Published in final edited form as:

Semin Nephrol. 2010 September ; 30(5): 443–454. doi:10.1016/j.semnephrol.2010.07.002.

Integrative Systems Biology for Data Driven Knowledge Discovery

Casey S. Greene¹ and Olga G. Troyanskaya²

¹Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ

²Department of Computer Science and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ

Abstract

Integrative systems biology is an approach that brings together diverse high throughput experiments and databases to gain new insights into biological processes or systems at molecular through physiological levels. These approaches rely on diverse high-throughput experimental techniques that generate heterogeneous data by assaying varying aspects of complex biological processes. Computational approaches are necessary to provide an integrative view of these experimental results and enable data-driven knowledge discovery. Hypotheses generated from these approaches can direct definitive molecular experiments in a cost effective manner. Using integrative systems biology approaches, we can leverage existing biological knowledge and large-scale data to improve our understanding of yet unknown components of a system of interest and how its malfunction leads to disease.

1. What is Integrative Systems Biology?

In the modern era of genome-scale experimental biology, researchers are capable of measuring genetic variation [1-2], transcript abundance [3-5], and protein concentration [6-7] on a broad scale in a cost effective manner. Technological advances have made the profiling of cellular states inexpensive and comprehensive. Unfortunately turning high-throughput experimental data into understanding, knowledge, and models of biological processes is very challenging. For example, in the field of human genetics the genome-wide association study held great promise for discovering genetic variations underlying complex traits such as common human diseases [8-9]. To this point results from these studies have often proven difficult to interpret and replicate [10-11]. Even findings that do replicate have had limited utility for assessing individual disease risk [12]. Systems biology is an approach designed to address this problem by treating biological components as part of a complex system.

The integrative systems biology approach we discuss is a data driven hypothesis generation approach. These approaches generate hypotheses for experimental follow-up based on high throughput experimental data. This is in contrast to the standard knowledge based approach, where scientists perform a literature review to develop a hypothesis. The knowledge based

approach can be made more efficient by high quality databases, but it is fundamentally limited by the completeness of the literature. The data driven approach can generate hypotheses based on the entirety of the high throughput data. This makes the data driven approach complementary to traditional knowledge driven hypothesis generation methods.

A. Genome-scale experimental techniques

Modern high throughput methods are capable of measuring many important biological molecules. Human genome-scale experimental techniques include microarrays [3-5], genome-wide association studies [8-9], proteomics studies [13-14], and RNA interference screens [15-16] among many other experimental designs [13-14, 17-21]. These experiments range from those targeted towards tissue specificity [22] to those targeted towards specific diseases such as cancer [23]. We briefly review modern methods for measuring the genome, transcriptome, and proteome.

Modern methods of measuring the genome include chip based technologies for measuring single nucleotide polymorphisms and copy number variations as well as next generation methods of directly sequencing the genome of interest. Chip based methods are capable of measuring more than one million single nucleotide polymorphisms [1, 24]. Because many polymorphisms are linked, this makes the indirect measurement of many more variations possible, although precise numbers are dependent on the population of interest [25]. Spencer et al. provide a thorough overview of considerations for genome-wide association studies [26]. Beyond these chip based technologies, next generation sequencing methods are capable of sequencing billions of bases per in a single run [27]. Currently available high throughput datasets are often from chip based methods, while datasets with next generation sequencing data are on the horizon. These methods, in the context of high quality studies, can provide important information about the genome for integrative approaches.

Modern methods of measuring RNA levels include chip based technologies such as microarrays, as well as methods based on next generation sequencing often called RNASeq [19]. RNASeq has now been applied to *S. cerevisiae* [28], *S. pombe* [29], mouse [30], HeLa [21], and stem cells [20] among an increasing list of cell types and organisms. These approaches seem to be as reproducible as modern array technology [31], and have an advantage over microarrays because they can examine a wider array of transcripts. At the moment, most existing datasets of RNA levels are still from microarray studies. The NCBI Gene Expression Omnibus [32], a database of microarrays alone, contains over 700 human datasets collected under diverse experimental conditions encompassing more than 8000 individual arrays.

