Development 134, 3227-3238 (2007) doi:10.1242/dev.001073

Automated data integration for developmental biological research

Weiwei Zhong and Paul W. Sternberg

In an era exploding with genome-scale data, a major challenge for developmental biologists is how to extract significant clues from these publicly available data to benefit our studies of individual genes, and how to use them to improve our understanding of development at a systems level. Several studies have successfully demonstrated new approaches to classic developmental questions by computationally integrating various genome-wide data sets. Such computational approaches have shown great potential for facilitating research: instead of testing 20,000 genes, researchers might test 200 to the same effect. We discuss the nature and state of this art as it applies to developmental research.

Introduction

The advent of high-throughput technologies (see glossary, Box 1) has greatly increased the amount of genetic information that is deposited in the public domain. In addition, the results of thousands of hard-won individual observations have been compiled in a consistent format (i.e. curated) in biological databases. Combined, these two sources have created genome-wide data sets that describe a wide range of biological processes (Fig. 1, see also Box 1).

As developmental biologists, we might be more interested in hypothesis-driven studies that focus on a small set of genes, rather than on generating genome-wide data. How can we best use all of these publicly available data to benefit our own research? Data integration has proven to be an effective strategy to extract biological meaning from heterogeneous data sets in both developmental research and other fields. It can be a powerful tool to identify candidate genes worthy of further study, and by automating the process it can allow the translation of genome-wide data into small-scale science.

It is highly likely that we have applied the principle of data integration in our research all along. When seeking regulators of biological processes or targets of gene functions, a common strategy is to compile a short list of candidate genes and then experimentally test them. We are probably all familiar with this type of simple 'filtering'. In an example of this approach, 766 *C. elegans* genes were selected as germ-line enriched judging from their microarray expression profiles, then the phenotypes of these genes were analyzed by RNA interference (RNAi) to identify new genes necessary for germ-line development (Piano et al., 2002). Along the same line, we might use RNAi to test for known or predicted signaling proteins (Lehner et al., 2006), transcription factors (e.g. Parrish et al., 2006; Fernandes and Sternberg, 2007), and so forth. Other 'filters' include a protein or transcript's potential phosphorylation sites, microRNA target sites or subcellular localization. Sometimes, multiple filters are applied. If we are looking for a transcriptional regulator active in a place and time of interest, we might look up all the papers on genes expressed at that time and place, look up all papers on transcription factors and determine if there is any overlap. Of course, these searches are easier if there is a comprehensive database that contains this information. The process of combining two or more data sets to identify their intersection is the simplest form of data integration.

Although data integration does not actually create new information, it can create new knowledge for the individual; as discussed above, it can limit the number of candidates a researcher should test, thereby allowing more time for an intensive analysis of each candidate. This strategy becomes even more desirable when a developmental process or the gene of interest does not have an easily screenable phenotype.

More-advanced techniques of data integration employ sophisticated statistical models to improve the extraction of meaningful information. These techniques may decrease false negatives, or attach a value of statistical confidence to each data point. Most of these techniques were developed in studies of *Saccharomyces cerevisiae* (e.g. Marcotte et al., 1999; Jansen et al., 2003), largely owing to the wealth of genome-wide data that is available for this simple organism. However, the impact of these techniques will be limited unless their applications can be extended to other organisms or can reach the large audience of small-scale

Box 1. Glossary

Bayesian network. A way of using Bayes theorem to calculate the probability of an outcome based on a network of information. The information consists of a set of observations correlated with the outcome, which allows the assignment of a probability of the outcome given the observation.

Controlled vocabulary. Defined as a set of terms that ensures their consistent use among many people. Controlled vocabulary terms can include synonyms. For example, FlyBase developed a vocabulary of body parts that covers many aspects of *Drosophila* anatomy. WormBase has a set of life-stages.

Genome-wide. Covering a large proportion of the genome; for example, 80% of 20,000 genes.

High-throughput. Rapid data collection, usually using automation such as robotics and image processing.

Likelihood ratio. The frequency in a positive training set divided by the frequency in the negative training set. The likelihood ratio is 1 if there is no difference between the training sets. A likelihood ratio of >1 indicates a positive predictor; a likelihood ratio of <1 indicates a negative predictor.

Ontology. A defined set of concepts with defined relationships among the concepts.

TILLING. The efficient identification of mutations in an organism by mutagenesis and the resequencing of genes of interest.

Training set. A set of data that can be used to teach ('train') a computer program. Such sets could range, for example, from a set of human faces and a set of other animal faces, to a set of genes that are known to interact and a set that are known not to interact. By using a training set, computer scientists can train computer programs to discriminate which of two sets of data an unclassified object (or gene) is likely to belong to.

HHMI and Division of Biology, Caltech, 1200 E. California Blvd, Pasadena, CA 91125, USA.

*Author for correspondence (e-mail: pws@caltech.edu)

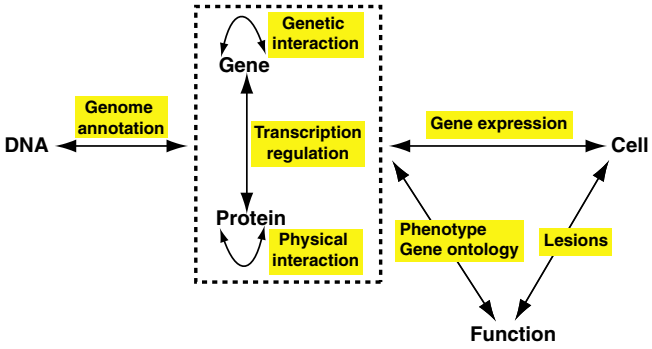

Fig. 1. Experiments that establish relationships between genes, proteins, cells and functions. Most genome-wide data sets describe biological entities or draw connections between entities. For example, DNA sequence is linked to genes by gene prediction and experimental annotation (e.g. cDNA sequencing). Genes are associated with other genes by genetic interactions. Proteins are related by physical binding, e.g. as detected in yeast two-hybrid assays. Proteins are shown to interact with DNA through chromatin-immunoprecipitation (ChIP) and yeast one-hybrid assays (e.g. Deplancke et al., 2006). Genes and protein are assigned functions based on perturbations (mutations, overexpression, RNAi). Cells are associated with genes and proteins by gene expression. Cells (or tissues) are associated with functions by mechanical (e.g. laser ablation) or genetic (e.g. mutation) lesion experiments or by generating genetic mosaics.

research. Recently, several studies have successfully applied genomics data integration strategies in metazoans, specifically to explore mechanisms of development (Gunsalus et al., 2005; Zhong and Sternberg, 2006). In this review, we highlight these examples to show how typical developmental biology laboratories can use more-advanced data integration techniques to benefit their own research. We first discuss the different types of genomic data that are publicly available, and then describe the resources and methods available to integrate these data. We also discuss the benefits and limitations of such approaches in studying developmental biology.

We focus on data integration rather than on approaches to mine individual genomic data sets because it provides the best use of the publicly available data. It is thus especially beneficial to developmental biologists, who might not be able to generate these genomic data themselves. We focus on the applications of these strategies in developmental biology investigations, rather than on the technical advances in data integration methods (for a review, see Joyce and Palsson, 2006). Therefore, rather than discuss in detail the bioinformatic techniques that are available in the data integration field, we introduce a few typical methods and key concepts that are of particular use to studies of developmental biology. This is also not a comprehensive review of the genomic data available for each organism (see Antoshechkin and Sternberg, 2007; Lee, 2005; Eppig et al., 2007; Crosby et al., 2007; Rhee et al., 2003).

What data to integrate: types of genome-wide data

In this section, we select five types of commonly used genome-wide data sets and briefly discuss the current state, strengths and weaknesses of these data. Although only experimental data are described here, it should be noted that non-experimental data can also be used in data integration. For example, the co-occurrence of gene names in the literature has been used, in addition to experimental data, in data integration to predict functional interactions (Lee et al., 2004).

Numerous databases have been developed to accommodate the growing amount of genomic data. Often, the data that we are interested in are covered by multiple databases. Which databases shall we use? Table 1 lists some popular, publicly accessible resources for the genomic data mentioned above. This is by no means a comprehensive list.

