Volume 325, number 1,2, 118–122 FEBS 12373 June 1993

Minireview

Use of reference libraries and hybridisation fingerprinting for relational genome analysis

Jörg D. Hoheisel and Hans Lehrach

Genome Analysis Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, London, WC2A 3PX, UK

Received 14 March 1993

The concept of relational genome analysis by hybridisation has been developed into a working system. Various genomic and cDNA libraries have been generated and are distributed via a reference system. Analysis procedures have been tested successfully in the mapping of the entire *Schizosaccharomyces pombe* genome. In another test-case for their refinement, analyses on the *Drosophila* genome are well under way. Human and mouse libraries are being studied on all levels, from generating YAC maps to partially sequencing representative cDNA libraries. The automation of the involved processes and the development of improved image detection and analysis are well advanced.

Genome analysis; Hybridisation; Reference library; Mapping; Sequencing

1. INTRODUCTION

In recent years, an international program has been established directed at the unravelling of the information stored in the genetic material of a variety of organisms, and particularly in the human genome. This (Human) Genome Project aims at a comprehensive molecular understanding of genome sequence, structure and organisation, of the functions that are encoded and their regulation. Landmarks to this end are the creation of physical clone maps of high resolution, the constitution of a gene inventory, a sequence analysis of coding sequences and, ultimately, a sequence determination of the genomic DNA. Large-scale mapping analyses in YACs (e.g. [1–3]), bacteriophage P1 clones and cosmids (e.g. [4–7]) are under way or have already been completed. Experiments on the determination of the human gene inventory have been started [8,9], and progress is being made in the field of large-scale sequencing (e.g. [10,11]).

For the analyses mentioned above, we use an approach that takes advantage of the basic characteristic of nucleic acid to form a double strand with complementary sequences. Hybridisation techniques reduce individual clone handling to a minimum, so that very large clone numbers (representative libraries of different types and organisms) can be studied simultaneously. Thereby the levels of analysis are directly related. Efficiency is further increased by the fact that basically any piece of DNA, from radiation hybrids and megabase YACs down to hexamer oligonucleotides, can be used both as a probe and a probant, so that experiments can be designed to fit their objectives.

Correspondence address: J. Hoheisel, Genome Analysis Laboratory, Imperial Cancer Research Fund, PO Box 123, Lincoln's Inn Fields, London, WC2A 3PX, UK. Fax: (44) (71) 269 3068.

2. CONCEPT

High-coverage clone libraries made from both genomic and transcribed DNA are generated and stored. Individual clones are spotted in ordered, high-density filter grids and the DNA is fixed in situ. Since large numbers of filter duplicates can be generated, the libraries as a whole are accessible for many hybridisation experiments, and the clones are nevertheless individually identifiable. The filters are made available as a resource to other laboratories in a reference system [12,13].

All hybridisation results (i.e. Fig. 1) and any information correlated to the probes, such as functions associated with consensus sequences or positional data of genetic markers, for example, are fed into a relational database. Thereby, the different levels of analysis are interrelated, which, among other things, assists in the determination of the clone order. A genomic clone map is established and aligned to other maps. The probes not only provide fingerprint information for the ordering, but are automatically positioned on the map, once completed. Thus, if cDNAs are used as probes, a transcriptional map is created simultaneously with the genomic map.

