TeaserCrisis of pharmaceutical industry prompts Research and Development (R&D) focus from blockbusters to niche busters; its clinical success depends on successful prediction of stratification biomarkers based on combining data and knowledge as an integrative model.

Challenges and opportunities for oncology biomarker discovery

Avisek Deyati[1,2], Erfan Younesi[2], Martin Hofmann-Apitius[2] and Natalia Novac[1]

[1] Knowledge Management, Merck Serono, 250 Frankfurterstrasse, Darmstadt 64293, Germany
[2] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Recent success of companion diagnostics along with the increasing regulatory pressure for better identification of the target population has created an unprecedented incentive for drug discovery companies to invest in novel strategies for biomarker discovery. In parallel with the rapid advancement and clinical adoption of high-throughput technologies, a number of knowledge management and systems biology approaches have been developed to analyze an ever increasing collection of OMICs data. This review discusses current biomarker discovery technologies highlighting challenges and opportunities of knowledge capturing and presenting a perspective of the future integrative modeling approaches as an emerging trend in biomarker prediction.

Since 1955 human life expectancy has improved worldwide from 48 years to 71 years [1] (Country comparison: life expectancy at birth: http://www.cia.gov/library/publications/the-world-fact-book/rankorder/2102rank.html). Two Noble laureates Joshua Lederberg and George Hitchings have claimed that the increase in life expectancy in past 50 years can be largely attributed to new medicines [2,3]. Their claim has been well supported by the success stories of pharmaceutical industries during this phase, which is reflected by the FDA (Food and Drug Administration, US) approval of ~1222 new drugs during this period [4]. However, in the past decade by and large all the drug companies are affected by crisis, characterized by increased expenditure, augmented pipeline attrition rate and patent expiry of major blockbusters. The success rate of late stage clinical trials has fallen by 10% for phase II studies in the recent years. At the same time the number of phase III terminations doubled in the past five years [5,6]. This devastating situation prompts pharmaceutical companies to search for new business models that would reduce time of a drug to reach the market and increase the clinical success rate, thereby satisfying regulatory authorities and patients' needs.

One of the paradigm shifts of pharmaceutical R&D in drug discovery is to move the focus from blockbusters to niche busters, that is therapies targeted towards specific patient populations coined as stratified medicine. The success of stratified medicine depends on accurate diagnostic

Avisek Deyati

Avisek Deyati is currently pursuing his PhD degree in Department of Bioinformatics at Fraunhofer Institute for Algorithms and Scientific Computing (SCAI) University of Bonn. He has received his MSc in Bioinformatics from the University of Sussex. Currently, he is involved in knowledge- and data-driven biomarker prediction and his main research interest is in the development of knowledge management workflows to aid identification of potential biomarkers from the scientific literature and further validation of the hypotheses using high throughput data from public and proprietary data repositories. To ensure the future applicability of the developed algorithms, Avisek is working in the industrial environment of pharmaceutical company Merck KGaA, Darmstadt, Germany.

Erfan Younesi

Erfan Younesi is a PhD candidate in Computational Life Sciences at University of Bonn and, since 2008, works as research associate at the Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany. His research interests revolve around translational bioinformatics including integrative knowledge- and data-driven modeling of disease mechanism with focus on brain disorders for drug target and biomarker discovery.

Martin Hofmann-Apitius

Martin Hofmann-Apitius studied Genetics and Microbiology at the University of Tuebingen. He holds a PhD in Molecular Biology, and for more than 10 years he worked in experimental molecular biology with a strong focus on metastatic behavior of tumor cells. He has experience in both, academic (Research Centre Karlsruhe; German Cancer Research Centre) and industrial research (BASF, Boehringer Ingelheim and LION bioscience). Since 2002 he has been leading the Department of Bioinformatics at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI) and also since July 2006 he is a Professor for Applied Life Science Informatics at B-IT. His current research areas are information extraction, in silico target validation and virtual screening, distributed and high performance computing.

Natalia Novac

Natalia Novac is currently working as a knowledge manager at Merck-Serono. She is responsible for driving knowledge management consultancy for R&D projects focused on candidate biomarkers prediction, compound repositioning, and mode of action modeling. Graduated as a biochemist at the State University of Moldova, Natalia has received her MSc at the University of Warwick, UK and PhD at the Institute of Genetics, Germany. Her 11 years of professional experience includes working in academic research (Terry Fox Laboratory, Canada and Georg-Speyer Haus Institute in Frankfurt) and pharmaceutical industry in the fields of high-content screening, scientific data and information extraction and analytics.

Corresponding author:. Novac, N. (natalia.novac@merckgroup.com)

1359-6446/06/$ - see front matter © 2013 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.drudis.2012.12.011

GLOSSARY

Stratification biomarker a biomarker that allows to identify the specific patient population that will benefit from the treatment
Knowledge base a database for information that is collected, organized, and structured in a computer-readable form
Disease map a connected network of molecular interactions representing the disease mechanism in graphs
Text-mining processing unstructured, free text into structured extractable information
Named entity recognition process of identification of biological entity names such as disease or gene names in running text
Integrative modeling linking heterogenous data types across different levels of a complex systems within a single framework

tests that identify patients who will benefit from targeted therapy. Today approximately 10% of drug labels approved by the FDA contain pharmacogenomic information reflecting a clear trend towards such customized therapy [7]. Pharmaceutical companies are actively perusing the biomarker driven approach as one of the next major reinventions in the field, which is evident through the FDA release of 'Pharmacogenomic Biomarkers in Drug Labels' summarizing 105 approved drugs with 112 associated biomarkers as of 27.07.2012 [8] (Table of Pharmacogenomic Biomarkers in Drug Labels: http://www.fda.gov/drugs/scienceresearch/research-areas/pharmacogenetics/ucm083378.htm). The trend is continuously rising with the most recent examples of Zelboraf (Vemurafenib) approved with the companion genetic test for the BRAF mutation for late stage melanoma and Xalkori (Crizotinib) approved in combination with the companion genetic test for the ALK gene for late stage lung cancer (FY 2011 Innovative Drug Approvals: http://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM278358.pdf) [7].

Despite these success stories of companion diagnostics, a significant gap exists between the R&D expenditure, number of biomarker-related research grants and available clinically validated biomarkers [8]. Better understandings of drugs' mode of action as well as the alternative pathways that can function to subside therapeutic effects are crucial to build a more productive therapy design. The technological advancements in the fields of genomics, transcriptomics, proteomics, metabolomics and commoditization of these technologies, enabling high-throughput profiling, lead to deluge of publicly available biomedical data. On the other hand, there is a vast amount of biomedical knowledge accumulated in the textual body of scientific literature and patents that could be of tremendous value for translational efforts. The lack of standardized translational algorithms allowing the use of OMICs data along with knowledge derived from scientific literature is one major reason behind the scarcity of biomarkers currently used in the clinic. To fill the gap of OMICs data interpretation, a number of systems biology approaches are suggested by the scientific community, however, there is no proof-of-concept methodology, which can lead to the successful biomarker prediction, and it is not clear how the success of OMICs technologies can be translated from research discovery to clinical biomarker.

