

Review

Keywan Hassani-Pak¹ / Christopher Rawlings¹

Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes

¹ Rothamsted Research, Computational and Analytical Sciences Department, Harpenden, AL5 2JQ, UK, E-mail: keywan.hassani-pak@rothamsted.ac.uk

Abstract:

Genetics and “omics” studies designed to uncover genotype to phenotype relationships often identify large numbers of potential candidate genes, among which the causal genes are hidden. Scientists generally lack the time and technical expertise to review all relevant information available from the literature, from key model species and from a potentially wide range of related biological databases in a variety of data formats with variable quality and coverage. Computational tools are needed for the integration and evaluation of heterogeneous information in order to prioritise candidate genes and components of interaction networks that, if perturbed through potential interventions, have a positive impact on the biological outcome in the whole organism without producing negative side effects. Here we review several bioinformatics tools and databases that play an important role in biological knowledge discovery and candidate gene prioritization. We conclude with several key challenges that need to be addressed in order to facilitate biological knowledge discovery in the future.

Keywords: Data integration, knowledge graph, knowledge discovery, genotype-to-phenotype, candidate gene prioritization

DOI: 10.1515/jib-2016-0002


Received: January 12, 2017; **Revised:** February 10, 2017; **Accepted:** February 16, 2017

1 Introduction

The discovery of causal genes and alleles that determine a particular biological phenotype in crops, animals or humans is referred to as the genotype to phenotype prediction-challenge [1], [2]. The use of genetics (e.g. genome wide association studies and quantitative trait mapping), and “omics” (e.g. RNA-sequencing) approaches can often identify large numbers of potential candidate genes, among which the causal genes are hidden. Experimental validation of candidate genes, e.g. from lab to greenhouse to field, is a slow process that can last several years. Following a wrong lead can waste significant effort, time and money. Scientists therefore need to prioritise genes that, when perturbed through potential interventions such as knock-down or gene editing approaches, might have a positive impact on the biological outcome for the whole organism without producing negative side effects. Because it is hard to undertake objective evaluation of large candidate gene sets, this choice is likely to be made subjectively, based on hunches or (potentially selective) prior experience and generally with limited scientific justification. The productivity and likelihood of success of genotype-phenotype mapping would be greatly improved if all candidate genes were to be thoroughly evaluated and only those with the highest level of confidence were considered for experimental validation.

A systematic prioritization of candidate genes needs to be based on the generation of hypotheses that explain how genotype might be linked to phenotype. This requires the consideration of multiple types of information that is very heterogeneous in nature such as: known records of gene-phenotype links, gene-disease associations, gene expression and co-expression, allelic information and effects of genetic variation, links to scientific literature, homology information from model species, protein-protein interactions, gene regulation, protein pathway memberships, gene-ontology annotations, protein-domain information and other domain specific information. The integration of such information into a knowledge network/graph combined with knowledge mining has considerable potential to improve the interpretation of complex genetic and omics experiments and help with the discovery of biological networks controlling phenotypes and diseases (Figure 1). However, it is not trivial

Keywan Hassani-Pak is the corresponding author.

 ©2017, Keywan Hassani-Pak, published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

to integrate and interrogate this information and obtain clear, objective answers that can be applied in practice. One of the key challenges is that the biological information is spread across many different databases and data formats [3].

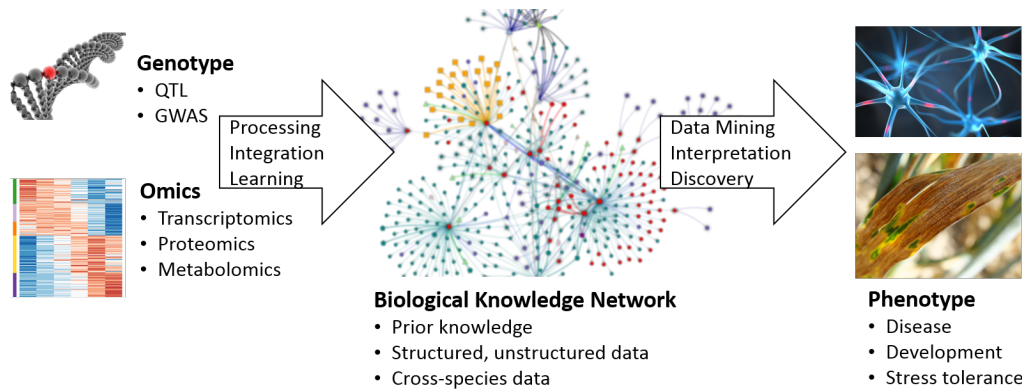


Figure 1: Using biological knowledge discovery to interpret genotype and omics experiments and establish links to phenotypes and diseases.

