D542–D547 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh107

CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature

Viviane Praz[1,2], Vidhya Jagannathan[2] and Philipp Bucher[1,2,*]

[1]Swiss Insitiute of Bioinformatics and [2]Swiss Institute for Experimental Cancer Research, Ch. des Boveresses 155, 1066 Epalinges s/Lausanne, Switzerland

Received August 15, 2003; Revised October 3, 2003; Accepted October 13, 2003

ABSTRACT

The main goal of CleanEx is to provide access to public gene expression data via unique gene names. A second objective is to represent heterogeneous expression data produced by different technologies in a way that facilitates joint analysis and cross-data set comparisons. A consistent and up-to-date gene nomenclature is achieved by associating each single experiment with a permanent target identifier consisting of a physical description of the targeted RNA population or the hybridization reagent used. These targets are then mapped at regular intervals to the growing and evolving catalogues of human genes and genes from model organisms. The completely automatic mapping procedure relies partly on external genome information resources such as UniGene and RefSeq. The central part of CleanEx is a weekly built gene index containing cross-references to all public expression data already incorporated into the system. In addition, the expression target database of CleanEx provides gene mapping and quality control information for various types of experimental resource, such as cDNA clones or Affymetrix probe sets. The web-based query interfaces offer access to individual entries via text string searches or quantitative expression criteria. CleanEx is accessible at: http://www.cleanex.isb-sib.ch/.

INTRODUCTION

Gene expression data obtained with a variety of different technologies are published in increasing amounts via local websites or public repositories, such as GEO (1), and are gradually becoming an information resource no less important than genome sequences or protein 3D structures. Their impact on biomedical research has nevertheless so far been limited by the absence of a coherent system to access and analyse all data concerning the same gene via a common interface. CleanEx attempts to fill this gap.

CleanEx is not the only ongoing effort to create a comprehensive public gene expression database, but none of these other projects has precisely identical goals or provides workable solutions to all problems addressed by our system. GEO (1) and ArrayExpress (2) are repositories that rely on an author submission scheme. The former is an archive already populated with an impressive amount of data. However, there is little cross-data set standardization, and advanced query interfaces are lacking. ArrayExpress offers powerful data access and analysis tools but contains few data. Moreover, it is restricted to profiles generated with microarray technology. The Stanford Microarray Database (SMD) (3) is currently restricted to data generated in-house with a particular technology [spotted arrays, ScanAlyze (4) software for signal processing], and is not meant to be an initiative to create a comprehensive public gene expression information resource. Nevertheless, it is perhaps the largest public database of this kind that is fully functional today. SOURCE (5) is a database which, like CleanEx, provides gene-based and clone-based information linked to expression data. The system retrieves expression data for each gene, but does not allow selection of clones via expression pattern comparison.

CleanEx was originally designed to provide links from the Swiss-Prot database (6) and the Eukaryotic Promoter Database (EPD) (7) to public gene expression data through unique and unambiguous gene names. For this purpose, we developed a system that allows dynamic remapping of static expression data to the growing and evolving catalogues of human and other organisms' genes. This is an important issue, not only because many genes are referred to by different names, but also because many gene expression measurements corresponded to unknown or partially characterized genes at the time when the experiments were carried out. The gene annotation provided with public gene expression data is usually static and consequently does not reflect subsequent progress towards functional characterization of all genes of the corresponding organism. When putting in place a coherent system for gene nomenclature, we realized that this is also a prerequisite for joint analysis of multiple data sets and cross-data set comparisons. Mainly for this reason CleanEx has grown in the meantime into an independent database project with a broader scope.

CleanEx is not a direct competitor of any other gene expression database. For one thing, it is very much focused on

*To whom correspondence should be addressed. Tel: +41 21 6925892/58; Fax: +41 21 652 6933; Email: Philipp.Bucher@isrec.unil.ch

