

Uncharacterized/hypothetical proteins in biomedical ‘omics’ experiments: is novelty being swept under the carpet?

Krzysztof Pawłowski

Advance Access publication date 19 July 2008

Abstract

Many ‘omics’ studies, gene expression microarray experiments in particular, aim at charting the molecular mechanisms of physiology, disease and drug response. This short review discusses the bias present in many such studies whereas the focus is set on the well understood and established molecular scenarios. The under-reporting rate of ‘hypothetical’ or uncharacterized genes and proteins, differentially regulated in disease context, is assessed here. Reasons for this bias are discussed. Particular examples from the genomics studies on respiratory diseases are presented. This review aims at increasing awareness of the unexplored genomics data and proposes remedies in order to refocus genomics studies on the less-charted territories of the genome, transcriptome and proteome. It is suggested that routine use of function prediction methods in conjunction with omics analyses may allow better interpretation of the data, and facilitate discovery of true novelty.

Keywords: *gene expression analysis; hypothetical proteins; uncharacterized genes; biomedicine; function prediction*

INTRODUCTION

The field of functional genomics can be viewed as an example of the broader omics domain, encompassing also proteomics. The advent of gene expression microarray technology [1–3] has provided the medical field with an extremely promising toolbox. Numerous studies, using patient material and model systems, both animal and human cell culture, have been conducted in recent years with the aim of charting yet unknown and unexpected molecular mechanisms of physiology, disease and drug response [4–7]. Still, such studies often focused on already well understood and established molecular scenarios. Thus, the exciting opportunity to expand molecular analyses beyond the known and expected has rarely been attempted.

Of note, hypothetical genes and proteins, uncharacterized as yet by experiment, comprise a substantial part of the human genome in general, and the ‘druggable’ genome in particular. This review probes recent microarray literature describing gene

expression regulation in human disease. The focus is set on ‘novel’ genes (coding uncharacterized/hypothetical proteins) identified in such studies as having biomedically relevant expression patterns. Estimates are attempted of the under-reporting rates of those genes, i.e. the gap between the number of author-reported and the actually measured uncharacterized genes of disease-regulated expression.

Biomedically relevant uncharacterized genes of expression varying in disease are often almost ‘willingly’ overlooked. This review aims at increasing awareness of those genes, and of potential benefits of routine exploration of their roles in disease.

WHAT CONSTITUTES NOVELTY IN BIOMEDICAL OMICS EXPERIMENTS? HOW MUCH ‘NOVELTY’ IS THERE WITHIN THE HUMAN GENE SET?

Many an article within the functional genomics trade boasts ‘novel genes’ implicated in a disease, condition,

Corresponding author. Krzysztof Pawłowski, Nencki Institute of Experimental Biology, PAS, ul. Pasteura 3, 02-093 Warszawa, Poland and Warsaw University of Life Sciences, Warszawa, Poland. Tel: +48 22 5892 222; Fax: +48 22 822 53 42; E-mail: k.pawlowski@nencki.gov.pl

Krzysztof Pawłowski is on the faculty of Nencki Institute of Experimental Biology in Warsaw, Poland’s major biological research institution. His group applies bioinformatics and systems biology to problems of potential medical interest. Previously he provided bioinformatics support to preclinical research at AstraZeneca.

disease model, etc. However, more often than not, such a statement actually refers to finding a novel functional relationship between an already well-known gene and a disease rather than discovery of unique functional facts for a hitherto uncharacterized gene.

