2216–2223 *Nucleic Acids Research, 1998, Vol. 26, No. 9*

Comparative gene expression profiling by oligonucleotide fingerprinting

Sebastian Meier-Ewert*, Jörg Lange, Helmut Gerst, Ralf Herwig, Armin Schmitt, Jan Freund, Thorsten Elge, Richard Mott[1], Bernhard Herrmann[2] and Hans Lehrach

Max-Planck-Institut für Molekulare Genetik, Ihnestraße 73, 14195 Berlin, Germany, [1]Bioinformatics, SmithKline Beecham Pharmaceuticals, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW, UK and [2]Max-Planck-Institut für Immunbiologie, Stübe Weg 51, 79108 Freiburg, Germany

Received November 7, 1997; Revised and Accepted March 8, 1998

ABSTRACT

The use of hybridisation of synthetic oligonucleotides to cDNAs under high stringency to characterise gene sequences has been demonstrated by a number of groups. We have used two cDNA libraries of 9 and 12 day mouse embryos (24 133 and 34 783 clones respectively) in a pilot study to characterise expressed genes by hybridisation with 110 hybridisation probes. We have identified 33 369 clusters of cDNA clones, that ranged in representation from 1 to 487 copies (0.7%). 737 were assigned to known rodent genes, and a further 13 845 showed significant homologies. A total of 404 clusters were identified as significantly differentially represented ($P < 0.01$) between the two cDNA libraries. This study demonstrates the utility of the fingerprinting approach for the generation of comparative gene expression profiles through the analysis of cDNAs derived from different biological materials.

INTRODUCTION

The genes expressed in a given stage of development, tissue or cell type, determine the molecular machinery available to carry out its biological functions. The identification of expressed genes and the determination of their expression level provides one of the most important indicators to characterise physiological states. To allow such an analysis, a number of different techniques have been proposed (1–3,5–7). The generation of expressed sequence tags (ESTs), short sequences from the ends of randomly selected cDNA clones, has been a particularly important strategy for the identification of new genes, and at least to some extent the characterisation of their expression levels. This technique does however have a number of inherent and important limitations (relatively high cost per sample, difficulties to correctly identify internal sequence changes, difficulties to identify motif sequences located outside the sequenced stretches).

Although a very large number of genes have been identified by classical gene sequencing (human and mouse), both the identification of genes giving rise to low abundance transcripts, and especially the exact quantitation of their levels of expression, will require an order of magnitude increase in the number of cDNA clones which can be analysed. Although most of the limitations remain, some approaches, particularly SAGE (8), have already gone some way toward this goal.

Based on an approach proposed (9), and tested (10), for the identification of overlapping clones by hybridisation with synthetic oligonucleotide probes, we (11) and others (1,2) have developed the hybridisation of short oligonucleotide probes under high stringency conditions to derive a sequence dependent 'fingerprint'. This fingerprint can identify new genes, as well as analyse their exact level of expression in different tissues. Over the past years we have established a set of automated procedures to facilitate large scale cDNA analysis by oligonucleotide hybridisation to large arrayed clone libraries immobilised on nylon membranes (11,12).

In order to test the use of oligonucleotide hybridisation as a tool for the characterisation of gene sequences and comparison between two cDNA libraries, hybridisations were performed with 110 pools of 16 decanucleotides each (Materials and Methods) to arrayed cDNA clones derived from two stages of mouse embryogenesis (9 and 12 day).

MATERIALS AND METHODS

cDNA library construction

Directionally cloned, oligo dT primed cDNA libraries were constructed from mouse embryos (9 and 12 day) and cloned into the plasmid pSV-SPORT. The average insert size was estimated to be 1400 bp, as assessed by PCR amplification of several hundred clones using primers flanking the plasmid cloning sites.

Using an automated picking robot (12), 38 783 and 56 832 primary clones from 9 and 12 day cDNAs respectively, were arrayed into 384-well microtitre plates.

Clone amplification by PCR

Amplification was carried out in 384-well microtitre plates in a volume of 30 µl containing: 5 pmol of each primer, 50 mM KCl, 10 mM Tris–HCl pH 8.55, 1.5 mM $MgCl_2$, 0.01% gelatine, 0.1 mM

*To whom correspondence should be addressed at present address: GPC Aktiengesellschaft, Lochhamerstrasse 29, 82152 Martinsried/Munich, Germany. Tel: +49 89 8996770; Fax: +49 89 89967710; Email: sebastian.meier-ewert@gpc-ag.com