It has been recognised that computational tools are needed that can systematically mine the wealth of biomedical information to boost candidate gene discovery [4]. The identification of patterns in such large structured, semi-structured and unstructured data is often referred to as “data mining”, or more broadly “knowledge discovery in databases”, or KDD [5]. In the last 25 years, many novel KDD approaches have been developed including methods to pre-process, integrate, analyse and interpret complex biomedical data with the aim of identifying testable hypotheses [6]. It has also been recognised that it is important to include the end user into the “interactive” knowledge discovery process with the goal of supporting human intelligence with machine intelligence [7]. Combined KDD-HCI (Human-Computer Interaction) approaches can significantly increase the capacity and efficiency of candidate gene discovery while reducing costs and time.

Here, we first provide a comprehensive review of biological information types and databases that play an important role in candidate gene prioritization. The second part of this article provides a short overview of bioinformatics tools for integrating information from selected databases, and an overview of interactive knowledge discovery approaches that can help to bridge the genotype to phenotype gap.

1.1 Information Types and Databases for Gene Discovery

Some key information types and databases for *in silico* genotype to phenotype discovery are described below, together with their value and importance for the discovery and prioritization of candidate genes.

1.1.1 Genotype and Genetics Data

Quantitative genetics uses natural populations or families (mapping populations) and applies statistical techniques to identify those regions in the genome that can explain the phenotypic variability in the population. These regions are referred to as quantitative trait loci (QTL) [8]. Genetic linkage studies show that typical QTLs in both plants and animals encompass quite sizeable parts of the genome – often several hundred genes. Genome-wide association studies (GWAS) associate phenotype with genotype at a genome-wide level using “unrelated” individuals [9]. The limitation of family-based mapping populations can be overcome by the use of unrelated genotypes that have accumulated much higher number of recombination events since their last common progenitor [10]. Although genetic intervals identified from GWAS encompass much smaller regions of the genome compared to QTLs from mapping populations, they are likely to identify many significant candidate genes.

Genetic variants that are linked to phenotypes via QTL mapping, GWAS or other genetic studies provide a key data resource for gene-phenotype discovery. Access to public databases that contain such information is invaluable, however, this information is often hidden in the literature in an unstructured manner; which makes it very hard to retrieve and integrate. An ideal resource for standardised QTL and GWAS data of livestock species is the AnimalQTLdb [11]. AnimalQTLdb contains 121,265 QTL for 1804 traits based on 1768 publications in seven species (Release 31, December 2016). In crop species, however, such structured genetics resources are

only slowly beginning to emerge. For example, GnpIS [12] and the Triticeae Toolbox [13] provide access to genetic information (e.g. markers, phenotype and pedigree data) for species of agronomic interest.

Genetic variants that do not have reported links to phenotypes might initially be considered less important to gene discovery. However, knowledge about published genetic variants and their effect on protein level can inform candidate gene prioritization, since variants of genes with major effects can be given higher weight than genes with no reported variants or minor variant effects. The European Variation Archive (EVA) provides access to all types of genetic variants, ranging from single nucleotide polymorphisms to large structural variants from any eukaryotic organism. EVA uses the variant effect predictor [14] of Ensembl to annotate variant consequences. The variant consequences are described using sequence ontology terms [15].

Reverse genetics approaches are based on disrupting genes of known sequence and studying the effect of the disruption on the phenotype [16]. Reverse genetics resources consist of plant material (i.e. seeds) with a certain knockout gene that can be grown and used for functional characterisation of the disrupted gene. For several plant species, e.g. Arabidopsis, rice and wheat, reverse genetics resources have been generated that allow scientists to study the function of many genes more effectively [17], [18], [19]. The phenotypic consequences of such genetic disruptions are recorded in several databases. The public database UniProt contains a subsection “disruption phenotype” that describes the *in vivo* effects caused by knockout or knockdown of a gene [20]. TAIR provides phenotypic information for unique genotypes with mutations in individual genes [21]. NCBI has the GeneRIF database [22] that contains concise phrases describing a gene function that is sometimes used to add phenotypic descriptions. The data from such resources can be used to rank candidate genes higher for which gene knockouts with associated phenotype data exist.

