Genet. Sel. Evol. 30 (1998) 521–541

Review

Expressed sequence tags for genes: a review

François Hatey *, Gwenola Tosser-Klopp, Catherine Clouscard-Martinato, Philippe Mulsant, François Gasser

Laboratoire de génétique cellulaire,
Institut national de la recherche agronomique,
BP 27, 31326 Castanet-Tolosan cedex, France

(Received 16 September 1997; accepted 7 July 1998)

Abstract – Expressed sequence tags (ESTs) are partial sequences from the extremities of complementary DNA (cDNA) resulting from a single pass sequencing of clones from cDNA libraries, and different ESTs can be obtained from one gene. Sequence information from ESTs can be used for deciphering the function and the organisation of the genome. From a functional viewpoint, they allow the determination of the expression profiles of genes in any particular tissue, in different conditions or status, and thus the identification of regulated genes. In order to identify genes involved in particular processes one can select a specific group of mRNAs. For such a selection, classical techniques include subtraction or differential screening and new techniques, using polymerase chain reaction (PCR) amplification, are now available. For studies on the organisation of the genome the main use of ESTs is the determination of chromosomal localisation of the corresponding genes using a somatic hybrid cell panel. This chromosomal localisation information is needed to identify genes or quantitative trait loci, according to the 'positional candidate' approach. ESTs also contribute to comparative genetics and they can help to decipher gene function by comparison between species, even genetically distant ones. Thus, combining sequence, functional and localisation data, ESTs contribute to an integrated approach to the genome.

expressed sequence tags / functional genomics / gene mapping / comparative genetics

Résumé – **Des étiquettes pour les gènes : une revue.** Les « étiquettes » correspondent aux séquences des extrémités des ADN complémentaires, obtenues de manière systématique à partir d'une seule réaction de séquençage. Cependant, à partir d'un

* Correspondence and reprints
E-mail: hatey@toulouse.inra.fr

seul gène plusieurs étiquettes différentes peuvent être obtenues : celles qui correspondent aux deux extrémités de l'ADN complémentaire, aux ADN complémentaires de tailles différentes synthétisés à partir d'un même ARN messager, et aux différents ARN messagers issus d'une même séquence d'ADN génomique. L'identification des gènes correspondants est faite par comparaison avec les séquences nucléiques ou protéiques contenues dans les bases de données publiques (GenBank ou EMBL, SwissProt), en utilisant des logiciels d'alignement automatique tels que FASTA ou BLAST. Les séquences annotées des étiquettes sont stockées dans une base de données particulière, dbEST, et soumises régulièrement à des tests de comparaison avec les bases de données citées. En raison de la présence d'une longue région non codante à l'extrémité 3' des ARN messagers, les étiquettes de l'extrémité 3' sont souvent non informatives. La comparaison des étiquettes entre elles permet d'essayer de regrouper celles qui peuvent appartenir à un même gène et de déterminer ainsi une séquence consensus, plus longue et donc plus informative. Au niveau fonctionnel, les étiquettes permettent d'établir les profils d'expression des gènes d'un tissu donné dans différentes situations physiologiques ou expérimentales et donc d'identifier les gènes qui sont régulés. Ces profils sont établis en utilisant les étiquettes pour mesurer la fréquence des différents ADNc dans une génothèque préparée à partir de ce tissu dans les différentes conditions étudiées. Dans une nouvelle stratégie, la SAGE (*Serial Analysis of Gene Expression*), des étiquettes d'une dizaine de nucléotides sont collectées, mises bout à bout et séquencées en série, ce qui permet d'accélérer l'acquisition de ces profils d'expression. Une autre approche est basée sur l'hybridation d'un grand nombre de clones déposés sur une même membrane en nylon « filtres haute densité », ou, dans un format miniature, sur une lame de verre, « *microarrays* ». Pour identifier les gènes impliqués dans des processus bien définis, différentes stratégies de soustraction ou de comparaison permettent de sélectionner une population particulière d'ARN messagers ; les techniques les plus récentes utilisent l'amplification par PCR. Au niveau de l'organisation du génome, les étiquettes contribuent au développement de la cartographie génique : les gènes correspondants sont localisés en utilisant un panel d'hybrides somatiques, les amorces nécessaires pour amplifier l'ADN des hybrides sont choisies grâce aux informations de séquence fournies par les étiquettes. Cette information de localisation chromosomique est indispensable pour identifier les gènes responsables des caractères étudiés par une stratégie de gène candidat positionnel. L'utilisation d'étiquettes d'une autre espèce peut également permettre d'effectuer ces localisations et donc de développer des cartes comparées entre espèces qui mettent en évidence une certaine conservation de l'organisation des gènes sur les chromosomes. Enfin, la conservation des gènes n'est pas limitée à la séquence et à l'organisation : grâce aux étiquettes, des analogies fonctionnelles de gènes appartenant à des espèces génétiquement éloignées ont été décrites et sont recherchées systématiquement pour identifier la fonction des gènes. Ainsi, en permettant de combiner des données de séquence, d'expression et de localisation chromosomique, les étiquettes participent au développement d'une approche intégrée du génome.

étiquette / génomique fonctionnel / cartographie des gènes / génétique comparée

1. INTRODUCTION

The identification of genes controlling economically important traits provides the basis for new progress in genetic improvement of livestock species, complementing traditional methods based only on measured performance. The identification of these genes, either major genes or quantitative trait loci (QTL), directly affecting variability in traits to be improved, is thus an objective to be

pursued, even though the use of linked genetic markers is an effective interim solution [36]. The search for such genes has long been based on fundamental knowledge of physiology, biochemistry or pathology which can lead to direct specification of 'candidate' genes. Today, thanks to the development of genetic maps, the genes controlling such characters can be located by the approach of 'positional cloning' based on the search for markers enclosing the gene more and more closely [26].

Such location assumes the establishment of a genetic map by study of segregation of markers over several generations, and of a cytogenetic map by determination of the positions of the markers on the chromosomes. The markers used are based on DNA polymorphisms: RFLP (restriction fragment length polymorphism) and repetitive sequence polymorphism (minisatellites and microsatellites). Microsatellites, highly polymorphic and distributed throughout the genome, have led to a remarkable advance in gene mapping: there were, in 1987, 42 markers in the pig, of which 20 gene markers were distributed in seven linkage groups and 22 genes were localised [31]. Less than 10 years later, the latest American map [79] which integrates the European [11] and Scandinavian [33] maps, covers the pig genome with an average interval of about 2 cM; it was established with 1 042 loci of which almost 1 000 were microsatellites. However, these microsatellites are without known function, and the sequences used to identify them are poorly conserved so that the information obtained with one species cannot be transposed to another, and sometimes not even from one population to another of the same species.

The combination of two pieces of information, that is the study of the colocation of a gene identified by genetic methods and of a candidate designated by knowledge of physiology or pathology, is defined as the 'positional candidate' approach [14]: if, for a particular trait, the genetic linkage data implicate a specific region of a chromosome, the genes located in this region are therefore candidates for the character. Their role in the variation of the character considered ought then to be analysed with, on the one hand, the identification of a genetic polymorphism in the populations and, on the other, a functional analysis.

