OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Perspective

From Bytes to Bedside: Data Integration and Computational Biology for Translational Cancer Research

Jomol P. Mathew*, Barry S. Taylor, Gary D. Bader, Saiju Pyarajan, Marco Antoniotti, Arul M. Chinnaiyan, Chris Sander, Steven J. Burakoff, Bud Mishra

Major advances in genome science and molecular technologies provide new opportunities at the interface between basic biological research and medical practice. The unprecedented completeness, accuracy, and volume of genomic and molecular data necessitate a new kind of computational biology for translational research. Key challenges are standardization of data capture and communication, organization of easily accessible repositories, and algorithms for integrated analysis based on heterogeneous sources of information. Also required are new ways of using complementary clinical and biological data, such as computational methods for predicting disease phenotype from molecular and genetic profiling. New combined experimental and computational methods hold the promise of more accurate diagnosis and prognosis as well as more effective prevention and therapy.

Introduction

Over the last two decades, our knowledge of cancer and its causes has increased greatly. However, we still have few examples of cures. This underscores the need for a clearer understanding of the alterations in the biological circuitry that lead to tumor development and growth. Sequencing of the human genome and biotechnological advances have led to the generation of large volumes of genome-scale data. Combining this genome-scale molecular data with clinical information provides new opportunities to discover how perturbations in biological processes lead to disease. This knowledge can be used to improve disease diagnosis, prognosis, prevention, and therapy. However, the large scale and diversity of both experimental and clinical data necessitate that they be well-organized and computationally accessible to research scientists for analysis and interpretation. This review focuses on the challenges and opportunities to combine clinical and genome-scale molecular data, using computational approaches, to better understand cancer biology and to translate this knowledge into improved disease prevention and therapy.

Translational cancer research to improve disease prevention and therapy. Cancer is a complex disease, involving multiple and specific changes at the DNA level that can be inherited or induced by environmental factors. There are many different types and subtypes of cancer marked by specific sets of molecular changes. Most of our current cancer treatment efforts are focused on surgery for a curative treatment and radiation and/or toxic drugs (chemotherapy) to induce remissions. Candidates for successful cancer therapy with surgery are few, and radiation and chemotherapy suffer from lack of target specificity, leading to serious side effects. Identifying cancer-specific molecular changes and discovering how they can be used to increase therapeutic specificity will lead to higher success rates and fewer side effects.

Translational research seeks to identify and understand the cause and effect of cancer-specific molecular defects and to translate this "bench" knowledge to the clinic to improve disease prevention and therapy. Examples of research questions include, from a clinical perspective: what are the molecular subtypes of cancer? What reliable molecular

Editor: Johanna McEntyre, National Center for Biotechnology Information, United States of America

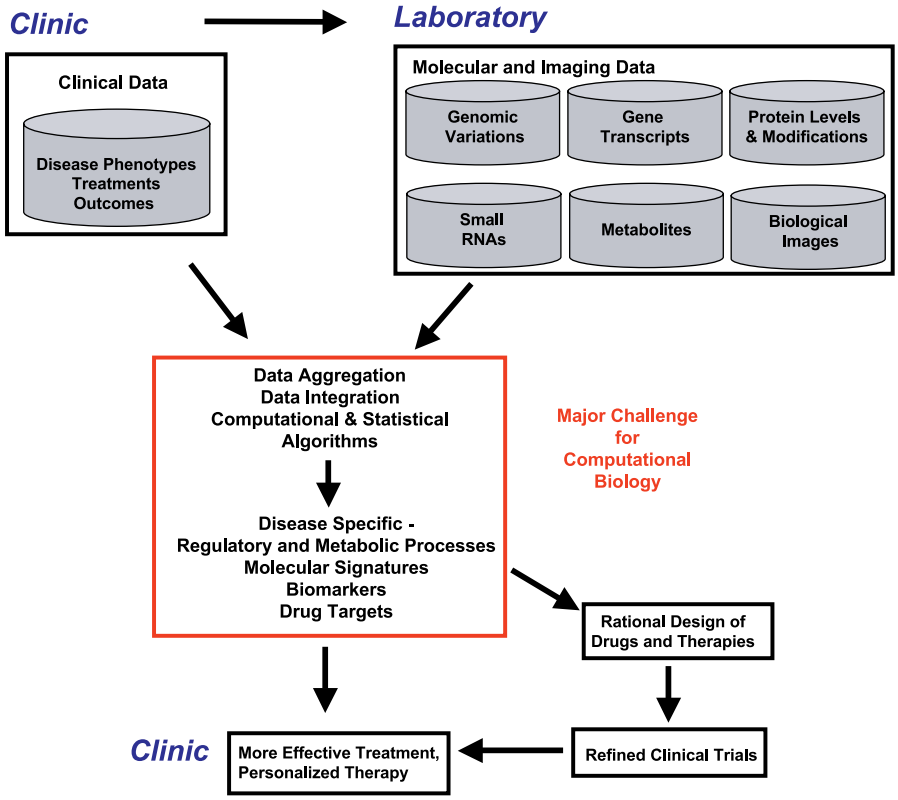
Citation: Mathew JP, Taylor BS, Bader GD, Pyarajan S, Antoniotti M, et al. (2007) From bytes to bedside: Data integration and computational biology for translational cancer research. PLoS Comput Biol 3(2): e12. doi:10.1371/journal.pcbi.0030012

Copyright: © 2007 Mathew et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: caBIG, Cancer Biomedical Informatics Grid; CGH, comparative genomic hybridization; FISH, fluorescent in situ hybridization; mRNA, messenger RNA; miRNA, microRNA; MITF, micropthalmia-associated transcription factor; NCI, National Cancer Institute; SNP, single nucleotide polymorphism

Jomol P. Mathew is with the Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America. Barry S. Taylor is with the Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York, United States of America, and the Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America. Gary D. Bader is with Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada. Saiju Pyarajan and Steven J. Burakoff are with the Skirball Institute of Biomolecular Medicine, New York, New York, United States of America, New York University Cancer Institute and the Department of Pathology at the New York University School of Medicine, New York, New York, United States of America. Marco Antoniotti is with the Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy. Arul M. Chinnaiyan is with the Department of Pathology (Urology), and the Bioinformatics Program at the University of Michigan Medical School, Ann Arbor, Michigan, United States of America. Chris Sander is with the Computational Biology Center at Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America. Bud Mishra is with the Courant Institute at New York University and the Department of Cell Biology at New York University School of Medicine, New York, New York, United States of America. At the time of preparation of this manuscript, Jomol Mathew was with the Department of Environmental Medicine, New York University School of Medicine, New York, New York, United States of America; Barry Taylor was at the Department of Pathology and Bioinformatics Program, University of Michigan Medical School, Ann Arbor, Michigan, United States of America; Gary Bader was at the Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America; and Marco Antoniotti was at the New York University Bioinformatics Group, Courant Institute, New York University, New York, New York, United States of America.