Modern methods for measuring the proteome consist of both chip based and mass spectrometry based experimental techniques. Chip based methods use antibodies spotted onto slides to measure relative protein concentrations [6]. Modern mass spectrometry based proteomic techniques often employ liquid chromatography and tandem mass spectrometry to identify peptides and determine their relative abundance [13-14]. The human PeptideAtlas [33], resource for mass spectrometry based proteomics experiments, currently contains almost 6.7 million MS/MS spectra representing almost 84,000 non-singleton peptides across 220 samples.

In addition to these high throughput experiments, there are databases of biochemical pathways [34], gene function [35], pharmacogenomics [36], and protein-protein interactions [37-39] designed provide large scale summary of experimental knowledge. These databases provide genome-scale information that can be used as part of an integration process to discover novel functional relationships between genes. Even in the age of genome-scale experimental biology, analyzing results from single experiments or experimental platforms can be limiting.

B. Systems biology

The systems biology approach involves placing components of interest in the context of all other cellular components to gain an understanding of the system as a whole [40]. While this seems relatively straightforward, the problem has been approached from a number of different directions, which can be classified as bottom-up or top-down approaches [41].

Bottom-up approaches focus on developing detailed models of biological components at a small scale. These measurements are then used to build a model of the system that can be probed computationally. These *in silico* experiments can suggest molecular experiments for follow-up, which can then be used to refine the model [42]. The goals of these approaches are to build a detailed understanding of the interactions of cellular components in the context of the surrounding biological system.

Top-down approaches focus on implicating previously unknown players in a process and analyzing data in the context of a specific pathway or function. An early example of this would be a pathway based microarray analysis, where gene ontology terms are used, for instance, to group genes during the analysis. In this framework the pathway becomes the unit of analysis and statistical testing as opposed to a gene. This can be used as a late analytical step to interpret a list of genes discovered in a high throughput experiment [43-44] or as a primary analysis method [45].

In general, bottom-up approaches are most useful for processes where many of the players are known although the direct mechanisms of action may not be understood. These approaches can benefit from detailed probing and measurement of specific pathways. At the moment these are usually well characterized pathways, often involved in cellular proliferation and cancer [46]. The top-down strategies are most useful when many components of a process may not yet be known. These can be broader processes or those which have not been as thoroughly studied. In yeast, for instance, computational top-down approaches were able to identify 100 new proteins involved in the process of mitochondrial biogenesis [47].

C. Data Driven Integration for Knowledge Discovery

Overview—Integrative systems biology, one top-down method of gaining understanding, utilizes the vast resources of previously acquired high throughput data to improve our understanding biological processes and functional relationships. Data from these approaches, even analyzed in isolation, can provide a great deal of information about the disease or process of interest. These data also provide a wealth of information about the cell

line or type, the metabolic state of the organism, and many other factors. This aspect of these data is often under-utilized. An integrative analysis can take advantage of these properties to develop a systems level understanding of complex processes such as disease.

The goal of such integrative analysis is to accurately and effectively summarize the entirety of these diverse experimental results, with the goal of expanding established knowledge-based networks in the context of human disease using data driven identification of novel pathway members. This approach is based on integrating available experimental data into functional relationship networks based on how likely it is that any two given proteins are interacting (in a physical or regulatory way) to accomplish a specific biological process. The integration is performed by a Bayesian network that is trained based on existing biological knowledge and can automatically weigh the accuracy and coverage of each input data set.

This approach takes into account the accuracy and coverage of each individual dataset and the relevance of each source of data to the biological question being addressed. Intuitively, a gene expression dataset would be weighted highly for an analysis of cytokines and growth factors mediating inflammatory responses that exhibit robust transcriptional alterations, but these same datasets would not be highly weighted for an analysis of oxidative stress response because this process is not reflected well in gene expression data. Networks built for oxidative stress relationships would instead up-weight physical interaction studies, which capture the biological relationships important to this process well. A single expression dataset from human renal tissue may be highly informative for chemokine signaling, but not for oxidative stress driven inflammatory responses.

Huttenhower et al. [48] employed this strategy to build functional networks for human data which are made available through a web-based tool called HEFaiMp. The datasets used for this integration encompassed more than 30,000 experiments from more than 15,000 publications and included more than 27,000,000,000 data points. This system summarizes available data to enable biological researchers to quickly and effectively navigate the entirety of these data and generate hypotheses in a data-driven fashion.

