



**CleanEx: a database of heterogeneous gene expression data
based on a consistent gene nomenclature and linked to an
improved annotation system.**

Thèse de doctorat en sciences de la vie (PhD)

Présentée à la

Faculté de Biologie et de Médecine
de L'Université de Lausanne

par

Viviane Praz

Diplômée en Biologie de l'Université de Genève

Jury :

Prof. Yves Poirier, Président du jury et Rapporteur

Dr. Philipp Bucher, Directeur de thèse

Dr. Laurent Duret, Expert

Dr. Richard Iggo, Expert

Dr. Pierre Gönczy, Expert

LAUSANNE 2005



Institut Suisse de Recherches Experimentales sur le Cancer
Institut Suisse de Bioinformatique



**CleanEx: a database of heterogeneous gene expression data
based on a consistent gene nomenclature and linked to an
improved annotation system.**

Thèse de doctorat en sciences de la vie (PhD)

Présentée à la

Faculté de Biologie et de Médecine
de L'Université de Lausanne

par

Viviane Praz

Diplômée en Biologie de l'Université de Genève

Jury :

Prof. Yves Poirier, Président du jury et Rapporteur

Dr. Philipp Bucher, Directeur de thèse

Dr. Laurent Duret, Expert

Dr. Richard Iggo, Expert

Dr. Pierre Gönczy, Expert

LAUSANNE 2005

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

<i>Président</i>	Monsieur Prof.	Yves Poirier
<i>Directeur de thèse</i>	Monsieur Dr	Philipp Bucher
<i>Rapporteur</i>	Monsieur Prof.	Yves Poirier
<i>Experts</i>	Monsieur Dr	Laurent Duret
	Monsieur Dr	Richard Iggo
	Monsieur Dr	Pierre Gönczy

le Conseil de Faculté autorise l'impression de la thèse de

Madame Viviane Praz

Biologiste diplômée de l'Université de Genève

intitulée

**CleanEx: a database of heterogeneous gene expression data based
on a consistent gene nomenclature
and linked to an improved annotation system**

Lausanne, le 9 décembre 2005

pour Le Doyen
de la Faculté de Biologie et de Médecine

Prof. Yves Poirier



Remerciements

Tout d'abord, tout ce projet n'aurait pas pu voir le jour sans le nécessaire concours de Rouaïda Périer, qui a su, comme toujours, trouver les bons arguments pour me permettre d'intégrer le groupe du Dr Bucher. Je vous remercie infiniment, Jacques et toi, pour votre gentillesse, votre amitié, et votre soutien de tous les instants.

Je tiens à remercier le Dr. Philipp Bucher, qui m'a donné l'opportunité de me lancer dans la bioinformatique et m'a accordé dès mon arrivée, et ce malgré mon expérience toute relative dans ce domaine, une très grande confiance.

Je remercie le Dr. Richard Iggo et le Dr. Pierre Farmer, dont les précieuses indications m'ont permis d'orienter le développement de CleanEx afin d'en faciliter l'utilisation et d'en améliorer les résultats.

Merci également à Monique Zahn, qui a accepté de corriger ce document.

Merci à tous mes collègues, passés et présents, de l'Institut Suisse de Bioinformatique d'Epalinges, pour leurs conseils, leur enthousiasme, leur soutien, ainsi que pour l'ambiance chaleureuse qui a régné au sein de ce groupe. J'y ai passé de très grands moments, et j'y ai rencontré des gens extraordinaires, tant au niveau scientifique qu'au niveau humain. Merci notamment à Marco Pagni et Thomas Junier de m'avoir transmis leur enthousiasme pour la programmation en perl. Merci à Christian Iseli, qui a toujours fait preuve d'une disponibilité sans bornes, en tant que collègue, mais aussi et surtout en tant qu'ami. Un merci tout particulier à Pascale Anderle, pour les discussions scientifiques, ainsi que pour le très apprécié et tout aussi important soutien moral. Et enfin, je remercie aussi Alix, Luli, Grégory, Thierry, Manu et Gautier pour m'avoir supportée, dans tous les sens du terme, tout au long de ces années.

Enfin, merci à ma famille, pour leur présence et leur appui constant.

“Forty-two !”

Deep Thought, in the “Hitch-hiker’s Guide to the Galaxy”

by Douglas Adams

Part of this work and related projects have been published in the following papers :

1. **Praz V, Perier R, Bonnard C, Bucher P.** (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.* **30**: 322-4.
2. **Ambrosini G, Praz V, Jagannathan V, Bucher P.** (2003) Signal search analysis server. *Nucleic Acids Res.* **31**: 3618-20.
3. **Praz V, Jagannathan V, Bucher P.** (2004) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res.* **32**: D542-7.

The CleanEx Database has also been presented during the following conferences :

1. Human Genome Meeting (HGM 2003), April 27-30 2003 Cancùn, Mexico.
CleanEx : a gateway to public gene expression data via officially approved gene names.
Viviane Praz, Philipp Bucher.
2. European Conference on Computational Biology (ECCB 2003), September 27-30, 2003, Paris, France.
CleanEx : a Gene Expression Reference Database.
Viviane Praz, Philipp Bucher.

TABLE OF CONTENTS

1. INTRODUCTION	19
1.1. <i>From sequence to gene expression : how static information becomes dynamic</i>	19
2. UNDERSTANDING RAW DATA	22
2.1. Genomic Data : storage and annotation	22
2.1.1. <i>Genome annotation</i>	23
2.1.2. <i>Sequence clustering</i>	24
2.2. Gene expression data generation	26
2.2.1. <i>Microarrays</i>	27
2.2.2. <i>Affymetrix</i>	29
2.2.3. <i>SAGE</i>	32
2.2.4. <i>MPSS</i>	34
2.2.5. <i>ESTs</i>	35
3. EXPRESSION DATABASES : HISTORY AND EVOLUTION	38
3.1. <i>Historical context : setup of MIAME standards</i>	38
3.2. <i>Emergence of expression databases</i>	39
3.3. Main expression data repositories	42
3.3.1. <i>SMD : the Stanford Microarray Database</i>	42
3.3.2. <i>CGAP and SAGEmap</i>	43
3.3.3. <i>ExpressDB</i>	44
3.3.4. <i>MGED recommended expression data repositories</i>	45
3.3.4.1. <i>GEO and other data repositories</i>	45
3.4. Genes oriented databases	47
3.4.1. <i>GeneCards</i>	48
3.4.2. <i>SOURCE</i>	49

3.4.3. <i>CleanEx</i>	50
4. THE <i>CleanEx</i> DATABASE : CONCEPT AND DATA ORGANIZATION	51
4.1. <i>CleanEx_exp</i>	52
4.2. <i>CleanEx_trg</i>	53
4.3. <i>CleanEx</i>	56
5. BUILDING <i>CleanEx</i>	57
5.1. <i>CleanEx_exp</i> files.....	57
5.2. <i>CleanEx_trg</i> files.....	68
5.3. <i>CleanEx</i> : link file between external databases and the <i>CleanEx</i> system.....	79
5.3.1. <i>Material</i> : source databases.....	79
5.3.2. <i>Data integration method</i>	84
6. RESULTS	88
6.1. <i>Survey of the most recent release</i>	88
6.2. <i>Database format</i>	89
6.2.1. <i>CleanEx</i>	89
6.2.2. <i>CleanEx_exp</i>	93
6.2.2.1. <i>Documentation entry</i>	93
6.2.2.2. <i>Expression data entries</i>	95
6.2.3. <i>CleanEx_trg</i>	96
6.2.4. <i>Additional format : XML version of <i>CleanEx</i> for Integr8</i>	100
6.2.5. <i>Specific formats for web applications</i>	101
6.3. <i>Indexes and retrieval system</i>	102
6.4. <i>Web-based interfaces</i>	103
6.4.1. <i>Entry search engines and viewers</i>	103
6.4.1.1. <i>Single entry search engines</i>	103
6.4.1.2. <i>CleanEx viewer</i>	104
6.4.1.3. <i>CleanEx_Exp</i> : expression viewer.....	105
6.4.1.4. <i>CleanEx_trg</i>	108
6.4.1.4.1. <i>Single entry retrieval</i>	108

6.4.1.4.2. Batch search for CleanEx_trg.....	109
6.4.2. <i>Cross dataset analysis</i>	110
6.4.2.1. Step-by-step expression pattern search.....	111
6.4.2.2. Common genes retrieval.....	111
6.4.2.3. By class expression pattern search.....	112
6.5. <i>Using CleanEx : examples and applications, a CleanEx tutorial</i>	114
6.5.1. <i>CleanEx single entries and multiviewer</i>	114
6.5.2. <i>Normal tissues : comparison of two dataset types</i>	122
6.5.2.1. <i>Astrocytomas and astrocytic gliomas comparison</i>	125
6.5.2.2. <i>Single dataset and sequence extraction : using SSA</i>	128
6.5.3. <i>By class expression pattern search</i>	128
6.5.4. <i>Finding common expression patterns in different datasets</i>	121
6.6. <i>CleanEx external applications</i>	130
6.6.1. <i>IO</i>	131
6.6.2. <i>DNA Chip Splice Machine</i>	131
7. DISCUSSION	133
7.1. <i>General considerations</i>	133
7.2. <i>Advantages and drawbacks of CleanEx</i>	136
8. FUTURE DEVELOPMENTS	140
8.1. <i>Web interfaces</i>	140
8.1.1. Single CleanEx entries.....	140
8.1.2. Targets and annotation retrieval data.....	140
8.2. <i>Expression data analysis</i>	140
8.3. <i>Update, database formats and database growth</i>	141
8.3.1. Update procedure : towards a new database format ?.....	141
8.3.2. MAGE-ML : giving access to raw data in standard exchange format.....	141
8.3.3. New datasets incorporation : adapting the GEO automatic procedure.....	141
9. REFERENCES	143
10. SUPPLEMENTARY TABLES	150

CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature and linked to an improved annotation system.

Viviane Praz,

Swiss Institute for Experimental Cancer Research and Swiss Institute of Bioinformatics

The automatic genome sequencing and annotation, as well as the large-scale gene expression measurements methods, generate a massive amount of data for model organisms. Searching for gene-specific or organism-specific information throughout all the different databases has become a very difficult task, and often results in fragmented and unrelated answers. The generation of a database which will federate and integrate genomic and transcriptomic data together will greatly improve the search speed as well as the quality of the results by allowing a direct comparison of expression results obtained by different techniques.

The main goal of this project, called the CleanEx database, is thus to provide access to public gene expression data via unique gene names and to represent heterogeneous expression data produced by different technologies in a way that facilitates joint analysis and cross-dataset comparisons. A consistent and up-to-date gene nomenclature is achieved by associating each single gene expression 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 genes from model organisms, such as human and mouse. 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 resources, such as cDNA clones or Affymetrix probe sets. The Affymetrix mapping files are accessible as text files, for further use in external applications, and as individual entries, via the web-based interfaces. The CleanEx web-based query interfaces offer access to individual entries via text string searches or quantitative expression criteria, as well as cross-dataset analysis tools, and cross-chip gene comparison. These tools have proven to be very efficient in expression data comparison and even, to a certain extent, in detection of differentially expressed splice variants.

The CleanEx flat files and tools are available online at: <http://www.cleanex.isb-sib.ch/>.

CleanEx: une base de données fédérant des expériences hétérogènes de mesure d'expression de gènes grâce à une nomenclature cohérente et à un système d'annotation efficace.

Viviane Praz,

Institut Suisse de Recherche Expérimentale sur le Cancer, Institut Suisse de Bioinformatique

L'automatisation du séquençage et de l'annotation des génomes, ainsi que l'application à large échelle de méthodes de mesure de l'expression génique, génèrent une quantité phénoménale de données pour des organismes modèles tels que l'homme ou la souris. Dans ce déluge de données, il devient très difficile d'obtenir des informations spécifiques à un organisme ou à un gène, et une telle recherche aboutit fréquemment à des réponses fragmentées, voir incomplètes. La création d'une base de données capable de gérer et d'intégrer aussi bien les données génomiques que les données transcriptomiques peut grandement améliorer la vitesse de recherche ainsi que la qualité des résultats obtenus, en permettant une comparaison directe de mesures d'expression des gènes provenant d'expériences réalisées grâce à des techniques différentes.

L'objectif principal de ce projet, appelé CleanEx, est de fournir un accès direct aux données d'expression publiques par le biais de noms de gènes officiels, et de représenter des données d'expression produites selon des protocoles différents de manière à faciliter une analyse générale et une comparaison entre plusieurs jeux de données. Une mise à jour cohérente et régulière de la nomenclature des gènes est assurée en associant chaque expérience d'expression de gène à un identificateur permanent de la séquence-cible, donnant une description physique de la population d'ARN visée par l'expérience. Ces identificateurs sont ensuite associés à intervalles réguliers aux catalogues, en constante évolution, des gènes d'organismes modèles. Cette procédure automatique de traçage se fonde en partie sur des ressources externes d'information génomique, telles que UniGene et RefSeq. La partie centrale de CleanEx consiste en un index de gènes établi de manière hebdomadaire et qui contient les liens à toutes les données publiques d'expression déjà incorporées au système. En outre, la base de données des séquences-cible fournit un lien sur le gène correspondant ainsi qu'un contrôle de qualité de ce lien pour différents types de ressources expérimentales, telles que des clones ou des sondes Affymetrix. Le système de recherche en ligne de CleanEx offre un accès aux entrées individuelles ainsi qu'à des outils d'analyse croisée de jeux de données. Ces outils se sont avérés très efficaces dans le cadre de la comparaison de l'expression de gènes, ainsi que, dans une certaine mesure, dans la détection d'une variation de cette expression liée au phénomène d'épissage alternatif.

Les fichiers et les outils de CleanEx sont accessibles en ligne (<http://www.cleanex.isb-sib.ch/>).

LIST OF ABBREVIATIONS

BAC : Bacterial Artificial Chromosome
BLAST : Basic Local Alignment Search Tool
BLAT : BLAST-Like Alignment Tool
CGAP : Cancer Genome Anatomy Project
DDBJ : DNA DataBank of Japan
DNA : deoxyribonucleic acid
DoTS : Database of Transcribed Sequences
EBI : European Bioinformatics Institute
EMBL : European Molecular Biology Laboratory
EPD : Eukaryotic Promoter Database
ERA: Estimated Relative Abundances
EST : Expressed Sequence Tag
GEO : Gene Expression Omnibus
GSS : Genome Survey Sequences
HGNC : Hugo Gene Nomenclature Committee
HTC : High Throughput cDNAs
HTG : High Throughput Genome
HUGO : Human Genome Organization
IFOM : Istituto FIRC di Oncologia Molecolare
IO : ISREC Ontologizer
ISREC : Institut Suisse de Recherches Experimentales sur le Cancer
MAGE : Microarray Gene Expression
MAGE-ML : MAGE Markup Language
MAGE-OM : MAGE Object Model
MAGE-stk : MAGE software toolkit
MAS4/MAS5 : Microarray Suite 4/5
MGED : Microarray Gene Expression Data
MGI : Mouse Genome Informatics
MGI : Mouse Genome Institute
MIAME : Minimum Information About a Microarray Experiment
MPSS : Massively Parallel Signature Sequencing

mRNA : messenger RNA
NCBI : National Center for Biotechnology Information
NCI : National Cancer Institute
ORF : Open Reading Frame
PAC : P1-derived Artificial Chromosome
PCR : Polymerase Chain Reaction
PM/MM : Perfect Match/MisMatch
POL II : Polymerase II
polyA: poly-Adenylation tail
RMA : Robust Multichip Average
RNA : RiboNucleic Acid
rRNA : ribosomal RNA
SAGE : Serial Analysis of Gene Expression
scRNA : small cytoplasmic RNA
SIB : Swiss Institute of Bioinformatics
SMD : Stanford Microarray Database
snoRNA : small nucleolar RNA
SNP : Single Nucleotide Polymorphism
snRNA : small nuclear RNA
SOM : Self-Organizing Maps
SQL : Structured Query Language
TPM : Tags Per Million
tRNA : transfert RNA
UCSC : University of California, Santa Cruz
UTR : UnTranslated Region
WHO: World Health Organization
XML : eXtended Markup Language

1. INTRODUCTION

1.1. From sequence to gene expression : how static information becomes dynamic

The recent emergence of very efficient and high-throughput techniques for either DNA sequencing, gene expression measurements, and protein structure determination or quantitation is producing an amount of data that is about to reach unexpected levels. The organization and retrieval of these data for the research community is a challenge that biologists, informaticians and bioinformaticians have to face together. The large volume of data generated is usually stored in specialized databases for each data type. For example, GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>), EMBL (<http://www.ebi.ac.uk/embl/>), and DDBJ (<http://www.ddbj.nig.ac.jp/>) are the three common resources for nucleotide sequence storage and access. Uniprot (<http://www.expasy.uniprot.org/>) is the universal protein sequence database. The large-scale sequencing projects, as well as the recent use of high-throughput expression measurements methods, have also generated specialized sequence or expression databases. Nowadays, more than five hundred different biological type-specific publicly accessible databases have officially been reported [1]. Amongst these, nearly half are of the genomic or gene expression type.

The “genomic” data type, as produced by the various genome projects, is a linear DNA (deoxyribonucleic acid) polymer consisting of four basic nucleotides (A, C, G, T) repeated non-randomly in strings of up to several hundreds of millions (the length of a chromosome). This linear genomic sequence information encodes, in smaller unit called genes, the range of responses that an organism can deploy to cope with its environment but is itself largely static. Indeed, apart from localized DNA mutations, the organism-specific genome itself does not change, but the information derived from it does. Each gene can be transcribed, or expressed, to produce mRNA (messenger ribonucleic acid). Regulation of gene transcription occurs through a complex feedback mechanism involving a number of pathways, ultimately being mediated by transcription factors, protein complexes that bind to short regulatory sequences of DNA near the start of transcription. The correct annotation of genomic sequences, for example the compilation of the genes positions together with their respective

regulatory elements along the linear DNA information, is the first step to ensure a correct analysis of gene expression measurements.

In contrast to the static genomic DNA, gene expression is the dynamic response of the genome to the cells' environment or specificity. In single-cell organisms, the function of gene control is mainly to adjust the enzymatic machinery of the cell in response to its immediate nutritional and physical environment.

For a multicellular organism, the morphological characteristics, as well as the different tissue functions, are mainly determined by the control of gene expression. Indeed, as the cells face a much more stable environment, the genes influenced by environmental changes represent a much smaller proportion than in single-cell organisms. Genes whose expression is controlled to take place at a precise time during the life cycle of an organism are said to be under temporal control, whereas genes expressed in a specific tissue or cell type are under spatial control (tissue-specific genes). Many genes are both temporally and spatially controlled, meaning that they are expressed in a specific tissue at a precise stage of development of the tissue. The organism's answer to different environmental signals, such as exposure to a chemical substance or physiological stress, also consists mainly of changes in gene expression.

Measurement of the gene expression level involves mainly two different steps. First, one has to isolate a unique mRNA in a complex sample that is harvested under specific biological conditions. Then, the respective quantity of each unique mRNA is measured and the behavior of the corresponding gene under these specific conditions can be evaluated. The techniques to measure quantities of mRNA are many, and range from the single-gene measurement (Northern blot [2]) to the large-scale analysis (microarrays [3], SAGE [4], MPSS [5], EST counts [6], Affymetrix GeneChips [7], and so on), capable of quantifying the expression level of all the genes present in one sample at once. One expression data experiment can thus be described as a biological sample's "screenshot", at a certain time, for a certain state, generated via one certain experimental protocol.

The storage of all these biological and experimental conditions, together with the numerical data engendered, is absolutely necessary for further comprehension and analysis. As such, gene expression data require more descriptive information (meta-data) to characterize it accurately, and this adds a new

dimension to the analysis. The resulting data are applicable to a very wide range of biological domains, according to the chosen protocol and the selected sample, such as biological network description, tissue or cancer type classification, effects of different treatments on gene expression, developmental gene expression, or even clinical prognosis based on differential expression patterns.

By combining gene expression measurements with genome annotation, one will reach the point where the whole genome of a certain organism is separated into functional units, namely genes, associated with their respective regulatory elements. For each of these units, the exact location on the genome, the structure, the function, as well as the precise expression level under different conditions will be defined. Merging all these data together in a gene-oriented way will then lead to a holistic view of the genetic mechanisms implied in the organism's response to different environmental changes, or in the organism's development. All these reflexions prompted us to generate the CleanEx [8] model database, as a way to solve this ambitious data integration and analysis problem.

The remaining part of this document is organized as follows. Chapter two gives a general description of the raw data which constitute the basis of the CleanEx system, namely genomic and gene expression data. In chapter three, a brief historical review of existing databases is presented, with the main concepts underlying such databases and the main points which distinguish them from CleanEx. Chapter four explains in detail the CleanEx database system and organization. This leads to chapter five, which describes the steps needed to build the CleanEx database. The source databases used in the procedure are detailed in this section. Chapter six presents the final version of the database format and the different tools which have been associated to CleanEx. A short tutorial gives some examples on how to use the information contained in this database via the web interfaces. The next chapter is a discussion about the advantages and drawbacks of the database, and the steps that could be taken to try to solve them. The last chapter gives some hints and new development ideas for the database format, data retrieval, and data representation.

2. UNDERSTANDING RAW DATA

2.1. Genomic data : storage and annotation

The nucleotide sequences which are stored in nucleotide databases differ mainly in the way they have been produced and can be roughly divided into three categories. First, one finds sequences which correspond to well-characterized genes, which have been individually sequenced and annotated according to the results of biological experiments. These sequences are of very good quality, but are usually short, and each one generally corresponds to a maximum of one gene, with or without its promoter region. The second sequence category (HTG and HTC, that is High Throughput Genome and High-Throughput cDNAs sequence respectively) contains pieces of DNA coming from high-throughput sequencing methods. These include long DNA sequences, like BAC or PAC clones. Most of these clones come from genome sequencing projects and are thus of very high quality, but they lack biological annotations. The last category represents the ESTs (Expressed Sequence Tags) or GSS (Genome Survey Sequences) which are generated by single pass sequencing of clones extracted from cDNA libraries prepared from specific cell samples, or genomic sequences, respectively. The EST sequences deposited in the databases come from a single-read sequencing procedure which mainly keeps the 3' or 5' ends of the transcript having a high reading quality score. This sequence type is thus usually short, and contains a sequencing error rate of about one percent. Nevertheless, these are now the most abundant sequences in the nucleotide databases.

The evolution of the ratio between these different sequence categories reflects the turn that has been taken these past decades in the genomic field. Indeed, the automation of the sequencing technique has allowed the release of complete genome sequences, from viral and bacterial genomes to the genome of higher eukaryote model organisms, like *Drosophila*, mouse, or even human [9]. With the release of whole genomes the new challenge is now to extract useful information from these raw DNA sequences by generating automatic sequence annotations.

The growing number of EST sequences in the databases and the multiple uses that these data can be put to have pushed forward new ways to annotate the genome, for example by using sequence clustering methods so as to determine the exact position of the genes on the genome sequence [10].

2.1.1. Genome annotation

The non-trivial task of finding the exact position of the genes on a genome can be based on two main principles [11].

First, one can make use of known sequence signals which occur at a certain distance from the gene to determine potential gene location. These signals could for example correspond to promoters, polyA sites, translation initiation, coding regions predicted on the basis of hexamer frequencies, codon usage, and for eukaryotic organisms, splice donors and acceptors. This represents the so-called *ab-initio* gene finding.

The second method is based on sequence similarity. One can find coding regions by aligning the raw sequence with for example known mRNAs, EST data, or protein sequences. A similarity search between different species can also be used to find orthologous genes.

Of course, each of these methods has its pitfalls, but the most important ones are common to both. The genes are not always absolutely linearly separated from each other on the DNA sequence. Genes are found on both strands in the same region, and even worse, some genes overlap. This problem occurs in prokaryotic organisms as well as in eukaryotic ones. Nevertheless, it is easier to deal with prokaryotes, as the gene density in these genomes is much higher than in eukaryotes and, as they have no splicing mechanism, the noise due to elements other than genes found in higher organisms, like repetitive elements, microsatellites, and so on, is reduced. The eukaryote-specific problem of splicing and alternative splicing mechanisms, which consists of cutting the pre-mRNA in introns which are discarded and joining exons together for further translation, can generate multiple transcripts from one single gene, sometimes in a tissue-specific way. This of course creates variable alignments for the second prediction method, and could also generate wrongly assigned exons to a specific transcript for the first method. The result is the same for small genes which are located in introns of longer genes, and

which could be missed, or interpreted as a part of the longer gene. For all these reasons, and because the genomic annotation tools as well as the gene prediction methods are far from being perfect, it is still quite difficult to estimate the number of genes for one organism. For example, since the release of the draft version of the human genome until now, the total number of human genes has been re-estimated from more than 100'000 genes to between 20 to 25'000 [12]. Of course, according to the method used, the gene number is different; the ab-initio methods usually overestimate the number of genes, while methods based on similarity tend to underestimate this number, as they use sequences which have already been annotated to identify genes.

Once the genes have been positioned correctly on the genome, one can then annotate the raw sequence with this information. Supplementary information integrated in the sequence typically include the presence of repetitive elements, low-complexity regions, specific protein-binding regions, polyA-signals, origin of replication, origin of transcription, tRNAs rRNAs, scRNAs snRNAs, snoRNAs, as well as regions with similarity to known proteins, and so on. For the sequences deposited in the EMBL database, a complete list of annotation features is available at http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html. Genome sequences and annotations are easily retrieved via web interfaces from different genome projects, such as Ensembl (<http://www.ensembl.org>), the common project between EMBL and the Sanger Institute, which provides a genome browser for sequenced and automatically annotated genomes of selected organisms. The UCSC genome browser (<http://genome.ucsc.edu/>) compiles annotation from different sources in one single viewer. The latest UCSC human assembly versions used for display are the RefSeq sequences, which represent the NCBI assemblies of genomic sequence data and the corresponding RNA and protein sequences. The NCBI displays all assembled sequences for selected organisms in the Map Viewer (<http://www.ncbi.nlm.nih.gov/mapview/>).

2.1.2. Sequence clustering

Though traditional nucleotide sequence repositories contain one entry per uploaded sequence, one could consider how useful it would be to determine which of these sequences actually belong to the same genes, or even to the same transcript variant. The process used to establish this link is sequence clustering. Usually based first on ESTs, the goal is to group sequences belonging to a same genomic

region together by aligning them on known mRNAs or on genomic data. Once these groups have been determined, the smaller groups are put together in gene clusters, and then mapped on gene exons, meaning the coding region of the gene. Again, this is not such an easy task. The fact that ESTs are enriched in 3' ends influences the results, as it happens that some genes are represented in the EST database only by their 3'UTR (UnTranslated Region). Some genes are thus not completely covered by EST sequences. Even worse, some weakly expressed genes are not covered at all by EST sequences. This particular problem can sometimes be solved by generating normalized libraries, where the sequence pool is enriched in weakly expressed genes, usually by using self-hybridization mechanism. The alternative splicing mechanism, which generates different transcripts represented by different ESTs, can also lead to fake transcripts builds, in particular in cases where sequences have been obtained from partially spliced RNAs. In the existing databases of clusters, these problems are treated in different ways. In Unigene (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>) [13] for example, no attempt at generating consensus sequences is made, and all the sequences from different splice variants are put together. Unigene is nevertheless tightly associated to RefSeq (<http://www.ncbi.nlm.nih.gov/RefSeq/>) [14], a reference sequence database which contains representative mRNAs and splice variants. The DoTS (<http://www.allgenes.org/>), Database of Transcribed Sequences, tries to first build different transcripts from the available sequences, and clusters them after to form a putative gene to which an identifier is given. The SIB transcriptome project, trome [15], is an attempt to map transcribed RNA from different sources to the current genome assemblies, and especially to RefSeq (NCBI genome assembly) sequences. The Unigene and RefSeq databases will be described in detail later, in the “source databases” chapter. For trome, the mapping of the transcribed RNA sources to the genome is a three-step process. One first builds pairwise alignments using megablast between all transcripts and the genomic data. Then local alignments are generated for each pair of matching RNA with sim4. Finally, alignments with too low e percentage of identity are removed. Trome also gives access to graphs representing all the putative splice variants for each gene.

These gene clusters are actually of great use for explaining and analyzing gene expression data, as it allows researchers to establish a correspondence between the measurement which is done on their support and the gene which has been transcribed. The last step needed to make this link is to find a unique name for each of the annotated genes. This step is taken care of by different groups, each

dedicated to a specific organism. For human genes, for example, the official group which deals with gene nomenclature is called HGNC (HUGO Gene Nomenclature Committee at <http://www.gene.ucl.ac.uk/nomenclature/aboutHGNC.html>). The mouse gene nomenclature is taken care by the MGI (Mouse Genome Informatics at <http://www.informatics.jax.org/>). These groups maintain a list of official symbols to be used for each known gene. Of course, these gene names catalogs are constantly evolving, even for fully sequenced genomes. First, from the nomenclature point of view, it happens that two different genes have been named with the same symbol, or that a gene name does not correspond anymore to its newly discovered function, and thus the nomenclature committee takes the decision to change the gene symbol, to maintain name uniqueness and coherence in the gene list. Then, even if the gene symbol lists were static, as fully sequenced genomes are at the present far from being fully annotated, and as the clustering and gene prediction procedures do not give definitive answers, the link between gene symbols and clusters evolve with the progress achieved in the genomes annotation process. The more precise the annotation, the more stable this link. In the meantime, all databases which try to make a link between gene symbols, gene clusters and sequences have to maintain a highly dynamic procedure with frequent updates to keep an accurate annotation.

2.2. Gene expression data generation

Gene expression for specific transcripts was traditionally analyzed using the Northern blot method. In this technique, sample RNA is separated by denaturing agarose gel electrophoresis, transferred to a solid support and immobilized. A radiolabeled RNA or DNA probe is then used to detect the molecule of interest. This is quite a straightforward procedure, but when one wants to study more than one molecule at a time, the process becomes time consuming and problematic. Since 1999, different new techniques have been setup which enable the simultaneous quantitative analysis of all the transcribed sequences in one sample. Some of them make use of the same principle as the Northern blot, meaning the ability of nucleotide sequences to hybridize with their complementary strand. Usually, the known complementary molecules are attached on a solid support, like glass-slides for microarrays and oligo-arrays, or nylon membranes. Some other new techniques are based on the direct sequencing of mRNA tags from the sample, like SAGE, MPSS, or even EST sequencing. Though very different in their conceptual aspects, all these methods have the same goal : determining at once the expression level of

as many genes as possible for one sample in one certain tissue, under precise biological conditions. By giving access to a global gene-centered information retrieval mechanism for all these different results and being able to compare them, a database will thus allow comparisons between these states regardless of the technique used for the data generation, and will thus avoid that researchers duplicate experiments which have already been done just because they are using a different technique. To generate such a database, one has to know how the different data are generated and what are the output formats which are accessible for analysis and integration. A short explanation for the major high-throughput expression techniques and their output formats is given below.

2.2.1. Microarrays

The first glass slide arrays were produced in Pat Brown's laboratory at Stanford [3]. There are three fundamental types of operations required in a cDNA microarray experiment. The first operation consists of printing the cDNA microarray itself. For Stanford-like microarrays, PCR products are purified and spotted onto poly-L-lysine coated microscope slides. The spotted sequence length can vary between 500 to 5000 nucleotides. For oligo-arrays, oligonucleotides can either be synthesized and spotted on the slides as it is done for cDNAs, or can be synthesized in situ directly on the glass surface. The expression level measurement with microarrays always takes place in the form of a comparison between two samples (see Figure 1). The messenger RNA from the two samples (the reference sample, usually the same one is used for all the experiments, and the sample to analyze), is extracted and reverse transcribed using two distinguishable fluorescent dyes to label the nucleotides. The two samples are then mixed and hybridized on a single chip. The chip is read with a scanner which measures fluorescence at two wavelengths, one for each sample. After image analysis, background subtraction and normalization, levels of transcripts in the two samples can be compared. The resulting transcript measurements for the reference and experiment sample are often called, channel one (or green channel) and channel two (or red channel) expression levels respectively.

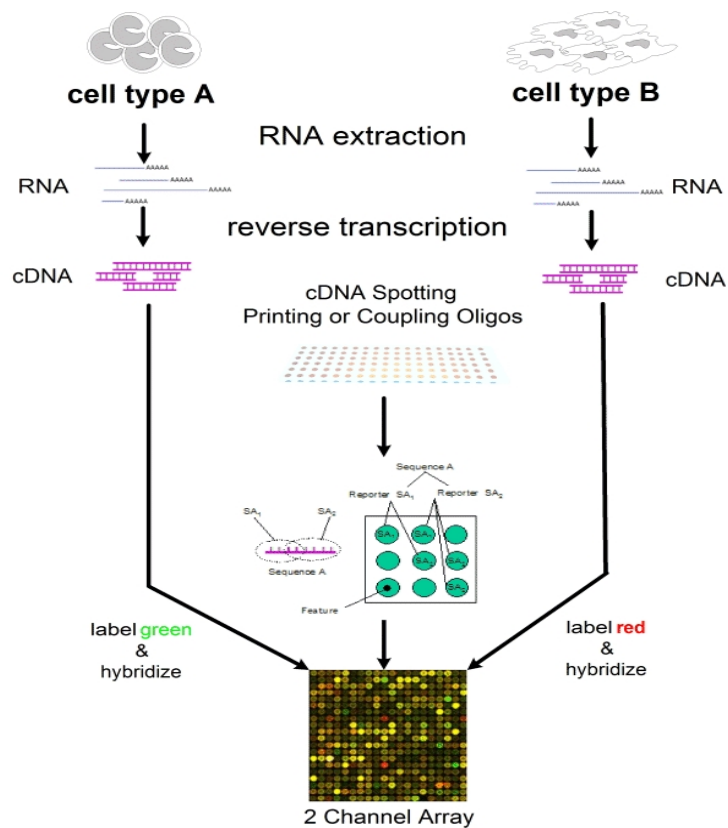


Figure 1 : Dual-channel experiment (from <http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Expression/>)

Microarray datasets are found in different flavours from different web sites. For example, data from Stanford can usually be obtained either via SMD, the microarrays specialized database from Stanford [16], or through specific project-based web pages.

In both cases, each experiment/chip of the dataset represents one file, most of the time Excel files, or tab-delimited text files. This file begins with a short description of the experiment and the expression data then follows. The first line of the data, called the header line, is the description of the output from the software used to read the chip (in most cases, Scanalyze software). It lists the different fields of the next lines. Common fields in all Stanford datasets are: Spot identifier, Sequence spotted (image clone or RNA from EMBL), description of the target gene, both channels intensities, backgrounds, background-subtracted intensities, ratio and $\log_2(\text{ratio})$ of the two channels, and a flag (0/1) for spots considered as good/bad. The numerical data which is used for further analysis is the $\log_2(\text{ratio})$ between the two channels.

After the header line, each following line contains all the described information for one spot on the chip. So the number of lines in the file, discarding the description lines, equals the number of spots on the chip.

Additional experiment description, clinical information, sample or biopsies names, treatments given, are provided in another file, usually a text file, or html page.