* To whom correspondence should be addressed. E-mail: Jomol_Mathew@dfci.harvard.edu



doi:10.1371/journal.pcbi.0030012.g001

Figure 1. Data Integration for Translational Cancer Research

Archival of clinical and molecular data in easily retrievable standardized formats, aggregation, integration, and data analysis will provide opportunities for the next-generation biomedical discoveries that can impact cancer research and treatment.

markers are available for early cancer detection (diagnostic) and for predicting the course of disease (prognostic)? How do we find better drugs and optimize therapy (development of more specific drugs with lower toxicity) to suit an individual patient's molecular profile? From a molecular biology perspective: can we accurately predict vulnerable point(s) in molecular pathways that are potential therapeutic targets? What specific drug or drug combination can target these vulnerable points in the pathway? Can genotype and pathway information be combined to predict the effect of a mutation on disease or therapy?

Advances in our understanding of cancer-specific molecular defects have led to improved cancer treatments. For example, the protein kinase inhibitor imatinib (Gleevec) was designed to treat chronic myelogenous leukemia (CML) based on knowledge of the causative molecular defect—translocation and dys-regulated BCR–ABL kinase. Protein kinase inhibitors such as gefitinib (Iressa) and erlotinib (Tarceva) are showing therapeutic promise by targeting known molecular abnormalities of non-small cell lung cancer (NSCLC). Similarly, antibody therapies such as Rituximab (Rituxan), an anti-CD20 monoclonal antibody for non-Hodgkin lymphoma; Cetuximab (Erbitux), an epidermal growth factor receptor (EGFR)-binding antibody for colorectal and head and neck cancer; Trastuzumab (Herceptin), a monoclonal antibody that allows targeted therapy in HER2 positive breast cancer; and Bevacizumab (Avastin), a recombinant humanized antibody against vascular endothelial growth factor (VEGF) for metastatic colorectal cancer are promising.

While these treatments based on molecular knowledge of the cancer show promise, major challenges remain. For instance, development of compensatory mutations induces resistance to Gleevec and limits its use, while humanization and effective delivery of antibodies is difficult [1]. Furthermore, the discovery and development of a new and effective drug can cost US$0.8–US$1.7 billion [2]. A new drug entering Phase I testing, where the drug is initially introduced to human subjects, is estimated to have only an 8% chance of reaching the market [2]. Failures can largely be attributed to poor target selection or poor candidate drug selection, leading to low drug effectiveness or toxicity. Development of safe and effective therapies, such as small molecule protein kinase inhibitors, at a reduced cost, requires better understanding of therapeutic interaction of the inhibitor with a range of targets and their effect on diverse cellular processes [3].

Cancer cells can now be profiled on a genome scale using new experimental techniques. We thus have an unprecedented opportunity to comprehensively study cancer-specific molecular processes. This study requires computational tools to handle the large volume and diversity of available information. Collection, standard organization, aggregation, storage, integration, and analysis of diverse genome-scale molecular data along with patient data collected in the clinic will broaden our understanding of how cancer-

Table 1. Genomic Variation Repositories

Type of Data	Public Data Source
Cytogenetic and array-CGH databases	Progenetix ([78]; http://www.progenetix.net), NCI and NCBI's SKY/M-FISH (spectral karyotyping/multiplex-FISH) and CGH Database ([79]; http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi
Chromosome aberrations databases	Mitelman database of chromosome aberrations [80] http://cgap.nci.nih.gov/Chromosomes/Mitelman, recurrent chromosome aberrations database http://cgap.nci.nih.gov/Chromosomes/RecurrentAberrations, and the NCBI Map viewer database
Human genomic polymorphism databases	TSC (The SNP Consortium) database http://snp.cshl.org of nearly 1.8 million SNPs, Database of Genomic Variants (DGV) from The Centre for Applied Genomic (TCAG) http://projects.tcag.ca/variation/index.html containing large-scale variations such as copy number polymorphisms (CNPs). dbSNP database of single base nucleotide substitutions and short deletion and insertion polymorphisms from NHGRI (National Human Genome Research Institute) and NCBI http://www.ncbi.nlm.nih.gov/SNP/index.html.
Methylation database	MethDB ([81,82] http://www.methdb.net for DNA methylation data
Somatic mutation in cancer	Catalogue of somatic mutations in cancer (COSMIC) [83] (http://www.sanger.ac.uk/genetics/CGP/cosmic)

doi:10.1371/journal.pcbi.0030012.t001

specific molecular defects affect clinical outcome and will lead to improved disease prevention and therapy (Figure 1).

Genome-Scale Molecular Data for Cancer Research

Genomic variation. Cancer is a genetic disease involving point mutations, translocations, segmental amplifications, and deletions in the genome that alter specific vulnerable molecular points in cellular regulatory pathways. Analysis of chromosomal changes by fluorescent in situ hybridization (FISH)-based cytogenetic approaches including comparative genomic hybridization (CGH), spectral karyotyping (SKY), and multiplex-FISH (M-FISH) [4] have led to the characterization of many cancer-associated chromosomal abnormalities. Microarray techniques, e.g., array-CGH or matrix-CGH, have become available to map regions of DNA sequence from the cancer tissue that are amplified or reduced compared to normal tissue [5]. Array-based technologies also allow genome-wide measurement of single nucleotide polymorphisms (SNPs) [6]. The international HapMap project has identified millions of SNPs in different populations. The data has been processed into haplotypes, sets of co-occurring SNPs, and tag SNPs (SNPs that distinguish a set of common haplotypes) that can be used to reduce the complexity of gene association studies (http://www.hapmap.org) [7].

Epigenetic changes such as DNA methylation, histone modification, and RNA silencing are involved in regulating many cellular processes, including development, via gene silencing (chromatin structure and transcription regulation) and genetic imprinting. Specific DNA methylation alterations have been identified in various neoplasms. For example, aberrant promoter methylation associated with transcriptional downregulation of tumor suppressor genes has been found in basal cell carcinoma (BCC), cutaneous squamous cell carcinoma (SCC), melanoma, and cutaneous lymphoma [8]. Though not exhaustive, Table 1 gives a list of publicly available cancer-relevant large genomic variation repositories.

Several projects attempt to comprehensively study genomic variation in cancer. The Cancer Genome Atlas (http://cancergenome.nih.gov/index.asp) and the Sanger Institute's Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP) aim to identify mostly somatic mutations in common tumor types using next generation DNA sequencing technology.