The definition of a ‘novel gene’, usually a loose one, often requires that the gene’s expression be not confirmed before. For example, the Ensembl system (www.ensembl.org) distinguishes ‘known’, ‘novel’ and ‘predicted’ genes [8], featuring varying amounts of experimental evidence supporting the correspondence of a given genomic region to functional protein-encoding transcripts. For the purpose of the current brief review, however, we shall focus on genes being ‘characterized’ as opposed to ‘known’ only. We shall consider a gene as ‘characterized’ if there is experimental evidence regarding the molecular function of its protein product, or, at least, experimental evidence regarding the product’s function within biological processes (cf. Gene Ontology main categories [9, 10]). Therefore, sole confirmation of expression of a gene will not be treated herein as functional characterization sufficient for the gene to be called ‘characterized’. Although it is straightforward to manually label a single gene as characterized or not, it is not as obvious how to perform such labeling automatically, on a genomic scale. Table 1 lists several estimates of uncharacterized gene numbers in the human genome, each inaccurate and possibly underestimated. Twenty-four percent of Ensembl ‘known’ genes have no links to the curated SwissProt protein database [11], however even among SwissProt proteins many are uncharacterized. Twenty-two percent of Entrez genes [12] have no links to PubMed articles, however many papers may only refer to validating expression,

not investigating function. Thirty-one percent of Entrez genes have no Gene Ontology annotations, however many functional annotations are automated and relying on predictions only. If the presence of a Pfam [13] structural domain linked to an Ensembl gene is used as a working definition of a ‘characterized’ gene, then a somehow higher proportion is obtained. However, one should bear in mind that a Pfam structural/functional domain present in protein sequence is by no means an indication that the gene has been characterized. Often, the Pfam (or other) domain present allows only a prediction of a function. Further, many Pfam domains are not linked to known functions, and often are denoted as ‘DUF’ (Domain of Unknown Function). Thus, a ‘guesstimate’ of the rate of uncharacterized ‘novel’ protein-coding genes in the human genome can be set to at least 40%. By extrapolation, any other biomedically relevant genome should have an even larger fraction of uncharacterized genes. Herein, we shall discuss how and *if* this definitely non-negligible portion of the human genome is explored in the omics experiments.

WHY SWEEPING NOVELTY UNDER THE CARPET? THE TYPICAL BIOMEDICAL OMICS EXPERIMENT

The typical biomedical omics experiment aims at relating the clinical manifestations of disease and pathology to the underlying molecular mechanisms. Such a study is usually a ‘joint venture’ involving the biologist, the medic and the statistician, and preferably, also the bioinformatician. As such, it is usually a compromise between the different scientific approaches. The statistician is interested in the numerical data analysis. The medic focuses on

Table 1: The ‘uncharacterized gene’ and ‘hypothetical protein’ concepts

Database queried	Approximate definition of ‘characterized’ protein coding genes	Number of genes	Approximate definition of ‘total’ protein coding genes	Number of genes	Uncharacterized (%)
Ensembl database at BioMart	As in Column 4 plus gene has associated Swissprot ID	16 283	Gene status = ‘known’ and gene has protein ID	21 418	24
Ensembl database at BioMart	As in Column 4 plus gene has associated Pfam ID	16 306	Gene status = ‘known’ and gene has protein ID	21 418	24
EntrezGene database at SRS–EBI	As in Column 4 plus gene has PubMed link	20 024	Gene status = ‘live’ and gene has protein description	25 635	22
EntrezGene database at SRS–EBI	As in Column 4 plus gene has Gene Ontology annotation	17 733	Gene status = ‘live’ and gene has protein description	25 635	31

Entrez at www.ncbi.nlm.nih.gov and SRS at srs.ebi.ac.uk queried on May 20, 2008; Ensembl at www.biomart.org queried on Oct 9, 2007.

relationships between the molecular measurements and the patient data or analogous medical data in model systems. The biologist is looking primarily at known biological pathways and relations emerging from the data analysis. None of these main stakeholders is strongly interested in investigating the molecular data that are hard to interpret in the light of the current biological knowledge, i.e. data on novel/uncharacterized genes and proteins. The bioinformatician, if present on the author team, is often charged with elucidating the expression relationships between known genes rather than analysing the less-known and unknown ones.