The identification of candidate genes influencing important traits is approached through complementary DNA (cDNA), 'copies' of messenger RNA (mRNA). Devoid of intronic and intergenic sequences whose biological significance is still obscure, these mRNAs represent only a small percentage of the total genome (about 3 % in mammals); by contrast, they contain the great majority of information since they correspond to the proteins expressed in different tissues, the proteins responsible for the identity of these tissues. The formal identification of genes proceeds by sequencing, but since there are some 50 000 to 100 000 genes in the mammalian genome this is still a tedious task. An alternative approach is to sequence only fragments of these cDNAs: new muscular proteins have thus been identified by sequencing 178 different cDNA fragments (approximately 250 bp) from a cDNA library of rabbit muscle [75]. The development of techniques of molecular biology, particularly in the field of sequencing, has helped to make this approach much more accessible. Thus the sequencing of the ends of cDNA from different libraries, on 200 to 400 bases – the average number of bases read on a sequencing gel – allows different

transcripts to be identified. Such expressed sequence tags (ESTs) can thus be obtained in a systematic manner [2].

We address properties of these tags in the first part of this review before considering, in the second part, their use in the functional domain; in the third part, we consider their use in genetics. Most results in this area have been obtained in man, and we will refer most often to this work. We will also use illustrations taken from animals, in particular the pig, since our laboratory has worked on the establishment of the genetic map and on the genetic analysis of ovarian function in that species.

2. TAGS TO IDENTIFY GENES

2.1. One gene, several tags

The mature messenger RNA molecule is asymmetric: the 5' end is characterised by a particular structure, the 'cap', and the 3' end is prolonged by a poly(A) sequence of 20 to 200 residues; this sequence is often used to bind, by hybridisation, a complementary oligo(dT), which serves as a primer for the synthesis of cDNA. In addition, a portion of variable length at each of the two ends occurs either upstream of the initiation codon or downstream of the stop codon and is not translated into protein; these are the untranslated, 'non-coding' regions (*figure 1A*).

2.1.1. Several complementary DNAs for one RNA

Because of the frequent presence of secondary structures which block reverse transcription to a variable extent, a messenger RNA can lead to different incomplete cDNAs: a single transcript can then give cDNAs of different lengths whose 5' end is in the coding region; however, these different cDNAs initiated at the poly(A) have the same 3' end, which allows recognition of those derived from the same messenger RNA. cDNA synthesis can also be initiated using random oligonucleotides which hybridise at different sites inside the coding sequence of the mRNA molecule; the different cDNAs obtained can thus overlap. Different tags can therefore be obtained starting from a single type of messenger RNA (*figure 1B, C*).

These tags permit identification of the corresponding messenger RNA by comparison with the sequences of public databases for nucleic acid sequences (GenBank/EMBL) or protein sequences (SwissProt). The identification is made essentially by linking the coding sequence with known proteins; the non-coding regions are thus a priori less useful. However, with such tags from the 3' end, Okubo and colleagues [69] were able to identify 22 % of their 984 clones, although their sequences were voluntarily short (270 bp on average). At the 5' end, the untranslated regions are shorter and the tags thus have a high probability of corresponding to coding sequences.

Since a single transcript can give several cDNAs and thus different tags, it is important to try to identify those which belong to the same clone or the same transcript in order to cluster them and to try to obtain a longer sequence (THC, Tentative Human Consensus, [7]; Unigene, [22]; Merck Gene Index, [1]). This clustering is achieved by means of the comparison program BLAST [9], and new

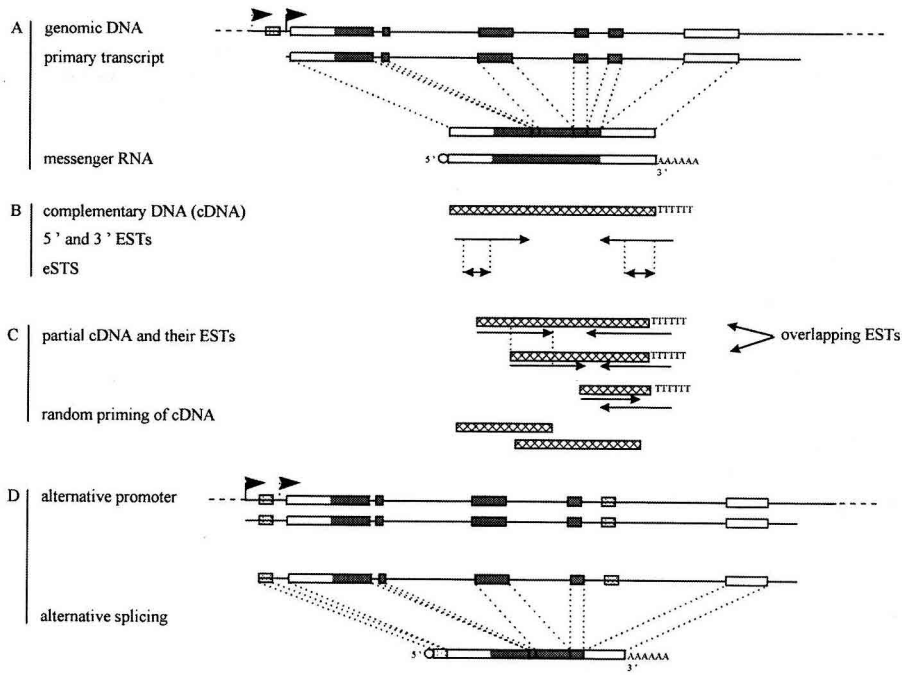

Figure 1. A) At the level of genomic DNA, a gene consists of exons (blank or grey rectangles) and of introns (solid lines). The primary transcript contains these two types of elements. It then undergoes a process of splicing in the course of which the introns are eliminated and the exons directly linked together. The mature messenger RNA undergoes two further modifications, the addition at the 5' end of a particular structure, the cap (open circle) and, at the 3' end, the addition of a sequence of several tens of A residues. B) Complementary DNA is synthesised starting with an oligo(dT) which hybridises to the poly(A) sequence of the messenger, and the sequencing of the extremities of this cDNA produces the ESTs, from which one can choose the sequences used as markers on the chromosomes (eSTS). C) Starting with a single messenger RNA, in particular because of the presence of secondary structures, incomplete cDNAs are often obtained, leading to different tags at the 5' end. These tags may overlap, enabling the construction of a consensus sequence. Similarly, cDNA can be synthesised starting with random primers, and many different cDNAs, perhaps overlapping, are obtained from the same transcript. D) From a single genomic DNA sequence, different messenger RNAs can be obtained, through use of alternative promoters, by alternative splicing, or by use of an alternative polyadenylation site (not shown). These different mechanisms may occur together.

computer programs have been developed [24, 48]. Similarly, various programs are available for the construction of consensus sequences [16]; Web server: see *table I*). However, manual validation is indispensible to take account of different possible sources of error, such as sequencing errors or alternative splicing (see later). Such an analysis has been carried out for the tags obtained from libraries of cDNAs from human muscle and brain [46]: about 19 000 sequences were clustered into nearly 7 000 families.

Table I. WEB sites.