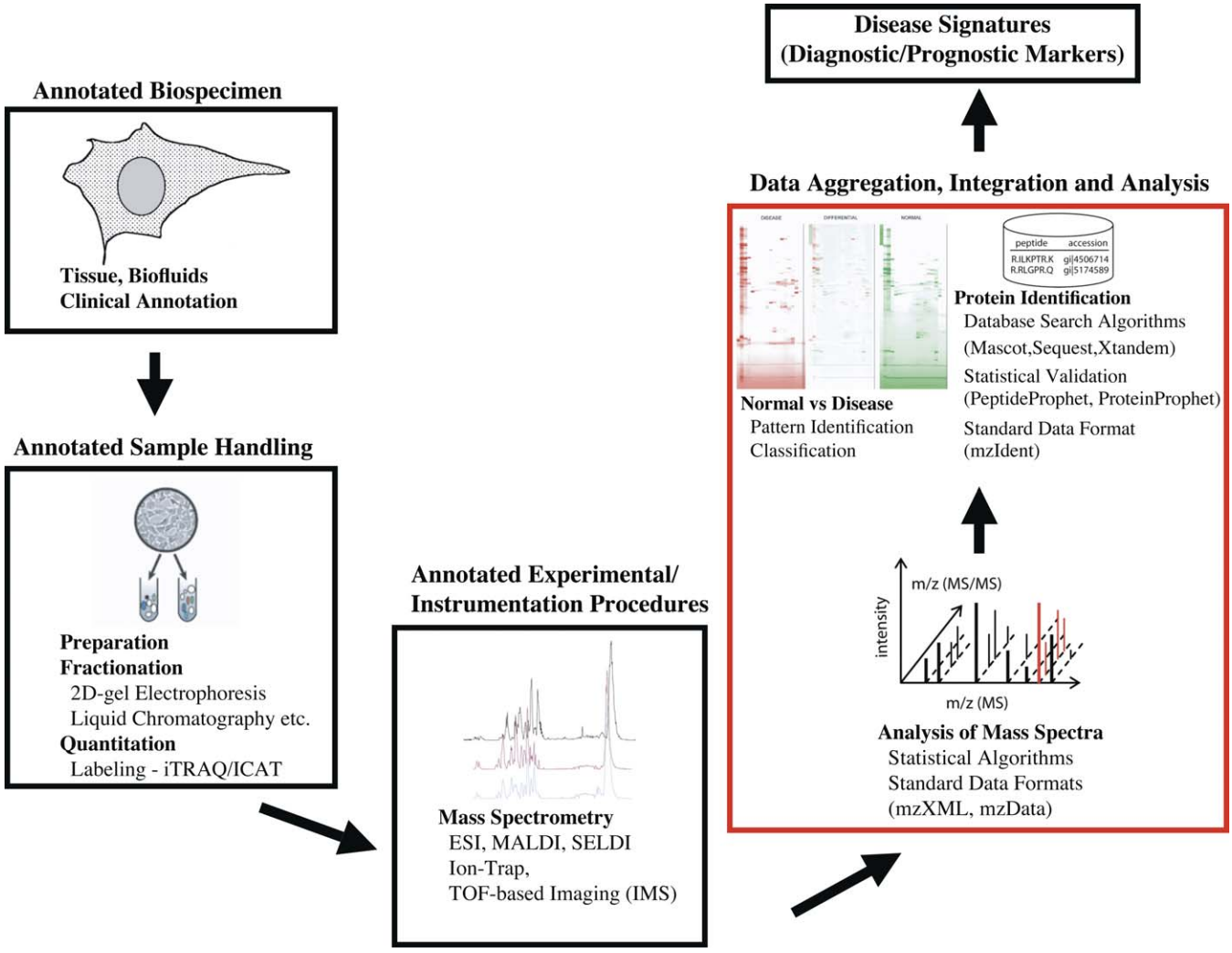
Gene transcript profiles. Global gene expression profiling with DNA microarrays [9–11] has furthered our understanding of the regulation of biological processes and has become an indispensable tool in the study and classification of human tumors. Semiquantitative profiles of gene expression have been measured for many cancer types

Table 2. Gene Expression Repositories

Public Data Source	Data	Description
Gene Expression Omnibus (GEO) [84]	4,439 experiments >100 organisms	Public data deposition and query platform for expression data[a]
ArrayExpress [85]	1,684 experiments; 1,170 arrays	Public data repository for well-annotated microarray data[a]
Stanford Microarray Database (SMD) [86]	1,1227 experiments; 43 organisms	Public data deposition and query platform for expression data[a]
Oncomine [68]	209 studies; 14,177 microarrays	Comprehensive meta-analysis and data mining platform specific to cancer biology

[a]MIAME-compliant and support the MAGE-ML submission format.

doi:10.1371/journal.pcbi.0030012.t002



doi:10.1371/journal.pcbi.0030012.g002

Figure 2. Protein Profiling for Cancer Diagnosis and Prognosis

Generation of protein profiles using mass spectrometry is an example of an experimental technique that produces massive amounts of data that is difficult to interpret without computational and statistical algorithms. For instance, comparison of disease versus control sample profiles can lead to identification of disease-specific protein expression signatures, which could be used as diagnostic or prognostic markers. Aggregation of such data from multiple sources and pooled analysis requires proper annotation of sample source, sample handling, and experiment information.

and subtypes [12,13]. Through unbiased comparative analysis of these profiles, a subset of genes can be found that correlate with tumor phenotype and can serve as diagnostic and prognostic markers of disease. Disease-specific regulatory programs can be studied using techniques such as chromatin immunoprecipitation (ChIP) of tumor biopsies [14]. Several large public compendiums of gene expression data generated from diverse experimental methods also exist, some examples of which are presented in Table 2.

Not all available cancer microarray data are generated from gold-standard tissues or primary cell culture. Often, immortalized cell line models of neoplastic disease are studied because they are easier to access than histopathologically characterized human tumors [15,16]. However, in vitro extended passaging of cell lines could lead to accumulation of alterations and yield less representative expression profiles and less stable disease phenotypes [17]. Additionally, the in vivo microenvironment affects tumor–host interactions and causes variations in gene expression and pathways. These changes make disease-specific expression patterns difficult to infer from cell line data alone. Nevertheless, mining and analysis of the large amount of available cell line gene expression data promises new insights into disease conditions.

Protein levels and modifications. Mass spectrometric instruments and protein chip technology allow large-scale analysis of proteins, their quantitative expression, interactions, post-translational modifications, and localization [18–25]. Proteomic profiling of clinical samples ranging from tissues to biofluids (e.g., urine, sera, plasma, whole blood, cerebrospinal fluid, and saliva) will help assess disease development and progression, generate diagnostic and prognostic disease markers, and predict patient response to intervention. This information can be used to identify regulatory networks and activated signaling events in biological pathways, and can help characterize pathologically benign and tumor tissue samples (Figure 2). Public proteomics repositories that collect this data are now available, and

Table 3. Proteomics Repositories

Public Data Source	Data	Description
Plasma Proteome Project (HUPO-PPP) [87]	9,504 high-confidence identifications	Identifications in the blood plasma proteome
Global Proteome Machine (GPM) [88]	3,411,277 proteins; 17,005,857 peptides; 16 organismal proteomes	Laboratory-submitted peptide mass spectra
Proteomics Identifications Database (PRIDE) [89]	215,621 proteins; 770,171 peptides; 160,788 spectra	Global repository for peptide/protein identification data
PeptideAtlas [90]	338,200 MS/MS spectra; 36,653 peptides (Human)	Compendium incorporating annotation of human genomic sequence with observed expressed peptides
Human Protein Atlas (HPR) [91]	1514 antibodies; 1,238,760 images	Expression and localization of proteins in normal human tissue and cancer cells

doi:10.1371/journal.pcbi.0030012.t003

relevant examples are listed in Table 3. Accurate protein identification involving processing and identification of mass spectral peaks, peptide sequencing, search algorithms, and statistical validation of correct assignment of peptides and proteins is challenging, often confounded by splice variants or other protein isoforms [26–30].

