From the Center for Genomics and Bioinformatics
Karolinska Institutet, Stockholm, Sweden

GENOMICS AND BIOINFORMATICS APPROACHES TO FUNCTIONAL GENE ANNOTATION

Danielle Kemmer

Stockholm 2006

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden

ISBN 91-7140-636-0

ABSTRACT

Biomedical research has been undergoing a quasi-revolution with the dawn of the genomics era. The flood of sequence data from the various genome projects, the task of cataloging the entire coding portion of a genome instead of identifying and characterizing individual genes, as well as technical demands accompanying these developments have posed great challenges to the research community. Although the entire human genome sequence has been virtually recorded, fundamental issues remain about the precise number of protein coding genes, as well as their functional characterization.

Available resources for the study of human gene function include large genome annotation pipelines, expression profiling data, and protein interaction screens. To gain biological insights from this maze of data, one must both find mechanisms to organize the information and assess the quality of the results.

This thesis focuses on the functional annotation of sparsely characterized human genes and their encoded proteins. The work includes four stages:

I. Gene expression profiling

II. Assessment of the level of characterization of human genes

III. Projection of protein networks from lower eukaryotes onto human

IV. Integration of computational and experimental results for data mining.

Initially, a cross-platform comparison for a set of gene expression profiling techniques was carried out to compare the performance of cutting-edge high-throughput methods and conventional approaches in terms of sensitivity, reliability, and throughput. In this study, we demonstrated that correlation between the different methods was poor and thus multi-technique validation was justified. Nonetheless, the strongest correlation between the new reference data in our report, i.e., a collection of traditional Northern blots, was observed with microarray-based technologies.

The assessment of the level of functional characterization of human genes was addressed in the second study, where we developed a scoring system to quantify the annotation status of each human gene. We created a metric to effectively predict the characterization status of human genes based on a set of predictors from the GeneLynx database[1]. This scoring function will not only assist the targeted analysis of groups of sparsely annotated genes and proteins, but will prove itself useful in the monitoring of long-term gene annotation efforts and the overall annotation status of the human genome.

Comparative genomics efforts to transfer gene annotation from proteins in amenable model organisms onto human proteins are currently restricted by the limited availability of experimental data. Nonetheless, we demonstrated how protein networks could be effectively projected from lower eukaryotes onto human and how the confidence in these projections increased with redundantly detected protein interactions. This so-called Interolog Analysis offers promise for reliable inference of protein function. The

bioinformatics system we created (Ulysses) provides a novel intuitive interface for biologists studying human proteins. As data depth and coverage will increase over time, this system will prove to be valuable in the extended prediction of high-confidence functional associations of a large portion of human genes.

The fusion of experimental data and computational predictions is a central goal of functional genomics. We constructed a bioinformatics workbench for the study of uncharacterized human gene families. By assembling bioinformatics resources and experimental results in a common space, the NovelFam3000 system facilitates functional characterization. Working with a collection of uncharacterized genes, we demonstrated how bioinformatics methods can lead to novel inferences about cellular function of specific protein families.

This thesis unites the identification of uncharacterized human genes, the assessment of genomics data quality, and the application of high-throughput data for the inference of protein function.

ORIGINAL PUBLICATIONS

I

Kemmer D., Faxén M., Hodges E., Lim J., Herzog E., Ljungström E., Lundmark A., Olsen M.K., Podowski R., Sonnhammer E.L.L., Nilsson P., Reimers M., Lenhard B., Roberds S.L., Wahlestedt C. Höög C., Agarwal P., and Wasserman W.W.
Exploring the Foundation of Genomics: A Northern Blot Reference Set for the Comparative Analysis of Expression Profiling Techniques.
*Comparative and Functional Genomics **5**, 584-595* (2004)

II

Podowski R.M., **Kemmer D.**, Brumm J., Wahlestedt C., Lenhard B., Wasserman W.W.
Gene Characterization Index: A Metric for Accessing How Well We Understand Our Genes.
Manuscript

III

Kemmer D., Huang Y., Shah S.P., Lim J., Brumm J., Yuen M.M.S., Xu T., Wasserman W.W., Ouellette B.F.F.
Ulysses – an Application for the Projection of Molecular Interactions Across Species.
*Genome Biology **6**, R106* (2005)

IV

Kemmer D., Podowski R., Lim J., Arenillas D., Hodges E., Roth P., Sonnhammer E.L.L., Höög C., Wasserman W.W.
NovelFam3000 – Uncharacterized Protein Domains Conserved Across Model Organisms.
***Submitted** to BMC Genomics (2005)*

CONTENTS

LIST OF ABBREVIATIONS

cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
E-Northern	Electronic Northern
EST	Expressed sequence tag
GCI	Gene characterization index
GFP	Green fluorescent protein
GPCR	G-protein-coupled receptor
mRNA	Messenger ribonucleic acid
MARS	Multivariate adaptive regression splines
MS	Mass spectrometry
NCBI	National Center for Biotechnology Information
ORF	Open reading frame
PCC	Pearson correlation coefficient
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction
SAGE	Serial analysis of gene expression
SNP	Single nucleotide polymorphism
SVM	Support vector machines
TAP	Tandem affinity purification

PREAMBLE

Current status of the human genome

In April of 2003 the international research community celebrates two milestone anniversaries: completion of the comprehensive sequence of the human genome[2] and the 50th year of the discovery of the double-helical structure of DNA[3]. The global significance of these scientific achievements has to be appreciated in the context of historical accomplishments that have lead to the genomics era (Figure 1): the discovery of the basic principles of heredity by Gregor Mendel[4], the identification of DNA as the physical carrier of heredity[5], the elucidation of DNA structure[3] and the genetic code[6], as well as the development of recombinant DNA technologies[7] and DNA sequencing methods[8-10].

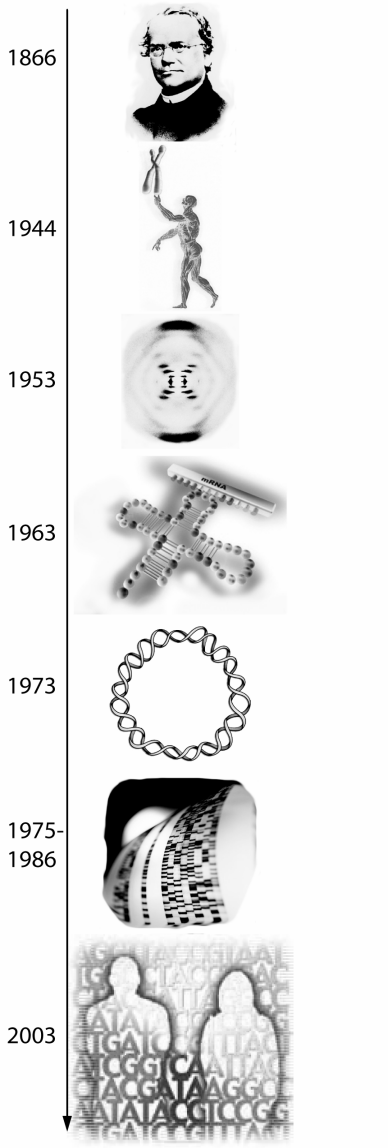

Figure 1. Path of gene research through history. 1866 - Mendel delineates the principles of heredity; 1844 - Avery and colleagues demonstrate that DNA is the material basis of heredity; 1953 - Watson, Crick, and Franklin discover the chemical structure of DNA; 1963 - the genetic code is cracked; 1973 - recombinant DNA technology is introduced; 1975 to 1986 - DNA sequencing techniques are developed; 2003 - the human genome is comprehensively sequenced.