The scope of this review is to understand which OMICs technologies are currently used for the identification of biomarkers on the one hand and which methodologies exist for the recovering of biomarker-related information from the text on the other hand. We will also shed light on what are the current standings of both and what is the future prospective in the integrations of these efforts. To examine the success of OMICs in biomarker discovery we did a retrospective analysis of currently approved biomarkers in oncology to search for technologies used for their discovery. We analysed the methodologies for biomarker-related information extraction from the literature based on publically and proprietary knowledge repositories for their strengths and weaknesses. We assessed the current state of the text-mining approaches representing the emerging trend in automated biomarker-related information extraction. Finally, to get a better understanding of the integrated approaches, we overview the methodologies combining OMICs data analysis with scientific text-derived knowledge and give a perspective of this method as an emerging hope of integration of OMICs data and text-derived knowledge that will contribute to the better biomarker prediction translated in successful clinical outcomes.

Biomarkers in current clinical practice: focus on oncology

Among the entire pharmaceutical R&D, the quest for therapeutic breakthroughs in the field of oncology accounts for 29% of total R&D expenditure. Cancer remains a major cause of human suffering with six million people dying every year and ten million new cases reported annually (Beyond the Blockbuster Drug: http://www.pharmatree.in/pdf/reports/Beyond%20the%20Blockbuster%20Drug_Strategies%20for%20nichebuster%20drugs,%20targeted%20therapies%20and%20personalized%20medicine.pdf). Given the fact that cancer is a highly heterogeneous disease not only in terms of histology and clinical outcome but also at the molecular level, it is not surprising that oncology is among the first indications moving towards targeted therapies. EGFR, Her2/neu, ALK, BRAF, Bcr-Abl, PIK3CA, JAK2, MEK, Kit and PML-RARα are targets of recently approved targeted therapies in cancer. These target molecules and their downstream effectors are often subject to various changes on genomic, transcriptomic, proteomic and epigenetic levels. Therefore, status of those molecules underlies diversified patient-specific clinical responses to targeted therapies [9].

According to the FDA, such molecules that can prospectively predict the probable response of a selected subpopulation of patients to therapy are defined as stratification or predictive biomarkers [10]. As cancer is one of the prime areas of targeted therapies, we present here Table 1 summarizing stratification biomarkers currently in clinical practice in oncology along with their approved treatment. As shown in Table 1, with the few exception such as KRAS, most of the biomarkers are direct drug targets of the respective therapies. Given the fact that the majority of stratification biomarkers have been approved after the therapy went to the market (i.e. derived from the retrospective analysis of late-stage clinical trials or postmarketing surveys), it is obvious that the biology of the target and its changes under the pathological conditions were not really apparent during the course of clinical development. Moreover, even in the stratified patient population, therapeutic response is far from being equally successful [11–13], suggesting that having a sound knowledge on changes that occur downstream of the therapeutic targets in the context of pathways and molecular interaction networks would be absolutely

TABLE 1

FDA-approved stratification biomarkers for targeted therapy in oncologyAdopted from: http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm

Functional class	Biomarker	Therapy
Kinase	EGFR	Cetuximab, Erlotinib, Gefitinib, Panitumumab
Kinase	Her2/neu	Lapatinib, Trastuzumab, Pertuzumab
Kinase	PDGFR	Imatinib
Kinase	Estrogen receptor	Fulvestrant, Exemestane
Kinase	ALK	Crizotinib
Kinase	KRAS	Cetuximab, Panitumumab
Kinase	BRAF	Vemurafenib
Immune cell surface receptor	CD20	Tositumumab
Immune cell surface receptor	CD25	Denileukin diftox
Immune cell surface receptor	CD30	Brentuximab vedotin
Immune cell surface receptor	C-Kit	Imatinib
Fusion gene	PML-RARα	Arsenic trioxide
Fusion gene	BCR-ABL	Dasatinib

crucial to move further in the direction of biomarker-driven stratified medicine. Efforts to elucidate such global downstream changes lead to the explosion of technologies for biomarker identification overviewed in the next section.

OMICs based technologies for biomarker identification

The rapid evolution of high-throughput technologies designed for screening of biomedical samples with the latest breakthroughs in whole genome sequencing and microRNA (miRNA) profiling gave birth to a number of biological disciplines devoted to generation and study of those multiple OMICs data. Figure 1 summarizes the latest technologies as well as diversification of the biomarker types and underlying data types depending on the nature of changes detected by the respective technology.

Genetic biomarkers

Genetic biomarkers are biomarkers derived from technologies assessing genomic changes, such as exome and whole genome sequencing, polymerase chain reaction (PCR) and fluorescence in situ hybridization (FISH). They can accurately identify single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and structural variations in the genome and delineate their functional significance in the pathophysiology of a defined phenotype. These technologies have been instrumental in finding stratification biomarkers in oncology and some of them are already in clinical practice. For example, KRAS sequencing and PCR were used to discover predictive and prognostic role of the KRAS mutation in colorectal cancer and lung cancer for anti EGFR-therapy resistance [14–17]. PCR/FISH analyses were used to reveal that translocation of BCR-ABL and PML-RARα may serve as predictive biomarkers conferring sensitivity to imatinib mesylate and resistance to arsenic oxide in leukemia [18,40]. The same technologies were used to identify mutation/amplification and translocation of ALK gene as biomarkers predicting the efficacy of crizotinib treatment in late stage lung cancer [20].

Expression biomarkers

Different to the traditional single biochemical and histopathological measurements, expression biomarkers (transcriptomics biomarker) represent a fingerprint containing multiple biomarkers,

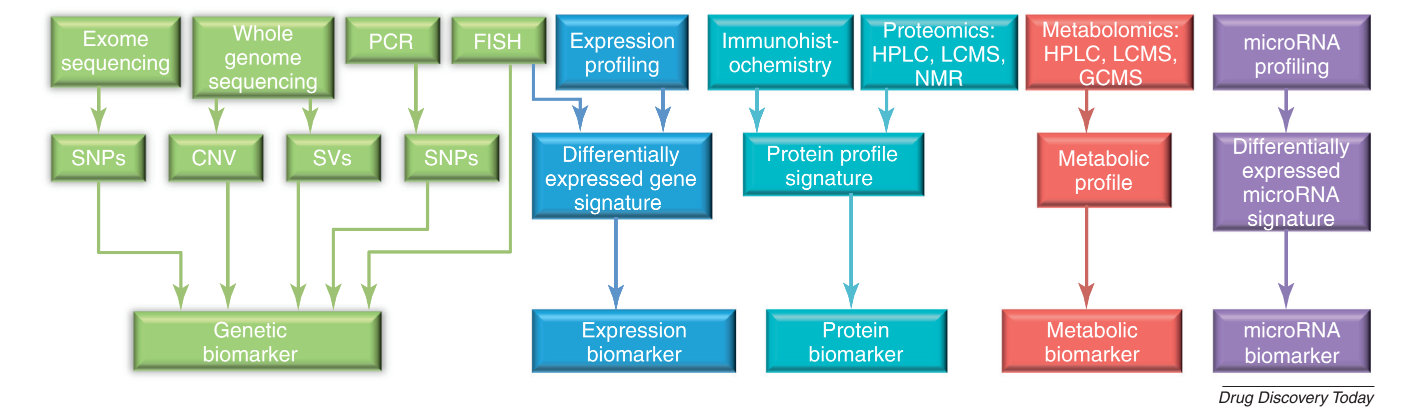

FIGURE 1

Current technologies and data types used for biomarker discovery in preclinical and clinical research. *Abbreviations*: CNV, copy number variations; FISH, fluorescence in situ hybridization; GCMS, gas chromatography mass spectrometry; HPLC, high-performance liquid chromatography; LCMS, liquid chromatography–mass spectrometry; NMR, nuclear magnetic resonance; PCR, polymerase chain reaction; SNPs, single nucleotide polymorphisms; SVs, structural variations.