Case Study—For this case study we have queried the functional relationship network in HEFaiMp for the gene BRCA1. BRCA1 encodes a tumor suppressor protein. The functional relationship network, shown in Figure 1, includes many cell cycle associated proteins such as MYC. Indeed the strongest functional relationships include genes such as MYC, JUNB, TP53, and E2F1. These relationships suggest that BRCA1 is involved in cell cycle progression. In addition to these well known genes, the list contains less characterized genes of interest. One such example is ZWINT which has a probability of a functional relationship with BRCA1 of 0.9793. ZWINT is a kinetochore protein which is known to interact with ZW10 but not known to have a functional relationship with BRCA1 [49-51]. This functional relationship prediction based on high throughput human data suggests that an investigation of the relationship between BRCA1 and ZWINT might be informative. Because this hypothesis comes not from published results but from the integrative analysis of high throughput data, this is an example of data driven hypothesis generation as opposed to the more frequently performed knowledge driven hypothesis generation. This process can be used to identify additional potential players in a process of interest, to find potential

functions for a protein discovered in a screen, or to find genes that have previously unknown functional relationships a specific protein of interest.

Bayesian Overview—The first step in the integration of heterogeneous genome-scale data has been to develop a common scale for measurements from diverse data types. Probabilities [52], or scores derived from probabilities [53-54], have often been used because each dataset can provide a likelihood which can then be integrated within a Bayesian framework. The Bayesian framework allows data from multiple sources to be integrated based on the strength of evidence from each data source. This strength of evidence can be expert annotated [52] or derived from known gold standards [55]. We and others have used integrative methods to discover novel biology [47-48, 56-58].

Intuitively the process begins with the development of a high quality gold standard. The gold standard contains pairs of genes that are known to have a relationship (gold standard positive relationships) and pairs of genes that are known to lack a relationship (gold standard negative relationships). The definition of “relationship” used in the creation of the gold standard determines the type of relationships that will be discovered during the integration process. If physical protein-protein interactions are used to develop the gold standard, these are the types of interactions that will be discovered. The work discussed here focuses on functional relationships, i.e. relationships where two genes are known to work together directly or indirectly to accomplish a certain biological task [59].

The next step is the collection of high throughput datasets and databases. Within each dataset, every pair of genes is evaluated and given a value. This value can be the correlation between genes, their mutual information, or simply the presence or absence of a relationship in a database. These values are placed into bins. Each of these bins is then evaluated for its proportion of gold standard positive and negative relationships. This proportion, when compared to the overall proportion of functional relationships, can be used to estimate the likelihood that unannotated gene pairs in that bin possess a functional relationship. Using the assumption of independence across datasets, the likelihood of a functional relationship can be calculated for genes outside of the gold standard. These probabilities of relationships can then be used to create a network in which edges represent the probability of a functional relationship (given the data) between genes.

Specifically a Bayesian integration process is performed. Myers et al. [55] describe the basic process in detail, but we will summarize it here. A discrete Bayesian network is constructed that relates evidence from different data types to the probability of a functional relationship. The critical elements of these networks are the structure and the conditional probability tables (CPTs). The structure determines relationships between evidence nodes. In this case, the network structure comes from earlier work by Troyanskaya et al. [52]. This network reflects relationships between evidence from different data types for the purpose of ensemble analysis and avoids double counting of evidence. The CPTs measure the reliability of each type of evidence. In this case, the CPTs were inferred using protein-protein relationships from the GO biological process ontology.

The CPTs, an example of which is shown in Table 1, can be constructed for each type of evidence using a gold standard. The gold standard should be carefully developed to capture the relationship of interest. In this case, the relationships of interest were functional relationships. A functional relationship is where a pair of genes work together to carry out a specific biological process.

For this work the Gene Ontology (GO) biological process ontology [35] provided a resource for development of the gold standard. Protein pairs whose most specific co-annotation occurs in GO terms of 300 total annotations or less were considered positives, while pairs whose most specific co-annotation occurs in GO terms of 1,000 total annotations or more were considered negatives.