Three criteria are useful when choosing your data sources. (1) Support for batch download. Batch download allows a user to download all data into one large file, so that the user does not have to send one query for each gene. Some databases – for example, the Nematode Expression Pattern Database (NEXTDB, http://nematode.lab.nig.ac.jp/index.html) – provide excellent in situ hybridization images for *C. elegans* genes, but without batch-downloadable data it is almost impossible to perform any bioinformatics analysis on these data. (2) Controlled vocabulary. Although detailed descriptions, such as phenotypic information in the *Saccharomyces* Genome Database (SGD, http://www.yeastgenome.org), or images, such as the in situ results in the Brain Gene Expression Map (BGEM, http://www.stjudebgem.org), are extremely valuable for obtaining information on single genes, they are difficult to process in computational applications. We chose data sets that are annotated in a controlled vocabulary (see glossary, Box 1) that can be readily processed by computers. (3) Comprehensive and up-to-date data collection. Some data sets can be accessed from multiple databases. For example, the global yeast two-hybrid mapping of *Drosophila* protein interactions can be obtained from the General Repository for Interaction Datasets (BioGRID, http://www.thebiogrid.org/index.php) or the Flynet Server (http://www.jhubiomed.org/perl/flynet.pl), among others. BioGRID also contains other interaction data and we have therefore chosen to access BioGRID because it is more comprehensive. The collection of data from various databases is tedious because different databases have different gene IDs, different data formats and different access methods. Therefore, for computational usages, it is highly preferable to obtain the same amount of data from the fewest possible sources and to avoid databases that only contain a subset of data from another database.

Expression data

There are two major types of gene expression data: (1) annotation-based data obtained from experiments such as reporter gene assays and in situ hybridization and (2) data from microarray experiments.

Annotation-based data indicate the developmental stage and the tissue of gene expression, often using consistent nomenclature. High-throughput studies in this category mostly employ RNA in situ hybridization or promoter-reporter gene fusion assays. Large-scale in situ hybridization data are available for: gene expression patterns during *D. melanogaster* embryogenesis (Tomancak et al., 2002); all stages in *C. elegans* development (NEXTDB, http://nematode.lab.nig.ac.jp/index.html); *Xenopus* embryogenesis (Pollet et al., 2005); mouse embryogenesis (Visel et al., 2004; Christiansen et al., 2006); and expression patterns in the adult mouse nervous system (Gray et al., 2004; Magdaleno et al., 2006; Lein et al., 2007). Large-scale reporter gene assays require transgenic strain construction; thus, it can be difficult to reach high-throughput efficiency in certain organisms. The technique was first successfully applied on a genome scale to study protein subcellular localization in *S. cerevisiae* (Kumar et al., 2002; Huh et al., 2003). Subsequently, several high-throughput projects have emerged to study developmental profiles in *C. elegans* (McKay et al., 2004; Dupuy et al., 2004; Dolphin and Hope, 2006) and gene expression in the mouse central nervous system (Gong et al., 2003).

Table 1. Publicly available sources of genomic data

Species	Database	URL
Expression		
C. elegans	WormBase	http://www.wormbase.org
Drosophila	FlyBase	http://www.flybase.org
	BDGP	http://www.fruitfly.org
Zebrafish	ZFIN	http://zfin.org
Mouse	MGI/GXD	http://www.informatics.jax.org/menus/expression_menu.shtml
Multiple	GEO	http://www.ncbi.nlm.nih.gov/geo
Multiple	SMD	http://smd.stanford.edu
Multiple	ArrayExpress	http://www.ebi.ac.uk/arrayexpress
Phenotype		
S. cerevisiae	CYGD	http://mips.gsf.de/genre/proj/yeast
C. elegans	WormBase	http://www.wormbase.org
	RNAiDB	http://www.rnai.org
Drosophila	FlyBase	http://www.flybase.org
	FlyRNAi	http://www.flyrnai.org
	GenomeRNAi	http://www.dkfz.de/signaling2/rnai/ernai.html
Arabidopsis	TAIR	http://www.arabidopsis.org
Zebrafish	ZFIN	http://zfin.org
Mouse	MGI/MPD	http://phenome.jax.org/pub-cgi/phenome/mpdcgi?rtn=docs/home
Human	OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Interaction		
S. cerevisiae	SGD	http://www.yeastgenome.org
	CYGD	http://mips.gsf.de/genre/proj/yeast
C. elegans	WormBase	http://www.wormbase.org
Drosophila	FlyBase	http://www.flybase.org
Multiple	IntAct	http://www.ebi.ac.uk/intact/site
Multiple	DIP	http://dip.doe-mbi.ucla.edu
Multiple	BioGRID	http://www.thebiogrid.org/index.php
Multiple	BIND	http://www.bind.ca
Multiple	MINT	http://mint.bio.uniroma2.it/mint/Welcome.do
Function		
Multiple	GO	http://geneontology.org

Abbreviations: BDGP, Berkeley *Drosophila* Genome Project; BIND, Biomolecular Interaction Network Database; BioGRID, General Repository for Interaction Datasets; CYGD, Comprehensive Yeast Genome Database; DIP, Database of Interacting Proteins; GEO, Gene Expression Omnibus; GO, Gene Ontology; MGI, Mouse Genome Informatics [which provides integrated access to several projects, including the Mouse Phenome Database (MPD) and the Mouse Gene Expression Database (GXD)]; MINT, a Molecular Interaction Database; OMIM, Online Mendelian Inheritance in Man; SGD, *Saccharomyces* Genome Database; TAIR, The *Arabidopsis* Information Resource; ZFIN, Zebrafish Information Network.

Microarray studies measure mRNA levels on a genome scale and have become a powerful tool with which to study development. Soon after the invention of the technique (Schena et al., 1995), developmental gene expression profiles were studied in several model organisms, including *C. elegans* (Hill et al., 2000; Kim et al., 2001) and *D. melanogaster* (White et al., 1999; Furlong et al., 2001). Now, microarray techniques have been extensively applied to the study of development in various mutant backgrounds and under different conditions. Accordingly, strategies for analyzing microarray data have also advanced tremendously. An interesting example is a cross-species microarray data integration study conducted by Stuart et al. (Stuart et al., 2003). Stuart et al. first identified conserved genes (meta genes) that have orthologs in humans, fruitflies, *C. elegans* and yeast. By examining expression data from these species, ~22,000 pairs of meta genes were found to be co-expressed across species, revealing conserved functional modules in core biological processes, such as the cell cycle, during transcription, and signaling. Their discoveries demonstrated the power of cross-species data integration.

Thanks to the establishment of data standards in the field (Quackenbush, 2004), microarray data have the advantage of being made available in a single format (per platform). However, recent studies suggest that more standards are needed to enable the results from different experiments, laboratories and platforms, to be more comprehensively and accurately compared (Irizarry et al., 2005; Larkin et al., 2005; Members of the Toxicogenomics Research Consortium, 2005). One disadvantage of microarray experiments is that the data often contain a high background. Data from in situ hybridization gives good spatial and temporal resolution, but there is still noise present resulting from sample preparation and processing. Reporter gene assays, using *lacZ* or *GFP* variants, give superb spatial and temporal resolution but might not reflect the endogenous gene owing to the inclusion of incomplete sequences in the transgene, or artifacts resulting from factors such as the stability of the reporter. In addition, reporter gene constructs often do not include the gene's 3′ UTR, a major source of post-transcriptional regulation – for example, by microRNAs (Ambros and Chen, 2007).

Interactome data

Interactome data provide an invaluable source to study molecular mechanisms that underlie development. Interactome data include both physical and genetic interactions. High-throughput studies on protein-protein physical interactions have become a rapidly developing field. In *S. cerevisiae*, there have been multiple genome-wide studies from different research groups, who have made use of different techniques, such as yeast two-hybrid assays (Ito et al.,

2001; Uetz et al., 2000) and affinity purifications coupled with mass spectrometry (Gavin et al., 2006; Gavin et al., 2002; Ho et al., 2002). Genome-wide yeast two-hybrid studies have also been conducted in flies (Giot et al., 2003), worms (Li et al., 2004) and humans (Rual et al., 2005; Stelzl et al., 2005).

