

Gene expression profiling in islet biology and diabetes research

Ernesto Bernal-Mizrachi
Corentin Cras-Méneur
Mitsuru Ohsugi
M. Alan Permutt*

*Washington University School of
Medicine, Division of Endocrinology,
Diabetes and Metabolism, St. Louis,
Missouri, USA*

*Correspondence to:
M. Alan Permutt, M.D. Washington
University School of Medicine
Division of Endocrinology, Diabetes
and Metabolism, 660 S. Euclid
Avenue, Campus Box 8127, Saint
Louis, MO 63110.
E-mail: apermutt@im.wustl.edu

Summary

Following the sequencing of most of the human and mouse genomes, the next task for physicians and scientists will be to assess the relative levels of expression of these genes during development, following exposure to various nutritional and pharmacological conditions, and in disease states such as diabetes and related metabolic disorders. This review provides an overview of the various methodologies available for monitoring global gene expression. Use of cDNA libraries, Expressed Sequence Tag (EST) sequencing projects and databases, differential display (DD), serial analysis of gene expression (SAGE), subtractive cloning, and both cDNA and oligo microarrays are discussed, along with their merits and limitations. The Endocrine Pancreas Consortium <http://www.cbil.upenn.edu/EPConDB/> has constructed mouse and human cDNA libraries from adult and various stages of embryonic development of endocrine pancreas. Over 100 000 ESTs have been deposited in public databases, and each clone is available through the IMAGE Consortium. A guide to Internet access is provided for future investigation. Copyright © 2002 John Wiley & Sons, Ltd.

Keywords cDNA library; SAGE; microarray; gene expression profile; endocrine pancreas consortium

Introduction

The discovery of the insulin gene, one of the first mammalian genes to be cloned, became an important part of biomedical history in 1977 [1]. Later the insulin-receptor gene was cloned in 1989 [2], followed shortly thereafter with the cloning of the glucokinase [3] and glucose-transporter genes [4]. Early diabetes/metabolic related gene-expression studies examined one or a few genes. At present, numerous genes have been identified in pancreatic islets and other metabolic tissues. For example, a survey of islet publications from the year 2001 to the present listed many studies describing a host of islet-transcription factors along with their effects on pancreatic islet development and function (e.g. PDX, HNF1 α , HNF4 α , HNF4 γ , HNF3 β , HNF3 γ , Imx1b, Nkx2.2, Nkx6.1, Neuro-D/Beta2, Neurogenin 3, Pax 4 and Pax 6, PPAR α and PPAR γ , c-Myc, c-Fos, Egr-1, Egr-2, Elk-1, SRF-1, CREB, NF- κ B, SMADs 1,2, and 4, HES-1, C/EBP, Id-1 and Id-3, RIPE3b1, and PBX, see Reference [5] for review). With the completion of the sequencing of the human genome, we now know of the existence of more than 30 000 genes. Analysis of expression of these genes has become a major focus of genome research, and for diabetes research in particular, the profiles of gene expression in critical tissues is of primary importance.

Gene expression studies for diseases such as diabetes will be challenging, as they must encompass various stages of embryonic development, different physiological and nutritional conditions, changes throughout the evolution of

Received: 6 June 2002

Accepted: 26 August 2002

the disease, and responses to pharmacotherapy. Understanding the regulation of expression of genes will predictably yield more insight into the biology. This does not presuppose that levels of regulation beyond transcription are not perhaps equally important, including posttranslational processing such as phosphorylation, glycosylation, acylation or myristoylation, as well as protein–protein interactions. Nevertheless, a thorough understanding of the genome at the level of gene expression is likely to markedly accelerate our appreciation of the biology of pancreatic islets and the tissues controlling lipid, protein, and carbohydrate metabolism.

It is still early in the development and utilization of expression profiling in diabetes research. The purpose of this review is to discuss the basic methods of monitoring gene expression, with a critical evaluation of the advantages and disadvantages of the methods most commonly employed today. Lastly, we will briefly present recent progress in the efforts of the Endocrine Pancreas Consortium (<http://www.cbil.upenn.edu/EPConDB/>) to identify and monitor genes expressed in developing and adult mouse and human endocrine pancreas.

Basic methods

cDNA libraries

The fundamental method of monitoring gene expression in tissues involves extraction of total RNA, further isolation of the mRNA or polyA RNA fraction, and then reverse transcription into cDNA, which is more resistant to degradation. When adaptors are added to the ends of the cDNA, it can be spliced into universal cloning sites in bacterial plasmids, and the whole spectrum of mRNAs are cloned in bacteria. The resulting collection of clones is called a cDNA library [6]. Individual clones contain a single cDNA representing one copy of an original mRNA. In general, all the cDNAs in a library together represent the frequency of expression of various genes, that is, the most abundantly expressed genes will yield the most copies of individual clones. When searching for less abundant genes, libraries are normalized, that is, the most abundant genes are subtracted and discarded, leaving in theory a single copy of each gene [7].

EST sequencing and electronic databases

cDNA clones are generally sequenced from their ends, generating an EST or an 'expressed sequence tag'. This generally encompasses up to 500 bp from either the 5'- or 3'-end of a cDNA clone. An EST sequencing project typically requires cDNA libraries, robotic clone pickers, automated DNA-sequencing capability, and bioinformatics support. A library is generally sequenced until the yield of novel clones is reduced to less than 10 to 20%. In this respect, normalized libraries have been

useful, as redundant clones are minimized at the outset. All clones once sequenced are made available through the IMAGE Consortium (<http://image.llnl.gov>), and the sequences are deposited in electronic databases. The National Center for Biotechnology Information (NCBI) <http://www.ncbi.nlm.nih.gov/> maintains a repository of ESTs.