A striking example of the general trend is provided by the usage of two sister microarrays from a leading manufacturer, Affymetrix, namely the HG-U133A and HG-U133B chips. Both were designed and manufactured at the same time and represented, colloquially speaking, two ‘halves’ of the human genome. HG-U133A contained probe sequences designed for the better characterized genes while HG-U133B contained probes for many ‘hypothetical genes’ and EST sequence clusters. As of writing this manuscript (May 2008), the public repository of microarray data, Gene Expression Omnibus at NCBI [14], contains 547 ‘gene expression series’ (GSE) records for the U133A chip and only 92 such entries for the U133B chip. It can be said then that the *omics* community is 6-fold more interested in the expression patterns of the known parts of the genome as compared to the unknown parts. Affymetrix microarrays are invoked here only because, for reasons specific to the design and manufacturing process of that particular manufacturer, they allow a separation of the genome representation on the microarray into two ‘halves’ annotated functionally in a very uneven way.

In practice, investigation of the *omics* data, e.g. microarray gene expression data, is skewed by the focus on *the known* molecular mechanisms. In a biomedical setting, it can be argued that it is more straightforward and useful to relate disease information to the known rather than *the unknown* biology. This means preferentially focusing on shorter term benefit which could be, for example, exploration of treatment ideas in a model or clinical setting, or looking for alternative drug targets in a well-studied signalling pathway, rather than slowly, incrementally and painfully building the understanding of molecular basis of the disease mechanisms.

WHY SWEEPING NOVELTY UNDER THE CARPET? REASONS FOR MISTRUST AND RESERVATION

A common reason for avoiding ‘novel’ findings in omics experiments is the so-called ‘biological scenario’ focus where the investigators decide to narrow down their study to some particular biological mechanism or signalling pathway. An analogous attitude is common in cases where immediate practical applications are desired. Then investigators would prefer to analyse well-understood potential drug targets or biomarkers.

On the other hand, some studies seem to be of the ‘*proof-of-concept*’ kind only. There, the researchers would validate the use of omics technologies in a particular biomedical setting by trying to reproduce the results obtained already before using more established techniques.

In the initial days of omics studies, reservations in the exploration and reporting of the results may have originated from the distrust in (automated) genomic annotations of the microarray probesets. In other cases, limitations in exploring novel genes may be explained by the lack of belief/trust in high-throughput experiments in general. Reliability and reproducibility issues in DNA microarray studies have since the beginnings of the field been of high concern [15, 16]. Only recently, however, the understanding of these has improved and confidence in cross-platform and cross-laboratory reproducibility has increased [17–19]. Thus, many authors, especially in earlier years, would rather keep their eyes set on those results that could be rationalized by previous knowledge than act on evidence about completely uncharacterized entities. Also, accuracy of targeted expression studies (e.g. PCR) is usually more trusted than that of genome-wide ones [20].

WHY SWEEPING NOVELTY UNDER THE CARPET? WAYS AND STATS OF REPORTING (OR HIDING) NOVELTY

As shown in an over-simplified literature survey performed in PubMed and summarized in Table 2, only 1.4% of gene expression-related literature highlights observation of novel/uncharacterized genes. This sounds extremely little considering the fact that uncharacterized genes constitute 40–50% of the human genome. This very inaccurate literature search is using only approximate definitions of

Table 2: Uncharacterized genes in omics literature

Query terms used for genes	Query terms used for omics	Number of articles	Uncharacterized/novel as fraction of omics (%)
Uncharacterized gene, novel gene and hypothetical protein	Gene expression profiling, gene expression, microarray analysis and transcription profiling	1591	1.4
—	Gene expression profiling, gene expression, microarray analysis and transcription profiling	116 821	
New target(s), new drug target(s), novel target(s) and novel drug target(s)	Gene expression profiling, gene expression, microarray analysis and transcription profiling	596	0.5
Uncharacterized gene, novel gene and hypothetical protein	—	4795	
New target(s), new drug target(s), novel target(s) and novel drug target(s)	—	3519	

PubMed queried October 9, 2007. Non-review articles published between 2000 and 2007 queried. 'gene expression profiling', 'gene expression' and 'microarray analysis' queried as MeSH terms, other expressions queried in All Fields.

omics literature. Also, since it searches titles and abstracts only, it does not account for uncharacterized genes that might be discussed in the main text or reported in the tables. Still, it captures the general trend. Of note, this quick survey also suggests that as much as one-third of all literature citing uncharacterized genes is related to genomics technologies. This means that genomics is still a very important tool for functional characterization of those genes. Incidentally, although genomics is expected to aid in drug target discovery, only 0.5% of gene expression-related literature reports novel drug targets (again, counted only if mentioned in abstracts/titles). Here, interestingly, only 17% of articles claiming new/novel drug targets in titles or abstracts correspond to expression profiling which reflects the prevalence of other, more important sources of drug target ideas. As a result, information regarding potential involvement of novel genes in biological processes is not 'hidden', but it is not highlighted and utilized, either.