For a hypothetical pair of genes i and j , it is possible to use the conditional probability table shown in Table 1 to calculate the probability of a functional relationship between the two genes $P(FR_{i,j})$ given an observed correlation state in the data, $D_{i,j}$. Mathematically we can write this statement as $P(FR_{i,j}|D_{i,j})$. Bayes' theorem says that

$$P(FR_{i,j}|D_{i,j}) = \frac{P(D_{i,j}|FR_{i,j}) P(FR_{i,j})}{P(D_{i,j})}.$$

Using the CPT, we can calculate the prior probability of a functional relationship, $P(FR_{i,j})$, from the proportion of positive and total relationships as

$$P(FR_{i,j}) = \frac{\text{Relationships}}{\text{Total in Standard}} = \frac{100}{500} = 0.2.$$

If we observe that our new gene pair is positively correlated, we can calculate that the probability of this state given that there is a functional relationship as

$$P(D_{i,j}|FR_{i,j}) = \frac{\text{Relationship and Positively Correlated}}{\text{Relationship}} = \frac{40}{100} = 0.4.$$

We can calculate the probability of observing a positively correlated gene pair as

$$P(D_{i,j}) = \frac{\text{Positively Correlated}}{\text{Total In Standard}} = \frac{160}{500} = 0.32.$$

The total probability of a functional relationship, given that gene i and j are positively correlated is then

$$P(FR_{i,j}|D_{i,j}) = \frac{P(D_{i,j}|FR_{i,j}) P(FR_{i,j})}{P(D_{i,j})} = \frac{0.4 \times 0.2}{0.32} = 0.25.$$

While this is the process for single evidence or independent evidence sources, the procedure can be more complicated when there are dependencies among evidence sources. Even in

those cases, these principles are the same and provide intuition for the Bayesian integration process.

The key components of this process are a high quality gold standard and useful data. For the Bayesian integration process it is not critical that all the data be pertinent, only that there be some subset of the data which is informative. A hypothetical gold standard of relationships based on phosphorylation would not allow for a successful integration when paired with datasets measuring mRNA levels. For this reason, all available high throughput data for the organism of interest are often employed.

The development of a gold standard is a difficult task. The ideal gold standard contains all positive relationships and negative relationships, but if all relationships are known there is little purpose to the integration of genome-scale data. In biological systems we are more often faced by the challenging situation of limited knowledge. In this situation, it is critical that the gold standard be of high quality, meaning that relationships listed as positive must be truly positive and relationships listed as negatives truly not be relationships. In the work we have been discussing, the gold standard was derived from the GO biological process ontology, but other methods have used sources such as cellular localization data [53-54, 60-61]. While these methods have generated promising results, the various gold standards can lead to very different findings [59]. Myers et al. [59] discuss the development of a gold standard that combines the GO biological process ontology with expert evaluation to create gold standards that address these issues.

2. Comparison of Databases and Data Driven Integration

There are a number of differing approaches to bringing systems level information to bear on biological problems. We discuss two database approaches, and compare these to data driven integration methods.

D. Databases of Interactions

The first approach centers on databases of interactions. There are numerous databases including the Database of Interacting Proteins (DIP) [37], the Mamalian Protein-Protein Interaction database (MIPS) [62], and the Biomolecular Interaction Network Database (BIND) [63]. Each of these databases collects various types of interactions from published work and datasets and makes them available to users through an interface. It is possible to use annotations in each of these to build networks of interactions, which some databases make available. These networks represent various types of interactions and thus an analysis of these networks requires a strong understanding of the underlying links. These databases provide information about what protein-protein links are currently known. By providing convenient access to known relationships, these databases can often facilitate knowledge driven hypothesis generation.

E. Functional Databases

In addition to databases of interactions, there are also functional databases such as DAVID [64-65]. These databases are often used to turn high throughput results into a systems level picture of what processes may be altered in the experimental condition. DAVID provides

tools to discover KEGG, GO, and other annotations that are overrepresented in a subset of genes of interest when compared to the experiment as a whole. DAVID is frequently updated and provides a user-friendly interface for researchers. DAVID uses current annotations to suggest functions of interest to researchers. In this way, DAVID allows for data driven hypothesis generation, but it is used to analyze results from individual high throughput experiments. It is also worth noting that DAVID highlights a process of interest but, in contrast to the Bayesian integration approach, does not provide specific predictions of functional relationships between genes.