Generating a cDNA library provides a highly redundant set of ESTs that then need to be merged to build up a set of unique genes. Sequences have to be aligned and assembled to produce a nonredundant set of sequences. Some amount of mismatching, owing to sequencing errors, has to be taken into account without taking the risk of clustering sequences corresponding to regions conserved between protein families. Several algorithms can be used for that purpose such as EZCluster or a greedy algorithm as used in UniGene [8]. UniGene is a repository of clustered sequences from different species that is directly accessible from the World Wide Web (Table 1).

dbEST is a division of NCBI maintaining a sequence database and other information on 'single-pass' cDNA sequences from a number of organisms. Examination of dbEST (release 041902) revealed 4310771 entries for *Homo sapiens* (human) and 2562293 entries for *Mus musculus* + *domesticus* (mouse). A search of the Unified Library Database <http://www.ncbi.nlm.nih.gov/UniLib/lb.cgi> with the query 'islet', listed 21 cDNA libraries with mouse and human ESTs. There are a number of ways that this EST data can be used for diabetes research. For instance, all this data is compiled by the UniGene database into sets of overlapping ESTs representing unique genes. Each of the 21 islet cDNA libraries can be queried to find the number of UniGene clusters represented in the library. For example, dbEST now lists 37 281 human islet ESTs. UniGene reports 8732 human islet or insulinoma ESTs for *Homo sapiens*. There are 26 869 mouse islet ESTs recorded, and in the UniGene 1779 records satisfy the query islet for the organism *Mus musculus*. Another resource that analyzes cDNA libraries is NCBI's National Cancer Institute Cancer Genome Anatomy Project (CGAP) <http://cgap.nci.nih.gov/>. These databases can be searched by gene, disease, tissue, chromosome, or function. For example, a search for human pancreatic islet genes lists 7 libraries, 50 510 ESTs, further separated into 279 genes unique to islets, and 8520 nonunique genes (6044 known and 2476 unknown). Additionally, the partial sequences of these ESTs can be used to design primers for polymerase chain reaction (PCR) amplification of full-length mRNA, or to provide probes for hybridization. The strength of high-throughput sequencing is the capacity to identify new genes. However, as complete genome sequences are obtained, few genes will remain unidentified, making high-throughput EST sequencing less productive.

High-throughput EST sequencing leaves the investigator with a large number of genes to screen for differential expression. With the discovery of PCR in the late 1980s,

Table 1. Links to Internet sites useful for monitoring gene expression profiling

NCBI homepage	http://www.ncbi.nlm.nih.gov/
Databases / libraries:	
Nucleotide search at NCBI	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
SRS	http://srs.ebi.ac.uk/ http://www.infobiogen.fr/srs/
DbEST	http://www.ncbi.nlm.nih.gov/dbEST/dbEST
UniLib browser	http://www.ncbi.nlm.nih.gov/UniLib/lb.cgi
CGAP	http://cgap.nci.nih.gov/
CGAP library finder	http://cgap.nci.nih.gov/Tissues/LibraryFinder
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/
SAGE	http://www.sagenet.org
SAGE tag to gene mapping	http://www.ncbi.nlm.nih.gov/SAGE/SAGETag.cgi
ESTs search	http://genome-www5.stanford.edu/cgi-bin/SMD/source/sourceSearch
allgenes.org RNA query	http://www.allgenes.org/gc/servlet?page=Booleans
RNA abundance database	http://www.cbil.upenn.edu/RAD2/query.htm
IMAGE	http://image.llnl.gov/
Software	
ESTCluster and IcaTool	http://www.jparsons.uklinux.net/bioinf/pub
Staden	http://www.mrc-lmb.cam.ac.uk/pubseq/staden_home.html
CAP3	http://genome.cs.mtu.edu/cap/cap3.html
GeneCluster	http://www-genome.wi.mit.edu/cancer/software/genecluster/genecluster.html
Xcluster	http://genome-www.stanford.edu/~sherlock/cluster.html
Gbuilder	http://industry.ebi.ac.uk/~muilu/GBuilder/
ESTate	http://fe.hgmp.mrc.ac.uk/gslater/www/webwebx.cgi?ViewPage=ESTatePage
Oligo array design	http://berry.engin.umich.edu/oligoarray/
WU BLAST	http://blast.wustl.edu/
Pancreas EST project home page and related sites	
EPConDB (Endocrine pancreas consortium web site)	http://www.cbil.upenn.edu/EPConDB/
BLAST on EPC databases	http://www.cbil.upenn.edu/epcondb/servlet?page=blast
Ranks for the different libraries	http://www.cbil.upenn.edu/EPConDB/pancreas.rank.current
Pancreas reports for the different libraries	http://www.cbil.upenn.edu/EPConDB/pancreas.report.current
Sequencing reports	http://genome.wustl.edu:8021/pub/gsc1/est/pancreas_wingz/

older methods have been improved and many new techniques have been developed that make discovery of differentially expressed genes more simple and allow the analysis of differential gene expression at the single cell level.

Subtractive cloning

The purpose of creating a subtraction library is considerable enrichment of the target cDNA clones. A subtracted library contains cDNA clones present in one cell or tissue type and not present in a second type [9]. This cDNA library is used to isolate a set of cDNA clones corresponding to a class of mRNAs that are differentially expressed, or to aid in the isolation of a cDNA clone where the screening procedure for the cDNA clone is laborious because a specific DNA or antibody probe is unavailable. In this technique, the tissue, library, RNA, or cDNA designated with a [+] contains the target or desired sequence(s), and that which is to be subtracted from the [+] is termed [-]. Two major disadvantages to this approach are that poly (A)⁺ RNA from both [+] and [-] source is required and that the hybridizations and library production can be technically difficult with a very

small amount of cDNA. Another disadvantage of this or any other subtraction protocol is that clones containing reiterated sequences would be eliminated from the library on that basis.

A good alternative to creating a subtracted library is differential screening of a library known to contain the target clone(s) (See Reference [10], Tedder *et al.*, for example). A potential drawback to the differential screening approach is that rare transcripts will have very low specific probe concentrations in the mixture and thus might not hybridize to the DNA from a target plaque in a reasonable period of time.

This technique has been applied successfully to identify important genes involved in the physiopathology of diabetes. Hutton *et al.* used this system to identify genes differentially expressed between insulinoma (β TC3) and glucagonoma (α TC2) cell lines [11]. Using difference analysis of cDNA, Kanwar *et al.* identified genes upregulated in the kidney of streptozotocin-induced diabetic mice [12]. A mammalian homologue of the translocase of mitochondrial inner membrane was isolated as being regulated in the kidney of diabetic newborn [12]. RT-PCR-based subtractive hybridization was used on islets from a patient who died at the onset of type 1 diabetes, and it identified a

type 1 diabetes-related cDNA encoding hepatocarcinoma-intestine-pancreas/pancreatic-associated protein (HIP/PAP) [13,14]. This protein has been proposed to be a target of the autoimmune response in the Non-Obese Diabetic (NOD) mouse [14]. However, because of the disadvantages described above, this technique is now used less frequently and has been replaced by some of the techniques described later.