General servers	
NCBI	http://www.ncbi.nlm.nih.gov/
Cancer Anatomy Genome Project	
dBest	
Online Mendelian Inheritance in Man (OMIM)	
Unigene	
TIGEM	http://www.tigem.it/TIGEM/INFOCORE/Infocore.htm
consensus sequence construction	
Drosophila-Related Expressed Sequences	
Large EST projects	
BodyMap	http://cookie.imcb.osaka-u.ac.jp/bodymap/
"IMAGE" project (Integrated Molecular Analysis of Genomes and their Expression)	http://www-bio.llnl.gov/bbrp/image/image.html
TIGR Human Gene Index (HGI)	http://www.tigr.org/tdb/hgi/hgi.html
Washington University-Merck EST project	http://www.merck.com/ http://genome.wustl.edu/est/esthmpg.html
Databases	
Laboratoire de Génétique Cellulaire INRA	http://www/lgc/lgc.html
MGD (Mouse Genome Database)	http://www.informatics.jax.org
XREF	http://www.ncbi.nlm.nih.gov/XREFdb/

2.1.2. Different messenger RNAs from the same gene

Starting from a single messenger RNA several complementary DNAs and several tags can thus be obtained. In addition, a single gene – a DNA sequence – can give several messenger RNAs through at least three mechanisms (*figure 1A, B*).

a) In eukaryotes the mature mRNA is not a direct copy of the sequence of the gene, since certain portions of the sequence, the introns, are suppressed by a process of 'splicing' leading to the joining of the other portions of the sequence, the 'exons', which are the only parts present in the final messenger RNA. Cells possess mechanisms which allow them to produce several different mRNAs starting from a single primary transcript: alternative splicing. This process consists in the incorporation or the exclusion of one or more exons

in the mature mRNA, a combinatorial strategy which greatly increases the possibilities of expression [85].

b) The location of the 3' poly(A) sequence is determined by a polyadenylation signal, and several signals can co-exist, possibly located in different exons, thus producing different messenger RNAs from the same gene.

c) Finally, the start of transcription occurs at a particular sequence, the promoter, and some genes have different promoters, possibly in different exons, so that from the same gene several different primary transcripts can be synthesised. These primary transcripts can of course be subjected to alternative splicing and/or alternative polyadenylation.

2.1.3. cDNA: a redundant representation of the genome

In a given cell, the different mRNAs are present in a variable number of copies. Molecular hybridisation studies on messenger RNAs of HeLa cells [19] or of different mouse tissues (liver, kidney, brain [40]) have shown that these RNAs fall into three abundance classes:

– very abundant messengers, of about ten different types, each present in the cells in thousands of copies;

– abundant messengers, some hundreds of molecular species, each present in some hundreds of copies;

– rare messengers, thousands of different molecular species, with only a few tens of copies per cell, or fewer than ten. This class can represent up to 50 % of the total mass of messengers.

The mRNA populations are thus complex and the presence of many copies of the same RNA entails the redundancy of cDNAs.

If the aim is to identify a large number of genes and thus to obtain a large number of different tags from cDNA libraries, the redundancy will quickly reduce the effectiveness of the search: the very abundant mRNAs will lead to repeated production of the same tag, while it will be very difficult to obtain tags corresponding to very rare messenger RNAs. For example, the determination of 3 000 ESTs allows the identification of at least one transcript of 99 % of very abundant messengers, more than 85 % of abundant messengers, but less than 5 % of rare messengers [53].

To alleviate this problem, a process of 'normalisation' can be used. Its objective is to obtain a library in which all the cDNAs are present in approximately equal quantities. The principle is to submit the population of cDNAs to a process of denaturation followed by reassociation in such a way that the most abundant molecular species reassociate more rapidly and can be partly eliminated; in the remaining 'normalised' population, the number of copies of each of the more abundant species will have decreased by a factor of 1 000 to 10 000. To facilitate the reassociation kinetics, and thus to increase the effectiveness of the treatment, the size of cDNAs is usually reduced. In the approach developed by the Soares group [86], the normalisation is conducted starting from cDNA cloned in single strand form. The synthesis of the second strand, of limited length, is started at the 3' end. This strategy thus conserves the advantage of short sequences, but the normalised gene library is composed of full length cDNAs already cloned. With some variations of this technique, 35 different libraries have been successfully constructed [23].

2.2. Identification and analysis of sequences

The identification of genes is carried out by comparison of their sequences with those contained in existing public databases such as GenBank or EMBL. In August 1998, GenBank contained 2 532 359 sequences representing 1 797 137 713 bases, where man represents about 50 % of the total. The EST sequences submitted to GenBank are kept in a specific database, dbEST, and compared (using the BLAST program) regularly with databases of protein and non-redundant nucleic acid sequences. The partial results of these searches (15 best matches) are included in the commentary on the EST considered [21]. At the end of August 1998, this database contained 1 785 394 entries. Among the 117 species present, the most closely studied species are man (1 086 919), mouse (353 450), *Caenorhabditis elegans* (72 569), the rat (57 274), Drosophila (37 848) and *Arabidopsis thaliana* (37 445) followed, in the animal kingdom, by the pig (2 365), the rabbit (1 659), the chicken (304), the goat (245), cattle (199), the dog (107) and the trout (89).

Sequence comparisons utilise different automatic alignment programs such as FASTA [71] or BLAST [9]. These analyses result in a list of more or less similar sequences, but, as we have already pointed out, these results require 'manual' interpretation. The 'best' results obtained by computer analysis must be validated; criteria such as the length and the percentage of similarity of these sequences, or even data on the evolution of the genes [32], are used to judge the pertinence of the identifications.

This identification of sequences would not be possible without the support and development of computer technology, which permits, in addition to the analysis of data, their management and access via the Internet. The role of this technology has become so important that one speaks of research 'in silico' ('in silicio' would be more correct) allowing new results to be derived from information already accumulated. The data on sequence and function provided by ESTs thus constitute primary material, and we cite here three examples of this research which will also be illustrated later, in the framework of comparative or integrated approaches.

ESTs permit a first approach to the analysis of the evolution of proteins or certain protein domains. Thus, tags obtained from libraries of the human brain [2, 3] and of the nematode *C. elegans* [95] have been compared systematically among themselves and with sequences of yeast and *Escherichia coli*. Regions conserved in the course of evolution, known as 'ancient evolutionary conserved regions' (ACR) have been detected in the corresponding proteins [38].

From another perspective, the analysis of redundant human sequences present in the database dbEST can allow the identification of new properties of genes already well characterised, such as new alleles or a specificity of expression. By comparing, for 15 genes, the genomic sequence with those of tags, Wolfsberg and Landsman were able to demonstrate new profiles of alternative splicing [98]. Similarly, the group of Pastan, starting from already available sequences, has identified genes expressed specifically in the prostate [92].

2.3. Applications

Developed extensively by the group of Venter [2–6], this systematic approach has culminated in the publication in 1995 of 174 472 human tags, derived from 300 cDNA libraries representing 37 different tissues or organs [7]. Another project has been conducted by the University of Washington, financed by the Merck company [1, 43]. Among other species, to cite only a few, studies are proceeding on the nematode *C. elegans* [60, 95], the model plant *A. thaliana* [28, 44, 66], the pig [88, 97] and the goat [52].

3. THE USE OF TAGS IN FUNCTIONAL STUDIES

In a particular tissue or cell type at a given time only some of the genes are expressed, determining the specificity of the tissue; two different tissues or cell types express different proteins whose synthesis is directed by different messenger RNAs. Also, for each tissue, gene expression varies with physiological, pathological or experimental conditions. Thus cDNA libraries mirror the set of genes expressed in a given tissue and provide access to various modes (tissue-specific, time-dependent, experimental) of gene expression.

3.1. Functional expression profiles

3.1.1. Numerical approach

The EST strategy shows here its full value: by simplifying sequencing it allows the rapid building of a catalogue, even a partial one. The 'expression profile' established using tags will permit appreciation of the activity levels of different genes, reflected in the frequency of appearance of the corresponding tag [69]. By analysing the 3' tags of about a thousand clones taken at random from a cDNA library of a hepatic cell line, Okubo and colleagues showed that 52 % of messenger RNAs were redundant, representing 173 genes. Among these redundant clones 55 corresponded to only three different species, in particular to serum albumin produced in great quantity by the liver (21 clones out of 982).