Some functional databases combine this technique with databases of interactions to build pathway databases. These tools allow investigators to put new discoveries into the context of existing knowledge. Many use human curation to annotate gene-gene relationships that are then put into pathways. One example of such a tool is Pathway Studio from Ariadne Genomics [66]. In some cases these tools may mine medical literature to make apparent relationships which can be learned from published literature but which may not have been explicitly stated [67]. These tools employ prior biological knowledge from publications to place discoveries into a systems context and are also often used to make sense of findings from high throughput methods. The results provided are similar to results from DAVID (e.g. that a specific pathway is likely to be of interest) but these tools also provide a user-friendly interface to see how up and down-regulated genes fit into pathways of interest. These tools show gene-gene relationships that are known, but do not predict new relationships of interest.

F. Comparison of Data Driven Discovery and Databases

The Bayesian integration method we describe is a distinct approach for developing new insight into biological systems. This method uses high throughput datasets to predict new relationships of interest for experimental testing instead of putting experimental results into the context of currently known relationships. This strategy has previously been used to direct experimental biology in yeast [47, 68]. Furthermore these results can be used iteratively, where discoveries from one round of prediction are used to predict the involvement of genes that may have initially been missed.

The difference between the knowledge gained integrative and data driven approaches and the knowledge stored in databases like DIP, MIPS, and the relationships stored in pathway databases is that the integrative approach is data driven discovery while these databases perform information summary. This approach is distinct from functional analysis tools like DAVID because it provides new predictions of relationships between genes. Furthermore, these databases are limited to hypotheses addressed in experiments from which entries are derived. The data integration method is not constrained by this limitation and indeed can and often should include databases as raw input to the knowledge discovery process.

3. Context Specific Data Integration

Often a global prediction of functional relationships is not as useful as a context specific prediction [69]. Knowing that genes i and j have a functional relationship in humans could be informative, but knowing specifically that these genes interact in a context, C_m , would be

much more informative from an experimental design standpoint. In addition, certain data types are better able to capture specific biological processes of interest than others.

The HEFaMP approach discussed earlier is also capable of producing context specific networks. The contexts in this case, consist of 229 GO biological process categories thought, by biologists, to be sufficiently informative to direct experiments. The functional networks from the integration were used to predict six genes with a role in macroautophagy, five of which were then experimentally shown to play a role through siRNA knockdowns.

Returning to our earlier example of BRCA1, this approach allows us to discover functional relationships within a specific context. For BRCA1, an informative context might be “response to DNA damage stimulus.” Here, as before, we have queried for BRCA1, but this time we have also included the context “response to DNA damage stimulus.” Figure 2 shows the resulting network. Many similar players are still included, but this time the strongest links also include genes involved in DNA repair such as RAD51 and BRCA2. This indicates that the context specific method is able to discover functional relationships that are important for specific protein functions. The gene, ZWINT, is again among the predicted relationships, this time with a probability of 0.9999. This increased probability suggests that ZWINT may share a functional relationship with BRCA1 through its role in response to DNA damage. In the general functional relationship network, one of the top three evidence sources was from SMART shared protein domains (see the bottom panel of Figure 1). For the context specific network, the SMART shared protein domain evidence is no longer in the top ten evidence sources. This provides information that examining the shared protein domains between BRCA1 and ZWINT may not be informative for the functional relationship between these genes in this context. The increased detail from this context specific data derived hypothesis may facilitate molecular follow-up.

Intuitively this approach is similar to the Bayesian example discussed previously where the independence between the input datasets was assumed. The difference here is that, when CPTs are built for each dataset, they are built separately for each biological context. Before, we discussed the calculation of the probability of a functional relationship between genes i and j given a single dataset, $D_{i,j}$, which we wrote as $P(FR_{i,j}|D_{i,j})$. Here we take a similar approach, but we are interested the probability of a functional relationship given the data and a context, C_m . We can write this as

$$P(FR_{i,j}|D_{i,j}, C_m) = \frac{P(FR_{i,j}|C_m) P(D_{i,j}|FR_{i,j}, C_m)}{P(D_{i,j}|C_m)},$$

which is the same as the earlier equations except for the addition of the context. By defining our gold standard per context, we are able to calculate the probabilities as before.