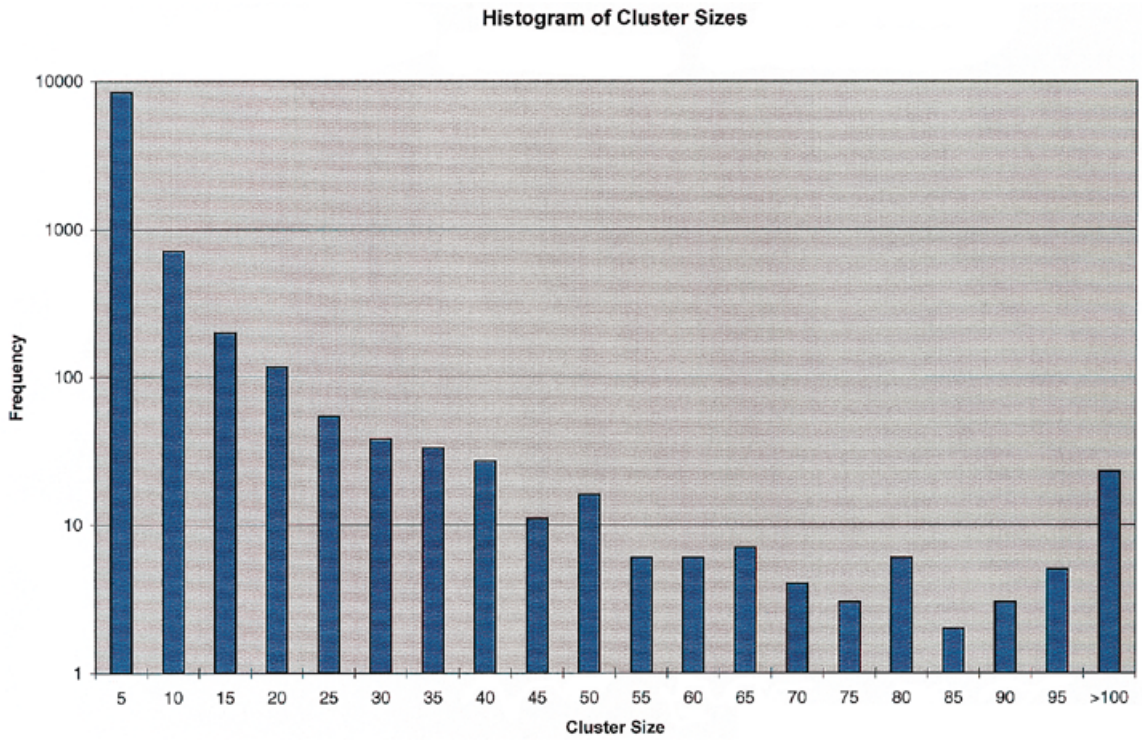

Figure 1. A histogram of the size distribution of the 9634 clusters generated with 110 oligonucleotide pool hybridisations.

dGTP, 0.1 mM dCTP, 0.1 mM dATP, 0.1 mM dTTP, 0.5 U *Taq* polymerase. The primers used were Sport/3 (20mer): 5′-CCGGTCCGGAATTCCCGGGT-3′ and Sport/5 (30mer): 5′-GCACGCGTACGTAAGCTTGGATCCTCTAGA-3′.
Reactions were inoculated with ~0.2 µl bacterial culture (phage lysate in the case of M13 control clones, see 'Oligonucleotide labelling and hybridisation' below) using a 384-pin transfer device (Genetix, Christchurch Dorset) and then heat sealed with a 45 µm bilaminar nylon/polypropylene film using a commercial plate-sealing device (Genetix, Christchurch, Dorset). The sealed 384-well microtitre plates were cycled automatically 30 times between waterbaths at 96°C for 3 min and 73°C for 5 min (11). Due to the high T_m of the amplification primers (86°C) it is possible to perform a two-step PCR amplification, which means that two instead of three waterbaths could be used. After cycling the plates were briefly centrifuged (Beckman J6/B) and the sealing film removed by re-heating in a plate-sealer and thus melting the surface of the plates. Samples were stored at –15°C.

Arraying of cDNA PCR products at high density

Nylon membranes carrying 25 344 PCR products in duplicate were generated using robotic spotting devices developed in house (12). Each PCR product was repeatedly spotted 10 times with a 400 µm diameter pin, thus transferring ~1 µl PCR product (up to 100 ng) (13). Hybond N+ membranes (Amersham, UK) were used as carriers and the DNA fixed according to the manufacturer's protocols.

Oligonucleotide labelling and hybridisation

Oligonucleotides used were pools of 16 decamers with a common octamer core (i.e. NXXXXXXXXN), obtained from Genosys Biotechnologies. These pools were used instead of simple octanucleotide probes, since the stability of a decamer duplex is significantly greater and therefore experimentally easier to detect. Since, for any given hybridisation signal it is not possible to determine which decamer in the pool has bound to the target, the sequence information for each signal is limited to the eight nucleotides common to all members of a pool. The fully deprotected oligonucleotides were diluted and labelled at their 5′ termini by phosphate transfer using T4 polynucleotide kinase. 30 pmol oligonucleotide was labelled in a 30 µl reaction containing: 3 µl 10× buffer (700 mM Tris–HCl pH 7.6, 100 mM $MgCl_2$, 50 mM dithiothreitol), 2 µl T4 polynucleotide kinase (10 U/µl; NEB) and 5 µl [γ-^{33}P]ATP (10 µCi/µl, 3000 Ci/mmol; Amersham). The reaction mixture was incubated at 37°C for 45 min and then terminated by the addition of 3.5 µl EDTA (0.5 M). Unless used immediately, labelled oligonucleotides were stored at –20°C.

For pre-hybridisation the nylon membranes were briefly placed in SSarc buffer [600 mM sodium chloride, 60 mM sodium citrate, 7.2% sodium lauroly sarcosinate (w/v) (Sarkosyl N30, BDH)] at room temperature (20–25°C). Oligonucleotides were hybridised at 4 nM in SSarc buffer, at 4°C for 3–16 h (adapted from Drmanac *et al.*, 14) (hybridisation equilibrium is reached after 3 h). Typically two 22 cm × 22 cm membranes were hybridised in one 300 mm glass bottle with 30 mm diameter (Hybaid) in a volume of 10 ml. Nylon mesh (Hybaid) was used to separate membranes that were hybridised together.

