Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues

Armin O. Schmitt*, Thomas Specht, Georg Beckmann, Edgar Dahl, Christian P. Pilarsky, Bernd Hinzmann and André Rosenthal

metaGen Gesellschaft für Genomforschung mbH, Ihnestraße 63, D-14195 Berlin, Germany

Received June 14, 1999; Revised and Accepted September 8, 1999

ABSTRACT

A four-step procedure for the efficient and systematic mining of whole EST libraries for differentially expressed genes is presented. After eliminating redundant entries from the EST library under investigation (step 1), contigs of maximal length are built upon each remaining EST using about 4 000 000 public and proprietary ESTs (step 2). These putative genes are compared against a database comprising ESTs from 16 different tissues (both normal and tumour affected) to determine whether or not they are differentially expressed (step 3; electronic northern). Fisher's exact test is used to assess the significance of differential expression. In step 4, an attempt is made to characterise the contigs obtained in the assembly through database comparison. A case study of the CGAP library NCI_CGAP_Br1.1, a library made from three (well, moderately, and poorly differentiated) invasive ductal breast tumours (2126 ESTs in total) was carried out. Of the maximal contigs, 139 were found to be significantly (α = 0.05) over-expressed in breast tumour tissue, while 13 appeared to be down-regulated.

INTRODUCTION

Expressed Sequence Tags (ESTs), single pass reads from randomly selected cDNA clones (1,2), have proved to be a valuable resource for genome research (3–7). In many branches of genetic research, EST libraries serve as a gateway for the detection and characterisation of new candidate genes. Virtually all EST libraries are tissue specific and their mode of preparation is documented. Database entries of EST sequences carry a label in their header which permits the identification of their source tissue, a feature which is of major interest in the present study. EST libraries of many human tissues of various developmental stages, both normal and diseased, have been and are still being established and partly made accessible to the public.

With the average size of EST libraries ranging typically between 1000 and 10 000 entries, an EST library cannot be regarded as faithfully representing the gene expression pattern of a tissue (8). It is estimated that, varying with the cell type, between 10 000 and 30 000 different genes are expressed in a cell, with an average of about 300 000 mRNA molecules per cell.

Therefore, an EST library cannot be more than a coarse grained snapshot of the mRNA composition of a certain tissue at a certain time. Especially, the representation of low abundance genes in an EST library cannot be taken for granted. However, the availability of many EST libraries derived from the same type and state of tissue mitigates this problem inasmuch as pools of equivalent EST libraries can be created. The EST numbers of such pools reaching tens of thousands, a proportional representation of all abundant and moderately expressed genes can be assumed.

Following the paradigm that a cell is, to a large extent, characterised by its transcript composition, and that the amount of a protein, the actual biochemical agent, is positively correlated with the abundance of its mRNA, such EST pools enable us to carry out a meaningful statistical expression analysis *in silico* (9,10). Of course, for mRNA expression analysis only non-normalised EST libraries are eligible for pooling; normalised or subtracted libraries have to be excluded.

It is the purpose of this work to present a method which permits exhaustive exploitation of the information relevant to biologists and pharmacologists hidden in EST libraries. We are especially interested in genes which are significantly up- or down-regulated in diseased tissue. The panel of genes resulting from such analysis may serve as candidates for therapeutic or diagnostic targets (11).

The basic method to hunt for differentially expressed genes would be to carry out a simple sequence comparison procedure with a standard sequence comparison program like BLAST (12). Each EST of a tumour library is searched against all other ESTs from that library and against a database of ESTs from the benign counterpart. Comparing the number of homologous sequences (hits) found in either of the two libraries, benign and tumour, (with consideration of the potentially different library sizes) would give a first clue if an EST represents a differentially expressed gene.

The protocol presented here goes beyond this first approach in two ways. (i) The counts obtained in the above sketched EST–EST comparison procedure will tend to remain low, because only direct matches of the query EST sequence can be

*To whom correspondence should be addressed. Tel: +49 30 8413 1668; Fax: +49 30 8413 1674; Email: armin.schmitt@metagen.de
Present address:
Georg Beckmann, Hoechst Marion Roussel, 102 Route de Noisy, F-93235 Romainville, France

detected. The observed hit number might not be sufficient to secure differential expression statistically. Since cDNA sequences can extend over several kilobases and since EST sequences rarely exceed a couple of hundred base pairs, many ESTs belonging to a gene will remain unnoticed. (For the results of a control study see Results.) In order to catch all available ESTs belonging to a gene it is desirable to use full-length sequences for the search process whenever possible. Therefore, we applied an iterative search and assembly procedure (see AUTEX: a protocol for the automatic extension of partial DNA sequences) to build *in silico* contigs of maximal length using all available ESTs. (ii) Building maximal contigs and pooling EST libraries still does not guarantee hit numbers high enough for easy and unequivocal identification of differential expression. The identification of changes in expression level based upon quotients of very small relative abundances is not very meaningful. Therefore, we compute *P* values by Fisher's exact test (see Statistical evaluation of EST occurrences), which allow statistical assessment of any observed hit distribution, even if the hit numbers are very low. A *P* value can be interpreted as the degree of certainty that an observed differential expression is not an artifact due to statistical fluctuations.

MATERIALS AND METHODS

Creating a set of non-redundant ESTs

Two similar EST sequences are very likely to match the same sequences in an EST database search, and, therefore, the *in silico* contigs built from their matches would also probably be almost identical. To save computer resources, it is advisable to extract a set of non-redundant sequences from the EST library under investigation. Only the sequences of this non-redundant set would serve as seeds (i.e. starting sequences) for the automatic extension procedure expounded below. We used an all-versus-all comparison to eliminate redundant ESTs from the initial set. The sequences of the library under investigation were sorted by decreasing lengths and were successively taken as query sequence for a BLAST search against all shorter sequences. Whenever the shorter sequence perfectly covered a third of the length of the longer one it was considered redundant and eliminated.

Identification of homologous sequences

We used the well-known basic alignment program BLAST (12) to find homologous sequences in a database. To increase sensitivity, we used a new BLAST implementation (v.2.0.5) which tolerates gaps in the sequence alignments (13). Our choice for the stringency parameters defining sequence homology were 10^{-4} for the *E* value and 95% sequence identity. For the characterisation of the contigs BLAST searches against the nucleotide database nt were run (14).

