Minireview

Complementary DNA sequence (EST) collections and the expression information of the human genome

Kousaku Okubo*, Kenichi Matsubara

Institute for Molecular and Cellular Biology, Osaka University, 1-3, Yamada-oka, Suita, Osaka 565, Japan

Received 7 January 1997

Key words: Expressed Sequence Tag; Gene expression; Genome analysis; Expression profile; Gene signature

1. Introduction

Since its inception in 1989/1990, the Human Genome Project has put major emphasis on structural genome analyses, resulting in elaborate genetic and physical maps. As a result of these achievements, important disease genes have been identified through positional cloning, and the focus of the project is now turning toward sequencing the entire genome. Complete sequences of genome DNA have already been determined for several single cell organisms [1–3] including *Saccharomyces cerevisiae* (yeast) [4], and vigorous analyses are being carried out on many other organisms [5,6]. Sequencing the human genome is expected to be complete around the year 2005.

The goal of genome analysis is to decode the entire genetic information carried in the genome. This means that, besides sequencing, information which cannot readily be read from structural data must be collected. Examples are the activities of hypothetical gene products, the expression control and the regulatory network of each gene. This requires systematic efforts which may be collectively called functional analysis of the genome. In this connection, a consortium has been established for functional analysis of the *S. cerevisiae* genome through systematic disruption of genes [7]

For higher eukaryotes with more genes, however, different strategies are needed, such as enlisting all (or nearly all) the active genes and studying the sites of their expression together with the extent of their activities. A list of active genes and their activities in the cells and tissues can be called an expression profile. Construction of such a list should be started with the representative cells or tissues of the body and expanded to cover the various stages of development or pathological conditions.

2. Expression profiles

Kohara et al. have initiated the collection of large numbers of in situ staining patterns with *Caenorhabditis elegans*, using probe cDNAs that represent novel genes as discovered by single run cDNA sequencing. This worm is transparent and is composed of only 965 cells, thus allowing for whole body analysis [8,9]. The facts that cell lineages are well established during development and that there is only a small number of genes for testing favor adoption of this procedure.

In the case of man, body size and the large number of genes preclude a similar approach. Thus, Okubo et al. have initiated efforts to identify active genes by single pass sequencing of cDNA obtained from a type of cell or tissue and quantified their activities in the mRNA population [10]. Although an adult human body consists of some 6 trillion cells, they can be categorized into 200 basic types [11]. Therefore, the number of cells or tissue species is within a reasonable range.

The cDNA libraries used by these authors contain only the 3′ terminal restriction fragments [12]. They were not amplified prior to the experiments so that they faithfully represent the composition of mRNA in the cellular source. In addition, care has been taken that the source materials are prepared to be as homogeneous as possible by using well-characterized cell lines [13], selective primary culturing [14], and purification with antibodies or through careful dissection [15]. The clones in the libraries were randomly selected and sequenced for identifying the genes and for measuring the abundance of their transcripts. Although they carry little amino acid sequence information, the 3′ sequences correspond to the respective genes, and are termed gene signatures (GS) [16]. The resulting lists, showing active genes and their relative activities, are called expression profiles, and some of these (parts of the profiles) are shown in Table 1.

3. Body map and ESTs

By compiling expression profiles from different parts of the body, the cells or tissues where any given gene is active can be identified. As genes are mapped in the body where they are active, this data set is called a ‘Bodymap’ (http://www.imcb.osaka-u.ac.jp/bodymap) [10,17]. As of June 1996, about 13 000 genes have been mapped on the body. A part of the bodymap which focuses on genes encoding cytoskeletal proteins is shown in Table 2.

In addition to the bodymap, there are two major collections of partial cDNA sequences, collectively called EST (expressed sequence tags), which are compiled in a dbEST (http://www.ncbi.nlm.nih.gov/dbEST) [18]. One set, constructed in order to identify new genes of commercial interest, consists of 174 000 ESTs from randomly primed cDNA fragments or the 5′ ends of conventional cDNAs of human organs [19] (for further references, see [20]). In this data set, several sequences from different regions of a single mRNA are collected indiscriminately. Another set containing some 280 000 ESTs [21] has been constructed to obtain probes for gene mapping [22].