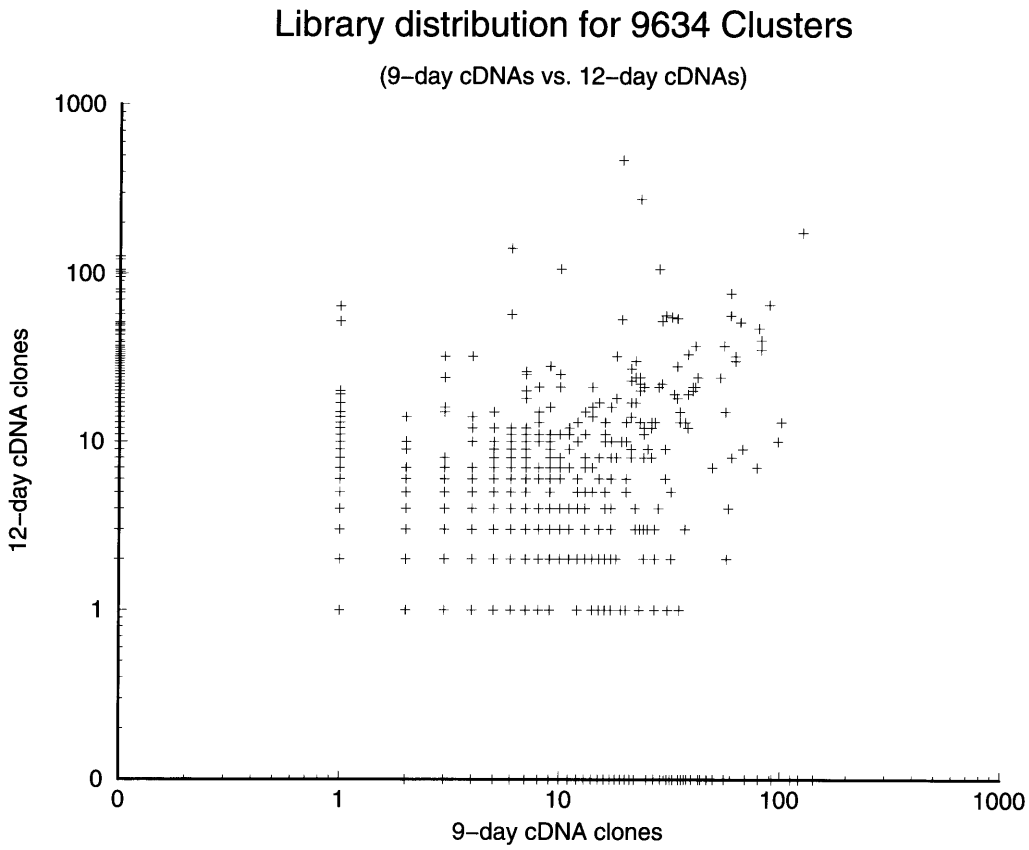

Figure 2. A plot of relative clone representations for 9 and 12 day mouse embryo cDNAs in the observed clusters. Each data point represents one cluster. The plot shows the differences between the two libraries, and that genes across the abundance spectrum are differentially represented.

After hybridisation, membranes were rinsed briefly in cold (4°C) SSarc buffer and washed together with the nylon meshes in 1 l SSarc buffer in a polypropylene lunch box at 10°C for 15–30 min. Up to eight membranes were washed together in 1 l SSarc even if they had been hybridised with different oligonucleotides. The reproducibility of each hybridisation signal was assessed through spotting each clone in duplicate. 27 648 duplicate pairs were spotted on each hybridisation membrane. The mean correlation factor of hybridisation signals indicates the reproducibility of each particular hybridisation. Additionally, PCR products of 1920 M13 phage clones, whose sequence had been previously determined by classical gel sequencing (15), were included on each hybridisation membrane. The sequenced control clones were used to assess the sequence specificity of each oligonucleotide hybridisation. Both reproducibility and sequence specificity varied significantly among the 110 pools of oligonucleotides used (data not shown).

To remove all bound radioactive oligonucleotide, up to 20 membranes were incubated twice in 1 l 0.1× SSarc at 65°C for 10 min. Membranes were used for 30 cycles of hybridisation and stripping without significant loss of signal strength.

Image capture and quantitation

After hybridisation and washing, the membranes were exposed to phosphor storage screens (Molecular Dynamics, Sunnyville, CA) for 3–16 h at room temperature. The screens were scanned at a resolution of 176 μm and the images captured in 16bit TIFF format with a phosphor imager (MD) and then transferred to a Digital 2000 server, with two CPUs, 512 Mb RAM, running the DEC-UNIX 4.0 operating system, for analysis with custom written software (to be described elsewhere).

RESULTS

Clone clustering and database comparisons

The pattern of hybridisation of short oligonucleotides to the DNA of a clone reflects its sequence and can therefore be used as a ‘fingerprint’ for its identification. Since, however, the actual signal intensity depends on a number of parameters which are often difficult to control (amount of DNA in each spot, exact hybridisation conditions, sequences surrounding the match), the hybridisation signal for each clone has to be determined quantitatively and normalised across all hybridisations.

One simple way to normalise the data, is to replace the intensity score for each hybridisation signal by its rank over all signals in the experiment. In this case values were generated by assigning the strongest hybridisation signal a score of 1 and the weakest a score of 0. All remaining signals were assigned scores between 1 and 0 according to their relative ranks in a sorted list of hybridisation signals. This procedure was applied to all signals in each hybridisation and a second time to all ranks across all hybridisations for each individual clone (double ranking). After the normalisation of the hybridisation data, fingerprints of all