Assembly of ESTs

We used the program GAP (15) for the assembly of ESTs. In a first round, 2% mismatches between the EST sequences were allowed; this relatively stringent choice was made to guarantee a stable backbone to the assembly. Since the sequencing error rate of ESTs can amount to up to 5% towards their 3′-ends, the stringency was lowered to 5% mismatches in a second round to collect as many ESTs as possible.

AUTEX: a protocol for the automatic extension of partial DNA sequences

We developed an iterative procedure, AUTEX, to build contigs of maximal length based upon arbitrary partial DNA sequences (seed sequences). An initial BLAST search of the seed sequence is performed against our repository of about 4 000 000 ESTs: the human division of dbEST (16) (~1 000 000) and a proprietary EST database (Incyte Pharmaceuticals, Palo Alto, CA) (~3 000 000). The sequences homologous to the seed sequence (parameters as in Identification of homologous sequences) are extracted from the databases and assembled as described in Assembly of ESTs. The consensus sequence is derived from the resulting multi-sequence alignment and taken as query sequence for another BLAST search against the EST databases. This procedure of alternating BLAST search and assembly of matching sequences is repeated until the consensus sequence has reached its maximal length. Approaches similar to AUTEX have been described by Ebeling *et al*. (20) and Gill *et al*. (21), and the feasibility of generating full-length gene sequences from ESTs using such a method was demonstrated by Prigent *et al*. (22).

Electronic northerns

Thanks to the specification of the source library in the headline of most EST database entries the source tissue of origin and related information, e.g. mode of cDNA library preparation or disease state of the source tissue, can be traced back. Therefore, a report generated in a BLAST search of a given DNA sequence against an EST database consisting of such 'pedigree' ESTs allows the inference of that gene's tissue distribution. In analogy to the corresponding laboratory method, this analysis is called electronic northern. The paucity of ESTs in a typical library makes it necessary to lump together EST libraries of common origin to create EST pools.

Demanding for a given tissue a pool of ESTs from both tumour and normal tissue and setting a minimum of 10 000 ESTs in a pool, we could create pool pairs for 16 tissues. One pool pair thus consists of all available ESTs from a tissue in both the normal and diseased states. Two hundred and sixty EST libraries from normal tissues and 172 EST libraries from tumour or cancerous tissues contributed to the EST pools. We mention in passing that pool pairs can be generated from public EST libraries alone for the following eight tissues: brain, breast, colon, kidney, lung, ovary, prostate and uterus.

In our electronic northern blot the tissue-specific relative abundance of a gene is defined as the ratio of number of homologous ESTs and total number of ESTs in the corresponding pool. Relative abundance figures were determined for normal and tumour pools separately, and the ratio of normal and tumour relative abundances, the so-called expression ratio or fold change, is used as a measure for the up- or down-regulation of a gene in tumour tissue with respect to normal tissue. For convenience, we introduce a classification scheme for the degree of differential expression comprising three levels: moderate, strong and very strong (see Table 1).

Statistical evaluation of EST occurrences

In order to assess the distribution of hits between normal and affected tissues observed in a BLAST search, we applied Fisher's exact test (17), a statistical test widely used for the evaluation of 2 × 2 contingency tables, i.e. representations of yes/no outcomes obtained from two disjoint samples. The outcome of Fisher's exact test is a significance value *P* ranging between 0 and 1 which describes the likelihood of the null hypothesis being true: 'The frequency of an event is the same in either of two samples' or, in our specific example 'The frequency of a gene is the same in normal and in diseased tissue'. The closer the significance value is to 1.0 the more the observations are compatible with the null hypothesis. A *P* value close to 0, on the other hand, is indicative of significant differential expression of the gene under consideration. Fisher's exact test is a conservative test as compared to other statistical tests (18), therefore selection of genes for further investigation based upon the criterion of small *P* values can be considered restrictive.

Table 1. A general classification scheme for degrees of differential expression and listing of all cases of differential expression found in our case study

Class	Degree	Expression ratio	α	Regulation in tumour		Σ
				Down	Up	
I	Very strong	>10	<0.001	0	15	15
II	Strong	>5	<0.01	2	17	19
III	Moderate	>2	<0.05	11	107	118
Σ				13	139	152

Two independent criteria, expression ratio and *P* value computed by Fisher's exact test are used to classify the degree of a gene's differential expression. In general, a gene is classified in the highest class (order: highest I, lowest III) whose criteria are both met. For example, a gene exhibiting an expression ratio of 12 and a *P* value of 0.0005 (<α = 0.001) would be class I. An expression ratio of 12 substantiated with a *P* value of 0.005 (<α = 0.01), however, would be classified II. If the expression ratio is not defined due to absence of hits in one of the EST pools, the *P* value serves as the only classification criterion. Our case study revealed that up-regulation of genes is 10 times more frequent in an EST library prepared from tumour tissue than down-regulation (139 versus 13 cases).

RESULTS

We performed a case study of the CGAP library NCI_CGAP_Br1.1 (provided by B. Soares and M.F. Bonaldo), a non-normalised EST library made from three pooled invasive ductal breast carcinomas (19) comprising 2126 entries.

Upon removal of 661 redundant ESTs (see above) we were left with a set of 1465 unique ESTs, which served as seed sequences for the automatic elongation by the software protocol AUTEX (see above). Almost all (1456/1465) of the seed ESTs found homologous ESTs in our concatenated EST database (Incyte and dbEST), and could, hence, be assembled and elongated. Nine seed ESTs did not find any matching sequences; they remained singletons. In 50 cases, elongated sequences were found to overlap with other elongated sequences. In other words, 50 non-overlapping pairs of ESTs were each derived from the same mRNA sequence and collapsed during the elongation process into one and the same contig. The median length grew from 418 bp for the initial EST set to 721 bp for the set of elongated sequences (Fig. 1).