1.1.2 Phenotype and Environment Data

Genotypic data is stable for a given plant or animal. In contrast, phenotypic characterisation requires environmental data because of the important role that environment has on the expression of a trait/phenotype. The development of standards for capturing phenotypic data has been challenging since “phenotype” is a broad concept that covers all observable traits stored as descriptive data, numeric observations including time series, molecular data and image data. Phenotypic information can be obtained from dedicated phenotyping platforms, from farmers’ fields, or from ecological diagnostics in natural environments. Phenotyping platforms measure a wide range of structural and functional plant traits at the same time as collecting accurate metadata on the environment and experimental setup [23]. Traits are measured at different spatial scales, from the field level (e.g. crop yield) to the cell (e.g. cell wall polysaccharide composition) and over widely varying temporal scales, from seconds (e.g. photosynthetic response) to months (e.g. whole season biomass). An important recent development is the publication of a minimal metadata standard for plant phenotyping experiments (MIAPPE).

Phenotype data itself (without being associated to genotype) is important in upstream processes involved in trait discovery and QTL mapping but has limited use to gene discovery *per se*. Once phenotype data can be related to genotype, gene or mutants then it becomes a relationship of high importance. The majority of phenotypic information is available in an unstructured form in the scientific literature and is therefore difficult to integrate with other knowledge resources such as ontologies. Text-mining techniques are required to identify and extract such information.

Due to the complexity of phenotype descriptions and the essential role of environmental information, a variety of ontologies have been developed to formalise their representation. Many of these are species-specific. For example, available ontologies for plants and crops include the Plant Ontology (www.plantontology.org), the Crop Ontology (www.croponontology.org), the Plant Trait Ontology (www.planteome.org) and the Environment Ontology (www.environmentontology.org). Although several new phenotype ontologies are emerging, not many plant genomes and experiments are yet annotated with these ontology terms. Even in model species such as Arabidopsis, most phenotypic descriptions are still in free text. The Drosophila phenotype ontology [24] is a good example of a phenotype ontology that is systematically used to annotate genes and alleles enabling more powerful search queries.

1.1.3 Gene Expression Data

Gene expression data can be used as evidence to confirm the expression of candidate genes in tissues, organs, during different developmental stages, under treatments of interest or in particular genotypes. For example, for human studies, the Genotype-Tissue-Expression resource can reveal correlations between genotype and tissue-specific gene expression levels and can help identify regions of the genome that influence whether and by how much a gene is expressed [25]. A similar baseline expression resource does not exist for most plant and

animal species. For example, identifying causal genes for a grain specific QTL would require any potential candidate gene to be expressed at some stage during grain development and potentially only expressed in certain individuals of a mapping population and not in others. Several other general gene expression databases exist such as the Gene Expression Atlas [26], the Gene Expression Omnibus [27] or the eFP Browser [28]. Reference-species resources such as The Arabidopsis Information Resource (TAIR) have annotated Arabidopsis genes with Plant Ontology [29] terms that describe in which tissues and during which developmental stages a gene is expressed. Other databases such as ATTED-II [30] analyse large amounts of expression datasets to compute clusters of coexpressed genes. Such co-expression data provides weak, speculative evidence that these genes are co-regulated and therefore could share a similar biological function or act together to control a phenotype.

1.1.4 Interaction Data

Protein-protein interaction (PPI) data provides very useful knowledge for candidate gene discovery. In contrast to co-expression data, PPI data provides evidence about the physical interaction of proteins in the cell. A large number of methods have been developed over the years to study protein-protein interactions, e.g. affinity-tagged proteins, the two-hybrid system and some quantitative proteomic techniques [31]. Measurable physical interaction implies that the proteins are involved in the same biological process and could contribute to higher-level traits although they might have different functions. Public PPI databases can be searched to identify previously reported interactions for a given bait protein. BioGRID [32] and IntAct [33] databases are populated by data either curated from the literature or from direct data depositions. Data access and download are provided for many species and in different data formats such as PSIMI-XML, PSIMI-TAB, BioPAX or RDF. Other types of interaction data such as protein-drug interactions [34] or pathogen-host interactions [35] can be considered for the discovery of genes relevant to human or plant disease.