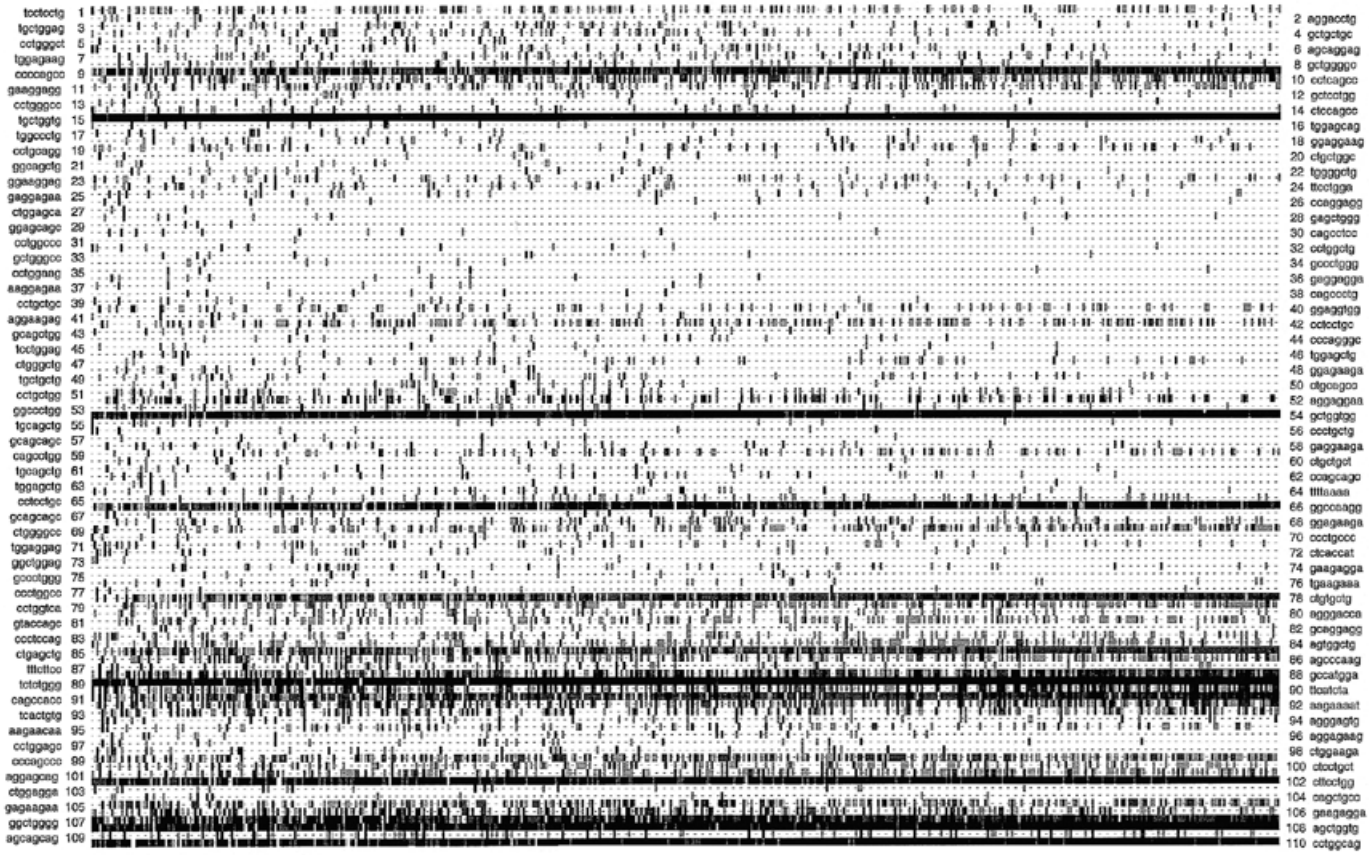

Figure 3. A graphical representation of the fingerprints from a cluster containing 487 clones. Listed vertically are all individual clones in the cluster and horizontally the oligonucleotide probes. The darker the intersection the stronger the interaction between the cDNA clone and the hybridisation probe. The figure shows 11 hybridisation probes that were positive for almost all clones, and form the basis of their similarities by which they were clustered. Additionally, there are a number of probes that hybridised to approximately half the clones in the cluster (e.g. probes 1 and 42), which are most likely due to the variation in clone length. The figure also gives some idea of the level of experimental noise (randomly distributed hybridisation signals) that this approach can tolerate.

clones were compared to each other, and a similarity score obtained for all possible pairs. A similarity score (S_{ab}) between clone *a* and clone *b* is calculated as follows:

$$S_{(ab)} = \sum_{h} g_{(ah)}\ g_{(bh)}\ w_{(h)}$$

where $g_{(ah)}$ and $g_{(bh)}$ are the hybridisation scores for clones *a* and *b* respectively for hybridisation *h* and $w_{(h)}$ is a weighting factor for that hybridisation that reflects the reliability based on duplicate correlations and sequence specificity.

Clones with similarity scores above six standard deviations from the mean were clustered as cliques, similar to methods described by Milosavljevic (16). Cliques were then merged if their member lists overlapped by 60% or more. For clones that were assigned to more than one clique, the assignment was made to that clique whose consensus fingerprint gave the highest similarity score. The resulting lists are considered as clusters of cDNA clones that share a high sequence similarity and are likely to be derived from the same gene. The size of each cluster gives a measure of the abundance of that particular transcript in the source tissue. Within each cluster, clones can be ranked according to their similarity to the consensus fingerprint of all cluster members. This ranking facilitates the identification of good representatives from clusters that can be used for further experiments.

All clones from both libraries were clustered together, and the clone representation of each library per cluster calculated. We obtained a total of 9634 clusters ranging in size from 2 to 487 members, containing 42 926 clones. 23 735 clones gave fingerprints which were unique and therefore remained as singletons. A histogram of gene cluster sizes is shown in Figure 1. The representation of clones from both libraries within each cluster yields information about their relative abundance. Figure 2 shows a plot of the number of clones from the two cDNA libraries in each of the clusters that were found. It shows that the majority of gene clusters are represented equally in both libraries (many of these clusters represent housekeeping genes that are ubiquitously expressed at both developmental stages). A total of 404 gene clusters were found to be represented significantly differently between the two libraries ($P < 0.01$) using a binomial test. For the purpose of this calculation only clusters with a minimum of 10 members were considered. As an example, Figure 3 shows a graphical representation of the largest cluster containing 487 clones which has been identified as ε-globin.

By calculating the expected hybridisation patterns with the set of oligonucleotides used, for all rodent and human database