2.2.2. *Affymetrix*

Affymetrix [7] technology is a combination of photolithographic technology adapted from microchip production and of a chemical “protection-deprotection” nucleotide synthesis method that is somewhat analogous to the Merrifield solid-state peptide synthesis method. The process begins with a quartz wafer coated with linker molecules. The linker molecules are protected by a chemical group that can be removed using UV light. Masks such as those used in microchip synthesis allow spatial selection of the regions where the linker is to be illuminated with UV light and hence deprotected. Next, a species of nucleotide (A, C, G or T) that is itself protected with the same group is chemically bound to the unprotected linkers. The illumination and binding steps are repeated for each of the remaining nucleotides. The deprotection and binding cycles can now be continued so as to extend the single nucleotides into oligonucleotide chains of up to length 25. The resulted probe array consists of a number of cells, each containing many copies of a unique probe. Probes are tiled in probe pairs consisting of a perfect match (PM) and a mismatch (MM). The sequence of PM and MM are the same, except for a base substitution in the middle of the MM probe sequence. A probe set includes a series of probe pairs and represents an expressed transcript (see Figure 2).

An mRNA sample is then reverse-transcribed and labeled as for Stanford microarrays. Only one sample is hybridized per chip. The scanner software returns values for all the individual probes of the sample. To measure the relative transcript concentrations, one takes into account the PM/MM intensity discrepancy. One would expect a given transcript to bind to its matching probes and no signal at all for the mismatch probes. In reality, the mismatch probes give an indication of the level of unspecific binding that takes place. The mismatch signal is subtracted from the match signal and an average is then taken over the set of probes in a given block. This “average-difference”, called “Signal”, is generally used as the indicator of transcript concentration. Other calculation methods have been proposed but are

not widespread, probably largely due to the fact that the software provided by Affymetrix has not implemented them. Affymetrix datasets give relative concentrations of transcripts in a single mRNA sample, and not a comparison between two samples (Figure 3).

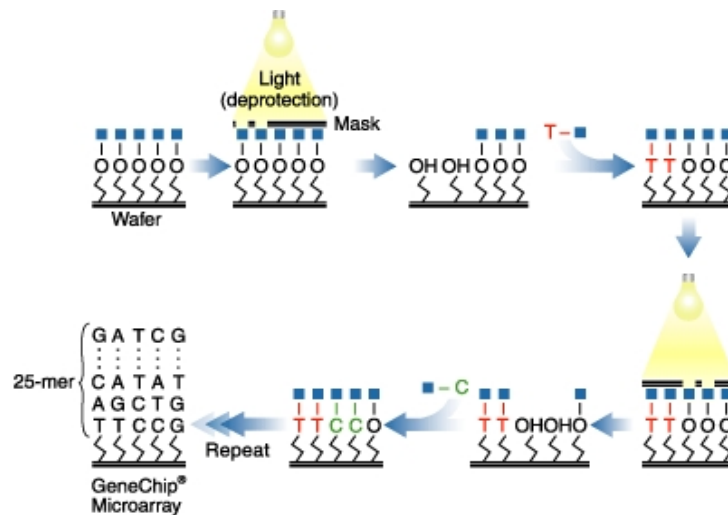


Figure 2 : GeneChip technology (from : <http://www.affymetrix.com>)

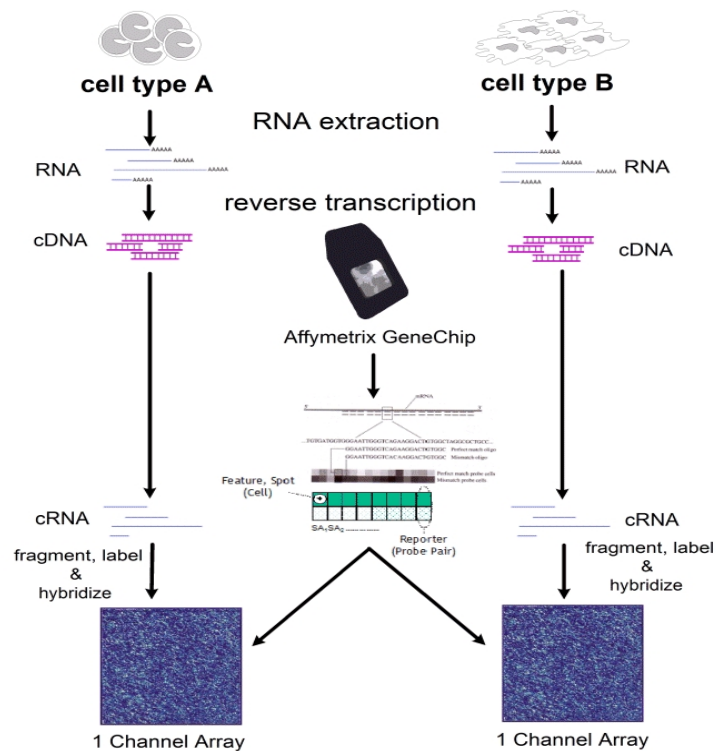


Figure 3 : Single channel experiment

Two very different output formats are available for Affymetrix-based experiments. The first one gives access to the direct output from the scanner. The resulting files are called CEL files. They contain expression values for each spot, meaning each separated tag of all probe sets. They need a quite extensive reformatting procedure. Though any Affymetrix-based experiment has to go through this step, the Affymetrix CEL files are not the most common format for publicly available datasets. Indeed, mainly due to space constraints, authors often prefer to give public access to data which have already been processed via a specific software in files which contain results per probe set, and not per tag.

This other format corresponds to the second chip analysis level. Once the chips have been read, CEL files are processed to create probe set-oriented information. Different softwares can be used to analyze CEL files. Amongst the published datasets that are in CleanEx, the most commonly used softwares are MAS4, MAS5 [17, 18], and RMA [19]. Some people also use in-house procedures. The usual output of these softwares is very basic, and consists mainly of two values :

- The “Signal”. Its measurement involves a comparison of all sequence-specific perfect match (PM) probe cells with their corresponding mismatch (MM) probe cells (see Figure 4) for each probe set using an estimate method that yields a robust weighted mean which is relatively insensitive to outliers. This Signal is calculated for each probe set and represents the relative level of expression of the transcripts. One important point to remember is that the given value in Affymetrix experiments is the intensity of one experiment, and not a ratio between a reference and the actual experiment (as in dual-channel experiments).
- The “detection call” value. The call value is a tag for assessing the reliability of the probe set' s intensity detection. To make the call, a first “discrimination value” is calculated by taking the median of $((PM-MM)*(PM+MM))$ for all the tags of one probe set, where PM and MM are respectively the “Perfect Match” tag and the “MisMatch” tag. A p-value is then calculated by applying a one-sided Wilcoxon's signed rank test to the discrimination value (note that this test is used in the MAS5 program. Other programs use a different statistical test to assess the call' s p-value). The call tag is then assigned according to two user-definable thresholds, which separate the probe sets in three

different categories : A, P or M, respectively for “Absent”, “Present” or “Marginal” call. A fourth tag (NC for “no call” is applied when all tags from one probe set are excluded from the analysis. This occurs for example when the MM tag is saturated.



Figure 4 : Affymetrix PM versus MM intensities

Depending on the software which is used by the dataset's authors, the calculation of the Signal and the call can differ. For example, the RMA software uses background subtraction instead of PM-MM discrepancy, and thus does not return the call value. To eliminate negative Signal values, the MAS5 software uses an adjusted MM value if this value is larger than the PM value.

This kind of data is usually provided as a single flat file containing all the experiments for all the probe sets on the chip. There is one line per probe set, which contains in the different columns the intensity and the call for all experiments. Metadata like experimentation protocols and descriptions are provided in a separate file, as for Stanford data.

2.2.3. SAGE

This technique was first developed by Velculescu in 1995 [4]. It's the first large-scale transcript abundance measurement method, and is based mainly on standard sequencing methods. First, mRNAs from the selected sample, or library, are extracted and reverse transcribed.

Double stranded cDNAs, bound to beads by the polyA tail, are then cut with a specific restriction enzyme called “anchoring enzyme”. Commonly used enzymes for SAGE are NlaIII or Sau3AI. The selected enzymes have a restriction site which leaves a “sticky end” on the cDNA. After the enzymatic step, each bead is left with one small specific fragment for each cDNA type. These fragments are then separated in two pools A and B and sequences from the two pools each receive one special linker (A or B). These linkers contain one sticky end which adapts to the anchoring enzyme site, as well as a

recognition site for a “tagging enzyme”, usually BsmF1. The particularity of the tagging enzyme is that it cuts a few bases in the 3' direction from its recognition site (10 to 14 for SAGE, and 25 for Long-SAGE), thus adding the "Tag" sequence to the linkers. This tag can be considered as a specific signature for one transcript.

After PCR, the linkers are removed to create “ditags” which are then concatenated in longer strands, cloned and sequenced. After the sequencing step, one is left with a total tag count per sample, as well as an individual tag count, allowing for the measure of the individual transcript concentration in the original sample (Figure 5). The next step is to attribute to each tag its specific gene. In CleanEx, this is done in the “target step”, explained later on.

The format resulting from this experiment type consists of a list of tags associated with their corresponding tag count. Often, added to the usual tag and tag-count information, the tag-to-gene correspondence that the authors used for data analysis is provided in each file.

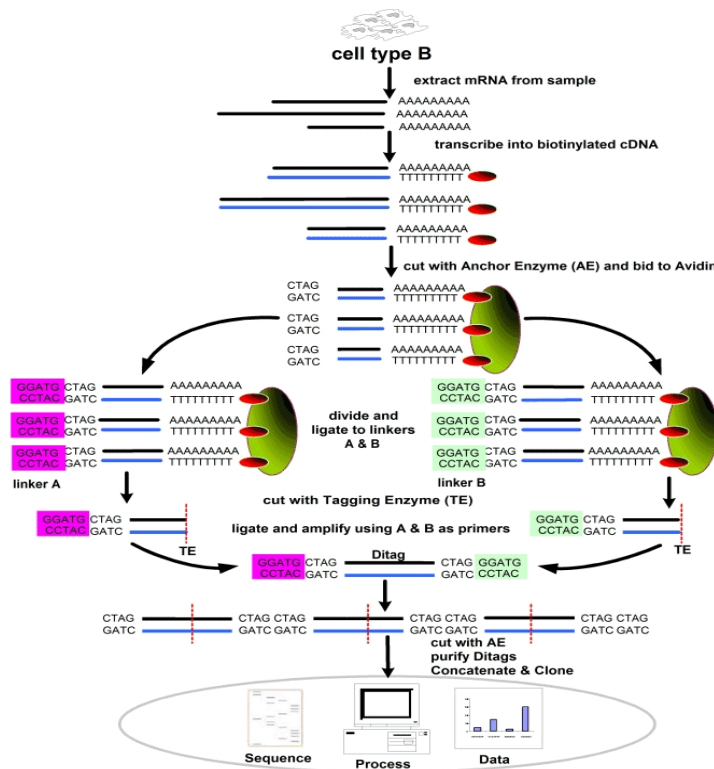


Figure 5 : SAGE technique (from <http://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Expression>)

2.2.4. MPSS

MPSS [5] is a relatively new technique developed at the Lynx Therapeutics Inc. One of its major technical advantages is that it eliminates the need for individual sequencing reactions and the physical separation of DNA fragments required by conventional sequencing methods. The procedure applied is as follows. From the sample, mRNAs are extracted and reverse transcribed in cDNAs which are then attached to unique 32-nucleotide long tags. The total number of possibilities with 32-nucleotide long tags means that each different cDNA receives one specific tag. Simultaneously, this combination tag-cDNA is amplified by PCR.

The tag-cDNA molecule population is then mixed with 5 micron micro-beads which are all coated with many copies of one unique anti-tag sequence, complementary to one unique tag of the tag-cDNA molecule. After hybridization, each micro-bead carries one type of cDNA found in the original sample.

The attached molecules are then cut with a restriction enzyme to create a 17- to 20-nucleotides long signature sequence for each cDNA (or micro-bead).

The micro-beads are fixed as a single layer array in a flow cell, solvents and reagents can be washed over the micro-beads in each cycle of the process. The protocol elicits sequence-dependent fluorescent responses from the micro-beads, which are recorded by a camera after each cycle. The 17- to 20-base-pair signature sequences are constructed through this process without requiring any separate sequencing reactions. A software is used to automate the delivery of reagents and solutions used in this sequencing process and to compile, from the images obtained at each cycle, the signature sequences that result from each experiment (Figure 6).

Once the process is done, the information obtained is the same as that obtained with the SAGE technique : individual signature count, and total sequence count, which will give the relative transcript abundance. The next step is to make a correspondence between the short signatures and the transcripts. The output format resembles that of the SAGE experiment.

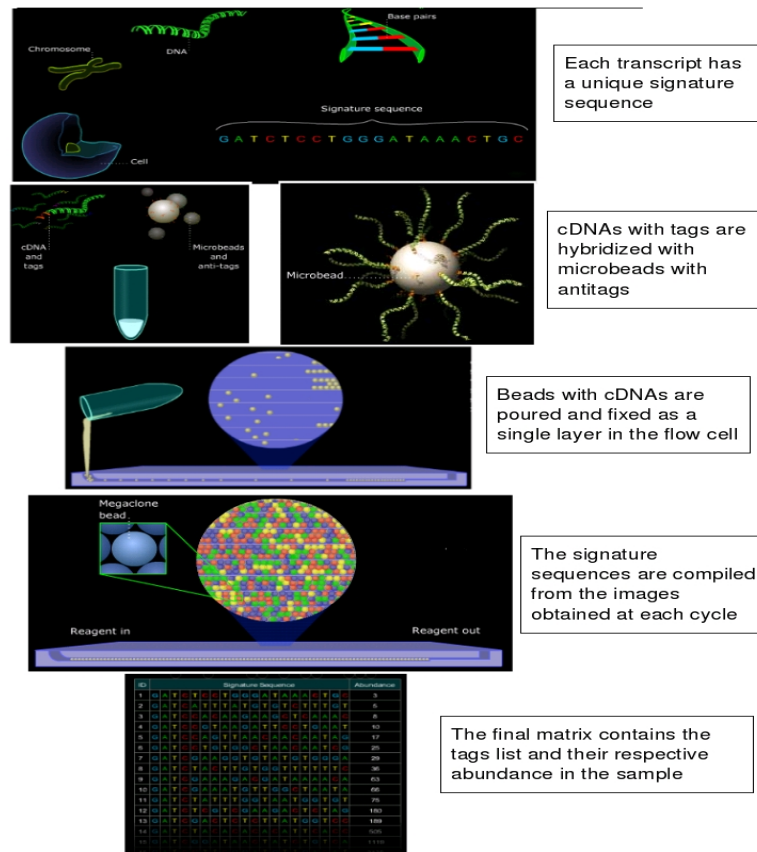


Figure 6 : MPSS technique (from <http://www.lynxgen.com>)

2.2.5. ESTs

Large-scale single-pass sequencing of cDNAs has been performed for approximately a decade in the form of expressed sequence tag (EST) projects. Though the original goals were to sample the transcriptome (transcribed portion of the genome) in order to discover new genes and to study exon-intron structures and to group together those ESTs that come from the same gene, today ESTs are used for a variety of other purposes. The actual clones can be spotted onto microarrays, and the sequences can be used to identify SAGE and MPSS tags. For the case when the generated cDNA libraries used are native (meaning neither normalized nor subtracted) and the actual sequences that are read can be considered a random sample from an mRNA population, the size of the clusters gives an indication of the transcript concentration in the respective sample [6]. Since libraries are usually generated from specific tissues, the data from EST frequency counts can be used to compare expression between tissues. This EST count is for example exploited with tools like DGED (Digital Gene Expression Displayer, <http://cgap.nci.nih.gov/Tissues/GXS>) or xProfiler (<http://cgap.nci.nih.gov/Tissues/xProfiler>)

by the CGAP group.

In conclusion, the huge volume of data produced by all these expression experiments has dictated the use of computerized data structures to store all of the information. Typically, authors who generate such a dataset also create an in-house very simple data management system to allow data retrieval and visualization. As high-throughput expression measurement experiments became a standard method used by an increasing number of laboratories, the need for more generalized repositories, capable of storage, retrieval comparison and analysis of heterogeneous expression data, became a priority [20]. The first problem to face when setting up such a repository, as several techniques can be used to generate this data type, is to coordinate standard and controlled procedures for the data integration, as well as the data retrieval. The definition of such procedures has been achieved by the MGED (Microarray Gene Expression Data) [21,22] (<http://www.mged.org>) Society via the acceptance of MIAME (Minimum Information About a Microarray Experiment) [23] rules by the scientific community. Such a well-defined format for expression experiments generated a burst of public data repositories [24], going from the single data-type repository to the complex automatic upload repository where authors of very different datasets could directly import their results in a standardized format. Amongst all databases created at that time, few appeared to be based on a strong enough architecture design to support the massive amount of data generated all over the world and to become a universal gene expression data repository and retrieval. Three of them, which we will describe in the next chapter, have been selected as official expression data repositories by the members of the MGED committee.

Expression data storage is one thing, but it is useless if there is no way to interpret the data. The meaning of the expression data comes from the link between the numerical results of the expression measurement and the biological data (namely the transcript which will be translated into a protein and the protein which will have an influence on the organism's behavior), corresponding to this measurement. This part of the analysis is of course strongly influenced by the evolution of the corresponding organism-specific genomic data annotation and gene discovery procedures. Though expression experiments are done once, the genome annotation evolves, and old data would need to be refreshed to keep some usefulness. For example, sequences which were undefined at the time when the expression measurements have been realized can be later classified in a gene cluster and thus become an

important piece of information. In other cases, some sequences were wrongly attributed to a known cluster, but then appear to be part of another newly discovered gene. The lack of a mechanism which will allow frequent re-annotation of expression data with up-to-date gene information has been one of the major reasons which prompted us to develop the CleanEx database.

3. EXPRESSION DATABASES : HISTORY AND EVOLUTION

3.1. Historical context : setup of MIAME standards

Already in November 1999, many of the major microarray users and developers, including Affymetrix (<http://www.affymetrix.com>), Stanford University (<http://genome-www5.stanford.edu/>) and the European Bioinformatics Institute (<http://www.ebi.ac.uk/>), founded the MGED Society, as a way to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. The two major outcomes which are now widely accepted as microarray database standards are MIAME and MAGE, which could be respectively considered as the data repository standard, and the data exchange standard.

MIAME stands for Minimum Information About a Microarray experiment. Its purpose is to make sure that the data published or submitted to a database are made publicly available in a format that enables unambiguous interpretation of the data and potential verification of the conclusions. The MIAME standards have now been accepted and followed by the three main expression data repositories. Since 2002, the major scientific journals require that data should be MIAME compliant to get published.

The MAGE (<http://www.mged.org/Workgroups/MAGE/mage.html>) group of MGED has now released a universally accepted language for expression data exchange. This language, called MAGE-ML, is based on XML (eXtensible Markup Language) and can describe microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results. MAGE-ML is derived from MAGE-OM (Microarray Gene Expression Object Model). MAGE-OM describes the structure and the links of all the entities which are to be stored in a database via MAGE-ML.

The use of the MGED standards thus allows easy and comprehensive data retrieval from any publicly accessible database.

3.2. Emergence of expression databases

Even before the MGED standards were set up, many solutions attempted to concatenate different datasets in one single database. These trials were sometimes directly linked to the efforts of one single laboratory which was producing a certain amount of expression data, like the Stanford Microarray Database [16] or ExpressDB [25]. Others were more related to the database management scientific community, like ArrayExpress [26], GEO [27, 28] or GeneX (<http://sourceforge.net/projects/genex/>). Some other repositories emerged from the needs of a specialized scientific community. These can be qualified as organism specific, tissue specific, disease specific, or even treatment specific databases.

Amongst all these databases, some have now been generally accepted as official data repositories [29]. New databases appear rapidly, and each has its own specifications. The databases can either be a single repository, or they can be linked to a plethora of analysis or format processing tools. An important characteristic which is not shared by all these databases is the acceptance of automatic upload procedures. Table 1 provides general information and the URL for some of the most used expression databases [1, 30]. It is not intended to give an exhaustive list of all expression databases, but points out the diversity existing amongst them, and the problem complexity when dealing with data coming from different sources. A very good description and comparison of gene expression databases can be found at : <http://ihome.cuhk.edu.hk/~b400559/array.html>.

Human databases		
GeneNote	Human genes expression profiles in healthy tissues	http://genecards.weizmann.ac.il/genenote
HugeIndex	Expression levels of human genes in normal tissues	http://hugeindex.org/
RefExA	Reference database for human gene expression analysis	http://www.lsbm.org/db/index_e.html
H-ANGEL	Human anatomic gene expression library	http://www.jbirc.aist.go.jp/hinv/index.jsp
BGED	Brain gene expression database	http://love2.aist-nara.ac.jp/BGED

emap Atlas	Edinburgh mouse atlas: an atlas of mouse embryo development and spatially mapped gene expression	http://genex.hgu.mrc.ac.uk/
EPConDB	Endocrine pancreas consortium database	http://www.cbil.upenn.edu/EPConDB
HemBase	Genes expressed in differentiating human erythroid cells	http://hembase.niddk.nih.gov/
PEDB	Prostate expression database: ESTs from prostate tissue and cell type-specific cDNA libraries	http://www.pedb.org/
Kidney DB	Kidney development and gene expression	http://golgi.ana.ed.ac.uk/kidhome.html
EpoDB	Genes expressed during human erythropoiesis http://www.cbil.upenn.edu/EpoDB/	http://www.genome.ad.jp/magest
Osteo-Promoter DB	Genes in osteogenic proliferation and differentiation	http://www.opd.tau.ac.il
Tooth Development	Gene expression in dental tissue	http://bite-it.helsinki.fi/

Mouse databases

GXD	Mouse gene expression database	http://www.informatics.jax.org/menus/expression_menu.shtml
GenePaint	Gene expression patterns in the mouse	http://www.genepaint.org/Frameset.html
Mouse SAGE	SAGE libraries from various mouse tissues and cell lines	http://mouse.biomed.cas.cz/sage
MAMEP	Gene expression data on mouse embryos	http://mamep.molgen.mpg.de/

Other organisms

rOGED	Rat ovarian gene expression database	http://web5.mccs.uky.edu/kolab/rogedendo.aspx
Axeldb	Gene expression in <i>Xenopus laevis</i>	http://www.dkfz-heidelberg.de/abt0135/axeldb.htm
FlyView	<i>Drosophila</i> development and genetics	http://pbio07.uni-muenster.de/

MAGEST	Ascidian (<i>Halocynthia roretzi</i>) gene expression patterns	http://www.genome.ad.jp/magest
MEPD	Medaka (freshwater fish <i>Oryzias latipes</i>) gene expression pattern database	http://www.embl.de/mepd/
GermOnline	Gene expression in mitotic and meiotic cell cycle	http://www.germonline.org/
NASCarrays	Nottingham Arabidopsis Stock Centre microarray database	http://affymetrix.arabidopsis.info
ExpressDB	Yeast and <i>E. coli</i> expression database from SAGE, microarrays and Affymetrix chips	http://salt2.med.harvard.edu/ExpressDB/

Data type specific databases

5'SAGE	5'-end serial analysis of gene expression	http://5sage.gi.k.u-tokyo.ac.jp/
SAGEmap	NCBI's resource for SAGE data from various organisms	http://www.ncbi.nlm.nih.gov/SAGE
SMD	Raw and normalized data from microarray experiments	http://genome-www.stanford.edu/microarray
GeneTrap	Expression patterns in an embryonic stem library of gene trap insertions	http://www.cmhd.ca/sub/genetrap.asp
TissueInfo	EST-based tissue expression profiles mapped on Ensembl transcripts	http://icb.med.cornell.edu/crt/tissueinfo/website.xml
CGAP	EST and SAGE-based expression profiling of normal, pre-cancer and cancer human or mouse tissues	http://cgap.nci.nih.gov/
BodyMap	Human and mouse EST based gene expression data	http://bodymap.ims.u-tokyo.ac.jp/

All-purpose expression databases

ArrayExpress	Public collection of microarray gene expression data	http://www.ebi.ac.uk/arrayexpress
CIBEX	Center for Information Biology gene EXpression database	http://cibex.nig.ac.jp/index.jsp
GEO	Gene expression omnibus: gene expression profiles	http://www.ncbi.nlm.nih.gov/geo/

RAD	RNA Abundance Database : will allow cross-data comparison, spot-to gene mapping is done via DoTS	http://www.cbil.upenn.edu/RAD/php/index.php
-----	--	---

Annotation and gene oriented databases		
NetAffx	Public Affymetrix probesets and annotations	http://www.affymetrix.com/
GeneAnnot	Revised annotation of Affymetrix human gene probe sets	http://genecards.weizmann.ac.il/geneannot/
GeneTide	A transcriptome-focused member of the GeneCards suite	http://genecards.weizmann.ac.il/genetide/
CleanEx	Expression reference database, linking heterogeneous expression data for cross-dataset comparisons	http://www.cleanex.isb-sib.ch/
LOLA	List of lists annotated: a comparison of gene sets identified in different microarray experiments	http://www.lola.gwu.edu/

Table 1 : List of existing expression databases

We will now describe in detail the emergence, development and major specificities of some of the most important gene expression databases. We will focus on historically important databases and on general purpose ones with heterogeneous data, as well as on MGED approved databases. We will then spend more time on describing the expression databases linked to gene annotation and cross-dataset analysis tools.

3.3. Main expression data repositories

3.3.1. SMD : the Stanford Microarray Database

Initiated in 1999, the Stanford MicroArray Database [16] is one of the first academic databases to be used on an institutional scale. Formerly developed as a research tool for Stanford scientists and their collaborators, it was restricted to dual-channel microarray data obtained via GenePix (http://www.axon.com/GN_GenePixSoftware.html) or Scanalyze (<http://graphics.stanford.edu/software/scanalyze/>) image analysis softwares until 2003 . The SMD now

also supports data generated with Custom arrays and Affymetrix chips [31]. The SMD, with the help of the MAGE-stk (MAGE software toolkit) has implemented a data translator which generates MAGE-ML expression files from the SMD format. These data can thus be directly uploaded in MIAME compliant data repositories. The SMD software can be downloaded and installed locally, but is only compatible with an ORACLE relational database on Solaris machines. To provide a database structure based on a fully open source system, a new version compatible with Linux and PostgreSQL, called the Longhorn Array Database [32], has been developed.

Though the majority of the analysis tools and the upload system are restricted to registered users (as for example upload and analysis of external data), the SMD also provides an impressive number of public search and analysis interfaces, as well as the possibility to generate on line plots with the selected public data. One can also select spots to be used for further analysis via filters as diverse as gene symbol or intensity of the spot. The possible analysis methods on these filtered data are SOM (Self-Organizing Maps), or hierarchical clustering. Another very interesting tool implemented in SMD is the spot history, which stores expression data for all the spots corresponding to the same feature and displays it in a histogram. This gives a nice general view of the spotted clone's behavior.

On May 2005, with a total of 54618 experiments, the SMD represents one of the largest collection of expression data. Amongst these experiments, 8979 are publicly accessible, the others come either from private datasets or from data which has not yet been published.

3.3.2. CGAP and SAGEmap

The Cancer Genome Anatomy Project (CGAP, <http://cgap.nci.nih.gov/>) [33] began in 1996. This program of the National Cancer Institute (NCI) is studying the molecular changes that occur when a normal cell is transformed into a cancer cell. It provides an impressive number of tools, from clone annotation or SNPs discovery, to library selection and annotation. The CGAP also consists of a huge collection of human or mouse ESTs classified according to their tissue origin and their disease state, either normal, pre-cancer, or cancer. From these different pools, CGAP provides digital expression analysis between tissues and disease states. After the achievement of the CGAP SAGE project, which allowed the assembly of over 5 million transcript tags from more than 100 human cell types, the ESTs digital analysis tools have been adapted to also accept SAGE data. These two data sources, namely the

ESTs and the SAGE tags, can not be used together; however, the digital display gives a fairly precise idea of the expression level in different tissues.

SAGEmap (<http://www.ncbi.nlm.nih.gov/projects/SAGE/>) [34] is a SAGE dedicated public expression repository. Initially designed to archive SAGE data from the CGAP, it now accepts SAGE type expression data from any source, via a tool called SAGEmap Submission Tool (SST). This tool not only allows a facilitated library annotation, but it is also designed to process the primary data product of the SAGE technique, which represents the concatenated tags, in pairs (ditags), separated by four base punctuation signals (e.g., NlaIII sites). Once processed by SST, the data are represented by a list of tags with their corresponding count values, and is thus a digital representation of cellular gene expression. The SAGEmap also provides a SAGE to gene assignment tool based on the sequences available in the Unigene (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>) clusters. Moreover, each SAGE to gene assignment is associated with a so-called “class”, representing the reliability of the tag mapping. The classes depend on the sequence type which is associated to the tag. Basically, The best quality class is attributed when the mapping occurs on a well-characterized mRNA sequence. If the mapping is done on an EST, the class varies according to the presence or absence of a polyA tail or signal, as well as the 3' or 5' annotation of the EST. Combining these two criteria gives four more classes, for a total of five different classes (including the mapping on well-known mRNAs). At the end, the tags are definitely mapped to the gene which gives the highest mapping quality, and real ambiguity (like tags which really map to more than one gene) is not taken into account for further analysis. This mapping tool is available on line so that any user can annotate his own SAGE tag collection.

SAGEmap also provides the “tag-” and “gene display” tools. The first one shows in shades of grey the relative abundance of the selected tag in all the currently hosted SAGE libraries. The gene display tool shows the gene's reliable tag assignments.

SAGEmap is, apart from CleanEx, the only repository which offers a quality control for the tag to gene annotation.

3.3.3. *ExpressDB*

ExpressDB (<http://salt2.med.harvard.edu/ExpressDB/>) [25] contains published and in-house expression

studies on Yeast and *E. coli*. Created in 1999, before the emergence of expression data standards or universal data repositories, ExpressDB has immediately been loaded with data from eleven yeast studies using three different kinds of high-throughput RNA level assays, namely SAGE, DNA microarray, and Affymetrix oligonucleotide array data. It is one of the first attempts to represent and manage data not only from multiple studies but also from multiple kinds of expression data types. The whole comparison process is built on the computation of so-called ERAs (Estimated Relative Abundances). An ORF (Open Reading Frame) ERA represents the fractional abundance of the ORF's RNA with respect to the total population of ORF RNAs in cells in a particular experimental condition. The computation of these ERAs is quite straightforward for Affymetrix or SAGE data types, which measure RNA abundance in one single sample. However, computing ERAs becomes much more problematic for dual-channel experiments, which give as output a ratio between two experiments. For this kind of experiments, the decision was thus taken to use the single so-called “red” channel, background subtracted, as a basis to calculate ERAs. This method, though, did not give as good results as for the other data types. Once all ERAs are computed, common ORFs were selected. Final ERAs for each experiment were calculated for each ORF by dividing each individual ORF ERA by the total sum of ERAs for all ORFs in that experiment. The ERAs produced can be used via a query interface instead of raw expression values for cross-dataset comparisons.

3.3.4. MGED recommended expression data repositories

As for nucleotide sequences repositories, the set up of common standards has prompted the MGED Society to recommend a few databases as official expression data repositories. These three selected databases are hosted by the same organizations than the three official nucleotide sequences databases, namely the EBI for ArrayExpress [26], the NCBI (<http://www.ncbi.nlm.nih.gov/>) for GEO [27], and the DDBJ for CIBEX [35]. There is anyway a huge difference in the management of sequence databases and expression databases. Indeed, if the three versions of the nucleotide repositories (EMBL, GenBank and DDBJ) are fully synchronized and contain the same entries in a slightly different format, the three expression databases are so far completely independent. They do not host the same data, and though they all follow the MGED standards, their design and implementation differ a lot.

3.3.4.1. GEO and other data repositories

The Gene Expression Omnibus (GEO) [27] at the National Center for Biotechnology Information (NCBI) is the largest fully public repository for high-throughput molecular abundance data, as well as a curated, on line resource for gene expression data browsing, query and retrieval. GEO became operational in July 2000. It has been populated with very heterogenous microarray-based experiments, done for very different purposes, like gene expression analysis by mRNA abundance measurements, genomic DNA arrays for linkage analysis, gene copy number studies, or protein arrays to monitor expression at the protein level. GEO also stores non-array-based technologies such as serial analysis of gene expression (SAGE) and mass spectrometry proteomic technology. Data can be submitted via interactive web-based forms. Bulk submissions in GEO SOFT specific format or MAGE_ML format are also accepted. The database is, as ArrayExpress, organized on the basis of three different levels, namely Platforms, Samples, and Series.

An instance of a *platform* is, essentially, a list of probes that define what set of molecules may be detected in any experiment utilizing that platform. For example, the platform data table may contain GEO-defined columns identifying the position and corresponding feature of each probe (spot) such as a GenBank accession number, open reading frame (ORF) name and clone identifier, as well as submitter-defined columns. It corresponds to the ArrayExpress “array” organization level. Platform accession numbers have a ‘GPL’ prefix.

An instance of a *sample* describes the gene expression level determined for a biological sample under one condition. It corresponds to the experiment level in ArrayExpress, or for example to the numerical output of one chip. A sample utilizes a specific platform to generate molecular abundance data. Each sample has only one parent platform which must be previously defined. For example, a sample data table could contain the output of Scanalyze realized with a specific dual-channel chip, as well as measurements for one experiment based on an Affymetrix chip (like absent/present call and intensity), or SAGE tags count for one specific sample. Each line of the table corresponds to the measured values for one spot or tag. Sample accession numbers have a ‘GSM’ prefix.

An instance of a *series* organizes samples into meaningful data sets which make up an experiment, and are bound together by a common attribute. Each series usually corresponds to one publication. Series accession numbers have a ‘GSE’ prefix, and could be compared to the “protocol” level in ArrayExpress

and to a dataset in CleanEx.

These three levels are accessible through web query interfaces.

A recently setup query system, linked to the NCBI Entrez database system (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=geo>), allows either so-called datasets or profiles retrieval. Interestingly, the profiles retrieval system provides a ‘gene-centric’ view of GEO data. The profiles output represents a histogram of expression measurements for one gene across each sample in a single GEO dataset. Other newly implemented features include the possibility of calculating an average rank or value differences between experimental subsets within a single dataset.

The ArrayExpress [26] public repository, hosted at the European Bioinformatics Institute, accepts any expression array data type, including Affymetrix GeneChips, but no SAGE data. In a way to facilitate the authors' submission procedure, data can be mainly submitted via the MIAMExpress (<http://www.ebi.ac.uk/miamexpress/>) on line data submission tool, which consists of a series of simple web forms to describe their experiment and upload the data files. Furthermore, each submitted dataset is then manually curated to ensure that data are MIAME compliant and well formatted. ArrayExpress also provides dedicated pipelines for specific users, like for example the SMD from Stanford. The query retrieval system gives access to three organization levels, the Array, *Experiment*, and Protocol, which correspond to the GEO *Platform*, *Sample* and *Series* respectively. ArrayExpress is linked to an on line data retrieval system as well as to an integrated on line visualization and analysis tool called Expression Profiler.

CIBEX [35] is a very new expression data repository which so far contains very few data, but as it has been recommended by the MGED Society as an official expression database, it will probably grow quite fast. For now, the CIBEX database allows only raw data retrieval and does not provide any analysis or visualizer tools. An on line submission system is now under development.