Similarly, to demonstrate genes regulated by a growth factor, Lee and colleagues have made a comparison of redundancy profiles obtained from treated and untreated cells, each profile being the result of sequencing more than 3 000 tags taken at random from the two libraries. The comparison of these two profiles enabled them to identify around 600 different regulated messengers, and the regulation of expression was checked for 15 sequences and validated for 12 of them [53].

In a remarkable strategy, SAGE (serial analysis of gene expression), Velculescu and colleagues [93] have developed this systematic determination of profiles by using short tags which they sequence in series. These authors have identified 289 transcripts whose expression level is different in normal and cancerous colon cells. This result was obtained by the analysis of 300 000 tags, corresponding to 49 000 different genes whose levels of expression varied from 1 to 5 300 copies per cell [100]. This technique has also been applied to yeast,

where it has enabled the analysis of the quasi-totality of transcripts [94] revealing 4 665 different genes starting from 60 000 tags.

3.1.2. Analogical approach

A different approach is based on hybridisation used in conditions where the intensity of hybridisation signals varies with the quantity of messenger RNAs and thus reflects the activity of the corresponding genes. The simultaneous hybridisation of a large number of clones permits information to be obtained on the expression of numerous genes.

In 'high density' filters the clones of a library are transferred to a single membrane with a density of the order of 25 to 50 colonies per cm^2 [54, 67, 73, 101]. The same principle, miniaturised, is used in 'microarrays' [82] where the density attains nearly 2 000 clones per cm^2, and the simultaneous use of two probes marked with different fluorochromes allows direct detection of the regulated genes [83]. This technique has been successfully used to analyse the expression of 6 000 genes of yeast, either in the course of growth or in mutant strains [29]. Still in the course of development, this approach has without doubt the promise of a great future.

3.2. Targeted research

As well as these global approaches, research can be targeted at a particular group of messenger RNAs to identify the genes involved in a given process. The selection of a messenger RNA subpopulation is classically performed by subtraction or by comparison; very promising new approaches using the polymerase chain reaction (PCR) are now available.

3.2.1. Subtraction

In the subtraction method, the aim is to enrich the cDNA population in species specific to a particular tissue or cell type by eliminating sequences common to several tissues or cell types and thus non-specific. The population of cDNAs to be studied (target cDNA) is mixed with an excess of cDNA (driver) in which the specific sequences are absent, the mixture is submitted to a process of denaturation followed by reassociation, and the two-stranded species are separated from the single-stranded species which have not rehybridised by chromatography on a hydroxylapatite column. The molecular species specific to the target are found among the single-stranded nucleic acids [80].

3.2.2. Comparison: differential screening

In the approach by differential screening, two replicates of a classic cDNA library are hybridised with two probes corresponding to the two populations of messenger RNAs which are to be compared. The variations in abundance of the messenger RNAs are deduced from the differences in intensity of the hybridisation signals. This approach allows direct access to the messenger RNAs which are regulated, but only from the abundant and very abundant classes in the cells from which they come, because molecular hybridisation favours

the species which are most frequent. By contrast, the clones corresponding to rare messenger RNAs give no hybridisation signal. This study of 'rare' clones has been applied to the mouse testis by Höög [45], by screening the library with cDNA probes prepared either from the testes or from other tissues (liver, kidney, heart) of the mouse: redundancy is reduced and the number of different sequences isolated is augmented.

3.2.3. New approaches

New strategies, having in common the use of PCR, now permit a large number of tags to be obtained without the need to isolate complete cDNAs. The amplification step also allows the use of small initial quantities of material.

With 'mRNA differential display' [56] and AFLP (amplified restriction fragment length polymorphism [13]) the cDNAs obtained by reverse transcription are amplified by PCR using different primers. They are then separated into different subgroups such that electrophoresis gives a profile of discrete bands. These techniques allow direct comparison of several messenger RNA populations by analysing on the same gel the profiles obtained from the different populations under study; thus, they give direct access to regulated messenger RNAs and permit identification of messengers of low abundance. A subtraction strategy is used in RDA (representational difference analysis [47]) and suppression subtractive hybridisation (SSH [30]) where the product subtracted is also normalised. In these two methods PCR enables specific amplification of the subtracted product.

3.3. An example of application

The value of the EST strategy is well illustrated by the work of Affara's group [8, 49] on the human testis, with the isolation of 359 clones among which 242 showed no similarity to known sequences. Analysis of the tissue specificity of expression showed that of 80 clones analysed 20 showed testicular specificity, either by exclusive expression or by the presence of a transcript of specific size. The chromosomal localisation was determined using a panel of somatic hybrids (see section 4.1). Among the ESTs identified by their similarity to genes of other species – indicated between brackets – there were at least three sequences related to testicular function: a heavy chain of dynein (sea-urchin) involved in spermatozoon mobility, a molecule for sperm adhesion (guinea pig) involved in the fertilisation of the oocyte by the spermatozoon, and a glycerol kinase (*Bacillus subtilis*). A deficiency of this enzyme is implicated in an X-linked hereditary disease, and a sequence homologous to that of the EST has been found in the region of chromosome X which contains the locus for glycerol kinase deficiency. This sequence is deleted in two patients suffering glycerol kinase deficiency, indicating the probable origin of the genetic defect [81].

ESTs thus permit the drawing up of a collection, more or less exhaustive, of genes expressed in a given tissue, the establishment of an expression profile indicating the level of activity of different genes, and the making of comparisons between different tissues or different states of the same tissue. This analysis of genes constitutes a first step towards the characterisation of the 'transcriptome' [94].

4. TAGS AS AIDS TO GENETIC STUDIES

The main genetic application of ESTs is the chromosomal localisation of genes, but, as well, information obtained from ESTs enables the approaches of comparative and integrated genetics to be developed.

4.1. Chromosomal localisation and mapping the genome

To determine the localisation and order of genes and the distance separating them, different approaches are used.

The genetic approach uses polymorphic markers to measure the frequency of recombination. This strategy has had an explosive development with the use of microsatellites, sequences of short repeated motifs (most often two or three nucleotides). Their value lies in their high degree of polymorphism (number of repeats) and their distribution, more or less homogeneous according to species, over the whole genome. The map resolution so obtained is of the order of a cM, that is, very approximately, one megabase.

The cytogenetic approach uses DNA probes to localise the corresponding sequences by hybridisation to chromosomes spread in metaphase (in situ hybridisation). The use of probes of large size (cosmids; YACs, yeast artificial chromosomes; BACs, bacterial artificial chromosomes; PACs, P1-derived artificial chromosomes) fluorescently labelled has allowed accelerated acquisition of data as compared with radioactive probes. The resolution is of the order of one chromosomal band or, very approximately, ten megabases.

New possibilities for mapping have appeared with the cytogenetic characterisation of interspecific somatic cell hybrids. In these hybrids between rodent cells and cells of the species of interest, a variable part of the chromosomes (entire or not) of the species to be analysed is lost in a random manner. The principle of localisation is then to seek cosegregation between a chromosome or fragment of chromosome and the gene studied. The presence of the gene in the different hybrids is shown by amplification of the DNA of the hybrid, using for the PCR reaction a couple of primers specific for this gene. One thus obtains an 'STS' (sequence tagged site [70]), that is, a landmark on a particular chromosome. The sequences of ESTs permit the design of such primers [96]. One can then speak of eSTS (expressed STS [12]).