Table 1. A list of the 100 largest clusters

Cluster	Size	9-day	12-day	Significance	GenBank ID	Description
1	487	19	468	4.037e-07	mmbgy2.gb_ro	Mouse germ line gene coding for beta-globin (Y2).
2	298	124	174	1.101e-11	museftu.gb_ro	Mus musculus protein synthesis elongation factor Tu (eEF-Tu).
3	298	23	275	3.411e-04	ratgffo.gb_ro	Rat basic fibroblast growth factor (FGF) mRNA, complete cds.
4	153	88	65	6.700e-06	musn038a.gb_ro	Mouse nucleolar protein N038 mRNA, complete cds.
5	147	6	141	4.442e-05	rnu05239.gb_ro	Rattus norvegicus opioid-receptor-like orphan receptor mRNA.
6	142	142	0	1.249e-04	muspai1.gb_ro	Mouse plasminogen activator inhibitor (PAI-1) mRNA, complete cds.
7	135	59	76	2.978e-05	mmarmep2.gb_ro	Mouse mRNA for 24kDa major androgen regulated protein, arMEP24.
8	134	28	106	8.963e-04	musferla.gb_ro	Mouse ferritin light chain, complete cds.
9	126	0	126	2.872e-04	mmu44942.gb_ro	Mus musculus recombinant quaking gene sequence.
10	126	79	47	7.433e-05	mmarppo.gb_ro	Mouse mRNA for acidic ribosomal phosophoprotein PO.
12	121	0	121			
11	122	122	0	2.396e-04	ratarget.gb_ro	Rattus norvegicus autoantigen p69 mRNA, complete cds.
13	121	81	40	1.993e-08	rnrps9.gb_ro	R.norvegicus mRNA for ribosomal protein S9.
14	116	65	51	5.647e-07	mmthymoa.gb_ro	Mouse mRNA for prothymosin alpha.
15	116	81	35	2.205e-05	rnppl7a.gb_ro	Rat mRNA for ribosomal protein L7a.
16	116	10	106	5.694e-06	mmbgy2.gb_ro	Mouse germ line gene coding for beta-globin (Y2).
17	115	59	56	4.606e-06	musrsa.gb_ro	Mouse LLRep3 protein mRNA from a repetitive element, complete cds.
18	113	100	13	9.616e-06	musgapdh.gb_ro	Mouse glyceraldehyde-3-phosphate dehydrogenase mRNA, complete cds.
19	112	112	0			
20	107	97	10	1.781e-04	ratpapa.gb_ro	Rat prostatic acid phosphatase (rPAP) mRNA, complete cds.
21	105	0	105	1.636e-04	musadrreca.gb_ro	Mouse alpha-2 adrenergic receptor gene, complete cds.
22	102	0	102	5.128e-05	musmyh.gb_ro	Mouse myosin heavy chain (MYH) gene.
23	99	0	99	1.414e-04	mmsrprtsa.gb_ro	M.musculus serine proteinase gene.
24	95	0	95	4.092e-04	ratmpmm.gb_ro	Rattus norvegicus matrilysin (MMP-7) mRNA, complete cds.
25	94	62	32	1.697e-04	rsgarb1.gb_ro	Rat mRNA for GABA(A) receptor beta-1 subunit.
26	92	55	37	4.167e-06	mml40kd.gb_ro	Mouse mRNA for translational controlled 40 kDa polyyeptide p40.
27	92	62	30	3.693e-05	musnptcc.gb_ro	Mouse mRNA for nuclear pore-targeting complex component of 58 kDa.
28	92	92	0	4.625e-05	musl3t4aa.gb_ro	Mouse T-cell differentiation antigen CD4 (L3T4) mRNA, complete cds.
29	88	34	54	2.045e-05	mabyglo.gb_ro	M.auratus mRNA for beta-like y-globin gene.
30	87	32	55	2.615e-04	ratriboi.gb_ro	Rat ribophorin I (Rpn-I) gene, 5'end.
31	86	30	56	8.830e-05	cgtubb1.gb_ro	Cricetulus griseus (chinese hamster) mRNA for beta tubulin .
32	85	78	7	2.243e-09	museftu.gb_ro	Mus musculus protein synthesis elongation factor Tu (eEF-Tu).
33	81	29	52	7.176e-06	rathdp.gb_ro	Rat helix-destabilizing protein mRNA, complete cds.
34	80	0	80	1.501e-05	ratbccapb.gb_ro	R.norvegicus beta-chain clathrin associated protein complex AP-2.
35	80	80	0	1.160e-06	musrsa.gb_ro	Mouse LLRep3 protein mRNA from a repetitive element, complete cds.
36	78	41	37	3.621e-06	rnrpl5.gb_ro	Rat mRNA for ribosomal protein L5.
37	77	53	24	1.028e-10	rrrpl8.gb_ro	R.rattus mRNA for ribosomal protein L8.
38	77	0	77	1.074e-04	mmdelta1.gb_ro	M.musculus mRNA for Delta-like 1 protein.
39	76	67	9	3.535e-06	ratcypdb2.gb_ro	Rat cytochrome P450-db2 mRNA, complete cds.
40	72	19	53	6.078e-05	mmmtmmp.gb_ro	M.musculus mRNA for membrane-type matrix metalloproteinase.
41	71	56	15	7.642e-09	mushspca.gb_ro	Mouse heat shock protein 70 cognate mRNA, complete cds.
42	71	38	33	2.908e-05	ratccasb.gb_ro	Rattus norvegicus pore-forming calcium channel alpha-1 subunit.
43	70	0	70	1.574e-04	mmig28.gb_ro	Mouse immunoglobulin variable gene V kappa-24 encoding amino acids.
44	70	70	0	4.792e-06	ratospa.gb_ro	Rat cell-binding bone sialoprotein mRNA, complete cds.
45	68	60	8	9.312e-04	rnu25137.gb_ro	Rattus norvegicus alternatively spliced signal transducer.
46	66	42	24	1.546e-15	musgpbsl.gb_ro	Mouse mRNA for G protein beta subunit homologue, complete cds.
47	65	1	64	1.901e-04	musant10a.gb_ro	Mouse cell surface antigen 114/A10 mRNA, complete cds.
48	63	6	57	2.279e-06	ratcg2a1.gb_ro	Rat prepro-alpha-1 type II cartilage collagen mRNA, 5' end.
49	62	58	4	8.086e-10	mmu13687.gb_ro	Mus musculus DBA/2J lactate dehydrogenase-A (LDH-A) mRNA.
50	62	34	28	3.313e-06	ratmtatpsa.gb_ro	Rat mitochondrial ATP synthase beta subunit mRNA, complete cds.
51	62	41	21	7.876e-06	rnu16245.gb_ro	Rattus norvegicus aquaporin-5 (AQP5) mRNA, complete cds.
52	61	40	21	2.939e-06	mmef1a.gb_ro	Mouse mRNA for elongation factor 1-alpha (EF 1-alpha).
53	61	40	21	2.959e-04	musgcanf.gb_ro	Mouse guanylate cyclase/atrial natriuretic factor receptor mRNA
54	60	40	20			
55	59	57	2	3.996e-04	ratosteo.gb_ro	Human osteocalcin gene, 5' end, and promoter region.
56	57	38	19	1.680e-10	mustuba2m.gb_ro	Mouse alpha-tubulin isotype M-alpha-2 mRNA, complete cds.
57	57	0	57	8.963e-04	mmu27106.gb_ro	Mus musculus clathrin-associated AP-2 complex AP50 subunit mRNA+G9
58	56	49	7	6.023e-05	mmu20107.gb_ro	Mus musculus synaptotagmin VIII mRNA, partial cds.
59	56	56	0	3.876e-04	rnu59486.gb_ro	Rattus norvegicus GDNF receptor alpha mRNA, complete cds.
60	53	1	52	1.117e-07	museftu.gb_ro	Mus musculus protein synthesis elongation factor Tu (eEF-Tu).
61	52	34	18	1.389e-04	mmu28726.gb_ro	Mus musculus protein kinase homolog (mess1) mRNA, complete cds
62	52	22	30	5.537e-05	rnu30788.gb_ro	Rattus norvegicus Tclone4 mRNA.
63	52	33	19	4.845e-10	mmu29402.gb_ro	Mus musculus acidic ribosomal phosphoprotein P1 mRNA, complete cds.
64	51	29	22	8.747e-04	musadhxx.gb_ro	Mouse alcohol dehydrogenase 1 (ADH-1) mRNA, 3' end.
65	51	0	51	2.335e-04	muskerda.gb_ro	Mouse keratin D mRNA, complete cds.
66	50	18	32	6.047e-05	ratdtr.gb_ro	Rat dihydropteridine reductase mRNA, complete cds.
67	50	0	50	6.251e-10	mustuba1m.gb_ro	Mouse alpha-tubulin isotype M-alpha-1 mRNA, complete cds.
68	50	38	12	7.309e-05	mmzonpel4.gb_ro	Mus musculus zona pellucida (Zp-1) gene, exons 10, 11, 12.
69	50	35	15	1.184e-05	muspsd95sp.gb_ro	Mouse mRNA for PSD-95/SAP90A.
70	50	37	13	1.273e-04	rnsdrpl13.gb_ro	R.norvegicus (Sprague Dawley) ribosomal protein L13 mRNA.
71	50	38	12	5.755e-06	musigkcni.gb_ro	Mouse Ig rearranged kappa-chain mRNA, clone AN09K.
72	50	50	0	3.495e-06	rnrpl13a.gb_ro	R.norvegicus mRNA for ribosomal protein L13a.
73	49	28	21			
74	49	0	49	3.642e-04	rrmap2.gb_ro	Rat mRNA for microtubule-associated protein 2.
75	48	35	13	4.724e-05	mmecadh.gb_ro	Mouse mRNA for E-cadherin.
76	48	21	27	9.640e-06	ratbccapb.gb_ro	R.norvegicus beta-chain clathrin associated protein complex AP-2.
77	47	23	24	3.208e-06	mmu16322.gb_ro	Mus musculus basic transcription factor MITF-2B mRNA, cds.
78	47	47	0	1.525e-06	ratthy.gb_ro	Rat prothymosin-alpha mRNA, complete cds.