In this section, we shall further take a look at a small number of recent biomedical functional genomics studies, selected by personal interest of the author, namely those related to respiratory diseases. We start with articles most reluctant towards exploiting uncharacterized genes.

The work of Crystal and co-workers [21] focused on gene expression in the airway epithelium of smokers. They used microarray experiment to assess 'levels of expression of 44 antioxidant-related genes' only. Strikingly, the authors focused on an *a priori* defined narrow gene group and pushed aside the data on remaining twenty plus thousand genes.

A similarly focused (or, as one might criticize, biased) approach has been undertaken by

Djukanovic *et al.* [22] who investigated oxidative stress in chronic obstructive pulmonary disease. These authors, however, broadened their pre-defined gene set under investigation by unbiased sequence-based functional predictions, to several hundred genes.

Follettie *et al.* [23] studied gene expression in a murine model of asthma. They discuss scores of genes with differential expression; however, no mention is made whatsoever of any uncharacterized genes. It can be assumed that such genes have been omitted from analysis and discussion.

Ross and colleagues [24] study mucociliary differentiation in human airway epithelium. In a data table, they cite 108 differentially expressed genes including 6 uncharacterized ones and mention 54 'genes/ESTs without known function or homology' while omitting any reference to the details of their expression changes.

In contrast to the work cited above, Woodruff and co-workers [25] adopt a highly recommendable approach. While analysing airway epithelial gene expression in asthma and in treatment response to corticosteroids, they report in an unbiased manner microarray probesets with differential expression, regardless of their annotation status. Thus, even probesets not linked to any official gene symbols or Entrez gene IDs are listed in a data table. Six out of twenty-two probesets listed there correspond to hypothetical proteins/uncharacterized genes. Also, while discussing the results, they bring up in the text a protein of unknown function with a peculiar expression pattern.

An approach keen on exploration of novelty may be exemplified by the work on the Fizz1

gene where the original microarray discovery of a hypoxia-inducible gene in lung [26] has been extended into discovery of an important inter-cellular signalling protein [27]. This may constitute an example as to what could be done to use the insight obtained from the large-scale omics experiments

Problems inherent in functional interpretation of microarray experiments are exemplified in an elegant recent paper reporting a meta-analysis of several asthma-related microarray datasets [28]. The authors utilize the module networks approach [29] to elucidate asthma-relevant gene expression themes recurring in different experiments. Such an approach provides more reproducible and generalized biomedically relevant observations, and strengthens the case for several genes with ‘novel role in asthma induction’. However, the focus of this interesting article which is seemingly *known* asthma mechanisms, makes it loose from sight uncharacterized genes that, although not explicitly excluded from analyses, are absent from the examples discussed.

A field more sympathetic to the uncharacterized genes is that of disease biomarker search where there is no strong requirement to focus on genes with clear functional link to the disease process. An example is a report that highlights a gene of poorly understood function, CUB-domain containing protein 1 (CDCP1) [30], where, microarray analysis identified CDCP1 as ‘a potential prognostic marker’ in renal cell carcinoma.

ARE SYSTEMS BIOLOGY APPROACHES HELPING TO EXPLORE NOVELTY? ANALYSIS OF GENE SETS AND PATHWAYS

Many systems biology tools, both academic and commercial [31–36], have appeared in recent years allowing mapping of gene sets emerging from *omics* experiments onto the existing body of biological knowledge, as derived from literature and biological pathway databases. These tools provide interpretation of differentially expressed gene sets in terms of biochemical and signalling pathways, regulatory modules and over-represented functional annotations. Systems biology approaches, although invaluable in the interpretation of the deluge of data, may in some hands pose a risk of diverting the researcher from analysis of the uncharacterized genes towards

only those that are backed by substantial literature-derived information.