This conceptually simple approach dramatically improves the resulting functional relationship predictions. In addition, this approach provides more useful information to a domain expert. Knowing that two genes share a functional relationship can be informative. Knowing that two genes share a functional relationship in the context of “response to DNA

damage stimulus” provides more information from an experimental design standpoint, particularly for multi-functional genes.

4. Tissue and Cell Lineage Specific Functional Networks

The complexity of multi-cellular organisms, which are made up of individual tissue types and cell lineages, combined with the substantially larger numbers and sizes of datasets in multicellular organisms present additional challenges to the application of integrated analysis methods to human biology. This is especially true in the context of human disease, where the understanding of molecular mechanisms and the development of personalized medicine will require the full utilization of thousands of genome-scale experimental results in specific biological and clinical contexts.

Tissue and cell lineage specific human functional networks would provide a potentially very powerful tool for researchers. In metazoans tissue specific contexts are likely to be critical because the functional relationships between molecular components can differ across tissues due to different expression [22], the presence or absence of binding partners [70], or tissue specific splice variants that affect the protein sequence [71]. Measuring tissue specific relationships experimentally can be difficult. For human tissues, ethical concerns often necessitate work in cell lines which may or may not recapitulate the biology of interest. In the case of cell-cell signaling processes it may even be difficult to determine if a cell line displays the phenotype of interest.

High quality computational predictions using integrative methods address many of these issues. Because these methods use data collected during high throughput experiments, they can take advantage of the diverse tissue types and cell lines represented. Additionally, tissue specific computational predictions could facilitate the choice of tissue or cell lines for a specific relationship of interest. Beyond helping tissue agnostic researchers discover useful experimental systems, these functional relationship networks can also provide tissue specific researchers with predicted relationships of interest. They could, for example, direct a researcher studying genes i and j to the podocyte, while directing a nephrologist to genes i and j .

Work in *C. elegans* [58] has shown that developing tissue specific functional relationship networks for metazoans is possible. The authors showed that microarray data from whole animals could yield high quality predictions of tissue specific expression with comparable accuracy to high throughput tissue specific expression studies. In humans, tissue specific integration is expected to pose additional challenges.

These obstacles arise due to the difficulty of performing tissue specific experiments in humans. Because tissue specific experiments are challenging and biological samples can be difficult to obtain, our understanding of tissue specific relationships is lacking. This presents a chicken-and-egg problem, as integrative approaches rely on standards developed from known relationships to predict undiscovered relationships.

Developing high quality functional networks for tissues as specific as a kidney podocyte is likely to require a two pronged approach. New methods capable of coping with sparser

standards are likely to be required. These methods will need to use new or additional sources of information to supplement or strengthen the gold standards used for integration. In addition to these new methods, new systems for building standards are likely to be required. These systems will need to facilitate input from domain experts. By using the collective wisdom of domain experts, it will be possible to build standards of the highest possible quality. Improvements in these areas can dramatically improve our understanding of the functional relationships between genes in tissues and cell types by addressing key challenges of the integrative biology approach.