In comparison to physical interaction studies, there have been relatively few high-throughput genetic interaction studies. The first genome-wide genetic interaction study was a synthetic lethality screen conducted in *S. cerevisiae* using a library of deletion mutants (Tong et al., 2001; Tong et al., 2004). These studies indicated the topology of the genetic interaction network (Boone et al., 2007), as well as specific discoveries such as genes involved in DNA replication in response to DNA damage (Budd et al., 2005). In multicellular organisms, RNAi technology (Fire et al., 1998) enables researchers to overcome the problem of using only a small number of available mutants and to design high-throughput (or at least genome-wide) genetic interaction screens (Baugh et al., 2005; Lehner et al., 2006; van Haaften et al., 2004; Dietzl et al., 2007). The scope of these studies has extended from studying a specific process, such as the DNA-damage response (van Haaften et al., 2004), to investigating a broad spectrum of multiple signaling pathways (Lehner et al., 2006). We have also seen the advent of sophisticated analyses of quantitative data to detect genetic interactions (Baugh et al., 2005; Schuldiner et al., 2005; Collins et al., 2007). Quantitative interaction data typically describe the fraction of organisms that have a wild-type phenotype (e.g. survival); thus, if wild type is 100% viable, mutant A is 80% viable and mutant B is 60% viable, a synergistic effect of A and B would be inferred if the A-B double mutant were significantly less viable than 48% (the product of 0.60 and 0.80), the expected value if A and B affect independent processes.

In addition to such high-throughput data sets, the extent of interaction data compiled from small-scale studies has also grown dramatically. In some model organisms, such as *S. cerevisiae* and *C. elegans*, the number of these interactions has reached the same scale as the number of interactions discovered by high-throughput methods (Reguly et al., 2006; Bieri et al., 2007).

However, interactome data are far from perfect. Only a small portion of physical interactions has been confirmed by more than one data set or by genetic interactions (von Mering et al., 2002; Reguly et al., 2006), suggesting that these data are far from complete. In addition, the yeast two-hybrid method can have high false-positive rates, up to 50% in some cases (Fields, 2005), but false positives are likely to vary between experiments and among individual results in a given screen. Several factors can contribute to false positives. For example, two proteins might interact in yeast nuclei but might never be expressed in the same cell in vivo. The particular construct used can also affect the results, so one domain might interact, but the full-length protein might not. However, large-scale analyses include estimates of error rates, and users of these data can have more confidence in them if they understand the methods used in each study. In addition, it is desirable to develop computational methods to explore potential interactions and to improve data reliability.

Transcriptional regulation data

There are many rich sources of data about transcriptional regulation from which developmental biologists would like to infer genetic regulatory networks. We will not review in detail the methods involved in data integration because this active area of bioinformatics deserves its own review, but instead will provide two examples of where automated data integration is useful. Much data integration in this area deals with the association of transcriptional regulators with the sites to which they bind and with gene expression. Individual types of analyses are often successful but have limits, which can be overcome partially by data integration.

An increasingly important technique, chromatin immunoprecipitation (ChIP), detects protein-DNA interactions in vivo and thus indicates where proteins bind to DNA. Genome-wide ChIP analysis has been achieved by detecting DNA fragments precipitated with a particular protein using whole-genome tiling microarrays, the conventional sequencing of fragment ends or, more recently, by sequencing the DNA fragments themselves (Johnson et al., 2007). However, proteins can bind to DNA without having a known functional consequence. These data can be integrated with sequence motifs, conservation of sequence alignment, and with gene expression data in order to make predictions about the genes regulated by a particular transcription factor.

Sandmann et al. (Sandmann et al., 2007) found ~2000 regions that bind the DNA-binding protein/transcriptional activator Twist in *Drosophila* using ChIP and a whole-genome tiling array. They narrowed down these results to ~500 candidate gene regions by integrating information about the co-expression of an associated gene and Twist, and genetic trans-regulation data. Six TWIST-bound regions were tested and all were sufficient to drive reporter gene expression in vivo. In a similar approach, Zeitlinger et al. (Zeitlinger et al., 2007) found hundreds of regions that bind three factors (Dorsal, Twist and Snail). Up to 80% of these regions had motifs for a given factor, as compared with ~35% of randomly chosen sequences. Conservation across 12 *Drosophila* species was used to gain further confidence in these enhancer regions, of which seven were shown to function as transcriptional enhancers in vivo. In both studies, ChIP data were combined with other data to accurately predict enhancer function.

Genomic DNA sequence comparisons of orthologous genes can identify conserved sequence alignments, as well as over-represented sequence motifs. Of course, a single important DNA site can be indistinguishable from a background of sequence if, for example, every gene on average has one copy of a sequence. However, with other information, a single site can be identified. For example, in *C. elegans* the RFX-family regulator DAF-19 binds a 14-mer site (called the Xbox) (Blacque et al., 2005). Blacque et al. compared mRNA from ciliated neurons with all neurons to identify transcripts enriched in ciliated neurons. They then selected those that had an Xbox near the start codon, and successfully predicted expression in ciliated neurons. However, only 42% of genes with correctly positioned Xboxes were expressed in ciliated neurons.

Phenotypic data

Phenotypic data are accumulating at an exponential rate, boosted by technologies such as RNAi and TILLING (Wienholds et al., 2003; Henikoff et al., 2004) (see glossary in Box 1). Genome-wide high-throughput phenotype characterizations have been conducted in yeast using a deletion library (Giaever et al., 2002), in worms by RNAi (Kamath et al., 2003; Simmer et al., 2003) and in *Drosophila* by RNAi on cell lines (Boutros et al., 2004) or by expression of inverted repeats in transgenic flies (Dietzl et al., 2007). Although there is no doubt that high-throughput in vivo studies have identified gene functions during various developmental events, studies in *Drosophila* and mammalian cell lines have also contributed to our knowledge of development by identifying new pathway components and by characterizing drug responses and cellular events (e.g. Berns et al., 2004; Eggert et al., 2004). In addition to high-throughput results, data from small-scale studies of individual genes are also being compiled through the literature curation that is provided by

model organism databases such as FlyBase and Mouse Genome Informatics (MGI). These data provide more-detailed information than do high-throughput data, whereas the high-throughput data have better consistency. Some high-throughput studies focus on a subset of genes instead of the entire genome, which allows results to be analysed in more detail. For example, the study by Piano et al. (Piano et al., 2002) described in detail 47 different early embryonic phenotypes created by the inactivation of 766 genes in *C. elegans*.

Phenotypes that can be associated with genes are often derived from studies in which gene function has been perturbed, such as in mutagenesis screens, targeted knockouts, RNAi screens, by using morpholinos, or in transgenic studies in which genes are overexpressed or modified to produce a constitutively active or dominant-negative protein. No one method is best. For example, the knockout of a gene might reveal only its earliest embryonic function, whereas RNAi or a particular missense mutation obtained in a genetic screen might identify other functions. The overexpression of a redundant gene can reveal its function even though the knockout of a single member of its family produces no discernible phenotype [e.g. alpha factor in yeast (Kurjan and Herskowitz, 1982)]. RNAi experiments may not effectively knockdown the function of many genes. For example, in genome-wide RNAi experiments in *C. elegans*, only 10% of tested genes displayed visible phenotypes on the standard wild-type laboratory strain (Kamath et al., 2003). When a more sensitive strain, *rrf-3*, was used, the number of genes associated with phenotypes increased to 23% (Simmer et al., 2003); however, this still leaves most genes without functional information.

Ontologies and functional annotation

The most common format of these genomic data is a list of genes that have been annotated with their expression patterns or associated with phenotypes in a controlled vocabulary. A more resourceful annotation method involves the use of ontologies, a more structured and controlled vocabulary (see glossary in Box 1 and Fig. 2). Ontology defines each annotation term and the relationship among terms, enabling us to associate genes with related, but not identical, functions (Fig. 2). For example, if one gene is annotated as causing embryonic lethality and another as defective gastrulation, they will not be associated unless a phenotype ontology is used that defines defective gastrulation as a specific case of embryonic lethality. Currently, the Mouse Phenome Database (MPD) and WormBase (see Table 1) have constructed mouse and worm phenotype ontologies, respectively, and provide annotations using these ontology terms. Anatomy ontologies have also been developed by the Zebrafish Information Network (ZFIN) (Sprague et al., 2006) and WormBase (see Table 1).