Despite the fact that near complete sequences have so far been obtained only for a few metazoans, including our own species[11, 12], a nematode[13], the fruit fly[14], and the mustard weed[15], draft sequences of a large number of species, including multiple eukaryotes and prokaryotes, have revolutionized biomedical research. Not only are we provided with a publicly accessible highly refined human sequence[16, 17], but we have access to the genetic, physical and transcript maps of many organisms. From a technical perspective, challenges of the new research era have resulted in the development of an array of genome-scale methods including high-throughput oligonucleotide synthesis[18], DNA microarrays[19], normalized and subtracted cDNA libraries[20], whole-genome knockouts (yeast)[21], and integrated yeast-two-hybrid mappings[22].

The emersion of the human genome sequence holds important promises for biomedical research. Through revelation of hereditary factors of diseases, the genome sequence leads humanity into an era of personalized medicine. More immediate for researchers, the high-quality sequence changes our focus from gene discovery to functional gene annotation.

Characterization of all proteins encoded by the human genome, the human proteome, is one of the key goals of functional genomics and proteomics. Despite the recent closure of the genome, a final gene count is still not definite, and numbers for protein-coding sequences are oscillating around ~ 25,000[23, 24]. Many of these genes are uncharacterized, and most encode proteins within which there are segments of unknown function.

Originally, a novel gene was defined as a protein that had not been reported in a database. The latest release statistics of Ensembl, the standard for automated gene annotation in eukaryotes, suggests coverage exceeding 95% known genes in the human genome[25]. Thus, novelty for many now refers to genes lacking holistic descriptions of their cellular function.

Gene annotation in functional genomics and proteomics

We have entered the next phase of the Human Genome Project (HGP) which is to catalogue, characterize, and understand the entirety of functional elements encoded in the human genome[26]. Protein-coding genes and their products are a principal class of functional elements studied in functional genomics. Since gene products don't execute their cellular tasks independently, but rather function in complex cellular networks, further study of biological pathways and protein complexes is essential for a more complete insight in their cellular function. By combining experimental and computational approaches, we control a set of tools to confirm active transcription of predicted human genes, to constrict their site of activity to specific cellular compartments, and to accelerate functional characterization by integrating experimental data from a wide range of model organisms.

Gene expression profiling

To fully comprehend a gene's activity and biological role by determining where and when it is expressed, and to exclude "dead" pseudogenes from functional characterization efforts, it is essential to confirm active cellular transcription of predicted genes. Gene expression profiling is often based on the assessment of mRNA levels of a gene in a specific tissue. Depending on the nature of the method applied, measured mRNA abundance can be representative for the strength of a gene's expression in a particular cell type.

During the genomics era, gene expression profiling techniques have undergone a quasi revolution. The traditional procedure for the analysis of single gene expression at the mRNA level are Northern blots[27, 28]. Quantitative polymerase chain reaction (PCR) has supplemented Northerns for such studies, but remains a low-capacity approach[29, 30]. With the dawn of expression analyses on a whole genome scale, high-throughput approaches to evaluate "transcriptomes", i.e., the collection of genes that are transcribed from genomic DNA, has become a priority.

Experimental methods supplemented with bioinformatics have proven effective for generating large volumes of gene expression data. The availability of genome sequences and the growing knowledge of all encoded genes have ignited the development of new approaches. The creation of tissue-specific cDNA libraries and the successive sequencing of the cDNAs generated expressed sequence tags (ESTs) representative of the expressed mRNA population[31, 32]. The original EST approach, i.e., the first sequencing-based method to measure gene expression on a large scale, was modified and improved in the creation of serial analysis of gene expression (SAGE)[33].

For SAGE, short sequence tags are used and their transcript sources are determined. Quantitation of the number of times a particular tag is observed reflects the expression level of the corresponding transcript. SAGE can be used to identify known genes as well as new genes and thus, allows for further gene discovery.

More recently, microarrays capable of providing expression profiles of tens of thousands of genes in parallel have become available. For hybridization, cDNA sequence portions of known genes are either spotted down[19, 34] or synthesized on the slide[35, 36]. DNA micorarrays allow for the exploration of patterns of gene expression on a global scale and have become commonplace across a variety of scientific fields. Challenges associated with this technology consist of collecting, managing, and analyzing the bulk of data generated from each hybridization experiment.

Sub-cellular protein localization

The progression from genomics to proteomics, i.e., the analysis of the protein complement within a cell, has been challenging, notably because of the biochemical differences between DNA and proteins. Each fragment of DNA behaves biochemically much like any other and the scaling of DNA sequencing and hybridization methods have been achievable. Proteins, on the other hand, have unique properties, and such individuality creates enormous obstacles in terms of standardized technologies.

Proteome analyses consist of the systematic study of all the proteins in a given cell, including all protein isoforms and modifications, the interactions occurring among them, structural descriptions, as well as their cellular localization. To date, sub-cellular localization of the entire proteome of a single eurkaryote, *Saccharomyces cerevisiae*, has been reported[37, 38]. Protein localization is assumed to be a strong indicator for gene function. Distinct compartmental organization of the eukaryotic cell delineates processes occurring within the boundaries of these membranes, and as such provides a means to delimit cellular function. Several bioinformatics tools have been developed to predict protein localization within a cell based on signal sequences for protein sorting, sequence homology, and phylogenetic profiles[39-41]. However, due to their deficient accuracy such methods can be suggestive at the most, and experimental approaches are required for reliable determination of sub-cellular localization.

A strategy to systematically localize proteins on a large scale uses microscopy to visualize the location of tagged forms of expressed recombinant proteins. Cloning of open reading frames (ORFs) of target genes for subsequent protein localization involves systematic amplification of full-length cDNAs by PCR. The PCR products are then inserted into appropriate expression vectors containing either a reporter gene[38, 42-44] or an epitope tag[37, 45]. Although the availability of high-capacity recombination-based cloning systems[46] has facilitated the tagging of ORFs for expression[47] (Figure 2), extending the approach to monitor localization of all proteins in metazoan cells remains a challenging task. For optimal analysis, proteins are ideally expressed in their cells of origin to place them in their natural environment and thus, subject them to native modifications and characteristic interaction partners. The "complexity" of multi-cellular animals exceeds the cellular organization of the yeast cell both in terms of multiple splice variants for a single gene and the diversity of cell types, each with its own specialized function and specific protein make-up.

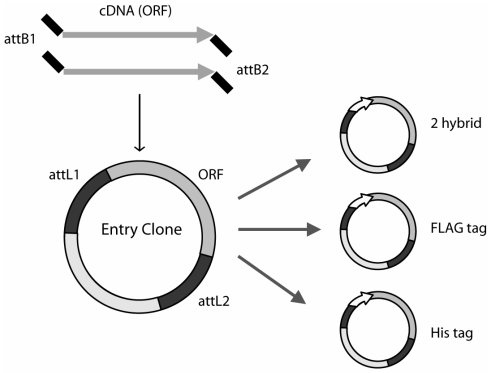

Figure 2. Gateway™ recombination-based cloning technology. The open reading frame (ORF) of a gene is PCR-amplified with primers containing recombinase recognition sites. Enzyme-mediated recombination inserts the gene of interest into an entry clone. The entry clone can then be used to transfer the insert into multiple expression vectors to facilitate the characterization of the protein product.

Protein networks

Protein interactions

A powerful method to deduce protein function is to identify interacting partners in the cell, as interacting proteins are often part of a common protein complex or pathway. Several high-throughput methods have emerged for delineation of interactions or associations including yeast two-hybrid systems[48, 49] and protein complex purification[50, 51]. These experimental approaches appear to be particularly powerful: the two-hybrid system detects binary interactions *in vivo*[48, 49]; protein complex purification uses affinity tags and, coupled with protein identification by mass spectrometry (MS), allows for the simultaneous detection of co-purified complex members[50, 51].