Small RNAs. Small RNAs add an additional layer of complexity to gene regulation [31]. Initially discovered in plants and *C. elegans*, at least four subfamilies exist: microRNAs (miRNAs), short interfering RNAs (siRNAs), tiny noncoding RNAs (tncRNAs), and small modulatory RNAs (smRNAs). miRNAs are small, ~21–24, nucleotide noncoding endogenous RNAs involved in translational repression and messenger RNA (mRNA) cleavage, which can potently downregulate translation of specific mRNAs by targeted 3′-UTR binding. An increasing number of miRNAs have been implicated in disease. For instance, a small cluster of miRNA genes on Chromosome 13, c13orf25, appears to be involved in B cell lymphoma [32]. Cell cycle regulation has been shown to involve a miRNA regulatory circuit where *c-Myc,* a known proto-oncogene upregulating E2F1 (a cell cycle regulator), also regulates the expression of six miRNAs on Chromosome 13, two of which have been shown to downregulate E2F1 [33]. In the case of acute lymphoblastic leukemia (ALL), in studies by Lu et al. (2005), miRNA transcriptional profiles have been shown to better classify tumors by type and developmental stage than mRNA profiles [34].

mRNA targets for several hundred miRNAs that are expressed in human have been computationally predicted, though few targets have been experimentally confirmed ([35,36], http://www.microrna.org). The lack of perfect base-pair complementarity between the miRNA sequence and its target and the short length of the miRNAs make accurate prediction of miRNA genes and targets difficult.

Pathways. Pathway information is vital for understanding biological processes and how they are disrupted or reprogrammed in disease. However, collecting complex pathway information in a usable form from diverse and heterogeneous sources, including more than 220 pathway databases (http://pathguide.org), is a major challenge [37]. A number of pathway database efforts seek to ameliorate the situation by making pathway data more accessible for computational analysis (Table 4). For instance, Memorial Sloan-Kettering Cancer Center and the Institute of Bioinformatics have collaboratively developed ten large cancer-focused signaling pathways (http://cancer.cellmap.org).

Metabolite profiles. Metabolomics involves measurement of metabolite concentrations and fluxes in cells and tissues [38]. Such measurements provide insight into the response of biological systems to genetic and environmental influences. Metabolites provide important markers for disease state and

Table 4. Human-Focused Pathway Repositories

Public Data Source	Data	Description
Kyoto Encyclopedia of Genes and Genomes (KEGG) [92]; HumanCyc [93]	Metabolic pathways	Biochemical reactions geared toward metabolite conversions
Reactome [94], PANTHER [95], BioCarta http://biocarta.com, INOH http://inoh.org, The cancer cell map http://cancer.cellmap.org	Signaling pathways	Protein–protein interactions and post-translational modifications
Biomolecular Interaction Network Database (BIND) [96], Human Protein Reference Database (HPRD) [97]	Molecular interactions/proteomics	Molecular interactions, such as protein–protein and protein–DNA, without much detail, but with extensive coverage [98]
Jasper, Transfac [99] (commercial)	Gene regulation networks	Links transcription factors and the genes they regulate
BioGRID [100]	Protein and Genetic interactions	Interactions between genes such as epistasis or synthetic lethality

doi:10.1371/journal.pcbi.0030012.t004

Table 5. Data Standardization Efforts in Biomedical Research

Type of Data	Standard	Developing Agency	Description
Gene annotation	Gene Ontology	GO Consortium	Controlled vocabulary describing genes and gene products in terms of their associated biological processes, cellular components, and molecular functions
Gene expression	Minimal Information About a Microarray Experiment (MIAME)	Microarray Gene Expression Data Society (MGED)	Checklist of information to provide as metadata for any microarray experiment http://www.mged.org/Workgroups/MIAME/miame.html
	MAGE-OM and MAGE-ML [101]	MGED	Data exchange model and XML data exchange format http://www.mged.org/Workgroups/MAGE/mage.html
Proteomics	PSI-OM, PSI-ML, Minimal Information About a Proteomic Experiment (MIAPE)	Proteomic Standard Initiative (PSI) of the Human Proteome Organization (HUPO)	Data exchange model, XML data exchange format, and checklist of information to provide metadata for any proteomic experiment
	mzData and mzIdent	PSI	XML format for exchange of mass spectrometry spectra (mzData) and protein identifications (mzIdent)
Pathways and networks	Biological Pathway Exchange (BioPAX)	The BioPAX Workgroup	XML representation of pathway information for metabolic pathways, molecular interactions, signaling and genetic regulatory pathways http://www.biopax.org
	CellML	CellML Project, Bioengineering Institute, The University of Auckland, New Zealand	XML-based standard for the storage and exchange of mathematical models of pathways http://www.cellml.org
	Systems Biology Markup Language (SBML)	The SBML Team	XML-based format for describing qualitative and quantitative models of biochemical networks http://www.sbml.org/index.psp
	PSI-MI (Molecular Interaction) [102]	PSI	XML standard for representing molecular interactions http://psidev.sourceforge.net
Biological images	The Open Microscopy Environment (OME)	Initiated in the labs of Jason Swedlow (University of Dundee) and Peter Sorger (MIT)	Open source informatics framework for microscopic imaging. The proposed OME data model and XML format, though currently focused on fluorescence microscopy, are extensible to any type of microscropic image http://www.openmicroscropy.org
Clinical data	Health Level Seven (HL7)	HL7 organization	Standard format and content for clinical and administrative healthcare data exchange between applications http://www.hl7.org
	CDISC	Clinical Data Interchange Standards Consortium (CDISC)	Standards for acquiring and regulatory reporting of clinical trial data primarily among pharmaceutical companies http://www.cdisc.org

doi:10.1371/journal.pcbi.0030012.t005

the pathways underlying drug metabolism. Metabolomic profiles can be used for classifying disease by type and stage, for prognosis, and for testing the effectiveness of therapeutics using statistical learning methods. For example, high-resolution magic angle spinning NMR (HRMAS–NMR), which can quantitatively identify a range of metabolites while leaving the tissue sample intact for further studies, has been used to profile and classify prostate tumors [39] and to detect drug efficacy in liposarcoma [40].