The ESTs obtained in the 3' non-coding region are valuable for this purpose because that region is less well conserved between species than the coding regions [59] which limits the possibilities of amplification of the DNA of the rodent (hamster or mouse). Furthermore, that region generally does not contain introns [41, 59] which allows amplification of a fragment whose size is known from the sequence of the EST. This strategy has been very quickly developed in man [2, 51, 74, 96]. A panel of this type has been developed in our laboratory for the pig [78, 99] and its use has enabled the location of numerous tags [25, 50, 89].

Obtained on the same principle, irradiated hybrids permit more precise localisation (resolution of the order of 100 kilobases) since irradiation of the cells before fusion causes the fracture of chromosomal DNA into fragments of several megabases. The outcome approaches that of genetic mapping since the frequency of chromosome breakages between two points is measured. A

final approach is that of libraries of large DNA fragments, of the order of a megabase, cloned in vectors of the artificial chromosome type (BAC, PAC, YAC), ordered and grouped into 'contigs'. These fragments have the advantage of giving access to the genomic DNA and thus to the complete structure of the gene (introns/exons, regulatory sequences, etc.).

These last approaches, YAC and irradiated hybrids, have produced a considerable advance in human gene mapping: an international consortium of 18 laboratories has in this way placed on the human map more than 16 000 genes – clusters (cf. section 2.1.1) – in relation to a frame of 1 000 genetic markers already mapped [84].

These different approaches and the maps which result from them are complementary: in an integrated mapping approach they allow in particular the localisation on the cytogenetic map of markers already placed on the genetic map. The chromosomal localisation of ESTs enables this map to be enriched to make a map of expression or of transcripts [22] which makes possible the strategy of the positional candidate gene [14, 27], that is, the identification of genes of interest whose presence in a particular region of the genome has been shown by genetic methods.

4.2. Comparative genetics

Comparative mapping is illustrated by the identification of the gene for halothane sensitivity in the pig with the gene for malignant hyperthermia in [58, 57]. It is based on the partial conservation between species not only of the sequence of genes but also of their organisation – linkage groups or syntenies – on the different chromosomes [10]. Also, the localisation in different species of mammals of a collection of genes covering the totality of the human genome has been proposed [68]. In species of agricultural interest, a systematic study of correspondences between different genomes, chromosome by chromosome, has been made between man and cattle [42] and pigs [35, 37, 77]. The information concerning the pig is available on the Web server of the laboratoire de Génétique Cellulaire (*table I*). A study of comparative mapping data is proposed for the site of the MGD database (Mouse Genome Database; *table I*). For each gene of the mouse for which a homologous locus has been identified in other species, the corresponding localisation data are indicated, and more than 50 different species are represented in the database [20, 65].

ESTs also contribute to this comparative mapping by allowing the localisation of the same gene in different species even if, at the 3' end, sequence conservation is poor; preliminary results obtained at the Genethon and in our laboratory indicate that primers established from the sequence of human 3' tags to permit their amplification by PCR are usable in other species, in a proportion of a few percent for pigs (G. Gyapay and Y. Lahbib-Mansais, pers. comm.).

It is then possible to know, for a given region of, for example, a pig chromosome, the genes located in the equivalent region in man and use these genes as 'positional candidates' in the pig.

As well as mapping, the comparison of genes of different model species can enable remarkable progress in understanding their function, benefiting from the knowledge of such model species as yeast and drosophila [90, 91]. Despite

the genetic distance between the species, certain genes have retained a function so close that mutations cause similar phenotypic changes in species as distant from one another as man, mouse and drosophila [63, 76].

This original approach has been well used by the group of Ballabio [15]. It consists of screening the dbEST database [21] to find mutant genes stored in the drosophila genetic database [18, 34]. The human ESTs corresponding to genes unknown in man, but known in drosophila, have been named DRES (drosophila-related expressed sequences). The corresponding clones have been isolated, sequenced and located precisely on the chromosomes; the genetic diseases whose genes are located in the corresponding regions have then been identified by interrogating the database MIM (Mendelian inheritance in man [61]; *table I*). Certain DRES clones then become promising positional candidate genes, in particular if the phenotype of the mutant in drosophila resembles that of the human disease. For example, DRES9 is homologous to the gene 'drosophila retinal degeneration B' and is located in 1q15 where at least three types of human retinopathy have been assigned. The human ESTs which show similarity to drosophila mutant genes can thus provide information on the possible phenotypic consequences of poor functioning of these genes. This systematic search has enabled identification of human ESTs corresponding to a given protein in drosophila ([39]; *table I*). Even if the sequence similarities and functional resemblances are not sufficient arguments to designate the gene(s) responsible, this information permits the designation of candidates, which a functional and genetic study will later allow to be accepted or rejected.

The same approach has been developed with respect to yeast, allowing the cloning of human mitochondrial RNA polymerase [87]. It is also used in the framework of the XREF project where the search for homology is made starting from a non-redundant database of proteins of different model organisms: *Mus musculus*, *Drosophila melanogaster*, *C. elegans*, *Saccharomyces cerevisiae* and *E. coli* [17]. This research, with regular updates, is also accessible on the World Wide Web (*table I*). Applied to genes identified by positional cloning, this approach has enabled the presence of orthologous genes to be demonstrated, in particular in *C. elegans* and, to a lesser extent, in *S. cerevisiae* [64]. The study of protein motifs has similarly enabled the identification in several of these genes of specific domains (an ATP binding site in a gene for colon cancer, an exonuclease domain in the protein of Werner's syndrome).

4.3. Integration of data

BodyMap is a database combining information, both qualitative and quantitative, concerning the expression of human genes, identified or not. In this expression map, the genes are assigned to the tissues in which they are active rather than to chromosomes. This database, fed by sequence data of 3' tags obtained from cDNA libraries [69], allows the determination of an expression profile of the gene of interest. In August 1998 it contained 10 896 entries, corresponding to 39 tissues or cell types (*table I*).

The 'Cancer Anatomy Genome Project' of the American National Cancer Institute has as its objective the identification of all the genes expressed in cancer cells, in order to produce a molecular characterisation of them

([72]; *table I*). The expression profiles will be correlated with the anatomical–pathological characteristics of the tissues with the aim of improving the diagnosis, prognosis and treatment of tumours. The programme proceeds by the construction of cDNA libraries for the five principal cancers (colon, ovary, lung, prostate, breast) from normal cells up to metastases. The gene libraries will then be sequenced following the EST strategy in order to establish 'digital differential displays', differential profiles indicating the relative expression levels of the different genes.

More ambitious still is the project 'IMAGE' (integrated molecular analysis of genomes and their expression; *table I* [55]) to put together resources and results with a view to the constitution of a universal gene library, that is, a collection of clones containing the cDNAs corresponding to each of the transcripts of the human genome. The resources consist of the cDNA libraries of different human tissues, ordered, from which it is possible to obtain either individual clones or high density filters. The results, sequences, chromosomal locations and expression profiles are collected in public domain databases. This strategy was first applied to muscle and to brain [12] and then developed for 35 different gene libraries [23]. The problem remains, in particular in the framework of this project, of the management of a larger and larger quantity of information, scattered among different sources or Internet sites. With a view to better access, a series of programs has been conceived to integrate the different pieces of information associated with the IMAGE clones [62].

5. CONCLUSION

Paraphrasing Henri Poincaré (Science and Hypothesis) "science is built with facts, just as a house is built with stones, but an accumulation of facts is no more a science than a pile of stones is a house", we can say that "an accumulation of sequences or markers is not a genome". The systematic sequencing and location of a very large number of genes constitutes a step which is indispensible, but insufficient, for the understanding of the organisation and functioning of the genome.