Thus, literature knowledge-based tools for omics data analysis and interpretation should preferentially be complemented by the use of advanced *in silico* tools for function prediction [37–39], in an iterative interplay of expression data analysis, literature data analysis and functional predictions. Individual genes of interesting expression patterns may be subject to individual experimental validation and ‘manual’ computational analyses and predictions. However, of more general interest are the methods that allow large-scale function prediction. Traditional homology-driven annotation transfer has been recently improved by annotation analysis of the homologues [40]. A new group of methods in the systems biology spirit, aims at exploiting various experimentally derived relationship networks (e.g. protein–protein interactions, co-expression, genetic interactions, regulatory) to elucidate functional predictions [41, 42]. Various classification methods allow general function predictions utilizing ‘homology-free’ protein sequence features [43–45]. A well-established function prediction approach relies on the analysis of a protein’s 3D structure [46, 47]. This approach is nowadays helped by the progress in protein structure prediction methods [48]. Many protein structure prediction methods are available as web servers [49], although this does not apply to some of the most resource-demanding methods. Systematic comparison of structure-based function predictions is now part of the CASP structure prediction competition [50]. Yet another dimension of complementary approaches to function prediction is offered by the attempts at direct inference of regulatory and signaling pathways from expression data [51]. These methods, if used routinely, should allow a more complete and unbiased interpretation of the omics data, leading to suggestions for experimental validation of predicted mechanisms, targets and biomarkers.

NOVELTY: UNDER-USED OR ABUSED?

An examination of the usage of the term ‘novel’ in titles of articles indexed in PubMed (Figure 1) allows for an interesting observation. The percentage of such articles rose from <0.01% before 1960 to >1% in the current decade. Undoubtedly, such a change reflects changing in the language usage rather than

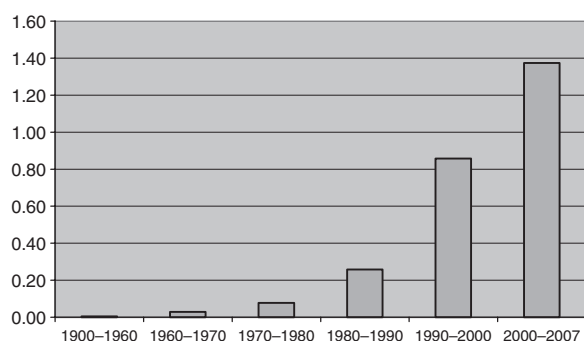


Figure 1: Percentage of PubMed articles claiming 'novel' in the title.

change in productivity or focus of the biomedical and life sciences. Still, it well reflects the authors' growing interest in having their research perceived as focused on novelty. Also this author is not guilt-free, since he has to admit authoring one article where the first word in the title says 'novel'. Having this novelty excitement in mind, it would be recommendable for the *omics* trade to really brave the unknown and 'novel' areas of the genome, transcriptome and proteome. This crude and naive chart (Figure 1) serves rather as an illustration that the 'novelty' term is somewhat abused while pursuance of novelty *per se* is not endaeuvoured as widely as it could be.

Remembering, that among the reasons to embark on *omics* experiments in bio-medicine essential is the search for novel drug targets, novel mechanisms of action of known drugs, novel diagnostic markers [52], it would be really reasonable not to miss a chance to find *really* novel targets, mechanisms and markers. The very appropriate concept of *inferential literacy* [53] should be extended to the understanding of opportunities the genome offers in terms of hypothetical proteins. The bias against novelty may be also remedied by the emerging approach to perform meta-analyses on repositories of expression data, often under-utilized by the original authors [54, 55]. There, recurring expression features of uncharacterized genes may demand closer inspection and follow-up. In conclusion, the intention here has not been to stigmatize the focus on known genes and proteins within 'omics' as wrong. Rather, it is argued, that it could be beneficial to complement the *omics* analyses with usage of various predictive tools allowing interpretation of high-throughput data also for the uncharacterized genes and proteins.