Biological images. Advances in optics, digital detectors, and automation have significantly improved biological imaging technology and have led to a large increase of quantitative information extracted from digital images. Fluorescent and confocal microscopy, and whole-body imaging of model organisms [41,42] can be used to test specific hypotheses of cellular function and disease. Deep tissue–penetrating infrared light and various alterations of two-photon laser scanning microscopy (2PLSM) have been successfully used to reveal the dynamic nature and spatio–temporal aspects of hematopoietic tissue [43], organ development [44], and neurobiology [45]. Fluorescent proteins and photo bleaching techniques enable visualization of protein localization, protein–protein interactions, and protein fate in vivo [46]. In clinical settings, particularly for solid tumors, better resolution and higher contrast dyes have allowed the use of magnetic resonance imaging (MRI), computed tomography (CT) scan, positron emission tomography (PET) scan, and ultrasound for diagnosis and tracking disease progression. As imaging data is highly context-specific, data aggregation from multiple sources is possible only if sufficient metadata on samples, microscope, and data derivation algorithms are available.

Clinical data. Clinical data is information about patients that is collected using surveys, during doctors' office visits, through administration of standard treatment procedures, or during clinical trials. Typical cancer clinical trials are conducted to determine the safety and efficacy of a drug in humans and depend on detailed patient information for accurate interpretation of results. Clinical trials range from pilot studies for feasibility assessment of the trial to more involved Phase I to IV trials [47]. Patient data collected during clinical trials includes family history (for example, if mother or sister had breast cancer), habits (for example, smoking/drinking), concomitant medications, alternate therapies, baseline characteristics preceding treatment, diagnostic parameters and clinical staging, treatment and procedural details, adverse events (toxicity), and clinical endpoints (for example disease recurrence or survival). For example, in solid tumors, tumor measurement is done at prespecified intervals

Box 1. Oncomine: A Case Study in Microarray Data Aggregation and Analysis

The Oncomine cancer microarray database is an integrated meta-analysis platform that overcame diverse data integration and normalization challenges to enable comprehensive analysis of complex multistudy disease datasets [52,53,68]. A software pipeline was developed to parse gene expression data, raw or log-transformed, from native formats into numeric matrices of reporter rows and sample columns excluding study-specific normalization. Comprehensive mapping of probe identifiers (IDs) from oligonucleotide arrays or IMAGE clone IDs from cDNA arrays to a common Unigene build, Genbank accession numbers, and other commonly used database identifiers that link to gene annotation was also critical. Samples were renamed and reassigned using NCI nomenclature for consistency across studies. Lack of adherence of the individual datasets to any common standards, such as MIAME, complicated the data aggregation process. Another major hurdle was the complexity and non-uniformity of sample description information. For example, diverse representations of clinical sample description make it difficult to compare histological data, such as Gleason score for prostate tumors or Estrogen receptor status for breast carcinomas. This problem was addressed by mapping to a common data format using parameter/value pairs. Finally, each study was independently normalized and archived in a relational database. A large amount of software engineering work was required to deal with the structure and large size of gene expression data and provide a robust query and analysis tool.

Software platforms such as Oncomine are important for discovery and algorithm development. When mined using appropriate algorithms, such as the cancer outlier profile analysis (COPA) method, they can supplement experiments to make fundamental contributions to cancer genetics [69]. The challenges encountered during the creation of Oncomine emphasize the need for data representation standards and public data warehouses that transcend a single community's needs to allow for integrative studies. Statistical normalization and analysis methods for such integrated datasets are also required. Further enriching transcriptome data with complementary information, such as quantitative proteomic data, will present new challenges resulting from even higher data-dimensionality and volume, concordance and discordance between mRNA and corresponding protein data, and potential for information conflict, but also will provide new opportunities for discovery [70–72].

for response assessment according to standards such as Response Evaluation Criteria in Solid Tumors (RECIST) [48,49]. Clinical data is then used in clinical research, for example, to relate exposure factors or treatment parameters to clinical outcome. As clinical data is often collected longitudinally at multiple visits, potentially by different health professionals, organization and storage of the data in standard formats is critical for analysis and interpretation.

Where Computational Biology Can Help

Data collection, organization, aggregation, and storage. To effectively use genome-scale molecular information, it must be collected, organized in a standard way, aggregated, and stored so that it is widely accessible to the research community. Aggregation is pooling data from multiple experiments of the same type. The advantages of data aggregation are: it can increase the sample size and lead to improved statistical power for comparisons, and it can improve coverage, for instance, over more cell types, different parts of the tumor, or from different populations. Public biomedical data repositories that organize, aggregate, and store data from genome-scale molecular experiments are increasingly available for diverse data types (Tables 1–4).

Data from these repositories support comprehensive molecular analysis of tumors. For instance, commonly activated gene signatures [50], coordinately regulated gene modules involved in a biological process [51], and regulatory programs that control disease development [52,53] in various cancers have been identified by combining data from multiple microarray datasets. This dataset pooling or meta-analysis helps draw inferences that may not be possible with a single study with a limited number of samples/observations.

However, data aggregation is difficult unless standard methods for data collection, organization, archiving, and exchange are developed and followed. Table 5 summarizes some of the standards for different types of biological data. Development and community-wide use of standards enhances the ability of research groups to exchange data and provides a strong foundation on which to build data storage, processing, and analysis software.

In Box 1 we present a case study on Oncomine—a microarray data aggregation platform—highlighting the challenges encountered in aggregating microarray data from multiple sources and the opportunities that it provides.

Data integration. Data integration is the combination of heterogeneous biological data encoded with different semantics. Integration of heterogeneous data is useful not only to validate and to improve confidence in experimental results but also to develop more complete models of biological systems. For instance, real time quantitative RT–PCR data are routinely used to validate cDNA array experiment results. Integration of gene expression and proteomics data, for example, could be used to identify post-transcriptional or post-translational modifications. It could also provide insights into the advantages and shortcomings of particular experimental methods.

The integration of diverse experimental data to build models of biological processes, or pathways, will boost our ability to identify clinical markers and therapeutic targets and to interpret genotype information. For instance, a marker such as prostate specific antigen (PSA) may be widely known and used clinically without much knowledge about its biology. Knowing the pathway involving the marker gene allows other pathway components, or the entire pathway, to serve as a more specialized marker.