Differential display

Although subtraction is quite sensitive and can detect fairly nonabundant mRNAs, the method recovers genes incompletely and selects genes in only one direction at a time during a two-way comparison between a pair of cells or tissues. This process is also laborious and time-consuming. Another method used to identify mRNA species for differentially expressed genes is differential display (DD). The DD technique was developed with the goal of identifying differentially expressed genes, detecting individual mRNA species that are changed in different sets of mammalian cells, then recovering and cloning the cDNA [15,16]. In this technique, each RNA sample is first reverse transcribed with a degenerate anchored oligo (dT) primer set that anneals at the start of the poly (A) tails of mRNAs. Each degenerate anchored oligo (dT) primer set (e.g. T₁₂MA) in theory, reverse transcribes one-fourth of the total mRNA population. In combination with a decamer oligonucleotide of arbitrary sequence, which in theory can hybridize to any mRNA, cDNA fragments representing the 3' termini of mRNAs defined by both primers, are amplified. Thus, this procedure allows amplification of an mRNA subpopulation without knowledge of sequence information (DD RT-PCR or DDRT-PCR). Because any given arbitrary decamer will not actually sample all mRNAs, different decamers can be used to permit sampling of, in theory, all mRNAs in a given tissue. One of the purposes of this technique is to provide a picture of mRNA composition of cells by displaying subsets of mRNAs as short cDNA bands. This mRNA fingerprinting is useful for observing alterations in gene expression. Secondly, these DNAs can be quickly reamplified, cloned, sequenced, and compared with sequences in databases. Finally, reamplified cDNAs can be used as probes for Northern or Southern blot hybridization and to isolate genes from genomic or cDNA libraries for further molecular characterization.

A particular difficulty with DD is a high rate of false positives because of the frequent failure to separate different cDNAs of similar size on gels [17]. DD also introduces an important number of false positives caused by the lack of specificity of the primers and the low stringency of the amplification conditions. A precise optimization of PCR parameters however, such as annealing temperature and additive contents, can help limit these aspects [18]. In addition, though this method has been outdated by newer technologies, DD

is still one of the most widely used methods for expression analysis because it can be performed in any laboratory equipped with standard molecular biology reagents and instrumentation, and the need for advanced bioinformatics for analysis is minimal. This technique has been applied *in vitro* to cell lines and used to identify genes differentially expressed in various animal models of diabetes and humans (See Table 2). Experiments using DD often result in the identification of novel or uncharacterized genes, as well as suggesting novel functions for known genes. For example, Eizirik *et al.*, performed DD on mRNA from isolated rat β cells treated with interleukin 1 β [19,20]. With the use of DD and most recently microarrays [21], this group has been able to identify the NF- κ B signaling pathway as an important mediator for the cytokine-induced β -cell dysfunction and death in type 1 diabetes.

In the analyses of differential regulation of various genes by hyperglycemia in different tissues, DD has been employed with a certain degree of success. For instance, with the use of DDRT-PCR, some of the differentially expressed genes in the hyperglycemic state were isolated from aortic smooth muscle cell [30,31] and heart [32]. Studies in humans have provided important information about gene expression in the muscle of type 2 diabetic subjects. Groop L *et al.* performed DD on mRNA isolated from skeletal muscle biopsies of normal and diabetic subjects [26]. The results of this study identified mitochondrial gene expression as a potential important element in the development of insulin resistance. Another study using omental fat from normoglycemic and obese diabetic subjects identified four candidate genes distinguishing these phenotypes [27]. DD was also employed to find beta cell-specific genes by comparing mRNA obtained from human islets and exocrine pancreas. This study identified 29 novel sequences and 19 of those were mapped to chromosomal locations by several methods [28].

SAGE

SAGE (Serial Analysis of Gene Expression) was initially described in 1995 by Velculescu [33]. SAGE involves the generation of short fragments of DNA, or tags, from a defined point in the sequence of all cDNAs in the sample analyzed. This short tag, because of its presence in a defined point in the sequence, is typically sufficient to uniquely identify every transcript in the sample. SAGE library construction involves anchoring mRNA molecules via their poly(A) tails to magnetic beads (MicroSAGE differs from conventional SAGE in that this anchoring at the 3' end takes place prior to cDNA synthesis rather than after cDNA synthesis). Double-stranded cDNA from the tissue(s) of interest is cleaved with a restriction endonuclease (anchoring enzyme) that is predicted to cut every transcript at least once. Such enzymes generally have 4-bp recognition sequences and cleave every 256 bp. The 3' ends of the resulting cDNA fragments are then

purified using streptavidin-coated magnetic beads and the resulting cDNA fragments divided into two populations, each of which is ligated to a different linker containing a type IIS-restriction endonuclease (tagging enzyme) recognition sequence. Such enzymes cleave DNA at a distance of up to 20 bp away from their recognition site. Digestion of the two cDNA populations thus results in the generation of a short sequence consisting of the linker and a short portion of its adjacent cDNA.

Following the creation of blunt ends, the two populations are ligated to each other and total cDNA amplified by PCR, resulting in the generation of products with two tags (a ditag), orientated tail to tail with an anchoring enzyme-recognition site at either end. Following cleavage at each anchoring enzyme-recognition sequence and concatenation of ditags via this site, products are cloned and individual clones consisting of at least 25 to 75 tags are selected for sequencing. An advantage of this

Table 2. Examples of differential display (DD) for monitoring gene expression in diabetes research¹

In vitro	Tissue	Method	Conditions	Differentially expressed genes	Genes	Reference
A subtractive cloning approach to the identification of mRNAs specifically expressed in pancreatic β cells	β TC3 insulinoma and α TC2 glucagonoma cell lines	Polymerase chain reaction-based subtractive hybridization (555 clones)		29 distinct sequences 17 were identical or homologous to known mammalian cDNAs or expressed sequence tags	Insulin islet amyloid polypeptide proinsulin convertase 1 neuropeptide Y	[22]
Genes expressed during the differentiation of pancreatic AR42J cells into insulin-secreting cells	Rat AR42J (pluripotent pancreatic cells)	DD (144 combinations used)	Dexamethasone-treated (Acinar phenotype) vs Activin A or Betacellulin (β cell phenotype)		Known genes β -Actin, Thymosin 10, PTHrP, SPP-24 precursor, Ca ²⁺ channel β subunit III, Carboxypeptidase E Homologues Tyrosine phosphatase, Transcovalamine II, Mouse secreted protein, β -tubulin, HNMP-1, Novel protein kinase PKN, Integrin 6 subunit, Phospholipase A2, Amyloid precursor-like protein, Keratin D. (See reference for more information)	[23]
Insulin secretion and differential gene expression in glucose-responsive and -unresponsive MIN6 sublines.	MIN6 insulinoma cell line	DD (216 combinations used)	Comparison between glucose responsive and -unresponsive MIN6 cell lines	10 differentially expressed genes	Upregulated Stanniocalcin, Delta-like protein precursor /Preadypocyte factor 1, EST(AA286583), EST(AU018846) Downregulated E3KARP, CCK-B receptor, EST(AU021144), EST (D28695), No match	[24]
Identification of IL-1 β -induced messenger RNAs in rat pancreatic beta cells by DD of messenger RNA.	β -cells isolated from Rat islets	DD	Interleukin 1 β (30U/ml) for 6 or 24 h.	8 differentially expressed genes	Adenine nucleotide translocator, islet antigen 2 and 2 β , Phospholipase D1, CINC 1 and 3, MCP-1, INOS, Serine protease inhibitor 3	[19, 20]
Alterations in skeletal muscle gene expression of ob/ob mice by mRNA DD.	Muscle	DD	Gastrocnemius muscle from ob/ob and ob/+	17 genes	Geranylgeranyl pyrophosphate synthase, Id-2, phosphoinositol glycan-specific phospholipase D, peroxisomal membrane protein, mouse ribosomal protein L3, Troponin T	[25]
Insulin-regulated mitochondrial gene expression is associated with glucose flux in human skeletal muscle.	Muscle	DD	Muscle biopsies from type 2 diabetic and healthy subjects	54 differentially expressed genes	Diabetics ND1, ND4 and ND5 (Mt), Myosin, α -actin, Ubiquitin hydrolase, D-loop (Mt), Ribosomal protein S15a, Titin, tRNA leu(Mt), COX1 (M t), germinal center kin. related kin. Nondiabetics Kiaa 0093, L-3-phosphoserine, Phosphatase homolog, Carbonic anhydrase III, exon7, ND2, ND4 and ND5 (Mt), 16s rRNA(Mt), Ribosomal protein S18, Titin, tRNAleu(Mt), COX1 (Mt), FXR 1	[26]