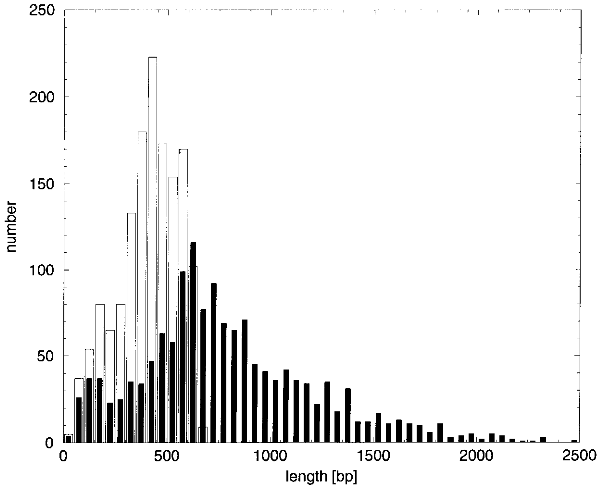

Figure 1. The length distributions for the initial set of ESTs (open bars) and for the contigs of maximal length (black bars) obtained in the automatic elongation procedure (see text) are shown. The median length nearly doubled from 418 to 721 bp. Binning width was 50 bp.

We performed a simple test to assess the validity of our assembly approach. Chimeric assemblies can be identified by inspection of the reports generated by a BLAST search against a database of known sequences. An assembly containing ESTs from two different known genes generates a contradictory BLAST report listing perfect matches to these two genes. Inspection of the BLAST reports suggests a rate of less than 5% cases of chimeric assemblies, a finding which is in accord with our experience gathered in other analyses.

Expression analysis was performed with the non-redundant set of 1406 contigs of maximal length. Of these, 152 proved to be differentially expressed in normal and tumour breast tissue, i.e. they exhibit an expression ratio of at least 2 and their *P* value computed by Fisher's exact test was below 0.05. Up-regulation in tumour tissue is found more than 10 times more frequently than down-regulation (139 versus 13 genes). This finding confirms our expectation that the EST library under investigation should be enriched with genes which are over-expressed in breast tumour tissue. The classification of the differentially expressed genes is listed in detail in Table 1 and the results of the nt database consultation for these sequences are given in Table 2.

One hundred and twenty-two genes matched database entries with an *E* value of 10^{-100} or better. Twelve of the 152 differentially expressed genes are of mitochondrial origin (plus one marginal hit), while one is a ribosomal gene. While it is well known that EST libraries can contain mitochondrial genes (23,24), the number can vary due to the mode of preparation. It is striking that most (eight out of 12) mitochondrial genes were classified as very strongly differentially expressed (class I). Two EST libraries in the breast tumour EST pool, the seed library NCI_CGAP_Br1.1 itself and NCI_CGAP_Br3, were identified as generating most of the hits. Differential expression of these mitochondrial genes might thus be an artifact of EST library preparation. On the other hand, up-regulation of mitochondrial genes in cancerous tissues has been reported

Table 2. Database comparison of 152 assembled sequences differentially expressed in breast

Accession Number	Description	Score	E-value	Similarity
Class I^+: Very strong over-expression in tumor tissue				
J01415	Human mitochondrion, complete genome	1154	0.0	608/618
M10546	Human mitochondrial DNA, fragment M1	942	0.0	506/515
AF006084	Homo sapiens Arp2/3 protein complex subunit	989	0.0	506/515
D83253	Homo sapiens genomic DNA, chromosome 21q22.2	42	0.23	21/21
J03801	Human lysozyme mRNA, complete cds with an Alu	961	0.0	519/525
X62996	H.sapiens mitochondrial genome	1439	0.0	757/770
X55654	H.sapiens mitochondrial coxII mRNA	926	0.0	491/498
AF042511	Homo sapiens isolate Asn2 cytochrome b	1003	0.0	546/554
J01415	Human mitochondrion, complete genome	1292	0.0	688/696
U85625	Homo sapiens ribonuclease 6 precursor, mRNA	1661	0.0	879/886
M10546	Human mitochondrial DNA, fragment M1	111	6e-23	62/64
X93334	H.sapiens mitochondrial DNA, complete genome	874	0.0	492/505
M19045	Human lysozyme mRNA, complete cds	1439	0.0	726/726
X93334	H.sapiens mitochondrial DNA, complete genome	1148	0.0	629/643
M94345	Homo sapiens macrophage capping protein mRNA	1201	0.0	609/610
Class II^-: Strong down-regulation in cancer tissue				
AC004439	Drosophila melanogaster DNA sequence P1 DS	44	0.062	22/22
D10040	Homo sapiens mRNA for long-chain acyl-CoA	2010	0.0	1024/1026
Class II^+: Strong over-expression in cancer tissue				
X62996	H.sapiens mitochondrial genome	1110	0.0	596/612
AC004222	Homo sapiens chromosome 17, clone HCIT499I2	44	0.053	28/30
L15203	Human secretory protein (P1.B) mRNA	950	0.0	479/479
D26077	Mouse mRNA for KIF3B protein, complete cds	44	0.047	31/34
J04823	Human cytochrome c oxidase subunit VIII	906	0.0	471/473
X13238	Human mRNA for cytochrome c oxidase subunit VIc	777	0.0	409/412
AF064603	Homo sapiens GA17 protein mRNA, complete cds	1215	0.0	656/673
X67951	H.sapiens mRNA for proliferation-associated	1423	0.0	852/892
D38048	Human mRNA for proteasome subunit z, complete cds	1092	0.0	583/591
U03398	Human receptor 4-1BB ligand mRNA, complete cds	440	e-121	236/243
Z69710	Human DNA sequence from cosmid L98A6	914	0.0	464/465
L47647	Homo sapiens creatine kinase B mRNA	1187	0.0	643/656
X82818	H.sapiens PTP1C/HCP gene	452	e-125	231/232
X63789	T.thermophila genes for snRNA U5-1, snRNA U5-2	48	0.004	24/24
X62996	H.sapiens mitochondrial genome	1122	0.0	566/566
M12075	Human estrogen receptor mRNA, partial cds	147	2e-33	83/86
M24470	Human glucose-6-phosphate dehydrogenase	129	1e-27	200/245

previously (25,26). Thirteen of the genes in class I are represented in public databases, while an additional two exhibit weak homologies to known genes.