The Gene Ontology (GO) Consortium (GO Consortium, 2006) provides one of the most widely used platforms of gene function annotations. The ontology contains three sets of controlled vocabularies to describe gene functions: biological process (e.g. embryonic development), cellular component (e.g. nucleus) and molecular function (e.g. protein kinase activity). To ensure consistency, the ontology (see Fig. 2C, for example) was defined by professional curators from multiple model organisms. Each gene is associated with these annotation terms and an evidence code (e.g. IEP, Inferred from Expression Pattern).

GO data are considered to be of high quality by bioinformaticians because they are manually inspected to ensure accuracy, and because they provide detailed information for a large number of genes. Although a developmental biologist might prefer more detail concerning gene function, there is a balance between granularity and coverage when conducting computational studies. In the context of genome-wide coverage, GO provides impressive data quality. Because of this, the GO data set has been used not only as a type of data to be integrated (Jansen et al., 2003), but also as a benchmark to test the performances of various data integration methods (Lee et al., 2004).

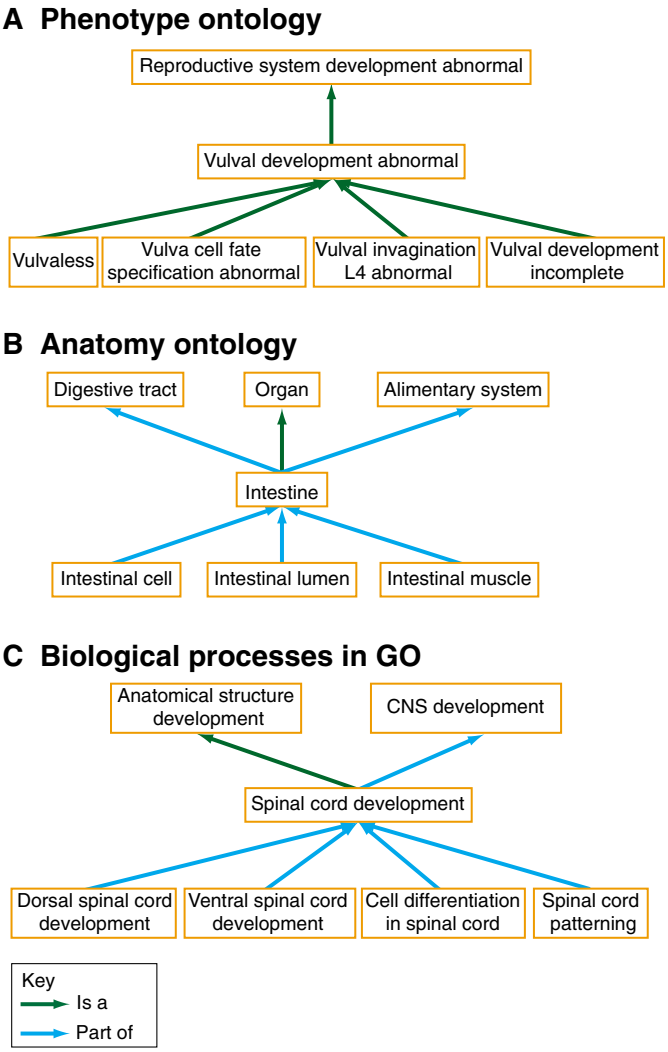

Fig. 2. Examples of bio-ontologies. An ontology captures relationships among terms and their definitions in a structured way. A structure used in many of the current ontologies is a 'directed acyclic graph' that differs from a tree or outline in that one term can connect to many terms but the connection is oriented (shown by arrows rather than by lines) and no cycles are allowed. Commonly used relationships are 'Is a' and 'Part of': term A *is an* example of term B; structure A *is part* of structure B (see www.geneontology.org or www.bioontology.org for more information). (**A**) Phenotype ontology. Reproductive system development defects include vulval developmental abnormalities, which include more-specific phenotypes, such as vulvaless and abnormal cell-fate specification. (**B**) Anatomy ontology. The intestine is part of the 'digestive tract' and 'alimentary system' and is an 'organ'. The intestine comprises intestinal cells, intestinal lumen and intestinal muscle. (A and B from WormBase WS180.) (**C**) Biological processes in the Gene Ontology (GO). 'Spinal cord development' is a case of 'anatomical structure development' and is part of 'central nervous system (CNS) development'. Spinal cord development comprises the development of sub-structures and includes both cell differentiation and patterning. (From GO Biological Process.)

When using GO data, it is important to remember that data with different evidence codes should be processed differently. For example, data with the IEA (Inferred from Electronic Annotation) code may be less reliable than manually curated data. Some GO data may overlap with other biological data sets. For example, in *C. elegans* the RNAi phenotypic data have been converted into GO biological process annotations (e.g. the phenotype 'embryonic lethal' has been converted to the GO process 'embryonic development'). The redundancy of these data sets should be noted if a computational method requires each data set to be independent.

Data integration: why?

Data integration is a useful tool for biologists to navigate through this seemingly overwhelming amount of sometimes contradictory genomic data. By combining various types and sources of information, we can achieve a better assessment of what is known about a protein or process, which will then facilitate our own research.

Take our own experience, for example. We integrated phenotypic, expression, interactome and GO data from yeast, flies and worms to predict genetic interactions in worms (Zhong and Sternberg, 2006). Among the results was a set of novel predicted interactors for *let-60*, which encodes a member of the RAS family and is a crucial regulator of the extent of vulval induction (Beitel et al., 1990; Han et al., 1990; Han and Sternberg, 1990). These interactions were tested by RNAi in a *let-60(gf)* background for suppression or enhancement of the multivulva phenotype. Twelve of 49 of the predictions yielded significant effects. As a striking confirmation of these results, the functional interaction between *let-60* and *tax-6*, one of our verified predicted interactions, was also identified in an independent classical genetic-interaction analysis (W. Johnson and M. Han, personal communication). In our study of *let-60* modifiers in vulva development, data integration facilitated our research by prioritizing which candidates to test.

To further illustrate the benefit of applying data integration to one's own research, we compare two scenarios in which the same study would be conducted with and without data integration. Without computational data integration, genetic screens would be undertaken by a conventional method to study genetic interactions. Lehner et al. (Lehner et al., 2006) conducted an RNAi screen for modifiers of ~30 genes. They applied RNAi to mutants and searched for synthetic phenotypes, such as lethality and growth defects. They found 345 genetic interactions from a total of ~65,000 pairwise tests, making their efficiency (measured as the ratio of discovery versus effort) 345/65,000=0.53%. In the second hypothetical scenario, the same study would be conducted but now with the data integration results. Our computational data predicted 83 interactions for the same genes at a stringent threshold, and 325 interactions at a lower threshold (Zhong and Sternberg, 2006). Had Lehner et al. tested only these 83 (or 325 using the lower threshold) pairs instead of the original 65,000 pairs, they would have recovered 14 (or 41 using the lower threshold) genetic interactions. The efficiency under this strategy is thus 14/83=16.9% (or 41/325=12.6% using the lower threshold). The efficiency improvement in using data integration is thus 32-fold (16.9/0.53) or 24-fold (12.6/0.53) for the two thresholds. Therefore, data integration can greatly decrease the screening and increase the efficiency of discovery, which would be highly desirable when studying complex phenotypes. However, this example also illustrates one limitation of this technique: only a subset of all real interactions (41/345=11.9%) would be recovered. In this section, we discuss what data integration can and cannot do.

Data integration can fill in missing data

Although ongoing literature curation by model organism databases is constantly extending our knowledge of new gene functions, most genome-wide data sets do not cover every gene in multicellular animals and plants. For example, in *C. elegans*, only ~20% (3802/20,000) genes are associated with an anatomical expression pattern in WormBase (WormBase WS166). Genes that are not associated with an annotation constitute a missing data problem. This problem severely limits the scope of our research if we rely on only one data set to conduct the computation (as in the example of *C. elegans* expression, we will miss out 80% of genes). A quick and effective solution to the missing data problem is to increase genome coverage by combining multiple data sets that cover different groups of genes. For example, in our study of genetic interaction predictions, we noticed that only 292 *C. elegans* genes were annotated for all three data types of phenotype, anatomical expression and GO process (Zhong and Sternberg, 2006). By adding cross-species data, we were able to expand our scope and predict genetic interactions for over 2200 genes. Integrating other data sets enabled us to 'borrow' information from other experiments and other species to piece together a more complete picture. Each type, source and quality of data will carry a different weight. For example, in our study, not surprisingly, anatomical-level expression data from *C. elegans* was more informative than similar data from *D. melanogaster*; however, *Drosophila* phenotypic information was more informative than similar data from *C. elegans* because there were only a limited number of phenotypic annotations in WormBase at that time.