(continued overleaf)

Table 2. (Continued)

In vitro	Tissue	Method	Conditions	Differentially expressed genes	Genes	Reference
Identification of novel genes differentially expressed in omental fat of obese subjects and obese type 2 diabetic patients.	Omental fat	DD and subtracted library techniques	Omental fat from lean and obese nondiabetic subjects and obese type 2 diabetic patients	2,078 cDNAs that showed potential differential expression in the omental fat of lean versus obese nondiabetic subjects versus obese type 2 diabetic patients	Differential display Novel Sequences 31, No matches 15 EST 16 Known genes 8 Subtractive hybridization Novel Sequences 1439 No matches 900 EST 539 Known genes 600 26 were confirmed	[27]
Mapping novel pancreatic islet genes to human chromosomes	Human islets	DD	Human islets vs Exocrine pancreas	42 differentially expressed genes	Novel sequences 29 No matches 15 Matches to EST 8 Known genes 13	[28]
Identification of differentially expressed genes induced in pancreatic islet neogenesis	Hamster pancreas	DD (80 primer combinations)	Cellophane wrapping of the hamster pancreas vs control pancreata	10 differentially expressed genes	Cytochrome c oxidase, Ubiquitin conjugating enzyme, Elastase I, Reg family of genes, Pancreatitis Assoc. Protein (PAP), Novel	[29]

¹This table serves as a representative rather than an exhaustive account of the research in this area.

method over sequencing of conventional cDNA clones is that each clone contains numerous tags, thus economizing on sequencing costs.

SAGE is a sensitive method for detecting the relative abundance of a transcript but as Zhang *et al.* pointed out [34], it would take the analysis of 300 000 tags to yield a 92% chance to detect a tag for a transcript expressed at three copies per cell. This implies the necessity of construction of large libraries in order to detect low abundance transcripts.

SAGE analysis has a number of unique advantages over hybridization-based measures of global gene expression, such as microarrays, subtractive hybridization, and DD methodologies. SAGE generates a tag for virtually every cellular mRNA, providing a level of coverage unequalled by any microarray yet available for humans or mice. This information is an important tool for gene discovery and allows comparing tag levels among libraries generated by different labs. SAGE data represent absolute expression levels based on the digital enumeration of transcript tags in a total population of transcripts. Therefore, results from any new experiment are directly comparable to existing gene expression databases. As more SAGE libraries are generated and made public, these data sets can be used to create a large-scale atlas of gene expression that is of great use to the whole scientific community. Such a resource is already available for human normal and malignant tissues at NCBI (<http://www.ncbi.nlm.nih.gov/SAGE>).

One of the negative aspects of the method is the occasional failure of a SAGE tag to match a predicted gene or to be long enough to easily isolate a full-length cDNA clone. While this happens at relatively low frequency for high-abundance transcripts in model organisms, it can limit the interpretation of the data in some cases. Another minor disadvantage is the absence of the restriction site used for the construction of the SAGE library in some of

the cDNAs. SAGE analysis does not provide information about alternative spliced genes and a tag can therefore correspond to several proteins.

For several years, SAGE has been used to provide comprehensive analysis of a variety of different tissue samples, each usually consisting of millions of cells. This approach has recently been extended to permit analysis of gene expression in substantially fewer cells, thereby allowing analysis of heterogeneous tissues such as pancreatic islets.

SAGE and the pancreas

Only one SAGE experiment on normal pancreatic tissues has been published [33]. One thousand tags were manually sequenced revealing a gene expression pattern characteristic of pancreatic function. We have also performed a SAGE analysis on human pancreatic islets, an insulinoma, and exocrine pancreas in order to identify the transcripts represented in this fraction of the pancreas and estimate their relative abundance (manuscript in preparation). Despite the limitations of this method, SAGE analysis reflects the absolute RNA expression levels in a tissue. As mentioned above, one of the most notable strengths of the SAGE method is that results from any new experiment are directly comparable to existing gene expression databases. SAGE data represent absolute expression levels on the basis of the digital enumeration of transcript tags in a total population of transcripts. The quantity of SAGE data that can be used for comparisons has grown considerably over the past 2 years, with more than 3 million transcript tags representing over 100 000 unique transcripts from over 70 libraries analyzed and available over the Internet [35]. Some Internet sites even offer *in silico* mRNA subtraction between tissues for

which databases are available. SAGE have been mostly used in oncology [36] and recently, the effect of different factors on kidney ductal cells has also been assessed with SAGE [37]. Comparison of SAGE tag frequencies from pancreatic islets with other sources could generate important information about differential gene expression and better understanding of the biology of β cells.

DNA microarrays

DNA microarrays have been extensively employed to monitor mRNA expression since this approach was first described in 1995 [38]. In this method, RNAs are labeled and hybridized onto cDNA or oligonucleotides spotted on a precise location on a suitable surface. The quantification of the intensity of the hybridization then allows an estimation of the relative expression of each transcript. The use of different labels can even permit the direct comparison of the intensity of all the transcripts in several RNA samples on the same DNA array. Using this approach, one can compare the transcriptional responses of a tissue when subjected to different conditions and stimuli, or compare gene expression between normal and pathological tissues (see Figure 1). Here we will discuss the different types of microarrays currently being employed to monitor gene expression.

cDNA arrays

Spotting small aliquots of many cDNAs at a high density was the first approach used for building DNA arrays to monitor expression profiles of thousands of genes consecutively [38]. cDNAs are first sequenced and clustered into unique transcripts before being directly spotted on a suitable surface. The advantage of a high-density array or microarray, usually on a microscope slide encompassing a total area of less than 1 cm², is that small

volumes of RNA can be hybridized. One limitation of this type of platform exists in the different hybridization stringencies that can result from cDNAs of different sizes and Guanine Cytosine (GC) content. The cDNAs can also hybridize to transcripts sharing a high similarity and induce false positives.