Cluster	Size	9-day	12-day	Significance	GenBank ID	Description
79	46	0	46	4.898e-04	mmu37720.gb_ro	Mus musculus CDC42 mRNA, complete cds.
80	46	0	46	7.505e-05	ratdpd.gb_ro	Rat cAMP phosphodiesterase mRNA, 3' end.
81	46	22	24			
82	45	24	21	9.271e-09	mmtax107.gb_ro	M.musculus mRNA for TAX responsive element binding protein 107.
83	45	23	22	1.911e-04	mmtb10ut5.gb_ro	M.musculus mRNA for testis-specific thymosin beta-10.
84	45	0	45	4.488e-04	mmspalttf.gb_ro	M.musculus mRNA for spalt transcription factor.
85	45	0	45	3.891e-05	mmu35142.gb_ro	Mus musculus retinoblastoma-binding protein (mRbAp46) mRNA.
86	44	21	23	2.731e-05	ratglytrn.gb_ro	Rattus norvegicus glycine transporter mRNA, complete cds.
87	44	44	0	1.840e-07	msalen.gb_ro	Mouse mRNA for alpha-enolase (2-phospho-D-glycerate hydrolase)
88	43	0	43	2.645e-05	mmu35249.gb_ro	Mus musculus CDK-activating kinase assembly factor p36/MAT.
89	43	23	20	2.295e-05	mmperf.gb_ro	M.musculus mRNA for perforin.
90	43	0	43	6.477e-05	rnu04808.gb_ro	Rattus norvegicus Sprague-Dawley putative G-protein coupled receptor.
91	43	0	43	1.932e-04	rnu28830.gb_ro	Rattus norvegicus RalBP1 mRNA, complete cds.
92	43	43	0	3.955e-07	mmtb10ut5.gb_ro	M.musculus mRNA for testis-specific thymosin beta-10.
93	40	37	3	1.962e-04	rat1433pa.gb_ro	Rat 14-3-3 protein mRNA for mitochondrial import stimulation factor.
94	40	40	0	3.903e-05	ratglytra.gb_ro	Rat glycine transporter mRNA, complete cds.
95	40	27	13	2.366e-06	mmu20611.gb_ro	Mus musculus thioredoxin-dependent peroxide reductase (tpx) mRNA.
96	39	22	17			
97	39	0	39	4.094e-04	musscd2.gb_ro	Mouse stearoyl-CoA desaturase (SCD2) gene, complete cds.
98	39	26	13	2.261e-05	muscx26a.gb_ro	Mouse connexin (Cx26) gene, partial exon 1.
99	39	30	9	1.068e-04	hamsgp2a.gb_ro	Hamster sulfated glycoprotein 2 mRNA, 3' end.
100	39	0	39	8.180e-06	mmbgy2.gb_ro	Mouse germ line gene coding for beta-globin (Y2).