From the perspective of 'genomics' the compilations published by Venter's group and by the International Consortium are undoubtedly major events, but to deepen our knowledge of the genome it is necessary to go further, towards 'functional genomics'. The approach through the use of ESTs, coupled with different strategies of random sampling, of selection, of subtraction or comparison, permits us to follow effectively the expression of numerous genes in different physiological conditions, such as cellular growth and organogenesis, or pathological conditions such as the development of cancers.

In the field of livestock improvement, selection would become more effective if the choice of animals were made on their genotypes and not only on their phenotypes, as shown by the example of halothane sensitivity, and it is desirable to extend this strategy to different traits of economic importance. To achieve this objective it is necessary to know the gene or genes responsible for the character under consideration, and the strategy of 'positional candidate' is currently the most promising. From this viewpoint, ESTs are the best method of obtaining the catalogue and localisations of genes implicated in a given function and, in consequence, of providing a list of candidate genes for major genes or

for QTLs identified by genetic methods. The strategy of establishment of such catalogues is a long-term project and is still poorly developed in livestock, but comparative mapping allows this handicap to be overcome: comparison with the human map where very numerous ESTs have been localised allows genes located in equivalent regions to be designated as candidates.

In a few years, ESTs have achieved considerable development in man and have made a major contribution to knowledge of the genome, both at the level of sequencing and at the level of the physical map. The integration of these data with data, still very fragmentary, concerning the places and modes of expression of genes will permit the understanding of the genome in its structure and its function. Such an approach has already led to the identification of the genes responsible for certain genetic diseases of mankind. Its application in domestic animals, in particular in the domain of identification of the genes implicated in economic traits, is rich in promise.

ACKNOWLEDGEMENTS

The authors thank Charles Auffray (CNRS, Villejuif), Claude Chevalet, Joël Gellin and Denis Milan (Inra, Castanet-Tolosan) for their comments and critical reading of the manuscript.

REFERENCES

[1] Aaronson J.S., Eckman B., Blevins R.A., Borkowski J.A., Myerson J., Imran S., Elliston K.O., Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data, Genome Res. 6 (1996) 829–845.

[2] Adams M.D., Kelley J.M., Gocayne J.D., Dubnick M., Polymeropoulos M.H., Xiao H., Merril C.R., Wu A., Olde B., Moreno R.F. et al., Complementary DNA sequencing: Expressed Sequence Tags and human genome project, Science 252 (1991) 1651–1656.

[3] Adams M.D., Dubnick M., Kerlavage A.R., Moreno R., Kelley J.M., Utterback T.R., Nagle J.W., Fields C., Venter J.C., Sequence identification of 2,375 human brain genes, Nature 355 (1992) 632–634.

[4] Adams M.D., Kerlavage A.R., Fields C., Venter J.C., 3,400 new expressed sequence tags identify diversity of transcripts in human brain, Nat. Genet. 4 (1993) 256–267.

[5] Adams M.D., Soares M.B., Kerlavage A.R., Fields C., Venter J.C., Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library, Nat. Genet. 4 (1993) 373–380.

[6] Adams M.D., Kerlavage A.R., Kelley J.M., Gocayne J.D., Fields C., Fraser C.M., Venter J.C., A model for high-throughput automated DNA sequencing and analysis core facilities, Nature 368 (1994) 474–475.

[7] Adams M.D., Kerlavage A.R., Fleischmann R.D., Fuldner R.A., Bult C.J., Lee N.H., Kirkness E.F., Weinstock K.G., Gocayne J.D., White O. et al., Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, Nature 377 (1995) 3–174.

[8] Affara N.A., Bentley E., Davey P., Pelmear A., Jones M.H., The identification of novel gene sequences of the human adult testis, Genomics 22 (1994) 205–210.

[9] Altschul S.F., Boguski M.S., Gish W., Wootton J.C., Issues in searching molecular sequence databases, Nat. Genet. 6 (1994) 119–129.

[10] Andersson L., Archibald A., Ashburner M., Audun S., Barendse W., Bitgood J., Bottema C., Broad T., Brown S., Burt D. et al., Comparative genome organization of vertebrates, The First International Workshop on Comparative Genome Organization, Mamm. Genome 7 (1996) 717–734.

[11] Archibald A.L., Haley C., Brown J.F., Couperwhite S., McQueen H.A., Nicholson D., Coppieters W., Van de Weghe A., Stratil A., Winterø A.K. et al., The PigMaP consortium linkage map of the pig (*Sus scrofa*), Mamm. Genome 6 (1995) 157–175.

[12] Auffray C., Behar G., Bois F., Bouchier C., Da Silva C., Devignes M.D., Duprat S., Houlgatte R., Jumeau M.N., Lamy B. et al., IMAGE: intégration au niveau moléculaire de l'analyse du génome humain et de son expression, C. R. Acad. Sci. III 318 (1995) 263–272.

[13] Bachem C.W., van der Hoeven R.S., de Bruijn S.M., Vreugdenhil D., Zabeau M., Visser R.G., Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development, Plant J. 9 (1996) 745–753.

[14] Ballabio A., The rise and fall of positional cloning, Nat. Genet. 3 (1993) 277–279.

[15] Banfi S., Borsani G., Rossi E., Bernard L., Guffanti A., Rubboli F., Marchitiello A., Giglio S., Coluccia E., Zollo M. et al., Identification and mapping of human cDNAs homologous to Drosophila mutant genes through EST database searching, Nat. Genet. 13 (1996) 167–174.

[16] Banfi S., Guffanti A., Borsani G., How to get the best of dbEST, Trends Genet. 14 (1998) 80–81.

[17] Bassett D.E., Boguski M.S., Spencer F., Reeves R., Kim S.H., Weaver T., Hieter P., Genome cross-referencing and XREFdb: Implications for the identification and analysis of genes mutated in human disease, Nat. Genet. 15 (1997) 339–344.

[18] Bellen H.J., Smith R.F., FlyBase: a virtual *Drosophila cornucopia*, Trends Genet. 11 (1995) 456–457.

[19] Bishop J.O., Morton J.G., Rosbash M., Richardson M., Three abundance classes in HeLa cell messenger RNA, Nature 250 (1974) 199–204.

[20] Blake J.A., Richardson J.E., Davisson M.T., Eppig J.T., The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. The Mouse Genome Informatics Group, Nucleic Acids Res. 25 (1997) 85–91.

[21] Boguski M.S., Lowe T.M., Tolstoshev C.M., dbEst – database for 'expressed sequence tags', Nat. Genet. 4 (1993) 332–333.

[22] Boguski M.S., Schuler G.D., ESTablishing a human transcript map, Nat. Genet. 10 (1995) 369–371.

[23] Bonaldo M.D.F., Lennon G., Soares M.B., Normalization and subtraction: Two approaches to facilitate gene discovery, Genome Res. 6 (1996) 791–806.

[24] Burke J., Wang H., Hide W., Davison D.B., Alternative native gene form discovery and candidate gene selection from gene indexing projects, Genome Res. 8 (1998) 276–290.

[25] Clouscard-Martinato C., Mulsant P., Robic A., Bonnet A., Gasser F., Hatey F., Characterization of FSH-regulated genes isolated by mRNA differential display from pig ovarian granulosa cells, Anim. Genet. 29 (1998) 98–106.

[26] Collins F.S., Positional cloning: Let's not call it reverse anymore, Nat. Genet. 1 (1992) 3–6.

[27] Collins F.S., Positional cloning moves from perditional to traditional, Nat. Genet. 9 (1995) 347–350.