Oligo arrays

Instead of cDNA, specifically designed oligomers aligned in arrays can be used to monitor gene expression profiles. Oligo arrays can be synthesized directly on a solid support using photochemical techniques [39,40] or ink-jet technology [41–43] or pre-synthesized and then spotted on the surface of a suitable material such as glass or nylon membranes. This method of monitoring gene expression is very sensitive, a single oligonucleotide per cell can be sufficient, and compares well to cDNA arrays [43,44]. The oligonucleotides spotted can vary in length between 25-mers up to several hundreds of bases long.

The quality of an oligo microarray depends entirely on the specificity of each oligomer arrayed. Each oligomer has to be able to distinguish its target from other members of the same family that could share similar domains. In addition, the affinity of each spot for the transcript it binds has to be similar to avoid preferential hybridization. Long sequences provide the advantage of high-detection sensitivity but they also bear a higher risk of overlapping sequences conserved in multiple genes and the induction of false positives [45].

Long oligo arrays

Short (25–35 nucleotides) and long oligo arrays (50–70 nucleotides) seem to share a similar sensitivity [44]. Long oligonucleotides provide a higher reproducibility but also bear a higher risk of hybridizing nontarget cDNA [44].

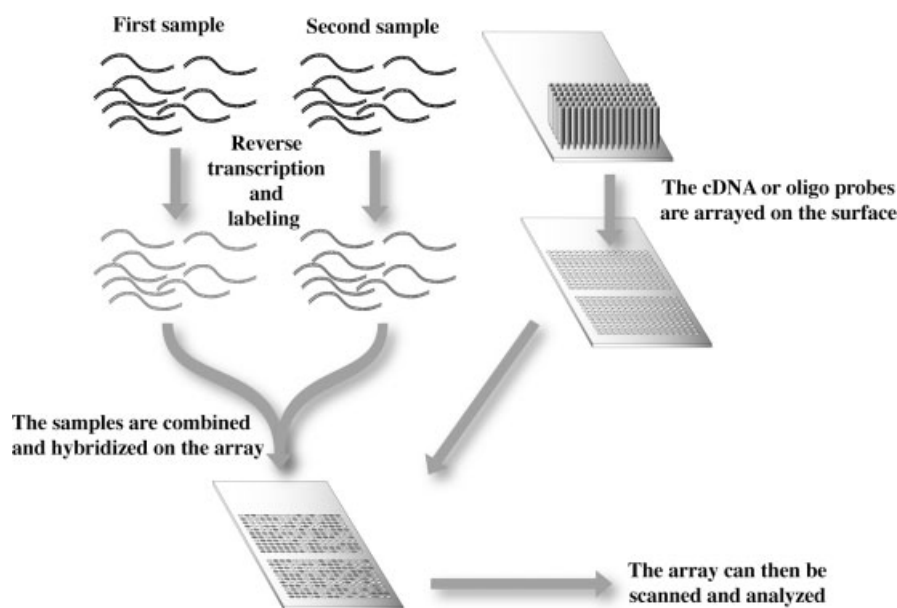


Figure 1. DNA microarray flowchart

Nevertheless, long oligonucleotides allow more stringent conditions to limit cross hybridization. About 50 or 60 oligomers seem to provide a higher reproducibility than shorter (25–35 mers) oligomers and allow a reproducible identification of the intensity variations of the transcripts.

Affymetrix

The Affymetrix GeneChip™ arrays use 20 to 25-mers with each probe corresponding to a different region of the mRNA of interest. The oligonucleotides are synthesized directly on the surface using light-directed combinatorial chemical synthesis in a spatially directed pattern. Only one set of mRNAs can be hybridized to a single microarray, so to compare transcript expression between different tissues, hybridizations have to be performed in several parallel experiments and normalized.

Limitations

Depending on the techniques used, DNA arrays can show a relatively important variability. The samples as well as the microarrays can vary depending on the microenvironment. The slightest differences, even the position of the cell cultures in the incubator or the position of the spots on the glass slide, can affect the quality of the spotted probed or the hybridization strength [46]. Comparing data obtained through different platforms can also show how difficult reproducing and validating the experiments performed through microarrays could be. The quality of the signal and the reproducibility of the hybridization entirely depend on the sequence chosen as a probe. Several studies have shown that the quality of the sequences used on the array can drastically affect the reproducibility of the results between platforms and their validations through other approaches. A poor probe used on one platform will not provide reproducible results when compared with a highly stringent probe used on another platform [47].

These platforms only allow monitoring the transcriptional variations between different conditions. Even though these levels can be correlated to the respective protein levels, [48] they do not reflect protein-level modifications such as phosphorylations, proteolysis, modifications of the tertiary structure or any other modification that can affect the function of proteins.

Arrays and the pancreas

High-density arrays have been used for the study of various aspects of pancreatic physiology or pathology. These platforms have allowed the identification of differentially regulated genes in animals subjected to different diets. For example, microarrays have been used to study the transcriptional modifications in the islets of suckling rats fed with high-carbohydrate milk [49], to study the effects of chronic lipid exposures [50] or to identify the factors involved in glucose regulation of secretion and metabolism in insulinoma cells in culture [51]. Microarrays have also permitted studies

of the effects of different agents on β cells, as for example, cytokine treatment [21] or the different pathways involved in the response to insulin or IGF-1 [52]. Finally, these different platforms have been used in the analysis of tumor markers in different pancreatic tumors [53]. Some of these experiments were performed with cDNA arrays [49,52] or oligo arrays (mostly Affymetrix platforms) [21,50,51,53], but the availability of new solutions such as long oligonucleotides array will offer more choices for future experiments.

Statistical analysis

Various sources of variations have been identified in the studies involving microarray data. Probe concentration, accuracy of spotting robots, variations in the hybridizations or background or imprecision of the array scanner for instance can affect the reproducibility of the measurements. To minimize these artifacts, several normalization methods can be followed prior to performing statistical analysis [54–56].