1.1.5 Functional Annotation Data

Functional annotation of genes and gene products provides a key resource to elucidate the biological processes and pathways controlling complex traits. Gene Ontology annotations capture the knowledge that we have about the molecular function of genes in a systematic and cross-species comparable manner. GO provides a controlled vocabulary to describe biological processes, molecular functions and cellular components. GO annotations require the provision of evidence codes that describe the experimental or computational methods used to establish the gene function. The Evidence and Conclusion Ontology (ECO) is used to describe the evidence in a formalised manner and help to distinguish high quality annotations (e.g. inferred through mutant phenotypes) from low quality annotations (e.g. inferred through electronic annotations). As the best studied plant species *Arabidopsis thaliana* has about 50,000 GO annotations of experimental evidence (25 % of total annotations). The majority of annotations in non-model species are electronically inferred through sequence based comparisons with model species. The common data type for functional gene annotations is the Gene Association Format (GAF). Many functional or structural bioinformatics databases provide mappings to GO terms e.g. EC2GO, Pfam2GO and InterPro2GO. Biological pathways provide a more fine-grained knowledge about the enzymes, chemical reactions and small molecules that form the elements of biosynthetic pathways. Popular pathways databases such as KEGG [36], Reactome [37] and BioCyc [38] provide curated pathway information for model species and computationally inferred pathways for non-model species.

1.1.6 Homology Data

The function of the vast majority of genes in non-model species remains uncharacterised. Any effort to prioritize candidate genes without any evidence about their function is difficult or even impossible. Genes that have been well characterised in other species provide a reliable source of putative evidence assuming this knowledge can be transferred from one species to another. The principal idea supporting cross-species annotation transfer is that the function of proteins is, to some extent, conserved through evolution. Thus, two orthologs in two closely related species are likely to share the same function. But the level of conservation of protein function across species largely depends on the evolution of these species, including the evolution of their proteins, of their biochemical pathways and of their higher level biological traits. Orthologous relationships can be established when comparing the genomes of two or more species. Identification of orthologous gene sets typically involves phylogenetic tree analysis, heuristic algorithms based on sequence conservation, synteny analysis, or some combination of these approaches [39], [40]. Some of the prominent databases of orthologous genes include

Ensembl [41], OrthoDB [42] OMA [43] and Phytozome [44]. The common data standard for orthology data provision is OrthoXML [45].

In addition to using orthology data for cross-species annotation transfer, a more direct approach exploiting sequence database search with the BLAST [46] or Smith-Waterman [47] algorithms can be used to infer putative gene function. This is a common shortcut taken by many scientists and bioinformatics tools such as Blast2GO [48]. Such data can be used for exploratory analysis but is prone to a high false positive rate. In the context of prioritizing genes it should be given a much lower importance than more accurate orthology inference methods.

2 Biological Knowledge Discovery for Gene Prioritization

Having identified various datasets and information types relevant to candidate gene discovery, the next step in the knowledge discovery process is the transformation of data into a suitable data structure. Biological data is typically highly connected, e.g. through common references to named biological entities, and semi-structured, e.g. because some data can be found in databases and other in free text. Furthermore, these data types are not static because new types of data are constantly emerging from advances in high-throughput experimental platforms. These characteristics of Life Science data make networks, consisting of nodes and links between them, a flexible data model that can capture much of the complexity and interconnectedness in the data [49]. In addition, networks are often considered as the layer that connects genotype to phenotype [50].

In contrast to homogeneous networks, where all nodes have the same type (e.g. protein-protein interaction networks), heterogeneous information networks, also referred to as knowledge graphs, are networks where nodes and links can have various types [51]. Biological knowledge networks are composed of nodes which represent biological entities such as genes, transcripts, proteins and compounds, as well as other entities such as protein domains, ontology terms, pathways, literature and phenotypes. The links in the network correspond to relations between entities and are described using terms which reflect the semantics of the biological or functional relationship such as *encodes*, *interacts*, *involved_in*, *expressed_in*, *published_in* etc.