*Corresponding author. Fax: (81) (6) 877-1922.
E-mail: kousaku@imcb.osaka-u.ac.jp

0014-5793/97/$17.00 © 1997 Federation of European Biochemical Societies. All rights reserved.
PII S0014-5793(97)00042-2

Table 1
Expression profiles of active genes in the adult liver [30], lung [31] and colon mucosa [32]

Adult liver			Lung			Colon mucosa		
GS#	f	Gene name	GS#	f	Gene name	GS#	f	Gene name
364	140	serum albumin	2894	67	pulmonary surfactant	196	37	fatty acid binding protein
2085	34	haptoglobin alpha 1S	2842	30	Clara cell protein	2546	18	
2174	16	α-1 acid glycoprotein	937	11	HLA-E heavy chain	4094	13	
476	13	apolipoprotein B	2105	9	fibronectin	2542	11	immunoglobulin λ light chain
2155	10	transferrin	211	5	ribosomal protein S8	2706	11	
277	9	apolipoprotein AII	314	5	ribosomal protein S11	2123	11	CD24 signal transducer
196	8	fatty acid binding protein	2148	4	α-2-macroglobulin	4116	9	carcinoma antigen GA733-2
2176	7	apolipoprotein C-II	1657	4	ribosomal protein L18	335	6	ribosomal protein L7a
2541	7	P-450 S-mephenytoin 4-OHase	292	4	ribosomal protein S16	223	6	cytokeratin 8
111	6	α1-antitrypsin	2542	4	immunoglobulin l chain	650	6	ribosomal protein S18
2093	6	aldolase B	1671	4	protein p68	1809	6	lysosomal glycoprotein CD63
2047	5	gamma fibrinogen	1720	4	lipocortin II	1766	5	metallothionein
2092	5	retinol binding protein	1929	4		285	5	ribosomal protein L21
2116	5	alcohol dehydrogenase β-1	689	3		273	5	tumor protein
2148	5	α-2-macroglobulin	293	3	ribosomal protein L27a	1657	4	ribosomal protein L18
2202	5	fibrinogen β-chain	1791	3	ubiquitin	162	4	ribosomal phosphoprotein P2
2525	5	prothrombin (F2)	583	3	ribosomal protein L3	155	4	thymosin β-4
732	4	ribosomal protein L38	1189	3		19	4	elongation factor 1-α
1766	4	metallothionein	114	3	γ-actin	565	4	β-2-microglobulin
2120	4	fibrinogen β-chain (short 3′UTR)	335	3	ribosomal protein L7a	917	4	
285	3	ribosomal protein L21	708	3	ribosomal protein L29	1670	4	
380	3		1367	3	hnRNP-E1	211	3	ribosomal protein S8
689	3		1382	3		96	3	alpha NAC
934	3	coxVIb	1786	3		304	3	ubiquitin
1305	3	vacuolar H$^+$ ATPase	1891	3	B4B	932	3	
2075	3	inter-α trypsin inhibitor	1919	3	calcyclin	2073	3	
2105	3	fibronectin	2025	3		2673	3	succinate dehydrogenase
2152	3	complement component C4A	2656	3	MHC class II antigen	1404	3	set
2423	3		2702	3		244	3	β-actin
2564	3		2728	3	c-fos	4070	3	carcinoembryonic antigen

Gene signatures (GS) that identify genes, composition of mRNA (f: frequency of appearance expressed per mil.), and the names of the genes are shown. Blanks under 'gene name' indicate novel genes. For purposes of clarity only the 30 most active genes are shown.

The majority of the source cDNA libraries for this work were 'normalized' in vitro in order to reduce the number of abundant, frequently appearing clones [23]. Regardless of the original purpose of construction, these data sets are useful gene pools for 'fishing' new members of gene families or human orthologs of genes of other species, because they contain a significant number of entries [24,25].