Once the normalization has been performed, statistical analysis can then proceed to help estimate whether the different conditions tested affect the transcriptional level in the tissues. The choice of the criterion for considering fold increases that may be relevant to biological responses is an important consideration prior to the statistical analysis itself [57]. Different methods of statistical analysis can then be used to assess the relevance of the differences seen through the microarray experiment: systematic statistical linear modeling [58] or multifactorial analyses [59–61], ANOVA tools like GeneANOVA [62] or mixture model-based approach to the clustering of microarray data through EMMIX-GENE [63] for instance.

Interpreting data

Different software can also help visualizing and organizing data in a graphical way suited for the analysis of multidimensional data. Self-organizing maps, such as GeneCluster [64] or visualization and classification software like GeneANOVA [62], help organizing and representing the data obtained through these techniques.

Once genes of interest have been identified, they can be organized and classified according to their implications into regulatory pathways or protein families, for example. The Gene Ontology, using a heuristic algorithm, allows defining the molecular functions of proteins based on sequence similarities [65]. Other tools allow tracing metabolic pathways and cellular interactions [66–68].

Commercial software and freely available packages are also available to perform these analyses. Links for obtaining some of this software are also given in Table 2.

Array validation

Despite a variety of computational methods to ensure the validity of microarray data mentioned above, concerns still exist regarding intra-chip and chip-to-chip variation.

Positive and negative data found in microarrays should be validated with other methods.

Conventional Northern blot analysis can be used, but it is not suitable for high-throughput data analysis because of its limitations. It requires relatively large amount of starting RNA and specific probes. Real-time quantitative PCR with reverse transcription (qRT-PCR) has several advantages to overcome the limitations of the Northern blot [69]. The starting material can be quite small in quantity. Indeed, combined with laser-capture microdissection (LCM) technology, real-time qRT-PCR can define quantitative gene responses in specific cell populations from frozen tissue sections (Readers can refer to <http://mecko.nichd.nih.gov/lcm/lcm.htm>). Using gene-specific TaqMan Probe or SYBR Green I method, real-time qRT-PCR continuously monitors target PCR product accumulation. This omits labor-intensive quantification steps in conventional quantitative PCR, such as gel electrophoresis or plate-capture hybridization. Real-time qRT-PCR provides accurate and reproducible quantification of transcript abundance compared with conventional quantitative RT-PCR. Other validation approaches such as Western blots, immunohistochemistry, and *in situ* hybridization are also useful.

Endocrine pancreas consortium (EPCon)

EPCon is part of the NIDDK-sponsored consortium on 'Functional Genomics of the Developing Endocrine Pancreas' with cosponsorship from the Juvenile Diabetes Foundation International. Its goal is to create pancreas cDNA libraries to allow identification of new transcripts expressed in this tissue. The Consortium has created 15 pancreatic mouse libraries and 7 pancreatic human libraries, listed in Table 3, along with the ESTs sequenced from each library. All the transcripts collected from these libraries are recorded in the public NCBI dbEST database. The EPCon web site also offers various information about the libraries as well as tools for analysis. The site can be searched for RNA or experimental information, and provides information about the clone sets and the IMAGE clone lists. Various BLAST and analysis tools can also be accessed to search these libraries.

As tabulated by UniGene, the EPCon libraries have provided the identification of over 753 previously unknown genes unique to these libraries to date. All the sequences, once clustered into nonredundant sets of sequences, will then allow building large microarrays that will permit gene expression profile experiments on pancreatic tissues. As seen in Table 3, as of May 23rd, 2002, the Consortium had sequenced over 104 000 mouse and human ESTs available in dbEST for public access.

Future directions

We are in a new era of applied biology generated by technologies such as genomics, proteomics, and

Table 3. Summary of the cDNA libraries constructed by the endocrine pancreas consortium

Mouse	
Total pancreas	
Mouse E10 5 12 5 pancreas cDNA library	1115
Melton amplified mouse E10 5 12 5 pancreas 1 M10S1-A	3 672
Melton mouse E16 5 pancreas library M16Z1	622
Melton amplified mouse E16 5 pancreas 3 M16S1 A	4 484
Melton mouse E16 5 pancreas library 2 M16B2	9 143
Melton mouse newborn pancreas	880
Melton mouse adult pancreas 1	1 036
Melton mouse adult pancreas 2	592
Kaestner ngn3 wt	5 424
Kaestner wt amplified	1 252
Melton normalized mixed mouse pancreas 1 N1-MMS1	23 036
Total	51 256
Islet	
Melton mouse islets MIZ1	431
Amplified Melton mouse islets 1 MIS1-A	3 018
Total	3 449
Transgenic	
Kaestner ngn3 - -	860
Kaestner ngn3 - - subtracted	1 159
Total	2 019
Total: 56 724 ESTs on May 23rd 2002	
Human	
Total pancreas	
Human fetal pancreas 1A	444
Human fetal pancreas 1B	924
Total	1 368
Islet	
Melton human islets HIZ1	560
Human pancreatic islets	1 204
Melton normalized human islet 4 N4-HIS 1	14 109
HR85 islet	21 015
Total	36 888
Insulinoma	
Human insulinoma	10 201
Total: 48 457 ESTs on May 23 rd 2002	

bioinformatics. This information can be used for diabetes research in many ways. Genetic alterations in expression profiles of pancreatic β cells, muscle, liver, and fat tissue result in phenotypic changes. Therefore, mapping the chromosomal location of the genes that encode differentially expressed transcripts found by SAGE or microarrays, may lead to the identification of underlying genetic differences accounting for diabetes. A recent analysis of 2.5 million SAGE tags derived from 12 different normal and cancerous human tissue types revealed that highly expressed genes seem to cluster in particular chromosomal regions. These regions were therefore named regions of increased gene expression (or RIDGES) [70]. The generation of transcriptome maps for tissues from diabetic and nondiabetic individuals could provide tools to identify candidate genes that are overexpressed or silenced in this disease. The hypothesis to be tested is that gene-expression data, in combination with chromosomal position information, may help identify potential genes that could have polymorphisms that alter

function or expression and contribute to the development of diabetes.

Although cell lines may not be suitable for determining gene-expression patterns associated with disease, they are ideal for the identification of downstream targets of a particular gene or signaling pathway that has been proven to be important in genetically modified mouse models. Hence, activation of signaling pathways identified by microarray technology can be used to provide new targets for drug discovery and understand the cellular response to the treatment. Microarrays, typically used to quantify transcriptional activity between samples, could allow identifying genes that are regulated by the cell cycle, stress, growth factor or nutrient stimuli. These methods can be used for example to identify factors implicated in the response to specific diets or metabolic disorders, or to compare transcriptional responses to low and high fat diets. Microarrays can also help identify factors downstream of a knocked out gene by subtracting the expression level represented in the wild type or control animals. These platforms can also help identify genes implicated in cancers and classifying pancreatic tumors. Microarrays allow identifying the signature of the state of the cells and their functional state. The main limitation for islet transplantation is the limited source of tissue. The application of SAGE and microarray technologies for the engineering of β cells will be useful. Identification of gene profiles during pancreas development will provide 'fingerprints' that could be used for the generation of pancreatic β cells *in vitro*. These applications can also be used to understand the different mechanisms involved in the differentiation of stem cells to mature β cells.