Key Points

- Omics studies, in particular functional genomics ones, are often biased towards *a priori* hypothesized scenarios.
- Large portions of omics datasets, involving uncharacterized genes, are often skipped and left unused.
- This bias against novelty may be overcome by increasing awareness of the problem and by emerging tools for function prediction and global dataset querying.

References

1. Spellman PT, Sherlock G, Zhang MQ, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998;**9**(12):3273–97.
2. Dehouck C, Goodfellow PN. DNA microarrays in drug discovery and development. *Nat Genet* 1999;**21**(Suppl 1):48–50.
3. Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov* 2002;**1**(12):951–60.
4. Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer* 2007;**7**(7):545–53.
5. Sears C, Armstrong SA. Microarrays to identify new therapeutic strategies for cancer. *Adv Cancer Res* 2007;**96**: 51–74.
6. Wang S, Cheng Q. Microarray analysis in drug discovery and clinical applications. *Methods Mol Biol* 2006;**316**:49–65.
7. Shai RM. Microarray tools for deciphering complex diseases. *Front Biosci* 2006;**11**:1414–24.
8. Curwen V, Eyraas E, Andrews TD, *et al.* The Ensembl automatic gene annotation system. *Genome Res* 2004;**14**(5):942–50.
9. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology consortium. *Nat Genet* 2000;**25**(1):25–9.
10. Thomas PD, Mi H, Lewis S. Ontology annotation: mapping genomic regions to biological function. *Curr Opin Chem Biol* 2007;**11**(1):4–11.
11. Boutet E, Lieberherr D, Tognolli M, *et al.* UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol Biol* 2007;**406**:89–112.
12. Maglott D, Ostell J, Pruitt KD, *et al.* Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007;**35**(Database issue):D26–31.
13. Finn RD, Mistry J, Schuster-Bockler B, *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;**34**(Database issue):D247–51.
14. Barrett T, Troup DB, Wilhite SE, *et al.* NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res* 2007;**35**(Database issue):D760–5.
15. Nimgaonkar A, Sanoudou D, Butte AJ, *et al.* Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* 2003;**4**:27.
16. Hess KR, Zhang W, Baggerly KA, *et al.* Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol* 2001;**19**(11):463–8.