In Table 3, distributions of ESTs among the human organs are shown. The same set of cytoskeletal gene transcripts as in Table 2 were selected for comparison. In accordance with the results in Table 2, actins and tubulins are distributed among a variety of tissues, as is already well known. Notice, however, that the multiple appearance of these gene transcripts in EST collections simply reflects incomplete normalization of the libraries, whereas those in the bodymap represent their gene activities. The high expression of tubulins in fetal neurons and fibroblasts, or the clear division of the site of cytokeratin gene expression into simple and stratified epithelial types, can be seen only in the bodymap because the histological resolution of gene expressions has been pursued. Overall, the dbEST should be regarded as a part of structural data, rather than functional data of the human genome.

As the number of genes collected and tissues analyzed in the bodymap is not yet large enough (the gene coverage is about one fourth of the dbEST), its usefulness remains limited. For this reason, a high throughput sequencing system is urgently needed. As a compromise, a method called SAGE has been proposed, in which 9 bp tags, resected from defined positions of cDNAs, are tandemly ligated and sequenced [26].

Methods other than nucleotide sequencing can also be employed for identifying active genes. Kato [27] has described 'molecular indexing', using the size of a restriction fragment of cDNA as an identifier: the cDNAs are cleaved by a type IIS enzyme and subjected to 64 different adapter-mediated PCRs. Altogether, 256 groups of the amplification products from a library have been fractionated in sequencing gels. By repeating this procedure with a few type IIS enzymes, most of the transcripts in source cells are displayed separately in the gel, one gene transcript being represented by a band of unique size, and its abundance by the band intensity. Another emerging technique is construction of arrays of oligonucleotides or unique fragments of cDNA at high density on solid support, which can be hybridized with uniformly labeled mRNA or cDNA [28,29], for detection of active genes and their relative activities by their intensities. At the moment, however, specificity and sensitivity are the problems, since hybridization parameters differ from sequence to sequence.

4. Conclusion

An average higher eukaryotic cell carries some 10 000 species of gene transcripts, and the total number of mRNA mol-

		blood cells						connective tissues									epithelial tissues									neural tissues														cancer			
Acc#	Gene	HL60	HL60/DMSO induced	HL60/TPA induced	granulocyte	CD4 Tcell	CD8 Tcell	aortic endothel	fibroblast	osteoblast	osteocyte	subcutaneous fat	visceral fat	itoh cell	mesangium	aortic media	hepG2	19Wliver	40Wliver	adult liver	lung	colon mucosa	keratinocyte	cornea	taste bud	retina	temporal lobe (1)*	temporal lobe (2)*	cerebellum	hippocampus	caudate nucleus	thalamus	putamen	corpus callosum	fetal neuron	fetal astrocytes	schwann cells	neuroblastoma	pituitary	small cell lung	adeno lung	squamous cell lung	total
actins																																											
X13839	α-actin								2	2		2		8		6												1	1												1		23
X00351	β-actin		14	1	4			1	11	2	2		1	2			2					3	4	1	1	1		3						3	1	3		4			7	2	73
X04098	γ-actin			1										3	1								2	1												3	1				1		13
tubulins																																											
K00558	α-tubulin	1		4		1		2	7			1		2	1						1		3			2	2	2	3	1	2	4	3	1	11	1		1	2	7	3	1	69
J00314	β-tubulin		2	3				1	6		2	1									1	2	4	2				3	1		1		1	1	5	5	1			1	1	1	45
M61764	γ-tubulin							1			1					1																											3
keratins(stratified epithelial type)																																											
X05421	cytokeratin 3																							9																			9
X07695	cytokeratin 4																							2																			2
M19723	cytokeratin 5																						2																				2
L42601	cytokeratin 6																						3		7																	4	14
X52426	cytokeratin 13																							1	6																	1	8
J00124	cytokeratin 14																						15	1	1																		17
	cytokeratin 15																																										0
Z19574	cytokeratin 17																						3																				3
keratins(simple epithelial type)																																											
	cytokeratin 7																																										0
X12882	cytokeratin 8												1		2		1		1		1	6		2							1	1									2	1	19
X12883	cytokeratin 18																1	1			1	1	1																		1	1	7
Y00503	cytokeratin 19																						1	2																	6	7	16
neural intermediate filaments																																											
J04569	GFAP																										3							10		2							15
X05608	neurofilament L																									1	1			2	1	3	1										9
Y00067	neurofilament M																											1				3	1										5
	total	866	1087	898	1186	990	1097	745	741	643	961	972	1100	934	1048	660	839	1288	1109	959	878	925	822	1158	293	887	527	941	1114	463	1081	927	873	950	1193	1112	976	1247	1026	854	1190	1203	38763

