
42. De Jager PL, Rossin E, Pyne S, *et al.* Cytometric profiling in multiple sclerosis uncovers patient population structure and a reduction of CD8low cells. *Brain* 2008;**131**:1701–11.
43. Bernt KM, Zhu N, Sinha AU, *et al.* MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L. *Cancer Cell* 2011;**20**:66–78.
44. Beroukhi R, Getz G, Nghiemphu L, *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci USA* 2007;**104**:20007–12.
45. Shain AH, Giacomini CP, Matsukuma K, *et al.* Convergent structural alterations define SWItch/Sucrose NonFermentable (SWI/SNF) chromatin remodeler as a central tumor suppressive complex in pancreatic cancer. *Proc Natl Acad Sci USA* 2012;**109**:E252–9.
46. Monticelli LA, Sonnenberg GF, Abt MC, *et al.* Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. *Nat Immunol* 2011;**12**:1045–54.
47. Weir BA, Woo MS, Getz G, *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* 2007;**450**:893–8.
48. Aiden AP, Rivera MN, Rheinbay E, *et al.* Wilms tumor chromatin profiles highlight stem cell properties and a renal developmental network. *Cell Stem Cell* 2010;**6**:591–602.
49. Hoshida Y. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PLoS One* 2010;**5**:e15543.
50. Tan IB, Ivanova T, Lim KH, *et al.* Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. *Gastroenterology* 2011;**141**:476–85. 485.e11–11.
51. Bellucci R, Nguyen H-N, Martin A, *et al.* Tyrosine kinase pathways modulate tumor susceptibility to natural killer cells. *J Clin Invest* 2012;**122**:2369–83.
52. Weigert O, Lane AA, Bird L, *et al.* Genetic resistance to JAK2 enzymatic inhibitors is overcome by HSP90 inhibition. *J Exp Med* 2012;**209**:259–73.
53. Hoshida Y, Villanueva A, Kobayashi M, *et al.* Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N Engl J Med* 2008;**359**:1995–2004.
54. De Keersmaecker K, Real PJ, Gatta GD, *et al.* The TLX1 oncogene drives aneuploidy in T cell transformation. *Nat Med* 2010;**16**:1321–7.
55. Feuerer M, Herrero L, Cipolletta D, *et al.* Lean, but not obese, fat is enriched for a unique population of regulatory T cells that affect metabolic parameters. *Nat Med* 2009;**15**:930–9.
56. Turcan S, Rohle D, Goenka A, *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* 2012;**483**:479–83.
57. Ebert BL, Galili N, Tamayo P, *et al.* An erythroid differentiation signature predicts response to lenalidomide in myelodysplastic syndrome. *PLoS Med* 2008;**5**:e35.
58. Shaw SY, Blodgett DM, Ma MS, *et al.* Disease allele-dependent small-molecule sensitivities in blood cells from monogenic diabetes. *Proc Natl Acad Sci USA* 2011;**108**:492–7.
59. Stransky N, Egloff AM, Tward AD, *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* 2011;**333**:1157–60.

60. Soemedi R, Wilson IJ, Bentham J, *et al.* Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am J Hum Genet.* 2012;**91**:489–501.
61. Corbett MA, Schwake M, Bahló M, *et al.* A mutation in the Golgi Qb-SNARE gene GOSR2 causes progressive myoclonus epilepsy with early ataxia. *Am J Hum Genet* 2011;**88**: 657–63.
62. Garcia CB, Shaffer CM, Eid JE. Genome-wide recruitment to Polycomb-modified chromatin and activity regulation of the synovial sarcoma oncogene SYT-SSX2. *BMC Genomics* 2012;**13**:189.
63. Ross-Innes CS, Brown GD, Carroll JS. A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC Genomics* 2011;**12**:593.
64. Cusanovich DA, Billstrand C, Zhou X, *et al.* The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum Mol Genet* 2012;**21**:2111–23.
65. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;**26**:589–95.
66. Piazza R, Valletta S, Winkelman N, *et al.* Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat Genet* 2013;**45**:18–24.
67. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.