17. Hayes DC, Secrist H, Bangur CS, *et al.* Multigene real-time PCR detection of circulating tumor cells in peripheral blood of lung cancer patients. *Anticancer Res* 2006; **26**(2B):1567–75.
18. Draghici S, Khatri P, Eklund AC, *et al.* Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* 2006; **22**(2):101–9.
19. Shi L, Perkins RG, Fang H, *et al.* Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr Opin Biotechnol* 2008; **19**(1):10–8.
20. Brazeau DA. Combining genome-wide and targeted gene expression profiling in drug discovery: microarrays and real-time PCR. *Drug Discov Today* 2004; **9**(19):838–45.
21. Hackett NR, Heguy A, Harvey BG, *et al.* Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. *Am J Respir Cell Mol Biol* 2003; **29**: 331–43.
22. Pierrou S, Broberg P, O'Donnell R, *et al.* Expression of genes involved in oxidative stress responses in airway epithelial cells of COPD smokers. *Am J Respir Crit Care Med* 2007; **175**(6):577–86.
23. Follettie MT, Ellis DK, Donaldson DD, *et al.* Gene expression analysis in a murine model of allergic asthma reveals overlapping disease and therapy dependent pathways in the lung. *Pharmacogenomics J* 2006; **6**(2):141–52.
24. Ross AJ, Dailey LA, Brighton LE, *et al.* Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am J Respir Cell Mol Biol* 2007; **37**(2):169–85.
25. Woodruff PG, Boushey HA, Dolganov GM, *et al.* Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. *Proc Natl Acad Sci USA* 2007; **104**(40):15858–63.
26. Teng X, Li D, Champion HC, *et al.* FIZZ1/RELMalpha, a novel hypoxia-induced mitogenic factor in lung with vasoconstrictive and angiogenic properties. *Circ Res* 2003; **92**(10):1065–7.
27. Chung MJ, Liu T, Ullenbruch M, *et al.* Antiapoptotic effect of found in inflammatory zone (FIZZ)1 on mouse lung fibroblasts. *J Pathol* 2007; **212**(2):180–7.
28. Novershtern N, Itzhaki Z, Manor O, *et al.* A functional and regulatory map of asthma. *Am J Respir Cell Mol Biol* 2008; **38**(3):324–36.
29. Segal E, Shapira M, Regev A, *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003; **34**(2):166–76.
30. Awakura Y, Nakamura E, Takahashi T, *et al.* Microarray-based identification of CUB-domain containing protein 1 as a potential prognostic marker in conventional renal cell carcinoma. *J Cancer Res Clin Oncol* 2008, May 16, [Epub ahead of print].
31. Troyanskaya OG. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform* 2005; **6**(1):34–43.
32. von Mering C, Jensen LJ, Snel B, *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 2005; **33**(Database issue):D433–7.
33. Sivachenko AY, Yuryev A, Daraselia N, *et al.* Molecular networks in microarray analysis. *J Bioinform Comput Biol* 2007; **5**(2B):429–56.
34. Ekins S, Nikolsky Y, Bugrim A, *et al.* Pathway mapping tools for analysis of high content data. *Methods Mol Biol* 2007; **356**:319–50.
35. Werner T. Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol* 2008; **19**(1):50–4.
36. Ganter B, Giroux CN. Emerging applications of network and pathway analysis in drug discovery and development. *Curr Opin Drug Discov Devel* 2008; **11**(1):86–94.
37. Ofra Y, Punta M, Schneider R, *et al.* Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov Today* 2005; **10**(21):1475–82.
38. Godzik A, Jambon M, Friedberg I. Computational protein function prediction: are we making progress? *Cell Mol Life Sci* 2007; **64**(19,20):2505–11.
39. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. *J Bioinform Comput Biol* 2007; **5**(1):1–30.
40. Gotz S, Garcia-Gomez JM, Terol J, *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 2008; **36**:3420–35.
41. Zhang S, Jin G, Zhang XS, *et al.* Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics* 2007; **7**(16):2856–69.
42. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007; **3**:88.
43. Jensen LJ, Gupta R, Staerfeldt HH, *et al.* Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* 2003; **19**(5):635–42.
44. Zhu F, Han LY, Chen X, *et al.* Homology-free prediction of functional class of proteins and peptides by support vector machines. *Curr Protein Pept Sci* 2008; **9**(1):70–95.
45. Lobley AE, Nugent T, Orengo CA, *et al.* FFPred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Res* 2008, May 7, [Epub ahead of print].
46. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* 2005; **15**(3):275–84.
47. Adams MA, Suits MD, Zheng J, *et al.* Piecing together the structure-function puzzle: experiences in structure-based functional annotation of hypothetical proteins. *Proteomics* 2007; **7**(16):2920–32.
48. Baker D. Prediction and design of macromolecular structures and interactions. *Philos Trans R Soc Lond B Biol Sci* 2006; **361**(1467):459–63.
49. Fischer D. Servers for protein structure prediction. *Curr Opin Struct Biol* 2006; **16**(2):178–82.
50. Lopez G, Rojas A, Tress M, *et al.* Assessment of predictions submitted for the CASP7 function prediction category. *Proteins* 2007; **69**(Suppl 8):165–74.
51. Lezon TR, Banavar JR, Cieplak M, *et al.* Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc Natl Acad Sci USA* 2006; **103**(50):19033–8.

52. Jayapal M, Melendez AJ. DNA microarray technology for target identification and validation. *Clin Exp Pharmacol Physiol* 2006;**33**(5,6):496–503.
53. Miron M, Nadon R. Inferential literacy for experimental high-throughput biology. *Trends Genet* 2006;**22**(2):84–9.
54. Hibbs MA, Hess DC, Myers CL, *et al.* Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 2007;**23**:2692–9.
55. Cahan P, Rovegno F, Mooney D, *et al.* Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene* 2007;**401**(1,2):12–8.