Mapping by the hybridisation of short oligomers (11–

Published by Elsevier Science Publishers B.V

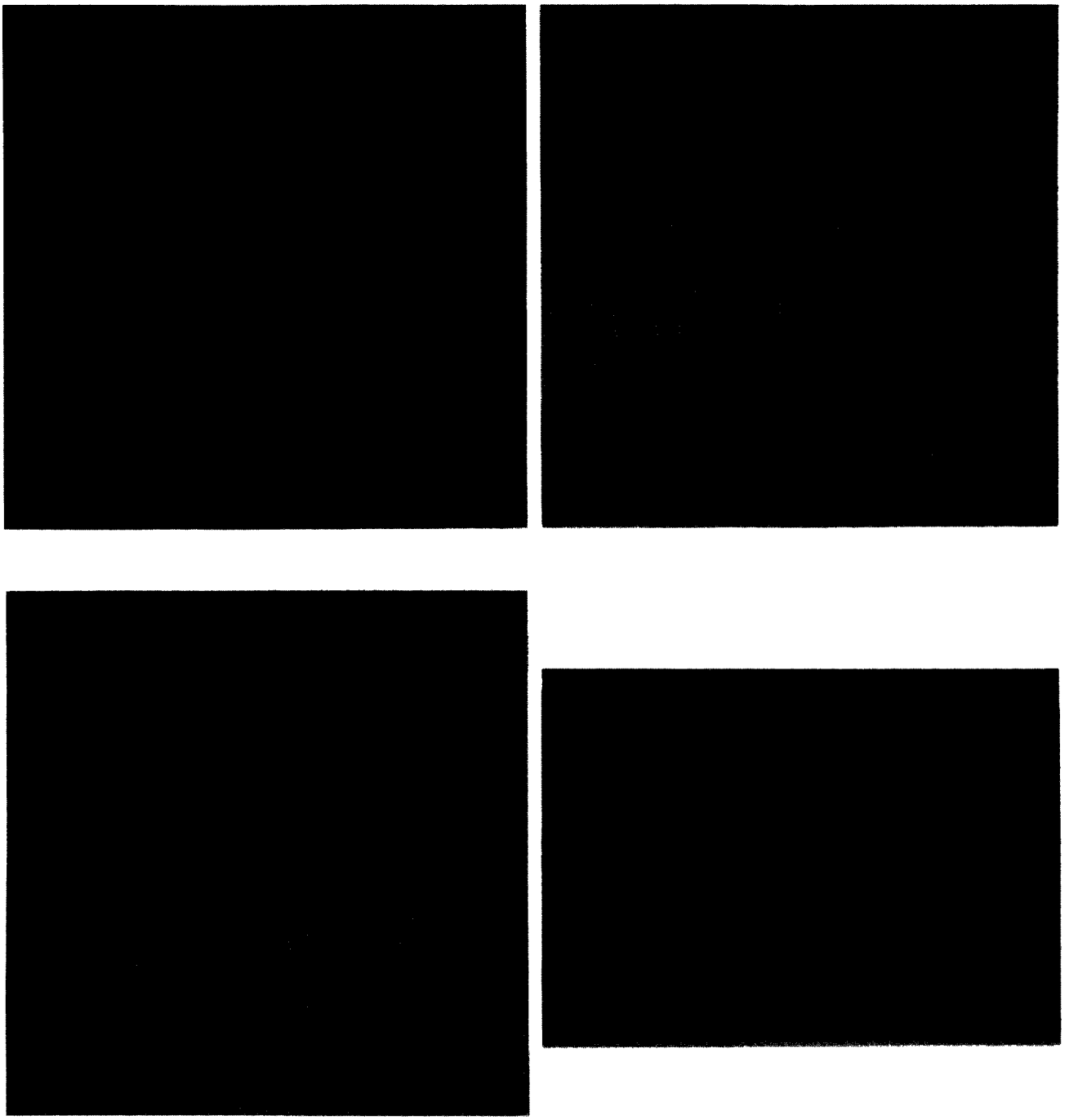

Fig. 1. Autoradiographs of probe hybridisations. (A) Unique probe hybridised to human YAC clones (M. Ross, unpublished). (B) Pattern produced by the oligomer AAAAAAAATAATTT on cosmids. (C) Probing with the oligomer $(ATT)_4$ of a *Drosophila* cDNA library. (D) Hybridisation of a clone pool to two identical filters containing *Drosophila* YACs.

15mers; Fig. 1B; [12]) has major advantages over other hybridisation procedures. The number of necessary probes is independent of the genome size and is very low (less than 200). Also, the hybridisations are unaffected by repeat sequences. Additionally, information on structural and organisational features can be acquired by using relevant sequences.