[28] Cooke R., Raynal M., Laudie M., Grellet F., Delseny M., Morris P.C., Guerrier D., Giraudat J., Quigley F., Clabault G. et al., Further progress towards a catalogue

of all Arabidopsis genes: Analysis of a set of 5 000 non-redundant ESTs, Plant J. 9 (1996) 101–124.

[29] DeRisi J.L., Iyer V.R., Brown P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[30] Diatchenko L., Lau Y.F., Campbell A.P., Chenchik A., Moqadam F., Huang B., Lukyanov S., Lukyanov K., Gurskaya N., Sverdlov E.D. et al., Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries, Proc. Natl. Acad. Sci. USA 93 (1996) 6025–6030.

[31] Echard G., The gene map of the pig (*Sus scrofa domestica* L.,). Genetic maps: a compilation of linkage and restriction maps of genetically studied organisms, S. J. O'Brien (1987) 490–493.

[32] Eisen J.A., Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis, Genome Res. 8 (1998) 163–167.

[33] Ellegren H., Chowdhary B.P., Johansson M., Marklund L., Fredholm M., Gustavsson I., Andersson L., A primary linkage map of the porcine genome reveals a low rate of genetic recombination, Genetics 137 (1994) 1089–1100.

[34] FlyBase, FlyBase – the Drosophila database, Nucleic Acids Res. 22 (1994) 3456–3458.

[35] Frönicke L., Chowdhary B.P., Scherthan H., Gustavsson I., A comparative map of the porcine and human genomes demonstrates ZOO-FISH and gene mapping-based chromosomal homologies, Mamm. Genome 7 (1996) 285–290.

[36] Gellin J., Grosclaude F., Analyse du génome des espèces d'élevage : projet d'établissement de la carte génique du porc et des bovins, Inra Productions Animales 4 (1991) 97–105.

[37] Goureau A., Yerle M., Schmitz A., Riquet J., Milan D., Pinton P., Frelat G., Gellin J., Human and porcine correspondence of chromosome segments using bidirectional chromosome painting, Genomics 36 (1996) 252–262.

[38] Green P., Lipman D., Hillier L., Waterston R., States D., Claverie J.M., Ancient conserved regions in new gene sequences and the protein databases, Science 259 (1993) 1711–1716.

[39] Guffanti A., Banfi S., Simon G., Ballabio A., Borsani G., DRES search engine: of flies, men and ESTs, Trends Genet. 13 (1997) 79–80.

[40] Hastie N.D, Bishop J.O., The expression of three abundance classes of messenger RNA in mouse tissues, Cell 8 (1976) 761–774.

[41] Hawkins J.D., A survey of intron and exons lengths, Nucleic Acids Res. 16 (1988) 9893–9908.

[42] Hayes H., Chromosome painting with human chromosome-specific DNA libraries reveals the extent and distribution of conserved segments in bovine chromosomes, Cytogenet. Cell Genet. 71 (1995) 168–174.

[43] Hillier L., Lennon G., Becker M., Bonaldo M.F., Chiapelli B., Chissoe S., Dietrich N., Dubuque T., Favello A., Gish W. et al., Generation and analysis of 280,000 human expressed sequence tags, Genome Res. 6 (1996) 807–828.

[44] Höfte H., Desprez T., Amselem J., Chiapello H., Rouze P., Caboche M., Moisan A., Jourjon M.F., Charpenteau J.L., Berthomieu P. et al., An inventory of 1 152 expressed sequence tags obtained by partial sequencing of cDNAs from Arabidopsis thaliana, Plant J. 4 (1993) 1051–1061.

[45] Höög C., Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy, Nucleic Acids Res. 19 (1991) 6123–6127.

[46] Houlgatte R., Mariagesamson R., Duprat S., Tessier A., Bentolila S., Lamy B., Auffray C., The genexpress index: A resource for gene discovery and the genic map of the human genome, Genome Res. 5 (1995) 272–304.

[47] Hubank M., Schatz D.G., Identifying differences in mRNA expression by representational difference analysis of cDNA, Nucleic Acids Res. 22 (1994) 5640–5648.

[48] Jiang, J., Jacob H.J., EbEST: An automated tool using expressed sequence tags to delineate gene structure, Genome Res. 8 (1998) 268–275.

[49] Jones M.H., Zhang Y., Tirosvoutis K.N., Davey P.M., Webster A.R., Walsh D., Spurr N.K., Affara N.A., Chromosomal assignment of 311 sequences transcribed in human adult testis, Genomics 40 (1997) 155–167.

[50] Jorgensen C.B., Winterø A.K., Yerle M., Fredholm M., Mapping of 22 expressed sequence tags isolated from a porcine small intestine cDNA library, Mamm. Genome 8 (1997) 423–427.

[51] Khan A.S, Wilcox A.S., Polymeropoulos M.H., Hopkins, J.A., Stevens T.J., Robinson M., Orpana A.K., Sikela J.M., Single pass sequencing and physical and genetic mapping of human brain cDNAs, Nat. Genet. 2 (1992) 180–185.

[52] Le Provost F., Lepingle A., Martin P., A survey of the goat genome transcribed in the lactating mammary gland, Mamm. Genome 7 (1996) 657–666.

[53] Lee N.H, Weinstock K.G., Kirkness E.F., Earle-Hughes J.A., Fuldner R.A., Marmaros S., Glodek A., Gocayne J.D., Adams M.D., Kerlavage A.R. et al., Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment, Proc. Natl. Acad. Sci. USA 92 (1995) 8303–8307.

[54] Lennon G.G, Lehrach H., Hybridization analyses of arrayed cDNA libraries, Trends Genet. 7 (1991) 314–317.

[55] Lennon G., Auffray C., Polymeropoulos M., Soares M.B., The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression, Genomics 33 (1996) 151–152.

[56] Liang P., Pardee A.B., Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction, Science 257 (1992) 967–971.

[57] MacLennan D.H., Phillips M.S., Malignant hyperthermia, Science 256 (1992) 789–794.

[58] MacLennan D.H., Duff C., Zorzato F., Fujii J., Phillips M., Korneluk R.G., Frodis W., Britt B.A., Worton R.G., Ryanodine receptor gene is a candidate for predisposition to malignant hyperthermia, Nature 343 (1990) 559–561.

[59] Makalowski W., Zhang J., Boguski M.S., Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences, Genome Res. 6 (1996) 846–57.

[60] McCombie W.R., Adams M.D., Kelley, J.M., FitzGerald M.G., Utterback T.R., Khan M., Dubnick M., Kerlavage A.R., Venter J.C., Fields C., *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues, Nat. Genet. 1 (1992) 124–131.

[61] McKusick V.A., Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders, 11th ed., Johns Hopkins University Press, Baltimore, 1994.

[62] Miller G., Fuchs R., Lai E., IMAGE cDNA clones, UniGene clustering, and ACeDB: An integrated resource for expressed sequence information, Genome Res. 7 (1997) 1027–1032.

[63] Mounkes L.C., Jones R.S., Liang B.C., Gelbart W., Fuller M.T., A Drosophila model for xeroderma pigmentosum and Cockayne's syndrome: haywire encodes the fly homolog of ERCC3, a human excision repair gene, Cell 71 (1992) 925–937.

[64] Mushegian A.R., Bassett D.E., Jr., Boguski M.S., Bork P., Koonin E.V., Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs, Proc. Natl. Acad. Sci. USA 94 (1997) 5831–5836.

[65] Nadeau J.H., Grant P.L., Mankala S., Reiner A.H., Richardson J.E., Eppig J.T., A Rosetta stone of mammalian genetics, Nature 373 (1995) 363–365.