Clinical data, securely and ethically accessed, can be integrated with molecular data from basic research to gain insight into disease state and lead to better treatments. Molecular and clinical data has been integrated for identification of clinically relevant subtypes of leukemia with 100% sensitivity and specificity [54]. Analysis of molecular profiles on biospecimens from patients before, during, and after therapy can lead to identification of drug-responsive or nonresponsive profiles that could be used to optimize choice of therapy. In the case of advanced non-small cell lung cancer (NSCLC), a significant difference in response to the kinase inhibitor gefitinib (Iressa) was observed for patients with mutations or amplifications in the *EGFR* gene [55,56]. Comparison of patient molecular profiles with poor and favorable outcomes can be used to predict disease outcome (prognosis). For example, in breast cancer patients, the estrogen-receptor status of the primary tumor and other clinical features have been used to construct nomograms that predict the likelihood of developing non-sentinel lymph node (non-SLN) metastases. Such information can be used to assess metastatic risk and the need for complete axillary lymph node dissection [57]. Additionally, a 32-gene expression signature distinguished p53-mutant and wild-type breast

Box 2. Data Integration for Biological Discovery: Case Studies

***LRPPRC* in Leigh syndrome.** Data integration was applied to identify one of the genes responsible for Leigh syndrome [73]. Classical Leigh syndrome is an early onset fatal neurodegenerative disorder characterized by bilateral lesions in the brain stem, basal ganglia thalamus, and spinal cord and is known to involve a cytochrome c oxidase (COX) deficiency mapped to Chromosome 9 [74,75]. The French Canadian form of Leigh syndrome (LSFC) is distinct and is known to involve a gene on Chromosome 2, where no known cytochrome c oxidase gene is localized [76]. However, clinical and biochemical data suggest that a mitochondrial respiratory chain disorder is involved. From a genome-wide association study that mapped LSFC to Chromosome 2p16–21, Mootha et al. collected 15 known and 15 predicted genes using the UCSC genome browser. Microarray databases from the Whitehead, RIKEN array database and the Genomics Institute of the Novartis research foundation were then used to identify genes that coexpressed with known mitochondrial genes. *LRPPRC,* which mapped to the LSFC candidate genomic region, showed the highest correlated gene expression with known mitochondrial genes. *LRPPRC* was also found in a list of mitochondrial-associated proteins identified by mass spectrometry. Neither method alone was enough to implicate the specific gene. Thus, data integration overcame the incomplete coverage, low sensitivity, or specificity limitations of the individual experimental approaches.

***MITF* in malignant melanoma.** Garraway et al. used an integrative approach to identify *MITF* as a "lineage survival" or "lineage addiction" oncogene required for development and maintenance of malignant melanoma [77]. The authors used the NCI60 panel, which is a collection of 59 human cancer cell lines derived from nine different types of tissues. SNP arrays of NCI60 cell lines were used to define genomic subclusters that were specifically amplified in the melanoma subset. This information was then integrated with the publicly available NCI60 gene expression data generated by the genomics institute of the Novartis foundation to correlate gene expression with the copy number gain. Remarkably, *MITF* was the only highly expressed gene in the amplicon identified in the SNP array analysis. This result was validated using FISH and automated quantitative analysis (AQUA) of MITF protein levels in patient samples. As before, public availability of gene expression data was instrumental for the authors to integrate expression data with their own SNP data in defining the function of a gene in the context of cancer.

tumors of different histologies, and this strategy outperformed sequence-based assessments of p53 in predicting prognosis and therapeutic response [58]. Combination of molecular profiles with clinical profiles can also help select patients for targeted treatment.

Effective integration of heterogeneous data is difficult since important information necessary to decipher data semantics, such as context, could be missing. It is often possible for the human brain to infer this information using prior knowledge, but such tasks have remained impossible to encode into a model or rule-based computational procedure. Thus, missing information can lead to errors during integration. For example, a query may identify relevant datasets labeled with the term *metastasis*. However, metastatic processes could be different among tissues, so tissue information is required to avoid errors in dataset selection. Also, datasets labeled with alternate descriptions of metastasis may be missed in the search. As another example, proper mapping of gene and proteins between microarray and mass spectrometry data is an important requirement for integrating the datasets. This task could be confounded if the measured gene transcript is a different splice variant than the measured protein product. Thus, correct data integration requires semantic compatibility among datasets and context resolution.

In Box 2 we present two successful examples of how integration of heterogeneous data from different sources can help in biological discoveries. In the first example, publicly available data was used to help identify *LRPPRC* as a specific gene involved in Leigh syndrome, a complex hereditary disease. In the other example, the discovery of the central role of micropthalmia-associated transcription factor *(MITF)* amplification to malignant melanoma was accomplished using experimental and computational approaches to integrate copy number data, with publicly available gene expression and genome information. These examples provide proof of principle that there already exists a wealth of biological data in the public domain, and data integration approaches can be used to better understand biology and development of disease, including cancer.

Software systems and algorithms for data analysis. The volume of the data generated by modern biomedical studies is too large to be processed by the human brain alone. Data storage, querying, and presentation software systems and computational algorithms are required for effective interpretation of large-scale experimental data. Automated methods are now available to find genes or pathways that are significantly differentially expressed using molecular profiles [59–63]. The GOALIE algorithm maps the temporal evolution of biological processes from time-course gene expression data [64]. An algorithm for general integration of heterogeneous data that differ in type and size has been proposed [65]. Further studies of pathways in the context of gene expression and complementary genome-scale data will lead to the discovery of new pathway components, some of which could be new vulnerable points and therapeutic targets [66].

Pathway simulations, requiring detailed cellular models, have been used in model organisms, such as *Escherichia coli* and budding yeast, to find pathway regulators and to design new experiments that test hypotheses about the function of the pathway [67]. Applying these methods to predict the result of a tumor-specific mutation in multicellular organisms using human pathway information is an important research direction and major challenge for computational biology.

Development of software systems that integrate diverse biomedical research data types promise to support the study of disease biology and development. For instance, the REMBRANDT (Repository for Molecular Brain Neoplasia Data) framework attempts to integrate clinical and molecular data from the Glioma Molecular Diagnostic Initiative (GMDI)—a collaborative effort of the NCI and the US National Institute of Neurological Disorders and Stroke (NINDS) (http://rembrandt.nci.nih.gov). REMBRANDT can be used to query and generate statistical reports across all component glioma (brain tumor) datasets.

Computational prediction of the biological effects of drugs based on structure–function relationships across many targets can help increase the success rate of clinical trials and may forewarn of possible adverse events associated with the small-molecule therapy. This strategy could also be used to develop drug combinations to target multiple vulnerable points to shut down tumor growth. ADME/Tox (absorption, distribution, metabolism, excretion, and toxicity) prediction based on molecular profiles can help eliminate candidate drugs that

have unacceptable toxicity early in the drug discovery process and thus reduce the cost of drug development.

Social challenges for computational approaches. Aggregating, integrating, and analyzing experimental data from multiple sources must overcome social as well as technical challenges. Critically, while archives of datasets from molecular studies are often publicly available, a public clinical counterpart remains largely unavailable due to patient privacy concerns. Securely providing de-identified patient data obtained with adequate patient consent, for example, as per the US Health Insurance Portability and Accountability Act (HIPAA) guidelines (http://www.hhs.gov/ocr/hipaa), is a viable solution.