Yeast two-hybrid method

The yeast two-hybrid system[52] is a genetic method for the identification and analysis of protein-protein interactions (Figure 3). It relies on the modular nature of eukaryotic transcription factors, which contain both a site-specific DNA-binding domain and a transcriptional-activation domain that activates the transcriptional machinery. The system uses ORFs fused either to the binding or activation domain of the GAL4 transcription factor. Increased transcription of a reporter gene results when proteins encoded by two ORFs interact in the nucleus of the yeast cell. In the typical practice, a protein of interest fused to the DNA-binding domain ("bait") is screened against a library of activation-domain hybrids ("preys"). Once the positive interaction is detected via the reporter gene product, the ORF is identified by sequencing the relevant clones.

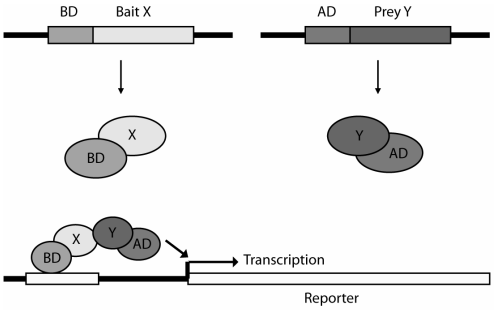

Figure 3. Yeast two-hybrid system. Protein-protein interactions are detected by measuring transcription of a reporter gene. If protein X and protein Y interact, the DNA-binding domain (BD) and activation domain (AD) combine to form a functional transcriptional activator, which proceeds to transcribe the reporter gene.

For these reasons, the yeast two-hybrid system is a simple and generic method amenable to high-throughput screening. Despite its great sensitivity and flexibility, the yeast two-hybrid system poses disadvantages, particularly non-negligible portions of false results (both negatives and positives). These include membrane and secretory proteins non-amenable to a nuclear-based detection system, proteins that activate transcription, proteins that fail to fold correctly in yeast, and interactions based on excluded domains and/or post-translational modifications.

Protein complex purification

MS-based protein interaction assays can be divided into three consecutive steps: 1) bait presentation, 2) affinity purification of the protein complex, and 3) analysis of the bound proteins (Figure 4). A generic strategy is to tag the proteins of interest with a sequence readily recognized by an antibody specific for the tag. We distinguish between tags supporting single-step purification[50], a convenient method with considerable yield, and tags supporting two sequential affinity steps (tandem affinity purification or TAP)[51, 53]. Identification of the eluted proteins involves mass spectrometric measurements[54]. Identified peptides, the typical output of a proteomic

experiment, are compiled in a list and further compared to matches with various sequence databases.

Although dual-step purification significantly reduces background noise, it probably results in the loss of transient and weak binding partners during the purification procedure. Another problem is the occurrence of false positives through non-specific binding to "sticky" proteins. Compared with the two-hybrid approach, complex purification strategies have the advantage to allow for full protein processing and modification, since the interactions take place in the native cellular environment. In addition, multi-component complexes can be analyzed in a single experiment. However, since many biological relevant interactions are of low affinity, transient and dependent on the environment, MS-based approaches will only detect a fraction of the actually occurring protein associations.

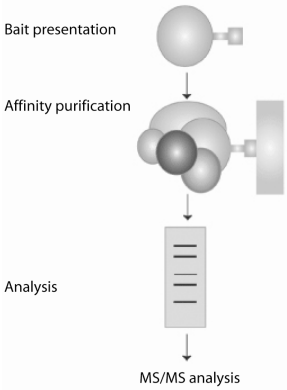

Figure 4. Mass spectrometry-based protein complex purification. A tagged, known protein binds to its unknown interacting protein partners. The resulting protein complex is recognized by antibodies attached to a solid support. After purification, the complex is released and its components are separated. Individual proteins are identified by mass spectrometry.

Data resources

With the internet as a means to globally share scientific knowledge, the access to scientific literature has become a topic of lively debate. The dawn of the open access movement[55] has resulted in requirements for large scale projects to publicly share their content. The exponential growth of biological knowledge over the past few years has lead to the development of critical bioinformatics resources (Table 1).

Table 1. Examples of leading biological databases.

Name	Description
PubMed[56]	Searchable compendium of biological literature
Ensembl[25]	Software system which produces and maintains automatic annotation on selected eukaryotic genomes
UCSC Genome Browser[57]	Genome annotation display for multiple species
Saccharomyces Genome Database (SGD)[58]	Database of the molecular biology and genetics of budding yeast
FlyBase[59]	Database of the *Drosophila* genome
WormBase[60]	The *Caenorhabditis elegans* model organism database
Gene Ontology (GO)[61]	Database of controlled vocabulary for molecular gene function, biological process, and cellular component

Besides leading species-specific databases containing mainly genetic and genomic information, a number of data repositories have been created collecting data from specific types of experiments. There have been several attempts to establish comprehensive interaction maps for principal model organisms like yeast[48-51, 62, 63], fly[64], worm[65], and recently human[66]. In consequence, a number of databases were created to manage the large influx of interaction information originating from high-throughput proteomics projects (Table 2).

Large-scale interaction datasets are scattered across various publicly available depositories and thus underutilized for gene characterization. Integrated analysis is

difficult, since datasets are not systematically linked. To alleviate this problem, the General Repository for Interaction Datasets (GRID)[67], a generic interaction database, was created[68]. Besides data collection in a common space, comprehensive analysis is greatly facilitated by data visualization. Interaction data maintained by the GRID can be graphically visualized and manipulated by the software tool Osprey[69]. Similar visualization tools have been created, including Cytoscape[70], VisANT[71], and TopNET[72].

Table 2. Leading protein interaction databases.

Name	Description
Biomolecular Interaction Network Database (BIND)[73, 74]	Information about biomolecular interactions, molecular complexes, and pathways including high-throughput and hand-curated interaction information from the literature.
Database of Interacting Proteins (DIP)[75, 76]	Report of binary protein-protein interactions curated both manually by expert curators and automatically using computational approaches.
Munich Information Center for Protein Sequences (MIPS-GSF)[77, 78]	Comprehensive database of protein-protein interactions in *S. cerevisiae* and high-quality protein interaction data collection from mammals.
The Molecular INTeraction database (MINT)[79, 80]	Stores interactions between biological molecules focusing on experimentally verified protein interactions in mammals.
IntAct[81, 82]	Toolkit for the storage, presentation, and analysis of protein interaction data.
Human Protein Interaction Database (HPID)[83, 84]	Assembles predicted protein interactions in human.

In addition to data repositories and individual graphics systems, more advanced applications for the integration of biomolecular interaction networks with other data types have been developed. Examples such as STRING[85], OPHID[86], and POINT[87] feature network predictions combined with experimentally observed protein interactions.

Network analysis

Data-driven genome-scale analyses of gene and protein networks have received ample attention with increased heterogeneous raw interaction data becoming available. The data collection phase has been accompanied by analysis of network topology, which resulted in the discovery of biological networks being small world, scale-free, and modular[88, 89]. In protein-protein interaction networks most proteins interact with few partners, while a small proportion interact with many partners forming biologically significant "hubs". These networks have a high degree of local clustering and biological processes or modules are organized by the "hubs".

Comparisons between different experimental large-scale approaches to monitor protein interactions have shown little overlap in their reported associations[90] pointing to the dynamic nature of the proteome and suggesting either poor specificity or poor coverage. To confirm observed protein associations from high-throughput yeast two-hybrid and complex purification experiments, different types of experimental data have been integrated, including correlated mRNA expression profiles[91, 92], data from genetic interactions[93, 94], as well as *in silico* (computational) interaction predictions[95-97].