On the level of DNA sequencing, a partial sequence analysis by the hybridisation of very short oligonucleotides (hexamers to octamers) is carried out to characterise and classify cDNAs and correlate them with a function by a comparison of their oligomer signature to known sequences [14]. This method has the potential to be expanded to a complete sequence determination by hybridising half the set of all 65,536 octamers to double-stranded DNA templates.

3. LIBRARIES

The libraries comprise the whole range of cloning vectors (YAC, P1, cosmid, λ phage, jumping and linking libraries as well as cDNA and exon-trap libraries). For the genomic libraries, usually between 10 and 30 genome equivalents are picked and stored, since, for genome mapping, library construction and data acquisition by hybridisation are less work intensive than a subsequent investigation of unresolved regions. Altogether, more than 900,000 individual clones, stored in several copies, are being worked with at the moment. Many more, particularly human cDNAs, will be added in the near future. After their characterisation, however, the number of clones can be reduced dramatically. A minimal set of 26 YACs, taken from a 47-fold coverage of 1,248 clones, for instance, spans the *S. pombe* genome [3].

Libraries have been made from *S. pombe*, *D. melanogaster*, pig, mouse and human DNA [5,13,15,16]. For the last, chromosome-specific as well as total human libraries exist. A list of available reference filters has been published recently [17] and can be requested from the authors. To date (January 1993), some 1,400 filters and 10,450 identified clones have been provided to 280 laboratories.

4. ROBOTICS

For the handling of the large clone numbers, a range of robotic devices has been designed and installed [9]. A picking robot transfers per h up to 2,000 randomly plated clones into 384-well microtiter dishes for growth and storage. For DNA purification, about 46,000 samples can be PCR-amplified in 3 h by a purpose-built machine. Using another device, 12 filter replicas of up to 36,864 clones or DNA samples each are gridded within 2.5 h, a time span that will soon be reduced to 40 min. Radioactive hybridisation signals are scored quantitatively by storage phosphor detection (PhosphorImager, Molecular Dynamics) and are directly transferred into a computer database for analysis. The automation of the hybridisation process is worked at using non-radioactive detection systems and rigid support matrices.

5. ANALYSIS PROCEDURES

In mapping hybridisations with unique probes, applied separately or in pooling schemes that allow the relation of each signal to a particular clone (e.g. [18]), better contigs were obtained by algorithms that order the *probes* first, producing a set of probe-tagged sites (PTSs; [3,5]) named in analogy to the STS concept [19], and subsequently fit the clones (Fig. 2). Totally anonymous DNA, without any sequence or positional information available, can be used as probes by this method. In comparison to the gel fingerprinting of *Caenorhabditis elegans* [4], about a quarter of the number of experiments sufficed in cosmid mapping identical distances in *S. pombe* [5] using single probe hybridisations.

As illustrated by the experiments on *S. pombe* [5], the main obstacle for an application of oligonucleotide fingerprinting is the accuracy with which this data can be harvested. Signal intensities have to be quantitatively detected and normalised for the quantity of target DNA present at each spot. This makes automated image processing a prerequisite. It is the focus of our current efforts on the process of data capture and is well advanced, as is work on programs generating maps from multilocus probe information [20,21].

6. TEST-CASES *(S. pombe, D. melanogaster)*

The genomes of *S. pombe* (14 Mbp) and *D. melanogaster* (160 Mbp) serve as systems for the testing and further development of the techniques, apart from being important model organisms. The *S. pombe* genome has been spanned in YACs [3] and P1 and cosmid clones (5). The simultaneous use of three libraries contributed substantially towards the resolution of the clone order. A wide range of probes, from entire YAC clones down to undecamer oligonucleotides, was applied successfully. From the raw hybridisation data alone certain structural features could be located, such as restriction (*Not*I) sites and (unknown) repeat sequences.