Acc#, accession number in GenBank. For other definitions, see Table 1. Entries for human actins, tubulins, cytokeratins, and representative neural intermediate filaments were culled from Swiss Prot and a cDNA or a gene sequence corresponding to each of them in GenBank/EMBL was compared with Gene Signatures (GS) using the FastA program [33]. GS with bases more than 90% identical to those of the cDNAs or gene sequences were counted. Genes for cytokeratins 1, 2, 9, 10, 11, 12, and 20 were not found among either GS or ESTs (see Table 3).

*Temporal lobes from an age matched pair of normal brain (1) and the brain of an Alzheimer disease patient (2).

ecules is estimated to be several hundreds of thousands. Expression profiling, by sequencing or by other means, to describe the composition of these mRNA populations provides the basis for future functional analyses of the genome. The method established for the global description of gene expression control and gene networks can be extended to a variety of other systems and organisms. Some of the expected applications may include sorting novel genes that are defined as cell or tissue specific and studying their possible medical, diagnostic and pharmaceutical applications. Obtaining markers for monitoring cell differentiation or detecting pathological changes of gene activities in cells for diagnostic purposes may be other interesting examples of the application of this method.

There is a great need for two major technological developments: high throughputs, as noted in the text, and down-scaling the analysis to a single cell level. Even though the system is still far from being technologically mature, the importance of the project demands enriching the databases in order to achieve the goal of describing how the approximately 100 000 genes in the human genome act in concert in the regulation of the whole body. Collaboration towards this goal is of crucial importance.

References

[1] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. et al. (1995) Science 269, 496–512.

[2] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M. et al. (1995) Science 270, 397–403.

[3] Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M. and Tabata, S. (1995) DNA Res. 2, 153–166.

[4] Goffeau, A., Barrel, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. et al. (1996) Science 274, 546–567.

[5] Waterston, R. and Sulston, J. (1995) Proc. Natl. Acad. Sci. USA 92, 10836–10840.

[6] Goodman, H.M., Ecker, J.R. and Dean, C. (1995) Proc. Natl. Acad. Sci. USA 92, 10831–10835.

[7] Oliver, S. (1996) Trends Genet. 12, 241–242.

[8] Birchall, P.S., Fishpool, R.M. and Albertson, D.G. (1995) Nature Genet. 11, 314–320.

		Washington-U / Merck																					other EST				
		non-neural tissues												neural tissues													
Acc#	gene	N.fetal liver/spleen	fetal spleen	ovary	N.ovarian tumor	N.breast (2NbH)	N.breast (3NbH)	N.placenta	placenta	N.placenta	liver	lung	olfact_epithelium	N.infant brain	N.adult brain(2Nb4H)	N.adult brain(2Nb5H)	N.m.s. Plaques	N.pineal gland	N.retina (N2b4HR)	N.retina (N2b5HR)	N.melanocyte	fetal cochlea	lymphocyte (gxp)	skeltal muscle (gxp)	fetal&infant brain(S)	others	total
actins																											
X13839	α-actin	4	1	1			7	6		2		2	1	1							2					1	28
X00351	β-actin	10	5	2			2		7	5	6	15	5	3	2	1				1	9	1			5		79
X04098	γ-actin	9	5	3			3	1	1	2	4	7	1	3			1	1	1	1	1	1			10		55
tubulins																											
K00558	α-tubulin	23	5				6	6	3	4	1	3	2	27		2					11	2			37	1	133
J00314	β-tubulin	6	2			1		1	2	1		1	1	23	2	1			3		5				13	3	65
M61764	γ-tubulin	9						1		1		1		5													17
keratins(stratified epithelial type)																											
X05421	cytokeratin 3																										0
X07695	cytokeratin 4											1															1
M19723	cytokeratin 5					1	1																				2
L42601	cytokeratin 6																										0
X52426	cytokeratin 13	1																								1	2
J00124	cytokeratin 14					5	1																				6
X07696	cytokeratin 15						4																				4
Z19574	cytokeratin 17				1	1	3						1														6
keratins(simple epithelial type)																											
X13353	cytokeratin 7	3			1				1	3		1														1	10
X12882	cytokeratin 8	4				1			2	1																	8
X12883	cytokeratin 18	1						2	5		3	3														1	15
Y00503	cytokeratin 19						2	1	3				2														8
neural intermediate filaments																											
J04569	GFAP													3	2	2	1									1	9
X05608	neurofilament L													3	1											2	6
Y00067	neurofilament M																	1									1
	total 3'-reads	34086	3018	1560	380	3440	3933	14822	1718	2753	3971	3893	1648	21043	1852	3758	922	640	1887	1067	10728	1239	1579	957	3472	2118	126484