One of the first attempts to predict functional relations between gene products and respective networks came from efforts using the genomic context of genes such as conservation and/or co-occurrence of genes, gene adjacency, and gene fusions[98-100]. Recently, refinements of prediction methods have taken into account phylogenetic trees shared by interacting proteins considering evolutionary interrelationships among species[101-103].

For functional associations derived from cellular pathways, anti-correlated distribution patterns of proteins substituting for each other can be detected[104]. This method identifies missing genes replaced by functionally equivalent ones that are otherwise unrelated. Newly identified genes may be functionally associated to all other members of the pathway. Other approaches for the prediction of protein associations are structure-based[105]. Target proteins containing domains occurring in frequently interacting proteins point to protein-protein interactions[106]. Entire protein complexes can be modeled based on the three-dimensional structure of protein sub-units[107].

Data quality

Integration and quality control of high-throughput datasets have become essential for reliable protein network reconstruction. A fundamental problem is that only a small fraction of interactions in networks are confirmed with certainty, and the number of true interactions is considerably larger than those reported. Benchmarks validating the accuracy of protein interactions are a prerequisite for successful data integration. High-quality subsets of protein interactions can sometimes be identified using supplemental criteria, such as the degree to which mRNAs of interacting proteins are co-expressed[108, 109], neighborhood cohesiveness properties[110, 111], shared pathways[112] or sub-cellular localization[113], or combinations of these approaches[114].

The ability to better predict protein interaction networks especially in yeast and increased data curation has led to the development of several benchmark data sets accessible via online databases[77, 115-118]. Using these benchmarks as reference associations, multiple approaches to integration have been tried. Both, simple intersections[97, 119] and unions[120] of separate data sets have been studied. In parallel, more sophisticated probabilistic approaches were developed to predict protein associations[85, 121]. Very recently, there have been attempts to use the context of a protein network to suggest additional interactions[122, 123].

Comparative genomics

Relationships between the genomes of different species can yield insights into many aspects of evolution, especially valuable for the identification of genes and regulatory regions. Comparative genomics supports the transfer of gene annotation among homologous proteins in different organisms. It has been shown that the identification of protein associations to known protein complexes or cellular pathways can give clues as to the cellular role of uncharacterized proteins[92]. The study of protein interactions is thus an important element of functional genomics.

The paucity of protein interaction data in human and many other organisms has naturally led to the question to which extent protein interactions can be transferred between species. The underlying concept is that sequence and structural similarities between conserved proteins suggest functional similarities and that functionally important interactions are maintained over evolution[124, 125]. This idea has been

implemented in several studies comparing protein networks between prokaryotes, lower eukaryotes[65, 126-128], and human[87, 129]. In another study it was shown that besides the projection of key protein interactions across species, evolutionary conserved co-expression patterns could be integrated to reconstruct genetic networks in multiple species[92].

The fundamental prerequisite to allow network comparisons is the identification of homologous proteins across organisms. There exist several phylogenetic classifications of proteins from complete genomes attempting to address orthology, i.e. proteins of conserved function connected through vertical evolutionary descent, and paralogy, i.e. homologous proteins following gene duplication[56, 130-134]. Despite ever more sophisticated algorithms, it remains a challenge to successfully distinguish between the one-to-one, one-to-many, and many-to-many homologous relationships between related proteins across organisms.

PRESENT INVESTIGATION

Functional genome annotation is a complex endeavor and occurs at the interface of a multitude of scientific specialties. Concerted efforts from the research community are rendered possible through fruitful communication between the various fields. While expert contributions are essential to achieve annotation depth, a prerequisite for successful genome annotation is the integration of various types of experimental results. Effective collaboration in the annotation pipeline is facilitated through both knowledgeable coordinators taking on a managing role as well as widely accessible data resources facilitating storage and retrieval of different data types.

From a technical perspective, functional genome annotation can be divided into successive stages that constitute a gene characterization process. The generation of experimental raw data for the bulk of protein-coding genes is a first step. A typical experiment is gene expression profiling to confirm active transcription of predicted genes on the mRNA level. Furthermore, proteomics methods may be employed to investigate proteins directly. In a second stage, the large volume of experimental data is analyzed. For this purpose, bioinformatics algorithms are applied to identify interesting components within the complex data that provide insights into partially characterized genes.

The present investigation focuses on the functional description of human genes and their encoded proteins. The publications presented address different aspects of the gene characterization process:

- Comparison of available gene expression profiling methods in terms of throughput, sensitivity, and reliability
- Assessment of the level of characterization of human genes
- Projection of protein networks from lower eukaryotes onto human
- Facilitated functional annotation by integration of computational resources and experimental data in an accessible common space.

Together these components constitute an in-depth study of human gene annotation.

Paper I: Exploring the Foundation of Genomics: A Northern Blot Reference Set for the Comparative Analysis of Expression Profiling Techniques.

In the initial phase of the thesis project we decided to select a set of scarcely characterized human genes for functional studies (paper IV). Since many of the potential candidate genes were solely based on computational gene predictions, an initial confirmation of active gene transcription was judged to be necessary to avoid the wasteful study of spuriously predicted genes. Our goal was to mimic a genome-scale situation where genes were taken in bulk for expression profiling rather than tested individually. At the time, there had appeared several new technologies for high-throughput gene expression screening[135-137], but few evaluation studies showed the efficacy of these newly developed approaches in terms of sensitivity and reliability compared to conventional methods[138]. There was notably a lack of studies conducting broad platform evaluations, instead pair-wise expression technology comparisons were the norm[139-145].

We therefore selected a diverse set of the most common gene expression profiling technologies to be compared in our laboratory, including both conventional low-throughput approaches and amenable large-scale methods. Since the *de facto* standard to evaluate a gene's expression in a set of tissues has been the traditional Northern blot, we created a database of published Northern results serving as a reference collection for our study[146].

To compare individual methods to the Northern results, we used RNA from the same source and generated expression profiles for selected sets of genes using multiple techniques. The correlation scores indicated that none of the tested methods agreed strongly with the Northern blot data, but that the highest correlations could be observed with the different microarray platforms. We concluded from the results that multi-technique validation was justified to obtain reliable gene expression profiles. Since completion of our work, several similar studies have been reported which support our results[147-149].

Currently, the dbMTN collection is a valuable resource for researchers to assess the performance of expression profiling technologies.

Paper II: Gene Characterization Index: A Metric for Accessing How Well We Understand Our Genes.

Selecting uncharacterized human genes from the completed genome for functional characterization was a prerequisite for critical parts of this thesis project. In order to target our characterization efforts we were interested in the level of functional annotation for both specific gene families (paper IV) and individual human genes as part of the coding body of the entire genome (paper III).

The present paper describes a first attempt to assign a novelty score to each human gene, i.e., to systematically assess the level of functional gene characterization on a large scale. The Gene Characterization Index (GCI) has been generated and applied to human genes via GeneLynx[150], a gene catalogue. The implementation of this scoring system was based on thorough selection of "training" sets of genes from the GeneLynx database and web-based evaluator rating by trained biologists.

To assess the scoring accuracy of our system, we evaluated the predictive value of several types of annotations from the GeneLynx database, including recorded SNPs, protein domains, gene ontology terms, and the number of PubMed abstracts. The predictive power of the chosen predictors was assessed using multivariate adaptive regression splines (MARS)[151] and support vector machines (SVM)[152] by comparing predicted to evaluated (curated) gene annotation scores. Both models performed adequately well in the sense that they reflected gene annotation ratings attributed by biologists. We observed the best performance for least and most characterized genes, while middle range scores were somewhat divergent from scientist ratings.