Finally the development of annotated databases for gene-expression data, to be shared freely by the diabetes research community, will become available. These databases can include functional annotation from sequence homologies and functional linkages to describe the regulatory nodes of the cellular network, thus allowing the construction of 'transcriptional regulatory networks' characterizing the cells and their conditions. Some tools remain to be created to facilitate data comparability. Standard data formats, for instance, would allow one to automatically compare the expression levels of numbers of transcripts between data sets [71]. Global databases centralizing the results of different experiments would also allow *in silico* analysis of transcription variation in response to cellular stimuli from experiments performed at different times and places [72–74].

Acknowledgments

We gratefully acknowledge the D. Melton lab (Harvard) and K. Kaestner and C. Stoeckert labs (Univ. of Pennsylvania) as well as Sandy Clifton, Hiroshi Inoue, Wes Warren, and the Genome Sequencing Center for their work with EPCOn. The authors would like to thank Mr. Gary Skolnick for preparation of the manuscript. This work was supported in part by N.I.H. grants DK16746, DK56954, and DK99007 (M.A.P.), and the Washington University DRTC.

References

1. Ullrich A, Shine J, Chirgwin J, *et al.* Rat insulin genes: construction of plasmids containing the coding sequences. *Science* 1977; **196**: 1313–1319.
2. Seino S, Seino M, Nishi S, Bell GI. Structure of the human insulin receptor gene and characterization of its promoter. *Proc Natl Acad Sci USA* 1989; **86**: 114–118.
3. Magnuson MA, Andreone TL, Printz RL, Koch S, Granner DK. Rat glucokinase gene: structure and regulation by insulin. *Proc Natl Acad Sci USA* 1989; **86**: 4838–4842.
4. Burant CF, Sivitz WI, Fukumoto H, *et al.* Mammalian glucose transporters: structure and molecular regulation. *Recent Prog Horm Res* 1993; **47**: 349–387.
5. Habener JF, Stoffers DA. A newly discovered role of transcription factors involved in pancreas development and the pathogenesis of diabetes mellitus. *Proc Assoc Am Physicians* 1998; **110**: 12–21.
6. Carninci P, Hayashizaki Y. High-efficiency full-length cDNA cloning. *Methods Enzymol* 1999; **303**: 19–44.
7. Bonaldo MF, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 1996; **6**: 791–806.
8. Zhang H, Stallock JP, Ng JC, Reinhard C, Neufeld TP. Regulation of cellular growth by the drosophila target of rapamycin dTOR. *Genes Dev* 2000; **14**: 2712–2724.
9. Ausubel FM, Brent R, Kingston RE, *et al.* (eds). Production of a subtracted cDNA library. In *Current Protocols in Molecular Biology*, John Wiley & Sons: New York, 2002; Section 25B1.
10. Tedder TF, Strueli M, Schlossman SF, Saito H. Isolation and structure of a cDNA encoding the B1 (CD20) cell-surface antigen of human B lymphocytes. *Proc Natl Acad Sci USA* 1988; **85**: 208–212.
11. PI Neophytou EM, Hutton JC. A subtractive cloning approach to the identification of mRNAs specifically expressed in pancreatic beta-cells. *Diabetes* 1996; **45**: 127–133.
12. Wada J, Kanwar YS. Characterization of mammalian translocase of inner mitochondrial membrane (Tim44) isolated from diabetic newborn mouse kidney. *Proc Natl Acad Sci USA* 1998; **95**: 144–149.
13. Conrad B, Weidmann E, Trucco G, *et al.* Evidence for superantigen involvement in insulin-dependent diabetes mellitus aetiology. *Nature* 1994; **371**: 351–355.
14. Werner Gurr RY, Wen Li, Shaw M, Mora C, Christa L, Sherwin RS. A reg family protein is overexpressed in islets from a patient with new-onset type 1 diabetes and acts as T-Cell autoantigen in NOD mice. *Diabetes* 2002; **51**: 339–346.
15. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992; **257**: 967–971.
16. Liang P, Averboukh L, Pardee AB. Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization. *Nucleic Acids Res* 1993; **21**: 3269–3275.
17. Martin KJP, Arthur B. Identifying expressed genes. *Proc Natl Acad Sci USA* 2000; **97**: 3789–3791.
18. Malhotra K, Foltz L, Mahoney WC, Schueler PA. Interaction and effect of annealing temperature on primers used in differential display RT-PCR. *Nucleic Acids Res* 1998; **26**: 854–856.
19. Chen MC, Schuit F, Eizirik DL. Identification of IL-1beta-induced messenger RNAs in rat pancreatic beta cells by differential display of messenger RNA. *Diabetologia* 1999; **42**: 1199–1203.
20. Chen MC, Schuit F, Pipeleers DG, Eizirik DL. IL-1beta induces serine protease inhibitor 3 (SPI-3) gene expression in rat pancreatic beta-cells. Detection by differential display of messenger RNA. *Cytokine* 1999; **11**: 856–862.
21. Cardozo AK, Kruhoffer M, Leeman R, Orntoft T, Eizirik DL. Identification of novel cytokine-induced genes in pancreatic beta-cells by high-density oligonucleotide arrays. *Diabetes* 2001; **50**: 909–920.
22. Kubo Y, Baldwin TJ, Jan YN, Jan LY. Primary structure and functional expression of a mouse inward rectifier potassium channel. *Nature* 1993; **362**: 127–133.
23. Mashima H, Yamada S, Tajima T, *et al.* Genes expressed during the differentiation of pancreatic AR42J cells into insulin-secreting cells. *Diabetes* 1999; **48**: 304–309.
24. Minami K, Yano H, Miki T, *et al.* Insulin secretion and differential gene expression in glucose-responsive and