3.4. Genes oriented databases

The databases that are here called “gene oriented expression databases” have a very different objective compared to expression data repositories. They in fact aim at giving access to any available expression

measurement corresponding to one gene under one single identifier. They are usually not made for raw expression data bulk retrieval, and for that reason they don't need to be MIAME compliant. The emergence of standards for expression data publication has anyway been of great use for these databases as it greatly facilitates the expression data integration in their system. The most well-known databases of this type are GeneCards [36], at <http://bioinfo1.weizmann.ac.il/genecards/index.shtml>, and SOURCE [37] , at <http://source.stanford.edu/cgi-bin/source/sourceSearch>. To a certain extent, with the new gene search that has recently been implemented, GEO can also be considered as a gene oriented database.

3.4.1. GeneCards

The GeneCards project [36], from the Weizmann Institute, began in 1997 and was first designed to integrate information about genes, proteins and diseases extracted from heterogeneous public databases. Over the years, it has evolved into a multi-purpose human-centered database, separated in different parts, according to their center of interests.

GeneLoc and GeneTide are transcript annotation databases. GeneLoc compares the genomic locations of different genome annotation sets to generate a unified location for each gene, while GeneTide integrates data from other resources, like Unigene [13], DoTS (<http://allgenes.org/>), AceView (<http://www.ncbi.nih.gov/IEB/Research/Acembly/>), and GeneAnnot.

GeneAnnot is a new revised annotation of three human Affymetrix chips, namely the HG-U95, HG-U133 and HG-U133 Plus2.0 chips. The mapping is done on full-length transcripts as well as on ESTs (Expressed Sequence Tags), via the BLAT [38] program. Transcripts to gene mapping was done directly to GeneCards entities whenever possible, and to Unigene as a second instance. To evaluate each probe set, two quality scores are provided, the sensitivity score which corresponds to the number of matching probes in the given probe set to a certain gene, and the specificity score, which lowers if some probes of one probe set match additional genes. The final display shows the two scores as well as the number of genes which had a hit for the individual probes. Access to the individual positions of each probe on the mapping sequences is not provided.

GeneNote compiles expression experiments of human healthy tissues performed on the Affymetrix HG-

U95 chips set from A to E at the Weizmann Institute.

The *GeneCards* part of the database system integrates all the data generated by the other members of the GeneCards Suite. It is also cross-referenced to a great number of external databases such as Unigene [13], Genew [39], Swissprot [40], OMIM [41], Ensembl and others. For our purpose, the most interesting viewers offered in one GeneCards gene entry are :

- A direct view and link to GeneAnnot and the mapping result for human chips
- An expression viewer, including the GeneNote healthy tissues expression profiles for each gene.
- A digital northern viewer generated from the Unigene ESTs, as well as another one created from the CGAP SAGE tags for the studied gene. The two additional viewers use the same tissue classification than the samples analyzed in GeneNote.

Access to other expression data is not provided yet.

3.4.2. SOURCE

SOURCE [37] is actually the database which resembles the most the CleanEx database ones. This gene oriented database, hosted by the Stanford University, links external resources like Unigene, chromosome location, Gene Ontology, Swissprot and many other universal databases to expression data from different datasets. It is structured on a backbone including two files types, the GeneReports and the CloneReports.

The GeneReports contain cross-links to features attributable to a single gene. The entry name is, whenever possible, built from the HUGO [42] official gene symbol, otherwise the associated Unigene cluster number is used.

The CloneReports store all annotation corresponding to the ESTs found in dbEST [43]. The whole structure of SOURCE consists of a series of links between these two files.

SOURCE stores and displays only expression data produced via cDNA or Affymetrix microarrays, but no SAGE data. In addition to these, it also provides, as CleanEx, the relative expression level of a gene

in different tissues based on the ESTs integrated in Unigene (the CleanEx method will be described later).

Both GeneReport or CloneReport can be accessed via common identifiers, like Unigene ID, clone ID, gene symbol or GenBank/EMBL accession numbers. A batch search allows retrieval of data for a whole gene list. This batch mode exists for both Genes and Clones. The single GeneReport also provides a link to all the clones associated to this entry, but not to SAGE tags or Affymetrix probe sets which also correspond to this gene symbol. SOURCE, as CleanEx, is updated on a very regular basis to ensure access to the most up-to-date information.

3.4.3. CleanEx

When the CleanEx [8] project began, none of the previously cited expression databases in the gene oriented category existed, and this appeared as an important problem to study for the general use and comprehension of expression data. Now that more databases have also chosen this way, CleanEx nevertheless still shows some unique and very useful features.

First of all, amongst all these databases, CleanEx is the only one which provides individual mapping and position of Affymetrix probes, and not probesets. Second, this feature has been extended to give SAGE tags positions on sequences. Moreover, the CleanEx system is able to retrieve not only all the clones common to one gene, but can show, for the selected gene, clones, potential SAGE tags, and Affymetrix probe sets altogether. Another feature which is very important in CleanEx is the cross-dataset comparison system which can deal with any data type. Lastly, though SOURCE also provides a direct link to a promoter sequence download system, in CleanEx, whenever possible, this link is done via the transcription start site position given in EPD, which gives much more reliability to the promoter elements' position.

The CleanEx expression database will now be described in detail. The first part will give an overview of the data organization in the database. The building processes will then be explained for each data type. The results part will explain the database content and the different possibilities of using CleanEx in a way to make heterogeneous gene expression results comprehensible in a gene-oriented way.

4. THE CleanEx DATABASE : CONCEPT AND DATA ORGANIZATION

The CleanEx (formerly called EPDEX) project began in 2001 as a companion database for EPD, the Eukaryotic Promoter Database [44, 45]. Its first aim was to map EPD promoters and Swissprot entries via genes symbols to expression profiles. As the companion database has grown and earned its independence, its main goal has also evolved and is now 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 catalogs of human genes and genes from model organisms. The completely automatic mapping procedure relies partly on external genome information resources such as UniGene [13] and RefSeq [14]. The central part of CleanEx is a gene index containing cross-references to all public expression data already incorporated into the system which is built on a weekly basis. In addition, the expression target database of CleanEx provides gene mapping and quality control information for various types of experimental resources, such as cDNA clones, Affymetrix probe sets and SAGE tags. The web-based query interfaces offer access to individual entries via text string searches or quantitative expression criteria.

So far, CleanEx contains human and mouse genes for which the symbol is approved by the representative organism nomenclature committee. For human genes, we use the approved Genew [39] gene symbols. The mouse gene index is based on the MGD (Mouse Genome Database) [46] nomenclature. There is one entry per gene name for each organism.

CleanEx is a flat file formatted database consisting of three different file types. Each of these files is

linked to the other through a defined accession number. The three file types are :

- CleanEx_exp
- CleanEx_trg
- CleanEx

4.1. CleanEx_exp

CleanEx_exp files contain public gene expression data in a slightly reorganized text file format and, if possible, equivalent to the original sources in terms of the information content. They are 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 measured expression levels for a set of target sequences and conditions, which is typically published and analyzed 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, which could be compared to the GEO *series* instance, provides general information about the data set including the number of spotted features, the number and the list of tissues or experiments for which expression values are provided, the organism, the associated published paper, and the type of associated reference sequences. A data entry contains expression values for a particular feature over all conditions. By feature we mean any molecule that is used to retrieve a certain transcript's abundance in an experiment, such as a clone or oligonucleotide spotted on a certain position of a dual-channel chip, an Affymetrix probe set, or a SAGE or MPSS tag. Note that this one feature/all experiments concept is very different from the one chosen in GEO, where each *sample* corresponds to the measurements of all features in one experiment. Each CleanEx_exp data entry's header line contains the CleanEx_target identifier linking the analyzed sequence to its target expressed sequence (and usually to the associated gene name). The word "target" stands for the transcript which is "targetted" by the so-called feature.

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 analyzed 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 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 (<http://rana.lbl.gov/EisenSoftware.htm>) or GenePix (<http://www.axon.com>). The new meta-entry is then usually generated by an *ad hoc* written perl script, as described below. If needed, new expression target entries are generated as well and will be added to `cleanex_trg`.

The `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 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 analyzed more than once in a gene expression profiling experiment (for instance if the same cDNA clone has been spotted twice on a microarray), or when different chip batches have been used for the same dataset, as sometimes the clones are not spotted on the same location across two different chip batches. This last case will be explained in the data integration part.

4.2. *CleanEx_trg*

The entries of this file type contain a physical description of the expression targets, linked to genes and quality control information. The `CleanEx_trg` does not correspond to the platform instance in GEO, in the sense that, to avoid redundancy, one entry could give information on more than one experiment set. For example, if two different datasets have used the same cDNA as a feature, there will only be one corresponding entry in `CleanEx_trg` for these two spots. Each spot will then be referenced in this entry. Nevertheless, for Affymetrix chips, or for custom arrays with specifically designed oligonucleotides, the GEO *platform* and `CleanEx_trg` concepts become similar, as these kinds of targets appear only once in

CleanEx_trg as well as in GEO *platforms*. The Affymetrix case could be discussed, in the sense that probe sets are quite often re-used in different arrays for the same organism. But in fact, even if the probe set identifier does not change, the number of tags corresponding to one probe set tends to lower, as the bad tags (the tags which match to more than one gene or which do not match any sequence) are eliminated in the most recent chips. So the decision has been made to create one CleanEx_trg entry per probe set AND per Affymetrix chip, even if this implies a certain redundancy in the annotation of individual tags.

As explained before, a CleanEx “target” stands for the sequence to which any nucleotide element which is spotted or sequenced for an expression experiment, corresponds. These elements can be either a spotted cDNA or oligo, an Affymetrix probeset, a SAGE or MPSS tag, and will be later on mentioned as a whole under the name “feature”. In short, a target entry in CleanEx_trg is an annotated feature with its corresponding gene name and possibly its position on the gene nucleotide sequence reference. Targets and features are tightly linked by an annotation procedure explained later.

The exact content of a target entry depends on the feature 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) SAGE or MPSS tags, (v) gene names and (vi) sequence database accession numbers. The latter two are not true physical descriptions of spotted features 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.

The CleanEx_trg entries consist of a stable part and a weekly updated dynamic part. The stable part is imported from external sources, such as the original feature names given by the experiment authors, or the probe set documentation files posted by Affymetrix (http://www.affymetrix.com/analysis/download_center.affx), and is used to generate the dynamic parts of CleanEx_trg via a weekly updating procedure. The primary purpose of the weekly update is to link targets to genes. This linking procedure also depends on the feature type. For public cDNA clones, sequence accession numbers and gene symbols, these links are established directly on the basis of the last available Unigene release. This is possible, because Unigene entries contain references to cDNA

clones, sequence accession numbers and gene names. This procedure is thus quite trivial, and consists of associating the Unigene accession number and its corresponding gene symbol, if exists, to each given clone number or accession number. For all feature types for which we have access to the sequence and whose relationship with Unigene is not direct, the procedure follows a different path, where the sequences given in the feature description are first mapped to mRNA sequences, for example RefSeq by Blast [47] or by in-house developed tag-matching software. Then the mRNA sequence identifiers are used to map the target via Unigene to the gene name. This indirect mapping procedure depends mainly on the type of sequence that we want to map, and individual methods for SAGE, MPSS, Affymetrix and Incyte clones (<http://www.incyte.com/>) will be developed later on in the CleanEx_trg building process part.

Of course, the link between features and their annotated part, the targets, is far from being a one-to-one relationship. Two main types of multiple relationships can be found. The first type is represented by the case when one target in CleanEx_trg represents multiple features in CleanEx_exp, either in the same dataset, or in different ones. For example, the same cDNA clone could correspond to more than one spot on the same chip. This kind of duplicate is quite frequent for the most important genes, as it serves as an internal control for the gene behavior according to the feature position on the chip, or to the feature position on the gene reference sequence. The second typical case of discrepancy between features and targets is found when one target in CleanEx_trg corresponds to more than one gene, or entry, in CleanEx. In this category, one can think of a wrongly designed Affymetrix probe, when the probe set matches different gene sequences. This also happens for shorter feature sequences, like SAGE tags. The multiple target match also occurs if the feature corresponds to a chimeric clone, a clone issued from the fusion of two pieces of ESTs coming from different genes, for example. 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. Typically, the quality tag significance and precision differs a bit according to the source of the target. It can thus take different values, according to the corresponding entry type or to the mapping protocol. Note that this quality tag reflects mainly hits on different genes, and does not take into account the splice variants problem. The significance of the quality tag will be explained and detailed for each mapping procedure. The target quality, as well as the alternative

splicing phenomena, greatly influence the results of the experiments. For example, two probesets designed for the same gene but for different splice variants could show a differential expression in two tissues or conditions. An example of a signal dilution due to the differential positions of two probesets on the gene sequence will be given in the results part.

Target entries are typically identified by the names of the corresponding reagents, e.g. an IMAGE (integrated molecular analysis of genomes and their expression) [48] clone number, a RefSeq accession number, or an Affymetrix chip and probe set name.

4.3. CleanEx

Cleanex is the catalog of officially approved genes from model organisms (for now : human and mouse) with cross-references to entries in `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 catalogs, Genew [39] for human and MGD [46] for mouse. These names are then used to establish cross-references to `cleanex_trg` entries and from there to expression data in `cleanex_exp` via the target unique identifier. The link between sequences and gene names is done via the Unigene database. To have a complete view of the transcript and its product, we also link each entry to the corresponding protein. We also provide the genomic position of the transcription start site from EPD [45], when available; otherwise we give the annotated start site position in Ensembl.

5. BUILDING CleanEx

The building procedure for the CleanEx system consists of regenerating from scratch the weekly updated files, namely CleanEx_trg, adding the dataset information contained in the stable files (CleanEx_exp) to this new version, and concatenating all the cross-references together in CleanEx. The following part will describe the building process of the stable part (CleanEx_exp), which occurs only once, and the updating procedures for the two other file types will then be described.

5.1. CleanEx_exp files : integrating expression datasets in the CleanEx database

The different platforms which have been integrated in the CleanEx system so far are :

- Dual channel chips from the Stanford Microarray Database (SMD)
- 60-mer oligoarray from the Rosetta institute (<http://www.rii.com/>).
- Nylon array from ClonTech (<http://www.clontech.com/>)
- Affymetrix experiments done with any commercially available Human or Mouse chip (<http://www.affymetrix.com>)
- EST counts
- SAGE tag counts
- The next experiment type which will be integrated soon is MPSS (Massively Parallel Signature Sequencing)

Though some features are similar between some datasets (for example the three first methods give as main output a ratio between a reference experiment and the tested condition, and the EST, SAGE, and MPSS methods all give a basic count of transcripts found), each type of dataset needs a specific protocol to be integrated in CleanEx, according to the kind of information which is provided by the dataset's authors.

Typically, the metadata for each dataset, which contains information like the type of experiment realized, organism, methods applied, paper reference and so on, give rise to the first entry of one dataset, namely the documentation file (DOC). This is the first part to generate for each dataset, regardless of its origin. This DOC entry is usually built by hand from the additional information provided by the authors and processed separately from the experiment files.

The next paragraphs are a detailed description of the procedures used to integrate expression data for all the different dataset types.

5.1.1. Stanford-like microarrays

The individual Stanford-like microarray experiment files (one file per experiment, as explained before), are processed via a perl script following this procedure (Figure 7) :

- 1) Check for the number of spots on all chips, check if the spotted clones come from the correct organism, and eliminate control or empty spots.
- 2) For each chip, extract the clone identifier and the spot number. During that step, we check the coordinates of the clones on each chip. The next process then depends on this check. If all experiments have been realized using the same chip batch, meaning having the same clone disposition for all the chips, there is no need to modify the files, and the procedure goes on with the extraction of the experiments' number (following step 3).

If there is any discrepancy between the chips, one needs to know which spot on one chip corresponds to the spots on the other chip, to be able to concatenate all the results for one feature under one entry in CleanEx_exp. To do so, the new CleanEx dataset needs to be adapted, each data entry, corresponding to one spot, will receive a new virtual number. This number will be the same across

all the chips, regardless of the spot's position on the chips. The position of this spot on each chip is kept in the new file, so that it will be possible to track this identifier on any chip. This is done by the following steps :

- a) Extract spot number and clone/sequence name from each chip.
 - b) Create an intermediate file with the following information : each line consists of well separated fields. The first field contains the clone identifier. The following ones correspond to its position or numerical identifier on all the chips. If the clone is spotted more than once on one chip, the line field corresponding to that chip will contain all the clone's position.
 - c) Consider duplicated spots, and create one line per spot. If the clone is not duplicated on all the chips, the second clone's spot will be considered as empty for the chips missing the duplicate.
 - d) Add the new spot number in each expression data file, and then proceed to the steps described below. Keep the old spot number for tracking reasons.
- 3) From the newly generated DOC, extract the experiment numbers.
 - 4) Add the experiment number to the corresponding data file.
 - 5) Put all expression data together, with the experiment number and, if needed, the new virtual spot number.
 - 6) Sort the file according to the spot number, and then sort each spot data according to the experiment number. In this manner one obtains so a complete file where all data corresponding to one spot are put together and ordered by experiment. We keep all the information given by the authors, to be able to generate statistical procedures on different fields (for example if one wants to work only with the RED channel, or if one makes use of the FLAG quality control tag).
 - 7) Last step : separate data per spot and add the entry header for each spot.

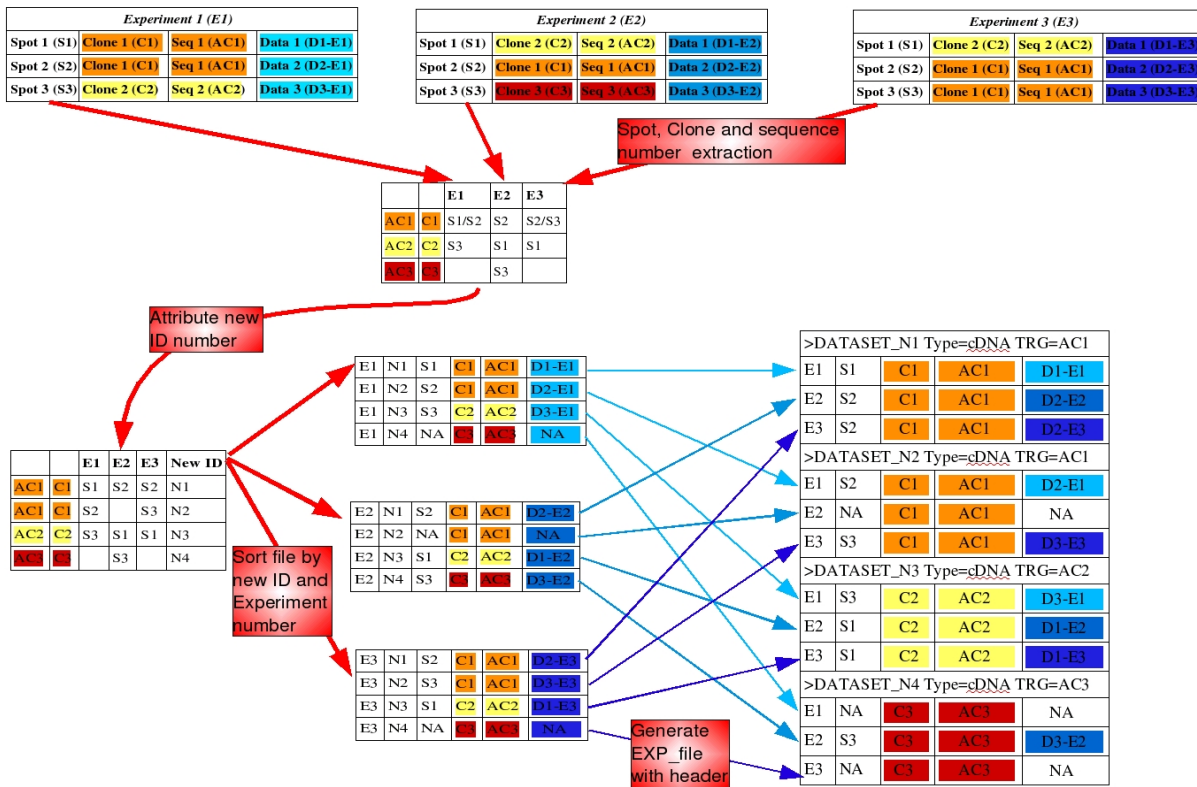


Figure 7 : Dual-Channel experiments integration. The three different blues indicate data coming from three different experiments. The red, orange and yellow colors each correspond to a different clone.

Note that, in principle, this procedure is applicable for any dual channel experiments, not only for the Stanford ones, as far as the authors give access to the output of the image analysis software, and to the experiment description.

5.1.2. Nylon membrane arrays

So far, there is only one nylon-membrane based dataset in CleanEx. Its format is very similar to the classical dual-channel chips. The reference to the “spotted feature” is usually an EMBL/GenBank nucleotide sequence accession number. If we have access to the sequences, we verify the annotation by applying the sequence-to-gene method used for INCYTE clones. This method will be explained later. If

the sequence is not provided by the authors, we rely on the authors' files, and just check that the accession numbers still exist in the reference database, and that they really correspond to human sequences. Otherwise, the method is the same as that described above for Stanford data.

The nylon-membrane cDNA array datasets are considered in CleanEx as “basic ratio”. This term designates experiments which are not based on the classical Stanford cDNA chips system, but which nevertheless measure the expression level as a ratio between a reference experiment and an analyzed sample. We provide the log₂ of the ratio in the final file as well.

5.1.3. Oligo-arrays

All the oligoarrays that are in CleanEx so far come from the Rosetta Inpharmatics Lab. They provide arrays which are spotted in-situ with 60-mer nucleotide sequences.

Usually, raw data files represent the results for each spotted oligonucleotide on one line. The line begins with the oligonucleotide identifier. It is followed by the gene name if existing, and then by the expression data. For each expression results, at least three values are given :

- Log₁₀(intensity) : The geometrical mean intensity for both red and green channels for the given probe.
- Log₁₀(ratio) : The mean ratio of the intensities of the red and green channels.
- P-value: The confidence level that a gene's mean ratio is significantly different from 1, or no change.

Unlike the other datasets, which use the log₂ ratio (so that a value of ± 1 corresponds to two-fold over- or under-expression), these oligoarrays give the ratio as a log₁₀. To facilitate the comparison between datasets, and also because the visualizer pages use the intensity instead of the ratio value, we extract the basic intensities of the two channels by applying the following formula (Figure 8):

$$I = \log_{10}(\sqrt{ChA \times ChB})$$

$$R = \log_{10}\left(\frac{ChA}{ChB}\right)$$

$$ChA = \sqrt{10^{2I+R}}$$

$$ChB = \sqrt{10^{2I-R}}$$

Figure 8: From Ratio to two channel values

In CleanEx, we thus keep the so-called “green channel intensity”, or ChB, and the “red channel intensity”, or ChA, as well as the original p-value given by the authors. The format of oligo-array based experiments is also defined as “basic ratio”.

The main issue for this dataset type remains the correlation between the original oligonucleotides identifiers and their corresponding EMBL or RefSeq accession number. This again is usually provided in a separate table, and needs to be checked for consistency before integration. Oligonucleotides for which the description does not correspond to the one given in the associated reference sequence, as well as those which have no associated reference sequence or which correspond to a sequence from another organism, are considered as “bad” oligonucleotides and are eliminated in this step. For the two oligo-array datasets which are in CleanEx from now on, nine oligos have been tagged as bad and eliminated from the files. If we once have access to the oligonucleotide sequences and not only to the identifiers, we will be able to run a procedure to retrieve the references by ourselves and maybe increase the annotation quality of these data.

5.1.4. Affymetrix dataset

From the two very different formats available for Affymetrix-based experiments, and since the CEL files are not always accessible for download, the effort has been concentrated on integrating the already processed Affymetrix data format type (earlier called second step chip analysis) which usually contains the raw intensity and the absent/present call per probe set, in CleanEx.

Incorporation into CleanEx of these datasets is facilitated because the Affymetrix chips are standards and the whole set is done with the same chip. As explained later, we provide anyway a gene-to-probe

set target file for all the available Affymetrix chips, so there is no need to check for the individual probe sets quality (as is done for dual-channel chips regarding the accession numbers), because this control will be managed in the TRG files. The integration procedure is as follows (Figure 9) :

- 1) Create the DOC file using the additional file containing experimentation protocols and descriptions, as for Stanford data.
- 2) For each line of the experiment results file, create one entry.
- 3) Create the entry header, containing the entry number, probe set identifier as the TRG reference, and the old name provided by Affymetrix.
- 4) For each experiment result, keep intensity and A/P call
- 5) For each entry, calculate the log-norm value. Log-norm value is the base 2 log of the intensity, mean-centered along the experiments, for each probe set. We use this individual probe set normalization procedure mainly because this log-norm value will then be the source value for the individual gene viewer web pages in CleanEx.

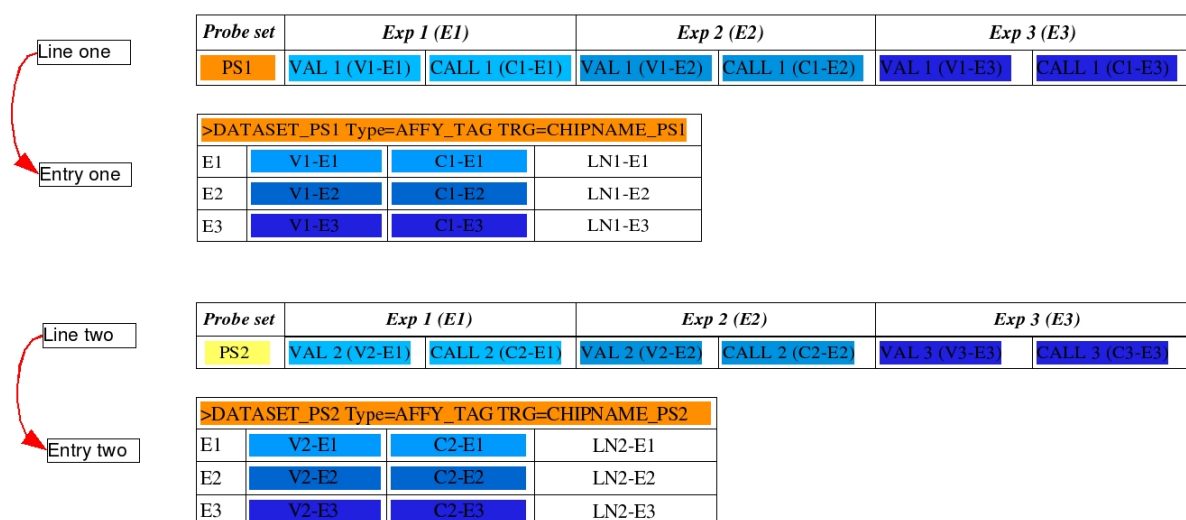


Figure 9 : Affymetrix experiments integration

5.1.5. ESTs

At the ISREC (Swiss Institute for Experimental Cancer Research, <http://www.isrec.ch>), most people working with expression data on human or mouse are often first trying to compare tumor versus healthy tissues. This prompted us to generate a new in silico expression dataset generated from a basic per-tissue split of ESTs from UniGene clusters according to the library from which they've been extracted (Figure 10). This will allow EST counts in healthy and tumor specific tissues to be compared with results obtained via other expression experiment protocols.

The tissue split is based on the library classification from CGAP (Cancer Genome Anatomy Project, <http://cgap.nci.nih.gov/>) at the NCBI. The decision use the CGAP classification came from the fact that it contains a precise description of tissue-specific libraries from the CGAP, MGC and ORESTES projects which are deposited in dbEST (<http://www.ncbi.nlm.nih.gov/projects/dbEST/>) and which can be classified as normal, precancer, or cancer. This type of classification is perfectly adapted to our need. The CGAP library classification contains fifty-five different tissue classes divided in three different histology classes. If one wants to make use of all the different sections, one obtains a very low count for some tissues. For that reason, and to be able to generate some statistically valuable data, we tried to keep a small number of tissue classes and to pool together subclasses in a way to obtain a reasonable amount of ETSs per class. We also eliminate the data coming from normalized libraries. Indeed, as these libraries are enriched in weakly expressed transcripts, they are not suitable for expression level comparison and will induce a bias in the analysis. Amongst the fifty-five tissue types, the different chosen classes which appear to contain a reasonable amount of ESTs are the following :

Colon, cancer
Colon, normal
Kidney, cancer
Kidney, normal
Lung, cancer
Lung, normal
Mammary Gland, cancer

Colon, cancer
Mammary gland, normal
Skin, cancer
Skin, normal
Cell-line, cancer
Cell-line, normal
Other tissues, cancer
Other tissues, normal

These classes have been retained according to the number of EST contents of their two respective histology types, but also according to the research interests at the ISREC.

At the update level, this dataset is a bit different from the other experiment datasets. Indeed, as it is primarily based on the UniGene database, it has to be re-generated for every CleanEx release. The procedure is fully automated and is described hereafter :

- 1) Refresh CGAP library classification from their web site : extract library identifier and full name, tissue type, tissue condition (tumor, normal).
- 2) From the Unigene Library info, extract the Unigene identifier and full name for each library found at the CGAP site.
- 3) From the Unigene clusters, classify ESTs according to their original library. Count all ESTs per tissue class, and then all ESTs per tissue class and per Unigene cluster.
- 4) Create entries : for each gene having an official gene symbol corresponding to a Unigene cluster, split cluster-related ESTs per category.
- 5) generate the EXP file. There is one entry per Unigene cluster. Each line in the entry corresponds to one experiment, meaning to one of the selected classes (tissue and condition). On each result line, three values are given :

1. The number of ESTs per class for this particular cluster.
 2. The total EST count for all genes for this class.
 3. The calculated relative amount of ESTs for this gene. This value is given as TPMs (Tags per Million).
- 6) Do all steps for both human and mouse data.

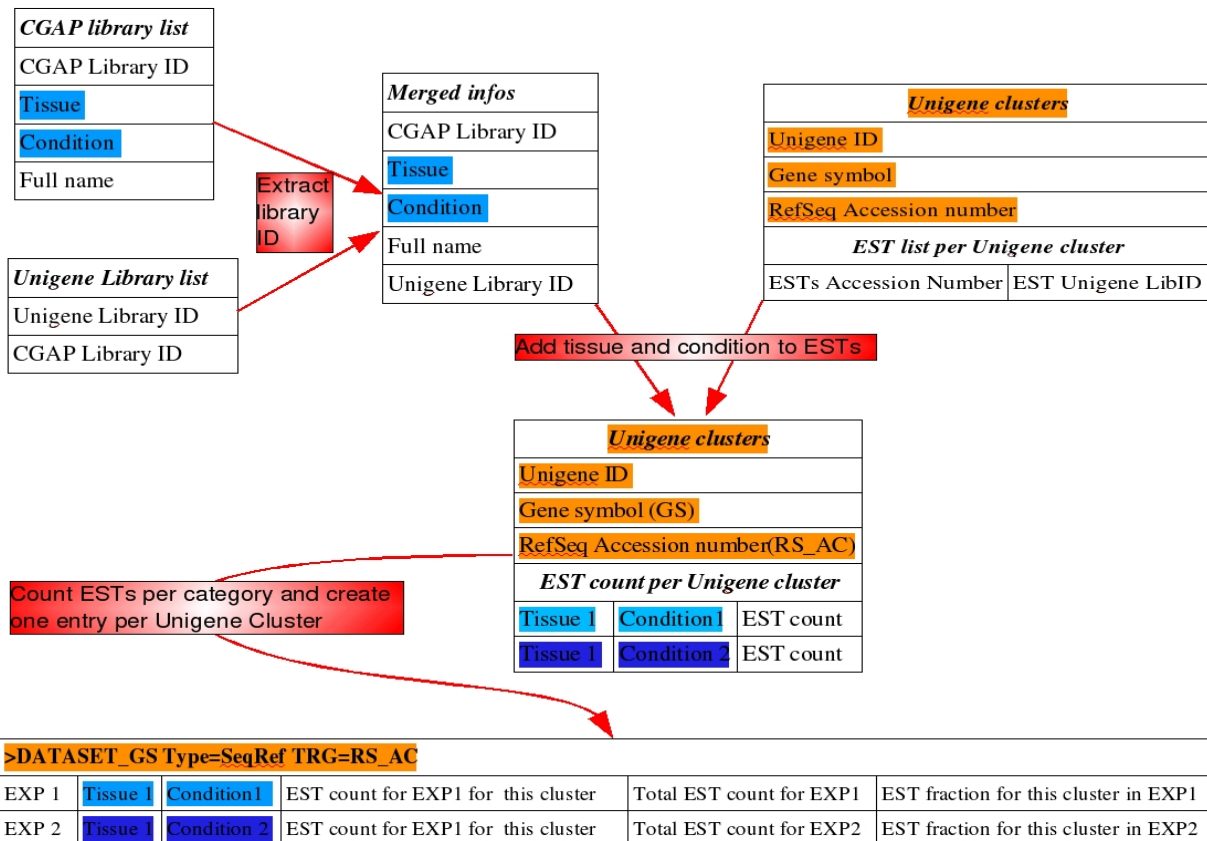


Figure 10 : EST dataset integration

5.1.6. SAGE and MPSS

As for the EST dataset, these expression data are also based on tag counts. The integration in CleanEx of this data type is facilitated by the fact that, as explained later, we provide anyway our own tag-to-gene correspondence for all SAGE tags in the CleanEx_trg file. The SAGE entries in CleanEx_trg have an identifier which is constructed by putting together the anchoring enzyme name and the tag itself. For

that reason, we can use this convention directly to automatically generate the link to the target file in the header line for each tag. The following operations consist only of the concatenation of all files to generate one entry per tag, as explained already in the Stanford-like chips part. As for the spotted arrays, it often happens that more than one entry in the EXP file corresponds to the same target in the TRG file. It will be shown later that the most difficult and time-consuming part of integrating these datasets in CleanEx is the construction of the TRG files, not of the EXP files.

5.1.7. Data from GEO : semi-automatic method

The number of publicly available data is growing quite fast, so a semi-automatic procedure has been created, which allows the direct creation of new CleanEx datasets from GEO (Gene Expression Omnibus), one of the three officially approved gene expression data repositories. As explained before, GEO has a very specific and well-designed format, including these three formerly described files types :

- 1) the *platform* used (the chip itself, like for example Affy HG_U133_PLUS)
- 2) the *series* made (all the experiments corresponding to one dataset, or in other words one publication).
It corresponds to a “meta-entry” in CleanEx.
- 3) the *sample* used, in independent formatted text files.

Each of these three components are attributed a unique identifier which allows data retrieval through the web via an in-house retrieval system for web-based documents called netfetch. As the GEO in-house format is MAGE-ML compatible, all the metadata can easily be retrieved and parsed automatically from the main *series* file. The datasets from GEO are thus the only ones for which the DOC file is also generated automatically.

This procedure appears to be especially efficient for SAGE and to a certain extent for Affymetrix data. The main reason is that for both of these data types, one does not need to make a spot-to-clone mapping, and the only information you need to extract from the platform file are, respectively, the chip name for Affymetrix data and the so-called anchoring enzyme for the SAGE data. Once these details are known, the link to the target file can easily be generated.