Apart from some genes associated with the increased metabolic activity of tumour cells, class II contains several genes which are known to be associated with various kinds of solid tumours. Two of three members of the trefoil gene family are represented in class II. Each of these secreted molecules is abundantly expressed in a broad but specific set of tumours (27). TFF1 (formerly pS2) was found to be overexpressed, for example, in carcinomas of the breast, pancreas, endometrium, bladder, prostate and lung, while TFF3 (formerly P1.B) is overexpressed in carcinomas of the intestine and invasive lobular and ductal carcinoma (28). Further examples of prominent tumour-associated genes found in class II are the proliferation-associated gene involved in breast cell proliferation (29) and the creatin kinase B gene, which is abundantly expressed in small cell lung carcinoma (30).

Class III can be considered a suitable source to find new tumour-associated genes since about 20 of the 118 genes in this class exhibit no or only very faint similarity to known genes. Several genes from class III identify tumour-associated genes which have been identified by ‘wet experiments’ only very recently. The transmembrane protein NET-4, for instance, belongs to the family of tetraspan proteins that have been implicated in the prognosis of different types of tumours such as lymphomas (31) and bladder cancer (32). The adult T cell leukemia-derived factor 1 (ATL-1), or thioredoxin, exhibits increased expression in gastric cancer cells that are resistant to the chemotherapeutic agent cis-diaminedichloroplatinum (33). Cathepsin D overexpression in breast cancer cells was shown to be associated with increased risk of early relapse (34) and

Table 2. *continued*

Accession Number	Description	Score	E-value	Similarity
Class III^-: Moderate down-regualtion in cancer tissue				
AF006515	Homo sapiens CHD3 mRNA, complete cds	190	9e-46	162/176
M34671	Human lymphocytic antigen CD59/MEM43 mRNA	1249	0.0	630/630
X07819	Human pump-1 mRNA homolog. to metalloproteina	1994	0.0	1019/1022
X04412	Human mRNA for plasma gelsolin	1558	0.0	805/809
J02874	Human adipocyte lipid-binding protein	1180	0.0	584/589
X53743	H.sapiens mRNA for fibulin-1 C	2200	0.0	1130/1134
V01512	Human cellular oncogene c-fos (complete sequence)	938	0.0	473/473
L23077	Rattus rattus zinc finger protein, complete	42	0.31	21/21
SPAC15A10	S.pombe chromosome I cosmid c15A10	40	0.58	20/20
X07696	Human mRNA for cytokeratin 15	1392	0.0	702/702
M12759	Human Ig J chain gene, exons 3 and 4.	420	e-115	225/228
Class III^+: Moderate over-expression in cancer tissue				
D21261	Human mRNA for KIAA0120 gene, complete cds	1643	0.0	842/845
X72580	H.sapiens mRNA for collagen X, exon 3	1152	0.0	584/585
X05344	Human mRNA for cathepsin D from oestrogen	422	e-116	327/357
U80030	Caenorhabditis elegans cosmid K12D9	42	0.14	21/21
U83115	Human non-lens beta gamma-crystallin	656	0.0	331/331
U11861	Human G10 homolog (edg-2) mRNA, complete cds.	1189	0.0	750/787
L16991	Human thymidylate kinase (CDC8) mRNA	1972	0.0	1024/1031
X77584	H.sapiens mRNA for ATL-derived factor	930	0.0	520/535
M23114	Homo sapiens calcium-ATPase (HK1) mRNA	767	0.0	447/472
K00558	human alpha-tubulin mRNA, complete cds	1300	0.0	730/748
AC004934	Homo sapiens PAC clone DJ0953B05 from 7p12	44	0.050	22/22
AB015476	Arabidopsis thaliana genomic DNA	44	0.057	22/22
J03607	Human 40-kDa keratin intermediate filament	1887	0.0	985/992
J05176	Human alpha-1-antichymotrypsin mRNA, 3' end	1011	0.0	533/538
Z11566	H.sapiens mRNA for Pr22 protein	1225	0.0	639/646
X95404	H.sapiens mRNA for non-muscle type cofilin	1063	0.0	624/656
L12168	Homo sapiens adenylyl cyclase-associated	706	0.0	372/380
J01415	Human mitochondrion, complete genome	1279	0.0	760/789
U03271	Human F-actin capping protein beta subunit	1544	0.0	833/843
AL021708	H.sapiens partial cDNA homologous to M.musculus JIP-1 gene	3150	0.0	1640/1653
AF086624	Rattus norvegicus serine threonine kinase	96	6e-18	106/124
AJ223352	Homo sapiens mRNA for for histone H2B	1503	0.0	781/786
D13666	Homo sapiens mRNA for osteoblast	1116	0.0	581/590
X52003	H.sapiens pS2 protein gene	952	0.0	480/480
AF086332	Homo sapiens full length insert cDNA	127	3e-27	106/120
L42088	Homo sapiens (subclone 3_e11 from P1 H17)	188	2e-45	234/277
AF052578	Homo sapiens androgen receptor associated	1275	0.0	665/670
L76200	Human guanylate kinase (GUK1) mRNA, complete cds	1618	0.0	836/840
M12670	Human fibroblast collagenase inhibitor mRNA,	1065	0.0	583/593
Z48605	H.sapiens partial mRNA for pyrophosphatase	680	0.0	343/343
S61826	hinge=OXPHOS system complex III mitochondrial	985	0.0	500/501

metastasis (35). Finally, overexpression of the human H19 gene has been described in breast adenocarcinoma (36). Imprinting and maternal expression of H19 is lost in a variety of cancers (37). Interestingly, an important function of the untranslated H19 RNA is the regulation of IGF-II expression (37), which in turn seems to regulate the routing of the cathepsin D gene product (38) mentioned above. This combined finding of genes which may belong to the same biochemical pathway demonstrates the value of our *in silico* approach and supports the expectation that further important cancer-related genes may be present within the cohort of the remaining 20 new candidates. However, since class III includes genes exhibiting an expression ratio of as low as 2, a factor which is commonly believed to be compatible with normal fluctuations, some of the genes will certainly be false positives.