Data collected from biological samples must be clearly annotated using standard representations, including descriptions of the sample and experimental conditions. Without such information data integration is significantly more difficult, inefficient, and error-prone. Effort must be spent to make data publicly available, to agree on and use community standards, and most importantly to make computational tools easy to use for biologists; these steps will significantly improve the effectiveness of translational cancer research. Computing infrastructure for facilitating data aggregation/integration can use either centralized systems wherein an investigator accesses a central computer system that holds all the data, or, alternatively, federated systems where an investigator sends a query and the system assembles pertinent information from where it exists. Two examples of research computer systems for data integration are caBIG and BIRN's cyber infrastructure. The Cancer Biomedical Informatics Grid (caBIG) is a network to enable sharing of data and software tools across individuals and cancer research institutions to improve the pace of innovations in cancer prevention and treatment (http://cabig.cancer.gov). The Biomedical Informatics Research Network (BIRN) is a distributed virtual community of shared resources that currently supports the sharing and analysis of neuroimaging data (http://www.nbirn.net).

Concluding Remarks

Computational biology is pivotal for to effectively use large and diverse data resources to provide insights into disease biology and to optimize treatment. Modeling and simulation techniques, standards, and software systems must be enhanced to deal with expanding molecular and clinical information. Making well-organized experimental datasets widely accessible will spur algorithm development, testing, and comparison, leading to the development of better computational methods. These new computational tools will allow us to effectively interpret available genome-scale datasets to improve disease diagnosis, prognosis, therapy, and prevention. ■

Acknowledgments

We apologize to those whose publications we were not able to refer to in this review because of space limitations.

Author Contributions. JPM conceived, contributed sections to, coordinated, and put together the complete manuscript. GDB, SP, AMC, CS, SJB, and BM discussed and contributed sections to the paper. BST and MA contributed sections to the paper.

Funding. The authors received no specific funding for this article.

Competing interests. The authors have declared that no competing interests exist.

References

1. von Mehren M, Adams GP, Weiner LM (2003) Monoclonal antibody therapy for cancer. Annu Rev Med 54: 343–369.
2. FDA (2004) Challenge and opportunity on the critical path to new medical products. US Federal Drug Administration. Available: http://www.fda.gov/oc/initiatives/criticalpath. Accessed 25 December 2006.
3. Fabian MA, Biggs WH III, Treiber DK, Atteridge CE, Azimioara MD, et al. (2005) A small molecule–kinase interaction map for clinical kinase inhibitors. Nat Biotechnol 23: 329–336.
4. McNeil N, Ried T (2000) Novel molecular cytogenetic techniques for identifying complex chromosomal rearrangements: Technology and applications in molecular medicine. Expert Rev Mol Med 2000: 1–14.
5. Baldwin C, Garnis C, Zhang L, Rosin MP, Lam WL (2005) Multiple microalterations detected at high frequency in oral cancer. Cancer Res 65: 7561–7567.
6. Rauch A, Ruschendorf F, Huang J, Trautmann U, Becker C, et al. (2004) Molecular karyotyping using an SNP array for genomewide genotyping. J Med Genet 41: 916–922.
7. Consortium TIH (2003) The International HapMap Project. Nature 426: 789–796.
8. Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6: 597–610.
9. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.
10. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14: 1675–1680.
11. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18: 630–634.
12. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286: 531–537.
13. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, et al. (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412: 822–826.
14. Blais A, Dynlacht BD (2005) Constructing transcriptional regulatory networks. Genes Dev 19: 1499–1511.
15. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24: 227–235.
16. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, et al. (2000) A gene expression database for the molecular pharmacology of cancer. Nat Genet 24: 236–244.
17. Creighton CJ, Bromberg-White JL, Misek DE, Monsma DJ, Brichory F, et al. (2005) Analysis of tumor–host interactions by gene expression profiling of lung adenocarcinoma xenografts identifies genes involved in tumor formation. Mol Cancer Res 3: 119–129.
18. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, et al. (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. Nat Biotechnol 20: 301–305.
19. Mann M, Ong SE, Gronborg M, Steen H, Jensen ON, et al. (2002) Analysis of protein phosphorylation using mass spectrometry: Deciphering the phosphoproteome. Trends Biotechnol 20: 261–268.
20. Wells L, Vosseller K, Cole RN, Cronshaw JM, Matunis MJ, et al. (2002) Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. MolCell Proteomics 1: 791–804.
21. Stover DR, Caldwell J, Marto J, Root K, Mestan J, et al. (2004) Differential phosphoprofiles of EGF and EGFR kinase inhibitor-treated human tumor cells and mouse xenografts. Clinical Proteomics 1: 69–80.
22. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1: 376–386.
23. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3: 1154–1169.
24. Chaurand P, Schwartz SA, Reyzer ML, Caprioli RM (2005) Imaging mass spectrometry: Principles and potentials. Toxicol Pathol 33: 92–101.
25. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, et al. (2005) Global analysis of protein phosphorylation in yeast. Nature 438: 679–684.
26. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551–3567.
27. Yates JR III, Eng JK, McCormack AL, Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 67: 1426–1436.
28. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383–5392.
29. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 4646–4658.