Using genomic YAC, P1 and cosmid libraries and representative, embryonic cDNA libraries, physical and transcriptional maps of the *Drosophila* genome are being generated [13]. The cDNAs have been screened with repeat sequences. Repeat-free clones are being pool-hybridised to the genomic clones (Fig. 1D). Since they are simultaneously hybridised back to the cDNA libraries, a strategy of sampling without replacement can be applied. Additionally, gene homologies can be identified this way. The cDNAs get arrayed in a PTS map to which the YAC clones are fitted. Anchor points are provided by about 1,000 YACs already positioned cytogenetically [22] and genetic markers. For the creation of a cosmid map, the PTS information is supplemented with data generated from complex probe hybridisations (oligomers; repeats; cosmid pools). The cDNAs are in the process of being characterised by pool hybridisations with tissue-specific, human and mouse cDNA libraries for the identification of gene expression patterns [23] and partial sequence analysis by oligomer hybridisation.

7. MAMMALIAN GENOMES (HUMAN, MOUSE, PIG)

YAC libraries representing the entire genomes of human (17 × coverage), mouse (7 ×) and pig (1.2 ×;

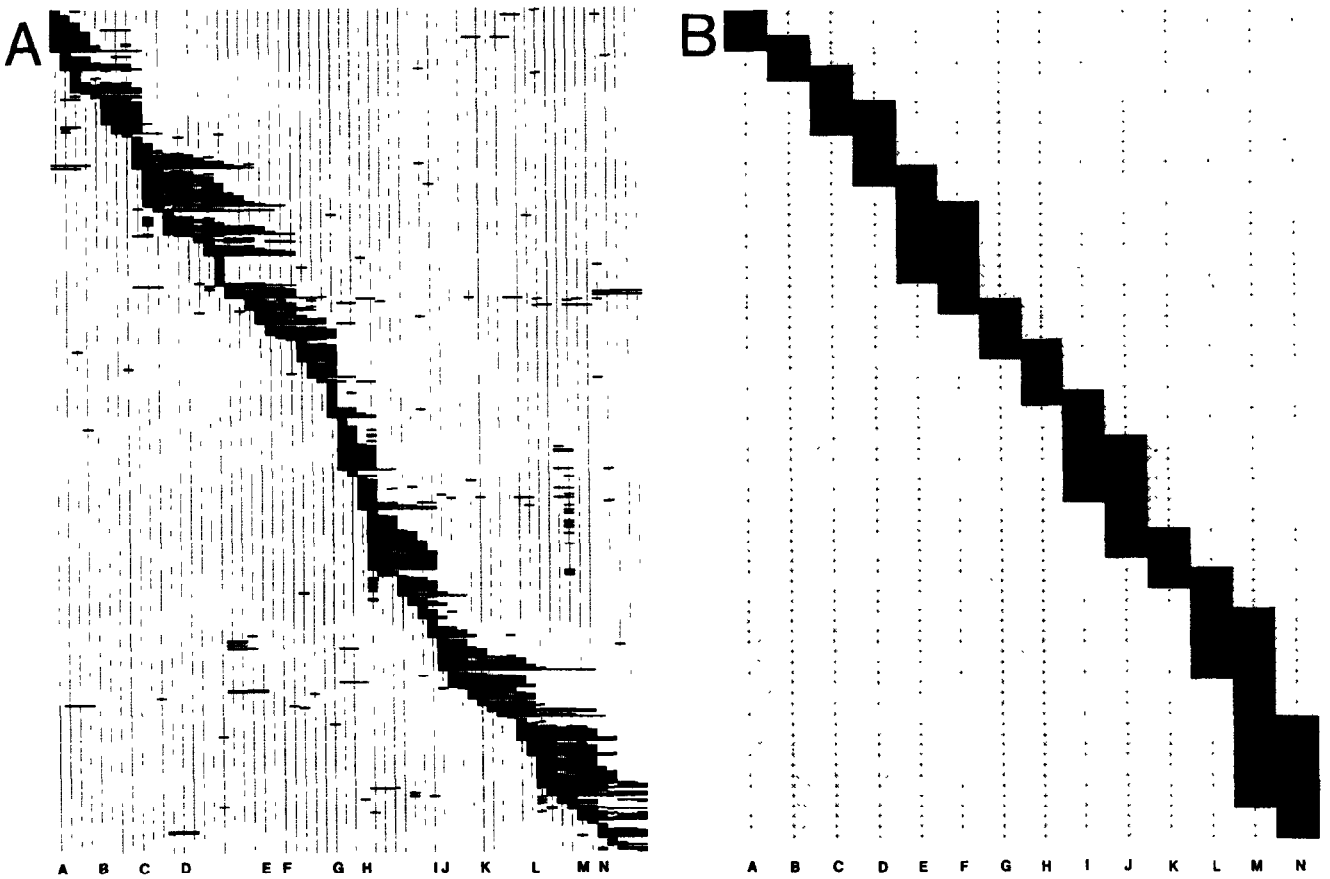

Fig. 2. YAC map of chromosome I of *S. pombe* [3] as constructed by programs using 'simulated annealing' algorithms (A) or heuristic rules (B) [21]. Each vertical line represents a probe, while clones are indicated by horizontal lines. The probes present in both panels are labelled A to N; their order is identical.

more clones in preparation) are being used for mapping [24]. DNA is individually purified from the human clones by *Alu-Alu* PCR [25]. The PCR products are arrayed on filters and also used as probes, following a strategy of sampling without replacement. *Alu* PCR products from radiation hybrid cells, genetic markers and pools made from chromosome-specific cosmid and plasmid libraries provide further placement information [26], both for YACs and cosmids. Utilising other repeat sequences, similar experiments are under way with mouse YACs.