Electronic northerns for three of the genes obtained in our assembly procedure, the TFF3 (formerly P1.B or ITF) gene coding for a secreted protein (39,40), the mammaglobin gene

Table 2. *continued*

Accession Number	Description	Score	E-value	Similarity
AL010260	Plasmodium falciparum DNA *** SEQUENCING	44	0.038	31/34
AF029890	Homo sapiens hepatitis B virus X interactin	1138	0.0	595/598
X97744	P.pygmaeus DNA for low affinity N-formyl	44	0.048	22/22
M26038	Human MHC class II DR beta mRNA, complete cds	1296	0.0	716/730
AC004981	Homo sapiens PAC clone DJ1159C10 from 7q34	42	0.19	24/25
AF072759	Mus musculus fatty acid transport protein	981	0.0	846/963
U32944	Human cytoplasmic dynein light chain 1	1261	0.0	639/640
M29873	Human cytochrome P450-IIB (hIIB3) mRNA	926	0.0	479/483
U46751	Human phosphotyrosine independent ligand	934	0.0	471/471
D50372	Homo sapiens mRNA for myosin regulatory light	1187	0.0	661/678
AF004877	Homo sapiens pro-alpha 2(I) collagen	1025	0.0	531/538
AF023268	Homo sapiens clk2 kinase (CLK2), propin1	458	e-127	285/303
AF087017	Homo sapiens H19 gene, complete sequence	585	e-165	311/315
AF065389	Homo sapiens tetraspan NET-4 mRNA, complete	121	5e-25	61/61
M29063	Human hnRNP C2 protein mRNA	1146	0.0	615/624
Z81588	Caenorhabditis elegans cosmid T07D10	42	0.10	27/29
J04164	Human interferon-inducible protein 9-27 mRNA	1348	0.0	683/684
U72063	Human immunoglobulin kappa chain constant	72	3e-11	39/40
Z98046	Human DNA sequence from clone 14O9	882	0.0	445/445
M37061	P.knowlesi Mbn-cutting sites	44	0.047	22/22
J03607	Human 40-kDa keratin intermediate filament	1090	0.0	560/562
X17644	Human GST1-Hs mRNA for GTP-binding protein	1211	0.0	624/627
M63138	Human cathepsin D (catD) gene, exons 7, 8	1170	0.0	613/618
X04968	E. coli miniF plasmid gene pifC for C protein	40	0.68	20/20
Y14551	Homo sapiens mRNA for DIF-2 protein	1043	0.0	538/542
AF080118	Arabidopsis thaliana BAC F8M12	40	0.43	20/20
AF028823	Homo sapiens Tax interaction protein 1 mRNA	1292	0.0	666/673
X65923	H.sapiens fau mRNA	666	0.0	339/340
U15008	Human SnRNP core protein Sm D2 mRNA	916	0.0	469/470
X78687	H.sapiens G9 gene encoding sialidase	1273	0.0	694/706
AF073298	Homo sapiens 4F5rel mRNA, complete cds	991	0.0	500/500
J03909	Human gamma-interferon-inducible protein	1070	0.0	626/652
U41060	Human breast cancer, estrogen regulated LIV-1	1538	0.0	824/832
L10284	Homo sapiens integral membrane protein	1134	0.0	591/596
AB014563	Homo sapiens mRNA for KIAA0663 protein	1094	0.0	558/560
AF038952	Homo sapiens cofactor A protein mRNA	755	0.0	381/381
L38939	Homo sapiens GT233 mRNA	979	0.0	500/502
X57522	H.sapiens RING4 cDNA	2026	0.0	1042/1046
X14420	Human mRNA for pro-alpha-1 type 3 collagen	1370	0.0	728/744
M24194	Human MHC protein homologous to chicken B	430	e-118	248/257
X79882	H.sapiens lrp mRNA	638	0.0	336/338
V00572	Human mRNA encoding phosphoglycerate kinase	1398	0.0	776/789
U70660	Human copper transport protein HAH1 (HAH1)	860	0.0	444/446

and a gene homologous to the mouse JIP-1 gene (accession nos L15203, U33147 and AL021708, respectively), are depicted in Tables 3, 4 and 5, respectively.

DISCUSSION

We have presented an integrated procedure for the mining of EST libraries for genes differentially expressed in normal and tumour tissues. This procedure comprises four steps: normalisation of an EST library of interest yielding a set of non-redundant ESTs (seeds); iterative assembly of all available ESTs homologous to the seeds (AUTEX); generation of electronic northerns for the resulting elongated sequences; and characterisation of the assemblies through database searches. Almost 10% (139/1406) of the assembled sequences were found to be significantly overexpressed (expression ratio at least 2, $\alpha = 0.05$) in breast tumour tissue and less than 1% (13/1406) were down-regulated in tumour tissue. This shows clearly that EST libraries obtained from tumour tissue are enriched with up-regulated genes. Both the set of up- and down-regulated genes contain candidate genes which could be involved in the development of tumours.

Table 2. *continued*

Accession Number	Description	Score	E-value	Similarity
M28204	Homo sapiens (clone pMF28) MHC class I	1263	0.0	706/725
Z21507	H.sapiens EF-1delta gene	1051	0.0	546/550
X98130	A.thaliana 81kb genomic sequence	48	0.008	27/28
M30684	Gorilla gorilla beta-2-microglobulin	1108	0.0	619/635
J03607	Human 40-kDa keratin intermediate filament	967	0.0	547/560
U61141	Mesocricetus auratus LIM-homeodomain protein	44	0.020	22/22
L27211	Human CDK4-inhibitor (p16-INK4) mRNA	963	0.0	556/574
L38939	Homo sapiens GT233 mRNA	979	0.0	500/502
AC002500	Human Cosmid g5129z101 from 7q31.3	40	0.63	23/24
X57351	Human 1-8D gene from interferon-inducible gene	1185	0.0	663/678
M57567	Human ADP-ribosylation factor (hARF5) mRNA	963	0.0	506/516
Y00503	Human mRNA for keratin 19	1328	0.0	766/782
U13665	Human cathepsin O (CTSO) mRNA, complete cds	1802	0.0	944/953
AC004472	Homo sapiens chromosome 9, P1 clone 11659	438	e-121	221/221
U90915	Human clone 23600 cytochrome c oxidase subunit	1041	0.0	566/576
X02761	Human mRNA for fibronectin (FN precursor)	1308	0.0	722/736
X63527	H.sapiens mRNA for ribosomal protein L19	1378	0.0	695/695
AC004686	Homo sapiens chromosome 17, clone hRPC.1073	1651	0.0	883/908
L26245	Human effector cell protease receptor-1	525	e-147	272/273
U09813	Human mitochondrial ATP synthase subunit 9	1281	0.0	646/646
AC004728	Drosophila melanogaster DNA sequence	82	1e-13	68/77
X14034	Human mRNA for phospholipase C	3172	0.0	1644/1656
AF053944	Homo sapiens aortic carboxypeptidase-like	607	e-172	306/306
M26252	Human TCB gene encoding cytosolic thyroid	1140	0.0	575/575
U70734	Human 38 kDa Mov34 isologue mRNA, complete cds	2490	0.0	1272/1276
U28249	Human 11kd protein mRNA, complete cds	1251	0.0	648/651
X87689	H.sapiens mRNA for putative p64 CLCP protein	2216	0.0	1140/1146
M12670	Human fibroblast collagenase inhibitor mRNA	1326	0.0	748/764
U33147	Human mammaglobin mRNA, complete cds	961	0.0	498/501
X13546	Human HMG-17 gene for non-histone	1318	0.0	665/665
Y00503	Human mRNA for keratin 19	1425	0.0	770/779
M68867	Human cellular retinoic acid-binding protein	1419	0.0	747/756
M58664	Homo sapiens CD24 signal transducer mRNA	2064	0.0	1119/1139