Bibliography

1. Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J*. 2007; 2(1):41–9. [PubMed: 17225249]
2. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clin Chem*. 2009; 55(4):641–58. [PubMed: 19246620]
3. Whitfield ML, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*. 2002; 13(6):1977–2000. [PubMed: 12058064]
4. Hegde P, et al. Identification of tumor markers in models of human colorectal cancer using a 19,200-element complementary DNA microarray. *Cancer Res*. 2001; 61(21):7792–7. [PubMed: 11691794]
5. Lock C, et al. Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nat Med*. 2002; 8(5):500–8. [PubMed: 11984595]
6. Stoevesandt O, Taussig MJ, He M. Protein microarrays: high-throughput tools for proteomics. *Expert Rev Proteomics*. 2009; 6(2):145–57. [PubMed: 19385942]
7. Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet*. 2009; 10(9):617–27. [PubMed: 19687803]
8. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–78. [PubMed: 17554300]
9. Schymick JC, et al. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*. 2007; 6(4):322–8. [PubMed: 17362836]
10. Shriner D, et al. Problems with Genome-Wide Association Studies. *Science*. 2007; 316(5833):1840–1841-1840–1841. [PubMed: 17600199]
11. Williams SM, et al. Problems with Genome-Wide Association Studies. *Science*. 2007; 316(5833):1841–1842-1841–1842.
12. Jakobsdottir J, et al. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet*. 2009; 5(2):e1000337. [PubMed: 19197355]
13. Link AJ, et al. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*. 1999; 17(7):676–682.
14. Opiteck GJ, et al. Comprehensive on-line LC/LC/MS of proteins. *Analytical Chemistry*. 1997; 69(8):1518–1524. [PubMed: 9109352]
15. Kittler R, et al. Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nat Cell Biol*. 2007; 9(12):1401–12. [PubMed: 17994010]
16. Krishnan MN, et al. RNA interference screen for human genes associated with West Nile virus infection. *Nature*. 2008; 455(7210):242–5. [PubMed: 18690214]
17. Ozsolak F, et al. High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*. 2007; 25(2):244–8.
18. Velculescu VE, et al. Serial Analysis of Gene-Expression. *Science*. 1995; 270(5235):484–487. [PubMed: 7570003]

19. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; 10(1):57–63.
20. Cloonan N, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*. 2008; 5(7):613–619. [PubMed: 18516046]
21. Morin RD, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*. 2008; 45(1):81. [PubMed: 18611170]
22. Su AI, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(16):6062–6067. [PubMed: 15075390]
23. Perou CM, et al. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747–52. [PubMed: 10963602]
24. Oliphant A, et al. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques. Suppl.* 2002;56–58. 60–61. 60-61-56-58. [PubMed: 12083399]
25. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nat Genet*. 2006; 38(6):659–62. [PubMed: 16715099]
26. Spencer CC, et al. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*. 2009; 5(5):e1000477. [PubMed: 19492015]
27. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*. 2008; 24(3):133–41. [PubMed: 18262675]
28. Nagalakshmi U, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008; 320(5881):1344–1349. [PubMed: 18451266]
29. Wilhelm BT, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008; 453(7199):1239–U39. [PubMed: 18488015]
30. Mortazavi A, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008; 5(7):621–628. [PubMed: 18516045]
31. Marioni JC, et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*. 2008; 18(9):1509–1517. [PubMed: 18550803]
32. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30(1):207–10. [PubMed: 11752295]
33. Desiere F, et al. The PeptideAtlas project. *Nucleic Acids Res*. 2006; 34(Database issue):D655–8. [PubMed: 16381952]
34. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000; 28(1):27–30. [PubMed: 10592173]
35. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; 25(1):25–9. [PubMed: 10802651]
36. Klein TE, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J*. 2001; 1(3):167–70. [PubMed: 11908751]
37. Xenarios I, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*. 2002; 30(1):303–5. [PubMed: 11752321]
38. Bader G, Betel D, Hogue C. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*. 2003; 31:248–250. [PubMed: 12519993]
39. Snel B, et al. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*. 2000; 28(18):3442–4. [PubMed: 10982861]
40. Kitano H. Systems biology: a brief overview. *Science*. 2002; 295(5560):1662–4. [PubMed: 11872829]
41. Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol*. 2007; 15(1):45–50. [PubMed: 17113776]
42. Di Ventura B, et al. From in vivo to in silico biology and back. *Nature*. 2006; 443(7111):527–33. [PubMed: 17024084]
43. Zeeberg BR, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*. 2003; 4(4):R28. [PubMed: 12702209]