which collectively indicate a particular pathophysiology [21]. High-throughput technologies like well-established microarray expression profiling and more recent RNAseq technologies are able to identify differential expression of an entire genome at any specific sample in any given time point. There are a number of reports on using these technologies to identify biomarkers in specific cancer subtypes. Two of expression biomarker tests are clinically approved for patient stratification in breast cancer. MammaPrint, a unique 70 gene expression profile, is a prognostic biomarker for distant recurrence of the disease following surgery in breast cancer patients [22]. Oncotype DX is another gene expression signature-based biomarker test containing16 cancer-related genes and five reference genes that predict the recurrence of breast cancer in tamoxifen-treated patients with node negative, estrogen positive tumors [23].

Protein biomarkers

Human plasma holds the largest source of the proteome and technologies that can measure slight changes of certain proteins are invaluable tools to identify protein biomarkers in blood. For example, mass spectrometry can capture minor changes of the protein levels and immunohistochemistry can accurately identify a specific protein in the living system. Application of proteomics in discovery of oncology biomarkers can be exemplified by immunohistochemistry-derived Her2, a prognostic, predictive biomarker for the sensitivity to trastuzumab therapy in breast cancer [25]. EGFR is another pharmacodynamic biomarker discovered by immunohistochemical assay in colorectal and lung cancer samples conferring sensitivity to cetuximab, panitumumab and gefitinib treatment [26–29]. Clinical acceptance of novel proteomic biomarkers suffers an anemic rate due to the lack of PCR-like amplification techniques for the vast number of analytes present in extremely small quantities in hugely dynamic plasma [30]. As a potential solution, a novel method of immuno-PCR using conjugations of specific antibodies and nucleic acids is suggested leading to 100–10 000-fold signal amplification thereby increasing sensitivity of protein biomarker detection [31,32].

Metabolic biomarkers

Ever since Otto Warburg hypothesized that altered metabolism (converting glucose carbon to lactate in oxygen rich condition) is specific to cancer cells due to mitochondrial defects, metabolic biomarkers have drawn the attention of researchers to be an effective biomarker for early cancer diagnosis and prognosis [33]. Since then, numerous efforts have been dedicated to the identification of metabolic biomarkers in oncology. Nuclear magnetic resonance (NMR) spectroscopy, high-performance liquid chromatography (HPLC), radioimmunoassay, liquid chromatography–mass spectrometry (LCMS), gas chromatography mass spectrometry (GCMS) and enzyme immunoassay are instrumental in analyzing the metabolite levels in response to pathophysiological change or treatment. Until now only two metabolic biomarkers have made it to clinical practice. Metanephrine and normetanephrine are two metabolites that are used to predict disease state associated to pheochromocytoma (510(k) substantial equivalence determination decision summary: http://www.accessdata.fda.gov/cdrh_docs/reviews/K032199.pdf). Despite the rapid technological advancement in metabolomics, it is still impossible to differentiate metabolites derived from different subcellular compartments and also current fractionation methods often lead to metabolite leakage between different layers making it even more difficult for metabolite identification [33].

microRNA biomarkers

The involvement of microRNAs (miRNAs) in key cellular processes such as proliferation and cell death and well known negative control over the expression of numerous oncoproteins make them a prime candidate as cancer biomarkers. It also has been reported that cancer-specific miRNAs are detected in the blood from the earlier stages of tumor development and increase in concentration as tumor progresses over time, making them an indicator of the tumor growth [34]. Moreover, unlike other types of biomarkers, miRNAs are remarkably stable in the circulation and formalin-fixed paraffin embedded tissue, making them potentially robust oncology biomarkers. Functional miRNA species have mostly been validated *in vitro* using luciferase reporter activity [34]. Microarray profiling is a powerful high-throughput technology capable of monitoring the expression of thousands of small noncoding RNAs at specific context. Mirage (SAGE), Stem-loop quantitative real time polymerase chain reaction (qRT-PCR) for mature miRNAs, qRT-PCR for precursor miRNAs and bead-based technologies are also frequently used for microRNA profiling [35]. However, no such biomarker exists in cancer clinical practice yet. It is noteworthy that the genetic biomarkers witnessed a remarkable rise in clinical acceptance after the human genome project characterized all the genes. Similar effort is needed to discover and characterize all the miRNAs in human cells in order to transform the potential of miRNAs as cancer biomarker into clinical success. Further understanding on how miRNAs compete with proteins to bind and control the expression of mRNA as well as the functional interaction networks through which miRNAs exert their tissue specific role is needed for future clinical translation [34].

Observing this enormous amplification of data points obtained from biomedical samples raises the question whether these technological advances and widespread, ever-increasing availability of the screening platforms lead to the clinical breakthroughs and which of them proved to be crucial for the discovery of approved biomarkers. To get a better understanding of the technologies that have contributed to the identification of currently approved stratification biomarkers in oncology, we present an OMICs wise overview on data generation platforms and types of data resulting from these platforms in Table 2.

As evident from Table 2, only a handful of stratification biomarkers derived from each technology are currently approved and are in clinical use for oncology. This reflects the hard and long way of developing sensitive, specific and highly predictive biomarkers relevant for clinical decision making from the initial high-throughput data. On the other hand, the potential of the OMICs technologies to discover future biomarkers is tremendous and the expectations have been high since the past 20 years, supporting the huge investments in the development of these technologies.

To understand the future potential of these high-throughput technologies, we compared the number of published candidate biomarkers that is, those reported in the scientific literature, clinical trials registries or scientific conferences with the number

TABLE 2

OMICs technologies for stratification biomarker discovery in oncology

OMICs	Technology	Biomarker	Associated no. of drugs	Refs
Genomics	Flurescence in situ hybridization	ALK	1	[36]
Genomics	Flurescence in situ hybridization	Her2/neu	1	[37]
Genomics	Polymerase chain reaction	BRAF	1	[38]
Genomics	Polymerase chain reaction	CD20	1	[39]
Genomics	Polymerase chain reaction	PML-RARα	1	[40]
Genomics	Polymerase chain reaction	KRAS	1	[19]
Genomics	Sequencing	EGFR	1	[24,41]
Genomics	Sequencing	KRAS	1	[28,29]
Genomics	Sequencing	C-Kit	1	[19]
Genomics	Sequencing	BCR-ABL	1	[42]
Genomics	Sequencing	PDGFR	1	[42]
Proteomics	Immunohistochemistry	EGFR	3	[25,43–45]
Proteomics	Immunohistochemistry	ER/PgR	1	[22]
Proteomics	Immunohistochemistry	CD25	1	[46]
Proteomics	Immunohistochemistry	Her2/neu	1	[47]
Proteomics	Western blot analysis	Her2/neu	1	[20]

of approved biomarkers for each technology described above. For this purpose, we retrieved all oncology-related candidate biomarkers (including disease, stratification, prognostic and diagnostic biomarkers) from GVK Bio Online Biomarker Database (GOBIOM). GOBIOM is an independent manually curated biomarker-related knowledgebase that uses the information derived from clinical reports, annual meetings and journal articles [48]. During the time of writing, GOBIOM possessed information on 15,732 biomarkers covering 16 therapeutic areas supported by 36,681 unique references.