A number of biological data warehousing (DW) systems have been constructed to facilitate data integration and information retrieval from diverse biological data sources [52]. Common requirements of such biological DW systems include: (i) to provide solutions for reproducible data acquisition and integration, (ii) to be flexibly extended to new species and new data types and (iii) to support complex queries using a powerful (semantic) search engine. InterMine [53], BioMart [54] and Ondex [55] are examples of such DW systems that provide tools (parsers) for integrating data from many common biological data sources and formats, and frameworks for adding custom user data in tabular format. Most biological DW use a relational database to store information and only a few systems such as Ondex use networks (graphs) as their internal data structure. Our group has developed genome-scale knowledge networks (GSKNs) for key plant and crop species using the Ondex platform [56]. For example, the wheat GSKN contains approximately 700,000 nodes of 20 different types and 3 Million links of 30 different types between them.

In order to expand knowledge networks with phenotypic information from unstructured free text such as scientific publications, automated approaches are needed that link trait descriptions to the cited genes and their corresponding nodes in the network. Such approaches will create novel, structured relationships between biological concepts and therefore improve the ability to reason over the data and make novel connections between previously unrelated biological concepts [57].

In recent years, several stand-alone text mining systems have been developed [58], mostly to support database curators finding evidence text for particular information of interest, such as protein-protein interactions or functional gene annotations [59], [60]. In addition to such user-centred systems, Java based libraries and frameworks have recently emerged providing APIs that enable language processing functionality to be embedded in diverse applications [61], [62]. Such frameworks allow text mining workflows to be created that consist of elementary components, for example text segmentation, sentence boundary detection, entity detection and relation extraction. For example, the Ondex data integration platform has been extended with easy to use text mining workflows that operate on the knowledge graph and include steps to filter associations with low scores [63].

Once the data has been transformed and integrated, the next step in the knowledge discovery process requires tools for knowledge mining, exploration and visualisation that help scientists to prioritize candidate genes and biological processes. A number of web-based resources for prioritizing candidate genes by exploiting multiple information types have been developed [4], [64]. For example, Endeavour [65] integrates 75 datasets from 6 model species including human and mouse into a local database, and uses basic machine learning techniques with a-priori known candidate genes to model the biological process under study and then to prioritize the candidate genes. Another tool named BioGraph is based on a graph data warehouse approach and

uses unsupervised data mining for the exploration and discovery of biomedical information [66]. In total, BioGraph contains 532,889 distinct relations among 71,042 biomedical concepts, supported by 61,570 literature references. The biological knowledge, which includes many indirect relationships, is used for gene prioritization and hypothesis generation. The main limitations of many gene prioritization tools, including Endeavour and BioGraph, are that they are restricted to the analysis of key model species and the data integration process is not easily reproducible and adaptable to other species. PosMed-Plus [68] was one of the first tools to prioritize candidate genes for two plant species (*Arabidopsis thaliana* and rice) using a knowledge-based approach and including literature co-occurrence and cross-species information. KnetMiner (<http://knetminer.rothamsted.ac.uk/>) is one of the first tools to provide a generic and easily configurable approach that works for model and non-model species. KnetMiner searches and evaluates millions of relations and concepts within biological knowledge networks (created using the Ondex data integration platform) in real-time to determine if direct or indirect links between genes and phenotypes, pathways, annotations etc. can be established using biologically plausible graph queries. KnetMiner accepts as user inputs: search terms in combination with a gene list and/or genomic regions. It produces tables of ranked candidate genes or evidence summaries, and allows users to explore the knowledge networks using interactive web-based tools. KnetMiner is currently available for several plant, crop and animal species such as *Arabidopsis*, wheat, maize, barley, camelina, potato, tomato, poplar, pig, cow and chicken. A benefit of the KnetMiner compared to other existing gene discovery tools is its generic and interactive approach.