The procedure consists of six main steps :

- 1) Extraction of all the series ID corresponding to one platform
- 2) If necessary, extract from the platform the correspondence between spots and sequences (spot to clone file).
- 3) For each series ID, create an automatic documentation file from the information contained in the GSE file. Extract also from this file the accession numbers of the corresponding samples.
- 4) For each sample, download the data. One now has the same format type as that of the Stanford data : one file for each experiment.
- 5) Apply a slightly modified version of the Stanford procedure to recreate a CleanEx EXP meta-entry.
- 6) Add the sequence number and value scales in the DOC file.

This new procedure ended up with the generation of more than one hundred new entries for the CleanEx_exp files.

5.2. *CleanEx_trg*

Amongst the two procedures used for annotation of features and integration of CleanEx_trg (TRG) entries, the indirect mapping method is by far the trickier, but it is also the one which gives the most precise and useful results. The indirect mapping method varies according to the feature type. The main difference between the data types is the length of the feature's nucleotide sequence. INCYTE clones are very long sequences compared to Affymetrix individual probes or SAGE and MPSS tags. For these clones, using a program which is taking into account possible mismatches and gaps between the clone and the reference sequence is indispensable. On the other hand, with shorter tags, introducing mismatches will only add noise into the results. For that reason, the mapping on INCYTE clones is done using MegaBLAST [47], an algorithm for the DNA sequence gapped alignment search, while the

shorter tags are mapped via a program called “tagger” which generates a list of perfect matches on the reference sequence database. The two different techniques for indirect mapping are detailed below.

5.2.1. INCYTE clones

So far, no datasets using INCYTE clones have been incorporated in CleanEx_exp. Nevertheless, some people at the ISREC were using these clones and were very interested in comparing the annotation of chips based on INCYTE clones with other chips. For that reason, a first remapping has been generated in-house, and the data were then incorporated in CleanEx despite the fact that there is no experiment linked to these targets.

For INCYTE clones, both 3' and 5' clone sequences are available. The mapping takes place in four steps. First, using megablast, the two sequences for each clone is compared against the Unigene consensus sequence database. Once the alignment on the reference database has been performed, the resulting output is parsed. The matches are kept only if they fill these two criteria :

- The matching similarity must be more than 95%
- The total alignment length should be as long as, or a maximum of 15 bases shorter than the original clone.

The Incyte clones are then annotated using the Unigene clusters description. Finally, the quality score is assigned to each of the clones. For INCYTE clones, the quality score depends on how many Unigene clusters the clones match.

The attribution of the quality criteria follows these rules :

- 1 : Both 3' and 5' ends of the clone are available and match the same Unigene cluster.
- 2 : Either 3' or 5' ends of the clone are available and give a statistically significant result.
- 3 : Both 3' and 5' ends of the clone are available, but only one is statistically significant.
- 4 : No statistically significant results have been found.

5 : Both ends of the clone match different genes.

6 : The sequence is not yet available.

It is interesting to note that this procedure is easily adaptable to any kind of nucleotide sequence. It could be used for example to re-annotate oligonucleotide-based arrays, as far as the access to the raw sequences is given.

5.2.2. Affymetrix probe sets

For each chip, Affymetrix releases annotation files, which link the probe set sequence to their corresponding transcript. Though renewed on a regular basis, there are anyway two interesting issues about Affymetrix annotations.

First, as we update CleanEx at the same time as its main resource, meaning Unigene, we have to perform a weekly control on the annotation. Affymetrix does not provide with such a regular update.

Second, the annotation files are given for so-called “consensus sequences”, which correspond to the whole sequence spanned by the individual probes of one probe set. The spotted features on the chip are 25 nucleotide long oligonucleotides, not a consensus sequence. As a consequence, the behavior of the hybridization process depends more on the probes than on the consensus sequence. For example, if one probe is found to share its sequence with two or more genes, its corresponding signal will be shared by all the target genes. Also, if one probe does not match the real targeted transcript, the total signal for this transcript will be diluted, even though outliers are minimized by the analysis softwares.

Thus, knowing the precise position of the probes allows experimenters to give a different weight to their results, depending on the real accuracy of the annotation (genes sharing probes, probes without matches, etc...), and even, to a certain extent to distinguish between differentially regulated splice variants of the same transcript.

Based on these considerations, we decided that using the Affymetrix annotation file might lower the accuracy of our target annotation quality. We thus introduced a new procedure to remap the individual tags on organism-based transcript databases, following the steps described below (see Figure 11) :

The whole process is built around a program developed at the SIB (Swiss Institute of Bioinformatics) by Christian Iseli, and maintained for the SIB by Giovanna Ambrosini. This open source program is called “tagger”, and it reports the complete list of perfect matches for a given list of short tags. For this particular problem of finding perfect matches for short tags of the same length, the tagger program is much more adapted than for example the BLAST program, and is much faster. The tagger source code is available via sourceforge (<http://sourceforge.net/projects/tagger>). It works as follows :

- Given a text file containing all the tags to test, one tag per line, and the sequence reference database to search (here we use as reference RefSeq, as well as mRNAs, HTCs, and ESTs from EMBL), tagger finds all occurrences of all tags within the specified list of reference sequences (the so-called reference database). To do so, it generates all the possible tags of the given length with the input reference database sequences, stores the sequence identifier as well as the positions of each tag generated with this sequence as an index, and then finds common features between the list of tags generated with the reference sequence database and the input tag list. By applying this technique, no match is missed, each tag to tag correspondence is stored, as well as the reference sequence(s) name(s) and the position of the match on the sequence(s).

- The following line is an example of the tagger' s output format :

```
GCCTCCCAAAGTGCTGGGATTACAG          NM_000367  +  NA          1423
CTGGGATTACAGGCGTGAGCCACTGCACCTGGCCTGACATTCTTTATGAA 2742
```

- There is one match report per line. The first field is the tag given as input. It is followed by the reference sequence identifier where the match occurs, the match orientation, the chromosome name (if available), the match start position, the target sequence directly following the 13 first nucleotides of the input tag, and the total length of the reference sequence.

From the tagger output, the extracted information (sequence and position) of all the matches are reintegrated in the primary tag file. If one tag has more than one match, these are concatenated and checked for discrepancies. Discrepancies, in this case, mean that one tag matches two different sequences, and that these two sequences correspond to two different genes. At that level, only the

individual probe discrepancy is taken into account.

After this check, all probes corresponding to one probe set are put together. At this point, a second verification step occurs, which checks for the whole probe set quality, by comparing the target sequences of all the corresponding probes.

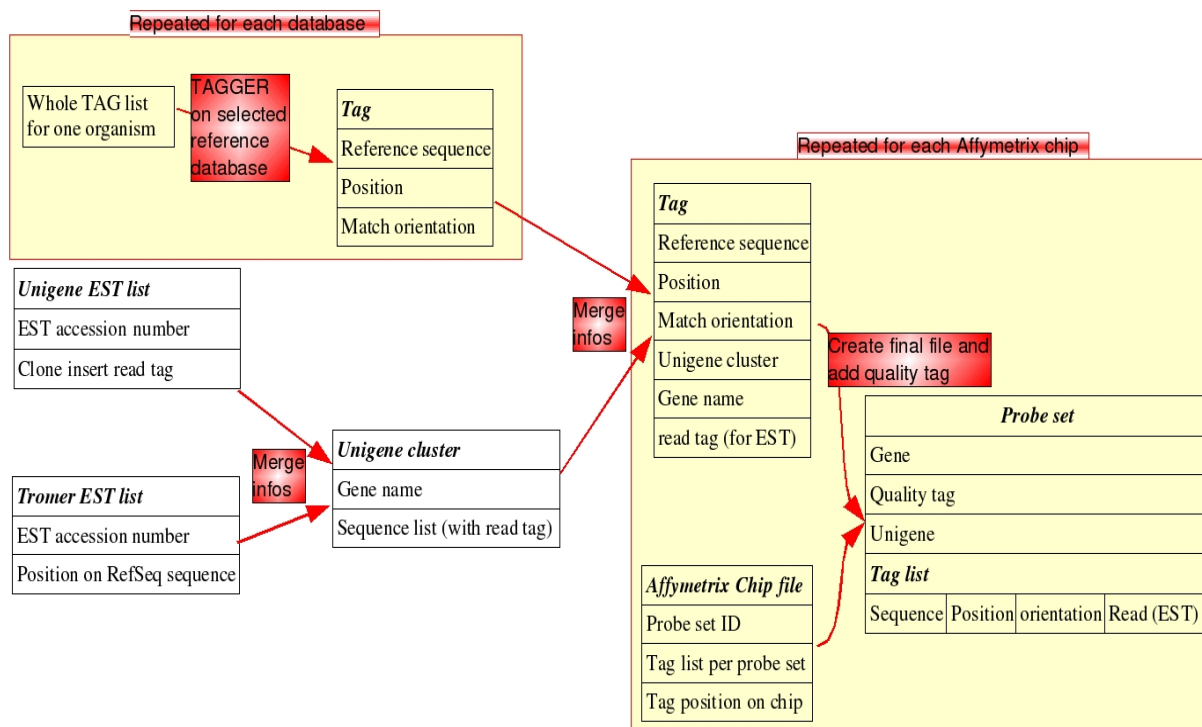


Figure 11: Creation of the target file for Affymetrix chips

In the provided file, each entry corresponds to one probe set. This entry also gives access to the position of each individual probe in the probe set, and includes a quality criteria based on the two integrity check steps. There are four different quality tags : High, Medium, Low and Unknown, attributed according to the matching procedure result.

The definition of the quality criteria follows these rules :

- A “High” quality probe set has a maximum of 2 Unigene identifiers that matched to it. All probes have to match all Unigene identifiers.

- A “Medium” quality probe set matches a maximum of 4 Unigene identifiers. In addition, a maximum of 3 “errors” were permitted. Errors were defined as probes that matched nothing, probes that failed to match a Unigene identifier or probes that matched an additional Unigene identifier.
- Anything below these criteria was considered to be of “Low” quality.
- The “Unknown” tag is given to probe sets for which absolutely no match on the selected mRNA databases was found.

A few comments on the criteria selection :

- The reason why we decided to take these criteria, and to allow for example a matching on two clusters instead of one for the “high” quality, is linked to the development stage of Unigene when we began this process. Indeed, three years ago, it often happened that probes matched two Unigene clusters which corresponded to the same gene, but had not yet been clustered together. Given these conditions, one of the clusters usually lacked the gene name, which was found in the other cluster. As the Unigene database has improved, these criteria could be oriented in a different way, by putting the probe set quality to “high” if a maximum of one tag does not match the same cluster. The quality threshold would be now more related to the output of the analysis software, which tend to lower the outliers influence on the result, and thus allows some flexibility in the annotation quality threshold.
- Several probe sets of the Low quality were found to match in excess of 700 different mRNA sequences, which in turn corresponded to several hundred Unigene identifiers. It was clear that the individual identifiers were of little relevance for these entries. Therefore, a limit as to the number of identifiers to be listed in CleanEx-trg final entries was set. A maximum of 4 Unigene identifiers is listed along with the corresponding RefSeq matches. All listed Unigene identifiers were required to match at least half of the probes in a probe set. The number of matches not explicitly listed is displayed in parentheses at the end of the list, as well as the number of corresponding gene symbols. A detailed description of this format, as well as an example of an entry (see Figure 17), is given in the results section.

Until quite recently, the CleanEx target files for Affymetrix were only based on the RefSeq [14]

database. Though growing quite fast, RefSeq is actually quite incomplete, and mapping the tags on RefSeq only creates a high number of losses from the quality point of view for CleanEx.

On the other hand, we were asked to also provide tag mappings on other mRNA databases. The same system has then been used to map all the Affymetrix chips on the following databases :

- RefSeq
- HTC subdivision of EMBL
- mRNA subdivision of EMBL
- dbEST, the EST subdivision of EMBL

The mapping on the first three databases does not cause orientation problems, as these databases contain sequences which have the same orientation as the original transcript, so the corresponding Affymetrix probe sets should also match on the same orientation. It is quite different regarding the EST database. An EST can be sequenced from both ends, and is then entered as is in the database. Sometimes the sequencing orientation is available in the EST description line, but this is not always the case. Moreover, the Unigene database keeps track of this orientation information, when present. So to be really consequent, we decided to apply an orientation filter on the tagger result coming from ESTs. We keep only results matching the same orientation as the one described for the corresponding EST. For example, if an EST is described as being “3' sequenced” in Unigene, we keep only tagger results which match the complementary strand of this EST.

We then realized that Unigene annotations concerning the mapping orientation of ESTs (usually described as 3' or 5') sometimes happen to be wrong, and then applied a new control step in our method. This step is based on the in-house transcriptome project called “trome” which has been described previously.

From trome, we extract the EST orientation regarding the mRNA reference sequence of the corresponding gene. We then compare this orientation with the Unigene tag, and correct it if necessary. We next apply this new orientation annotation instead of the Unigene one if it exists, otherwise we keep

the Unigene orientation description.

We provide annotation for all the main Affymetrix chips and organisms on all the formerly cited databases, as far as these databases exist. If there is no specialized database for this organism, we try to extract the organism's specific sequences from the upper taxonomy level in the sequence files. For example, for the bovine chip we extract the cow sequences from the mammalian division of EMBL mRNAs.

The updated annotation files are available on the SIB ftp server (ftp://ftp.isrec.isb-sib.ch/pub/databases/CleanEx/Affy_mapping/). Each subdirectory in this site contains organism-specific chip annotation files corresponding to the mapping on the four databases. For the EST database matches, ESTs with tags 5' or 3' are accessible in two different files, respectively flanked with the extension “_PLUS” or “_MINUS”.

These mapping files are formatted as follows :

Each line contains one match for one individual probe. Supplementary information included are : UniGene accession number, gene symbol and LocusLink (now Entrez GeneID, <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch19>) [49] accession number, if it exists.

By comparing the mapping results with these different databases, we realized that introducing these mappings in CleanEx will indeed increase the number of probe sets with a high quality tag. After having checked the contribution of these different databases in the quality tag improvement, we found out that the two most useful databases, apart from RefSeq, are indeed mRNAs and HTCs division from EMBL. The ESTs did not increase the quality that much. Moreover, using the ESTs as a first reference might induce errors, for example because EST sequencing produces a high error rate and that tagger only deals with perfect matches. So we decided to integrate the mRNAs and HTCs mappings in CleanEx. Supplementary Tables 3 and 4 show the quality gain for each human and mouse chip at each new database integration step. The gain obtained with the matches on the EST database was considered insufficient, and hence these were not integrated.

As we wanted to keep the RefSeq high quality entries anyway, we used a step-by-step procedure which compares the other mappings to the RefSeq ones, and which integrates in the final file the entries having the highest quality tag, but only if these new entries give a better quality information than the RefSeq ones.

5.2.3. SAGE tags

The traditional SAGE protocol makes use of 10 nucleotide long tags. Adding the anchoring enzyme sequence gives a final tag length of 14 nucleotides. To be able to extract the corresponding possible gene from this tag, one has to take into account the fact that the usual size of the sequence which is cut by the enzyme is not longer than 500 nucleotides. One then has to search for a tag occurrence in the last 500 nucleotides of the 3' end of the predicted genes for the organism. Though, to allow alternative splicing or if the enzyme has more than one restriction site on the sequence, the search length is often extended to 1000 nucleotides.

To complete the SIB trome project, this mapping has already been done by Christian Iseli for SAGE tags, LongSAGE tags, as well as for MPSS tags. The mapping result is given in flat files for the predicted SAGE tags for the NlaIII, long NlaIII, Sau3AI, and Sau3AI through MPSS enzymes respectively. The format of the files is a tab-delimited list with the following elements:

- 1 - stable identifier
- 2 - tag sequence
- 3 - tag ordinal number (from 1, for the 3'-most tag, to 3)
- 4 - gene symbol (can have multiple, separated by "; ")
- 5 - Swissprot AC (can have multiple, separated by "|")
- 6 - descriptions (can have multiple, separated by "; ")
- 7 - associated 3'tag (can have multiple, separated by "|")

8 - 3'tag ordinal number (same order as column 7) number 1 is 5' most

9 - PolyA signal flag (1 if present, 0 if not)

10 - mRNA sequences used as evidence (separated by "|")

11 - minimal observed distance from end of sequence

12 - maximal observed distance from end of sequence

13 - D for genomic based, R for RNA only based

14 - genomic contig where the tag is located

15 - strand of the genomic contig where the tag is located

16 - position of the genomic contig where the tag is located. The position is the first nucleotide of the restriction site

17 - N for normal, S when the tag spans a splice junction.

For the CleanEx_trg files, there is one more piece of information that we need to extract from the tag/reference sequence alignment which is not in trome files : the exact position of the tag on the corresponding expressed sequence. We obtain this position by using tagger on the SAGE tags.

As a way to gain a considerable amount of time during the CleanEx release, the SAGE tags mapping is done on a trome-based pre-filtered reference sequence database. The filter consists in creating a temporary reference database by selecting only the sequences which are considered to contain a SAGE tag in trome. This reduces the search space for the tagger program and thus makes the release much faster and much more accurate, as we keep only matches which correspond to a restriction enzyme site close to the 3' end of the gene (Figure 12). Note that as the mapping procedure occurs on the RefSeq sequences as well as on mRNAs or ESTs from EMBL, this allows the retrieval of tags corresponding to different variants of the expressed gene.

The tagger program is then used on this temporary database as for Affymetrix individual probes. In the second part of the procedure, the integration of these data in CleanEx_trg, we use a slightly modified version of the Affymetrix probe set quality control for CleanEx. Indeed, for SAGE tags, there is only one control step which is necessary, as one single tag is meant to represent one transcript. Of course, ambiguous tags exist. For example in the last CleanEx_trg version, 1'160 SAGE tags, NlaIII- or Sau3AI-based, have a Low quality criteria, meaning that they match on more than three Unigene clusters, 38'491 have a Medium quality criteria (they match on three Unigene clusters), and 192'267 have a High quality criteria (they match on a maximum of two Unigene clusters). Here are a few examples of the genes which are related to NlaIII SAGE tags with different quality annotation results.

<i>Gene name</i>	<i>High tags</i>	<i>Medium tags</i>	<i>Low tags</i>
TP53	10	2	0
EGFR	17	10	0
ABCB1	1	3	0
ERBB2	11	1	1
TNF	3	1	0
FN1	22	11	0
WNT	1	2	0

Considering this, one may well ask whether tags matching more than one gene should be eliminated from the analysis. This will obviously result in a bias in the analysis, as all the tags, including the ones with medium or low quality, could very well come from the studied gene, and eliminating them will lower the real gene expression measurement. On the other hand, getting rid of the bad tags will have a smaller influence on the comparison of same tags across different experiments, as this resembles the traditional dual-channel “ratio” procedure and measures a relative change of expression. The solution chosen by the people who generated SAGEmap is to choose for each tag only the gene giving the highest score according to the criteria explained previously. In CleanEx, all of the information is given to the users. The quality criteria applied for SAGE tags is the same as the one used for Affymetrix. As

the CleanEx database not only contains the most 3' end tags, the tag position on the reference sequence is given, and might also help deciding whether to keep or to discard the suspicious tag.

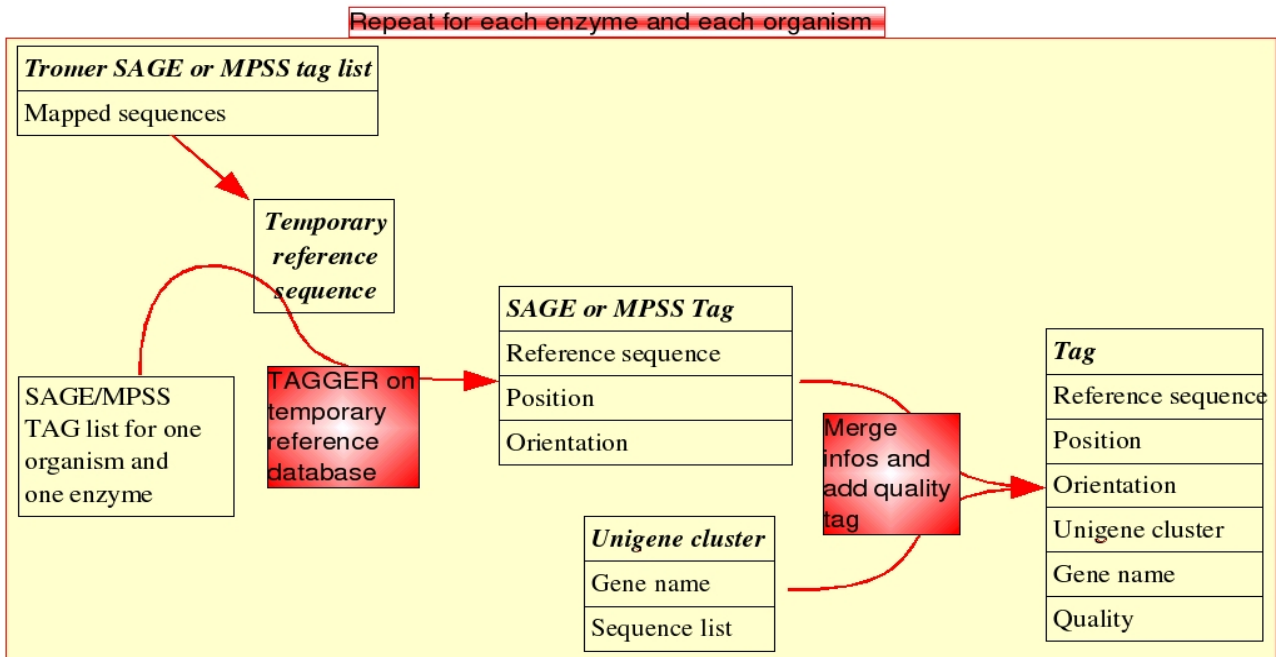


Figure 12: Creation of the target file for SAGE or MPSS tags

5.3. CleanEx link file between external databases and the CleanEx system

To begin the description of the CleanEx building process, a short presentation of the databases which are used during this procedure, and which represent the source material used for each CleanEx release, will first be provided. The next part will focus on the different steps which allow to combine the useful information in all these databases into CleanEx, together with gene expression information stored in CleanEx_exp, via the link files CleanEx_trg.

5.3.1. Material : source databases

5.3.1.1. Genew

Genew [39], the Human Gene Nomenclature Database, is the primary resource that provides data for all human genes which have approved symbols based on specific nomenclature guidelines (<http://www.gene.ucl.ac.uk/nomenclature/guidelines.html>). It is managed by the HUGO Gene

Nomenclature Committee (HGNC) [42], and contains a rapidly growing number of records. The data in Genew are highly curated by HGNC editors. Data are integrated with other human gene databases, e.g. GDB, LocusLink and SWISS-PROT, and approved gene symbols are carefully co-ordinated with the Mouse Genome Database (MGD).

The different fields in Genew which are used by CleanEx are the following ones :

- HGNC ID - A unique numeric ID provided by the HGNC.
- Approved Symbol - The official gene symbol that has been approved by the HGNC and is publicly available. This will be used as the CleanEx unique entry identifier for human data.
- Status - Indicates whether the gene is classified as:
 - *Approved* - these genes have HGNC-approved gene symbols
 - *Approved non-human* - these entries have been approved in order to maintain the orthologous gene symbol in the human gene family series. It is quite likely that most of these genes will ultimately be found in the human genome
 - *Entry withdrawn* - these previously approved genes symbols no longer exist

In CleanEx, only entries with the “*Approved*” tag are integrated.

- Previous Symbols - Symbols previously approved by the HGNC for this gene.
- UniProt ID - The UniProt identifier, provided by the EBI (<http://www.ebi.ac.uk/>).

The UniProt ID is derived from external sources and as such are not subject to HGNC strict checking and curation procedures. We will use this information only in cases where we can not link the gene symbol to the Swissprot database.

5.3.1.2. MGD

MGD is the mouse official gene database from the Jackson laboratory. It includes data on gene characterization, nomenclature, mapping, gene homologies among mammals, sequence links, phenotypes, allelic variants and mutants, and strain data.

The different fields in MGD which are used by CleanEx are the following :

- MGI Marker Accession ID - A unique numeric ID provided by the Mouse Genome Informatix database.
- Marker Symbol - The official mouse gene symbol. This will be used as the CleanEx unique entry identifier for mouse data. We exclude the withdrawn entries from integration in CleanEx
- Status - Indicates whether the gene is classified as:
 - *O* - these genes have MGD-approved gene symbols
 - *W* – withdrawn. These previously approved gene symbols no longer exist
- Secondary Accession IDs – MGI Accessions previously used by MGD for this gene.
- SWISS-PROT Protein Accession IDs - The Swissprot identifier.
- RefSeq ID – The Reference Sequence accession number
- Entrez GeneID – Locus number from the NCBI

As for the UniProt ID in Genew, the three last fields are not internally curated by MGD, so they will be used only in cases where we do not have access to the original information.

5.3.1.3. Unigene

UniGene [13] is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

The Unigene clustering process is done in several stages, with each stage adding less reliable data to the results of the preceding stage. Builds are either genome-based or transcript-based. The main transcript-based clustering steps include :

- Elimination of contaminants

- Alignment and clustering of mRNA sequences
- Alignment and clustering of ESTs
- Removal of bad clustered sequences and of low quality clusters, like clusters without polyA tail.

For the genome-based clustering, the process begins with the identification of transcript boundaries on the genomic sequence, and makes use of the intron-exon boundaries to segregate, for example, overlapping genes on opposite strands, or genes located within introns of other genes.

Currently, sequences from the animals human, rat, mouse, cow, zebrafish, clawed frog, fruitfly and mosquito, as well as from plant organisms like wheat, rice, barley, maize and cress have been processed.

Each Unigene release includes amongst others, the following files, for each organism :

- lib.info file - Additional information regarding the LID (Library ID) field. This file is used to generate the EST dataset

- data file – Unigene clusters. Each cluster entry contains links to the following features, which are reported in CleanEx :

- Unigene accession number
- Gene Symbol, if it exists, as well as the gene description.
- Cytological band
- Entrez GeneID (formerly called LocusLink)
- Concatenation of all the mRNA sequences which cluster together, including RefSeq sequences, mRNAs from GenBank/EMBL, as well as the list of all the clustered ESTs. The EST description includes the clone identifier and the clone insert read.

The EST list is used to create the per-gene split EST count dataset. From the EST description, we also keep the insert read for the mapping of Affymetrix chips on ESTs procedure.

The clone identifier is used to generate the direct clone-to-Unigene mapping procedure, and the list of clustered mRNAs to map the tags on the other databases (HTC, RefSeq and mRNA from EMBL).

The Unigene database is a work in progress, and is updated weekly. As this is the main information source to link CleanEx expression data with the official gene symbols, we try to keep the same update timing for CleanEx.

5.3.1.4. RefSeq

RefSeq [14] is the NCBI curated Reference Sequence collection. It aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.

The main features of the RefSeq collection include:

- non-redundancy
- explicitly linked nucleic acid and amino acid sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators, with reviewed records indicated

The RefSeq mRNA entries serve as the major source for tag mapping. They are also used as target identifiers, when more precise information is missing.

5.3.1.5. Swissprot

Swissprot [40] is a curated protein sequence database which provides a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.) and a minimal level of redundancy. To provide a good integration with other databases (nucleic acid sequences, protein sequences and protein tertiary structures), SwissProt is currently cross-referenced with about 60 different databases. Amongst all these databases, Swissprot provides a link to genew (for human entries) and to MGI (for mouse entries) accession numbers. We use this reference to link Swissprot to CleanEx. Note that this link is reciprocal, as the Swissprot database uses the same system to link human and mouse entries to CleanEx.

5.3.1.6. EPD

The Eukaryotic Promoter Database (EPD) [44, 45] is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally. Access to promoter sequences is provided by pointers to positions in nucleotide sequence entries. The annotation part of an entry includes a description of the initiation site mapping data, cross-references to other databases, and bibliographic references. Cross-referenced databases include Swissprot and the genomic position of the transcription start site, when available.

In CleanEx, we link EPD via its Swissprot cross-reference. We also include the transcription start site determined in EPD.

5.3.2. *CleanEx: data integration method*

For each new Unigene release, the CleanEx files to be updated are rebuilt from scratch (Figure 13). As explained before, in the CleanEx_exp file type, only the meta-entry extracted from the EST count needs updating.

The CleanEx_trg file type is rebuilt as well for each target type (clones, Affymetrix, SAGE..). Concerning the TAGs type targets, as their mapping includes a sequence comparison part, the procedure depends on the updating rhythm of the mapping database. RefSeq is updated every week, whereas the EMBL release occurs every three months. Between EMBL releases, we keep the sequence matches positions already found for RNAs, ESTs, and HTCs sections, and redo only the sequence accession number mapping on Unigene clusters. For the RefSeq database, we use the complete procedure, with the tagger part, for every CleanEx release. Once these updates are ready, the _trg files are formatted for CleanEx :

- 1- From the previously updated Affymetrix and SAGE annotation files, format the CleanEx Affymetrix and SAGE target database, and include links to the individual experimental data.
- 2- From the Unigene new release, extract clones and their corresponding cluster number and gene symbol (if exists). This will lead to the generation of the CleanEx_target database for clones-based

experiments. Links to individual experimental data are also included.

Once the CleanEx_trg files are ready, the following part concerns the update of the CleanEx file itself. This update is a step-by-step integration process, going from the Unigene database and the official organism gene symbol database, extracting the information from all the different cross-linked databases, and putting all the information, expression data included, in gene-oriented entries. All the integration steps are described below. They consist in integrating first all the references linked to Unigene and sequence clusters, and then all references linked to the approved gene symbols database.

5.3.2.1. Unigene-related steps

1- From the new Unigene release, extract the following fields, and store them in a temporary file :

- Unigene cluster accession numbers (line ID)
- Gene description (line TITLE)
- Gene symbol (line GENE)
- Entrez GeneID (line LOCUSLINK)
- Locus position (line CYTOBAND)
- RefSeq associated sequences (in the lines SEQUENCE, only the RefSeq entries)

The temporary file now contains one Unigene cluster per line. Each extracted field is separated by a common field separator (we use “|”, as this symbol is absent from all the extracted lines).

If there is more than one information per field, for example if the gene's position is not yet well determined, or if it is duplicated, each sub-field is then separated via a new separator.

2- From the EMBL database, extract the mRNAs list, search in Unigene for their corresponding cluster, and add this RNA list to each corresponding line in the former temporary file.

3- Sort this file via gene symbols, and concatenate all the references corresponding to Unigene clusters which have the same gene symbol. Having two Unigene entries for the same gene happens sometimes when the clustering procedure is unable to generate one single cluster for one gene, due to the lack of a

total gene sequence coverage by ESTs. In this case, one single gene is usually represented by two clusters, one in the 3' and one in the 5' region of that gene. It could also be that two clusters correspond to the same gene, but have not been merged yet in the database.

5.3.2.2. Gene nomenclature-related steps

4- From the Swissprot database, extract the Swissprot accession number and identifier, as well as the corresponding gene symbol.

5- From the EPD database, extract the EPD accession number, the EPD identifier, as well as the corresponding Swissprot accession number. Merge the EPD information to the Swissprot file. Sort this file via gene symbols again, and merge this new information (Swissprot and EPD) to the former file.

6- From the newly generated CleanEx_trg files, extract the target unique identifier as well as all the references to individual expression experiments, per target entry. Put together all the information related to one single gene symbol, and again add these expression references to the former file.

7- From the officially approved list of gene symbols, extract the database accession number. For human, this number corresponds to the Genew unique identifier. From Genew, the MIM accession number is also extracted. For the mouse, the identifier is the MGI accession number. Make a table between these identifiers and the corresponding gene symbols. Eliminate all the old or withdrawn entries, but keep track of the old gene names, if they exist.

8- Concatenate the Unigene-related construction file and the symbol-related construction file. Keep only lines having an official symbol.

9- Additional step : adding the genomic position of the gene's transcription start site, if known. As the CleanEx online tools allow retrieval of a pool of genes which share some expression features, we thought that it could be useful to provide a new link to the genomic position of the transcripts. Having this reference in CleanEx means a huge gain of time in the promoter sequence retrieval for further 5' sequence analysis. To give access to a relatively precise transcription start site position, we extract this information from the Eukaryotic Promoter Database whenever possible. Otherwise, we rely on the position given by the NCBI through the Entrez GeneID genomic annotation file. However, to keep track

of the origin of the given position, the line is tagged with the word “EPD” or “ANNOTATION” for EPD-based position or NCBI-based annotation respectively.

External resources

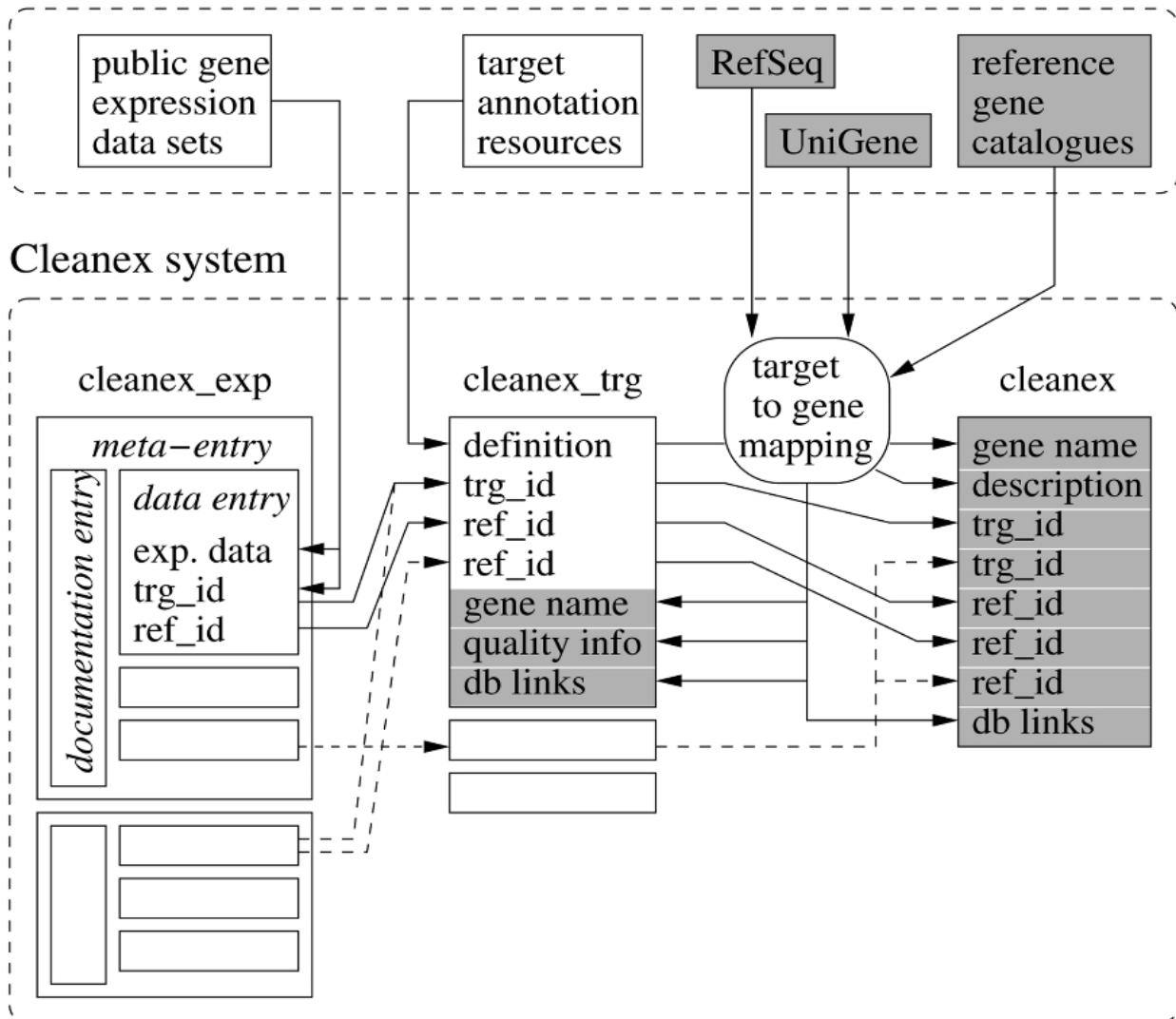


Figure 13: CleanEx data integration protocol

6. RESULTS

6.1. Survey of the most recent release

The content of the last CleanEx database is given in a table showing all the references to other databases and to specific expression datasets. Datasets are pooled together according to their type.