As evident in Fig. 2, although transcriptomics technology, that is, microarray analysis, is one of the oldest and widely used high-throughput technologies, most of the candidate biomarkers are reported to be coming from genomic research followed by proteomics. Stability of the signal coming from genomic analysis as well as higher stability of the protein versus mRNA might be the reason for those biomarkers overweighting the transcriptomics derived biomarkers. Comparing the number of approved biomarkers to those mentioned in public domain reveals that majority of candidate biomarkers either failed or did not reach the clinic yet.

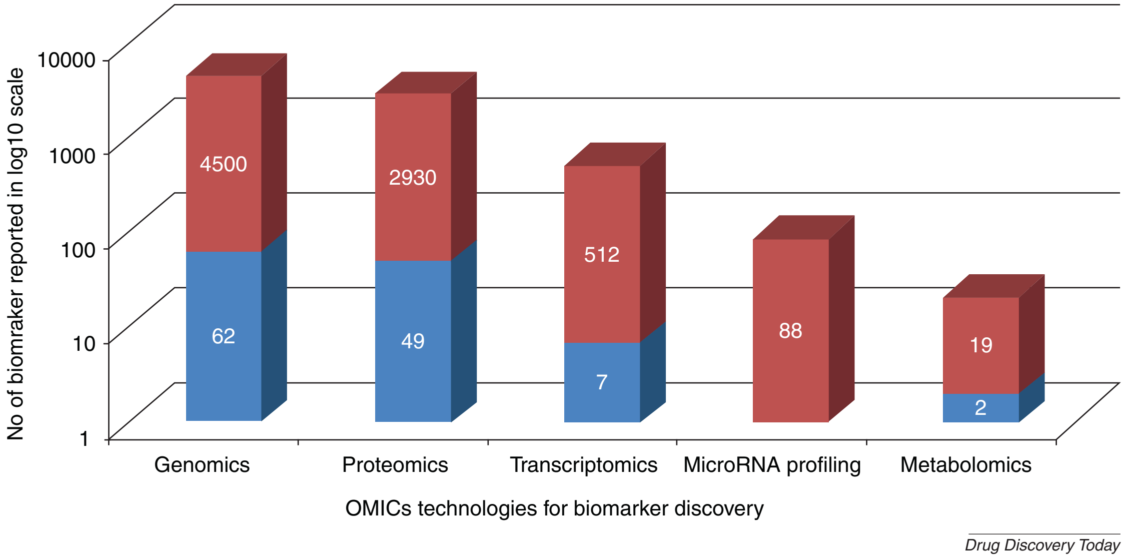

FIGURE 2

Current contribution of OMICs technologies in oncology biomarker discovery extracted from the Gobiom database. In red: total number of candidate biomarkers reported in the public domain. In blue: number of FDA approved biomarkers in current clinical practice for oncology.

Even in the case of strong signal derived from high-throughput technologies, its conversion to clinical practice meets a number of challenges associated, in the first instance, with its functional interpretation. Interpretation of the high-throughput data in the context of molecular pathophysiology of underlying disease and specific treatment is the current rate-limiting step in the biomarker identification and validation. If properly identified, extracted and interpreted, OMICs data sets can provide valuable biological insights. Functional analysis of OMICs data requires knowledge on the molecular interactions and pathways underlying pathophysiology of diseases and treatment mode of actions. Accumulated biological knowledge across different systems levels, therefore, needs to be collected, annotated, transformed into computer-readable form, and stored in a semantically enhanced knowledge base. Such knowledge bases can then be used for knowledge-based analysis of OMICs data sets through integrative approaches that aim at finding key biological processes, pathways, interaction modules or causative network signatures that could be used as candidate biomarkers. In the next section, we provide an analytic view on existing biomarker knowledge bases and their applications in interpretation of high-throughput data.

Knowledge bases dedicated to the interpretation of OMICs data for biomarker discovery

The mission of knowledge bases is to collect and systematize biomedical information through manual information extraction from primary publications in so-called curation process. The curation process organizes knowledge via mapping of extracted information to an underlying ontology. Such knowledge bases provide a number of features for analysis of OMICs data allowing for overlaying OMICs data onto the known pathways, identification of the key pathways underlying the changes and providing network analysis algorithms for identification of the key regulatory molecules behind the respective gene signatures. During the past several years, several public and commercial knowledge bases have been introduced that offer an integrated environment consisting

TABLE 3

Knowledge bases to analyze OMICs data leading to biomarker discovery

Knowledge base	Features	Applications	Refs
Metacore	Metacore is an integrated commercial knowledge base from Thomson Reuters (previously GeneGo) which can support functional analysis (pathways, networks and maps) of span of OMICs data including microarray, sequence based gene expression, SNPs and CGH (comparative genomic hybridization) arrays, proteomics and metabolomics Ranking of the affected pathways and networks from the experimental data based on proprietary algorithms and common functional gene expression interpretation analysis i.e. using gene ontology (GO) Filters based on disease, tissue, sub cellular localization and functional processes to capture specific network The toxicology application of Metacore is specifically designed to discover safety, efficacy and toxicity biomarker to a chemical compound See: http://www.genego.com/metacore.php	Brentnall *et al.* in collaboration with Institute of Systems Biology completed a quantitative proteomic analysis to investigate differentially expressed proteins associated with ulcerative colitis (UC) neoplastic progression. Functional analyses of the differentially expressed proteins with Metacore software identified Sp1 and c-MYC as biomarkers of early and late stage of UC tumorigenesis The same collaborative group made an ICAT-based quantitative proteomics research to analyze protein expression in chronic pancreatitis in comparison with normal pancreas. Metacore assisted pathway analysis revealed that c-MYC as a prominent regulator in the networks of differentially expressed proteins common in pancreatic cancer and chronic pancreatitis Another collaborative group with Bayer Schering Pharma discovered the functional link between the KRAS mutation and Erlotinib resistance in non-small cell lung carcinoma (NSCLC). The functional analysis of the RNA expression data with Metacore indicated a possible correlation between differential expressions of cell adhesion proteins to NSCLC	[49–51]
IPA (A software developed by Ingenuity Systems)	IPA is a manually curated commercial knowledge base from Ingenuity systems Its biomarker filter is specialized to prioritize the molecular biomarker based on species specific connection to diseases, detection in body fluid, expression in specific cell type, cell line, clinical samples and also in stratification biomarker discovery based on disease state or drug response The tool also can produce functional annotation of the biomarker including pathway association See: http://www.ingenuity.com/science/knowledge_base.html	Using Ingenuity pathway analysis Merck & Co. predicted and then experimentally validated that phospho-PRAS40 (Thr246) positively correlates with PI3K pathway activation and AKT inhibitor sensitivity in PTEN deficient mouse prostate tumor model and triple-negative breast tumor tissues Bristol-Myers Squibb has analyzed gene expression signature of responders and non responders to neoadjuvant ixabepilone therapy in breast cancer. Functional analysis of the data with IPA has indicated that significant deregulation of certain proliferation and cell cycle control genes can potentially predict treatment sensitivity Cleveland clinic reported a functional analysis with IPA of the genes carrying non synonymous SNPs that may be associated with the severity of sunitinib-induced toxicity in metastatic clear cell renal cell carcinoma. As per the functional analysis those genes clustered around biological processes like interferon-γ, TNF β, TGF β 1 and amino acid metabolism molecular pathways	[52–54]