GCI will prove to be useful in the estimate of the general level of annotation of groups of genes like nuclear receptors, G-protein-coupled receptors (GPCRs), or kinases and will enable the targeted selection of subsets of genes for experimental studies. GCI can also survey total genome novelty by monitoring average levels of functional characterization of the coding sequences of the human genome, thus reflecting efficacy of long-term annotation efforts (Figure 5).

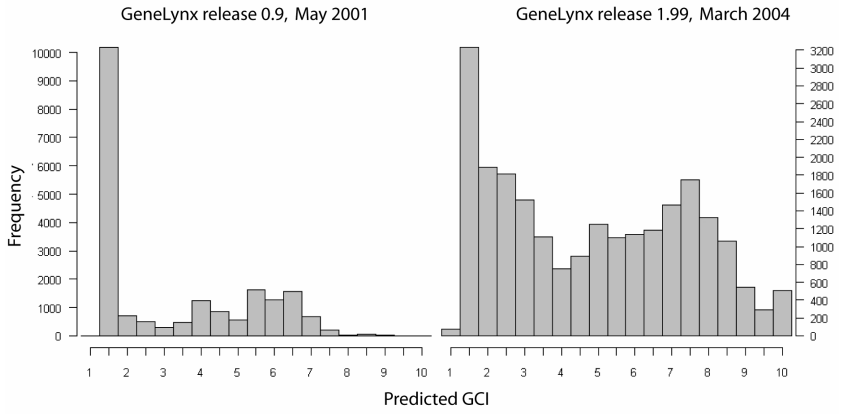

Figure 5. Histogram of GCI scores for GLIDs with cDNAs. Predicted gene characterization index (GCI) scores for human genes with observed cDNAs reported in GeneLynx. Two GeneLynx releases are compared. In March 2004, a significant portion of genes scored higher compared to May 2001 reflecting extended overall functional annotation.
(GLID = GeneLynx identifier)

Paper III: Ulysses – an Application for the Projection of Molecular Interactions across Species.

The mapping of gene-gene relationships across organisms has proven to be a powerful means to predict gene function[128, 129, 153, 154]. It has become possible to project functional gene networks from lower eukaryotes onto human by capitalizing on the body of functional gene annotation from model organisms[49-51, 64, 65] combined with homology mapping across species[130, 132].

In order to efficiently apply core data mining components, data from various resources need to be unified in a manageable common space. Here we report the implementation of the first annotation system for human genes based on the projection of gene networks detected in yeast, worm, and fly. We integrated homology mapping through HomoloGene[56], data management with the data warehouse Atlas[155] and developed a data visualization platform to facilitate biological interpretation. The Ulysses system can be accessed via a web interface[156].

The performance of the underlying algorithm to successfully predict human protein associations was assessed against published reference collections[115, 117, 157]. As data coverage in the different resources is deficient and data overlaps for homologous proteins occur rarely[90], we assessed performance based on the common sub-cellular co-localization of putative interacting proteins. Applying these criteria we could show that genes could be effectively linked to correct networks and that confidence in these associations was considerably increased with redundantly occurring protein interactions. Capitalizing on existing networks, we demonstrated that our system had the capacity to extend previously described cellular pathways and complexes with novel protein associations. We also confirmed its ability to discover discrete networks by reconstructing cellular complexes responsible for biological core functions like mRNA processing, DNA replication, and protein degradation.

As a successful strategy in the functional characterization of human genes on a large scale we described a computational framework to transfer gene associations from the leading model organisms onto human. While we established that this kind of network projection effectively predicts human protein associations and thus confers biological function, the limiting factor for functional inference is the sparse coverage of interactions in publicly available resources. We therefore encourage deepened coverage of genomics data by the research community to take full advantage of the predictive power of the Ulysses system.

Paper IV: NovelFam3000 – Uncharacterized Protein Domains Conserved Across Model Organisms.

While functional gene characterization efforts are often conducted on a gene-by-gene basis[158-160], there exist several efforts to group genes according to shared structural subunits, each carrying out a distinct biochemical function[161-169]. Depending on their biological significance, structural and functional protein domains are often conserved across long evolutionary time and can therefore be found in a number of distantly related species[170]. Most metazoan proteins consist of several distinct domains, and each gene product can be considered as a composite of modular building blocks.

In this study, we systematically grouped gene products according to shared domains. Based on the Pfam database[164], we extracted highly reliable protein domains of unknown function (DUFs) predicted by hidden Markov models, as well as less trusted automatically generated protein domains that were both conserved in at least three organisms (human, worm, and fly). We constructed a database composed of the selected domains and their corresponding proteins in a number of organisms[171].

Our goal was to facilitate gene annotation by providing bioinformatics resources and experimental results for different members of a protein domain family. Through active research community participation and by gathering biological knowledge for a wide variety of proteins sharing a specific domain, we showed how NovelFam3000 could be used to delineate family-specific traits transferable to further family members.

Depending on the species, we provided links to various bioinformatics resources[25, 56, 57, 158, 172]. We combined bioinformatics and experimental annotation strategies to comprehensively characterize sets of genes. We integrated the gene characterization index (GCI, paper II) (Figure 6), species-specific resources[38, 58-60, 150, 173], array-based gene expression results[174], gene association information[175], and links to the Ulysses system[176] (paper III). For a selected set of meagerly annotated family members, we performed targeted laboratory experiments. We focused on RT-PCR-based expression profiling to individually confirm a predicted gene's expression and sub-cellular protein localization to define compartmentally organized protein activity.

We showed that transferable family-consistent results can be obtained with our approach and that the combination of high- and low-throughput bioinformatics and experimental annotation strategies has great potential in the accelerated elucidation of human gene function.

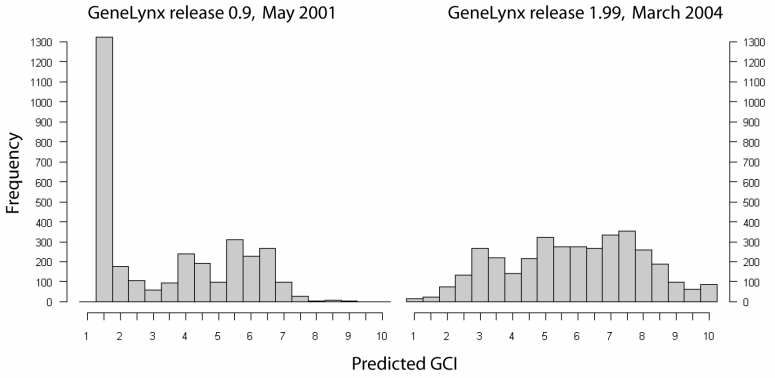

Figure 6. Histogram of GCI scores for NovelFam3000 genes. Predicted gene characterization index (GCI) scores for human genes included in the NovelFam3000 system. Two GeneLynx releases are compared. In May 2001, a large portion of genes scored very low (GCI = 0-2) indicating their sparse functional characterization. In March 2004, most of these novel genes obtained higher GCI scores reflecting successive functional annotation. (GLID = GeneLynx identifier)

CONCLUDING REMARKS

This work developed during an intense and highly dynamic period of the genomics era. We set out when the first coarse draft of the human genome including large gaps and errors was published in 2001 and went through the near complete human genome sequence now available. Our endeavors touched on different downstream efforts to functionally annotate coding portions of the genome by combining experimental approaches and computational strategies. We were challenged by the modification of conventional annotation strategies such as gene expression profiling to adapt to large-scale requirements and a constantly changing genome depiction. Here we will discuss our achievements regarding different topics covered in this thesis, suggest further improvements, as well as point to future directions.