The columns contain the clusters numbered consecutively, the total number of members, the number of members from the 9 and 12 day mouse embryo cDNA libraries respectively, the significance of database matches, the GenBank ID of the matched entry and its description. Significance values are categorised as follows: <e-07, positive assignment; >e-07 and <e-06, tentative assignment; >e-5 and <e-4, significant homology; >e-4, no match.

sequences in GenBank corresponding to transcribed sequences, theoretical fingerprints were generated. For each oligonucleotide pool, the common core octamer sequence was used (Materials and Methods) to calculate the theoretical fingerprint. Oligonucleotides were assigned a score of 1 if they matched a given database sequence and a score of 0 if they did not. Using an algorithm based on that used in the BLAST program (17), we compared all observed fingerprints for experimental clones to the theoretical fingerprints, and thereby were able to assign the identity of many of the gene clusters found. In total, 374 clusters were assigned to known rodent genes with very high confidence. Additionally, of the clones that remained as singletons (i.e. whose fingerprints were found only once in the two libraries), 363 were matched to known rodent database sequences. When compared with predicted fingerprints from human DNA sequences an additional 141 clusters and 270 singletons were matched at very high confidence. A total of 793 cDNA clones from both libraries were analysed by a single pass sequence from the 5′-end ('tag-sequencing') in order to validate the clustering, check the accuracy of database matches predicted from fingerprints, and to analyse new genes (this data is freely available from the Resource Centre of the German Human Genome Project, RZPD; URL: http://www.rzpd.de/). Out of 129 clones that were sequenced from clusters with significant database matches ($P < 10^{-7}$), 117 (91%) showed the same database match using the BLAST (17) program. At lower significance values ($P < 10^{-5}$), fingerprints can be tentatively assigned to known sequences but frequently there are multiple matches of equivalent significance, that cannot at present be resolved. At this level an accuracy of 89% was estimated by gel-sequencing (171 sequences), and a total of 3486 database matches to rodent and a further 3167 to human sequences were found. Similarities to database sequences that cannot be classified as positive assignments were found in 13 845 cases. Table 1 contains a list of the 100 largest clusters found, their distribution across the two libraries and corresponding database matches. From the database matches of the clusters it is clear that some genes have been falsely split into multiple clusters. This can be verified by back hybridisation of individual clones from these clusters to the entire library. In a small number of cases that were analysed in detail we found that clones from the same genes that were placed into separate clusters were of different lengths (data not shown), and as a result gave significantly different fingerprints. Mostly this occurred when the fingerprints did not contain many positive oligonucleotide signals. Clearly, this will lead to an overestimate of the complexity of the cDNA libraries. Some of the database matches highlight that with the fingerprints generated here it is difficult to distinguish database sequences accurately, which have a high homology to one another. For example Cluster 1 which has a match to β-globin, should in fact match ε-globin (as confirmed by sequencing).

In order to evaluate this technology as a means of identifying previously unknown genes, 77 clones were sequenced from clusters that had shown no significant database match by fingerprint analysis. Of these, 57 (74%) were sequences not present in the GenBank databases.

Representatives from each of the 33 369 clusters are being re-arrayed into a normalised library, which will be used in large scale sequencing and whole mount *in situ* hybridisation (18) projects. This library will also be available from the Resource Centre of the German Human Genome Project.

DISCUSSION

Comparative gene expression profiling is emerging as one of the most promising approaches to large scale functional gene analysis. A number of different methods have been developed in recent times. One class of techniques, based essentially on counting the number of transcripts for each gene in the corresponding cDNA libraries, all rely on some form of sequence determination to identify clones originating from the same gene. Such a sequence determination can either involve end-sequencing (8,19), the determination of a short indicator sequence by gel techniques (8), or in our case the use of oligonucleotide hybridisation to identify a sequence dependent fingerprint of each

clone. Based on counting, these techniques can be made very sensitive (only dependent on the number of clones analysed) and in addition to providing information on transcript abundance, can also give information on potential sequence changes.

Alternative to this approach, a number of techniques have been developed, by which transcript abundance levels are ultimately determined by some signal intensity, inherently an analogue approach. Examples are the use of complex probes to hybridise to clone or oligonucleotide grids (1,2,20,21), and techniques such as differential display (22) and others (8,23), in which changes in the abundance of each member of small groups of transcripts are identified by changes in the intensity of PCR products. There are some disadvantages and technical limitations associated with all methods to date. Methods that are based on reassociation or hybridisation of complex mixtures of nucleic acids will frequently suffer from the problem that minor sequence variants of abundant genes, such as splice variants and rare gene family members, will either be lost or not distinguished. The use of oligonucleotide arrays that cover various regions of all known genes, can in principle overcome this limitation. However, the use of complex mixtures of probes presents a sensitivity problem, in that low abundance transcripts are difficult to detect and that very large amounts of labelled probe are required. Some of these issues have recently been addressed (5), although the problem of sensitivity remains (large quantities of RNA are required per experiment). Also, the sensitivity of detection will vary from gene to gene, since the level of crosshybridisation is sequence dependent and affected by the representation of homologous genes in the mixture. In some cases, where amounts of source material are limited, such as in this case with mouse embryos, the requirement for large amounts of RNA cannot be fulfilled. Additionally, oligonucleotide arrays designed from database sequences are by necessity limited to the analysis of known genes. As more and more genes are identified, this becomes a less serious limitation. However, the analysis of important model organisms for which there is less complete sequence coverage still requires an alternative approach.

Oligonucleotide fingerprinting can overcome many of the above limitations, and therefore holds some promise as a technology to complement existing methods. Compared to the commonly used partial ('EST') sequencing strategy, the costs per clone are very significantly reduced, since large numbers of clones are analysed in parallel. In addition, the sequence fingerprint generated covers statistically the entire sequence of the clone, in contrast to the short EST sequences. On the one hand, this allows considerably greater success in identifying similarities or identities in clones with different ends, since the majority of the fingerprint will be shared, and on the other hand, internal sequence changes have a much higher chance of being identified allowing, for example, the identification of internal deletions or splice variants represented in different clones derived from the same gene.

The sensitivity in this case is determined by the number of clones that are arrayed as targets. As automation technologies improve, arraying densities increase so that it is now possible to routinely use libraries with 100 000 or more clones. Since no prior knowledge of gene sequences is required, the technology can be applied to the study of genes from most organisms. Conversely, when sufficient gene sequences are known, oligonucleotide probes selected from specific motif sequences can be hybridised in order to identify clones corresponding to members of gene families of interest, such as tyrosine kinases, TGF-β family, or G-protein coupled receptors. This approach combines effective sequence classification with a targeted selection of genes of interest.

These considerations make oligonucleotide fingerprinting an attractive approach to large scale and in depth expression profiling using large cDNA libraries of >100 000 clones. In addition, the fingerprinting strategy offers a highly efficient strategy to identify genes not yet represented in previously sequenced cDNA collections.