TABLE 3 (*Continued*)

Knowledge base	Features	Applications	Refs
Pathway Studio	Pathway Studio is commercial software from Elsevier for pathway analysis as well as analysis of high-throughput OMICs data. Algorithms for analysis of the differential expression data such as Gene Set Enrichment Analysis (GSEA) or network analysis algorithm (NEA) allow detection of weak but consistent expression changes across the pathway genes It is based on the proprietary databases ResNet, DiseaseFx, ChemEffect, Mamalian and Plant database containing relationships between biological molecules, chemicals, diseases and adverse events The databases are built based on proprietary Natural Language Processing (NLP)-based relationship extraction from scientific literature The software suit also provide state of the art network algorithm to pinpoint important nodes from the network perspective. The researcher can also visualize weight of each relationship in the pathways based on the number of literature evidence See: http://www.pathwaystudio.com/	A group from Harvard Medical School published functional connection of 117 highly differentially expressed genes to endometrial cancer. Pathway Studio assisted analysis of the data predicted that many of these genes are correlated to angiogenesis, cell proliferation and chromosomal instability. Further more they also reported ten key differentially regulated genes to be associated to tumor progression Xiao ***et al.*** published functional analysis of EGFR regulated phosphorproteome in nasopharyngeal carcinoma (NPC) to shed light on EGFR downstream signaling. They first identified 33 unique phospho proteins by 2 dimensional difference gel electrophoresis (2D-DIGE) and mass spectrometry. Based on the proteomic data the group built EGFR signaling in NPC by using Pathway Studio and also validate GSTP1 as one of the key EGFR-regulated proteins which is involved in chemoresistance in NPC cells	[55,56]
Compendia Bioscience (Oncomine)	DNA copy number browser: identifying focal amplification across multiple cancer clinical data sets to identify any associated pattern Gene expression browser: to browse differential expression of genes across multiple cancer type covering multiple data sets Mutation browser: discovering cancer association of certain mutations by looking at the frequency of certain gene mutation OncoScore: based on the gene expression data to stratify the patient population based on disease prognosis and response to a therapeutic intervention. At the moment the service is limited to breast and colon cancer See: http://www.compendiabio.com/	Using Oncomine a group from the University of Michigan predicted that decreased protein expression of Raf kinase inhibitor protein (RKIP) is a prognostic biomarker in prostate cancer Another group of the same university predicted that the high expression of EZH2 and ECAD was statistically significantly associated with prostate cancer recurrence after radical prostatectomy	[57,58]
NextBio	*NextBio Clinical*: Semantic based integration of the proprietary OMICs data with public knowledge to get better insight leads to discovery of drug targets and biomarkers Discover and validate stratification biomarker to a therapy accessing genomic data from cell lines, stem cells, animal models and retrospective analysis of clinical trials *NextBio Research*: Identifying crucial pathways leads to a disease phenotype supported by cross studies and multiple data points Identification of disease biomarker and analysis of pharmacokinetic profiles or toxicity indications It uses proprietary algorithms to rank the search outcomes based on the statistical significance of the correlation supported by bioset data points See: http://www.nextbio.com/b/nextbioCorp.nb	Using the NextBio platform Walia *et al.* reported that loss of breast epithelial marker hCLCA2 (chloride channel accessory protein) promotes higher risk of metastasis	[59]
Selventa	Discovery of predictive response biomarkers by reverse engineering disease mechanisms a priori from molecular patients data (OMICs data) It utilizes an extensive and manually curated knowledge base containing literature-derived triples encoded into BEL It identifies disease- and tissue-specific biomarker content that can match targeted therapies to subpopulation of patients Reverse Causal Reasoning (RCR) algorithm is used for identification of master regulators	Very recently, Selventa has introduced its openBEL framework for biomarker discovery based on mechanistic causal reasoning and demonstrated its application in stratifying responders to ulcerative colitis drug, infliximab, from non-responders based on identification of IL6 as the biomarker for alternative disease mechanisms in non-responders	[60]

TABLE 3 (*Continued*)

Knowledge base	Features	Applications	Refs
tranSMART	A knowledge management platform enabling integration of the OMICs data with published literature, clinical trial outcome and established knowledge from Metacore, Ingenuity IPA, National Laboratory of Medicine, US (NLM) The applications of this platform include making novel hypothesis, validating them, disease association of certain pathways, genes, SNPs and biomarker discovery http://www.transmartproject.org/	Analysis of transcriptomic data from melanoma patients using k-means clustering facility in tranSMART showed that the expression levels of ***cyclin D1*** increase from benign to malignant, whereas in metastatic melanomas the expression level decreases, clearly delineating multiple subgroups of samples in the presumably homogenous metastatic melanoma cohort	[61]
KegArray	A microarray gene expression and metabolomics data analysis tool from KEGG Able to map OMICs data to KEGG Pathways, Brite and genome maps See: http://www.kegg.jp/kegg/download/kegtools.html	KegArray was used to investigate metabolic pathways associated with the marker metabolites that were detected by 2D gas chromatography mass spectrometry in tissues from 31 patients with colorectal cancer. The results led to the identification of chemically diverse marker metabolites and metabolic pathway mapping suggested deregulation of various biochemical processes	[62]

TABLE 4

Summary of cons and pros of biomarker-related knowledge bases

Advantages	Disadvantages
Evidence-supported data content	Poor annotation of metadata and incompleteness
Structured data representation	Lack of standard representation model
Enhanced retrieval and retention of information	Lack of flexible filtering criteria
Focused semantic context	Divergent in content and subject focus

of an annotated knowledge base and analytical tools so that performing a full-fledged functional analysis is facilitated.

Table 3 provides an overview of knowledge bases summarizing their main features as well as published examples of their application in biomarker discovery.

Although all these databases contain manually curated knowledge, their differences in the coverage and granularity of the information reflects underlying differences in methodology of information retrieval, variability of the resources used for knowledge extraction as well as the difference in interpretation of the experimental results by the annotators. Shmelkov *et al.* have recently carried out a comparative analysis on quality and completeness of human regulatory pathways among ten public and commercial pathway knowledge bases and found out that surprisingly there is little overlap in the knowledge content of these databases [63]. The authors reported that the only exception was the MetaCore pathway database whose content was validated in 84% of the cases with experimental results, compared to the low overlap of 24% obtained from the KEGG database.