30. Anderson DC, Li W, Payan DG, Noble WS (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J Proteome Res 2: 137–146.
31. Ruvkun G, Wightman B, Ha I (2004) The 20 years it took to recognize the importance of tiny RNAs. Cell 116: S93–S96, and 92 pages following.
32. He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, et al. (2005) A microRNA polycistron as a potential human oncogene. Nature 435: 828–833.
33. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT (2005) c-Myc-regulated microRNAs modulate E2F1 expression. Nature 435: 839–843.
34. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, et al. (2005) MicroRNA expression profiles classify human cancers. Nature 435: 834–838.
35. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116: 281–297.
36. John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. (2004) Human MicroRNA targets. PLoS Biol 2(11): e363.
37. Cary MP, Bader GD, Sander C (2005) Pathway information for systems biology. FEBS Lett 579: 1815–1820.
38. Nicholson JK, Wilson ID (2003) Opinion: Understanding "global" systems biology: Metabonomics and the continuum of metabolism. Nat Rev Drug Discov 2: 668–676.
39. Cheng LL, Burns MA, Taylor JL, He W, Halpern EF, et al. (2005) Metabolic characterization of human prostate cancer with tissue magnetic resonance spectroscopy. Cancer Res 65: 3030–3034.
40. Chen JH, Enloe BM, Weybright P, Campbell N, Dorfman D, et al. (2002) Biochemical correlates of thiazolidinedione-induced adipocyte differentiation by high-resolution magic angle spinning NMR spectroscopy. Magn Reson Med 48: 602–610.
41. Kupfer A, Kupfer H (2003) Imaging immune cell interactions and functions: SMACs and the immunological synapse. Semin Immunol 15: 295–300.
42. Reinhardt RL, Jenkins MK (2003) Whole-body analysis of T cell responses. Curr Opin Immunol 15: 366–371.
43. Shakhar G, Lindquist RL, Skokos D, Dudziak D, Huang JH, et al. (2005) Stable T cell–dendritic cell interactions precede the development of both tolerance and immunity in vivo. Nat Immunol 6: 707–714.
44. Bousso P, Robey EA (2004) Dynamic behavior of T cells and thymocytes in lymphoid organs as revealed by two-photon microscopy. Immunity 21: 349–355.
45. Zuo Y, Yang G, Kwon E, Gan WB (2005) Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. Nature 436: 261–265.
46. Presley JF (2005) Imaging the secretory pathway: The past and future impact of live cell optical techniques. Biochim Biophys Acta 1744: 259–272.
47. Piantadosi S (2005) Clinical trials: A methodological perspective. 2nd edition. New York: Wiley. 720 p.
48. Duffaud F, Therasse P (2000) New guidelines to evaluate the response to treatment in solid tumors. Bull Cancer 87: 881–886.
49. Therasse P, Le Cesne A, Van Glabbeke M, Verweij J, Judson I (2005) RECIST versus WHO: Prospective comparison of response criteria in an EORTC phase II clinical trial investigating ET-743 in advanced soft tissue sarcoma. Eur J Cancer 41: 1426–1430.
50. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101: 9309–9314.
51. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098.
52. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, et al. (2005) Mining for regulatory programs in the cancer transcriptome. Nat Genet 37: 579–583.
53. Rhodes DR, Chinnaiyan AM (2005) Integrative analysis of the cancer transcriptome. Nat Genet 37: S31–S37.
54. Haferlach T, Kohlmann A, Schnittger S, Dugas M, Hiddemann W, et al. (2005) Global approach to the diagnosis of leukemia using gene expression profiling. Blood 106: 1189–1198.
55. Paez J, Sellers WR (2003) PI3K/PTEN/AKT pathway. A critical mediator of oncogenic signaling. Cancer Treat Res 115: 145–167.
56. Taron M, Ichinose Y, Rosell R, Mok T, Massuti B, et al. (2005) Activating mutations in the tyrosine kinase domain of the epidermal growth factor receptor are associated with improved survival in gefitinib-treated chemorefractory lung adenocarcinomas. Clin Cancer Res 11: 5878–5885.
57. Van Zee KJ, Manasseh DM, Bevilacqua JL, Boolbol SK, Fey JV, et al. (2003) A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy. Ann Surg Oncol 10: 1140–1151.
58. Miller LD, Smeds J, George J, Vega VB, Vergara L, et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 102: 13550–13555.
59. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31: 19–20.
60. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, et al. (2003) MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 4: R7.
61. Pradines J, Rudolph-Owen L, Hunter J, Leroy P, Cary M, et al. (2004) Detection of activity centers in cellular pathways using transcript profiling. J Biopharm Stat 14: 701–721.
62. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. (2005) A network-based analysis of systemic inflammation in humans. Nature 437: 1032–1037.
63. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18: S233–S240.
64. Antoniotti M, Ramakrishnan N, Mishra B (2005) GOALIE, a common lisp application to discover "Kripke Models": Redescribing biological processes from time-course data. Proceedings of the International Lisp Conference; 19–22 June 2005; Stanford, California, United States. Available: http://bioinformatics.nyu.edu/~marcoxa/publications/TLBIO/ARM-ILC-2005.pdf. Accessed 25 December 2006.
65. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, et al. (2005) A data integration methodology for systems biology. Proc Natl Acad Sci U S A 102: 17296–17301.
66. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085–1094.
67. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, et al. (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol Biol Cell 11: 369–391.
68. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) ONCOMINE: A cancer microarray database and integrated data-mining platform. Neoplasia 6: 1–6.
69. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310: 644–648.
70. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, et al. (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 8: 393–406.
71. Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, et al. (2002) Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 1: 304–313.
72. Chen G, Gharib TG, Wang H, Huang CC, Kuick R, et al. (2003) Protein profiles associated with survival in lung adenocarcinoma. Proc Natl Acad Sci U S A 100: 13537–13542.
73. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, et al. (2003) Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci U S A 100: 605–610.
74. Shoubridge EA (2001) Cytochrome c oxidase deficiency. Am J Med Genet 106: 46–52.
75. Zhu Z, Yao J, Johns T, Fu K, De Bie I, et al. (1998) SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome. Nat Genet 20: 337–343.
76. Lee N, Daly MJ, Delmonte T, Lander ES, Xu F, et al. (2001) A genomewide linkage-disequilibrium scan localizes the Saguenay–Lac-Saint-Jean cytochrome oxidase deficiency to 2p16. Am J Hum Genet 68: 397–409.
77. Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, et al. (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 436: 117–122.
78. Baudis M, Cleary ML (2001) Progenetix.net: An online repository for molecular cytogenetic aberration data. Bioinformatics 17: 1228–1229.
79. Knutsen T, Gobu V, Knaus R, Padilla-Nash H, Augustus M, et al. (2005) The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: Linkage of chromosomal aberrations with the genome sequence. Genes Chromosomes Cancer 44: 52–64.
80. Mitelman FJB, Mertens F, editors (2006) Mitelman Database of Chromosome Aberrations in Cancer. Available: http://cgap.nci.nih.gov/Chromosomes/Mitelman. Accessed 3 January 2007.
81. Grunau C, Renault E, Rosenthal A, Roizes G (2001) MethDB—A public database for DNA methylation data. Nucleic Acids Res 29: 270–274.
82. Amoreira C, Hindermann W, Grunau C (2003) An improved version of the DNA Methylation database (MethDB). Nucleic Acids Res 31: 75–77.
83. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, et al. (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91: 355–358.
84. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
85. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, et al. (2005) ArrayExpress—A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33: D553–D555.
86. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, et al. (2001) The Stanford Microarray Database. Nucleic Acids Res 29: 152–155.
87. Omenn GS (2005) Exploring the human plasma proteome. Proteomics 5: 3223, 3225.
88. Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3: 1234–1242.
89. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: The proteomics identifications database. Proteomics 5: 3537–3545.

90. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, et al. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 6: R9.
91. Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics 4: 1920–1932.
92. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–D280.
93. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6: R2.
94. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. (2005) Reactome: A knowledgebase of biological pathways. Nucleic Acids Res 33: D428–D432.
95. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions, and pathways. Nucleic Acids Res 33: D284–D288.
96. Bader GD, Betel D, Hogue CW (2003) BIND: The Biomolecular Interaction Network Database. Nucleic Acids Res 31: 248–250.
97. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363–2371.
98. Bader GD, Heilbut A, Andrews B, Tyers M, Hughes T, et al. (2003) Functional genomics and proteomics: Charting a multidimensional map of the yeast cell. Trends Cell Biol 13: 344–356.
99. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles. 31: 374–378.
100. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: A general repository for interaction datasets. Nucleic Acids Res 34: D535–D539.
101. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3: RESEARCH0046.
102. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al. (2004) The HUPO PSI's molecular interaction format—A community standard for the representation of protein interaction data. Nat Biotechnol 22: 177–183.