[66] Newman T., de Bruijn F.J., Green P., Keegstra K., Kende H., McIntosh L., Ohlrogge J., Raikhel N., Somerville S., Thomashow M. et al., Genes galore: a summary

of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones, Plant Physiol. 106 (1994) 1241–1255.

[67] Nguyen C., Rocha D., Grandjeaud S., Baldit M., Bernard K., Naquet P., Jordan B.R., Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones, Genomics 29 (1995) 207–216.

[68] O'Brien S., Womack J.E., Lyons L.A., Moore K.J., Jenkins N.A., Copeland N.G., Anchored reference loci for comparative genome mapping in mammals, Nat. Genet. 3 (1993) 103–112.

[69] Okubo K., Hori N., Matob R., Niiyama T., Fukushima A., Kojima Y., Matsubara K., Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, Nat. Genet. 2 (1992) 173–179.

[70] Olson M., Hood L., Cantor C., Botstein D., A common language for physical mapping of the human genome, Science 245 (1989) 1434–1435.

[71] Pearson W.R., Lipman D.J., Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444–2448.

[72] Pennisi E., Genomics – A catalog of cancer genes at the click of a mouse, Science 276 (1997) 1023–1024.

[73] Pietu G., Alibert O., Guichard V., Lamy B., Bois F., Leroy E., Mariage-Sampson R., Houlgatte R., Soularue P., Auffray C., Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array, Genome Res. 6 (1996) 492–503.

[74] Polymeropoulos M.H., Xiao H., Glodek A., Gorski M., Adams M.D., Moreno R.F., Fitzgerald M.G., Venter J.C., Merril C.R., Chromosomal assignment of 46 brain cDNAs, Genomics 12 (1992) 492–496.

[75] Putney S.D., Herlihy W.C., Schimmel P., A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing, Nature 302 (1983) 718–721.

[76] Quiring R., Walldorf U., Kloter U., Gehring W.J., Homology of the *eyeless* gene of Drosophila to the *Small eye* gene in mice and *Aniridia* in humans, Science 265 (1994) 785–789.

[77] Rettenberger G., Klett C., Zechner U., Kunz J., Vogel W., Hameister H., Visualization of the conservation of synteny between humans and pigs by heterologous chromosomal painting, Genomics 26 (1995) 372–378.

[78] Robic A., Riquet J., Yerle M., Milan D., Lahbib-Mansais Y., Dubut-Fontana C., Gellin J., Porcine linkage and cytogenetic maps integrated by regional mapping of 100 microsatellites on somatic cell hybrid panel, Mamm. Genome 7 (1996) 438–445.

[79] Rohrer G.A., Alexander L.J., Hu Z.L., Smith T.P.L., Keele J.W., Beattie C.W., A comprehensive map of the porcine genome, Genome Res. 6 (1996) 371–391.

[80] Sargent T.D., Isolation of differentially expressed genes, Methods Enzymol. 152 (1987) 423–432.

[81] Sargent C.A., Affara N.A., Bentley E., Pelmear A., Bailey D.M., Davey P., Dow D., Leversha M., Aplin H., Besley G.T. et al., Cloning of the X-linked glycerol kinase deficiency gene and its identification by sequence comparison to the Bacillus subtilis homologue, Hum. Mol. Genet. 2 (1993) 97–106.

[82] Schena M., Shalon D., Davis R.W., Brown P.O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[83] Schena M., Shalon D., Heller R., Chai A., Brown P.O., Davis R.W., Parallel human genome analysis: microarray-based expression monitoring of 1000 genes, Proc. Natl. Acad. Sci. USA 93 (1996) 10614–10619.

[84] Schuler G.D., Boguski M.S., Stewart E.A., Stein L.D., Gyapay G., Rice K., White R.E., Rodriguez-Tome P., Aggarwal A., Bajorek E. et al., A gene map of the human genome, Science 274 (1996) 540–546.

[85] Smith C.W., Patton J.G., Nadal-Ginard B., Alternative splicing in the control of gene expression, Annu. Rev. Genet. 23 (1989) 527–577.

[86] Soares M.B., Bonaldo M.F., Jelene P., Su L., Lawton L., Efstratiadis A., Construction and characterization of a normalized cDNA library, Proc. Natl. Acad. Sci. USA 91 (1994) 9228–9232.

[87] Tiranti V., Savoia A., Forti F., Dapolito M.F, Centra M., Racchi M., Zeviani M., Identification of the gene encoding the human mitochondrial RNA polymerase (h-mtRPOL) by cyberscreening of the expressed sequence tags database, Hum. Mol. Genet. 6 (1997) 615–625.

[88] Tosser-Klopp G., Benne F., Bonnet A., Mulsant P., Gasser F., Hatey F., A first catalogue of genes involved in pig ovarian follicular differentiation, Mamm. Genome 8 (1997) 250–254.

[89] Tosser-Klopp G., Mulsant P., Yerle M., Regional localisations of VIM, HSD3b, ACTA1 and PGM1 in pigs, Anim. Genet. 29 (1998) 23–26.

[90] Tugendreich S., Boguski M.S., Seldin M.S., Hieter P., Linking yeast genetics to mammalian genomes: identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base, Proc. Natl. Acad. Sci. USA 90 (1993) 10031–10035.

[91] Tugendreich S., Bassett D.E., Jr., McKusick V.A., Boguski M.S., Hieter P., Genes conserved in yeast and humans, Hum. Mol. Genet. 3, Spec. No. (1994) 1509–1517.

[92] Vasmatzis G., Essand M., Brinkmann U., Lee B., Pastan I., Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis, Proc. Natl. Acad. Sci. (USA) 95 (1998) 300–304.

[93] Velculescu V.E., Zhang, L, Vogelstein, B, Kinzler, K.W., Serial analysis of gene expression, Science 270 (1995) 484–487.

[94] Velculescu V.E., Zhang L., Zhou W., Vogelstein J., Basrai M.A., Bassett D.E., Jr, Hieter, P., Vogelstein B., Kinzler K.W., Characterization of the yeast transcriptome, Cell 88 (1997) 243–251.

[95] Waterston R., Martin C., Craxton M., Huynh C., Coulson A., Hillier L., Durbin R., Green P., Shownkeen R., Halloran N. et al., A survey of expressed genes in Caenorhabditis elegans, Nat. Genet. 1 (1992) 114–123.

[96] Wilcox A.S., Khan A.S., Hopkins J.A., Sikela J.M., Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome, Nucleic Acids Res. 19 (1991) 1837–1843.

[97] Winterø A.K., Fredholm M., Davies W., Evaluation and characterization of a porcine small intestine cDNA library: Analysis of 839 clones, Mamm. Genome 7 (1996) 509–517.

[98] Wolfsberg T.G., Landsman D., A comparison of expressed sequence tags (ESTs) to human genomic sequences, Nucleic Acids Res. 25 (1997) 1626–1632.

[99] Yerle M., Echard G., Robic A., Mairal A., Dubut-Fontana C., Riquet J., Pinton P., Milan D., Lahbib-Mansais Y., Gellin J., A somatic cell hybrid panel for pig regional gene mapping characterized by molecular cytogenetics, Cytogenet. Cell Genet. 73 (1996) 194–202.

[100] Zhang L., Zhou W., Velculescu V.E., Kern S.E., Hruban R.H., Hamilton S.R., Vogelstein B., Kinzler K.W., Gene expression profiles in normal and cancer cells, Science 276 (1997) 1268–1272.

[101] Zhao N., Hashida H., Takahashi N., Misumi Y., Sakaki Y., High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression, Gene 156 (1995) 207–213.