Beside the issues of coverage and quality, the lack of consistent standard schemas for biomarker classification and biomarker knowledge representation has hampered literature searches for information about biomarkers. The fact that qualification of translational biomarkers requires a wide range of information on the level of sensitivity, specificity, the mechanisms of action, toxicity and clinical performance, emphasizes the need for standardization of biomarker vocabularies and classification. Recently, a prototypical process has been suggested for creating evidentiary standards for biomarkers and diagnostics to ensure qualification of biomarkers based on seven types of scientific evidence [64]. Similarly, the Pistoia Alliance, which was initially established by information experts from several pharmaceutical companies, has launched a project focusing on development of an ontological and data standards for integrating biomarker assay data and handling different endpoints [Pistoia Alliance: http://www.pistoiaalliance.org/]. Although in preliminary stages, such developments can form the basis for future biomarker standardization efforts. Therefore, next-generation knowledge bases should address above challenges by introducing efficient information retrieval/extraction tools as well as biomarker data standards. Taken all together, there are both advantages and disadvantages associated with existing knowledge bases, which are summarized in Table 4.

The resolution and quality of knowledge bases are largely dependent on the granularity of underlying ontology, quality of data retrieval and experience of annotators. Creation and maintenance of manually curated knowledge bases is becoming a tremendous task in times of ever accelerating speed of publication growth different from the steady slow process of manual curation. To exemplify an effort, a recent report shows that assembling a compendium of potential biomarkers for pancreatic cancer, which was carried out by systematic manual curation of the literature, took over 7000 person hours [65]. In the absence of automated methods for retrieval of biomarker information, the slow pace of manual curation cannot guarantee that the current content of knowledge bases is comprehensive and sufficient for functional interpretation of OMICs data. Novel high throughput text-mining approaches are absolutely essential for automated biomarker knowledge processing. In the next section we describe automated biomarker information retrieval methods that can be used in support of systematic update of knowledge bases and acceleration

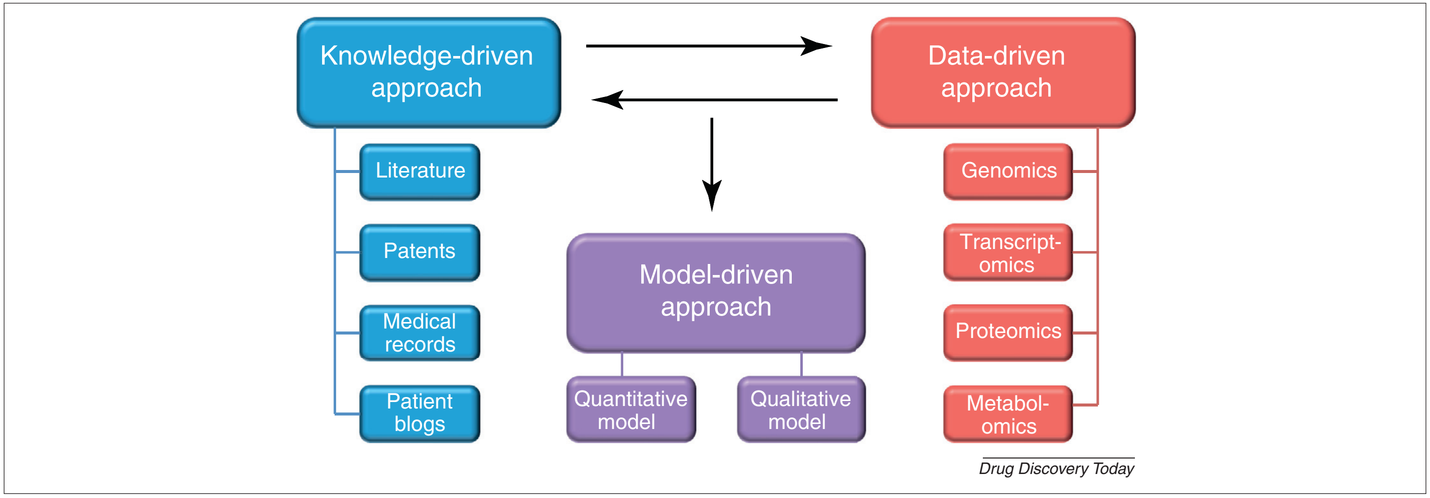

FIGURE 3

Model driven approach combining data-driven and knowledge-driven approaches for the identification of candidate biomarkers.

of the biomarker-related information extraction from the unstructured text.

Text-mining strategies for retrieval and extraction of biomarker information

To accelerate the speed of the curation process, emerging state-of-the-art information retrieval and extraction technologies are under active development. Behind such tools, there are text-mining algorithms that automatically recognize potential biomarkers, such as genes and proteins in text in a process called 'named entity recognition' or NER [66,67]. However, existing NER approaches are not sufficiently selective for the retrieval of biomarker-related content information (such as its association with drug or disease) from the literature. Consequently, studies on biomarker relation extraction from text have considered the extraction of semantic relations between named entities such as the relation between diseases and genes or proteins [68,69]. Some efforts have been recently dedicated to mining and extraction of such relationships using semantically enhanced methods [70,71]. One limitation of these approaches is that they do not take into consideration additional textual features, besides disease and gene names, that represent quality properties of candidate biomarkers such as measurement evidence and technique. In an attempt to overcome this limitation, Ongenaert and Dehaspe (2010) have employed different keyword lists containing terms that specify methylation biomarkers in cancer and used them in conjunction with gene names from GeneCards to generate the methylation database in cancer, PubMeth [72].

As a step in this direction, we have recently developed a dedicated biomarker terminology organized in six proposed classes and used it for information retrieval and extraction of biomarker knowledge embedded in the literature [73]. It was demonstrated that the application of this dedicated biomarker terminology could enhance the retrieval performance significantly through combined search for cancer-related genes and selected classes of the biomarker retrieval terminology. Further evaluation of this terminology in an independent disease area, namely Alzheimer's disease, showed that not only well-known biomarkers were retrieved successfully but also new biomarker candidates could be identified. Integration of such terminologies into search tools supporting semantic and ontological search can reduce the high number of unspecific search results and improve the retrieval rate of informative documents.

Ultimately, context-sensitive biomarker information extracted from literature can be used for automated enrichment of knowledge bases and/or combined with OMICs data may generate a basis for integrated models of disease or drugs mode of action with the aim of prospective prediction of candidate biomarkers [74].

Discussion

Currently pathway analytics and knowledge bases represent a useful tool for the interpretation of OMICs data, identification of upstream mechanistic drivers as well as visualization of OMICs data assisting scientific understanding of the underlying biological processes. Certain limitations of conventional pathway analytics hinder the use of knowledge bases as predictive tools for biomarker discovery and were recently reviewed by Butte [75]. The majority of pathways accumulated in the knowledge bases represent a mixture of findings described in different healthy and pathological conditions in various biological systems and tissues. Creation of the tissue, treatment or condition specific pathways is a challenge and is currently in focus of many commercial knowledge bases providers. Since today's knowledge bases transform multiple transcripts and SNPs to Entrez Gene ID in pathway representation, granularity of the pathways should be further improved for the analysis of RNA and DNA-sequencing-derived OMICs data. Finally, existing knowledge bases contain only static information that represents 'snapshots' of the systems behavior for particular condition under which the data has been obtained. Pathway interdependencies reflecting the sequence of events in pathological processes is not really captured thereby limiting their use for modeling and prediction. A number of systems biology approaches based on quantitative modeling has been suggested to be of use for biomarker prediction reviewed by Kreeger *et al.* [76]. However, labor-intensive collection of quantitative data as well as limitation of current computational power to model complex biological systems containing over hundred molecules hinders current use of quantitative modeling for biomarker prediction.