References from UniGene Build #183 Homo sapiens

Number of entries	18792
Number of RNA cross-references	84602
Number of Entrez GeneID cross-references	15622
Number of Unigene cross-references	14485
Number of Genew cross-references	18778
Number of RefSeq cross-references	19055
Number of EPD cross-references	1389
Number of SWISS-PROT cross-references	10052
Number of cross-references to EST counts	13606
Number of cross-references to dual channel experiments	79290
Number of cross-references to Affymetrix experiments	742012
Number of cross-references to SAGE experiments	124554

References from UniGene Build #146 Mus musculus

Number of entries	32982
Number of RNA cross-references	34639
Number of Entrez GeneID cross-references	25293
Number of Unigene cross-references	18834
Number of MGD cross-references	32421
Number of RefSeq cross-references	19305
Number of EPD cross-references	94
Number of SWISS-PROT cross-references	8468
Number of cross-references to EST counts	13173
Number of cross-references to Affymetrix experiments	6993
Number of cross-references to SAGE experiments	6832

6.2.Database format

6.2.1. CleanEx

CleanEx entries are presented in a similar format as EMBL, SWISS-PROT, or EPD entries. Each line starts with a two or three letters line code identifying the type of information presented. The current line types and line code are shown below:

ID - IDentification.

DE - DEscription.

ON - Old gene Name.

RNA - RNA sequence in EMBL.

DR - Databases cross-links.

EXP - EXPression cross-references.

// - Termination line.

Spacer lines (XX) are inserted in order to make the database easier to read by eye. Some line types occur many times in a single entry. Each entry must begin with an identification line (ID) and end with a terminator line (//). An example of a CleanEx text entry is given in Figure 14.

```
ID    HS_FN1    2q34.
XX
DE    Fibronectin 1.
ON    none.
OS    Homo sapiens.
XX
RNA   EMBL; X02761.1; HSFIB1.
XX
DR    GENOME; NC_000002; -(2); 216126296
DR    Entrez GeneID; 2335.
DR    Unigene; Hs.203717.
DR    MIM; 135600.
DR    Genew; HGNC:3778; FN1.
DR    RefSeq; NM_002026.
DR    SWISSPROT; P02751; FINC_HUMAN.
DR    EPD; EP16038; HS_FN1.
XX
EXP   A0001; A0001_1321; RNA_X02761.
EXP   GDS420; GDS420_210495_x_at; AFFY_HG-U133A_210495_x_at.
EXP   GSE953; GSE953_TGTCCTTACCA; SAGE_NlaIII_TGTCCTTACCA.
EXP   HSEST; HSEST_FN1; NM_212482.
EXP   LYMPHOMA1; L0001_15953; IMAGE_139009.
EXP   NCI60; NCI60_136798; IMAGE_136798.
EXP   NCI60; NCI60_151144; IMAGE_151144.
EXP   PEROU1; P0001_269203; IMAGE_269203.
EXP   PEROU1; P0001_296556; IMAGE_296556.
EXP   PEROU1; P0001_60846; IMAGE_60846.
EXP   ROSETTA; R0001_20907; RNA_X02761.
EXP   R0002; R0002_24261; RNA_X02761.
EXP   SERUM1; S0001_136798; IMAGE_136798.
EXP   T0001; T0001_16455; IMAGE_1870935.
//
```

Figure 14 : Example of a CleanEx text entry

A detailed description of each line type follows.

The ID line

The identification line is always the first line of an entry. The general form of the ID line is:

ID GENE_NAME genetic_locus.

GENE_NAME is the species code followed by the gene identifier which obeys the organism-specific nomenclature rules.

The genetic_locus field is the cytogenetic location of the gene. It is cross-linked with the NCBI's genome map viewer.

The DE line

DE Fibronectin 1.

The description lines contain general descriptive information about the gene. It is extracted from the corresponding Unigene entry. The description is in English and is free-format. In some cases, more than one DE line is required; in this case, the text is divided only between words.

The ON line

ON STGD1, ABCR, RP19, STGD.

The ON line describes the history of the gene nomenclature. It lists all the previous gene symbols which have been attributed to the specific gene.

The RNA line

It contains cross-references to the mRNA entries for this gene. These mRNAs are found in the EMBL <<http://www.ebi.ac.uk/embl/index.html>> database. The RNA lines can refer to partial mRNAs. The format is a three-field line separated by “;”:

RNA EMBL; EMBL_SV; EMBL_ID.

- The first field is the target database code
- EMBL_SV is the EMBL sequence version number.
- EMBL_ID is a secondary identifier or name for the EMBL entry.

The DR lines

The DR lines contain cross-references to entries from other databases. So far, we have incorporated links to SWISS-PROT, LocusLink, RefSeq, Unigene, GeneCards and EPD. The precise format of these lines depends on the target database. The format of the DR line is shown by the following examples :

- DR GENOME; NT_005403; -(2); 66510206; ANNOTATION
- DR Unigene; Hs.339722.
- DR Genew; HGNC:3778; FN1.
 - DR SWISSPROT; P02751; FINC_HUMAN.

The first item on the DR line is the abbreviated name of the data collection to which reference is made.

The currently defined databank identifiers are the following:

GENOME	Genomic contigs from RefSeq.
Entrez GeneID	A single query interface to curated sequence and descriptive information about genetic loci.
Unigene	NCBI RNA clusters
MIM	The Mendelian Inheritance in Man Database, a catalog of human genes and genetic disorders (only for human).
Genew and MGI	The respective gene symbol catalogs
RefSeq	The NCBI Reference Sequence project.
SWISSPROT	Protein sequence database.
EPD	The eukaryotic promoter database.

The second item is the primary accession number (or an equivalent unique identifier of another data bank) of the entry to which reference is made.

The meaning of the third item (if present) is database-dependent. In most cases, it is a secondary identifier or name for the cross-referenced database entry. For Genew, this number is the HGNC (Hugo Gene Nomenclature Committee) unique identifier of the gene. A very special case is the GENOME line. Fields after the first unique identifier are the orientation of the gene on the given genomic sequence, followed by the chromosome number. The next field is the position of the transcription start site. The last field gives the origin of the data (EPD or ANNOTATION).

The EXP line

The EXP line contains cross-references to the publicly available data on human gene expression. Currently, 122 published data sets are integrated in CleanEx. An exhaustive list of these datasets as well as a short description of the experiments realized is accessible through the CleanEx web pages (<http://www.cleanex.isb-sib.ch/datasets.html>).

The format of the EXP line is a period-delimited fields line shown by the following example.

- EXP AFFY001; AFFY001_1575_at; AFFY_HC-G110_1575_at; High.

The first field of the EXP line is the abbreviated in-house name of the data collection to which reference is made.

The second field is the local identifier of the corresponding expression entry. It is cross-referenced with the CleanEx_exp entries.

The third field is the reference to the CleanEx_trg corresponding entry.

The last field gives the quality tag associated with this CleanEx_trg entry. This allows a direct evaluation of the experiment results for that CleanEx entry.

The // (terminator) line designates the end of an entry.

6.2.2. *CleanEx_exp*

The CleanEx_exp files contain two differently formatted parts : the documentation entry, and the expression entries.

6.2.2.1. Documentation entry

The DATASET_DOC entry itself contains two kinds of information. The first one is the content description of the experiments performed by the authors. The second consists in precise and highly-formatted values and tags about the expression measurements of the dataset which are then extracted and used by the CleanEx web interfaces. Its general format is as follows :

The first line is the documentation entry identifier. It begins with a ">", and is followed by the dataset's code and the extension "_DOC". The documentation entry for a dual channel experiment is shown in Figure 15.

The ID line contains the code name of the dataset, followed by the number of experiments and the total features per experiment.

The OS line stands for the organism

The TI line is a short description of the dataset's contents

The DE lines are a detailed description of this dataset

The FM lines stand for ForMat. They are used by the web interfaces. They differ according to the dataset's type, as described below :

- Minimum and maximum log ratio, as well as individual (red and green) channel scales for dual channel chips. The same information is provided for Basic_Ratio datasets.
- Minimum and maximum log ratio for Counts, that is for SAGE data or EST counts.
- Minimum and maximum log intensities for Affymetrix data.

The RN (Reference Number), RX (Electronic reference), MA (Authors), RT (Title) and RL (Journal reference) lines describe the paper published about this dataset.

The HP line give the main URL provided by the authors. Usually this leads to a local search page for the dataset.

FM lines are also used by the web interfaces. The first FM line contains a code for the data type. So far, the data type codes are : Stanford_Scanalyze for Stanford-like dual channel chips read with ScanAlyze, Intensity for cases where we could only obtain the raw intensity for each spot, Affy_probeset for Affymetrix data, Basic_Ratio for the oligonucleotides datasets, and Counts for the SAGE datasets. The second FM line gives indications on the threshold that should be used to flag spots considered as unreliable.

The FD line describes the target types of the spotted features, as well as the identifier used in CleanEx_trg for the mapping on genes. Target types are :

- Probeset for Affymetrix experiments
- EST, IMAGE, for clone-based experiments
- X-mer oligonucleotides for oligo arrays, where X represents the length of the spotted oligos.

Usual reference identifiers are RefSeq (code "RefSeq") or EMBL accession number (code "AC").

The EX lines are the experiment description lines. Each of these represents one experiment. Each field on the line is separated by a semi-colon. The first field gives the experiment number, which is reproduced later on in each expression entry. The second field is the short chip name usually given by

the authors. It is completed by a short experiment description in the third field.

If possible, the experiments are ordered in a way so as to reproduce the cluster or the associations found in the paper figures.

The documentation entry ends with a “//” termination line.

```
>P0001.DOC
ID PEROU1; 26 Experiments; 5531 sequences;
OS Homo Sapiens (human).
TI Gene expression in human mammary epithelial cells and breast cancers
DE Distinctive gene expression patterns in human mammary
DE epithelial cells and breast cancers.
PA Minimum log ratio : -7.37301408828019;
PA Maximum log ratio : 6.1746831304838;
PA Red channel log scale : from 5.51690215908742 to 15.629384712127;
PA Green channel log scale : from 6.15723948123096 to 15.5893740233065;
RN [1]
RX MEDLINE=10430922; PubMed=99362737;
MA Perou C.M., Jeffrey S.S., van de Rijn M., Rees C.A., Eisen M.B.,
MA Ross D.T., Pergamenschikov A., Williams C.F., Zhu S.X., Lee J.C.,
MA Lashkari D., Shalon D., Brown P.O., Botstein D.;
RT "Distinctive gene expression patterns in human mammary epithelial
RT cells and breast cancers.";
RL Proc. Natl. Acad. Sci. U. S. A. 96:9212-9217(1999).
HP "http://genome-www.stanford.edu/sbcnp/";
FM Stanford_Scanalyze.
FM Quality="FLAG"; Comparison=""; Threshold="1";
FD CloneRef=IMAGE; SeqRef=AC;
EX 1; "0164-5k"; "HMEC X HB2";
EX 2; "0376-5k"; "HMEC X MCF7";
EX 3; "0174-5k"; "HMEC X Hs578t";
EX 4; "0134-5k"; "HMEC X BC24";
EX 5; "0132-5k"; "HMEC X BC790";
EX 6; "0166-5k"; "HMEC X BC1257";
EX 7; "0167-5k"; "HMEC X BC2";
EX 8; "0129-5k"; "HMEC X BC17";
EX 9; "0095-5k"; "HMEC X BC16";
EX 10; "0097-5k"; "HMEC X BC1369";
EX 11; "0090-5k"; "HMEC X BC23";
EX 12; "0087-5k"; "HMEC X BC1498";
EX 13; "0096-5k"; "HMEC X BC14";
EX 14; "0111-5k"; "HMEC X BC5";
EX 15; "0088-5k"; "HMEC X BC4";
EX 16; "0130-5k"; "HMEC X BC21";
EX 17; "0165-5k"; "HMEC X Normal Breast Pool 1";
EX 18; "0374-5k"; "HMEC X Normal Breast Pool 2";
EX 19; "0119-5k"; "HMEC X HMEC plus TGF-beta 24 hours";
EX 20; "0106-5k"; "HMEC Minus EGF X HMEC Plus EGF 90 Minutes";
EX 21; "0386-5k"; "HMEC X HMEC grown on Matrigel 24 hours";
EX 22; "0387-5k"; "HMEC X Senescent HMEC";
EX 23; "0083-5k"; "HMEC X HMEC Minus EGF 48 hours";
EX 24; "0175-5k"; "HMEC X Confluent HMEC";
EX 25; "0183-5k"; "HMEC X HMEC Plus Interferon-alpha 24 hours";
EX 26; "0182-5k"; "HMEC X HMEC Plus Interferon-gamma 24 hours";
//
```

Figure 15: CleanEx_exp documentation entry for a dual-channel dataset

6.2.2.2. Expression data entries

The entries format varies according to the type of dataset, as described earlier. The only line which is standardized throughout all the different experiment types is the header line of each “feature”, or sequence for which an expression measure has been done. This line always begins with a “>”, and is

then followed by four different fields. The first field is the “feature” identifier, which is built by concatenating the dataset's identifier and the feature's identifier itself. The second field indicates the type of the corresponding target for this feature. The third field is the target identifier, and the last field gives the original name, given by the authors, of that feature, if provided. Figure 16 shows examples of headers for different data types.

```
>P0001_547224 Type=cDNA_clone; TRG=IMAGE_547224; Name="SID fibronectin 1";
>GSE10_TGGTTGCTGG Type=SAGE_Tag; TRG=SAGE_NlaIII_TGGTTGCTGG; Name="SLK";
>GDS181_1001_at Type=Probeset; TRG=AFY_HG-U95Av2_1001_at; Name="X60957";
>HSEST_FN1 Type=Seq_Ref; TRG=NM_212482; Name="Fibronectin 1";
```

Figure 16: Examples of CleanEx_exp headers, for respectively dual-channel, SAGE, Affymetrix and EST count datasets

Following the header line, each entry contains the former header provided in the raw data files, as for example the ScanAlyze output header. This will be the guide for further specific expression measurements extraction. If no header is provided for the dataset, one creates such a line with the different measurement fields indicated. All the other lines of one entry correspond to the measurements results for that spot and for each experiment.

6.2.3. CleanEx_trg

Each CleanEx_trg entry corresponds to one "target" (or "expression feature") used in an expression measurement experiment. Identifiers are composed of a code which describes the target type followed by an underscore and the target accession number. Types could be, for example, IMAGE clone (IMAGE), Affymetrix probe set (AFFY), SAGE tags (SAGE), or EMBL RNA or DNA sequences (RNA,DNA).

The format of CleanEx_trg resembles that of CleanEx. Each CleanEx_trg entry contains the following information :

- * ID CleanEx_trg ID
- * OS Organism Species

- * GC Gene Count
- * GN Official Gene Symbol from the organism catalog
- * OA Original Annotation (if existing)
- * QU Quality tag
- * SR Sequence Reference
- * FN Feature Number
- * UG UniGene release
- * F[1-25] Feature
- * DR CleanEx_exp ID
- * //

Description of the line formats :

The ID line

ID TRG_ID Type

The identification line is always the first line of an entry.

TRG_ID is the internal identifier for the entry. The first part of the ID is a target type identifier. The second part is built with the original target name (image clone identifier, Affymetrix chip and probeset name,...)

The Type field is a description of the target's provenance. Type could be for example "Seq_Ref" (for a sequence in EMBL or in RefSeq), "cDNA_clone", "Affy_Tag", "SAGE_Tag" or "MPSS_Tag".

The OA line

This line contains either the target's Original Annotation found in the corresponding description files, for example the Affymetrix chips annotation, or the description of the sequence given in the corresponding EMBL entry. It exists only for CleanEx_trg entries corresponding to Affymetrix tags.

The GN line

GN TIE

The GN line lists the official gene symbols which correspond to that entry. For the Affymetrix entries type, if more than four genes match the target, only the first four are listed. The total number of matched genes is mentioned in parentheses.

The GC line

GC 1

The GC line gives the total count of genes having an approved symbol which match that target entry

The QU line

QU High

The QU line is the quality tag based on the precision of the mapping of the target. As explained in the part describing the construction of the target files, this tag can take different values, according to the corresponding entry type or to the mapping protocol.

The SR line

SR Unigene=Hs.21330;

The SR line stands for Sequence Reference and gives the associated Unigene Cluster for the whole target.

The FM line

FM Tag;

This line describes the format of the features for the target.

The FN line

FN 16

The FN line gives the number of features belonging to that target. For cDNA clones, this number is typically one. For Affymetrix probe sets, it can vary between eleven to twenty-five.

The UG line

UG UniGene Build #160

The UG line shows the Unigene Release which has been used to map the target sequences to its corresponding cluster.

The F1-F25 lines

F1 TGTCCAGGCTGGAACAAAGCGCCAG:283-105; Refseq=NM_000927(+);

These lines show the individual mapping for all the features of the corresponding target. Fields are

separated by a ";". The first field is the name of the feature. The second field contains the RefSeq or EMBL accession number of the sequences which map the feature. If the F line corresponds to a tag and has been mapped via the two-steps procedure, the second field contains more detailed information. The sign in parenthesis indicates if the tag mapped to the positive or to the negative strand of the corresponding sequence. The numbers in square brackets show the exact position of the tag on this sequence, and the last number after the square brackets indicate the total length of the sequence on which the tag has been mapped. If more than one sequence had a match for this tag, the sequences are listed in that same format, and are separated with a "|". For Affymetrix, we write down up to four sequences, then the total number of sequences with a match is indicated in parentheses at the end of the line.

The DR line

```
DR AFFY001_1575_at;
```

DR lines in CleanEx_trg are cross-links to the expression data found in CleanEx under the line type "EXP". The link is done via the expression data local identifier found in the CleanEx_exp files.

In Figure 17, target entry examples for Affymetrix, SAGE and clones are shown.

```

ID AFFY_HC-G110_1000_at Type=Affy_Tag
OA X60188; Human ERK1 mRNA for protein serine/threonine kinase
OS Homo sapiens.
GN MAPK3
GC 1
QU High
SR RefSeq=NM_002746 Unigene=Hs.861;
SN 1
FM Tag;
FN 16
UG UniGene Build #183
F1 TCTCCTTGGCTGAGGCCCTCCAGCTT:137-179; RefSeq=NM_002746(+) [1355..1379]1866;
F2 AGGCCTCCAGCTTCAGGCAGGCCAA:138-179; RefSeq=NM_002746(+) [1367..1391]1866;
F3 CCAGCTTCAGGCAGGCCAAGGCCCT:139-179; RefSeq=NM_002746(+) [1373..1397]1866;
F4 AGCTCAGTGGGCCCCAGTTCAATCT:140-179; RefSeq=NM_002746(+) [1433..1457]1866;
F5 AGTTCTGGAAATGGAAAGGGTTCTGGC:141-179; RefSeq=NM_002746(+) [1511..1535]1866;
F6 TAGGGACTCAGGGCCATGCCTGCCC:142-179; RefSeq=NM_002746(+) [1583..1607]1866;
F7 TTCCTGAAAGGAACATTCCTTAGTC:143-179; RefSeq=NM_002746(+) [1637..1661]1866;
F8 GAAGGAACATTCCTTAGTCTCAAAG:144-179; RefSeq=NM_002746(+) [1643..1667]1866;
F9 CTTAGTCTCAAAGGCCCTAGCATCCCT:145-179; RefSeq=NM_002746(+) [1655..1679]1866;
F10 CTCAAGGGCTAGCATCCCTGAGGAG:146-179; RefSeq=NM_002746(+) [1661..1685]1866;
F11 GGCTAGCATCCCTGAGGAGCCAGGC:147-179; RefSeq=NM_002746(+) [1667..1691]1866;
F12 CTGTCAAAGCTGTCACTTGGCGTGC:148-179; RefSeq=NM_002746(+) [1709..1733]1866;
F13 AAGCTGTCACTTGGCGTGGCCTCGC:149-179; RefSeq=NM_002746(+) [1715..1739]1866;
F14 CGCCTGGCCCTCGCTGCTTCTGTGTG:150-179; RefSeq=NM_002746(+) [1727..1751]1866;
F15 CCCCTGGTGTCTTGTGTGTGGTGA:151-179; RefSeq=NM_002746(+) [1733..1757]1866;
F16 CTGCTTCTGTGTGTGTGGTGGCAGAA:152-179; RefSeq=NM_002746(+) [1739..1763]1866;
DR AFFY001_1000_at;
//
ID SAGE_NlaIII_AAAAAAAAAAC Type=SAGE_Tag
OS Homo sapiens (human).
GN SLC25A3|KRT18
GC 2
QU Medium
SR Sequence=BE618231 Unigene=Hs.290404;
SR Sequence=BQ712279 Unigene=Hs.406013;
SN 2
FM Tag;
FN 1
UG UniGene Build #182
F1 AAAAAAAAAAC; Sequence=BE618231(+) [333..342]379;BQ712279(+) [728..737]932;
DR GSE10_AAAAAAAAAAC;
DR GSE17_AAAAAAAAAAC;
DR GSE31_AAAAAAAAAAC;
DR GSE41_AAAAAAAAAAC;
DR GSE506_AAAAAAAAAAC;
DR GSE507_AAAAAAAAAAC;
DR GSE508_AAAAAAAAAAC;
DR GSE514_AAAAAAAAAAC;
DR GSE545_AAAAAAAAAAC;
DR GSE608_AAAAAAAAAAC;
DR GSE953_AAAAAAAAAAC;
//
ID IMAGE_1000208 Type=cDNA_clone
OS Homo sapiens (human).
GN CDKL3
GC 1
QU High
SR RefSeq=NM_016508 Unigene=Hs.105818;
SN 1
FM cDNA_clone;
FN 1
UG UniGene Build #183.
F1 IMAGE_1000208; EMBL=AA533975;
DR T0001_51;
//

```

Figure 17: Examples of CleanEx_trg entries for Affymetrix, SAGE, and IMAGE clone data types

6.2.4. Additional format : XML version of CleanEx for Integr8

Integr8 (<http://www.ebi.ac.uk/integr8/>) [50] is a European project which aims to develop an integrative layer in database services to facilitate the synthesis of related information. Integr8 will be an automatically populated database which will :

- Maintain stable identifiers for biological entities

- Describe their relationships with each other
- Store equivalences between identified entities in the source databases

Typically, the goal of Integr8 is not to mirror all the European databases, but more to provide a stable link between these. For that reason, only core data (database unique identifiers, links to Unigene, Swissprot, Ensembl, RefSeq) are stored. The full entry information is retrieved via web links to the source database.

To integrate CleanEx in this European project, we created a minimized XML version of CleanEx, which contains only the core data. This file is also available via our ftp server.

6.2.5. Specific formats for web applications

As a way to increase the speed of online expression data retrieval and analysis, the CleanEx system also includes a few specific internal formats which are generated at the same time as the original three file types. For example, the cross dataset tool makes use of a matrix type file containing expression measurements by experiments, and not by gene. By reformatting the expression files in an “experiment-centered” way, the retrieval speed is nearly five-fold faster. Indeed, in this new very specific file, only the expression measurements values are kept. One line is created for each experiment, which contains the values for all the spots on that chip. The first field of that line is the experiment identifier, created by concatenating the dataset's code and the experiment number found in the documentation file of CleanEx_exp. This file is then indexed via this first field.

Another analysis specific file which is provided for the web interface is the so-called “classes for experiments” file. In this file, the first field is the same as in the special file described above, and allows experiment-centered data retrieval. Each field corresponds to a description of the different classes to which this experiment belongs. By class, we mean for example : organism, cell type, tissue, disease, treatment, and so on. All the main class divisions found in CleanEx are listed in a separate file, and a number is attributed to each of these. Then, in the classes per experiment file, a Boolean number is attributed to each class number for each experiment line, indicating if this specific experiment belongs to this class or not. For example, the class “human” has the class number 1. If the experiment has been

done with human material, the human field for this line will be “1:1”. If not, the human field will then be “1:0”. This system allows a very fast parsing mechanism and specific expression values retrieval for further analysis of experiments belonging to the same classes, but not to the same dataset. It is the basis for a real cross-dataset comparison system.

The last web specific file provided is a direct link between the datasets features (tags, probe sets, or spotted clones, for example) and their corresponding gene symbols. Let's call this file `Dataset_to_gene`. The format resembles the two others in the sense that each line contains the relative information for all the so-called features of one dataset. The first field is the dataset's code. It is followed by the gene symbol for each of the chip's feature. There is one line per dataset already included in CleanEx. The position of the gene symbol on the line corresponds to the position of the corresponding feature in the dataset.

6.3. Indexes and retrieval system

All the CleanEx files are indexed and retrieved by the fetch system, which allows fast and easy one-by-one entry retrieval, given a specific identifier. The fetch system is an in-house utility which is used only on our site and which works on the basis of an index file which contains three features per line. The first feature is the accession key for the entry. It is typically the entry unique identifier. The second feature is the start position of the corresponding entry, and the third one is the length of the entry.

The CleanEx entries can be retrieved via : CleanEx identifier, EXP line (the corresponding experiments), gene symbols, and Genew accession number. The CleanEx_exp entries are retrieved via the CleanEx_exp identifier. The targets can be retrieved via CleanEx_trg identifier, Unigene cluster, reference sequence accession number, gene symbols, or corresponding CleanEx_exp entries.

Fetch also allows multiple entry retrieval. This could be really useful, for example, to retrieve all the target entries corresponding to the same gene.

The fetch system is also used for all the data retrieval which occur via the web-based interfaces.

6.4. Web-based interfaces

6.4.1. Entry search engines and viewers

6.4.1.1. Single entry search engines

The CleanEx and CleanEx_trg files are accessible either as flat files on the ftp server (<ftp://ftp.isrec.isb-sib.ch/pub/databases/CleanEx/>), or via a web-based entry search and retrieval system at : <http://www.cleanex.isb-sib.ch>.

There are two ways to extract single CleanEx entries via the web interfaces. The first one implies that one already knows the entry identifier. For the CleanEx part, this is made much easier by the fact that CleanEx identifiers are built with the organism abbreviation followed by the official gene symbol. Though having a little bit more complex identifiers, target data can also be retrieved via the quick search interface (Figure 18). A detailed explanation of the target identifier format is provided on the same page. The advantage of this query is speed, as there is no need to search in the whole file. The fetch system will retrieve the queried entry at once.

The second search method is used when one does not know the exact entry identifier. This query form can be filled with information as diverse as gene name, description, Unigene accession number, organism, RefSeq sequence, Swissprot or EPD identifiers, or even the clone accession numbers, or the expression experiment's identifiers. As the search is done on the whole file, this takes a bit longer than the quick search system. It works as follows :

From the selected fields in the query page, it extracts the lines to search in the corresponding database. This process is facilitated by the fact that each line type begins with a specific two or three letter code. Then the different words and conditions given for search are transformed in a perl regular expression. The program then reads the CleanEx file entry by entry and tries to match this expression to the entries. For every entry which corresponds to the given criteria, the CleanEx identifier is stored in an array. Once the whole file has been read, the entries selected are shown, and one can then select the data to retrieve.

Again, this search engine works for CleanEx as well as for CleanEx_trg data.



CleanEx Expression Reference Database



CleanEx is a database which provides access to public gene expression data via unique approved gene symbols and which represents heterogeneous expression data produced by different technologies in a way that facilitates joint analysis and cross-dataset comparisons. [[More details](#) / [Survey of most recent release](#)].

Current release is based on Unigene database available on Mar-14-2004.

Access to CleanEx

<input type="text" value="hs_add3"/> <input type="button" value="Quick Search"/> in CleanEx database by AC, ID or documentation text <input type="text"/> <input type="button" value="Quick Search"/> in CleanEx Target database by AC, ID or documentation text	<ul style="list-style-type: none"> • Browse CleanEx database • Browse CleanEx Target database • Make a Batch Search on the CleanEx Target database • View list of datasets in CleanEx • Expression query form for AFFY001 • Expression query form for AFFY002 • Expression query form for P0001 • Expression query form for C0001 • Expression query form for R0001 • Expression query form for L0001 • Expression query form for S0001 • Expression query form for NCI60 • Expression query form for HSEST • Download CleanEx files by FTP
---	---

Figure 18: CleanEx quick search page

6.4.1.2. CleanEx viewer

All the previously cited external databases which are cross-linked in the CleanEx entries are accessible via the specific database identifier. Besides these cross-references, the CleanEx entry also provides access to :

1- The list of all potentially spotted features, meaning for example, clones, RNAs, probe sets or tags, which correspond to this gene. As for the creation of the target file, sequences like clone sequences, or RNAs are directly extracted from the corresponding Unigene cluster(s). The tags and probe sets information come from the CleanEx_trg file. The final list includes the gene symbol, the Unigene

cluster, the sequence accession number (if it exists) the target identifier (if it exists), and the type of the sequence. This sequence type corresponds to the one which is attributed to the target entries in CleanEx.

2- The possibility to extract the sequence region around the determined transcription start site (TSS) for further promoter sequence analysis. The 5' and 3' distance from the TSS can be chosen by the user. The output of this query is a FASTA formatted text file. The FASTA header contains different fields separated by a "|". The first field is the gene symbol. It is followed by the genomic contig accession number. The third field shows the genomic region which has been extracted, and the last one is a tag to determine the origin of the information (either EPD or ANNOTATION, as mentioned in the CleanEx construction paragraph).

3- The most important part of a CleanEx entry is, of course, the link between the gene and the heterogeneous expression data. The list of all the datasets which have a target corresponding to this gene entry is provided in the "Expression Data References" sub-table. Data can then be explored in very different ways :

The first field, called Dataset's name in this sub-table, links to the local CleanEx_exp documentation file for the corresponding dataset. The Target ID column is the direct link to the CleanEx_trg entry. The last part, named "Expression Data", links to the CleanEx_exp entry. There, access is given to the text entry, or to a local expression viewer which will be explained in the following part. Note that if the dataset possesses more than one feature for this particular gene, this local expression viewer is called to visualize all the features at the same time using the button "View all dataset experiment".

6.4.1.3. CleanEx_Exp : expression viewer

The local expression viewer is a color representation of the gene's expression across the different experiments of one dataset. The first part of the viewer remains the same across all the data types. It describes the origin of the data, and gives a direct link to the target entry. A short description of each experiments in this dataset is also provided. The second part of the viewer provides the color-based display of the expression data. According to the dataset type, the color display can vary in the format as well as in the color scale. This is also a way to distinguish between the different origins of the datasets included. The different color codes are detailed in the next paragraph. Figure 19 shows the viewers for

different experiment types.

1. Counts type datasets

For all the datasets based on counts, namely the EST counts, SAGE or MPSS datasets, the local viewer is based on the estimated TPM (Tags Per Million). For each experiment, the number of tags for the represented gene is divided by the total number of tags for that experiment. This ratio is then converted into the TPM value. To lower the impact of cases where no tags or ESTs are found, pseudo-counts are added to both values before obtaining the ratio. The fraction is represented by a scale from white (low expression level) to black (high expression level).

2. Dual channel (Stanford_like) and Basic_Ratio datasets

If the original dataset contained enough information, the viewer shows two different representations of the data. In the first column, the color represents the log in base two of the ratio between the two channels (green and red). The color display goes from light green (underexpressed) to light red (overexpressed). This is the traditional expression representation given by programs used to analyze dual-channel chip (for example Michael Eisen's Treeview software). The second column displays the superposition of both channels. This typically gives an idea of the intensity level of the spot, and corresponds to the reconstructed image of the chip with both scanned values shown together. When the original data provides only the final ratio, this second column is omitted. The color scale is built according to the ratio and channels extreme values stored in the documentation file of the corresponding dataset under the PM (ParaMeter) lines.

The ratio color range goes from 1 to 256 and the value is scaled according to the following formula :

$((\log_{\text{ratio}} - \log_{\text{min}}) / (\log_{\text{max}} - \log_{\text{min}} / 256))$, where \log_{ratio} is the log in base two of the ratio, \log_{min} is the minimum log ratio found in the dataset and \log_{max} is its maximum. The color which is displayed is the selected according to the common representation of over-expression (red) and under-expression (green).

For the display showing the sum of both channels, the color is obtained by superposing the intensity of

both channels. The color value for each channel is scaled in the same way as for the color ratio value :

$((\log_{\text{channel}} - \log_{\text{minchannel}}) / (\log_{\text{maxchannel}} - \log_{\text{minchannel}} / 256))$ where “logchannel” is the log in base two of the channel intensity, logminchannel is the minimum value of the log in base two of the channel intensity, and logmaxchannel is its maximum.

For some Basic_ratio datasets, the authors also provide a P-value, indicating if the spot is reliable or not. They give the usual threshold defined for that dataset as being the highest acceptable p-value for a spot. As this information is also stored in the CleanEx_ref documentation file, the viewer considers spots with a bigger p-value as flagged, and shows them with a grey color. The p-value is also shown. The same grey flag is applied for Stanford data, according to the “FLAG” tag given for each spot in the raw expression file.

3. Affymetrix datasets

The colors chosen to represent Affymetrix datasets varies a bit compared to the usual expression displays. From the expression entry, it makes use of the LOG_NORM value, in other words the log in base two of the intensity for that probe set, but with the mean centered over the experiments. Values below zero are shown in a blue scale, and values greater than zero are shown in a pink scale. Darker colors in both scales indicate the most under- or overexpressed cases. Again, the colors are scaled according to the maximum and minimum log values stored in the documentation file, as described earlier. If provided, each color spot contains also the Absent/Present call generated by the analysis software. This replaces the flag defined for dual-channel datasets.

For all these data types, the multiple expression viewer which is accessible via the CleanEx entry page is based on the same criteria. For space reasons, though, this view shows only the ratio column, and not the sum column. This gives a more compact and readable view of the different features. Having all the features corresponding to one single gene on the same view is a good way to have a first fast control on the internal chip reproducibility. To some extent, it could give also a first clue on the differential expression pattern of transcript variants along the different experiments realized in the dataset. An example will be shown in the CleanEx tutorial.