A number of improvements need to be made in order to increase the sensitivity of the method. Although theoretical calculations predict that 110 hybridisation probes should be sufficient to generate unique fingerprints, this pilot study shows that more hybridisations are required in order cluster a collection of cDNA clones close to completion. Given the experimental noise and the fact that the fingerprints generated are not all equally significant due to statistical variations, it is clear that the sensitivity of the method can be increased by the use of ~100–200 more oligonucleotide probes. The resolution in terms of sequence homologies that will lead to clones being clustered together, varies according to the number of positive (informative) hybridisations events that make up a fingerprint. On average we estimate that currently homologies of >70% result in clones being assigned to the same cluster. Increasing the number of oligonucleotide probes, will mean that smaller differences can be discriminated such that it should be possible even to detect splice variants of a single gene using this approach.

An analysis of clustering error rates showed that 90% of clones in clusters are truly derived from the same gene, and that ~30% of the clones that should be clustered remain as singletons. The false negative rate is high and reduces the sensitivity of the method in two ways. It will lead to an under-estimate of the expression of genes and thus increases the size of arrayed libraries required to detect medium to rare transcripts, and reduces the sensitivity of expression difference detection.

The accuracy of database comparisons needs to be increased by the selection of oligonucleotides that hybridise more sequence specifically, or the adaptation of hybridisation conditions to improve specificity. For hybridisation conditions it might be fruitful to perform washes at varying temperatures according to the predicted stability of the duplexes. Varying the time of post hybridisation washes can also increase the specificity in some cases. Since it is not presently possible to predict accurately the specificity of hybridisation for short oligonucleotides, we continually select sets of more specific probes by assessing the quality of hybridisations empirically through the use of control clones of known sequence.

As long as not all human genes are known one cannot exclude the possibility that some genes will not give an informative fingerprint with the current set of oligonucleotide probes. The sets of probes are continually updated with the aim that almost all genes available in public databases can be discriminated (some very short genes are difficult to capture with this approach).

We feel that this study demonstrates much of the potential of oligonucleotide fingerprinting as a tool for in depth comparative expression profiling, while highlighting some of its present limitations.

It is anticipated that further developments to streamline the processes, increase the throughput and miniaturise the arrays, will increase the number of clones that can be characterised at equivalent effort by a further order of magnitude.

ACKNOWLEDGEMENTS

We would like to thank Sabine Thamm and Richard Reinhardt for the sequencing work. Thanks also to Elmar Maier, David Bancroft and Igor Ivanov for many stimulating discussions during the course of this work.

REFERENCES

1 Milosavljevic,A., Zeremski,M., Strezoska,Z., Grujic,D., Dyanov,H., Batus,S., Salbego,D., Paunesku,T., Soares,M.B. and Crkvenjakov,R. (1996) *Genome Res.*, **6**, 132–141.
2 Drmanac,S., Stavropoulos,N.A., Labat,I., Vonau,J., Hauser,B., Soares,M.B. and Drmanac,R. (1996) *Genomics*, **37**, 29–40.
3 Shoemaker,D.D., Lashkari,D.A., Morris,D., Mittmann,M. and Davis,R.W. (1996) *Nature Genet.* **14**, 450–456.
4 Hacia,J.G., Brody,L.C., Chee,M.S., Fodor,S.P. and Collins,F.S. (1996) *Nature Genet.*, **14**, 441–447.
5 Lockhart,D.J, Dong,H., Byrne,M.C, Follettie,M.T, Gallo,M.V, Chee,M.S, Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) *Nature Biotechnol.*, **14**, 1675–1680.
6 Sikela,J.M. and Auffray,C. (1993) *Nature Genet.* **3**, 189–191.
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.* (1995) *Nature*, **377** (suppl. S), 3–174.
8 Velculescu,V.E, Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) *Science*, **270**, 484–487.
9 Poustka,A., Pohl,T., Barlow,D.P., Zehetner,G., Craig,A., Michiels,F., Ehrich,E., Frischauf,A.M. and Lehrach,H. (1986) *Cold. Spring. Harb. Symp. Quant. Biol.*, **51 Pt 1**, 131–139.
10 Hoheisel,J.D., Maier,E., Mott,R., McCarthy,L., Grigoriev,A.V., Schalkwyk,L.C., Nizetic,D., Francis,F. and Lehrach,H. (1993) *Cell*, **73**, 109–120.
11 Meier Ewert,S., Maier,E., Ahmadi,A., Curtis,J. and Lehrach,H. (1993) *Nature*, **361**, 375–376.
12 Maier,E., Meier Ewert,S. and Lehrach,H. (1994) *J. Biotechnol.* **35**, 191–203.
13 Maier,E., Meier-Ewert,S., Bancroft,D. and Lehrach,H. (1997) *Drug Discovery Today*, **2**, 315–324.
14 Drmanac,R., Strezoska,Z., Labat,I., Drmanac,S. and Crkvenjakov,R. (1990) *DNA Cell Biol.* **9**, 527–534.
15 Beck,S., Kelly,A., Radley,E., Khurshid,F., Alderton,R.P. and Trowsdale,J. (1992) *J. Mol. Biol.* **228**, 433–441.
16 Milosavljevic,A., Strezoska,Z., Zeremski,M., Grujic,D., Paunesku,T. and Crkvenjakov,R. (1995) *Genomics*, **27**, 83–89.
17 Altschul,S.F., Gish,W., Miller,W., Meyers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
18 Herrmann,B.G. (1991) *Development*, **113**, 913–917.
19 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.* (1991) *Science*, **252**, 1651–1656.
20 Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) *Science*, **270**, 467–470.
21 Gress,T.M., Muller-Pillasch,F., Geng,M., Zimmerhackl,F., Zehetner,G., Friess,H., Buchler,M., Adler,G. and Lehrach,H. (1996) *Oncogene*, **13**, 1819–1830.
22 Liang,P. and Pardee,A.B. (1992) *Science*, **257**, 967–971.
23 Lisitsyn,N. and Wigler,M. (1993) *Science*, **259**, 946–951.