Data integration can reduce noise

Data sets are often noisy, containing false positives and false negatives. Also, a data set may be a good predictor of one type of prediction, but a weak predictor for another type of prediction. Data integration improves data reliability by confirming a conclusion with several independent experiments. Therefore, we can filter out erroneous information and increase the overall predictive strength by combining several weak predictions.

In one example, Natarajan et al. (Natarajan et al., 2006) integrated a set of data obtained by the Alliance for Cell signaling in which they treated a macrophage cell line with single and multiple ligands and measured multiple readouts, including calcium and cAMP dynamics, cytokine expression and phosphorylation of signaling proteins. The data for each cell biological readout was converted to a z-score, which is essentially the number of standard deviations from the relevant control, and then the z-scores summed to obtain an integrated view of each ligand treatment. This integration allowed them to compare treatments, learning, among other things, that receptor-stimulated calcium mobilization increases cAMP production stimulated by a distinct signal transduction pathway.

It is, however, debatable whether increasing the number of data sets will necessarily improve data quality in terms of reducing false positives and false negatives. When predicting protein-protein interactions, a study of yeast genomic data integration has suggested that there is a limit to improving performance by the integration of more data sets (Lu et al., 2005). At some point, the utility of adding more data sets saturates and thus clustering additional data sets, especially those constituting weak predictors and data of poor quality, introduces confusion instead of further reducing noise (Lu et al., 2005). With current multicellular organism data sets, we are probably still far from this saturation point, and the benefit of reducing missing data might outweigh the concern of data saturation.

Caveats of data integration

As shown by our example of comparing brute force and prediction-guided tests, although data integration can promote discovery, it often can recover only a subset of all hits. The scope of data integration is limited to the genes that exist in data sets. As a result of the biased genome coverage of some data sets (especially those compiled from small-scale studies), data integration tends to favor genes that have been well studied and genes that are conserved in multiple organisms. Therefore, data integration cannot replace genetic screens for predicting gene function and interactions because genetic screens provide unbiased coverage of the genome: genes can be identified without any prior assumption or knowledge.

In most cases, data integration requires experimental validation. To various degrees, most data integration techniques rely on certain assumptions and simplifications of biological data. For example, cross-species data integration assumes that gene functions are conserved in different organisms. We thus have to apply other methods to cross-validate these computational results. Experimental verification is a direct and convincing way to detect exceptions to these assumptions. When computational predictions are proved to be correct, experiments can also provide more-detailed biological information.

Because of these limitations, we consider data integration as a means of prioritizing experiments rather than dictating which experiments to pursue.

Data integration: how?

We explain in this section, with a few examples, the basic steps of data integration. In brief, the steps are: to obtain data in a consistent, structured form that can then be fed into a computer program; to define orthologs when data are to be integrated across species; to select a statistical model with which to integrate data; and to choose a threshold by which to interpret the output of the computer program.

Make data computable

Although numeric data are desirable because they can be used directly, data in the format of annotated text can be converted into a form that is amenable to information extraction and mathematical computation. One simple way to convert text annotations to numbers is to use a binary code (0, 1) to denote the absence and presence of an annotation term. Piano et al. (Piano et al., 2002) used this method to convert *C. elegans* embryonic phenotype annotations into a string of binary numbers. This enabled them to compare the similarity in phenotype of two genes by computing the correlation of the two strings of 0s and 1s (shown as yellow and blue in Fig. 3).

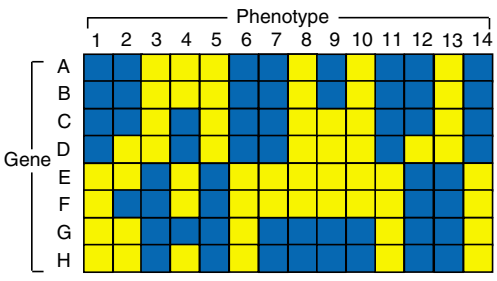

Fig. 3. Correlating a spectrum of phenotypes. A set of 14 phenotypes for eight genes is indicated by the presence (blue) or absence (yellow) of the phenotype. In this example, genes A and B are perfectly correlated (14 of 14 phenotypes), genes C and D are tightly correlated (12 of 14 phenotypes), and genes A-D are more correlated with each other than with E-H. This data representation allows genes and phenotypes to be clustered and calculations of pairwise correlation coefficients to be made.

Annotation statistics can provide additional information, especially when an ontology is used. For example, in GO, the general term 'developmental process' is associated with as many as 17,281 genes, whereas the more specific term 'embryonic development' is associated with only 4826 genes (http://www.geneontology.org, as of July 2007). The number of genes associated with an annotation term thus indicates how specific the term is. Such information can prove to be very useful when predicting functional interactions because if two genes share a more specific term then they are more likely to interact than genes that share a general term (Jansen et al., 2003; Zhong and Sternberg, 2006).

Another method is to convert heterogeneous data into weighted scores. A popular scoring scheme uses likelihood ratios (see glossary, Box 1) (Jansen et al., 2003; Lee et al., 2004; Rhodes et al., 2005; Zhong and Sternberg, 2006). This requires a training set (see glossary, Box 1) with known positives and negatives. The likelihood ratio is the value of the frequency of a feature appearing in the positive training set divided by the frequency in the negative training set. A high likelihood ratio indicates that more positives than negative have the feature. If an unknown gene has this feature, a high likelihood ratio is awarded indicating that the gene is more likely to have a positive outcome.

Map orthologs

Orthologs are homologs that diverged concomitant with the divergence of species. When performing cross-species data integration, one important step is to select a good method to identify orthologs among species. Most of us are familiar with assessing orthology for individual genes of interest, but when the process is scaled up to ~20,000 genes, some automation is necessary.

A number of automated strategies to detect orthologs have been proposed, and several databases have been devoted to ortholog mapping (Fig. 4). Many methods detect reciprocally best-matching proteins from BLAST searches (Altschul et al., 1997) as orthologs; for example, NCBI KOG [euKaryotic Orthologous Groups, http://www.ncbi.nlm.nih.gov/COG/ (Tatusov et al., 2003)], InParanoid [http://inparanoid.sbc.su.se (Remm et al., 2001)] and its multi-species extensions, such as OrthoMCL [http://orthomcl.cbil.upenn.edu/cgi-bin/OrthoMclWeb.cgi (Li et al., 2003)] and MultiParanoid [http://www.sbc.su.se/~andale/multiparanoid/html/index.html (Alexeyenko et al., 2006)]. Using a different approach, TreeFam (http://www.treefam.org) constructs phylogenetic trees and uses manual curation to annotate orthologs (Li et al., 2006).

A good ortholog mapping method should identify all members in an ortholog group and avoid large groups of paralogs. For example, in the hypothetical protein family of Fig. 4C, applying a two-species InParanoid analysis of species B and C will correctly identify two ortholog groups, (B1, C1) and (B2, C2, C3); but including a distant species A in a KOG analysis might erroneously lead to all the proteins being grouped together as one group. Based on its phylogenetic approach and manual supervision, TreeFam probably provides the highest data quality.

A good ortholog mapping method should also be updated with the latest gene model and annotation changes. Since gene models are updated continually as genome sequence, assembly and gene finding improves, ortholog analysis needs to be based on the latest sequence data. InParanoid, OrthoMCL and MultiParanoid provide open

source codes at their websites and thus can be installed to analyze any desired genome sequences. However, it might require a substantial computational resource to conduct these analyses when a large number of genomes are included.