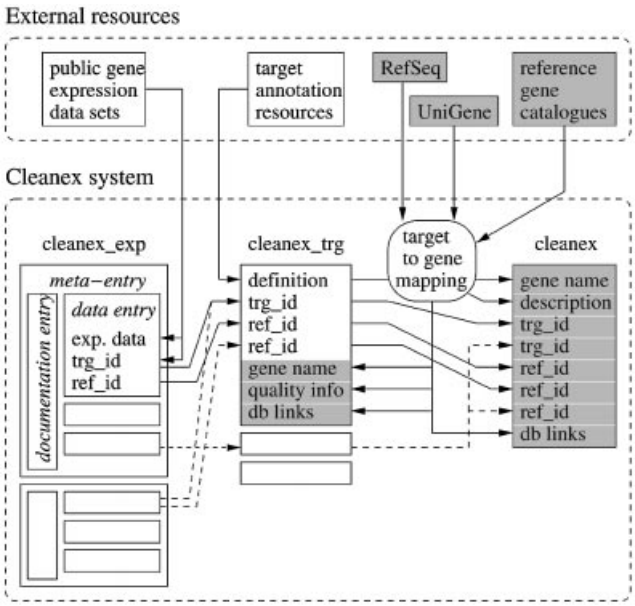

Figure 1. Sources and data flow in the CleanEx system. Shaded boxes represent dynamic data, which are updated on a regular basis, either by external suppliers or by automatic procedures of the CleanEx system. The empty boxes representing additional CleanEx entries, and the broken arrows serve to indicate one-to-many or many-to-one relationships. Note that target entries can also be related to multiple gene entries, but this is not the normal desirable case and thus not explicitly shown in the diagram.

one particular problem not adequately addressed by other databases, namely that of mapping expression data to evolving gene catalogues. Furthermore, it is designed such that it can synergically be used with other resources. For instance, we plan to import raw data from GEO (1) in the future. Moreover, our dynamic annotation of RNA sequence tags and hybridization targets could be imported into other gene expression databases in order to provide up-to-date gene annotation.

OVERVIEW

CleanEx is compiled from a variety of public sources (Fig. 1): (i) electronically published expression data, (ii) information on expression profiling reagents provided by commercial suppliers, (iii) official gene nomenclature databases such as Genew (8) for human and (iv) genome-related public databases such as UniGene (9), RefSeq (10), EPD (7) and Swiss-Prot (6). It is a manually curated database not relying on a direct submission mechanism. The data copied over from public web or FTP sites undergo stringent quality control procedures before they are imported into the CleanEx system by *ad-hoc*-developed data set-specific Perl scripts.

CleanEx attempts to be flexible with regard to data representation in order to remain close to the original source. A variety of formats is used to store different kinds of expression profiles obtained with different technologies. We try to preserve as much information as possible from the original sources. For instance, output files generated by the microarray image-processing program ScanAlyze (4), which contain a rich numerical representation of the signal including many quality control indicators, are kept in a very similar and information-wise equivalent format, allowing for exploitation of all information in subsequent data normalization steps. More basic formats are used when only single expression values or ratios are made available by the authors.

As explained before, dynamic gene nomenclature and annotation are perhaps the most original and important aspects of CleanEx. The mechanism currently in place ensures that all data concerning a particular gene can be retrieved at any time via the same officially approved gene name. This is achieved through indirect linking of expression data to genes via so-called 'expression targets'. The concept of an 'expression target' is central to CleanEx. The term stays for a physical description of an RNA population or the biochemical reagents used to measure the abundance of an RNA population. In practice it consists of something like a sequence tag, a clone name or a set of oligonucleotide sequences from an Affymetrix probe set. The most important property of a gene expression target is that it is stable. Expression data can thus be permanently associated with expression targets (stored in a separate file) which in turn are dynamically linked every week to genes using sequence matching algorithms and other frequently updated information resources such as UniGene (9).

ORGANIZATION, DATA ACQUISITION AND MAINTENANCE PROCEDURES

CleanEx is organized as three interconnected flat files: (i) cleanex_exp containing the expression data, (ii) cleanex_trg providing target information and (iii) cleanex itself, which links genes to entries of the other files. A clear distinction is made between stable and dynamic data fields (Fig. 1). The stable data fields are generated manually or semi-automatically whenever new data are imported from external sources. The dynamic data fields are automatically generated or updated by a weekly procedure which remaps all targets to genes. Note that the cleanex file itself consists exclusively of dynamic fields and thus is rebuilt from scratch every week.