The 152 elongated sequences which were found to be differentially expressed in normal and tumourous breast tissue were compared against the non-redundant NCBI database nt. The accession no., abbreviated description, BLAST score, and *E* value of and similarity to the highest scoring sequence are given. The order of presentation is: sequences with very strong (class I), strong (class II) and moderate differential expression (class III). There is no apparent order within the classes.

We introduced a classification scheme of differential expression comprising three levels, moderate (class III), strong (class II) and and very strong (class I), to describe the degree of differential expression. Interestingly, many genes of class I turned out to be involved in cell metabolism or tumour infiltration, whereas many well-known tumour-related genes were classified in classes II and III. Thus, on the one hand, the degree of over-expression does not seem to reflect the significance of a gene in tumourigenesis. On the other hand, there is good reason to believe that, apart from the known tumour-associated genes, our set contains further, tumour-associated genes.

Due to the enormous variance in size of tissue-specific EST pools (ranging from about 10 000 to almost 180 000 ESTs), absolute numbers of matching ESTs found in BLAST searches do not suffice to assess whether or not a gene is differentially expressed. Even the expression ratio, i.e. the quotient of the relative abundances in normal and tumour tissues, does not correct these circumstances entirely. Expression ratios derived from low hit numbers do not allow statistically solid conclusions, whereas the same expression ratio can be statistically significant if derived from higher hit numbers. For example, a hit ratio of 5:1 observed in two pools each of 10 000 ESTs is associated in Fisher's exact test (two-tailed) with a *P* value of 0.2187, while a hit ratio of 10:2 is statistically substantiated by a *P* value of 0.0224. Thus, one and the same expression ratio 5 is statistically non-significant by traditional criteria in the first case, whereas it is statistically significant in the second case at the level $\alpha = 0.05$. We found the *P* value to be especially helpful when one of the two pools was completely devoid of hits. Then the expression ratio is not defined, but the *P* value computed by Fisher's exact test

Table 3. Electronic northern for the TFF3 gene

Tissue	Benign		Tumour		*P* value	Expression ratio	Class
	Abundance	Pool size	Abundance	Pool size			
Bladder	0	25 643	0	42 556			
Brain	2	178 865	0	100 254	0.540		
Breast	11	120 733	37	67 587	7.4×10^{-9}	6.0^{+}	II^{+}
Colon	49	52 193	22	35 113	0.117	1.5^{-}	
Endocrine tissue	4	62 286	9	60 780	0.175	2.3^{+}	
Kidney	0	44 687	0	20 743			
Liver	3	21 521	6	15 763	0.181	2.7^{+}	
Lung	2	102 767	14	54 087	1.8×10^{-5}	13.3^{+}	I^{+}
Muscle/skeleton	0	58 355	0	27 075			
Ovary	13	33 707	9	41 943	0.200	1.8^{-}	
Pancreas	0	60 518	3	21 810	0.019		
Prostate	10	106 113	43	76 774	9.0×10^{-9}	5.9^{+}	II^{+}
Stomach/oesophagus	0	13 801	2	12 120	0.219		
Testis	0	24 906	0	16 900			
T lymphoma	1	60 743	0	17 318	1.000		
Uterus	2	67 669	3	21 740	0.096	4.7^{+}	

The expression levels in 16 normal and tumour tissues are shown for the TFF3 gene. The abundance is given as number of homologous ESTs found in a pool of tissue- and state-specific ESTs. The expression ratio is defined as ratio of the relative abundances (number of homologous ESTs divided by pool size). The greater of the two relative abundances (normal or tumour) is divided by the smaller so that expression ratios are always ≥1. The plus sign indicates up-regulation in tumour tissue, the minus sign down-regulation in tumour tissue. The *P* value computed by Fisher's exact test expresses the statistical validity of the differential expression (see text). For the definition of classes see Table 1. The TFF3 gene is an example of a gene which is up-regulated in several tumour tissues: about 5-fold in uterus, 6-fold in breast and prostate tumour, and 13-fold in lung. The differential expression in lung has, to our knowledge, not been reported so far.

still allows statistical assessment and classification. For example, the *P* values for the two hit distributions 5:0 and 6:0 are 0.0625 (not significant, unclassified) and 0.0312 (significant at the level $\alpha = 0.05$, class III), respectively (pools of 10 000 ESTs).

We were also interested in whether differential expression is detectable without prior assembly and to this end calculated electronic northerns from all 1465 ESTs of the non-redundant seed set. One hundred and fifty-five ESTs showed differential expression (expression ratio greater than 2, *P* value below 0.05) in breast tissue. However, only 100 of those differentially expressed ESTs were found among the seed ESTs which finally led to the set of 152 differentially expressed elongated sequences. This means that differential expression determined on the basis of non-elongated ESTs leads to rates of about 30% false positives and 30% false negatives.