Select a statistical model

The simplest statistical model to use for integrating multiple data sets is a voting system (Fig. 5A). For example, one data set gets one vote, and the total number of votes is then added up for final decisions. The voting system makes it easy to vary the stringency with which data are selected. At one extreme, when the threshold vote number is set to one, the system selects the union of all data sets. At the other extreme, when the threshold is as high as the total number of data sets, then only the intersection of all data sets is allowed and the system becomes a filtering model (intersection). Because of its simplicity and no requirement of any training data set, the voting method has been extensively used in various developmental studies. For example, in *C. elegans*, the strategy has been applied to integrate phenotype, expression and protein-protein interaction data to study germline development (Walhout et al.,

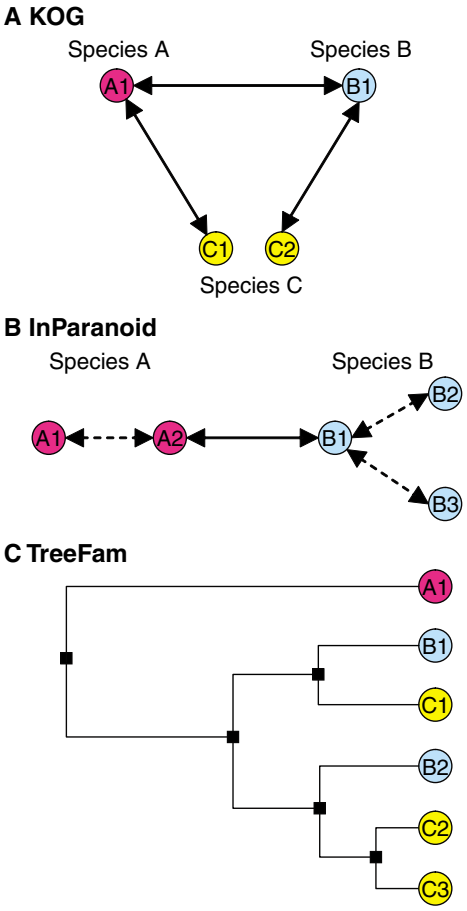

Fig. 4. Three methods of assigning orthology relationships. Species are designated by letters and paralogs by numbers. (**A**) KOG. The NCBI KOG detects reciprocally best-matching proteins from BLAST searches as orthologs. An ortholog group is thus defined as the union of best BLAST hits among all pairwise comparisons of multiple species. In the example shown, species A and species B each have a 1:1 ortholog, but species C has two orthologous proteins. (**B**) InParanoid. Since inter-genome reciprocal best BLAST analysis forces a one-to-one relationship, InParanoid also detects intra-genome best BLAST hits as co-orthologs. Solid arrows, inter-genome BLAST; dashed arrows, intra-genome BLAST. (**C**) TreeFam. In this approach, the relationships among proteins are defined by phylogenetic analysis.

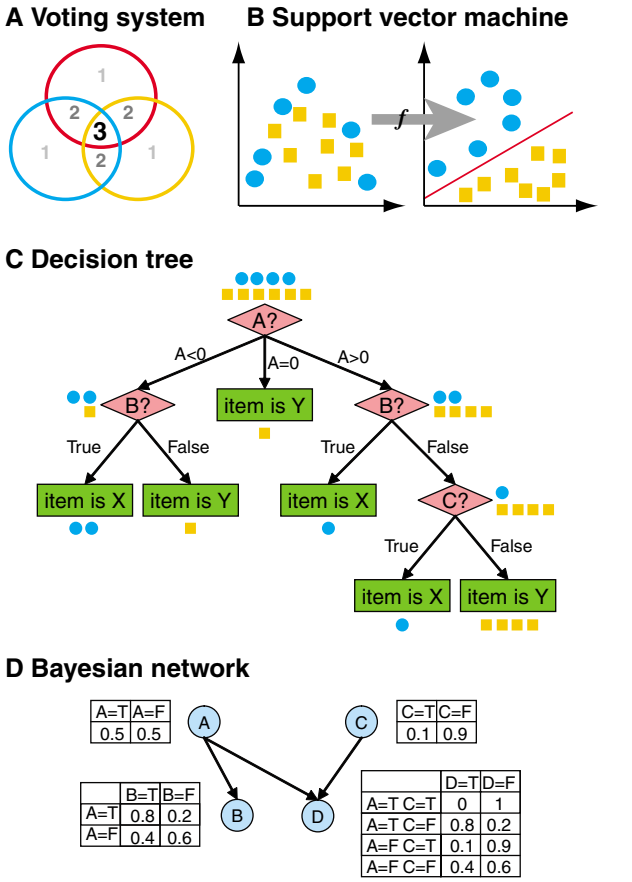

Fig. 5. Four examples of statistical models for data integration. (**A**) Voting system. Each circle represents one data set and has one vote. Gray numbers indicate total votes. Data that are confirmed by multiple data sets have multiple votes. In this example, there are three data sets; thus, three is the maximum number of possible votes. (**B**) Support vector machine. Blue circles indicate positives in the training set and yellow squares represent negatives. In this example, there are two attributes (as represented by the *x*- and *y*-axes) for each data point. The data are plotted based on the values of these attributes. A function *f* is used to convert the data points so that they become linearly separable. The training set is used to derive the one-dimensional plane (red line) that separates positives from negatives. (**C**) Decision tree. In this hypothetical tree, the goal is to classify the input items into two categories, X and Y, which are denoted as blue circles and yellow squares, respectively. The category of each item is hidden, but we know the values of its three attributes (A,B,C). We use a set of conditions (represented by pink diamonds) to evaluate these attributes. Based on their values, we separate the items into subsets. The separation continues until the final outcome of the items (leaf nodes, represented by green boxes) is reached. (**D**) Bayesian network. In Bayesian networks, nodes represent variables and edges represent variable dependencies. Here, each node represents a Boolean variable, the value of which is denoted as true (T) or false (F) in conditional probability tables. The edges indicate that the value of B is dependent on the value of A and that the value of D is influenced by both the values of A and C. The conditional probability tables detail such dependency. For example, the probability of B being true is 0.8 if A is true; the probability drops to 0.4 if A is false. This network enables us to derive probabilities from different attribute values – for example, the probability of A being true given that B is true.

Box 2. Using data integration to study *C. elegans* embryogenesis: a success story

In 2005, Gunsalus et al. integrated RNAi phenotypic data, microarray profiles and large-scale yeast two-hybrid results from *C. elegans* to predict a gene function network that regulates *C. elegans* early embryogenesis. In this study, they constructed a graph in which each of 661 genes is a node. Edges are drawn in this graph if the protein products of genes interact, or if the spectrum of phenotypes associated with the perturbation of two genes is highly correlated, or if the expression of two genes is highly correlated. By doing this, they obtained 31,173 connections among the 661 genes. A multiple-support graph was then developed in which two genes were linked if there were two or more edges linking them in the previous graph. This graph contains 305 nodes and 1,036 edges. It was this multiple-support graph that was analyzed computationally to identify highly interconnected sets of genes (sub-graphs), which can be interpreted as sets of gene products that act together, possibly as macromolecular machines. Indeed, the tight sub-graphs identified in this study correspond to known machines, such as the anaphase promoting complex (APC). Gunsalus et al. tested the predictive power of their approach by examining the subcellular localization dynamics of ten novel proteins using GFP translational fusions. In most cases, the localization was highly consistent with their predictions. For example, proteins connected to those involved in DNA synthesis licensing factors localized to condensed chromosomes at metaphase. The success of this study is due in part to its focus on one aspect of *C. elegans* development, and the authors' great expertise in the biological detail of this process.

2002) and embryogenesis (Gunsalus et al., 2005) (see Box 2).

When training sets are available, more-sophisticated statistical models can be used. *S. cerevisiae* has been a test-bed for such bioinformatic methods. For example, yeast protein and genetic interactions have been inferred by Bayesian networks (Troyanskaya et al., 2003), decision trees (Wong et al., 2004) and kernel methods (Ben-Hur and Noble, 2005) (see below). Whereas some of these methods require substantial amounts of programming, most of the statistical models can be constructed with the assistance of freely available software packages. The computational principle behind these methods is to formalize different data types into a number of attributes, to detect the patterns of these attributes in known positives and negatives, and to define a process (function) to differentiate positives from negatives based on their attribute values.

For example, if there are a total of n attributes, then each data point in the training set is a vector of n values, and is associated with an outcome. Kernel methods, in particular support vector machines (SVMs), use a kernel function (often non-linear) to transform these attribute values so that they are linearly separable, and then use a regression (often a simple linear regression) to derive a hyperplane that achieves the maximum separation between positives and negatives (Fig. 5B). We can then use this hyperplane to classify an unknown item into a positive or a negative outcome based on its attribute values.