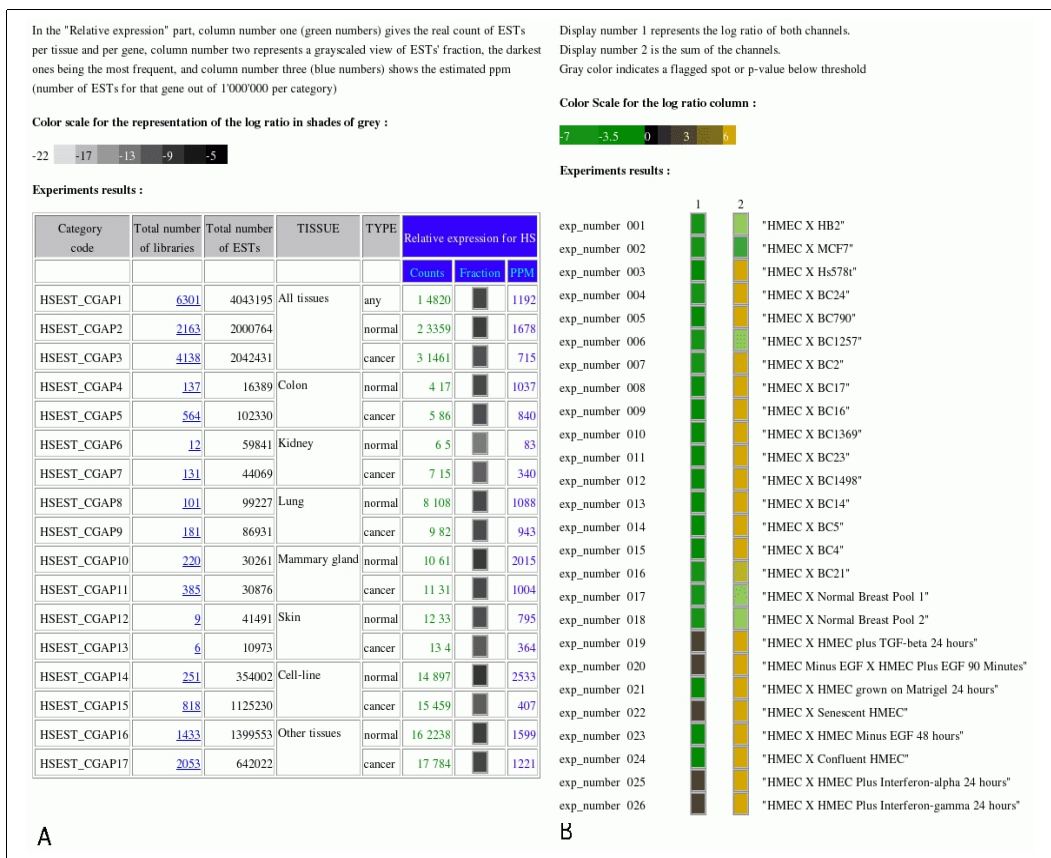


Figure 19: Two examples of the expression single viewer. A : EST count representation. B : Dual-channel representation

6.4.1.4. CleanEx_trg


6.4.1.4.1. Single entry retrieval

The CleanEx target entries can be retrieved individually with the same search engine that retrieves CleanEx entries. The CleanEx_trg entry viewer resembles the CleanEx one. It also gives access to the original cross-references URLs via sequence accession number, and it also provides a direct link to the expression data targeted by the corresponding entry. In the case of a target which has been determined via the two-steps procedure, meaning a target for which the sequence was known and which has been mapped to RNAs via the tagger program, the exact position of all the tags on the reference sequence is provided. This information could be useful in two ways. First of all, for SAGE tags, one can verify the distance between the site where the restriction enzyme has cut the sequence and the 3' end of the gene sequence. For Affymetrix, this position could be even more useful. According to the constraints which limit the number of choices regarding the 25 nucleotides tags choice per probe set, it is sometimes

impossible to have eleven tags which do not overlap. Knowing the position of each individual tag allows to know the real portion of the transcript which is spanned by the probe set. Moreover, if one gene is represented by more than one probe set, it might be worth checking, if the different probe sets span the same region of the gene, or for example if they could represent two different transcript variants.

6.4.1.4.2. Batch search for CleanEx_trg (http://www.cleanex.isb-sib.ch/trg_batch_search.html)

Though the single data retrieval can be quite useful if you just need information on one specific target, as for example one Affymetrix probe set, researchers are more interested in the correspondence between the features coming from different platforms. Moreover, they often want to have access to a pool of features and their corresponding genes, instead of retrieving this information one gene at a time. This practical problem prompted us to generate a so-called “batch search” service for the target files (Figure 20). The system will retrieve all the CleanEx Target entries corresponding to the given input identifier list and the given organism. One can obtain targets correspondence via a choice of different identifiers, like Unigene, RefSeq, EMBL accession numbers, as well as gene symbol or even the CleanEx target identifier. A link to a detailed description of the possible input formats is given on the query page. One can also select the organism for which one wants to retrieve targets. A combination of sequences coming from the two organisms presently in CleanEx is also possible. The data output is then classified according to the user's choice, either by gene symbol, Unigene cluster, sequence identifier, or targets identifier. One can also select the type of features to be kept in the result page, for example if one just wants to compare different Affymetrix chips and is not interested in ESTs or other RNAs. The result can be retrieved in HTML or in text format. The HTML format has the advantage of providing direct links to the other databases, and to give a nice human-readable view. It consists of a table which contains one type of information per column. The table header gives the column content type as well as the total number of features found in the database. The information provided is : CleanEx target, Gene symbol, Refseq, Unigene, Sequence accession numbers, and experiments found in CleanEs. The text format is an easy-to-parse file, which contains well-separated entries having one type of information per line (gene, Unigene, RefSeq, target, experiment found in CleanEx), with space-separated features on each line.



CleanEx Target batch search

Online target retrieval for human and mouse expression experiments

Use : The system will retrieve all the CleanEx Target entries corresponding to the given input identifier list and the given organism. Possible queries include target-to-target retrieval as well as gene-to-target, RefSeq-to-target or Unigene-to-target retrieval.
A detailed description of the CleanEx database can be found [here](#).

Input settings

Select Type of Sequence Identifier in your Input List
See [HERE](#) for the input format.

Gene symbols

Paste your identifier list in the box below

Or upload your identifier list from a local file

Browse...

Select Organism

Human

Mouse

Output settings

Select Key identifier for output classification

Gene symbols

Select entry type in CleanEx Target

Any targets

DNA sequences

RNA sequences

RefSeq sequences

ESTs

Affymetrix probesets from HG-U133A chip

Affymetrix probesets from HG-U133B chip

Affymetrix probesets from HG-U133_Plus_2 chip

Affymetrix probesets from HG-U95A chip

Affymetrix probesets from HG-U95Av2 chip

Affymetrix probesets from HG-U95B chip

Affymetrix probesets from HG-U95C chip

Affymetrix probesets from HG-U95D chip

Affymetrix probesets from HG-U95E chip

Affymetrix probesets from HuGeneFL chip

Affymetrix probesets from HC-G110 chip

SAGE tags

IMAGE clones

INCYTE clones

ResGen clones

Retrieve Output in :

HTML
 TEXT

Figure 20: The CleanEx_trg batch search web interface

6.4.2. Cross dataset analysis

In the beginning, the possibility of analyzing data coming from different platforms was the first objective for CleanEx. This task is not that easy, due to major differences between the data sources. The most critical difference is that some data are the result of a comparison between two experiments, and give an expression value which is a ratio. These are typically the dual-channel chips. Other ones, namely Affymetrix, EST counts, SAGE and MPSS are done with only one experiment, and hence yield the relative abundance of transcripts in one sample. To bypass this problem, the first cross-dataset comparison system which has been created is a step-by-step procedure, which treats one dataset at a time. Later on, a second version has been implemented, which is able to directly compare chips coming from different sources. The two models are explained below.

6.4.2.1. Step-by-step expression pattern search (http://www.cleanex.isb-sib.ch/step_by_step_analysis.html)

This procedure allows the search results in one dataset to be combined with a new search step in another

dataset. The principle is as follows. One first selects a first dataset where one wants to extract genes having a common behavior. In this dataset, a selection of two experiments pools is done. One can select the experiments to put in each pool, and the analysis which has to occur between these two pools. Comparison can be over-expression in either the first or the second pool compared to the other one, or co-expression levels in the two pools. Once ready, one can choose the number or percentage of genes to keep. Currently, the comparison is based on a general mean difference ranking process. The mean expression is calculated for each gene and for each experiment pool. Then the difference between the first pool mean and the second pool mean is calculated, again for each gene. The result is then ranked, and it's this rank which is taken into account to display the genes. The results page shows the two groups to compare, and the list of features which satisfy the given criteria. Below the features list, a table gives the number of common genes, amongst the retrieved features, in other datasets. To go to the next level, one just selects the dataset in which one wants the new comparison to occur. The next page then provides the dataset-specific page, as the first one, and experiment selection can be done again. The main difference is that for the second step, the analysis will be done only on the features corresponding to the common genes' list, and not on the whole new datasets. The output is the list of features which share the given criteria in the two datasets explored. A practical example will be developed in the CleanEx tutorial part.

6.4.2.2. Common genes retrieval (http://www.cleanex.isb-sib.ch/compare_dataset_genes.html)

One way for researchers to make sense of their data is often to compare the results they obtained with previously published corresponding experiments. The problem is that datasets to compare are often issued from different techniques and platforms. In that case, knowing to which gene corresponds each feature in the two datasets to compare is the first step. To facilitate such an analysis, CleanEx provides a direct common genes retrieval system, which works on all the datasets already integrated in the database. The principle is to make use of the special formats cited before, meaning the Dataset_to_gene file, which contains the gene symbol for each feature of one dataset on one line. Once the datasets to compare are selected, the system extracts the corresponding lines in this file, as well as the line containing the list of the datasets' features. The features and their corresponding genes are indexed, then the genes in both datasets are compared, and common genes extracted in both datasets. The gene symbol position on the line is then traced back and gives access to the original features on the two

compared chips.

6.4.2.3. By class expression pattern search

This feature is quite recent on CleanEx. It has been possible to set up such a process only since the database contains a sufficient number of data to analyze. The main goal of this method is to be able to compare two chip pools coming from different datasets which have been generated by very different methods. By using data from different sources, one will be able to generate much wider comparisons between different classes. Indeed, published datasets are often very specific and relate to a single question, like gene expression in normal tissues, or the effect of a drug on one tissue type, or cancer classification in one tissue. If one wants to compare gene expression between normal and cancer tissues, for example, it will be of great use to be able to use data from more than one dataset, as it will increase the number of experiments as well as the range of possible comparisons. This will lead to the discovery of discriminant genes between classes, or even to a more accurate class prediction basis.

To facilitate the use of such a tool, the way that has been chosen is to compare two class pools, instead of two chips pools. The generation of the file which maps experiments on the different classes has been described before. The process begins with the selection of the two classes to compare (Figure 21). One could for example select in the first pool all experiments done with normal mammary gland tissue, and in the second pool all experiments done with tumor mammary gland biopsies. The program, will then, via the “classes for experiment” file, generate two lists of experiments which correspond to the asked conditions. At the next level, these experiments are shown to the user accompanied with a brief description, so that one can reselect the desired experiments. Once this is done, the real comparison takes place. Of course, to be able to compare expression values in different datasets, the first step is to find common features in all the datasets included in the search. This part is controlled by the former described Common Genes Retrieval system. Once the common genes have been extracted, the correspondent expression values for all the chosen experiments are extracted. The method and fields used as raw values are so far the same than for the ExpressDB ERAs (Estimated Relative Abundances). According to the data type, the value extracted is : red channel background subtracted for dual-channel type experiments, intensity for Affymetrix data, and counts for Counts type data. The normalization is then done as in ExpressDB. The analysis can be performed using the mean difference ranking already

used for the step-by-step analysis procedure.

Class selector module

Select the characteristics of the chips that you want to compare. Comparison will occur between chips in first set and chips in second set.

First set	Second set
<ul style="list-style-type: none"> ● <input type="checkbox"/> Tissues/cells origin <ul style="list-style-type: none"> ● <input type="checkbox"/> cell line ● <input type="checkbox"/> stem cells ● <input type="checkbox"/> mammary gland ● <input type="checkbox"/> ovary ● <input type="checkbox"/> uterus ● <input type="checkbox"/> fetal annexes ● <input type="checkbox"/> embryo ● <input type="checkbox"/> testis ● <input type="checkbox"/> prostate ● <input type="checkbox"/> bladder ● <input type="checkbox"/> kidney ● <input type="checkbox"/> spleen ● <input type="checkbox"/> pancreas ● <input type="checkbox"/> brown fat ● <input type="checkbox"/> adipose tissue ● <input type="checkbox"/> liver ● <input type="checkbox"/> gall bladder ● <input type="checkbox"/> intestine ● <input type="checkbox"/> stomach ● <input type="checkbox"/> brain ● Cells treatment/state information <ul style="list-style-type: none"> ● <input type="checkbox"/> normal/untreated/validation ● <input type="checkbox"/> treated ● <input type="checkbox"/> irradiated ● <input type="checkbox"/> infected (virus) ● <input type="checkbox"/> infected (other) ● <input type="checkbox"/> tumour ● <input type="checkbox"/> primary ● <input type="checkbox"/> low-grade ● <input type="checkbox"/> high-grade ● <input type="checkbox"/> secondary ● <input type="checkbox"/> metastatic ● <input type="checkbox"/> non-metastatic ● <input type="checkbox"/> recurrent ● <input type="checkbox"/> adenoma ● <input type="checkbox"/> carcinoma ● <input type="checkbox"/> hormone sensitive ● <input type="checkbox"/> hormone resistant ● <input type="checkbox"/> surgery ● Experiment type <ul style="list-style-type: none"> ● <input type="checkbox"/> time-course ● <input type="checkbox"/> classification ● <input type="checkbox"/> survival ● <input type="checkbox"/> survival>5 yr ● <input type="checkbox"/> survival<5 yr 	<ul style="list-style-type: none"> ● <input type="checkbox"/> Tissues/cells origin <ul style="list-style-type: none"> ● <input type="checkbox"/> cell line ● <input type="checkbox"/> stem cells ● <input type="checkbox"/> mammary gland ● <input type="checkbox"/> ovary ● <input type="checkbox"/> uterus ● <input type="checkbox"/> fetal annexes ● <input type="checkbox"/> embryo ● <input type="checkbox"/> testis ● <input type="checkbox"/> prostate ● <input type="checkbox"/> bladder ● <input type="checkbox"/> kidney ● <input type="checkbox"/> spleen ● <input type="checkbox"/> pancreas ● <input type="checkbox"/> brown fat ● <input type="checkbox"/> adipose tissue ● <input type="checkbox"/> liver ● <input type="checkbox"/> gall bladder ● <input type="checkbox"/> intestine ● <input type="checkbox"/> stomach ● <input type="checkbox"/> brain ● Cells treatment/state information <ul style="list-style-type: none"> ● <input type="checkbox"/> normal/untreated/validation ● <input type="checkbox"/> treated ● <input type="checkbox"/> irradiated ● <input type="checkbox"/> infected (virus) ● <input type="checkbox"/> infected (other) ● <input type="checkbox"/> tumour ● <input type="checkbox"/> primary ● <input type="checkbox"/> low-grade ● <input type="checkbox"/> high-grade ● <input type="checkbox"/> secondary ● <input type="checkbox"/> metastatic ● <input type="checkbox"/> non-metastatic ● <input type="checkbox"/> recurrent ● <input type="checkbox"/> adenoma ● <input type="checkbox"/> carcinoma ● <input type="checkbox"/> hormone sensitive ● <input type="checkbox"/> hormone resistant ● <input type="checkbox"/> surgery ● Experiment type <ul style="list-style-type: none"> ● <input type="checkbox"/> time-course ● <input type="checkbox"/> classification ● <input type="checkbox"/> survival ● <input type="checkbox"/> survival>5 yr ● <input type="checkbox"/> survival<5 yr
<input type="button" value="Submit"/>	

Figure 21: Class selection page model. The page has been shortened for clarity

6.5. Using CleanEx : examples and applications, a CleanEx tutorial

6.5.1. CleanEx single entries and multiviewer

Fibronectins bind cell surfaces and various compounds including collagen, fibrin, heparin, DNA, and actin. Fibronectins are involved in cell adhesion, cell motility, wound healing, and maintenance of cell shape. Interaction with TNR mediates inhibition of cell adhesion and neurite outgrowth. They consist mostly of heterodimers or multimers of alternatively spliced variants, connected by 2 disulfide bonds near the carboxyl ends. This gene is submitted to a high degree of alternative splicing, and there are nowadays twelve different known fibronectin isoforms. This high number of transcript variants makes fibronectin a very good case study for CleanEx. Let's look at the HS_FN1 fibronectin CleanEx entry(http://www.cleanex.isb-sib.ch/cgi-bin/get_doc?db=cleanex&format=nice&entry=HS_FN1, Figure 22).

On the top of the page, direct access to the corresponding list of clones, tags, of Affymetrix probe sets for this gene is provided (Figure 23). On the same line, one can also extract the promoter sequence for this gene (Figure 24).

Nice View of CleanEx: [HS_FN1](#)

[\[General\]](#)
[\[RNA sequences\]](#)
[\[Cross-references\]](#)
[\[Expression data\]](#)

[View list of associated clones or tags](#)

[Extract corresponding genomic sequence](#)

General information about the entry		
Entry name	HS_FN1	
Locus	2q34	
Description of the gene	Fibronectin 1.	
Cross-references		
Entrez GeneID	2335	
Unigene	Hs.203717	
MIM	135600	
Genew	HGNC:3778; FN1	
GeneCards	FN1	
Ensembl	FN1	
RefSeq	NM_002026	
RefSeq	NM_054034	
RefSeq	NM_212474	
RefSeq	NM_212475	
RefSeq	NM_212476	
RefSeq	NM_212478	
RefSeq	NM_212482	
SWISSPROT	P02751 ; FINC_HUMAN	
EPD	EP16038 ; HS_FN1	
Expression Data References		
Dataset's Name	Target ID	Expression Data
GDS181 (Original web site)	AFFY_HG-U95Av2_31719_at	GDS181_31719_at [Entry (text) / Local viewer]
View all GDS181 experiments	AFFY_HG-U95Av2_31720_s_at	GDS181_31720_s_at [Entry (text) / Local viewer]
	AFFY_HG-U95Av2_AFFX-HUMRGE/M10098_3_at	GDS181_AFFX-HUMRGE/M10098_3_at [Entry (text) / Local viewer]
GDS505 (Original web site)	AFFY_HG-U133A_210495_x_at	GDS505_210495_x_at [Entry (text) / Local viewer]
View all GDS505 experiments	AFFY_HG-U133A_211719_x_at	GDS505_211719_x_at [Entry (text) / Local viewer]
	AFFY_HG-U133A_212464_s_at	GDS505_212464_s_at [Entry (text) / Local viewer]
	AFFY_HG-U133A_214701_s_at	GDS505_214701_s_at [Entry (text) / Local viewer]
	AFFY_HG-U133A_214702_at	GDS505_214702_at [Entry (text) / Local viewer]
	AFFY_HG-U133A_216442_x_at	GDS505_216442_x_at [Entry (text) / Local viewer]
	AFFY_HG-U133A_AFFX-HUMRGE/M10098_3_at	GDS505_AFFX-HUMRGE/M10098_3_at [Entry (text) / Local viewer]
	AFFY_HG-U133A_AFFX-r2-Hs18SrRNA-3_s_at	GDS505_AFFX-r2-Hs18SrRNA-3_s_at [Entry (text) / Local viewer]
PEROU1 (Original web site)	IMAGE_139009	P0001_139009 [Entry (text) / Original viewer / Local viewer]
View all PEROU1 experiments	IMAGE_268091	P0001_268091 [Entry (text) / Original viewer / Local viewer]
	IMAGE_269203	P0001_269203 [Entry (text) / Original viewer / Local viewer]
	IMAGE_296556	P0001_296556 [Entry (text) / Original viewer / Local viewer]
	IMAGE_60846	P0001_60846 [Entry (text) / Original viewer / Local viewer]
T0001 (Original web site)	IMAGE_1870935	T0001_16455 [Entry (text) / Local viewer]
View all T0001 experiments	IMAGE_233277	T0001_21421 [Entry (text) / Local viewer]
	IMAGE_256820	T0001_23539 [Entry (text) / Local viewer]
	IMAGE_296556	T0001_26630 [Entry (text) / Local viewer]
	IMAGE_415682	T0001_30517 [Entry (text) / Local viewer]
	IMAGE_502201	T0001_34343 [Entry (text) / Local viewer]
	IMAGE_139009	T0001_3455 [Entry (text) / Local viewer]
	IMAGE_139009	T0001_3456 [Entry (text) / Local viewer]
	IMAGE_840726	T0001_44462 [Entry (text) / Local viewer]
	IMAGE_1554962	T0001_6984 [Entry (text) / Local viewer]
RNA sequences according to Unigene available on May-2-2004		
EMBL	AB191261.1 ; AB191261 . [EMBL / GenBank / DDBJ]	
	AF312399.1 ; AF312399 . [EMBL / GenBank / DDBJ]	

Figure 22 : Nice view representation model of the CleanEx entry for the human fibronectin 1. The real HS_FN1 entry has been shortened.

Gene name	Unigene Cluster	Sequence AC	Probe name	Clone end	Probe type
FN1	Hs.418138	-	HG-U133 Plus 2 1558199_at		AFFY
FN1	Hs.418138	-	HG-U133 Plus 2 212464_s_at		AFFY
FN1	Hs.418138	-	HG-U133 Plus 2 214701_s_at		AFFY
FN1	Hs.418138	-	HG-U133 Plus 2 214702_at		AFFY
FN1	Hs.418138	-	HG-U133 Plus 2 216442_x_at		AFFY
FN1	Hs.418138	-	HG-U95A 1008_f_at		AFFY
FN1	Hs.418138	-	HG-U95B 45557_r_at		AFFY
FN1	Hs.418138	-	HuGeneFL X05276_at		AFFY
FN1	Hs.203717	AA033511.1			EST
FN1	Hs.203717	AA089528.1		5'	EST
FN1	Hs.203717	AA089787.1		5'	EST
FN1	Hs.203717	AA092761.1		5'	EST
FN1	Hs.203717	AA092804.1		5'	EST
FN1	Hs.203717	AA093454.1		5'	EST
FN1	Hs.203717	AA094608.1		5'	EST
FN1	Hs.203717	AA095858.1		5'	EST
FN1	Hs.203717	AA095876.1		5'	EST
FN1	Hs.203717	CR749316.1			HTC
FN1	Hs.203717	CR749317.1			HTC
FN1	Hs.203717	AA564214.1	IMAGE:1016221	3'	IMAGE
FN1	Hs.203717	AA564220.1	IMAGE:1016306	3'	IMAGE
FN1	Hs.203717	AA852772.1	NHTBCae16a12		OTHER_CLONE
FN1	HS.119878	-	SAGE NLAIII AAAGAAATCA		SAGE
FN1	HS.119878	-	SAGE NLAIII AAAGAAATCA		SAGE
FN1	HS.482017	-	SAGE SAU3AI ATGCTGCTGG		SAGE
FN1	HS.482017	-	SAGE SAU3AI ATGCTGCTGG		SAGE
FN1	Hs.203717	X02761.1			mRNA

Figure 23 : Clones and Tags list output for the human fibronectin. This table shows all the SAGE tags, Affymetrix probe sets, or clones from the corresponding Unigene cluster which have a match with the FN1 sequence. For SAGE and Affymetrix, a link to the CleanEx_trg entry gives access to the position of the tags on the sequence. The entry has been shortened

Retrieve Sequence Form

FN1 on contig NT_086634

Extract sequence from :

5' distance from TSS

3' distance from the TSS

>FN1|NT_086634|upstream length=1000, downstream length=100 |ANNOTATION
CTGCCTCCACGCTGAGTTATCCGATGCTGAAATGTCACAGCACTTAGTCTTACTCTTCTATGGCCTACTTTCTACTG
CTATTTGGTACTCATGCTACCCATCTTATCTCCCTCAGTGTGTGAGACGCTGGCAGATTGGCCTCTCCACACA
CTCAACATTATGTGTGCACACAGTAGGTAACAATACATGCAAGTTTCTGAAATAGATATTTCTAGTCATCTGTGC
ACCTGCTATCTACTGAAAATTACAAAATGCAATTAATCTCAATTTTACATTTGGGATTTACAGAAAATACTCTCT
CTCCAAGAAATGCATAACAATTTAGCTAGGGCAATGCCAGGTCCGAGTTAAGACATTAATGCGCTTCGATCGGATAAG
GATTTATCCTTATCCCATCCTCATCTTTCTGCGCTGCTAATTCAGTTAGGTCAGTAAAGGAAACCTTTTCGTTTTAG
CAACCAATCTGCTCCCTTCTTCTGCGCTCTTTCTCTCTTTTGTGGTAGACGACTCTCAGCCTCTGTCCTTAATTTTA
AAGTTTATGCCCACTGTACCCTCTGCTTTTGGTGATTTAGAGATTTTCAAAGCTGCTCTGCACACAGACTCTTCTT
GGATTGCAACTTCTACTTTGGGTTGGAAACGGCTTCTCCGTTTTGAAACGCTAGCGGGGAAAAATGGGGAGAAAGT
TGAGTTAAACTTTTAAAGTTGAGTCACGGCTGGTTGCGCAGCAAAAGCCCGCAGTGTGGAGAAAGCTAAACGTGGT
TTGGTGGTGGGGGTTGGGGGGGTTGGGGGGGACTTTGGGGGATAAGGGGCGTGGAGCCAGGGAATGCCAAAGCCCTGCC
CGGCCTCCGACGGCGCCGCCCGCCCTCGCTCTCCCGCCCGCAGTGAGGCCGGGCTCCCGCCGACTGATGTC
GCGCGCTTGGTGTGTGGCAACCGCAACTCAGAGGCCGGCCAGAAAACCGAGCGAGTAGGGGGCGGCGCGCA
GGAGGGAGGAGAATGGGGCGCGGGAGGCTGGTGGTGTGGGGGGTGAGATGAGAAGA

A

B

Figure 24: Sequence retrieval query box. This box allows users to select the length of the sequence to retrieve by changing the 5' and 3' distance from the Transcription Start Site.

The first point to mention about this gene is that it has seven different RefSeq references. This is an

evidence for alternative splicing. While looking down in the Expression Data Reference, one can see that this gene corresponds to more than one feature in most of the datasets. By clicking on the “View all DATASET experiments” one has access to all the features corresponding to the fibronectin 1 on that specific dataset. While selecting the PEROU1 dataset, one can see that, on the viewer page, one of the clones spotted (clone number 4, namely the IMAGE clone number 296556 , see Figure 25) does not behave in exactly the same way as the others. Selecting the multiple viewer for the T0001 dataset, which is a DNA microarray survey of gene expression in normal tissues, gives even more hints. The previous clone also shows a different expression pattern , together with the clone number 502201. This could push the user to go further and compare the positions of these respective clones on the genomic or RefSeq sequences. Doing this leads to the following conclusion : compared to the four other ones, these two clones map to a different place on the genomic sequence for fibronectin.

Now let's have a look at another dataset's results, namely the GDS505 (Figure 26). This dataset compares cell carcinoma samples with their corresponding adjacent normal tissue samples from the same patient. By comparing the expression patterns of the six first probe sets, on the multiviewer, it is evident that two probe sets (in column 4 and 5) have a different expression pattern. Looking at the corresponding targets shows that these two probe sets match a different set of RefSeq sequences, though each of them has a high quality tag. Moreover, there is a “gradient pattern” between the probes sharing some, but not all, the RefSeq reference sequences.

By using the multiviewer for determining alternatively spliced transcripts which are differentially expressed, one always has to keep in mind that the results for all the targets, regardless their quality tag, are shown. This means that it's always very wise to check for the real mapping of each target before going on with a deeper sequence analysis of the different targets. For example, in the Affymetrix dataset shown before, the two last probe sets also show a slightly different behavior. Going from there to the respective target page, one will mention that this probe set has a low quality tag, and that only a single tag out of all the tags compiled for this probe set matches a sequence. This obviously means that the expression data for this probe set should not be taken into account for further analysis.

Column number	Experiment ID	Target ID
1	T0001_16455	IMAGE_1870935
2	T0001_21421	IMAGE_233277
3	T0001_23539	IMAGE_256820
4	T0001_26630	IMAGE_296556
5	T0001_30517	IMAGE_415682
6	T0001_34343	IMAGE_502201
7	T0001_3455	IMAGE_139009
8	T0001_3456	IMAGE_139009
9	T0001_44462	IMAGE_840726
10	T0001_6984	IMAGE_1554962

Color Scale for the log ratio display :



Experiments results :



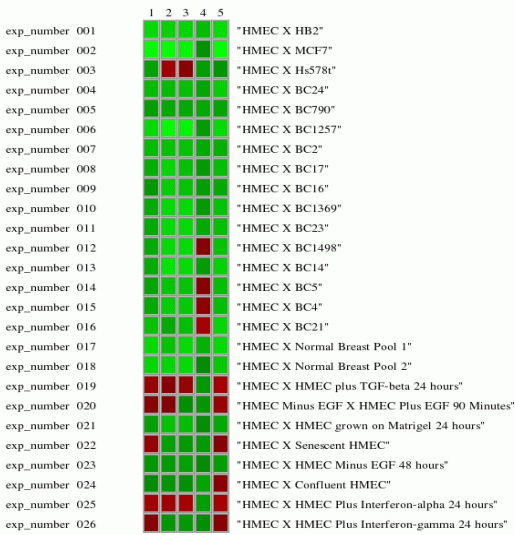
5 clones found

Column number	Experiment ID	Target ID
1	P0001_139009	IMAGE_139009
2	P0001_268091	IMAGE_268091
3	P0001_269203	IMAGE_269203
4	P0001_296556	IMAGE_296556
5	P0001_60846	IMAGE_60846

Color Scale for the log ratio display :



Experiments results :



A

B

Figure 25: Expression representation for P0001 (A) and T0001 (B) datasets. All spots corresponding to fibronectin 1 are shows on the same image for each dataset.

Column number	Experiment ID	Target ID
1	GDS505_210495_x_at	AFFY_HG-U133A_210495_x_at
2	GDS505_211719_x_at	AFFY_HG-U133A_211719_x_at
3	GDS505_212464_s_at	AFFY_HG-U133A_212464_s_at
4	GDS505_214701_s_at	AFFY_HG-U133A_214701_s_at
5	GDS505_214702_at	AFFY_HG-U133A_214702_at
6	GDS505_216442_x_at	AFFY_HG-U133A_216442_x_at
7	GDS505_AFFX-HUMRGE/M10098_3_at	AFFY_HG-U133A_AFFX-HUMRGE/M10098_3_at
8	GDS505_AFFX-r2-Hs18SrRNA-3_s_at	AFFY_HG-U133A_AFFX-r2-Hs18SrRNA-3_s_at

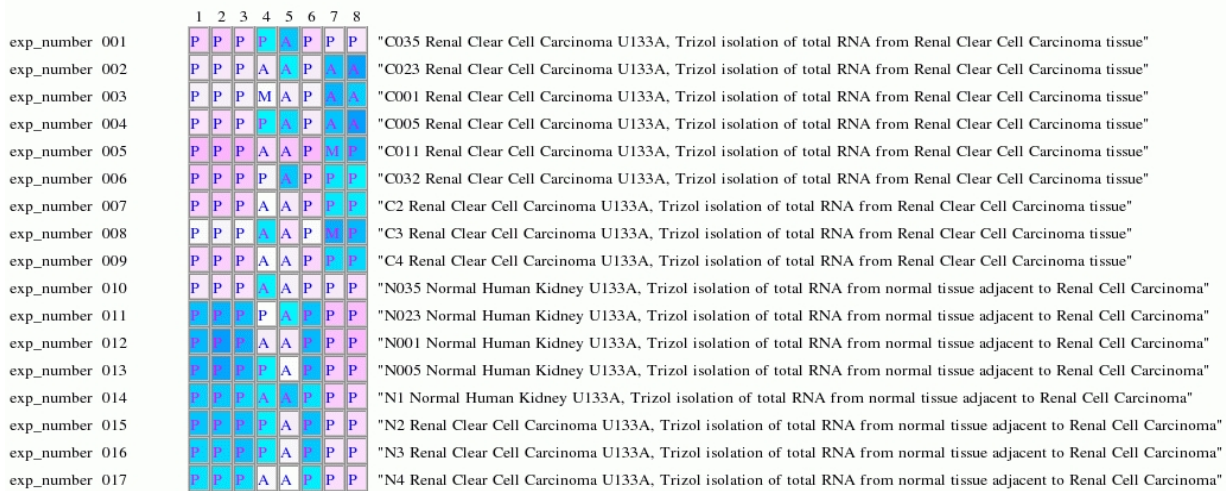


Figure 26 :Multiviewer representation of the Affymetrix dataset GDS505. Experiment results can be influenced by the individual probe position on the gene sequence, as well as by the number of individual probes matching this sequence.

Back to the T0001 multiviewer page for the fibronectin, one can mention that the same clone (IMAGE number 139009, corresponding to columns seven and eight on the viewer) has been spotted twice for this dataset. Looking at the two columns together, it is quite reassuring to see that the two spots behave exactly the same way along all the experiments. This viewer can thus also provide a visual approximation of the data quality.

Another nice example of good internal dataset quick control is given with the entry HS_KLK3. This protein is highly specific to prostate tissue. The T0001 dataset, again, shows a higher expression level in prostate tissue. By then selecting the multiview for the GDS181 dataset, which is an Affymetrix-based survey of expression in normal tissues, one can see a very good correlation between the three probe sets pertaining to the same gene, as well as a high difference in the expression level between prostate tissues and other tissues. Interestingly, when opening the corresponding targets pages, one can mention that one of the three targets matches on EMBL sequences, but not on the RefSeq which is the reference for the two other targets. This is due to a minor discrepancy (deletion of two nucleotides in the EMBL RNA sequence) between the EMBL RNA and the RefSeq. The Affymetrix probe sets must have been

designed by using the two types of sequences. The fact that CleanEx now takes also into account the tags which match on the EMBL RNAs thus allows one to retrieve this target as a high quality one, quality which would have been lost otherwise.

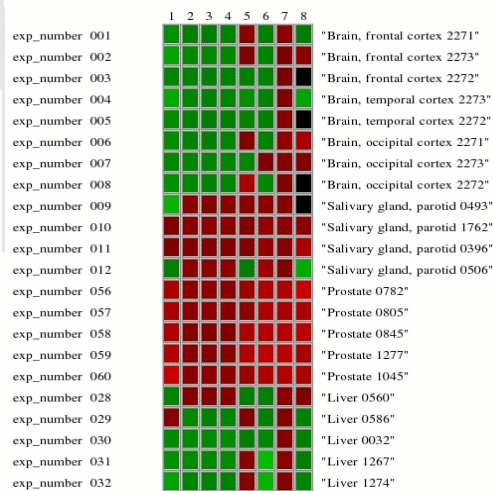
Looking at these two datasets showing expression in normal tissues (Figure 27), one can see an obvious correlation. For example, though prostate-specific, this gene also seems to be expressed in salivary gland in both datasets. Nevertheless, the Affymetrix data appears to be more precise. This could be due to the fact that the three Affymetrix probe sets come from the very same region of the gene, and thus will exhibit a very similar expression pattern. If one considers the alignment of the genomic sequence for KLK3 with the clones from T0001 and the individual tags from the Affymetrix dataset shown in Figure 28, one can easily see that the clones span a much extended region of the gene compared to the Affymetrix tags, and the relative fuzziness in the microarray dataset could be attributed to this.