Together with the *Drosophila* cosmids, the cosmid library specific for the human chromosome 21 has been hybridised to a large number of oligomers and other complex probes. Apart from the mapping information gained, such experiments allow comparisons between two genomes. The more data that accumulates, the more accurately the degree of homology between different libraries can be examined, because the growing oligomer map is in principle a very partial sequence determination of the genomes.

A wide variety of human and mouse cDNA libraries from different tissues and developmental stages have been generated and are in the process of being picked. They are PCR-amplified and the DNA is transferred onto filters. Hybridisations with octamer nucleotides are used to characterise and subsequently catalogue their sequences [9]. The system is based on the statistical probability of sequence identity of clones sharing a given number of hybridisation events. This partial sequence information is obtained from the entire length of the clones, rather than from one end only as by the 'tag sequencing' technique [8]. Also, oligomers of unique or degenerate sequence can be selected to identify certain features, such as homeobox sequences, for instance. Computer simulation on all known human gene sequences suggest that less than 100 hybridisations could be sufficient for clone classification (R. Mott and S. Meier-Ewert, personal communication). Once the clones are classified, a comparison to already sequenced genes will identify homologies, and thereby relate clones to associated functions. The technique has the potential for a complete sequence analysis of transcribed and, eventually, genomic DNA [27], with either very large clone numbers or the complete set of, e.g. octamers bound to a hybridisation matrix [28,29].

Acknowledgements: We thank Richard Mott and Andrei Grigoriev for the provision of Fig. 2, and Mark Ross and Elmar Maier for suggestions on the manuscript.