A decision tree judges one condition at a time and reduces the problem at each step (Fig. 5C). Starting from the root node, a condition (e.g. do the two genes have the same phenotype?) is applied, and the unknown item is classified into one of the daughter nodes based on its values. The decision-making step continues until a leaf node is reached where the unknown item is classified into a final category.

Bayesian networks (see glossary, Box 1) use a directed acyclic graph to represent variables and their relationships. The network is composed of nodes that represent variables, directed edges that link two dependent nodes, and probability tables (Fig. 5D). The training set tells us the conditional probability of an outcome given the values of all variables. Bayesian networks can use this information together with the prior probabilities to compute the inverse problem, that is, given a set of values, what the probability of an outcome is (Eddy, 2004).

Box 3. Putting it together: an application of advanced data integration techniques

Jansen et al. (Jansen et al., 2003) have successfully applied Bayesian networks (see glossary, Box 1) to integrate various genomic data sets to predict protein-protein interactions in yeast. Four data types were used: yeast mRNA expression, biological function, phenotype (viable or not), and protein physical interaction data. Jansen et al. constructed a training set using hand-curated protein-protein interactions as positives and proteins located in separate subcellular compartments as negatives. Equipped with this training set, they computed the likelihood ratios (see glossary) for each data set to convert them into weighted scores. The scores were computed at a detailed level. For example, different scores were assigned to GO terms with different annotation statistics so that a GO term describing a specific biological process had a higher score than a broadly defined GO term. These scores were combined using a Bayesian networks approach to produce a final, probabilistic protein-protein interaction network. Mass spectrophotometric analysis of 98 affinity-purified protein complexes (tandem affinity purification 'TAP-tagging' experiments) verified 424 predicted interactions, validating the accuracy of their predicted network.

A simplified Bayesian network model, naïve Bayes classifier, has become one of the most popular methods for integrating data to predict gene functions (Jansen et al., 2003; Lee et al., 2004; Rhodes et al., 2005) (see example in Box 3). This method assumes all variables to be independent, and thus simplifies the computation down to an easy product calculation: the final score is the product of the likelihood ratios of all features of the input gene(s). The higher the scores, the more likely that the genes are positives. A cut-off value is then applied to the final score.

Although numerous approaches have been developed, there has been no systematic study of the different statistical models. However, the bottleneck in data integration is likely still to be the data collection step. The impact from statistical models seems relatively minor considering the severity of the missing data problem and the unreliable nature of the data quality in multicellular organisms. When the performance of a logistic regression model to predict *C. elegans* genetic interactions was compared with that of a simple naïve Bayes classifier, only minor improvements were observed (Zhong and Sternberg, 2006), suggesting that improving data quality should be of higher priority than finding better statistical models.

Choose a threshold

The threshold stringency directly affects the ratios of false positives and false negatives in the final computational predictions. As developmental biologists, we care more about individual genes and proteins than about statistical properties of the genome and data sets, but such properties provide important clues. False positives can only be eliminated by experiment. If 80% of inferences are false positives, then on average we have to test five inferences to obtain one true result. If the experiments are relatively straightforward (double mutants; examining gene expression), then such rates might be acceptable. If the experiments are less straightforward (making a conditional knockout mouse; determining a crystal structure), then those rates are probably not acceptable.

The extent of false negatives depends on the threshold set for predictions. If one followed up on all predictions, then there would be no false negatives, but then one would gain nothing from the computation. There is no correct threshold, but rather it depends on the value and costs placed on following predictions.

Conclusions

We have discussed how data integration, pioneered in yeast, has been extended to metazoans, especially *C. elegans*. Data integration is likely to be especially important in studies of mammalian development, where the interest, complexity and cost of experiments are typically higher than in invertebrate systems. The desire to maximize the yield per experiment should lead a developmental biologist to want to use automated data integration.

This review has focused on integrating large-scale data sets to learn about individual genes, but the analyses of features of whole interaction networks promises additional, global insights (reviewed by Albert, 2005). Features of networks, such as the degree of sub-structure and the average connectedness of genes, are likely to be related to functional features. In addition, stereotyped features of networks (network motifs) raise general questions about relationships among genes. Protein-protein interactions can be analyzed at the level of individual domains, and many genetic interactions can be analyzed in terms of specific alleles that affect parts of proteins. Taken together, these data might provide a higher resolution view of interaction networks.

Data integration is not an isolated field. Further improvements in data integration are closely tied to the progress of source databases, such as the synchronization of data annotation with gene model changes, the standardization of ortholog mapping techniques and the development of bio-ontologies and their consistent compilation, often by a professional curator.

Negative data are relatively underreported in the published literature. As discussed above, sophisticated data integration methods often use positive and negative training sets. Authors, reviewers and editors could help the cause by encouraging the inclusion of negative observations ('gene A does not interact with gene B for phenotype C') in papers, for example as supplemental data.

Many of the model organism databases and the Gene Ontology Consortium that are crucial to data integration are supported by the National Human Genome Research Institute (e.g. WormBase grant HG02223 to P.W.S. and Gene Ontology Consortium grant HG002273 to Judith Blake). We thank Min Han for communicating unpublished results, and Alok Saldanha, Erich Schwarz, Xiaodong Wang and anonymous reviewers for comments on the manuscript. Relevant genetics research in our laboratory is supported by the Howard Hughes Medical Institute, with which P.W.S. is an Investigator and W.Z. an Associate.

References

Albert, R. (2005). Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947-4957.
Alexeyenko, A., Tamas, I., Liu, G. and Sonnhammer, E. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-e15.
Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
Ambros, V. and Chen, X. (2007). The regulation of genes and genomes by small RNAs. *Development* **134**, 1635-1641.
Antoshechkin, I. and Sternberg, P. W. (2007). The versatile worm: genetic and genomic resources for Caenorhabditis elegans research. *Nat. Rev. Genet.* **8**, 518-532.
Baugh, L. R., Wen, J. C., Hill, A. A., Slonim, D. K., Brown, E. L. and Hunter, C. P. (2005). Synthetic lethal analysis of *Caenorhabditis elegans* posterior embryonic patterning genes identifies conserved genetic interactions. *Genome Biol.* **6**, R45.
Beitel, G. J., Clark, S. G. and Horvitz, H. R. (1990). *Caenorhabditis elegans* ras gene *let-60* acts as a switch in the pathway of vulval induction. *Nature* **348**, 503-509.
Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21** Suppl. 1, i38-i46.
Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., Kerkhoven, R. M., Madiredjo, M., Nijkamp, W., Weigelt, B. et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway, *Nature* **428**, 431-437.
Bieri, T., Blasiar, D., Ozersky, P., Antoshechkin, I., Bastiani, C., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P. et al. (2007). WormBase: new content and better access. *Nucleic Acids Res.* **35**, D506-D510.
Blacque, O. E., Perens, E. A., Boroevich, K. A., Inglis, P. N., Li, C., Warner, A., Khattra, J., Holt, R. A., Ou, G., Mah, A. K. et al. (2005). Functional genomics of the cilium, a sensory organelle. *Curr. Biol.* **15**, 935-941.
Boone, C., Bussey, H. and Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.* **8**, 437-449.
Boutros, M., Kiger, A. A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S. A., Heidelberg Fly Array Consortium, Paro, R. and Perrimon, N. (2004). Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832-835.
Budd, M. E., Tong, A. H., Polaczek, P., Peng, X., Boone, C. and Campbell, J. L. (2005). A network of multi-tasking proteins at the DNA replication fork preserves genome stability. *PLoS Genet.* **1**, e61.
Christiansen, J. H., Yang, Y., Venkataraman, S., Richardson, L., Stevenson, P., Burton, N., Baldock, R. A. and Davidson, D. R. (2006). EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res.* **34**, D637-D641.
Collins, S. R., Miller, K. M., Maas, N. L., Roguev, A., Fillingham, J., Chu, C. S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M. et al. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-810.
Crosby, M. A., Goodman, J. L., Strelets, V. B., Zhang, P., Gelbart, W. M. and The FlyBase Consortium (2007). FlyBase: genomes by the dozen. *Nucleic Acids Res.* **35**, D486-D491.
Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A. M., Grove, C. A., Martinez, N. J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J. S., Hope, I. A. et al. (2006). A gene-centered *C. elegans* protein-DNA interaction network. *Cell* **125**, 1193-1205.
Dietzl, G., Chen, D., Schnorrer, F., Su, K. C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S. et al. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448**, 151-156.
Dolphin, C. T. and Hope, I. A. (2006). *Caenorhabditis elegans* reporter gene fusion genes generated by seamless modification of large genomic DNA clones. *Nucleic Acids Res.* **34**, e72.
Dupuy, D., Li, Q. R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I. A. et al. (2004). A first version of the *Caenorhabditis elegans* promoterome. *Genome Res.* **14**, 2169-2175.
Eddy, S. R. (2004). What is Bayesian statistics? *Nat. Biotechnol.* **22**, 1177-1178.
Eggert, U. S., Kiger, A. A., Richter, C., Perlman, Z. E., Perrimon, N., Mitchison, T. J. and Field, C. M. (2004). Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol.* **2**, e379.
Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E. and Mouse Genome Database Group (2007). The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.* **35**, D630-D637.
Fernandes, J. S. and Sternberg, P. W. (2007). The tailless ortholog *nhr-67* regulates patterning of gene expression and morphogenesis in the *C. elegans* vulva. *PLoS Genet.* **3**, e69.
Fields, S. (2005). High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.* **272**, 5391-5399.
Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806-811.
Furlong, E. E., Andersen, E. C., Null, B., White, K. P. and Scott, M. P. (2001). Patterns of gene expression during *Drosophila* mesoderm development. *Science* **293**, 1629-1633.
Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.
Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B. et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-636.
Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B. et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-391.
Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E. et al. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-1736.