Qualitative modeling approaches can provide an alternative for prospective biomarker prediction. Quite a few qualitative modeling approaches are based on boolean networks and able to simulate the dynamics of signaling pathways. They have been employed for the discovery of novel oncological biomarkers as well as used to develop robust clinical treatment decisions [77]. An example of another type of qualitative modeling is BEL (Biological Expression Language)-based causal network modeling approach that integrates the literature-derived 'cause and effect' relationships into an integrated biomarker discovery platform [60,78]. Because none of the today's approved biomarkers were predicted in prospective studies there is no current proof-of-concept modeling approach for biomarker prediction. With the emerging technologies and ever developing computational biology approaches the field has an immense opportunity for future development (Fig. 3).

Concluding remarks

A variety of OMICs technologies have been developed in recent years with the aim to contribute detailed understanding of disease pathophysiology and drug mode of action. However neither OMICs data nor the knowledge accumulated in the text can be automatically translated into clinical advances. Knowledge capturing technologies combined with pathway analytics provide a great framework for OMICs data interpretation. The lack of standardized translational algorithms allowing the use of OMICs data along with the knowledge derived from the scientific literature hampers endeavours to predict biomarkers with greater confidence. Thus, any improvement of the current situation depends on the improvements in knowledge representation standards enabling us to present dynamic interconnectivity of the molecular pathways supported by integration of strong signal from experimental data and enriched granular knowledge.

Given recent initiatives to address the issue of biomarker data quality and exchange, it is expected that next-generation biomarker knowledge bases, with enhanced data quality standards and improved data interoperability, play a major role in the future efforts on integrative biomarker identification. The overall trend indicates that there is a drive away from correlative biomarkers towards causative biomarkers. Therefore, the aim of next-generation integrative models is to capture causal relationships between the candidate biomarker and clinical outcome. This will in the near future lead to a new paradigm that engages quantitative and qualitative modeling for prospective prediction of biomarkers.

References

1 Triggle, D.J. (2003) Medicines in the 21st century OR pills, politics, potions, and profits: where is public policy? *Drug Dev. Res.* 59, 269–291

2 Lederberg, J. (1996) Medicine's old battle against the bugs isn't over at all. *Int. Herald Tribune*

3 Hitchings, G.H. (1993) Health care and life expectancy. *Science* 262, 1632

4 Munos, B. (2009) Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* 8, 959–968

5 Arrowsmith, J. (2011) Phase II failures: 2008–2010. *Nat. Rev. Drug Discov.* 10, 328–329

6 Elias, T. *et al.* (2006) Why products fail in phase III. *In Vivo* 24

7 Frueh, F.W. *et al.* (2008) Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy* 28, 992–998

8 Ptolemy, A.S. and Rifai, N. (2010) What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. *Scand. J. Clin. Lab. Invest. Suppl.* 242, 6–14

9 Majewski, I.J. and Bernards, R. (2011) Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nat. Med.* 17, 304–312

10 Trusheim, M.R. *et al.* (2011) Quantifying factors for the success of stratified medicine. *Nat. Rev. Drug Discov.* 10, 817–833

11 Sawyers, C. (2004) Targeted cancer therapy. *Nature* 432, 294–297

12 Paez, J.G. *et al.* (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500

13 Kreitman, R.J. (2006) Immunotoxins for targeted cancer therapy. *AAPS* 8, 532–551

14 Moroni, M. *et al.* (2005) Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study. *Lancet Oncol.* 6, 279–286

15 Lièvre, A. *et al.* (2006) KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res.* 66, 3992–3995

16 Eberhard, D.A. *et al.* (2005) Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *J. Clin. Oncol.* 23, 5900–5909

17 Amado, R.G. *et al.* (2008) Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. Clin. Oncol.* 26, 1626–1634

18 Druker, B.J. *et al.* (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* 344, 1031–1037

19 Heinrich, M.C. *et al.* (2003) Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J. Clin. Oncol.* 21, 4342–4349

20 Kwak, E.L. *et al.* (2010) Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* 363, 1693–1703

21 Bhattacharya, S. and Mariani, T.J. (2009) Array of hope: expression profiling identifies disease biomarkers and mechanism. *Biochem. Soc. Trans.* 37, 855–862

22 van 't Veer, L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536

23 Paik, S. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* 351, 2817–2826

24 Moroni, M. *et al.* (2005) Somatic mutation of EGFR catalytic domain and treatment with gefitinib in colorectal cancer. *Ann. Oncol.* 16, 1848–1849

25 Phillips, G.D.L. *et al.* (2008) Targeting HER2-positive breast cancer with trastuzumab-DM1, an antibody-cytotoxic drug conjugate. *Cancer Res.* 68, 9280–9290

26 Saltz, L.B. *et al.* (2004) Phase II trial of cetuximab in patients with refractory colorectal cancer that expresses the epidermal growth factor receptor. *J. Clin. Oncol.* 22, 1201–1208

27 Vanhoefer, U. *et al.* (2004) Phase I study of the humanized antiepidermal growth factor receptor monoclonal antibody EMD72000 in patients with advanced solid tumors that express the epidermal growth factor receptor. *J. Clin. Oncol.* 22, 175–184

28 Lynch, T.J. *et al.* (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* 350, 2129–2139

29 Freeman, D.J. *et al.* (2009) Activity of panitumumab alone or with chemotherapy in non-small cell lung carcinoma cell lines expressing mutant epidermal growth factor receptor. *Mol. Cancer Ther.* 8, 1536–1546

30 Rifai, N. *et al.* (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* 24, 971–983

31 Niemeyer, C.M. *et al.* (2005) Immuno-PCR: high sensitivity detection of proteins by nucleic acid amplification. *Trends Biotechnol.* 23, 208–216

32 McDermed, J.E. *et al.* (2012) Nucleic acid detection immunoassay for prostate-specific antigen based on immuno-PCR methodology. *Clin. Chem.* 58, 732–740

33 Ward, P.S. and Thompson, C.B. (2012) Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell* 21, 297–308

34 Krutovskikh, V.A. and Herceg, Z. (2010) Oncogenic microRNAs (OncomiRs) as a new class of cancer biomarkers. *Bioessays* 32, 894–904

35 Chang-Gong, L. *et al.* (2008) MicroRNA expression profiling using microarrays. *Nat. Protoc.* 3, 563–578

36 Janoueix-Lerosey, I. (2008) Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* 455, 967–970

37 Bekaii-Saab, T. *et al.* (2009) A multi-institutional phase II study of the efficacy and tolerability of lapatinib in patients with advanced hepatocellular carcinomas. *Clin. Cancer Res.* 15, 5895–5901

38 Chapman, P.B. *et al.* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* 364, 2507–2516

39 Kaminski, M.S. *et al.* (2005) 131I-tositumomab therapy as initial treatment for follicular lymphoma. *N. Engl. J. Med.* 352, 441–449

40 Niu, C. *et al.* (1999) Studies on treatment of acute promyelocytic leukemia with arsenic trioxide: remission induction, follow-up and molecular monitoring in 11 newly diagnosed and 47 relapsed acute promyelocytic leukemia patients. *Blood* 94, 3315–3324