3 Conclusion

Mining information across different biological databases has the potential to discover new knowledge that was hidden before. For example, linking a GWAS dataset that contains statistical associations between SNPs and phenotypes, with genomic information about genes and proteins, and protein-protein interaction data, can reveal new insights into the regulation of complex traits. In this article we have reviewed several biological databases and information types that can be used to provide evidence for the discovery of genotype to phenotype relationships. Creating a complete knowledge base of gene functions, interaction networks and trait biology is technically challenging because the relevant data are dispersed in myriad databases in a variety of data formats with variable quality and coverage. Innovative approaches are often needed to infer implicit relationships between concepts in a knowledge network. For example, linking SNP to gene can be based on genomic coordinate information, linking gene to phenotype can be based on sentence-level co-occurrence of names. Building knowledge networks in non-model species is even more challenging as the majority of genes are not well studied and have unknown names or function.

In this article, we also reviewed a small set of tools for biological knowledge discovery and candidate gene mining. Although many candidate gene mining tools already exist, there is still an urgent need for tools that improve the efficiency and interactivity of gene discovery using new approaches from the KDD-HCI field. In the following we identify some of the challenges we consider to be key to improving.

3.1 Key Challenges for Data Integration

Ontologies play an important role into data integration and allow us to unify different terminologies. It is important that data providers increase their use of ontologies and metadata standards as much as possible to facilitate data integration. In recent years, linked data principles and Semantic Web standards (RDF) have further contributed to the integration of heterogeneous data sources. Making more data available in such linked form, will significantly simplify data integration processes and improve capturing most aspects of data provenance. The Monarch Initiative [67] is an outstanding example of how RDF and semantic web technologies can be harnessed to build analytical tools that connect genotype to phenotype across species. Furthermore, recent developments in this field have been using innovative approaches to address the problem of interoperability between different ontologies and data models [69]. Finally, more synergistic approaches will be needed that can effectively integrate information from structured databases with facts extracted from semi-structured and unstructured data [70], [71].

3.2 Key Challenges for Inference over Integrated Knowledge Networks

Once the heterogeneous information has been transformed into a standard data structure such as a knowledge graph or network, tools are need for interrogating the network and for analysis and inference steps, for example

to prioritise genes based on an evaluation of quality of the supporting evidence. One of the challenges that need to be addressed by biological graph mining approaches is to distinguish between high and low confidence links, for example, links that are based on poor alignments, weak associations or insufficient evidence need to be treated differently to high-quality curated links. In the future, we hope to see applications similar to the Google Knowledge Graph Search to be developed for the Life Sciences that utilise the strength of biological knowledge graphs.

3.3 Key Challenges for Interactive Knowledge Discovery

For users, the visual representation of complex biological information and navigating it to find new knowledge or testable hypotheses is relevant. Networks provide the means for interactive knowledge discovery. While networks are intuitive for biologists, there remain challenges in terms of usability of the current generation of network visualisation tools. Key challenges are the representation of many different information types, uncertainty of relationships and linked quantitative data such as time series or dose response. It is important that a new generation of interactive knowledge discovery tools are developed that allow human intelligence to play a major role in candidate gene discovery and decision making.

Funding

Our research was and is funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) award BB/F006039/1 and BB/N022874/1. We have also received additional support from Rothamsted Research 20:20 Wheat[®] Institute Strategic Program (BBS/E/C/00005202 and BBS/E/C/00005203).