General information about the selected dataset and gene

Gene	KLK3
Gene description	Kallikrein 3, (prostate specific antigen).
Dataset	T0001
Dataset description	DNA microarray survey of gene expression in normal tissues
Dataset format	Stanford_Scanalyze
Web access	http://genome-www5.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=426
Reference	Genome Biology 2005, 6:R22;

8 clones found

Column number	Experiment ID	Target ID
1	T0001_121	IMAGE_1008836
2	T0001_156	IMAGE_1010026
3	T0001_168	IMAGE_1010103
4	T0001_39283	IMAGE_745615
5	T0001_41215	IMAGE_782957
6	T0001_46173	IMAGE_914588
7	T0001_46214	IMAGE_916250
8	T0001_46524	IMAGE_953487



A

General information about the selected dataset and gene

Gene	KLK3
Gene description	Kallikrein 3, (prostate specific antigen).
Dataset	GDS181
Dataset description	Large-scale analysis of the human transcriptome
Dataset format	Affy_probeset
Web access	http://expression.gnf.org
Reference	Proc Natl Acad Sci U S A. 2002 Apr 2;99(7):4465-70. Epub 2002 Mar 19

3 clones found

Column number	Experiment ID	Target ID
1	GDS181_1804_at	AFFY_HG-U95Av2_1804_at
2	GDS181_1805_g_at	AFFY_HG-U95Av2_1805_g_at
3	GDS181_40794_at	AFFY_HG-U95Av2_40794_at



B

Figure 27: Expression level for KLK3 in two different dataset types. Each dataset compares expression in normal human tissues. KLK3 is over expressed in salivary gland and in prostate in the two datasets.

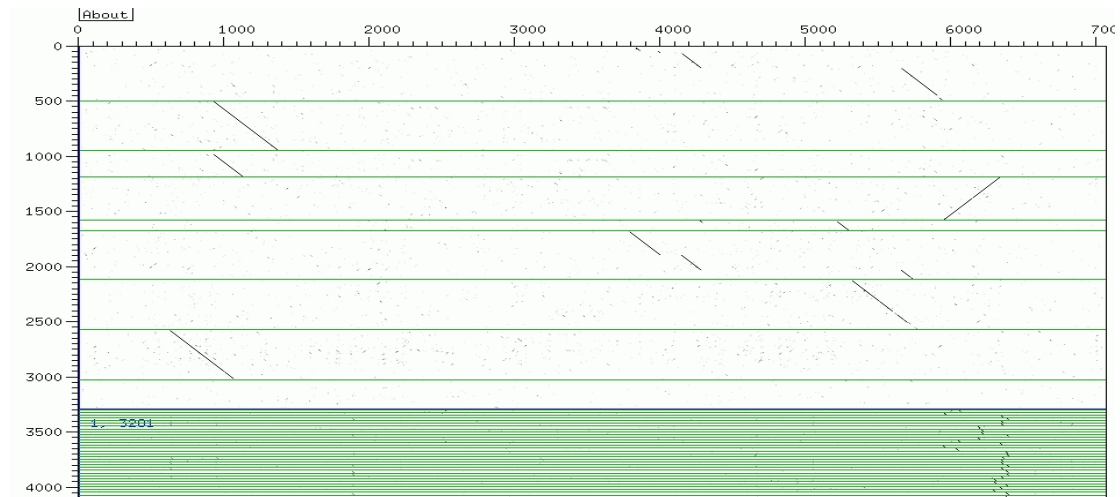


Figure 28: Representation of the alignment of the *KLK3* genomic sequence (horizontal) against the clones from experiment T0001 (wide rows), and individual tags from the two GDS181 dataset's probe sets (narrow rows). Clones span a much wider region of the transcript.

6.5.2. Finding common expression patterns in different datasets

To show the accuracy of the step-by-step cross-dataset analysis procedure, we will present two different examples of analysis. The first one will focus on the two datasets previously cited, namely the T0001 and GDS181, which are two experiments done with normal tissue, one with microarray, and one with Affymetrix chips. The second case study will show the comparison of expression patterns in two datasets comparing expression in astrocytic gliomas and astrocytomas. The first one, C0001, has been realized with nylon-membrane arrays, the second dataset, AFFY002, is an Affymetrix-based experiment.

6.5.2.1. Normal tissues : comparison of two dataset types

With the example of the Kallikrein 3 gene, we have seen that this gene is highly prostate-specific. We will now try to find genes which share the same criteria. By selecting as first pool all other normal tissues, and as second pool prostate tissues in the GDS181 query form, we extract the first 500 genes showing over-expression in prostate (Figure 29A). On the results page, it is reassuring to see that the kallikrein 3 probe sets are all amongst the first ones listed (Figure 29B). Including the prostate cancer tissues in the second pool introduces a very small bias towards cancer-specific genes. Now we can push onto the second step. On the results page, the table below the feature's list contains the number of common features, among the list reported, in other datasets. By choosing to go on with the analysis in the T0001 dataset, we want to confirm the results found in GDS181. This will extract the genes which show the same expression characteristics in both datasets. And the result page confirms that kallikrein 3 is indeed overexpressed in prostate, along with other genes like KLK2, ACPP (Figure 29C). These results are confirmed by going back to the CleanEx entry and checking the respective datasets multiviewers. Interestingly, among all the top genes of the list, most of them also show over-expression in salivary gland.

Expression Query Form for GDS181 dataset

Find all genes which show in group2 compared with group1

Category code	Experiment	GROUP 1	VS	GROUP 2
EXP 1	H9LMS00102615, whole brain	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 2	H9LMS00102605, whole brain	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 3	H9DKG00061403, cortex	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 4	H9LMS00111501, cerebellum	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 5	H9LMS00111502, cerebellum	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 6	H9RGV00102602, amygdala	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 7	H9RGV00110701, amygdala	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 8	H9RGV00102612, fetal brain	<input type="checkbox"/>		<input type="checkbox"/>
EXP 9	H9RGV00111403, fetal brain	<input type="checkbox"/>		<input type="checkbox"/>
EXP 10	H9RGV00110702, caudate nucleus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 11	H9RGV00110703, caudate nucleus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 12	H9LMS00111701, spinal cord	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 13	H9LMS00111505, spinal cord	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 14	H9RGV00102610, thalamus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 15	H9RGV00111502, thalamus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 16	H9RGV00110705, corpus callosum	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 17	H9RGV00111404, corpus callosum	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 18	H9RGV00111002, DRG	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 19	H9RGV00111001, DRG	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 20	H9LMS00092711, thymus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 21	H9LMS00092712, thymus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 22	H9DKG00061402, retinoblastoma	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 23	H9LMS00061601, lung	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 24	H9LMS00102609, lung	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 25	H9LMS00092709, testis	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 26	H9LMS00092710, testis	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 27	H9LMS00102606, heart	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 28	H9LMS00112901, heart	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 29	H9LMS00102607, kidney	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 30	H9LMS00111002, kidney	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 31	H9RGV00061304, kidney	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 32	H9LMS00102612, spleen	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 33	H9LMS00121802, spleen	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 34	H9LMS00111004, pancreas	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 35	H9RGV00061306, pancreas	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 36	H9LMS00102614, thyroid	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 37	H9LMS00111007, thyroid	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 38	H9LMS99102622, ovary pool	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 39	H9LMS00111503, ovary-pool	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 40	H9LMS99102618, ovary pool	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 41	H9LMS99102620, OVR278E	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 42	H9LMS99102621, OVR278S	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 43	H9LMS00111506, uterus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 44	H9LMS00102617, uterus	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 45	H9LMS00102610, placenta	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 46	H9LMS00111005, placenta	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 47	H9RGV00102601, trachea	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 48	H9LMS00111601, trachea	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 49	H9RGV00102608, pituitary gland	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 50	H9RGV00112206, pituitary gland	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 51	H9LMS00102604, adrenal gland	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 52	H9LMS00111001, adrenal gland	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 53	H9LMS00111006, salivary gland	<input type="checkbox"/>		<input type="checkbox"/>
EXP 54	H9LMS00102611, salivary gland	<input type="checkbox"/>		<input type="checkbox"/>
EXP 55	H9RGV00102611, fetal liver	<input type="checkbox"/>		<input type="checkbox"/>
EXP 56	H9RGV00111402, fetal liver	<input type="checkbox"/>		<input type="checkbox"/>
EXP 57	H9LMS00102608, liver	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 58	H9LMS00111003, liver	<input checked="" type="checkbox"/>		<input type="checkbox"/>
EXP 59	H9LMS99081101, prostate cancer	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 60	H9LMS99081601, prostate cancer	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 61	H9LMS99072808, prostate cancer	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 62	H9LMS99081103, prostate cancer	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 63	H9LMS99080701, prostate cancer	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 64	H9LMS99090506, prostate	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 65	H9LMS99090504, prostate	<input type="checkbox"/>		<input checked="" type="checkbox"/>
EXP 66	H9LMS99090508, prostate	<input type="checkbox"/>		<input checked="" type="checkbox"/>

Show entries which figure amongst the

First sequences

OR

First percent of the data (Total sequences in dataset : 12655)

Figure 29A: GDS181 dataset selection form. The two experiments pools will be compared according to the user's criteria. Here, we search for genes over expressed in prostate compared to other tissues.

Display the first 500 sequences (out of 12655) corresponding to the given comparison

CleanEx_ref ID	Gene name	Reference Sequence	Description	Diff value	Rank	
GDS181_40794_at	KLK3	AFFY_HG-U95Av2_40794_at	Kallikrein 3, (prostate specific antigen).	12646	1	<input type="checkbox"/>
GDS181_32200_at	ACPP	AFFY_HG-U95Av2_32200_at	Acid phosphatase, prostate.	12637	2	<input type="checkbox"/>
GDS181_41721_at	KLK2	AFFY_HG-U95Av2_41721_at	Kallikrein 2, prostatic.	12635	3	<input type="checkbox"/>
GDS181_217_at	KLK2	AFFY_HG-U95Av2_217_at	Kallikrein 2, prostatic.	12625	4	<input type="checkbox"/>
GDS181_1514_g_at	Not in CleanEx	AFFY_HG-U95Av2_1514_g_at	None	12623	5	<input type="checkbox"/>
GDS181_41172_at	RDH11	AFFY_HG-U95Av2_41172_at	Retinol dehydrogenase 11 (all-trans and 9-cis).	12610	6	<input type="checkbox"/>
GDS181_32149_at	MSMB	AFFY_HG-U95Av2_32149_at	Microsminoprotein, beta-.	12604	7	<input type="checkbox"/>
GDS181_617_at	ACPP	AFFY_HG-U95Av2_617_at	Acid phosphatase, prostate.	12602	8	<input type="checkbox"/>
GDS181_39278_at	TGM4	AFFY_HG-U95Av2_39278_at	Transglutaminase 4 (prostate).	12591	9	<input type="checkbox"/>
GDS181_1804_at	KLK3	AFFY_HG-U95Av2_1804_at	Kallikrein 3, (prostate specific antigen).	12562	10	<input type="checkbox"/>
GDS181_32112_s_at	AIM1	AFFY_HG-U95Av2_32112_s_at	Absent in melanoma 1.	12500	11	<input type="checkbox"/>
GDS181_32609_at	HIST2H2AA	AFFY_HG-U95Av2_32609_at	Histone 2, H2aa.	12450	12	<input type="checkbox"/>
GDS181_1805_g_at	KLK3	AFFY_HG-U95Av2_1805_g_at	Kallikrein 3, (prostate specific antigen).	12423	13	<input type="checkbox"/>
GDS181_1071_at	GATA2	AFFY_HG-U95Av2_1071_at	GATA binding protein 2.	12418	14	<input type="checkbox"/>
GDS181_38763_at	SORD	AFFY_HG-U95Av2_38763_at	Sorbitol dehydrogenase.	12395	15	<input type="checkbox"/>
GDS181_37141_at	FOXA1	AFFY_HG-U95Av2_37141_at	Forkhead box A1.	12390	16	<input type="checkbox"/>
GDS181_32113_at	AIM1	AFFY_HG-U95Av2_32113_at	Absent in melanoma 1.	12350	17	<input type="checkbox"/>
GDS181_37023_at	LCP1	AFFY_HG-U95Av2_37023_at	Lymphocyte cytosolic protein 1 (L-plastin).	12323	18	<input type="checkbox"/>
GDS181_32134_at	TES	AFFY_HG-U95Av2_32134_at	Testis derived transcript (3 LIM domains).	12316	19	<input type="checkbox"/>
GDS181_36898_r_at	PRIM2A	AFFY_HG-U95Av2_36898_r_at	Primase, polypeptide 2A, 58kDa.	12315	20	<input type="checkbox"/>

Common genes in other datasets

Dataset	Sequences	Corresponding Gene Symbols	
AFFY001	184 sequences	95 corresponding gene symbols	<input type="checkbox"/>
AFFY002	670 sequences	385 corresponding gene symbols	<input type="checkbox"/>
GDS422	670 sequences	385 corresponding gene symbols	<input type="checkbox"/>
GDS426	101 sequences	69 corresponding gene symbols	<input type="checkbox"/>
HSEST	425 sequences	380 corresponding gene symbols	<input type="checkbox"/>
R0001	580 sequences	363 corresponding gene symbols	<input type="checkbox"/>
T0001	1170 sequences	365 corresponding gene symbols	<input type="checkbox"/>
GDS425	64 sequences	48 corresponding gene symbols	<input type="checkbox"/>
L0001	585 sequences	168 corresponding gene symbols	<input type="checkbox"/>
P0001	257 sequences	151 corresponding gene symbols	<input type="checkbox"/>
GDS182	253 sequences	174 corresponding gene symbols	<input type="checkbox"/>
GDS424	80 sequences	54 corresponding gene symbols	<input type="checkbox"/>
NCI60	343 sequences	228 corresponding gene symbols	<input type="checkbox"/>
S0001	343 sequences	228 corresponding gene symbols	<input type="checkbox"/>
GDS423	106 sequences	70 corresponding gene symbols	<input type="checkbox"/>
C0001	78 sequences	69 corresponding gene symbols	<input type="checkbox"/>

Extract sequence in the checked subset(s)

5' distance from TSS

3' distance from the TSS

Continue analysis for this gene subset in the selected dataset :

Figure29B : Results of the first step. These are genes over expressed in prostate in the GDS181 dataset. 1170 sequences, corresponding to 365 different genes are also found in the T0001 dataset. We will then go on in analyzing the 100 most over expressed genes in prostate in T0001 for this new sequence pool.

Expression Query Form for T0001 subset of 1170 genes

Find all genes which show **OVEREXPRESSION** in group2 compared with group1

Genes in group 2 are OVEREXPRESSED compared to group1

Result extracts the sequences which are shared by previous analysis step with GDS181 and T0001 and which belong to the first 100 sequences (out of 47018) corresponding to the given criteria

CleanEx_ref ID	Gene name	Reference Sequence	Description	Diff value	Rank	
T0001_1677	ACPP	IMAGE_1203949	Acid phosphatase, prostate.	44093	9	☐
T0001_120	ACPP	IMAGE_1008791	Acid phosphatase, prostate.	42379	13	☐
T0001_46714	ACPP	IMAGE_984952	Acid phosphatase, prostate.	40463	26	☐
T0001_30534	ANGPT1	IMAGE_415769	Angiopoietin 1.	40119	32	☐
T0001_1636	ACPP	IMAGE_1203125	Acid phosphatase, prostate.	39999	33	☐
T0001_121	KLK3	IMAGE_1008836	Kallikrein 3, (prostate specific antigen).	39635	38	☐
T0001_21296	AGR2	IMAGE_2321113	Anterior gradient 2 homolog (Xenopus laevis).	38761	47	☐
T0001_95	KLK2	IMAGE_1007855	Kallikrein 2, prostatic.	38515	48	☐
T0001_118	MSMB	IMAGE_1008751	Microsminoprotein, beta-.	38483	49	☐
T0001_40927	PLA2G2A	IMAGE_77915	Phospholipase A2, group IIA (platelets, synovial fluid).	38360	50	☐
T0001_4888	KRT15	IMAGE_1474900	Keratin 15.	37272	62	☐
T0001_20282	ALOX15B	IMAGE_2118808	Arachidonate 15-lipoxygenase, second type.	37194	64	☐
T0001_33241	MALT1	IMAGE_462778	Mucosa associated lymphoid tissue lymphoma translocation gene 1.	36607	68	☐
T0001_34958	AGR2	IMAGE_510576	Anterior gradient 2 homolog (Xenopus laevis).	36530	70	☐
T0001_26605	DMXL1	IMAGE_296210	Dmx-like 1.	36467	72	☐
T0001_34929	FXYD3	IMAGE_510336	FXYD domain containing ion transport regulator 3.	36148	77	☐
T0001_19438	TMPRSS2	IMAGE_2028487	Transmembrane protease, serine 2.	35995	78	☐
T0001_46599	CLDN4	IMAGE_967930	Claudin 4.	35915	81	☐
T0001_38031	TFAP2C	IMAGE_725680	Transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma).	35833	82	☐
T0001_9581	CDC2L6	IMAGE_1619693	Cell division cycle 2-like 6 (CDK8-like).	35745	84	☐

Figure 29C : Result of the re-analysis of the sequence pool in T0001. The first part is the header of the T0001 query form. The gene list includes the *KLK3* gene, as well as other genes known to be over expressed in prostate tissue, like *KLK2*, *ACPP*, or *ANGPT1*.

6.5.2.2. Astrocytomas and astrocytic gliomas comparison

Two other datasets generated with different platforms but including relatively close studies are the AFFY002, which compares low-grade and high-grade astrocytomas, and C0001, which classifies three groups of astrocytic gliomas according to gene expression.

Beginning with the Affymetrix-based dataset, we first extract genes which are overexpressed in high-grade astrocytomas compared with the low-grades (Figure 30A). As before, we continue the analysis by using the resulting gene list as input for the second step. Though separated in three categories by the authors, to keep as close as possible to the first dataset, we will use a two classes separation based on the WHO classification for tumors of the nervous system. This includes WHO grade II for the low-grade astrocytomas, consisting of our first pool, and WHO grade IV for the second pool, including

primary and secondary glioblastomas. The result from Figure 30B shows a great correlation with the papers describing expression changes in these two tumor categories.

Criteria :

GROUP 1	GROUP 2
Low-grade astrocytoma sample 1	High-grade astrocytoma sample 1
Low-grade astrocytoma sample 2	High-grade astrocytoma sample 2
Low-grade astrocytoma sample 3	High-grade astrocytoma sample 3
Low-grade astrocytoma sample 4	High-grade astrocytoma sample 4
Low-grade astrocytoma sample 6	High-grade astrocytoma sample 5
Low-grade astrocytoma sample 8	High-grade astrocytoma sample 6
Low-grade astrocytoma sample 9	

Genes in group 2 are OVEREXPRESSED compared to group1

Display the first 1262 sequences (out of 12625) corresponding to the given comparison

CleanEx_ref ID	Gene name	Reference Sequence	Description	Diff value	Rank	
AFFY002_32874_at	TCF3	AFFY_HG-U95Av2_32874_at	Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47).	5326	1	<input type="checkbox"/>
AFFY002_709_at	TUBB	AFFY_HG-U95Av2_709_at	Tubulin, beta polypeptide.	5307	2	<input type="checkbox"/>
AFFY002_37258_at	TMEFF1	AFFY_HG-U95Av2_37258_at	Transmembrane protein with EGF-like and two follistatin-like domains 1.	5279	3	<input type="checkbox"/>
AFFY002_32751_at	ILF3	AFFY_HG-U95Av2_32751_at	Interleukin enhancer binding factor 3, 90kDa.	5253.5	4	<input type="checkbox"/>
AFFY002_41188_at	LAPTM4B	AFFY_HG-U95Av2_41188_at	Lysosomal associated protein transmembrane 4 beta.	5190.5	5	<input type="checkbox"/>
AFFY002_1373_at	TCF3	AFFY_HG-U95Av2_1373_at	Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47).	5147.5	6	<input type="checkbox"/>
AFFY002_41275_at	E2F5	AFFY_HG-U95Av2_41275_at	E2F transcription factor 5, p130-binding.	5082.5	7	<input type="checkbox"/>
AFFY002_1287_at	PARP1	AFFY_HG-U95Av2_1287_at	Poly (ADP-ribose) polymerase family, member 1.	5001.5	8	<input type="checkbox"/>
AFFY002_32825_at	HRMT1L2	AFFY_HG-U95Av2_32825_at	HMT1 hnRNP methyltransferase-like 2 (S. cerevisiae).	5000.5	9	<input type="checkbox"/>
AFFY002_39113_at	Not in CleanEx			4950	10	<input type="checkbox"/>
AFFY002_1942_s_at	CDK4	AFFY_HG-U95Av2_1942_s_at	Cyclin-dependent kinase 4.	4899.5	11	<input type="checkbox"/>
AFFY002_1044_s_at	E2F5	AFFY_HG-U95Av2_1044_s_at	E2F transcription factor 5, p130-binding.	4896	12	<input type="checkbox"/>
AFFY002_37739_at	SSRP1	AFFY_HG-U95Av2_37739_at	Structure specific recognition protein 1.	4891	13	<input type="checkbox"/>
AFFY002_1792_g_at	CDK2	AFFY_HG-U95Av2_1792_g_at	Cyclin-dependent kinase 2.	4876	14	<input type="checkbox"/>
AFFY002_37393_at	HES1	AFFY_HG-U95Av2_37393_at	Hairy and enhancer of split 1, (Drosophila).	4867	15	<input type="checkbox"/>
AFFY002_40422_at	IGFBP2	AFFY_HG-U95Av2_40422_at	Insulin-like growth factor binding protein 2, 36kDa.	4858.5	16	<input type="checkbox"/>

Common genes in other datasets

Gene ID	Sequences	Corresponding gene symbols		Extract sequence in the checked subset(s)
AFFY001	373 sequences	211 corresponding gene symbols	<input type="checkbox"/>	
GDS181	1626 sequences	968 corresponding gene symbols	<input type="checkbox"/>	
GDS422	1626 sequences	968 corresponding gene symbols	<input type="checkbox"/>	
HSEST	1044 sequences	949 corresponding gene symbols	<input type="checkbox"/>	
L0001	1301 sequences	398 corresponding gene symbols	<input type="checkbox"/>	
NCI160	807 sequences	578 corresponding gene symbols	<input type="checkbox"/>	
P0001	587 sequences	357 corresponding gene symbols	<input type="checkbox"/>	
R0001	1383 sequences	927 corresponding gene symbols	<input type="checkbox"/>	
S0001	807 sequences	578 corresponding gene symbols	<input type="checkbox"/>	
T0001	2359 sequences	891 corresponding gene symbols	<input type="checkbox"/>	
GDS182	615 sequences	470 corresponding gene symbols	<input type="checkbox"/>	
GDS426	202 sequences	146 corresponding gene symbols	<input type="checkbox"/>	
GDS423	202 sequences	132 corresponding gene symbols	<input type="checkbox"/>	
C0001	237 sequences	195 corresponding gene symbols	<input type="checkbox"/>	
GDS425	168 sequences	129 corresponding gene symbols	<input type="checkbox"/>	
GDS424	176 sequences	117 corresponding gene symbols	<input type="checkbox"/>	

5' distance from TSS

3' distance from the TSS

Continue analysis for this gene subset in the selected dataset : [AFFY001](#)

Figure 30A : Low-grade versus high-grade astrocytomas comparison, first step. Analysis of genes in AFFY002 dataset.

Criteria :

GROUP 1	GROUP 2
Low grade astrocytoma WHO grade II, sample 898	Secondary glioblastoma WHO grade IV, sample 735
Low grade astrocytoma WHO grade II, sample 676	Primary glioblastoma WHO grade IV, sample G226
Low grade astrocytoma WHO grade II, sample 528	Secondary glioblastoma WHO grade IV, sample 772
Low grade astrocytoma WHO grade II, sample 355	Secondary glioblastoma WHO grade IV, sample 413
Low grade astrocytoma WHO grade II, sample 374	Primary glioblastoma WHO grade IV, sample G216
Low grade astrocytoma WHO grade II, sample 551	Primary glioblastoma WHO grade IV, sample 1621
Low grade astrocytoma WHO grade II, sample 510	Secondary glioblastoma WHO grade IV, sample 633
Low grade astrocytoma WHO grade II, sample 210	Primary glioblastoma WHO grade IV, sample 1284
Low grade astrocytoma WHO grade II, sample 875	Primary glioblastoma WHO grade IV, sample 1437
Low grade astrocytoma WHO grade II, sample 501	Primary glioblastoma WHO grade IV, sample 1316
Low grade astrocytoma WHO grade II, sample 589	Primary glioblastoma WHO grade IV, sample 1399
Low grade astrocytoma WHO grade II, sample 552	Primary glioblastoma WHO grade IV, sample G204
Low grade astrocytoma WHO grade II, sample 421	Primary glioblastoma WHO grade IV, sample 1430
Low grade astrocytoma WHO grade II, sample 698	Primary glioblastoma WHO grade IV, sample G197
Low grade astrocytoma WHO grade II, sample 416	Primary glioblastoma WHO grade IV, sample 1419
Low grade astrocytoma WHO grade II, sample 92	Primary glioblastoma WHO grade IV, sample 1308
Low grade astrocytoma WHO grade II, sample 289	Primary glioblastoma WHO grade IV, sample 1453
Low grade astrocytoma WHO grade II, sample 1070	Primary glioblastoma WHO grade IV, sample 1317
Low grade astrocytoma WHO grade II, sample 460	Primary glioblastoma WHO grade IV, sample 1297
Low grade astrocytoma WHO grade II, sample 80	Primary glioblastoma WHO grade IV, sample 1303
Low grade astrocytoma WHO grade II, sample 736	Primary glioblastoma WHO grade IV, sample 1360
Low grade astrocytoma WHO grade II, sample 635	Secondary glioblastoma WHO grade IV, sample 749
Low grade astrocytoma WHO grade II, sample 246	Secondary glioblastoma WHO grade IV, sample 946
Low grade astrocytoma WHO grade II, sample 328	Secondary glioblastoma WHO grade IV, sample 809
	Secondary glioblastoma WHO grade IV, sample 978
	Oligoastrocytoma WHO grade III, sample 1357

Genes in group 2 are OVEREXPRESSED compared to group 1

Result extracts the sequences which are shared by previous analysis step with AFFY002 and C0001 and which belong to the first 200 sequences (out of 1185) corresponding to the given criteria

CleanEx_ref ID	Gene name	Reference Sequence	Description	Diff value	Rank	
C0001_A08f	IGFBP2	RNA_M35410	Insulin-like growth factor binding protein 2, 36kDa.	1048	2	☑
C0001_E14a	TGFB2	RNA_M19154	Transforming growth factor, beta 2.	869	6	☑
C0001_E14m	IGHM	RNA_X57086	Immunoglobulin heavy constant mu.	827	8	☑
C0001_F06n	KLF10	RNA_S81439	Kruppel-like factor 10.	789	10	☑
C0001_B11i	SOCS2	RNA_AB004903	Suppressor of cytokine signaling 2.	754	16	☑
C0001_A11k	CKS2	RNA_X54942	CDC28 protein kinase regulatory subunit 2.	744	18	☑
C0001_F05d	LDHA	RNA_X02152	Lactate dehydrogenase A.	681	20	☑
C0001_E09e	GDF15	RNA_AF019770	Growth differentiation factor 15.	546	30	☑
C0001_F02n	TLK1	RNA_D50927	Tousled-like kinase 1.	496	34	☑
C0001_B06e	MELK	RNA_D79997	Maternal embryonic leucine zipper kinase.	410	48	☑
C0001_C04f	RFC4	RNA_M87339	Replication factor C (activator 1) 4, 37kDa.	408	49	☑
C0001_C07f	PRIM1	RNA_X74330	Primase, polypeptide 1, 49kDa.	404	51	☑
C0001_C06d	SUMO1	RNA_U83117	SMT3 suppressor of mif two 3 homolog 1 (yeast).	371	57	☑
C0001_C07b	CASP3	RNA_U13737	Caspase 3, apoptosis-related cysteine protease.	365	59	☑
C0001_D13a	CHAF1A	RNA_U20979	Chromatin assembly factor 1, subunit A (p150).	353	62	☑
C0001_C12e	TOP2A	RNA_J04088	Topoisomerase (DNA) II alpha 170kDa.	351	64	☑
C0001_A10k	CKS1B	RNA_X54941	CDC28 protein kinase regulatory subunit 1B.	339	67	☑
C0001_C05f	MCM2	RNA_D21063	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae).	339	68	☑
C0001_C13e	PCNA	RNA_M15796	Proliferating cell nuclear antigen.	320	77	☑
C0001_F04b	HSPG2	RNA_M85289	Heparan sulfate proteoglycan 2 (perlecan).	317	79	☑
C0001_E04i	PDGFA	RNA_U41745	PDGFA associated protein 1.	315	80	☑
C0001_D07b	HMGXB2	RNA_X62534	High-mobility group box 2.	301	85	☑
C0001_F12d	UMPS	RNA_J03626	Uridine monophosphate synthetase (orotate phosphoribosyl transferase and orotidine-5'-decarboxylase).	295	87	☑
C0001_E09n	COL4A2	RNA_X05610	Collagen, type IV, alpha 2.	291	90	☑
C0001_A04i	CCNA2	RNA_X51688	Cyclin A2.	290	91	☑
C0001_A11e	MYBL2	RNA_X13293	V-myb myeloblastosis viral oncogene homolog (avian)-like 2.	280	100	☑
C0001_F11b	XPO1	RNA_Y08614	Exportin 1 (CRM1 homolog, yeast).	274	103	☑
C0001_A05i	CCNB1	RNA_M25753	Cyclin B1.	262	108	☑
C0001_F11b	PRPS1	RNA_D00860	Phosphoribosyl pyrophosphate synthetase 1.	256	111	☑
C0001_A02e	MYB	RNA_M15024	V-myb myeloblastosis viral oncogene homolog (avian).	225	124	☑
C0001_B02b	DYRK3	RNA_Y12735	Dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 3.	220	129	☑
C0001_C07e	SIAH2	RNA_U76248	Seven in absentia homolog 2 (Drosophila).	204	140	☑
C0001_D03h	GNAI3	RNA_M27543	Guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3.	202	143	☑
C0001_C06f	MCM4	RNA_X74794	MCM4 minichromosome maintenance deficient 4 (S. cerevisiae).	196	149	☑
C0001_E03d	IFRD1	RNA_Y10313	Interferon-related developmental regulator 1.	185	158	☑
C0001_F14n	MKI67	RNA_X65550	Antigen identified by monoclonal antibody Ki-67.	184	160	☑
C0001_A01i	DLEU1	RNA_Y15227	Deleted in lymphocytic leukemia, 1.	183	161	☑
C0001_A02j	CDC2	RNA_X05360	Cell division cycle 2, G1 to S and G2 to M.	171	166	☑
C0001_B11b	GRB10	RNA_U69276	Growth factor receptor-bound protein 10.	160	175	☑
C0001_B10b	ABI2	RNA_U23435	Abl interactor 2.	158	176	☑
C0001_A01k	WEE1	RNA_U10564	WEE1 homolog (S. pombe).	149	185	☑
C0001_E08a	BMP6	RNA_M60315	Bone morphogenetic protein 6.	149	187	☑
C0001_F02c	MTHFD1	RNA_J04031	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methylenetetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase.	141	194	☑
C0001_F06i	THOC1	RNA_L36529	THO complex 1.	139	195	☑
C0001_E08h	JAG1	RNA_AF028593	Jagged 1 (Alagille syndrome).	137	198	☑
C0001_F10d	RRM1	RNA_X59543	Ribonucleotide reductase M1 polypeptide.	137	199	☑
C0001_A13k	AURKB	RNA_AF008552	Aurora kinase B.	136	200	☑

Figure 30B : Genes showing over expression in the two examined datasets (AFFY002 and C0001).

6.5.3. *Single dataset and sequence extraction : using SSA*

The next example will show how to extract common sequence features in a set of co-expressed genes. As source dataset, we will use the in-house built per-tissue breakdown of EST counts. From the human EST dataset, we select as first pool the normal tissues, and as second pool the cancer tissues. As cell lines sometimes show an extreme behavior, the experiments based on cell-lines will be discarded. The first result extracts genes which are overexpressed in cancer tissues compared to normal tissues. Once the gene list is provided, instead of going on with the expression analysis, we can extract the promoter region of each gene from the given list. This sequence list will be in FASTA format. Another option is to use this sequence set in further promoter sequence analysis via SSA, the Signal Search Analysis server (<http://www.isrec.isb-sib.ch/ssa/>) [51]. This on-line tool has been created on the basis of former sequence analysis tools developed at the SIB by Philipp Bucher [52] and is now maintained by Giovanna Ambrosini. This server provides different search tools, amongst which one finds the Oprof, or Signal Occurrence Profile generation. Oprof scans a set of fixed-length DNA sequences aligned with respect to a functional site, for example the transcription start site, in a sliding window in order to determine the frequency with which a particular sequence motif (signal) defined by a particular signal occurs. With the sequence set extracted from the previous expression analysis, one can for example search for the frequency of TATA-boxes, or also of CpG islands, in the promoter region. repeating this analysis with genes overexpressed in cancer tissues or with genes underexpressed in cancer tissues will lead to a striking conclusion. In general, TATA-boxes are more frequent in cancer-specific genes than in other genes.

6.5.4. *By class expression pattern search*

The by-class expression analysis tool is currently being developed. The principle is as follows :

The first step in the by-class web interface allows the extraction of experiments corresponding to specific biological conditions. These data are separated in two pool which will then be compared. The set of extracted experiments as well as a short experiment description is then displayed. At this point, one can decide to get rid of some chips which appear not to be relevant for the case under study. Searching for gene names in the remaining experiments leaves us with a certain number of common genes. Once expression values for each gene and for each selected experiment in the two pools have

been extracted, one can compare the expression level between the two selected pools. The comparison criteria has to be determined at that level. The normalization step is then performed on the extracted expression data, and the analysis of genes showing a different expression level in the two selected pools goes on.

Here is a practical example of application for this tool. Suppose that we want to find genes which are overexpressed in metastatic samples but not in non-metastatic samples. After having selected the non-metastatic class as our first pool and the metastatic class as our second pool, the corresponding experiments are extracted. Figure 31A shows the experiments pools retrieval. This step leaves us with a set of 1199 common genes (see Figure 31B) across the different selected experiments. If then one choses to classify these genes so that the first ones on the list will show overexpression in the metastatic pool, one obtains the result shown in Figure 31. It is striking to see, for example, that the first and third candidates in that list are the MUC2 (Mucin 2, whose expression is associated with aggressive tumor behavior [53]) and the MTA1 (Metastasis associated 1) genes. The matrix metalloproteinases, which are also involved in tumor invasion, are also high in this gene list. This first trial for real cross-dataset comparison via biological classes is thus quite promising.