44. Zeeberg BR, et al. High-Throughput GoMiner, an ‘industrial-strength’ integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*. 2005; 6:168. [PubMed: 15998470]
45. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102(43):15545–50. [PubMed: 16199517]
46. Orton RJ, et al. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. *Biochem J*. 2005; 392(Pt 2):249–61. [PubMed: 16293107]
47. Hess DC, et al. Computationally Driven, Quantitative Experiments Discover Genes Required for Mitochondrial Biogenesis. *PLoS Genet*. 2009; 5(3):e1000407. [PubMed: 19300474]
48. Huttenhower C, et al. Exploring the human genome with functional maps. *Genome Research*. 2009; 19(6):1093–106. [PubMed: 19246570]
49. Obuse C, et al. A conserved Mis12 centromere complex is linked to heterochromatic HP1 and outer kinetochore protein Zwint-1. *Nature Cell Biology*. 2004; 6(11):1135–U37. [PubMed: 15502821]
50. Starr DA, et al. HZWint-1, a novel human kinetochore component that interacts with HZW10. *Journal of Cell Science*. 2000; 113(11):1939–1950. [PubMed: 10806105]
51. Wang HM, et al. Human Zwint-1 specifies localization of zeste white 10 to kinetochores and is essential for mitotic checkpoint signaling. *Journal of Biological Chemistry*. 2004; 279(52):54590–54598. [PubMed: 15485811]
52. Troyanskaya OG, et al. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100(14):8348–8353. [PubMed: 12826619]
53. Jansen R, et al. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*. 2003; 302(5644):449–453. [PubMed: 14564010]
54. Lee I, et al. A Probabilistic Functional Network of Yeast Genes. *Science*. 2004; 306(5701):1555–1558. [PubMed: 15567862]
55. Myers C, et al. Discovery of biological networks from diverse functional genomic data. *Genome Biology*. 2005; 6(13):R114–R114. [PubMed: 16420673]
56. Covert MW, et al. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*. 2004; 429(6987):92–96. [PubMed: 15129285]
57. Zhu J, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*. 2008; 40(7):854–861. [PubMed: 18552845]
58. Chikina MD, et al. Global prediction of tissue-specific gene expression and context-dependent gene networks in *Caenorhabditis elegans*. *PLoS Comput Biol*. 2009; 5(6):e1000417. [PubMed: 19543383]
59. Myers CL, et al. Finding function: evaluation methods for functional genomic data. *BMC Genomics*. 2006; 7:187. [PubMed: 16869964]
60. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol*. 2004; 7(5):535–45. [PubMed: 15451510]
61. Patil A, Nakamura H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*. 2005; 6(1):100. [PubMed: 15833142]
62. Pagel P, et al. The MIPS mammalian protein-protein interaction database. *Bioinformatics*. 2005; 21(6):832–4. [PubMed: 15531608]
63. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*. 2003; 31(1):248–50. [PubMed: 12519993]
64. Dennis G Jr, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 2003; 4(5):P3. [PubMed: 12734009]
65. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*. 2009; 4(1):44–57. [PubMed: 19131956]
66. Nikitin A, et al. Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics*. 2003; 19(16):2155–2157. [PubMed: 14594725]

67. Chen H, Sharp B. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*. 2004; 5(1):147. [PubMed: 15473905]
68. Hibbs MA, et al. Directing Experimental Biology: A Case Study in Mitochondrial Biogenesis. *PLoS Comput Biol*. 2009; 5(3):e1000322. [PubMed: 19300515]
69. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. *Bioinformatics*. 2007; 23(17):2322–2330. [PubMed: 17599939]
70. Kabe Y, et al. The role of human MBF1 as a transcriptional coactivator. *Journal of Biological Chemistry*. 1999; 274(48):34196–34202. [PubMed: 10567391]
71. Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*. 2002; 30(17):3754–3766. [PubMed: 12202761]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. The top panel indicates the search used to generate this network. The middle panel shows the functional relationship network from HEFalmP for BRCA1. Red edges indicate high confidence links and green edges indicate lower confidence links. The bottom panel shows the top sources of evidence for the link between BRCA1 and ZWINT. The top sources include a dataset of pre-mRNA splicing indicating a potential mechanism of interest.

Figure 2.

The context specific functional relationship network from HEFaiMp for BRCA1 in the context of “Response to DNA damage.” Red edges indicate high confidence links and green edges indicate lower confidence links.

Table 1

A conditional probability table for a hypothetical data source.

	Negatively Correlated	Uncorrelated	Positively Correlated	Total (Relationships)
Relationship	10	50	40	100
No Relationship	170	110	120	400
Total (Correlation)	180	160	160	

Gene pairs in the data source can have one of three correlation values. The numbers of true and false relationships are constructed from a gold standard. These values can be used to predict the probability of a functional relationship for a gene pair with unknown relationship status.