Assessment of gene expression profiling techniques

The comparison of gene expression profiling platforms is limited by a number of factors, such as consistent RNA sources, representation of analogous genes within different technologies, public accessibility of raw expression data, uniform controls across diverse methods, and lack of a universal standard widely accepted in the research community. While there exist publicly accessible repositories of microarray results[177-182], high-throughput gene expression results generated with less common techniques and low-throughput data are independently generated in individual laboratories under a wide range of experimental conditions and protocols.

In our investigation, we introduced a curated collection of published Northern blot results for evaluation and compared conventional small scale methods and accessible large scale approaches to this standard. While we were able to capture the performance of the selected techniques by measuring the expression of defined sets of genes compared to our standard, it was beyond the scope of this study to comprehensively include the multitude of currently available methods discussed in the preamble. For future comparisons, the research community will have to widely share its data to greatly improve the thorough comparison of available gene expression platforms, a process recently initiated and echoed by the open access initiative[183-186]. In close cooperation with this development, universal data standards, as they are currently available for microarray expression results[187], will need to be developed for a wide variety of data types allowing the unlimited exchange of gene expression experiments.

Genome annotation status

We examined gene annotations from two different angles, first by developing a scoring system to assess functional annotation of human genes (paper II), and second by constructing an annotation system for scarcely characterized genes containing novel domains evolutionary conserved across eukaryotes (paper IV).

In paper II, we showed that the human perspective for the level of gene annotation can be effectively estimated by a statistical model. In particular, the number of annotations for each human gene in the GeneLynx database[150] were used to predict the functional annotation score for each gene. Moderately and more extensively annotated genes displayed greater divergence from biologists' ratings and the algorithm may be improved by additional data. As more annotation sources become available, our definition of functional annotation will change over time. In

consequence, the scoring system will need to be adjusted accordingly by conducting further surveys of biologists and retraining the statistical model.

In paper IV, we created NovelFam3000, a database of highly conserved protein domain families with minimal functional annotation. The focus on conserved genes minimizes the risk of spurious computational predictions. We implemented a knowledge framework for integrating various bioinformatics resources combined with user-submitted experimental evidence for individual members of novel domain protein families. Compared to existing data mining tools[158, 188, 189] NovelFam3000 builds on active research community participation and thus, represents a unique platform for gene annotation.

With the example of selected members of domain families, we showed how the combination of user-submitted experimental results and bioinformatics resources may elucidate the cellular function of specific proteins thus providing clues as to the subordinate biological function of the domain family as a whole. In our study, we focused on sub-cellular localization and tissue-specific expression, and compiled a set of computational resources for functional characterization.

To fully capitalize on the range of available functional genomics tools and bioinformatics resources for the comprehensive characterization of a protein domain family, additional experimental approaches and links to external resources need to be implemented. For instance, low-throughput protein interaction experiments and functional assays may further resolve and confirm suggested cellular function. As gene-specific experiments are beyond the scope of large-scale annotation efforts, joining forces between individual laboratories, including extensive data sharing, will be the key to successful systematic genome annotation.

Comparative genomics and network projection

Evolutionary conservation of biologically significant portions of the genome is the underlying principle of comparative genomics[190]. While in theory it is possible to compare an infinite number of more or less distantly related species to each other, current efforts are limited by experimental data coverage to a selected set of species. The direct comparison of proteins remains a challenging task mainly due to the definition of homologous sequences across species. In many cases, it is not trivial to distinguish between paralogous and orthologous sequences[191].

With the Ulysses system (paper III) we were able to show how to project protein networks from distantly related species onto human by capitalizing on the body of experimental evidence generated for favorite model organisms. We explicitly demonstrated the functionality of such a system. However, its current utility is limited by the depth and coverage of protein interaction data. As the most reliable associations were those either confirmed by multiple approaches in a single species or repeatedly detected across organisms, increased amounts of protein interaction data will further allow for ample high-confidence network projections. Additionally, the inclusion of further model organisms will extend projections to currently undetectable protein networks.

As outlined in the preamble, there exist a number of systems to seize the homologous relationship between gene products across a variety of organisms, and each of them has its inherent advantages and drawbacks. In the Ulysses system, we chose NCBI's

HomoloGene to delineate homologous proteins between yeast, worm, fly, and human. Even though we made this choice in the first release of the Ulysses system, future versions are intended to include options for the use of other homology mappings.

This thesis introduced tools for bioinformatics analysis. Thus, results are reflected in both scientific publications and internet-based software systems. It is my belief that both components have accelerated the global effort to understand the human genome.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who went along with me during this journey and who contributed to this work.

Special thanks to:

My thesis advisor Wyeth W. Wasserman
It is almost impossible to find the right words to express all my gratitude to this special and impressive person. I would like to thank Wyeth for excellent scientific guidance and for giving me endless opportunities, especially for offering me to join him to Vancouver; for all his support, patience, and understanding; for his never failing sense of humor and his common sense.

My thesis co-advisor Christer Höög
I would like to thank Christer for initially taking me on the project at the Karolinska Institute; for his firm and continuous guidance and especially for his support and contributions to ease the continuation of my graduate studies from a distance.

There are countless people to thank, colleagues, friends, and family, who supported me during these years, and who helped me, both on a professional and a personal level. I am using this space to present my thanks to colleagues in particular. For all those who shared personal moments of joy and gloom, I hope that I will be given the opportunity to give back and acknowledge every single one of you.

Ample thanks to:

The members of the former Pegasus team for sharing and instructing me in my first experiences with molecular biology, in particular Margareta Faxén, and Annika Eriksson.

Claes Wahlestedt, for help and support as a prefect.

Former and present members of Christer's group, for assisting me in the lab, especially Mary-Rose Hoja, Emily Hodges, and Eva Brundell.

Former members of Wyeth's group at the Karolinska Institute, in particular Elena Herzog, Albin Sandelin, and Boris Lenhard for letting me in on the mysteries of bioinformatics.

Present members of Wyeth's group in Vancouver, including Raf Podowski, Jonathan Lim, Jochen Brumm, and David Arenillas, for being such a great team to translate all these biological ideas into computer-understandable jargon and statistical models; Shannan Ho Sui and Elodie Portales-Casamar for critical reviews.

Former and present members of the UBiC team, especially Francis Ouellette, Stefanie Butland, Graeme Campbell, Jessica Sawkins, Julie Stitt, and Scott McMillan for a very

warm welcome at the CMMT in Vancouver; and former Ulysses team members including Sohrab Shah, Yong Huang, Mack Yuen, John Ling, and Tao Xu for excellent team work on the Ulysses project.

Michael Hayden for accomodating me at the CMMT, and Elisabeth Simpson for providing me with lab space.

Colleagues from the Karolinska Intitute, former Pharmacia Corp., Stockholm University, Royal Institute of Technology, and GlaxoSmithKline, including Erik Sonnhammer, Elsebrit Ljungström, Anders Lundmark, Mark Reimers, Peggy Roth, Peter Nilsson, Mary Olsen, Steven Roberds, and Pankaj Agarwal, for concerted efforts in the expression platform comparison and NovelFam3000 projects.

Administrative personnel, both at the Karolinska Institute and CMMT, including Miroslav Hatas, Jonathan Falkowski, Bent Terp, and Pierre Dubitskij, for dealing with diverse computer issues; and Gitt Elsén, Britt-Marie Uppgren, and Elisabeth Grenmyr for helping me with the administrative issues related to graduate work and thesis defense preparation from a distance.

Very special thanks to my friend Barbara Albiger, there are just very few of you in a lifetime.