68. Lank SM, Golbach BA, Creager HM, *et al.* Ultra-high resolution HLA genotyping and allele discovery by highly multiplexed cDNA amplicon pyrosequencing. *BMC Genomics* 2012;**13**:378.
69. Culhane AC, Schwarzl T, Sultana R, *et al.* GeneSigDB—a curated database of gene expression signatures. *Nucleic Acids Res* 2010;**38**:D716–25.
70. Liberzon A, Subramanian A, Pinchback R, *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**: 1739–40.
71. Pico AR, Kelder T, van Iersel MP, *et al.* WikiPathways: pathway editing for the people. *PLoS Biol* 2008;**6**:e184.
72. Ho Sui SJ, Begley K, Reilly D, *et al.* The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acids Res* 2012;**40**:D984–91.
73. Lanzén A, Oinn T. The Taverna Interaction Service: enabling manual interaction in workflows. *Bioinformatics* 2008;**24**:1118–20.
74. Goble CA, Bhagat J, Alekseyevs S, *et al.* myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res* 2010;**38**:W677–82.
75. Wolstencroft K, Haines R, Fellows D, *et al.* The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* 2013;**41**:W557–61.
76. Fisher P, Hedeler C, Wolstencroft K, *et al.* A systematic strategy for large-scale analysis of genotype-phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Res* 2007;**35**:5625–33.
77. Kasprzyk A. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* 2011;**2011**:bar049.
78. Tan W, Madduri R, Nenadic A, *et al.* CaGrid Workflow Toolkit: a Taverna based workflow tool for cancer grid. *BMC Bioinformatics* 2010;**11**:542.
79. Bian X, Gurses L, Miller S, *et al.* Data submission and curation for caArray, a standard based microarray data repository system. *Nat Precedings* 2009. Available from <<http://dx.doi.org/10.1038/npre.2009.3138.1>>
80. De Bruin JS, Deelder AM, Palmblad M. Scientific workflow management in proteomics. *Mol Cell Proteomics* 2012;**11**:M111.010595.
81. Reich M, Liefeld T, Gould J, *et al.* GenePattern 2.0. *Nat Genet* 2006;**38**:500–1.
82. Mesirov JP. Computer science. Accessible reproducible research. *Science* 2010;**327**:415–16.
83. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;**11**:R86.
84. Karasavvas K, Wolstencroft K, Mina E, *et al.* Opening new gateways to workflows for life scientists. *Stud Health Technol Inform* 2012;**175**:131–41.
85. Ludäscher B, Altintas I, Berkley C, *et al.* Scientific workflow management and the Kepler system. *Concurr Comput* 2006;**18**:1039–65.
86. Stropp T, McPhillips T, Ludäscher B, *et al.* Workflows for microarray data processing in the Kepler environment. *BMC Bioinformatics* 2012;**13**:102.
87. Altintas I. Distributed workflow-driven analysis of large-scale biological data using biokepler. In: *Proceedings of the 2nd International Workshop on Petascale Data Analytics: Challenges and Opportunities*. Seattle, WA, 2011, pp. 41–2.
88. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
89. Jensen LJ, Bork P. Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS Biol* 2010;**8**:e1000374.
90. Kohl M. Standards, databases, and modeling tools in systems biology. *Methods Mol Biol* 2011;**696**:413–27.
91. Smith B, Ashburner M, Rosse C, *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**: 1251–5.
92. Rosse C, Mejino JLV Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform* 2003;**36**:478–500.
93. Smith CL, Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med* 2009;**1**:390–9.
94. Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet* 2010;**77**:525–34.
95. Antezana E, Venkatesan A, Mungall C, *et al.* ONTO-Toolkit: enabling bio-ontology engineering via Galaxy. *BMC Bioinformatics* 2010;**11**:S8.
96. Noy NF, Shah NH, Whetzel PL, *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.
97. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. *Summit on Translat Bioinforma* 2009;**2009**: 56–60.
98. Gimenez JAM, Aranguren ME, Tomas J, *et al.* NCBO-Galaxy: bridging the BioPortal web services and the Galaxy platform. *ICBO, volume 897 of CEUR Workshop Proceedings* 2012.