REFERENCES

[1] Chumakov, I. Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G., Gervy, P., LeGall, I., Soularue, P., Grinas, L., Bougueleret, L. et al. (1992) Nature 359, 380–387.
[2] Foote, S., Vollrath, D., Hilton, A. and Page, D.C. (1992) Science 258, 60–66.
[3] Maier, E., Hoheisel, J.D., McCarthy, L., Mott, R., Grigoriev, A.V., Monaco, A.P., Larin, Z. and Lehrach, H. (1992) Nature Genet. 1, 273–277.
[4] Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J. and Waterston, R. (1991) BioEssays 13, 413–417.
[5] 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.
[6] Merriam, J., Ashburner, M., Hartl, D.L. and Kafatos, F.C. (1992) Science 254, 221–225.
[7] Stallings, R., Torney, D.C., Hildebrand, C.E., Longmire, J., Deaven, L., Jett, J., Dogget, N. and Moyzis, R. (1991) Proc. Natl. Acad. Sci. USA 87, 6218–6222.
[8] 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., Kervalage, A.R., McCombie, W.R. and Venter, J.C. (1991) Science 252, 1651–1656.
[9] Meier-Ewert, S., Maier, E., Ahmadi, A.R., Curtis, J. and Lehrach (1993) Nature 361, 375–376.
[10] Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P.G., Benit, P. et al. (1992) Nature 357, 38–46.
[11] Sulston, J., Du, Z., Thomas, K., Wison, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, R., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., Waterston, R. (1992) Nature 356, 37–41.
[12] Lehrach, H., Drmanac, R., Hoheisel, J.D., Larin, Z., Lennon, G., Monaco, A.P., Nizetic, D., Zehetner, G. and Poustka, A. (1990) in: Genome Analysis, vol. 1: Genetic and Physical Mapping (Davies, K.E. and Tilghman, S. eds.), pp. 39–81, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[13] Hoheisel, J.D., Lennon, G.G., Zehetner, G. and Lehrach, H. (1991) J. Mol. Biol. 220, 903–914.
[14] Drmanac, R., Lennon, G., Drmanac, S., Labat, I., Crkvenjakov, R. and Lehrach, H. (1991) in: Proc. First Int. Conf. on Electrophoresis, Supercomputing and the Human Genome (Cantor, C.R. and Lim, H.A. eds.) pp. 60–74, World Scientific, Singapore.
[15] Larin, Z., Monaco, A.P. and Lehrach, H. (1991) Proc. Natl. Acad. Sci. USA 88, 4123–4127.
[16] Nizetic, D., Zehetner, G., Monaco, A.P., Gellen, L., Young, B.D. and Lehrach, H. (1991) Proc. Natl. Acad. Sci. USA 88, 3233–3237.
[17] Hoheisel, J.D., Ross, M.T., Zehetner, G. and Lehrach, H. (1993) J. Biotechnol. (submitted).
[18] Evans, G.A. and Lewis, K.A. (1989) Proc. Natl. Acad. Sci. USA 86, 5030–5034.
[19] Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) Science 245, 1434–1435.
[20] Michiels, F., Craig, A.G., Zehetner G., Smith, G.P. and Lehrach, H. (1987) Comput. Appl. Biosci. 3, 203–210.
[21] Mott, R., Grigoriev, A.V., Maier, E., Hoheisel, J.D. and Lehrach, H. (1993) Nucleic Acids Res. 21, 1965–1974.
[22] Ajioka, J.W., Smoller, D.A., Jones, R.W., Carulli, J.P., Vellek, A.E.C., Garza, D., Link, A.J., Duncan, I.W. and Hartl, D.L. (1991) Chromosoma 100, 495–509.
[23] Gress, T.M., Hoheisel, J.D., Lennon, G.G., Zehetner, G. and Lehrach, H. (1992) Mammalian Genomes 3, 609–619.
[24] Ross, M.T., Hoheisel, J.D., Monaco, A.P., Larin, Z., Zehetner, G. and Lehrach, H. (1992) in: Techniques for the Analysis of Complex Genomes (Anand, R. ed.) pp. 137–154, Academic Press.
[25] Monaco, A.P., Lam, V.M.S., Zehetner, G., Lennon, G.G., Douglas, C., Nizetic, D., Goodfellow, P.N. and Lehrach, H. (1991) Nucleic Acids Res. 19, 3315–3318.
[26] Ross, M.T., Nizetic, D., Nguyen, C., Knights, C., Vatcheva, R., Burden, N., Douglas, C., Zehetner, G., Ward, D.C , Baldini, A. and Lehrach, H. (1992) Nature Genet. 1, 284–290.
[27] Strezoska, Z., Paunescu, T., Radosavljevic, D., Labat, I., Drmanac, R. and Crkvenjakov, R. (1991) Proc. Natl. Acad. Sci. USA 88, 10089–10093.
[28] Krapho, K.R., Lysov, Y.P., Khorlin, A.A., Ivanov, I.B., Yersov, G.M., Vasilenko, S.K., Florentiev, V.L. and Mirzabekov A.D. (1991) DNA Sequence 1, 375–388.
[29] Maskos, U. and Southern, E.M. (1992) Nucl. Acids Res. 20, 1679–1684.