My deepest gratitude to the two most special persons in my life, my husband Rob Cassidy, for his never failing support and cheerfulness, even during the most challenging moments, and our daughter Sara, who helped me to put everything into perspective, and who fills our live with joy.

REFERENCES

1. GeneLynx - A portal to mammalian genomes (http://www.genelynx.org)
2. Pennisi, E. Human genome. Reaching their goal early, sequencing labs celebrate. *Science* **300**, 409 (2003).
3. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
4. Mendel, G. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn* **4**, 3-47 (1866).
5. Avery, O.T., MacLeod, C.M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J Exp Med* **79**, 137-158 (1944).
6. Nirenberg, M.W. The genetic code. II. *Sci Am* **208**, 80-94 (1963).
7. Cohen, S.N., Chang, A.C., Boyer, H.W. & Helling, R.B. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70**, 3240-3244 (1973).
8. Sanger, F. & Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-448 (1975).
9. Maxam, A.M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564 (1977).
10. Smith, L.M. et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).
11. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
12. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
13. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
14. Adams, M.D. et al. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).
15. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
16. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).
17. Schmutz, J. et al. Quality assessment of the human genome sequence. *Nature* **429**, 365-368 (2004).
18. Pease, A.C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* **91**, 5022-5026 (1994).
19. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* **6**, 639-645 (1996).
20. Bonaldo, M.F., Lennon, G. & Soares, M.B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* **6**, 791-806 (1996).
21. Wattler, S., Kelly, M. & Nehls, M. Construction of gene targeting vectors from lambda KOS genomic libraries. *Biotechniques* **26**, 1150-1156, 1158, 1160 (1999).
22. Walhout, A.J. et al. Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr Biol* **12**, 1952-1958 (2002).
23. Southan, C. Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics* **4**, 1712-1726 (2004).
24. Harrison, P.M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res* **30**, 1083-1090 (2002).
25. Hubbard, T. et al. The Ensembl genome database project. *Nucleic Acids Res* **30**, 38-41 (2002).
26. Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. A vision for the future of genomics research. *Nature* **422**, 835-847 (2003).

27. Alwine, J.C., Kemp, D.J. & Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A* **74**, 5350-5354 (1977).
28. Thomas, P.S. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc Natl Acad Sci U S A* **77**, 5201-5205 (1980).
29. Becker-Andre, M. & Hahlbrock, K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Res* **17**, 9437-9446 (1989).
30. Wang, A.M., Doyle, M.V. & Mark, D.F. Quantitation of mRNA by the polymerase chain reaction. *Proc Natl Acad Sci U S A* **86**, 9717-9721 (1989).
31. Okubo, K. et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* **2**, 173-179 (1992).
32. Adams, M.D. et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3-174 (1995).
33. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).
34. Schena, M. et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* **93**, 10614-10619 (1996).
35. Chee, M. et al. Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-614 (1996).
36. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-1680 (1996).
37. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-719 (2002).
38. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691 (2003).
39. Nakai, K. & Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24**, 34-36 (1999).
40. Marcotte, E.M., Xenarios, I., van Der Bliek, A.M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A* **97**, 12115-12120 (2000).
41. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**, 1005-1016 (2000).
42. Tsien, R.Y. The green fluorescent protein. *Annu Rev Biochem* **67**, 509-544 (1998).
43. Ding, D.Q. et al. Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library. *Genes Cells* **5**, 169-190 (2000).
44. Morin, X., Daneman, R., Zavortink, M. & Chia, W. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in Drosophila. *Proc Natl Acad Sci U S A* **98**, 15050-15055 (2001).
45. Ross-Macdonald, P., Sheehan, A., Roeder, G.S. & Snyder, M. A multipurpose transposon system for analyzing protein production, localization, and function in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **94**, 190-195 (1997).
46. Walhout, A.J. et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287**, 116-122 (2000).
47. Simpson, J.C., Wellenreuther, R., Poustka, A., Pepperkok, R. & Wiemann, S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep* **1**, 287-292 (2000).
48. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-4574 (2001).
49. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).
50. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-183 (2002).
51. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147 (2002).

52. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246 (1989).
53. Rigaut, G. et al. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**, 1030-1032 (1999).
54. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
55. Butler, D. Scientific publishing: who will pay for open access? *Nature* **425**, 554-555. (2003).
56. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **33 Database Issue**, D39-45 (2005).
57. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
58. "Saccharomyces Genome Database" (http://www.yeastgenome.org/)
59. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **31**, 172-175 (2003).
60. Chen, N. et al. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res* **33 Database Issue**, D383-389 (2005).
61. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
62. Marcotte, E.M. et al. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-753 (1999).
63. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288 (1999).
64. Giot, L. et al. A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-1736 (2003).
65. Li, S. et al. A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-543 (2004).
66. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).
67. The General Repository for Interaction Datasets (http://biodata.mshri.on.ca/grid/servlet/Index)
68. Breitkreutz, B.J., Stark, C. & Tyers, M. The GRID: the General Repository for Interaction Datasets. *Genome Biol* **4**, R23 (2003).
69. Breitkreutz, B.J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biol* **4**, R22 (2003).
70. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504. (2003).
71. Hu, Z., Mellor, J., Wu, J. & DeLisi, C. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* **5**, 17. (2004).
72. Yu, H., Zhu, X., Greenbaum, D., Karro, J. & Gerstein, M. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* **32**, 328-337. Print 2004. (2004).
73. Alfarano, C. et al. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33 Database Issue**, D418-424 (2005).
74. Biomolecular Interaction Network Database (http://www.bind.ca)
75. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451 (2004).
76. Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/)
77. Mewes, H.W. et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**, D41-44 (2004).
78. Munich Information Center for Protein Sequences (http://mips.gsf.de/)
79. Zanzoni, A. et al. MINT: a Molecular INTeraction database. *FEBS Lett* **513**, 135-140 (2002).
80. The Molecular Interaction Database (http://mint.bio.uniroma2.it/mint/)

81. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**, D452-455 (2004).
82. IntAct (http://www.ebi.ac.uk/intact/index.jsp)
83. Han, K., Park, B., Kim, H., Hong, J. & Park, J. HPID: the Human Protein Interaction Database. *Bioinformatics* **20**, 2466-2470 (2004).
84. Human Protein Interaction Database (http://www.hpid.org)
85. von Mering, C. et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31**, 258-261 (2003).
86. Brown, K.R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076-2082. Epub 2005 Jan 2018. (2005).
87. Huang, T.W. et al. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics* **20**, 3273-3276 (2004).
88. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113 (2004).
89. Han, J.D. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93. Epub 2004 Jun 2009. (2004).
90. von Mering, C. et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403 (2002).
91. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126 (2000).
92. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-255 (2003).
93. Tong, A.H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-2368 (2001).
94. Tong, A.H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-813 (2004).
95. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).
96. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901 (1999).
97. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).
98. Galperin, M.Y. & Koonin, E.V. Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* **18**, 609-613 (2000).
99. Valencia, A. & Pazos, F. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* **12**, 368-373 (2002).
100. Huynen, M.A., Snel, B., von Mering, C. & Bork, P. Function prediction and protein networks. *Curr Opin Cell Biol* **15**, 191-198 (2003).
101. Goh, C.S. & Cohen, F.E. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol* **324**, 177-192 (2002).
102. Ramani, A.K. & Marcotte, E.M. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* **327**, 273-284 (2003).
103. Gertz, J. et al. Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* **19**, 2039-2045 (2003).
104. Morett, E. et al. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat Biotechnol* **21**, 790-795 (2003).
105. Russell, R.B. et al. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol* **14**, 313-324 (2004).
106. Ng, S.K., Zhang, Z. & Tan, S.H. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* **19**, 923-929 (2003).
107. Aloy, P. et al. Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026-2029 (2004).

108. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-486 (2001).
109. Kemmeren, P. et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133-1143 (2002).
110. Saito, R., Suzuki, H. & Hayashizaki, Y. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics* **19**, 756-763 (2003).
111. Goldberg, D.S. & Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* **100**, 4372-4376 (2003).
112. Date, S.V. & Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21**, 1055-1062 (2003).
113. Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein-protein interaction data? *J Mol Biol* **327**, 919-923 (2003).
114. Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* **22**, 78-85 (2004).
115. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-280 (2004).
116. Camon, E. et al. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* **13**, 662-672 (2003).
117. Luc, P.V. & Tempst, P. PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics* **20**, 1413-1415 (2004).
118. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-2371 (2003).
119. Strong, M., Mallick, P., Pellegrini, M., Thompson, M.J. & Eisenberg, D. Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biol* **4**, R59 (2003).
120. Yanai, I. & DeLisi, C. The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol* **3**, research0064 (2002).
121. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453 (2003).
122. Schlitt, T. et al. From gene networks to gene function. *Genome Res* **13**, 2568-2576 (2003).
123. Tornow, S. & Mewes, H.W. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res* **31**, 6283-6289 (2003).
124. Wilson, C.A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**, 233-249 (2000).
125. Hegyi, H. & Gerstein, M. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res* **11**, 1632-1640 (2001).
126. Kelley, B.P. et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* **100**, 11394-11399 (2003).
127. Wojcik, J., Boneca, I.G. & Legrain, P. Prediction, assessment and validation of protein interaction maps in bacteria. *J Mol Biol* **323**, 763-770 (2002).
128. Yu, H. et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* **14**, 1107-1118 (2004).
129. Lehner, B. & Fraser, A.G. A first-draft human protein-interaction map. *Genome Biol* **5**, R63 (2004).
130. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052 (2001).

131. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36 (2000).
132. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
133. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003).
134. Lee, Y. et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res* **12**, 493-502 (2002).
135. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome Res* **7**, 986-995 (1997).
136. Freeman, W.M., Robertson, D.J. & Vrana, K.E. Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques* **29**, 1042-1046, 1048-1055 (2000).
137. Madden, S.L., Wang, C.J. & Landes, G. Serial analysis of gene expression: from gene discovery to target identification. *Drug Discov Today* **5**, 415-425 (2000).
138. Taniguchi, M., Miura, K., Iwao, H. & Yamanaka, S. Quantitative assessment of DNA microarrays--comparison with Northern blot analyses. *Genomics* **71**, 34-39 (2001).
139. Ishii, M. et al. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* **68**, 136-143 (2000).
140. Wang, T. & Brown, M.J. mRNA quantification by real time TaqMan polymerase chain reaction: validation and comparison with RNase protection. *Anal Biochem* **269**, 198-201 (1999).
141. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. & Kohane, I.S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405-412 (2002).
142. Gnatenko, D.V. et al. Transcript profiling of human platelets using microarray and serial analysis of gene expression. *Blood* **101**, 2285-2293 (2003).
143. Tan, P.K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684 (2003).
144. Huminiecki, L., Lloyd, A.T. & Wolfe, K.H. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* **4**, 31 (2003).
145. Barczak, A. et al. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res* **13**, 1775-1785 (2003).
146. dbMTN (http://www.cisreg.ca/dbMTN)
147. Yauk, C.L., Berndt, M.L., Williams, A. & Douglas, G.R. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res* **32**, e124. (2004).
148. Mah, N. et al. A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol Genomics* **16**, 361-370. (2004).
149. Irizarry, R.A. et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**, 345-350. Epub 2005 Apr 2021. (2005).
150. Lenhard, B., Hayes, W.S. & Wasserman, W.W. GeneLynx: a gene-centric portal to the human genome. *Genome Res* **11**, 2151-2157 (2001).
151. Friedman, J.H. & Roosen, C.B. An introduction to multivariate adaptive regression splines. *Stat Methods Med Res* **4**, 197-217. (1995).
152. Pontil, M. & Verri, A. Properties of support vector machines. *Neural Comput* **10**, 955-974. (1998).
153. von Mering, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33 Database Issue**, D433-437 (2005).
154. Matthews, L.R. et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* **11**, 2120-2126 (2001).
155. Shah, S.P. et al. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics* **6**, 34. (2005).
156. Ulysses (http://www.cisreg.ca/ulysses)

157. Peri, S. et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, D497-501. (2004).
158. Bairoch, A. et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**, D154-159. (2005).
159. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370. (2003).
160. Wu, C.H. et al. The Protein Information Resource. *Nucleic Acids Res* **31**, 345-347. (2003).
161. Mulder, N.J. et al. InterPro, progress and status in 2005. *Nucleic Acids Res* **33**, D201-205. (2005).
162. Falquet, L. et al. The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**, 235-238. (2002).
163. Attwood, T.K. The PRINTS database: a resource for identification of protein families. *Brief Bioinform* **3**, 252-263. (2002).
164. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141. (2004).
165. Corpet, F., Servant, F., Gouzy, J. & Kahn, D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* **28**, 267-269. (2000).
166. Ponting, C.P., Schultz, J., Milpetz, F. & Bork, P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* **27**, 229-232. (1999).
167. Haft, D.H., Selengut, J.D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**, 371-373. (2003).
168. Wu, C.H., Xiao, C., Hou, Z., Huang, H. & Barker, W.C. iProClass: an integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res* **29**, 52-54. (2001).
169. Andreeva, A. et al. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-229. (2004).
170. Rubin, G.M. et al. Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215. (2000).
171. NovelFam3000 (http://www.cisreg.ca/cgi-bin/NovelFam/novelfam)
172. Hoffmann, R. & Valencia, A. A gene network for navigating the literature. *Nat Genet* **36**, 664. (2004).
173. Safran, M. et al. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* **31**, 142-146. (2003).
174. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067. Epub 2004 Apr 6069. (2004).
175. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-250. (2003).
176. Kemmer, D. et al. Ulysses - an application for the projection of molecular interactions across species. *Genome Biol* **6**, R106 (2005).
177. Killion, P.J., Sherlock, G. & Iyer, V.R. The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics* **4**, 32. (2003).
178. Parkinson, H. et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33**, D553-555. (2005).
179. Lee, J.K. et al. GeneX Va: VBC open source microarray database and analysis software. *Biotechniques* **36**, 634-638, 640, 642. (2004).
180. Barrett, T. et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* **33**, D562-566. (2005).
181. Greene, J.M. et al. The NCI/CIT microArray database (mAdb) system - bioinformatics for the management and analysis of Affymetrix and spotted gene expression microarrays. *AMIA Annu Symp Proc*, 1066. (2003).
182. Manduchi, E. et al. RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinformatics* **20**, 452-459. Epub 2004 Jan 2022. (2004).

183. Gruss, P. Open access to science and culture. *Science* **303**, 311-312. (2004).
184. Tamber, P.S., Godlee, F. & Newmark, P. Open access to peer-reviewed research: making it happen. *Lancet* **362**, 1575-1577. (2003).
185. Giles, J. Trust gives warm welcome to open access. *Nature* **432**, 134. (2004).
186. Check, E. NIH open-access plans draw fire from both sides. *Nature* **433**, 561. (2005).
187. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-371. (2001).
188. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* **14**, 656-664. (1998).
189. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33**, D54-58. (2005).
190. Haubold, B. & Wiehe, T. Comparative genomics: methods and applications. *Naturwissenschaften* **91**, 405-421. Epub 2004 Jun 2025. (2004).
191. Fitch, W.M. Homology a personal view on some of the problems. *Trends Genet* **16**, 227-231. (2000).