Chips selector module

Please untick the chips that you want to discard for your analysis

FIRST SET : 109 experiments	SECOND SET : 16 experiments
Human AND Mammary gland OR Ovary OR Uterus OR Testis OR Prostate OR Intestine OR Brain AND Non-metastatic	Human AND Mammary gland OR Ovary OR Uterus OR Testis OR Prostate OR Intestine OR Brain AND Metastatic
<input checked="" type="checkbox"/> AFFY001_EX1 : "non-metastatic (M0) primary medulloblastoma sample 2" <input checked="" type="checkbox"/> AFFY001_EX2 : "non-metastatic (M0) primary medulloblastoma sample 3" <input checked="" type="checkbox"/> AFFY001_EX3 : "non-metastatic (M0) primary medulloblastoma sample 4" <input checked="" type="checkbox"/> AFFY004_EX9 : "Infiltrating ductal breast adenocarcinoma, non-metastatic, test sample 24" <input checked="" type="checkbox"/> AFFY004_EX10 : "Infiltrating ductal breast adenocarcinoma, non-metastatic, test sample 30" <input checked="" type="checkbox"/> AFFY004_EX172 : "Prostate adenocarcinoma, non-metastatic, sample 9"	<input checked="" type="checkbox"/> AFFY001_EX15 : "metastatic (M+) primary medulloblastoma sample 1" <input checked="" type="checkbox"/> AFFY001_EX16 : "metastatic (M+) primary medulloblastoma sample 2" <input checked="" type="checkbox"/> AFFY001_EX17 : "metastatic (M+) primary medulloblastoma sample 3" <input checked="" type="checkbox"/> AFFY004_EX173 : "Prostate adenocarcinoma, metastatic, test sample 40" <input checked="" type="checkbox"/> AFFY004_EX174 : "Prostate adenocarcinoma, metastatic, test sample 41"
<input type="button" value="Extract Common genes"/>	

Figure 31A : The Chips Selector Module. The two pools have been extracted from two different datasets.

Comparison module

PLEASE WAIT...

Finding common genes in CleanEx for selected datasets

Found 1199 common genes for all the selected chips

Figure 31B : Gene extraction in the two pools of experiments. A total of 1199 common genes have been found.

Genes in experiments from pool 2 are OVEREXPRESSED compared to experiments from pool 1

	Rank	Diff value	CleanEx ID	Description
<input checked="" type="checkbox"/>	1	1169	HS_MUC2	Mucin 2, intestinal/tracheal.
<input checked="" type="checkbox"/>	2	1126	HS_FGF8	Fibroblast growth factor 8 (androgen-induced).
<input checked="" type="checkbox"/>	3	1100	HS_MTA1	Metastasis associated 1.
<input checked="" type="checkbox"/>	4	1095	HS_MTHFD1	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methylenetetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase.
<input checked="" type="checkbox"/>	5	1085	HS_THRA	Thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian).
<input checked="" type="checkbox"/>	6	1073	HS_MT1G	Metallothionein 1G.
<input checked="" type="checkbox"/>	7	1055	HS_XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining; Ku autoantigen, 80kDa).
<input checked="" type="checkbox"/>	8	1042	HS_MYBL1	V-myb myeloblastosis viral oncogene homolog (avian)-like 1.
<input checked="" type="checkbox"/>	9	1041	HS_THRB	Thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian).
<input checked="" type="checkbox"/>	10	1027	HS_TCTEL1	T-complex-associated-testis-expressed 1-like 1.
<input checked="" type="checkbox"/>	11	1022	HS_MST1R	Macrophage stimulating 1 receptor (c-met-related tyrosine kinase).
<input checked="" type="checkbox"/>	12	1020	HS_MPHOSPH1	M-phase phosphoprotein 1.
<input checked="" type="checkbox"/>	13	1017	HS_CYP4F2	Cytochrome P450, family 4, subfamily F, polypeptide 2.
<input checked="" type="checkbox"/>	14	998	HS_MSH2	MutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli).
<input checked="" type="checkbox"/>	15	994	HS_EEF1A1	Eukaryotic translation elongation factor 1 alpha 1.
<input checked="" type="checkbox"/>	16	984	HS_MUTYH	MutY homolog (E. coli).
<input checked="" type="checkbox"/>	17	982	HS_TCEB1	Transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C).
<input checked="" type="checkbox"/>	18	971	HS_TCEA2	Transcription elongation factor A (SII), 2.
<input checked="" type="checkbox"/>	19	969	HS_TBX5	T-box 5.
<input checked="" type="checkbox"/>	20	960	HS_MMP15	Matrix metalloproteinase 15 (membrane-inserted).
<input checked="" type="checkbox"/>	21	950	HS_MMP14	Matrix metalloproteinase 14 (membrane-inserted).
<input checked="" type="checkbox"/>	22	949	HS_CYP27A1	Cytochrome P450, family 27, subfamily A, polypeptide 1.
<input checked="" type="checkbox"/>	23	920	HS_WT1	Wilms tumor 1.

Figure 31C : The top genes found to be overexpressed in metastatic samples compared with non-metastatic samples.

6.6. CleanEx external applications

As mentioned above, all the CleanEx_trg and CleanEx files are available via our ftp server. This, together with the fact that the files are renewed on a regular basis, prompted some people to make use

of CleanEx for external applications. The files which are usually of interest for other applications are the mapping files, especially the ones generated for Affymetrix probe sets. The main reason for using these files is that it is an easy way to retrieve, or concatenate, very precise matches of independent tags on sequences available in general databases, like EMBL or GenBank. Moreover, the exact match positions on the reference sequences are given, together with the orientation of the match. This allows one to have a nice and complete view of each probe set, tag-by-tag, and to have a more precise control on the annotation of the probes. The CleanEx mapping files have been so far successfully used in two independent projects : The ISREC Ontologizer, and the DNA Chip Splice Machine.

6.6.1. IO

IO (ISREC ontologizer : <http://www.io.isb-sib.ch/>) [54, 55] is a program for classifying microarray results in the Gene Ontology. The strong point of the program is that it allows a fine evaluation of the results based on various quality thresholds, in particular on annotation quality. For a given list of differentially regulated probe sets, IO shows their distribution over all GO (Gene Ontology, <http://www.geneontology.org/>) [56] classes subdivided by classes of probe set quality and evaluates the statistical significance of over-representation of a GO class. The advantage of IO compared to most GO classification programs, such as MappFinder or OntoExpress, is that this evaluation can be done not only with the probe sets as individual entities but also by pooling in groups those that represent the same Unigene cluster. This allows one to study the degree of agreement between probe sets of the same cluster. Moreover, IO provides a confidence assessment regarding the significantly regulated functional classes.

IO is actually delivered with a reevaluated set of annotation files for the Affymetrix mouse and human chips. The reevaluation is based on the CleanEx Affymetrix mapping files. The confidence assessment is based on the quality tag given in the CleanEx_trg file.

IO is developed by Thierry Sengstag in collaboration with Pascale Anderle.

6.6.2. DNA Chip Splice Machine

The DNA Chip Splice Machine (<http://bio.ifom-firc.it/AffyDB/>) is a tool which has been developed by Alessandro Guffanti and Davide Rambaldi at the IFOM (Italy). It's main goal is to allow users to

visualize, in a gene-structure orientation, the Affymetrix individual probes of one probe set. The tool makes use of the CleanEx_trg file for Affymetrix. It then links the tags positions on the RefSeq to the mapping of the RefSeq on the genome. It then concatenates the results on a viewer which shows the Affymetrix independent tags mapped on gene's exons. It so far considers only matches on RefSeq, but a new version will be available soon which will include the matches on EMBL RNA sequences, also extracted from the CleanEx database. The authors are in the process of applying this tool to the interpretation of experimental results and linking it directly to GeneSpring, a statistical analysis software for DNA chips.

7. DISCUSSION

7.1. General considerations

7.1.1. CleanEx development decisions

While building an expression database like CleanEx, a few important points have to be taken into account, that could bring major changes to the usability of the database. For example, one needs a reliable system to link data to gene names. When CleanEx began, only a few databases contained sufficiently reliable information about gene names, gene localization, and corresponding sequences. The most famous one at that time was the Unigene database. Another possibility would have been to use the in-house built version of the human transcriptome, called Trome. The final decision to use Unigene was based on the following major points. First, Unigene was very well known all over the world. Having as first reference a universally used database allows users to hang on known information while using CleanEx. Second, the Trome database was built only for human at that time, which would have been problematic for the expansion of CleanEx to other organisms. Nowadays, Trome exists also for other major model organisms. Third, the first goal of CleanEx was not to deal with alternative splicing. Using Unigene was thus easier, as each entry corresponds to one gene, and not to one transcript. In Trome, all the possible transcripts are represented, and a supplementary procedure would have been required to concatenate the information about one gene. This is now also provided in the recent versions of Trome. The other important point is that the Trome database is updated at the same time than the EMBL database, meaning one release every three months, whereas the Unigene database follows a weekly update system. As this is a major point in CleanEx, it seemed important to follow a system with a very frequent update. At present time, a few other organism-specific genomic databases exist, which provide even more detailed information than the Unigene database, and which are also updated very frequently. One could for example consider using the Ensembl database, or directly the Entrez GeneID system

from the NCBI as gene references. This would probably improve the update procedure. Indeed, as the information about a gene is concentrated at one spot, the update will take much less time.

Despite the emergence of brand new gene expression databases, all based on more and more precise international recommendations, the ideal solution has not yet been found. The creation and structure of each such database depends mainly on the use that one intends to make of it. It is thus very difficult to obtain a consensus definition of a good and useful expression database. For example, though designed at the same time, it is striking to see that the structure of databases like GEO or ArrayExpress differ significantly from the CleanEx one. Indeed, though also split in three major file systems, the GEO structure is directly linked to its repository function. The CleanEx design is much more specific to gene information retrieval. The first design attempt for CleanEx was to generate only two file types. The first one contained the expression experiments, while the other one was intended to serve as gene data retrieval and link to expression data. This kind of simple design was efficient enough for dual channel experiments. The integration of other data types raised new problems, which were not easily solvable by keeping that kind of structure. For example, in the Affymetrix experiments, one probeset is represented by more than one sequence on the chip. This change in the relationship between spotted sequences and the numerical values obtained, going from a “one to one” relationship to a “many to one”, needed a new way of storing the data. The decision was thus taken to generate an intermediate file for target information (the CleanEx_target file type). This structure change was also justified by the fact that, while adding new datasets in CleanEx, the same sequences are sometimes reused in new experiments. Keeping the two-files system would then have greatly increased the redundancy in the database, while the intermediate file can contain one entry per sequence, regardless of the number of datasets which have used it. The creation of this third file also allowed us to store more precise information for sequences which were mapped with the tagger, like the position of the spotted sequence on the reference sequence, the cases of multiple hits, as well as the quality tag for each sequence.

Actually, the major databases are mainly data repositories, meaning storage facilities, linked to a few analysis tools. GEO, ArrayExpress, or even the Stanford Microarray Database are the most famous amongst these ones. From the structure of this kind of databases, consisting of a split between series, samples, and platforms, it is quite difficult to retrieve gene-centered expression information, or to compare the expression level of a few genes at the same time. Moreover, the data annotation is not

trivial and the link between numerical data of expression measurements and the actual gene which is over- or under-expressed in the specific tissue is not always well defined. In that sense, the CleanEx database structure is an alternative solution which generates coherent and biologically understandable information about gene expression. The table 2 gives a comparison between the features present in CleanEx and in the other major expression databases.

	<i>ArrayExpress</i>	<i>GEO</i>	<i>GeneCards</i>	<i>SOURCE</i>	<i>CleanEx</i>
<i>Dataset Upload</i>	😊	😊	😞	😞	😞
<i>Dual-channel</i>	😊	😊	😊	😊	😊
<i>Affymetrix</i>	😊	😊	😊	😊	😊
<i>SAGE and/or</i>	😞	😊	😞	😞	😊
<i>ESTs</i>	😞	😞	😊	😊	😊
<i>Single dataset</i>	😊	😊	😊	😊	😊
<i>cross dataset</i>	😞	😞	😞	😊	😊
<i>Genes-oriented</i>	😞	😊	😊	😊	😊
<i>Show clones per</i>	😞	😞	😞	😊	😊
<i>Show tags per</i>	😞	😞	😞	😞	😊
<i>Sequence</i>	😞	😞	😞	😊	😊

Table 2 : Database comparison. Pink “smileys” indicate that the corresponding database has the selected feature.

Blue frowns show a lack of this feature.

7.1.2. Linking expression data to promoter analysis

Nowadays, as the amount of publicly available data is increasing, new databases appear, which try to link biological interpretation with heterogeneous expression or genomic data. As expression data become also more precise, researchers want to push the analysis towards the discovery of new regulatory elements. The idea of finding common promoter elements in co-expressed genes is very tempting [57, 58, 59], and this feature is now in a trial phase in some newly generated expression databases [60, 61, 62, 63]. Anyway, finding motifs in promoter sequences is a huge problem *per se*. For

example, one should not forget that the transcription start site is quite often not well defined, or even unknown, for many genes. Sometimes, genes have also alternative transcription start sites, which are used in different tissues, or under different conditions. It is thus very difficult to determine the exact position of a motif, relative to the transcription start site, on the genomic sequence. Moreover, it might be the relationship between different motifs, and not always the position of the motif itself, which could influence the transcription level of the gene. One thus needs a tool which would be able to recognize “metamotifs” instead of single motifs, meaning a sequence, or a suite, of conserved motifs, in different promoter regions. The solution chosen in the CleanEx system, the link to the SSA server, was dictated by two major points. First, all the SSA tools have been generated and set up in-house, and they can thus be easily tailored for our purpose. Second, these tools are based on a precise alignment of the sequences around a defined site (here the transcription start site), information which is available and easily retrievable for a few genes in the other in-house database called Eukaryotic Promoter Database. By using the position information of EPD to align the genomic sequences of co-expressed genes, we are able to study some basic information regarding the promoter sequence, like extracting regions showing a non random distribution of nucleotides, or finding the percentage of sequences containing known binding motifs. This is a first step in promoter analysis. If one would like to go ahead in that way, one would need to adopt a more general view on the existing motifs and their relationships (like position and distance regarding the other motifs, number of motif occurrence...) between the promoters of the co-expressed genes. One way to do it would be to link the first results obtained by SSA about non-random sequences in the promoters with a new metamotifs tool. One would also need to take into account all the possible alternative transcription start sites of the co-expressed genes as an attempt to determine which one is used under which condition or in which sample.

7.2. Advantages and drawbacks of CleanEx

CleanEx combines the use of sequence annotation and expression data by linking a precise and up-to-date target annotation database with a powerful expression data retrieval system.

CleanEx is indeed a very powerful tool for gene-oriented expression data retrieval and analysis. By using a simple web-based tools system, the user can directly access a complete expression viewer

showing very heterogeneous experiments and various conditions for the gene of interest. This viewer will provide the user with general information on his gene, like it's name, description, corresponding genomic, transcriptomic or proteomic sequences, as well as gene expression information coming from very heterogeneous experiment types at a glance. Moreover, the multiviewer will then lead the user to a possible comparison of different clones corresponding to the same gene in one experiment, and will thus give a hint on either the quality of the experiment, or even the possibility of alternative splicing occurrences for this gene. This multiviewer for expression data can also give users quick clues on how to pursue their researches. Though this approach has been already used in other databases (S.O.U.R.C.E, GeneCards), CleanEx is so far the only database which includes results coming from protocols as different as SAGE, Affymetrix, Dual-channel, and EST counts. It is one of the rare databases which allows not only expression data retrieval from these heterogeneous techniques, but which also provides gene-centered information about all possible features from these heterogeneous sources (SAGE tags, Affymetrix probe sets, as well as clones from Unigene).

Regarding the comparison of a gene set between heterogeneous expression datasets, the two tools accessible via the CleanEx web server have two significantly different functions, though they are both based on cross-dataset comparison. The first one, the step-by-step tool, allows comparison of expression levels in different datasets, meaning in data which have been generated using different techniques, but which address a closely related question, like comparable tissue types. The question raised by this tool is : how coherent and how comparable are expression results if they come from different sources ? Do we retrieve the same genes in the over-expressed set ? By applying this step-by-step method, one obtains a first clue on the comparability level of the selected datasets. One could also use this kind of tool to orient or refine the design of a new experiment. This comparison tool is fully functional. As it is based on mean difference ranking, it works especially well for comparing highly differentiated expression levels in two experiment pools, for example high versus low expression. It still needs some more powerful statistical tools for the retrieval of genes which share a common expression pattern in the two experiments pools.

The second comparison tool, the comparison of expression levels between two different biological categories, gives accurate results and allows the discovery of highly specific genes via a very simple

interface. This tool is meant to provide expression measurements analysis, but can also be used to retrieve common genes from different platforms. The experiment classification also allows the retrieval of the name and original dataset of comparable experiments in a big pool of heterogeneous datasets. The by-class cross-dataset analysis tool still needs further developments. The major problem is to generate a list of biological classes and to attribute these classes to the integrated datasets. This takes time, as the CleanEx database does not provide a list of keywords for each experiments. In fact, the creation of a biological keywords list, based on a universal controlled vocabulary, as for example the GO (Gene Ontology) system, and its integration and indexation for each dataset would probably increase the analysis capacity as well as the accuracy of this tool.

The basic structure of CleanEx, meaning the split by of the data in three different files according to their type, not only allows partial update of the database, but also increases the search speed and data retrieval via the common unique identifier which links these three files. The semi-automatic procedure for GEO datasets already increased considerably the number of integrated SAGE and Affymetrix datasets. This procedure will be run regularly, as a way to retrieve newly uploaded datasets, and integrating new data will shortly become a fully automatic procedure, either from GEO or from other web sites, with different options according to the raw data format.

The most important and yet most useful part in CleanEx is still the CleanEx_trg file, which provides links between genes and features found in the expression experiments. This link clearly appears to be missing or incomplete for many features. The CleanEx procedure, which provides precise and individual mapping results, is an easy and fast way to solve this lack of information. The Affymetrix re-annotation files are probably the best example of this kind of information. The use of these files by external developers is a very encouraging step for maintaining this procedure in CleanEx. In fact, the probe-to-gene files could even be used to discriminate between splice variants spotted on the chips.

Still there are two major drawbacks linked to CleanEx. The first one is, as mentioned above, that the cross-dataset tool needs much more solid statistical tools for more precise comparison. The other important feature to increase the cross-experiment comparison precision is to do the analysis on a bigger data sample. To achieve this, the dataset integration system in CleanEx has to be improved. On

one side, the automatic procedure will help solving this problem, but on the other side, if the number of integrated datasets will of course increase, the experiments range, meaning the number of different experiment types in CleanEx, will increase as well. This will in the end only generate more experiment's classes, and will not increase the number of experiments per class. In that sense, this automatic procedure will not increase the accuracy of the statistical comparisons. One way to bypass this problem could be to filter the datasets' integration by keeping only a few classes representing very specific problems (like cancer and tissue type, or survival analysis), and to temporarily leave the other datasets aside. To achieve this, one should take into account the further described proposition of creating an indexed keywords system based on a controlled vocabulary.

8. FUTURE DEVELOPMENTS

8.1. Web interfaces

8.1.1. Single CleanEx entries

To improve the CleanEx entry interface and to facilitate data extraction from this page, the accent will be put on classifying the expression links. So far, it could be quite difficult to understand and retrieve the appropriate information, due to the increasing number of expression links for each entry. A good solution to remedy this lack of clarity will be to put the expression links on another page. These results could be pre-selected by class on the CleanEx gene entry, and then displayed. This system will enable a short description of the linked datasets to be added to this intermediate page, when keeping the results page short and readable.

8.1.2. Targets and annotation retrieval data

The target page itself does not need to be modified significantly. However, numerical results of the tags and probes mapping, though very useful, are not that easy to read. Linking these numbers to a basic viewer, as done by the DNA Chip Splice Machine for Affymetrix probes, will help users to interpret the data. This will be the page corresponding to the CleanEx_exp single viewer. If one thinks of a correspondence for the CleanEx_exp multiviewer, one could end up with the creation of a general viewer giving the accurate position of any possible feature present in any CleanEx dataset on the gene and, further on, on the chromosome. This kind of representation already exists, to a certain extent, for example on the Ensembl gene viewer, though it does not give access to either individual probes positions or SAGE or MPSS tags. Having a general view with clones, Affymetrix probe sets, as well as SAGE tags on one single page could lead to more detailed interpretation of discrepancies in the results obtained by different techniques, as previously shown with the T0001 and GDS181 datasets.

8.2. Expression data analysis

So far, the step-by-step analysis is done without any normalization procedure, and as said before, with quite weak statistical tools. One way to improve this method will be to apply a more powerful analysis

procedure, like a Student's t-test, to the two datasets pools. Of course, using more complex tools implies a loss in the procedure speed. As one could be interested in quickly generated and general results, in the future, a choice of different statistical methods to compare the two pools of experiments will be given to the user.

8.3. Update, database formats and database growth

8.3.1. Update procedure : towards a new database format ?

The split into three different files for CleanEx is already a very helpful step to decrease the time taken by the update procedure. The whole database has to be rebuilt entirely at each release anyway . There is a file structure which avoids this so-called “from scratch” build, which is a relational database system. By using such a database format, one would be able to use an incremental update system, thus updating only entities which have changed in between two releases. This will be especially efficient for the most time-consuming parts of the update, meaning the Affymetrix and single tags mapping procedure. The relational database interface also comprises a very complex and fast query language system which will allow an even faster entry retrieval from the CleanEx tables. Moreover, it will be easy as well to rebuild the three original flat files from these tables, so that one can still have access to the old CleanEx format. This could be useful for local batch experiments, for example. Given all these considerations, it would be very interesting to make a trial relational version of CleanEx to see how much time would be gained, for the update as well as for the on-line query retrieval system.

8.3.2. MAGE-ML : giving access to raw data in standard exchange format

As mentioned before, the first function of CleanEx is not raw expression data retrieval, and thus does not need to be fully MGED compatible. Nevertheless, in a near future, we plan to give public access to the raw expression data so that people could redo their own analysis. As standard formats now exist, we will have to think about creating an interface capable of recreating the standard format, as implemented in the Stanford Microarray Database. The MGED Society, via the MAGE_stk, provides a great number of scripts to allow this procedure, and it should thus be quite feasible to build such a tool and to provide raw data in an MAGE_ML compatible format.

8.3.3. New datasets incorporation : adapting the GEO automatic procedure

As a way to increase the number of available data in CleanEx, a new automatic procedure will be set up. As all data published in major journals are now available in one of the three official data repositories, the automatic data incorporation implemented for GEO will be modified to fit the two other officially approved databases, namely ArrayExpress and CIBEX. The formerly explained procedures to integrate Series, Samples, and Platforms from GEO will be adapted to the Experiment, Array, and Protocol level organizations of the ArrayExpress database respectively . To avoid fuzziness in the data, a filter will allow data selection, based on the dataset description, according to the chosen centers of interest for CleanEx.

9. REFERENCES

1. **Galperin MY.** (2005) The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res.* **33**: D5-24.
2. **Alwine, JC, Kemp, DJ, and Stark, GR** (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization of DNA probes. *Proc. Natl. Acad. Sci. USA* **74**: 5350-5354.
3. **Schena, M., Shalon, D., Davis, R. W., and Brown, P. O.** (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.
4. **Velculescu VE, Zhang L, Vogelstein B, Kinzler KW.** (1995) Serial analysis of gene expression. *Science* **270**: 484-487.
5. **Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K.** (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18**: 630-634.
6. **Ewing, R. M., and Claverie, J. M.** (2000). EST databases as multi-conditional gene expression datasets. *Pac Symp Biocomput*, 430-442.
7. **Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ.** (1999) High density synthetic oligonucleotide arrays. *Nat Genet.* **21**: 20-24.
8. **Praz V, Jagannathan V, Bucher P.** (2004) CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. *Nucleic Acids Res.* **32**: D542-7.
9. **Lander ES et al.** (2001) Initial sequencing and analysis of the human genome.", *Nature* **409**: 860-921
10. **Jongeneel CV.** (2000) Searching the expressed sequence tag (EST) databases: panning for genes.

Brief Bioinform. **1**: 76-92.

11. **Wang Z, Chen Y, Li Y.** (2004) A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics* **2**: 216-21.
12. **Pennisi E.** (2003) Human genome. A low number wins the GeneSweep Pool. *Science* **300**: 1484.
13. **Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L.** (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**: 28-33.
14. **Pruitt KD, Tatusova T, Maglott DR.** (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**: D501-504.
15. **Sperisen P, Iseli C, Pagni M, Stevenson BJ, Bucher P, Jongeneel CV.** (2004) trEMBL, trEST and trGEN: databases of predicted protein sequences. *Nucleic Acids Res.* **32**: D509-11.
16. **Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM.** (2001) The Stanford Microarray Database. *Nucleic Acids Res.* **29**: 152-155.
17. **Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP.** (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics.* **18**: 1593-9.
18. **Hubbell E, Liu WM, Mei R.** (2002) Robust estimators for expression analysis. *Bioinformatics.* **18**: 1585-92.
19. **Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP.** (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249-64.
20. **Brazma A, Robinson A, Cameron G, Ashburner M.** (2000) One-stop shop for microarray data. *Nature* **403**: 699-700.
21. **Brazma A.** (2001) On the importance of standardisation in life sciences. *Bioinformatics* **17**: 113-114.

22. **Causton HC, Game L.** (2003) MGED comes of age. *Genome Biol.* **4**: 351.
23. **Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M.** (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet.* **29**: 365-71.
24. **Gardiner-Garden M, Littlejohn TG.** (2001) A comparison of microarray databases. *Brief Bioinform.* **2**: 143-58.
25. **Aach J, Rindone W, Church GM.** (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.* **10**: 431-445.
26. **Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA.** (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**:68-71.
27. **Edgar R, Domrachev M, Lash AE.** (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**: 207-210.
28. **Boyle J.** (2005) Gene-Expression Omnibus integration and clustering tools in SeqExpress. *Bioinformatics* **21**: 2550-2551
29. **Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N.** (2004) Submission of microarray data to public repositories. *PLoS Biol.* **2**: E317.
30. **Anderle P, Duval M, Draghici S, Kuklin A, Littlejohn TG, Medrano JF, Vilanova D, Roberts MA.** (2003) Gene expression databases and data mining. *Biotechniques Suppl*: 36-44
31. **Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G.** (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic*

- Acids Res. **33**: D580-582.
32. **Killion PJ, Sherlock G, Iyer VR.** (2003) The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics* **4**: 32.
 33. **Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD.** (2000) The cancer genome anatomy project: building an annotated gene index. *Trends Genet.* **16**: 103-106.
 34. **Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF.** (2000) SAGEmap: a public gene expression resource. *Genome Res.* **10** :1051-1060.
 35. **Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y.** (2003) CIBEX: center for information biology gene expression database. *C R Biol.* **326**: 1079-1082.
 36. **Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D.** (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* **14**: 656-664.
 37. **Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA.** (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31**: 219-223.
 38. **Kent WJ.** (2002) BLAT--the BLAST-like alignment tool. *Genome Res.* **12**:656-64.
 39. **Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S.** (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.* **32**: D255-257.
 40. **Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS.** (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**: D154-159.
 41. **Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA.** (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**: D514-517.
 42. No authors listed. (2003) HUGO--a UN for the human genome. *Nat Genet.* **34**: 115-116.
 43. **Boguski MS, Lowe TM, Tolstoshev CM.** (1993) dbEST--database for "expressed sequence tags".

- Nat Genet. **4**:332-333.
44. **Praz V, Perier R, Bonnard C, Bucher P.** (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.* **30**: 322-4.
45. **Schmid CD, Praz V, Delorenzi M, Perier R, Bucher P.** (2004) The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res.* **32**: D82-85.
46. **Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM, Boddy WJ, Bradt DW, Burkart DL, Butler NE, Campbell J, Cassell MA, Corbani LE, Cousins SL, Dahmen DJ, Dene H, Diehl AD, Drabkin HJ, Frazer KS, Frost P, Glass LH, Goldsmith CW, Grant PL, Lennon-Pierce M, Lewis J, Lu I, Maltais LJ, McAndrews-Hill M, McClellan L, Miers DB, Miller LA, Ni L, Ormsby JE, Qi D, Reddy TB, Reed DJ, Richards-Smith B, Shaw DR, Sinclair R, Smith CL, Szauter P, Walker MB, Walton DO, Washburn LL, Witham IT, Zhu Y; Mouse Genome Database Group.** (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res.* **33**: D471-475.
47. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.
48. **Lennon G, Auffray C, Polymeropoulos M, Soares MB.** (1996) The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **33**: 151-152.
49. **Maglott D, Jim Ostell J, Pruitt KD, Tatusova T.** (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**: D54-8.
50. **Pruess M, Kersey P, Appweiler R.** (2004) The Integr8 project - a resource for genomic and proteomic data. *In Silico Biol.* **5**: 0017
51. **Ambrosini G, Praz V, Jagannathan V, Bucher P.** (2003) Signal search analysis server. *Nucleic Acids Res.* **31**: 3618-20.
52. **Bucher P, Trifonov EN** (1986) Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res.* **14**: 10009-10026.
53. **Rakha A E, Boyce W G R, Abd El-Rehim D, Kurien T, Green A R, Paish E C, Robertson J F**

- R, O Ellis Y.** (2005) Expression of mucins (MUC1, MUC2, MUC3, MUC4, MUC5AC and MUC6) and their prognostic significance in human breast cancer.
54. **Sengstag T, Anderle P, Praz V, Bucher P, Delorenzi M.** Io: A graphical Gene Ontology classifier with a quality-based reannotation of Affymetris chips (In preparation)
55. **Anderle P, Sengstag T, Mutch DM, Rumbo M, Praz V, Mansourian R, Delorenzi M, Williamson G, Roberts MA.** (2005). Changes in the transcriptional profile of transporters in the intestine along the anterior-posterior and crypt-villus axes. *BMC Genomics* **6**: 69.
56. **Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium.** (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**: D258-61.
57. **Roth FP, Hugues JD, Estep PW, Church GM.** (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol.* **16**: 939-45.
58. **Brazma A, Jonassen I, Viol J, Ukkonen F.** (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* **8**:1202-15.
59. **Park PJ, Butte AJ, Kohane IS.** (2002). Comparing expression profiles of genes with similar promoter regions. *Bioinformatics.* **18**:1576-84.
60. **Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R.** (2005). EXPANDER--an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**: 232.
61. **Wrobel G, Chalmel F, Primig M.** (2005). goCluster integrates statistical analysis and functional

interpretation of microarray expression data. *Bioinformatics* **21**: 3575-7.

62. **Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, Engelen K, Glenisson P, Moreau Y, Mathys J, De Moor B.** (2003). INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res* **31**: 3468-70.
63. **Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B.** (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res* **31**: 1753-64.

10. SUPPLEMENTARY TABLES

CHIP	Total number of probesets	Matches from RefSeq			Supplementary matches from RNA			Supplementary matches from HTC			Supplementary matches from EST_PLUS			Supplementary matches from EST_MINUS			Probesets with no match
HC-G110	1887	1494			337			26			15			15			89
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		1254	48	103	320	6	11	16	8	2	6	5	4	0	1	2	
HG-Focus	8793	7261			1219			110			123			80			172
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		6680	208	201	1180	25	14	92	16	2	79	22	22	0	21	11	
HG-U133A_2	22277	16606			4363			434			478			396			1545
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		13915	487	659	3940	155	268	356	57	21	291	85	102	0	108	76	
HG-U133A	22283	16853			4420			446			568			0			1723
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		13915	505	710	3940	177	299	356	63	27	291	122	155	0	0	0	
HG-U133B	22645	13569			6171			908			2001			0			7683
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		5145	389	352	4798	778	591	666	174	68	733	534	734	0	0	0	
HG-U133_Plus_2	54675	33994			15620			2297			2764			0			10945
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		21006	937	1106	13379	1139	1102	1924	260	113	1118	693	953	0	0	0	
HG-U95A	12454	9570			2236			313			335			0			799
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		7519	552	700	1919	142	175	241	49	23	180	64	91	0	0	0	
HG-U95Av2	12453	9569			2238			312			334			0			799
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		7517	553	700	1921	142	175	240	49	23	180	64	90	0	0	0	
HG-U95B	12620	7781			2793			418			709			919			3479
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		3517	490	295	2119	420	254	313	71	34	352	134	223	0	225	172	
HG-U95C	12646	8532			2243			387			690			794			5304
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		2231	360	637	1431	342	470	260	69	58	253	144	293	0	165	219	
HG-U95D	12644	9149			1679			272			610			934			7457
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		850	238	604	872	317	490	146	51	75	212	119	279	0	206	302	
HG-U95E	12639	7964			2361			419			834			1061			4888
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		1952	472	652	1468	403	490	245	84	90	337	161	336	0	253	352	
U133_X3P	61359	36277			17138			2526			2455			2967			10855
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		23873	662	887	15555	777	802	2258	172	96	1518	399	538	0	682	561	

Table 3 : Quality gain by introduction of matches on other databases for Human chips

CHIP	Total number of probesets	Matches from RefSeq			Supplementary matches from RNA			Supplementary matches from HTC			Supplementary matches from EST_PLUS			Supplementary matches from EST_MINUS			Probesets with no match
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
MG-U74A	12654	9067			1732			741			810			307			2962
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		5167	313	625	1484	85	160	525	75	141	534	81	195	0	48	76	
MG-U74Av2	12488	8754			1907			770			761			299			1733
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		5707	342	972	1619	94	191	592	78	100	581	80	100	0	45	49	
MG-U74B	12636	6608			1442			1688			1777			1124			3538
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		2414	335	321	1192	148	99	1172	238	278	1062	315	400	0	216	187	
MG-U74Bv2	12477	5842			1726			1841			1849			1222			1715
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		3089	414	624	1402	177	144	1330	253	258	1191	332	326	0	226	173	
MG-U74C	12728	9112			321			869			809			1620			8341
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		459	66	246	207	33	78	490	93	286	500	63	246	0	50	116	
MG-U74Cv2	11934	7163			636			1516			790			1832			5204
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		932	186	841	351	68	214	744	172	600	536	64	190	0	69	140	
MOE430A	22690	15678			3766			1662			1169			418			1188
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		13380	627	483	3486	145	132	1427	134	101	846	171	152	0	86	66	
MOE430B	22575	9680			1796			7237			2020			1845			5808
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		3173	290	409	1556	120	117	5822	1010	405	961	593	466	0	540	335	
Mouse430A_2	22690	15678			3766			1661			1169			419			1188
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		13380	628	482	3486	146	131	1427	133	101	846	170	153	0	87	66	
Mouse430_2	45101	25225			5547			8888			3183			2261			6952
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		16475	917	881	5028	265	251	7242	1139	507	1803	763	617	0	627	400	
Mu11KsubA	6584	4546			1059			374			454			154			804
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		3125	232	385	919	72	65	240	51	83	322	46	86	0	31	37	
Mu11KsubB	6002	4724			738			166			339			38			1388
		High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low	
		2041	267	1028	579	62	94	81	29	56	181	59	99	0	7	12	

Table 4 : Quality gain by introduction of matches on other databases for Mouse chips