Cleanex_exp

Cleanex_exp contains public gene expression data in a weakly standardized text file format and, if possible, information-wise equivalent to the original sources. It is formatted as a hierarchically structured file which consists of so-called meta-entries, which in turn contain entries. A meta-entry contains a matrix of expression levels for a set of targets and conditions, the data structure, which is typically published and analysed at once, and referred to by a common name. Each meta-entry consists of a documentation entry plus one data entry for each expression target. The documentation entry provides general information about the data set including the list of tissues or conditions for which expression values are provided. A data entry contains expression values for a particular target over all conditions.

The first step in generating a new meta-entry consists of downloading a public data set from an external FTP or website. The source files are archived in a local repository but are not considered to be part of the CleanEx system. The data are then first analysed by the curator and subjected to a number of consistency and quality checks. A decision has to be made at this stage as to what kind of target identifier and expression

Nice View of CleanEx Target : AFFY_HC–G110_1003_s_at

Entry based on UniGene Build #161

[General] [Features] [Expression data]

General information about the entry	
Entry name	**AFFY_HC–G110_1003_s_at**
Feature Type	Affy_Tag
Original annotation	X68149; HSBLR1A Homo sapiens BLR1 gene for Burkitt's lymphoma receptor 1
Organism	Homo sapiens
Gene name(s)	BLR1
Number of matching gene(s)	1
Quality	High
Reference Sequence(s)	RefSeq=NM_001716; Unigene=Hs.113916
	RefSeq=NM_032966; Unigene=Hs.113916

Features of this target : Tag			
TCCATCCACATCACCTGTGGGACCA:	185–179 NM_001716(+)[577..601]	NM_032966(+)[646..670]	Hs.113916
TGGGCTTCCTCCTTGCCTTGCCAGA:	186–179 NM_001716(+)[611..635]	NM_032966(+)[680..704]	Hs.113916
CTTGCCAGAGATTCTCTTCGCCAAA:	187–179 NM_001716(+)[627..651]	NM_032966(+)[696..720]	Hs.113916
CTCTTCGCCAAAGTCAGCCAAGGCC:	188–179 NM_001716(+)[640..664]	NM_032966(+)[709..733]	Hs.113916
TGCCACGTTGCACCTTCTCCCAAGA:	189–179 NM_001716(+)[680..704]	NM_032966(+)[749..773]	Hs.113916
TCCCAAGAGAACCAAGCAGAAACGC:	190–179 NM_001716(+)[697..721]	NM_032966(+)[766..790]	Hs.113916
ACAGGTTGCGCCAGGCCCAGCGGCG:	191–179 NM_001716(+)[812..836]	NM_032966(+)[881..905]	Hs.113916
TCAGGGTGGCCATCCTGGTGACAAG:	192–179 NM_001716(+)[857..881]	NM_032966(+)[926..950]	Hs.113916
CCTGGTGACAAGCATCTTCTTCCTC:	193–179 NM_001716(+)[870..894]	NM_032966(+)[939..963]	Hs.113916
CGTGGACAATACCTGCAAGCTGAAT:	194–179 NM_001716(+)[951..975]	NM_032966(+)[1020..1044]	Hs.113916
CTCCCCGTGGCCATCACCATGTGTG:	195–179 NM_001716(+)[982..1006]	NM_032966(+)[1051..1075]	Hs.113916
CACCATGTGTGAGTTCCTGGGCCTG:	196–179 NM_001716(+)[996..1020]	NM_032966(+)[1065..1089]	Hs.113916
CATGCTCTACACTTTCGCCGGCGTG:	197–179 NM_001716(+)[1041..1065]	NM_032966(+)[1110..1134]	Hs.113916
TTCGCCGGCGTGAAGTTCCGCAGTG:	198–179 NM_001716(+)[1054..1078]	NM_032966(+)[1123..1147]	Hs.113916
AGCTCTTCCCTAGCTGGCGCAGGAG:	199–179 NM_001716(+)[1133..1157]	NM_032966(+)[1202..1226]	Hs.113916
TGGCGCAGGAGCAGTCTCTCTGAGT:	200–179 NM_001716(+)[1147..1171]	NM_032966(+)[1216..1240]	Hs.113916

Expression Reference(s)		
AFFY001_1003_s_at	Gene expression in medulloblastomas	Nat. Genet. 2001 Oct;29(2):143–52.