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

- [1] Burga A, Lehner B. Predicting phenotypic variation from genotypes, phenotypes and a combination of the two. *Curr Opin Biotechnol.* 2013;24:803–9.
- [2] Willet CE, Wade CM. From the phenotype to the genotype via bioinformatics. *Methods Mol Biol.* 2014;1168:1–16.
- [3] Rigden DJ, Fernández-Suárez XM, Galperin MY. The 2016 database issue of nucleic acids research and an updated molecular biology database collection. *Nucleic Acids Res.* 2016;44:D1–6.
- [4] Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet.* 2012;13:523–36.
- [5] Holmes JH. Knowledge discovery in biomedical data: theory and methods. *Methods in Biomedical Informatics.* 2014;179–240.
- [6] Sacchi L, Holmes JH. Progress in biomedical knowledge discovery: a 25-year retrospective. *Yearb Med Inform.* 2016;S117–29.
- [7] Holzinger A, Jurisica I. Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions. *Lect Notes Comput Sci.* 2014;1–18.
- [8] Kearsey M. The principles of QTL analysis (a minimal mathematics approach). *J Exp Bot.* 1998;49:1619–23.
- [9] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet.* 2005;6:95–108.
- [10] Sonah H, O'Donoghue L, Cober E, Rajcan I, Belzile F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol J.* 2015;13:211–21.
- [11] Hu Z-L, Park CA, Reecy JM. Developmental progress and current status of the animal QTLdb. *Nucleic Acids Res.* 2016;44:D827–33.
- [12] Steinbach D, Alaux M, Amselem J, Choisne N, Durand S, Flores R, et al. 2013. GnpIS: an information system to integrate genetic and genomic data from plants and fungi. *Database (Oxford).* 2013;2013:bat058.
- [13] Blake VC, Birkett C, Matthews DE, Hane DL, Bradbury P, Jannink J-L. The triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding. *Plant Genome.* 2016;9:1–10.
- [14] Yourshaw M, Paige Taylor S, Rao AR, Martín MC, Nelson SF. Rich annotation of DNA sequencing variants by leveraging the ensembl variant effect predictor with plugins. *Brief Bioinform.* 2015;16:255–64.
- [15] Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.* 2005;6:R44.
- [16] Gilchrist E, Haughn G. Reverse genetics techniques: engineering loss and gain of gene function in plants. *Brief Funct Genomics.* 2010;9:103–10.
- [17] Kleinboelting N, Huet G, Kloetgen A, Viehoveer P, Weisshaar B. GABI-Kat SimpleSearch: new features of the Arabidopsis thaliana T-DNA mutant database. *Nucleic Acids Res.* 2012;40:D1211–15.

- [18] Chen L, Huang L, Min D, Phillips A, Wang S, Madgwick PJ, et al. Development and characterization of a new TILLING population of common bread wheat (*Triticum aestivum* L.). *PLoS One*. 2012;7:e41570.
- [19] An G, Gynheung A, Dong-Hoon J, Ki-Hong J, Sichul L. Reverse genetic approaches for functional genomics of rice. *Plant Mol Biol*. 2005;59:111–23.
- [20] “Disruption Phenotype”. 2015. Available from: http://www.uniprot.org/help/disruption_phenotype. Accessed September 5.
- [21] “Website”. 2015. Available from: ftp://ftp.arabidopsis.org/home/tair/User_Requests/Locus_Germplasm_Phenotype_20130122. Accessed September 5.
- [22] “About Gene RIF – Gene – NCBI”. 2015. Available from: <http://www.ncbi.nlm.nih.gov/gene/about-generif>. Accessed September 5.
- [23] Fiorani F, Schurr U. Future scenarios for plant phenotyping. *Annu Rev Plant Biol*. 2013;64:267–91.
- [24] Osumi-Sutherland D, Marygold SJ, Millburn GH, McQuilton PA, Ponting L, Stefancsik R, et al. The drosophila phenotype ontology. *J Biomed Semantics*. 2013;4:30.
- [25] GTEx Consortium. Human genomics. the genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348:648–60.
- [26] Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, et al. Expression Atlas update – a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2014;42:D926–32.
- [27] Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207–10.
- [28] Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. An ‘Electronic Fluorescent Pictograph’ browser for exploring and analyzing large-scale biological data sets. *PLoS One*. 2007;2:e718.
- [29] Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, et al. Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res*. 2014;42:D1193–99.
- [30] Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K. ATTED-II provides coexpressed gene networks for arabidopsis. *Nucleic Acids Res*. 2009;37:D987–91.
- [31] Berggård T, Tord B, Sara L, Peter J. Methods for the detection and analysis of protein–protein interactions. *Proteomics*. 2007;7:2833–42.
- [32] Chatr-aryamontri A, Breitkreutz B-J, Oughtred R, Boucher L, Heinicke S, Chen D, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res*. 2014;43:D470–78.
- [33] Orchard S, Ammari M, Aranda B, Breuzza L, Briganti L, Broackes-Carter F, et al. The MIntAct project – IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*. 2014;42:D358–63.
- [34] Wishart DS. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34:D668–72.
- [35] Urban M, Cuzick A, Rutherford K, Irvine A, Pedro H, Pant R, et al. PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database. *Nucleic Acids Res*. 2016;45:D604–10.
- [36] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;27:29–34.
- [37] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2016;44:D481–87.
- [38] Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2013;42:D459–71.
- [39] Trachana K, Forslund K, Larsson T, Powell S, Doerks T, von Mering C. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One*. 2014;9:e111122.
- [40] Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for gene orthology inference. *Brief Bioinform*. 2011;12:379–91.
- [41] Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. *Database (Oxford)*. 2016;2016. DOI:10.1093/database/bav096.
- [42] Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res*. 2015;43:D250–56.
- [43] Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I, et al. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res*. 2015;43:D240–49.
- [44] Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2011;40:D1178–86.
- [45] Schmitt T, Messina DN, Schreiber F, Sonnhammer EL. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform*. 2011;12:485–88.
- [46] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
- [47] Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 1981;147:195–97.
- [48] Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 2008;36:3420–35.
- [49] Huber W, Carey VJ, Long L, Falcon S, Gentleman R. Graphs in molecular biology. *BMC Bioinformatics*. 2007;8:S8.
- [50] Carter H, Hofree M, Ideker T. Genotype to phenotype via network analysis. *Curr Opin Genet Dev*. 2013;23:611–21.
- [51] Sun Y, Han J. Mining heterogeneous information networks: principles and methodologies. Morgan & Claypool Publishers, 2012.
- [52] Triplet T, Butler G. A review of genomic data warehousing systems. *Brief Bioinform*. 2014;15:471–83.
- [53] Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*. 2012;28:3163–65.
- [54] Yates A, Akanni W, Ridwan Amode M, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. *Nucleic Acids Res*. 2016;44:D710–16.
- [55] Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, et al. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*. 2006;22:1383–90.
- [56] Hassani-Pak K, Castellote M, Esch M, Hindle M, Lysenko A, Taubert J, et al. Developing integrated crop knowledge networks to advance candidate gene discovery. *Appl Transl Genom*. 2016;11:18–26.