The same set of cytoskeletal gene sequences were compared with ESTs. Only EST entries in GenBank (re93, June96) annotated as 3′ were used to avoid multiple counting of ESTs representing the same mRNA molecule. The libraries subjected to normalization procedures are denoted as N. The numbers represent multiple appearances in the same library, reflecting insufficient normalization or excess amplification.

[9] Lynch, A.S., Briggs, D. and Hope, I.A. (1995) Nature Genet. 11, 309–313.

[10] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. (1992) Nature Genet. 2, 173–179.

[11] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) Molecular Biology of the Cell, 3rd edn., p. 34, Garland Publishing, New York.

[12] Okubo, K., Hori, N., Matoba, R., Niiyama, T. and Matsubara, K. (1991) DNA Seq. 2, 137–144.

[13] Okubo, K., Itoh, K., Fukushima, A., Yoshii, J. and Matsubara, K. (1995) Genomics 30, 178–186.

[14] Kita, H., Okubo, K. and Matsubara, K. (1996) DNA Res. 3, 1–7.

[15] Nishida, K., Adachi, W., Shimizu-Matsumoto, A., Kinoshita, S., Mizuno, K., Matsubara, K. and Okubo, K. (1996) Invest. Ophthalmol. Vis. Sci. 37, 1800–1809.

[16] Matsubara, K. and Okubo, K. (1993) Curr. Opin. Biotechnol. 4, 672–677.

[17] Matsubara, K. and Okubo, K. (1993) Gene 135, 265–274.

[18] Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) Nature Genet. 4, 332–333.

[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] 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, 173–174.

[21] Hiller, L., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W. et al. (1996) Genome Res. 6, 807–828.

[22] Schuler, G.D., Boguski, E.A., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E. et al. (1996) Science 274, 540–546.

[23] Soares, M.B., Bonaldo, M.F., Jelene, L., Su, P., Lawton, L. and Efstratiadis, A. (1994) Proc. Natl. Acad. Sci. USA 91, 9228–9232.

[24] Nicolaides, N.C., Papadopoulos, N., Liu, B., Wei, Y.F., Carter, K.C., Ruben, S.M., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D., Fraser, C.M. et al. (1994) Nature 371, 75–80.

[25] The European Polycystic Kidney Disease Consortium (1994) Cell 77, 881–894.

[26] Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Science 270, 484–487.

[27] Kato, K. (1995) Nucleic Acids Res. 23, 3685–3690.

[28] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Science 270, 467–470.

[29] Drmanac, S., Stavropoulos, N.A., Labat, I., Novau, J., Hauser, B., Soares, M.B. and Drmanac, R. (1996) Genomics 37, 29–40.

[30] Kawamoto, S., Matsumoto, Y. Mizuno, K., Okubo, K. and Matsubara, K. (1996) Gene 174, 151–158.

[31] Itoh, K., Okubo, K., Yosii, J., Yokouchi, H. and Matsubara, K. (1994) DNA Res. 1, 279–287.

[32] Okubo, K., Yoshii, J., Yokouchi, H., Kameyama, M. and Matsubara, K. (1994) Dna Res. 1, 37–45.

[33] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA. 85, 2444–2448.