Figure 2. Example of a CleanEx target entry.

data format will be used. As mentioned above, CleanEx supports a number of different formats for representing gene expression data, from simple sequence tag counts to the rich numerical representation of microarray images produced by programs like ScanAlyze (4). The new meta-entry is then usually generated by an *ad hoc* written Perl script. Sometimes, new expression target entries need to be generated as well and will be added to cleanex_trg.

Cleanex_exp meta-entries are in principle static, meaning that the original data are downloaded once and reformatted once. Exceptions to this rule occur when the authors modify their own data. Another exception to this rule is the meta-entry that contains the tissue distribution of public ESTs, which is derived from Unigene (9) and regenerated from scratch whenever the original source is updated.

Cleanex_exp meta-entries have short alpha-numeric strings as identifiers. Expression data entries have composite identifiers consisting of the meta-entry ID followed by an underscore character and a second identifier. The second identifier is often identical to the corresponding target entry ID. Exceptions occur when the same target has been analysed more than once in a gene expression profiling experiment (for instance if the same cDNA clone has been spotted twice on a microarray).

Cleanex_trg

The entries of this file contain a physical description of the expression targets, links to genes and quality control information. The physical description consists of either a sequence tag or information on a hybridization target, e.g. a cDNA clone or an Affymetrix probe set. The exact content of an entry depends on the target type. Currently we distinguish between: (i) public cDNA clone names included in UniGene, (ii) cDNA clones from private suppliers, e.g. Incyte, (iii) Affymetrix probe sets, (iv) gene names and (v) sequence database accession numbers. The latter two are not true physical descriptions of the expression targets and serve as substitutes when more precise information is lacking. For instance for some data sets generated with commercial oligonucleotide microarrays, we were unable to access the corresponding oligonucleotide sequences and therefore used the sequence accession numbers provided by the authors instead.

Nice View of CleanEx: HS_FGF2

[General] [RNA sequences] [Cross-references] [Expression data]

General information about the entry	
Entry name	**HS_FGF2**
Locus	4q26-q27
Description of the gene	fibroblast growth factor 2 (basic).
Old gene names	FGFB
Cross-references	
LocusLink	2247
Unigene	Hs.284244
MIM	134920
Genew	HGNC:3676; FGF2
GeneCards	FGF2
Ensembl	FGF2
RefSeq	NM_002006
SWISSPROT	P09038; FGF2_HUMAN
Expression Data References	
HSEST	NM_002006; ; HSEST_FGF2. [Entry (text) / Local viewer]
LYMPHOMA1 (Original web site)	R38539; 16115; L0001_16115. [Entry (text) / Original viewer / Local viewer]
NCI60 (Original web site)	W44677; 323776; NCI60_323776. [Entry (text) / Original viewer / Local viewer]
View all NCI60 experiments	W53020; 325559; NCI60_325559. [Entry (text) / Original viewer / Local viewer]
PEROU1 (Original web site)	H19554; 172276; P0001_172276. [Entry (text) / Original viewer / Local viewer]
View all PEROU1 experiments	H45381; 176537; P0001_176537. [Entry (text) / Original viewer / Local viewer]
	T65362; 21775; P0001_21775. [Entry (text) / Original viewer / Local viewer]
	R24315; 34191; P0001_34191. [Entry (text) / Original viewer / Local viewer]
	H12092; 47810; P0001_47810. [Entry (text) / Original viewer / Local viewer]
	H28814; 49657; P0001_49657. [Entry (text) / Original viewer / Local viewer]
ROSETTA (Original web site)	NM_002006; ; R0001_22668. [Entry (text) / Local viewer]
SERUM1 (Original web site)	W44678; 323776; S0001_323776. [Entry (text) / Original viewer / Local viewer]
RNA sequences according to Unigene available on Mar-3-2003	
EMBL	AF086208.1; AF086208. [EMBL / GenBank / DDBJ]
	AJ250952.1; HSA250952. [EMBL / GenBank / DDBJ]
	J04513.1; HSGFB. [EMBL / GenBank / DDBJ]
	M17599.1; HSGFBF. [EMBL / GenBank / DDBJ]
	M27968.1; HSFGFB. [EMBL / GenBank / DDBJ]
	S47380.1; S47380. [EMBL / GenBank / DDBJ]

Figure 3. Example of a CleanEx entry.