- [57] Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet.* 2012;13:829–39.
- [58] Leitner F, Florian L, Martin K, Valencia A. BioCreative meta-server and text-mining interoperability standard. *Encyclopedia of Systems Biology.* 2013;8401:106–10.
- [59] Lu Z, Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database.* 2012;bas043–bas043.
- [60] Mao Y, Van Auken K, Li D, Arighi CN, McQuilton P, Thomas Hayman G, et al. Overview of the gene ontology task at BioCreative IV. *Database.* 2014;2014. DOI:10.1093/database/bau086.
- [61] Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE’s full lifecycle open source text analytics. *PLoS Comput Biol.* 2013;9:e1002854.
- [62] “Apache UIMA – Apache UIMA”. 2015. Available from: <http://uima.apache.org/>. Accessed September 9.
- [63] Hassani-Pak K, Legaie R, Canevet C, van den Berg HA, Moore JD, Rawlings CJ. Enhancing data integration with text analysis to find proteins implicated in plant stress response. *J Integr Bioinform.* 2010;7. DOI:10.2390/biecoll-jib-2010-121.
- [64] Bornigen D, Tranchevent L-C, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P, et al. An unbiased evaluation of gene prioritization tools. *Bioinformatics.* 2012;28:3081–88.
- [65] Tranchevent L-C, Ardeshirdavani A, ElShal S, Alcaide D, Aerts J, Auboeuf D, et al. Candidate gene prioritization with endeavour. *Nucleic Acids Res.* 2016;44:W117–21.
- [66] Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, Del-Favero J. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol.* 2011;12:R57.
- [67] Mungall Christopher J., et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research.* 2016 11 29;45:D712–D722. DOI:10.1093/nar/gkw1128.
- [68] Makita Y, Kobayashi N, Mochizuki Y, Yoshida Y, Asano S, Heida N, et al. PosMed-plus: an intelligent search engine that inferentially integrates cross-species information resources for molecular breeding of plants. *Plant Cell Physiol.* 2009;50:1249–59.
- [69] Deus HF, Prud’hommeaux E, Miller M, Zhao J, Malone J, Adamusiak T, et al. Translating standards into practice – one Semantic Web API for Gene Expression. *J Biomed Inform.* 2012;45:782–94.
- [70] Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, van Ommen G, et al. The value of data. *Nature Genet.* 2011;43:281–83.
- [71] Hellmann S, Lehmann J, Auer S, Brümmer M. Integrating NLP using linked data. *Lecture Notes Computer Science* 2013:98–113.