GO Consortium (2006). The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34**, D322-D326.
Gong, S., Zheng, C., Doughty, M. L., Losos, K., Didkovsky, N., Schambra, U. B., Nowak, N. J., Joyner, A., Leblanc, G., Hatten, M. E. et al. (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917-925.
Gray, P. A., Fu, H., Luo, P., Zhao, Q., Yu, J., Ferrari, A., Tenzen, T., Yuk, D. I., Tsung, E. F., Cai, Z. et al. (2004). Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science* **306**, 2255-2257.
Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J. D., Hao, T., Berriz, G. F., Bertin, N., Huang, J., Chuang, L. S. et al. (2005). Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861-865.
Han, M. and Sternberg, P. W. (1990). *let-60*, a gene that specifies cell fates during *C. elegans* vulval induction, encodes a ras protein. *Cell* **63**, 921-931.
Han, M., Aroian, R. V. and Sternberg, P. W. (1990). The *let-60* locus controls the switch between vulval and nonvulval cell fates in *Caenorhabditis elegans*. *Genetics* **126**, 899-913.
Henikoff, S., Till, B. J. and Comai, L. (2004). TILLING. Traditional mutagenesis meets functional genomics. *Plant Physiol.* **135**, 630-636.
Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G. and Brown, E. L. (2000). Genomic analysis of gene expression in *C. elegans*. *Science* **290**, 809-812.
Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K. et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183.
Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691.
Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G. et al. (2005). Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345-350.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569-4574.
Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453.
Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502.
Joyce, A. R. and Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* **7**, 198-210.
Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M. et al. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 220-221.
Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. and Davidson, G. S. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087-2092.
Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y. et al. (2002). Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707-719.
Kurjan, J. and Herskowitz, I. (1982). Structure of a yeast pheromone gene (MF alpha): a putative alpha-factor precursor contains four tandem copies of mature alpha-factor. *Cell* **30**, 933-943.
Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R. and Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nat. Methods* **2**, 337-344.
Lee, I., Date, S. V., Adai, A. T. and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* **306**, 1555-1558.
Lee, R. (2005). Web resources for *C. elegans* studies. In *WormBook* (ed. The *C. elegans* Research Community), WormBook, doi/10.1895/wormbook.1.7.1, http://www.wormbook.org.
Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A. G. (2006). Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* **38**, 896-903.
Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J. et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168-176.
Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J. K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L. et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572-D580.
Li, L., Stoeckert, C. J., Jr and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178-2189.
Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T. et al. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540-543.
Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* **15**, 945-953.
Magdaleno, S., Jensen, P., Brumwell, C. L., Seal, A., Lehman, K., Asbury, A., Cheung, T., Cornelius, T., Batten, D. M., Eden, C. et al. (2006). BGEM: an *in situ* hybridization database of gene expression in the embryonic and adult mouse nervous system. *PLoS Biol.* **4**, e86.
Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86.
McKay, S. J., Johnsen, R., Khattra, J., Asano, J., Baillie, D. L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E. et al. (2004). Gene expression profiling of cells, tissues and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 159-169.
Members of the Toxicogenomics Research Consortium (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* **2**, 351-356.
Natarajan, M., Lin, K. M., Hsueh, R. C., Sternweis, P. C. and Ranganathan, R. (2006). A global analysis of cross-talk in a mammalian cellular signalling network. *Nat. Cell Biol.* **8**, 571-580.
Parrish, J. Z., Kim, M. D., Jan, L. Y. and Jan, Y. N. (2006). Genome-wide analyses identify transcription factors required for proper morphogenesis of *Drosophila* sensory neuron dendrites. *Genes Dev.* **20**, 820-835.
Piano, F., Schetter, A. J., Morton, D. G., Gunsalus, K. C., Reinke, V., Kim, S. K. and Kemphues, K. J. (2002). Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans. Curr. Biol.* **12**, 1959-1964.
Pollet, N., Muncke, N., Verbeek, B., Li, Y., Fenger, U., Delius, H. and Niehrs, C. (2005). An atlas of differential gene expression during early Xenopus embryogenesis. *Mech. Dev.* **122**, 365-439.
Quackenbush, J. (2004). Data standards for 'omic' science. *Nat. Biotechnol.* **22**, 613-614.
Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A. et al. (2006). Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5**, 11.
Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041-1052.
Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M. et al. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31**, 224-228.
Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* **23**, 951-959.
Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N. et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178.
Sandmann, T., Jakobsen, J. S. and Furlong, E. E. (2006). ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. *Nat. Protoc.* **1**, 2839-2855.
Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.
Schuldiner, M., Collins, S. R., Thompson, N. J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J. F. et al. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507-519.
Simmer, F., Moorman, C., van der Linden, A. M., Kuijk, E., van den Berghe, P. V., Kamath, R. S., Fraser, A. G., Ahringer, J. and Plasterk, R. H. (2003). Genome-wide RNAi of *C. elegans* using the hypersensitive *rrf-3* strain reveals novel gene functions. *PLoS Biol.* **1**, E12.
Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Mani, P., Ramachandran, S. et al. (2006). The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.* **34**, D581-D585.
Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S. et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968.
Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-255.
Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya,

A. N. et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.
Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E. et al. (2002). Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **3**, research0088.1-0088.14.
Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H. et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-2368.
Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M. et al. (2004). Global mapping of the yeast genetic interaction network. *Science* **303**, 808-813.
Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA* **100**, 8348-8353.
Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627.
van Haaften, G., Vastenhouw, N. L., Nollen, E. A. A., Plasterk, R. H. A. and Tijsterman, M. (2004). Gene interactions in the DNA damage-response pathway identified by genome-wide RNA-interference analysis of synthetic lethality. *Proc. Natl. Acad. Sci. USA* **101**, 12992-12996.
Visel, A., Thaller, C. and Eichele, G. (2004). GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res.* **32**, D552-D556.
von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.
Walhout, A. J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K. C., Schetter, A. J. et al. (2002). Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**, 1952-1958.
White, K. P., Rifkin, S. A., Hurban, P. and Hogness, D. S. (1999). Microarray analysis of *Drosophila* development during metamorphosis. *Science* **286**, 2179-2184.
Wienholds, E., van Eeden, F., Kosters, M., Mudde, J., Plasterk, R. H. and Cuppen, E. (2003). Efficient target-selected mutagenesis in zebrafish. *Genome Res.* **13**, 2700-2707.
Wong, S. L., Zhang, L. V., Tong, A. H., Li, Z., Goldberg, D. S., King, O. D., Lesage, G., Vidal, M., Andrews, B., Bussey, H. et al. (2004). Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. USA* **101**, 15682-15687.
Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A. and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev.* **21**, 385-390.
Zhong, W. and Sternberg, P. W. (2006). Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311**, 1481-1484.