The stable parts of a cleanex_trg entry are usually imported from external sources, such as the probe set documentation files posted by Affymetrix (http://www.affymetrix.com/analysis/download_center.affx). The dynamic parts are generated by the weekly updating procedure whose primary purpose is to link targets to genes. For public cDNA clones, sequence accession numbers and gene symbols, these links are established directly on the basis of Unigene. This is possible, because Unigene entries contain references to cDNA clones, sequence accession numbers and gene names. For all other target types, the sequences given in the target description are first mapped to mRNA sequences of RefSeq (10) by Blast (11) or by in-house-developed tag-matching software. Then the RefSeq identifiers are used to map the target via Unigene to the gene name.

Target-to-gene mapping is not always successful. As long as the reference gene catalogue of a model organism is incomplete, some targets will not match any gene. Other targets may hit multiple genes. The latter happens for instance with chimeric cDNA clones or ill-designed Affymetrix probe sets. In such cases, the cleanex_trg entry lists all corresponding genes found but adds a quality-control flag to indicate that the mapping is ambiguous. The weekly target-to-gene mapping procedure thus also serves to add quality-control information to the target entry. Note in this context, that >12% of the probe sets in the latest human HG-U133A gene chip from Affymetrix hit multiple genes, and 26% do not match any sequence in RefSeq.

Target entries are typically identified by the names of the corresponding reagents, e.g. an IMAGE clone number or an Affymetrix probe set name. An example of a target entry is shown in Figure 2.

Cleanex

Cleanex is a catalogue of officially approved genes from model organisms with cross-references to entries in

Table 1. CleanEx accessibility

General pages	
home page	http://www.cleanex.isb-sib.ch/index.html
user manual	http://www.cleanex.isb-sib.ch/current/CleanEx_manual.html
list of data sets	http://www.cleanex.isb-sib.ch/datasets.html
Query forms	
for gene entries	http://www.cleanex.isb-sib.ch/cleanex_query_form.html
for target entries	http://www.cleanex.isb-sib.ch/cleanex_trg_query.html
expression query form	http://www.cleanex.isb-sib.ch/AFFY002_expression_form.html
Hyperlinks to individual entries	
gene entry	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex&format=nice&entry=HS_FGF2
expression data set documentation entry	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_ref&format=nice&entry=P0001_DOC
expression data, type Basic-Ratio	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_ref&format=html&entry=R0001_437
expression data, type Counts	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_ref&format=html&entry=HSEST_TP53
expression data, type Affy_Probesets	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_ref&format=html&entry=AFFY001_1593_at
expression data, type Stanford_Scanalyze	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_ref&format=html&entry=S0001_325559
expression data multiviewer, type Basic_Ratio	http://www.cleanex.isb-sib.ch/cgi-bin/exp_query_result.pl?experiment=R0001&org=HS&gene=CBX6&desc=chromobox_homolog_6
expression data multiviewer, type Affy_Probeset	http://www.cleanex.isb-sib.ch/cgi-bin/exp_query_result.pl?experiment=AFFY002&gene=ABCC5&org=HS&desc=ATP-binding_cassette_C5
expression data multiviewer, type Stanford_Scanalyze	http://www.cleanex.isb-sib.ch/cgi-bin/exp_query_result.pl?experiment=P0001&gene=FGF2&org=HS&desc=Fibroblast_growth_factor_2
target entry, type Clone	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_trg&format=nice&entry=IMAGE_325559
target entry, type Affy_Probeset	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_trg&format=nice&entry=AFFY_HC-G110_1593_AT
target entry, type Sequence AC	http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex_trg&format=nice&entry=NM_002006

cleanex_trg and cleanex_exp, and links to external databases. There is one entry per gene, regardless of whether there are corresponding expression data in cleanex_exp. This file is completely rebuilt from scratch every week synchronously with the remapping of expression targets to genes. The process starts with a compilation of officially approved gene names from the reference gene catalogues, e.g. Genew (8) for human. These names are then used to establish cross-references to cleanex_trg entries and from there to expression data in cleanex_exp. The gene names are further used to establish links to external databases, currently Unigene (9), LocusLink and RefSeq (10), MIM (12), Swiss-Prot (6) and EPD (7).