41 Hazarika, M. *et al.* (2008) Tasigna for chronic and accelerated phase Philadelphia chromosome–positive chronic myelogenous leukemia resistant to or intolerant of imatinib. *Clin. Cancer Res.* 14, 5325–5331

42 Takei, H. *et al.* (2008) Multicenter phase II trial of neoadjuvant exemestane for postmenopausal patients with hormone receptor-positive, operable breast cancer: Saitama Breast Cancer Clinical Study Group (SBCCSG-03). *Breast Cancer Res. Treat.* 107, 87–94

43 Tsao, M.S. *et al.* (2005) Erlotinib in lung cancer – molecular and clinical predictors of outcome. *N. Engl. J. Med.* 353, 133–144

44 Addo, S. *et al.* (2002) A phase I trial to assess the pharmacology of the new oestrogen receptor antagonist fulvestrant on the endometrium in healthy postmenopausal volunteers. *Br. J. Cancer* 87, 1354–1359

45 Hochhaus, A. *et al.* (2008) Dasatinib induces durable cytogenetic responses in patients with chronic myelogenous leukemia in chronic phase with resistance or intolerance to imatinib. *Leukemia* 22, 1200–1206

46 Dang, N.H. *et al.* (2007) Phase II trial of denileukin diftitox for relapsed/refractory T-cell non-Hodgkin lymphoma. *Br. J. Haematol.* 136, 439–447

47 Slamon, D.J. *et al.* (2001) Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* 344, 783–792

48 Jagarlapudi, S.A. and Kishan, K.V. (2009) Database systems for knowledge-based discovery. *Methods Mol. Biol.* 575, 159–172

49 Brentnall, T.A. *et al.* (2009) Proteins that underlie neoplastic progression of ulcerative colitis. *Proteomics Clin. Appl.* 3, 1326–1339

50 Chen, R. *et al.* (2007) Quantitative proteomics analysis reveals that proteins differentially expressed in chronic pancreatitis are also frequently involved in pancreatic cancer. *Mol. Cell Proteomics* 6, 1331–1342

51 Fichtner, I. *et al.* (2008) Establishment of patient-derived non-small cell lung cancer xenografts as models for the identification of predictive biomarkers. *Clin. Cancer Res.* 14, 6456–6468

52 Andersen, J.N. *et al.* (2010) Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors. *Sci. Transl. Med.* 2, 43–55

53 Chang, H. *et al.* (2011) Effect of neoadjuvant ixabepilone (ixa) on cell cycle genes and tumor-initiating cell (TIC) signature in breast cancer (BC). *J. Clin. Oncol.* 29

54 Faber, P.W. *et al.* (2008) Potential non-synonymous single nucleotide polymorphisms (nsSNPs) associated with toxicity in metastatic clear cell renal cell carcinoma (MCCRCC) patients (pts) treated with sunitinib. *J. Clin. Oncol.* 26

55 Wong, Y.F. *et al.* (2007) Identification of molecular markers and signaling pathway in endometrial cancer in Hong Kong Chinese women by genome-wide gene expression profiling. *Oncogene* 26, 1971–1982

56 Ruan, L. *et al.* (2011) Analysis of EGFR signaling pathway in nasopharyngeal carcinoma cells by quantitative phosphoproteomics. *Proteome Sci.* http://dx.doi.org/10.1186/1477-5956-9-35

57 Fu, Z. *et al.* (2006) Metastasis suppressor gene Raf kinase inhibitor protein (RKIP) is a novel prognostic marker in prostate cancer. *Prostate* 66, 248–256

58 Rhodes, D.R. *et al.* (2003) Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. *J. Natl. Cancer Inst.* 95, 661–668

59 Walia, V. *et al.* (2012) Loss of breast epithelial marker hCLCA2 promotes epithelial-to-mesenchymal transition and indicates higher risk of metastasis. *Oncogene* 31, 2237–2246

60 Laifenfeld, D. *et al.* (2012) Early patient stratification and predictive biomarkers in drug discovery and development: a case study of ulcerative colitis anti-TNF therapy. *Adv. Exp. Med. Biol.* 736, 645–653

61 Szalma, S. *et al.* (2010) Effective knowledge management in translational medicine. *J. Transl. Med.* http://dx.doi.org/10.1186/1479-5876-8-68

62 Mal, M. *et al.* (2012) Metabotyping of human colorectal cancer using two-dimensional gas chromatography mass spectrometry. *Anal. Bioanal. Chem.* 403, 483–493

63 Shmelkov, E. *et al.* (2011) Assessing quality and completeness of human transcriptional regulatory pathways on a genome-wide scale. *Biol. Direct* 6, 15

64 Altar, C.A. *et al.* (2008) A prototypical process for creating evidentiary standards for biomarkers and diagnostics. *Clin. Pharmacol. Ther.* 83, 368–371

65 Harsha, H.C. *et al.* (2009) A compendium of potential biomarkers of pancreatic cancer. *PLoS Med.* http://dx.doi.org/10.1371/journal.pmed.1000046

66 Pennings, J.L. *et al.* (2009) Discovery of novel serum biomarkers for prenatal Down syndrome screening by integrative data mining. *PLoS One* http://dx.doi.org/10.1371/journal.pone.0008010

67 Deng, X. *et al.* (2006) Link test – a statistical method for finding prostate cancer biomarkers. *Comput. Biol. Chem.* 30, 425–433

68 Bundschus, M. *et al.* (2008) Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* http://dx.doi.org/10.1186/1471-2105-9-207

69 Elkin, P.L. *et al.* (2009) BioProspecting: novel marker discovery obtained by mining the bibleome. *BMC Bioinformatics* http://dx.doi.org/10.1186/1471-2105-10-S2-S9

70 Islam, M.T. *et al.* (2010) Biomarker Information Extraction Tool (BIET) development using natural language processing and machine learning. In *Proceedings of the International Conference and Workshop on Emerging Trends in Technology* pp. 121–126, ACM

71 Jessen, W. *et al.* (2012) Mining PubMed for biomarker-disease associations to guide discovery. *Nat. Proc.* http://dx.doi.org/10.1038/npre.2012.6941.1

72 Ongenaert, M. *et al.* (2008) PubMeth: a cancer methylation database combining text mining and expert annotation. *Nucleic Acids Res.* 36, D842–D846

73 Younesi, E. *et al.* (2012) Mining biomarker information in biomedical literature. *BMC Med. Inform. Decis. Making* 12, 148

74 Butcher, E.C. *et al.* (2004) Systems biology in drug discovery. *Nat. Biotechnol.* 22, 1253–1259

75 Khatri, P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* http://dx.doi.org/10.1371/journal.pcbi.1002375

76 Kreeger, P.K. and Lauffenburger, D.A. (2010) Cancer systems biology: a network modeling perspective. *Carcinogenesis* 31, 2–8

77 Sahoo, D. (2012) The power of boolean implication networks. *Front Physiol.* http://dx.doi.org/10.3389/fphys.2012.00276

78 Martin, F. *et al.* (2012) Assessment of network perturbation amplitude by applying high throughput data to causal biological networks. *BMC Syst. Biol.* http://dx.doi.org/10.1186/1752-0509-6-54