For example, the EST which served as seed in the assembly resulting in the murine JIP-1 gene homologue did not collect any breast EST apart from itself ($P = 0.40$, unclassified), while it collected eight ESTs from normal brain libraries and one EST from a brain cancer library ($P = 0.17$, expression ratio 4.5, unclassified). This gene thus would not have been classified as differentially expressed by the stringent criteria used in this study. For comparison, the elongated sequence matched four ESTs in breast cancer and none in normal breast tissue ($P = 0.017$, class III^{+}), and 25 ESTs in normal brain and five ESTs in brain cancer tissue ($P = 0.036$, expression ratio 2.8, class III^{-}; see Table 3). As in Tables 3–5, the superscripts + and – indicate up- and down-regulation in tumour tissue, respectively.

We used for this case study our full repository of about 4 000 000 ESTs, 3 000 000 of which were from a proprietary database. The use of exclusively public ESTs permits the creation of EST pools large enough (at least 10 000 ESTs) for the calculation of electronic northerns for eight tissues, among them tissues of great pharmacological interest, like prostate, breast and lung. Since many (122 out of 152) of our assembled sequences represent known genes with an *E* value of 10^{-100} or better in BLAST searches against nt (Table 2), the procedure could be modified by restricting the assembly step to ESTs with no clear annotation and using the matching full-length gene sequence for the calculation of electronic northerns. Building assemblies for the remaining 30 unknown ESTs from only public ESTs is, according to our experience, possible, but results in poorer coverage of the assembly, hence in a somewhat poorer sequence quality and shorter sequence length.

The software for this analysis, essentially shell scripts controlling BLAST searches and EST sequence assemblies, was not optimised for speed; the time necessary for complete analysis of one EST depends very much on the number of

Table 4. Electronic northern for the human mammaglobin gene

Tissue	Benign		Tumour		*P* value	Expression ratio	Class
	Abundance	Pool size	Abundance	Pool size			
Bladder	0	25 643	0	42 556			
Brain	0	178 865	0	100 254			
Breast	49	120 733	98	67 587	4.0×10^{-14}	3.6^{+}	III^{+}
Colon	0	52 193	0	35 113			
Endocrine tissue	0	62 286	0	60 780			
Kidney	0	44 687	0	20 743			
Liver	0	21 521	0	15 763			
Lung	0	102 767	0	54 087			
Muscle/skeleton	0	58 355	0	27 075			
Ovary	0	33 707	0	41 943			
Pancreas	0	60 518	0	21 810			
Prostata	0	106 113	0	76 774			
Stomach/oesophagus	0	13 801	0	12 120			
Testis	0	24 906	0	16 900			
T lymphoma	0	60 743	0	17 318			
Uterus	0	67 669	0	21 740			

The expression levels in 16 normal and tumour tissues are shown for the mammaglobin gene. The abundance is given as number of homologous ESTs found in a pool of tissue- and state-specific ESTs. The expression ratio is defined as ratio of the relative abundances (number of homologous ESTs divided by pool size). The greater of the two relative abundances (normal or tumour) is divided by the smaller so that expression ratios are always ≥ 1. The plus sign indicates up-regulation in tumour tissue, the minus sign down-regulation in tumour tissue. The *P* value computed by Fisher's exact test expresses the statistical validity of the differential expression (see text). For the definition of classes see Table 1. The human mammaglobin gene is very specific to breast cells. Its moderate overexpression in breast cancer has been described by Watson and Fleming (41).

cycling steps taken to build a contig of maximal length and on the number of ESTs involved in the assembly step, but was estimated to be of the order of 10 min on a 250 MHz Sun Ultra SPARC processor.

In future studies, various steps could be undertaken to accelerate and improve the suggested protocol, e.g. a filtering procedure at the very beginning to avoid housekeeping genes being analysed. Further characterisation of the set of assembled sequences could include Pfam searches and translating searches against protein databases.

ACKNOWLEDGEMENTS

We thank Ralf Herwig (Max Planck Institute for Molecular Genetics, Berlin) and Henrik Seidel (Schering AG, Berlin) for valuable contributions, and Daryl Faulds (Berlex Biosciences, Richmond, CA), Robert Grützmann and Birgit Weber (metaGen) for critical review of the manuscript.

REFERENCES

1. Adams,M., Kelley,J., Gocayne,J., Dubnick,M., Polymeropoulos,H., Xiao,H., Merril,C., Wu,A., Olde,R., Moreno,R. *et al.* (1991) *Science*, **252**, 1651–1656.
2. Okubo,K., Hori,N., Matoba,R., Niiyama,T., Fukushima,A., Kojima,Y. and Matsubara,K. (1992) *Nature Genet.*, **2**, 173–179.
3. Sikela,J.M. and Aufray,C. (1993) *Nature Genet.*, **3**, 189–191.
4. Gautheret,D., Poirot,O., Lopez,F., Audic,S. and Claverie,J.-M. (1998) *Genome Res.*, **8**, 524–530.
5. Rawlings,J.R. and Searls,D.B. (1997) *Curr. Opin. Genet. Dev.*, **7**, 416–423.
6. Banfi,S., Borsani,G., Rossi,E., Bernard,L., Guffanti,A., Rubboli,F., Marchitiello,A., Giglio,S., Coluccia,E., Zollo,M., Zuffardi,O. and Ballabio,A. (1996) *Nature Genet.*, **13**, 167–174.
7. Zweiger,G. and Scottm,R.W. (1997) *Curr. Opin. Biotechnol.*, **8**, 684–687.
8. Vingron,M. and Hoheisel,J. (1999) *J. Mol. Med.*, **77**, 3–7.
9. Lander,E.S. (1996) *Science*, **274**, 536–539.
10. Anderson,L. and Seilhammer,J. (1997) *Electrophoresis*, **18**, 533–537.
11. Fannon,M.R. (1996) *Trends Biotechnol.*, **14**, 294–298.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
13. Altschul,S.F., Stephen,F., Madden,T.L., Schaffer,A.A., Zhang,J., Thang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
14. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
15. Bonfield,J.K., Smith,K.F. and Staden,R. (1995) *Nucleic Acids Res.*, **23**, 4992–4999.
16. Boguski,M.S. and Schuler,G.D. (1995) *Nature Genet.*, **10**, 369–371.
17. Fisher,R.A. (1973) *Statistical Methods and Scientific Inference*, 3rd Edn. Macmillan Hafner, New York, NY.
18. Audic,S. and Claverie,J.-M. (1997) *Genome Res.*, **7**, 986–995.
19. National Cancer Institute Cancer Genome Anatomy Project (CGAP) (1997) Tumour Gene Index, http://www.ncbi.nlm.nig.gov/ncbicgap
20. Ebeling,M., Ernst,P., Falkenhahn,M., Glatting,K.-H., Hotz-Wagenblatt,A., Kühl,A., Thang,G. and Suhai,S. (1998) In Zimmermann,O. and