The identifier of a cleanex entry consists of the species code followed by an underscore character and the gene name from the reference catalog. An example of a cleanex entry is shown in Figure 3.

ACCESS AND INTERFACES

All information in CleanEx can be accessed via web interfaces. The cleanex and cleanex_trg flatfiles can also be downloaded from our FTP site (ftp://ftp.epd.unil.ch/pub/databases/CleanEx/). Cleanex_exp is not redistributed by FTP due to copyright concerns. Table 1 contains a comprehensive list of useful URLs and hyperlinks to interesting examples. The expression data viewers use different colour scales for displaying expression patterns based on different technologies. For instance, ratio measurements generated with the microarray technology developed at Stanford, are visualized with the same green–black–red colour scheme used by the SMD (3) web interface. A grey-scale is used to represent a digital expression pattern, for instance a tissue breakdown of EST sequences pertaining to a particular gene. There is also a multitarget viewer capable of displaying multiple expression patterns for the same gene from the same Cleanex_exp meta-entry.

For cleanex and cleanex_trg, there is a text-based query form helping the user to find the desired documents. Expression data sets (meta-entries) from cleanex_exp can be searched for genes satisfying certain expression criteria via data-type-specific expression query forms. These web pages are interconnected in a way that makes it easy to use multiple data sets for selecting genes of interest. For instance, it is possible to search two breast cancer sets in succession, e.g. those from Perou *et al.* and van't Veer *et al*. (13,14), for genes that are expressed specifically in tumours, and subsequently combine the resulting gene lists.

CURRENT LIMITATIONS AND FUTURE PROSPECTS

CleanEx is still in a development phase. The most serious current limitation is that so far relatively few data sets have been incorporated. Nevertheless, CleanEx is currently the only web-based system offering an easy way to make hyperlinks from gene or protein sequences to multiple expression patterns. Such hyperlinks are currently provided by the ExPaSy (http://www.expasy.org) and EPD (http://www.epd.isb-sib.ch) servers and they are already useful, despite the fact that the cross-referencing to public expression data is not

comprehensive. The target database is more advanced and at that moment perhaps the most useful part of the CleanEx system. Currently, we offer up-to-date gene annotation and quality control information for all human Affymetrix probe sets, and for a variety of cDNA clone resources, including clones from Incyte and from the IMAGE consortium (15).

Future efforts will focus primarily on the incorporation of new data sets. We hope to be able to speed up this process by importing data from raw data repositories such as GEO (1) rather than from the original sources. We further plan to improve the query mechanisms based on expression data to allow for more sophisticated types of cross-data set comparison. Finally we intend to add interfaces for downloading numerical expression data from cleanex_exp in various formats appropriate for subsequent import into statistical analysis software packages. These exporting and reformatting mechanisms will probably be combined with different data standardization methods.

ACKNOWLEDGEMENTS

CleanEx is funded by grants from the Swiss government and the Swiss National Sciences Foundation (31-063933).

REFERENCES

1. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
2. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
3. Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Weng,S., Jin,H., Ball,C.A. *et al.* (2001) The Stanford Microarray Database. *Nucleic Acids Res.*, **29**, 152–155.
4. Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.
5. Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D., Brown,P.O. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.
6. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
7. Schmid,C.D., Praz,V., Delorenzi,M., Périer,R. and Bucher,P. (2004) The Eukaryotic Promoter Database EPD. The impact of *in silico* primer extension. *Nucleic Acids Res.*, **32**, D82–D85.
8. Wain,H.M., Bruford,E.A., Lovering,R.C., Lush,M.J., Wright,M.W. and Povey,S. (2002) Guidelines for human gene nomenclature. *Genomics*, **79**, 464–470.
9. Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
10. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
13. Perou,C.M., Jeffrey,S.S., van de Rijn,M., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.
14. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
15. Lennon,G.G., Auffray,C., Polymeropoulos,M. and Soares,M.B. (1996) The I.M.A.G.E. Consortium: An Integrated Molecular Analysis of Genomes and their Expression. *Genomics*, **33**, 151–152.