- unresponsive MIN6 sublines. *Am J Physiol Endocrinol Metab* 2000; **279**: E773–781.
25. Vicent D, Piper M, Gammeltoft S, Maratos-Flier E, Kahn CR. Alterations in skeletal muscle gene expression of ob/ob mice by mRNA differential display. *Diabetes* 1998; **47**: 1451–1458.
 26. Huang X, Eriksson KF, Vaag A, *et al.* Insulin-regulated mitochondrial gene expression is associated with glucose flux in human skeletal muscle. *Diabetes* 1999; **48**: 1508–1514.
 27. Corominola H, Conner LJ, Beavers LS, *et al.* Identification of novel genes differentially expressed in omental fat of obese subjects and obese type 2 diabetic patients. *Diabetes* 2001; **50**: 2822–2830.
 28. Ferrer J, Wasson J, Schoor KD, Mueckler M, Donis-Keller H, Permutt MA. Mapping novel pancreatic islet genes to human chromosomes. *Diabetes* 1997; **46**: 386–392.
 29. Rafaeloff R, Qin XF, Barlow SW, Rosenberg L, Vinik AI. Identification of differentially expressed genes induced in pancreatic islet neogenesis. *FEBS Lett* 1996; **378**: 219–223.
 30. Nishio Y, Aiello LP, King GL. Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB J* 1994; **8**: 103–106.
 31. Aiello LR, Robinson GS, Lin Y, Nishio Y, King GL. Identification of multiple genes in bovine retinal pericytes altered by exposure to elevated levels of glucose by using mRNA differential display. *Proc Natl Acad Sci USA* 1994; **91**: 6231–6235.
 32. Nishio YW, Buczek-Thomas CE, Rulfs JA, *et al.* Identification and characterization of a gene regulating enzymatic glycosylation which is induced by diabetes and hyperglycemia specifically in rat cardiac tissue. *J Clin Invest* 1995; **96**: 1759–1767.
 33. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995; **270**: 484–487.
 34. Zhang L, Zhou W, Velculescu VE, *et al.* Gene expression profiles in normal and cancer cells. *Science* 1997; **276**: 1268–1272.
 35. Velculescu VE, Vogelstein B, Kinzler KW. Analysing uncharted transcriptomes with SAGE. *Trends Genet* 2000; **16**: 423–425.
 36. Polyak K, Riggins GJ. Gene discovery using the serial analysis of gene expression technique: implications for cancer research. *J Clin Oncol* 2001; **19**: 2948–2958.
 37. Robert-Nicoud M, Flahaut M, Elalouf JM, *et al.* Transcriptome of a mouse kidney cortical collecting duct cell line: effects of aldosterone and vasopressin. *Proc Natl Acad Sci USA* 2001; **98**: 2712–2716.
 38. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**: 467–470.
 39. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet* 1999; **21**: 20–24.
 40. Lockhart DJ, Dong H, Byrne MC, *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996; **14**: 1675–1680.
 41. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol* 1998; **16**: 301–306.
 42. Okamoto T, Suzuki T, Yamamoto N. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat Biotechnol* 2000; **18**: 438–441.
 43. Hughes TR, Mao M, Jones AR, *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001; **19**: 342–347.
 44. Kane MD, Jatke TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* 2000; **28**: 4552–4557.
 45. Everts EM, Au-Young J, Ruvolo MV, Lim AC, Reynolds MA. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques* 2001; **31**: 1182, 1184, 1186.
 46. Lander ES. Array of hope. *Nat Genet* 1999; **21**: 3, 4.
 47. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 2002; **18**: 405–412.
 48. Gygi SP, Rochon Y, Franz BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999; **19**: 1720–1730.
 49. Song F, Srinivasan M, Aalinkel R, Patel MS. Use of a cDNA array for the identification of genes induced in islets of suckling rats by a high-carbohydrate nutritional intervention. *Diabetes* 2001; **50**: 2053–2060.
 50. Busch AK, Cordery D, Denyer GS, Biden TJ. Expression profiling of palmitate- and oleate-regulated genes provides novel insights into the effects of chronic lipid exposure on pancreatic beta-cell function. *Diabetes* 2002; **51**: 977–987.
 51. Webb GC, Akbar MS, Zhao C, Steiner DF. Expression profiling of pancreatic beta cells: glucose regulation of secretory and metabolic pathway genes. *Proc Natl Acad Sci USA* 2000; **97**: 5773–5778.
 52. Dupont J, Khan J, Qu BH, Metzler P, Helman L, LeRoith D. Insulin and IGF-1 induce different patterns of gene expression in mouse fibroblast NIH-3T3 cells: identification by cDNA microarray analysis. *Endocrinology* 2001; **142**: 4969–4975.
 53. Iacobuzio-Donahue CA, Maitra A, Shen-Ong GL, *et al.* Discovery of novel tumor markers of pancreatic cancer using global gene expression technology. *Am J Pathol* 2002; **160**: 1239–1249.
 54. Schuchhardt J, Beule D, Malik A, *et al.* Normalization strategies for cDNA microarrays. *Nucleic Acids Res* 2000; **28**: E47.
 55. Yang YH, Dudoit S, Luu P, *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002; **30**: E15.
 56. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol* 2000; **7**: 819–837.
 57. Tsien CL, Libermann TA, Gu X, Kohane IS. On reporting fold differences. *Pac Symp Biocomput* 2001; 496–507.
 58. Chu TM, Weir B, Wolfinger R. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math Biosci* 2002; **176**: 35–51.
 59. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
 60. Vingron M. Bioinformatics needs to adopt statistical thinking. *Bioinformatics* 2001; **17**: 389, 390.
 61. Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev* 2000; **14**: 963–980.
 62. Didier G, Brézellec P, Remy E, Hénaut A. GeneANOVA-gene expression analysis of variance. *Bioinformatics* 2002; **18**: 490, 491.
 63. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 2002; **18**: 413–422.
 64. Tamayo P, Slonim D, Mesirov J, *et al.* Interpreting patterns of gene expression with self-expression maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999; **96**: 2907–2912.
 65. Schug J, Diskin S, Mazzarelli J, Brunk BP, Stoeckert Jr CJ. Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res* 2002; **12**: 648–655.
 66. van Someren EP, Wessels LF, Reinders MJ. Linear modeling of genetic networks from experimental data. *Proc Int Conf Intell Syst Mol Biol* 2000; **8**: 355–366.
 67. Koza JR, Mydlowec W, Lanza G, Yu J, Keane MA. Reverse engineering of metabolic pathways from observed data using genetic programming. *Pac Symp Biocomput* 2001; 434–445.
 68. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput* 2001; 422–433.
 69. Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 2000; **25**: 169–193.
 70. Caron H, van Schaik B, van der Mee M, *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 2001; **291**: 1289–1292.
 71. Brazma A, Hingamp P, Quackenbush J, *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001; **29**: 365–371.
 72. Kellam P. Microarray gene expression database: progress towards an international repository of gene expression data. *Genome Biol* 2001; **2**: 4011.
 73. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**: 207–210.
 74. Sherlock G, Hernandez-Boussard T, Kasarskis A, *et al.* The Stanford microarray database. *Nucleic Acids Res* 2001; **29**: 152–155.