Table 5. Electronic northern for the human homologue of the murine JIP-1 gene

Tissue	Benign		Tumour		*P* value	Expression ratio	Class
	Abundance	Pool size	Abundance	Pool size			
Bladder	0	25 643	0	42 556			
Brain	25	178 865	5	100 254	0.036	2.8^{-}	III^{-}
Breast	0	120 733	4	67 587	0.017		III^{+}
Colon	0	52 193	0	35 113			
Endocrine tissue	0	62 286	0	60 780			
Kidney	0	44 687	0	20 743			
Liver	0	21 521	0	15 763			
Lung	0	102 767	0	54 087			
Muscle/skeleton	0	58 355	0	27 075			
Ovary	0	33 707	0	41 943			
Pancreas	0	60 518	0	21 810			
Prostata	1	106 113	0	76 774	1.0		
Stomach/oesophagus	0	13 801	0	12 120			
Testis	0	24 906	0	16 900			
T lymphoma	0	60 743	0	17 318			
Uterus	0	67 669	0	21 740			

The expression levels in 16 normal and tumour tissues are shown for the human homologue of the murine JIP-1 gene. The abundance is given as number of homologous ESTs found in a pool of tissue- and state-specific ESTs. The expression ratio is defined as ratio of the relative abundances (number of homologous ESTs divided by pool size). The greater of the two relative abundances (normal or tumour) is divided by the smaller so that expression ratios are always ≥1. The plus sign indicates up-regulation in tumour tissue, the minus sign down-regulation in tumour tissue. The *P* value computed by Fisher's exact test expresses the statistical validity of the differential expression (see text). For the definition of classes see Table 1. The overexpression in breast cancer tissue of the human homologue to the mouse JIP-1 gene, an inhibitor of c-Jun N-terminal kinase (42), is accompanied by down-regulation in brain cancer tissue.

Schomburg,D. (eds), *Proceedings of the German Conference on Bioinformatics.* University of Cologne, Cologne, Germany.

21. Gill,R.W., Hodgman,T.C., Littler,C.B., Oxer,M.D., Montgomery,D.S., Taylor,S. and Sanseau,P. (1997) *Comput. Appl. Biosci.*, **13**, 453–457.
22. Prigent,C., Gill,R., Trower,M. and Sanseau,P. (1998) *In Silico Biol.*, **2**. http://www.bioinfo.de/isb/1998/01/0011
23. Adams,M.D., Kelley,J.M., Gocynes,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al*. (1991) *Science*, **252**, 1651–1656.
24. Welle,S., Bhatt,K. and Thornton,C.A. (1999) *Genome Res.*, **9**, 506–513.
25. Chen,J., Gokhale,M., Li,Y., Truch,M.A. and Yager,J.D. (1998) *Carcinogenesis*, **19**, 2187–2193.
26. Schiemann,S., Schwirzke,M., Brunner,N. and Weidle,U.H. (1998) *Clin. Exp. Metastasis*, **16**, 129–139.
27. Williams,G.R. and Wright,N.A. (1997) *Virchows Arch.*, **431**, 299–304.
28. Poulsom,R., Hanby,A.M., Lalani,E.N., Hauser,F., Hoffman,W. and Stanp,G.W.H. (1997) *J. Pathol.*, **183**, 30–38.
29. Prosperi,M.T., Apiou,F., Dutrillaux,B. and Goubin,G. (1994) *Genomics*, **19**, 236–241.
30. Kaye,F.J., McBride,O.W., Battey,J.F., Gazdar,A.F. and Sausville,E.A. (1987) *J. Clin. Invest.*, **79**, 1412–1420.
31. Ferrer,M., Yunta,M. and Lazo,P.A. (1998) *Clin. Exp. Immunol.*, **113**, 346–352.
32. Finch,J.L., Miller,J., Aspinall,J.O. and Cowled,P.A. (1999) *Int. J. Cancer*, **80**, 533–538.
33. Yamada,M., Tomida,A., Yoshikawa,H., Taketani,Y. and Tsuruo,T. (1996) *Clin. Cancer Res.*, **2**, 427–432.
34. Scorilas,A., Yotis,J., Pateras,C., Trangas,T. and Talieri,M. (1999) *Clin. Cancer Res.*, **5**, 815–821.
35. Rochefort,H. and Liaudet-Coopman,E. (1999) *Acta Pathol. Microbiol. Immunol. Scand.*, **107**, 86–95.
36. Adriaenssens,E., Dumont,L., Lottin,S., Bolle,D., Leptretre,A., Delobelle,A., Bouali,F., Dugimont,T., Coll,J. and Curgy,J.J. (1998) *Am. J. Pathol.*, **153**, 1597–1607.
37. Feinberg,A.P. (1999) *Cancer Res.*, **59**, 1743s–1746s.
38. De Leon,D.D., Issa,N., Nainani,S. and Asmerom,Y. (1999) *Hormone Metab. Res.*, **31**, 142–147.
39. Theisinger,B., Seitz,B., Dooley,S. and Welter,C. (1996) *Breast Cancer Res. Treat.*, **38**, 145–151.
40. May,F.E.B. and Westley,B.R. (1997) *J. Pathol.*, **183**, 4–7.
41. Watson,M.A. and Fleming,T.P. (1996) *Cancer Res.*, **56**, 860–865.
42. Mooser,V., Maillard,A., Bonny,C., Steinmann,M., Shaw,P., Yarnall,D.P., Burns,D.K., Schorderet,D.F., Nicod,P. and Waeber,G. (1999) *Genomics*, **55**, 202–208.