Application of hybridization techniques to genome mapping and sequencing

JÖRG D. HOHEISEL

Establishment of a variety of hybridization techniques for the analysis of large genomic areas has paved the way for a parallel examination of genomes on many levels within the framework of the various genome projects. Here, I discuss some recent achievements in the application of DNA hybridization techniques, in particular oligomer hybridization.

It seems that molecular genome analysis has at last come of age, a conclusion reinforced by the wealth of data presented at recent meetings on this subject. The measure of this success is perhaps that progress in the field no longer seems remarkable: advances are reported on such a wide front that no single one seems exceptional. Even the identification of genes that underlie human diseases has become almost routine: the intense interest surrounding the recent isolation of the Huntington's disease gene[1] was perhaps as much due to the technical difficulty of the endeavour as to its biological significance. The extent of mapping studies and the numbers of genomic regions being analysed have increased considerably; for example, a very coarse map of the human genome derived from sequences cloned in yeast artificial chromosomes (YAC) is now approaching completion[2], and identification of cDNAs has been facilitated by methods that allow tagged sequences to be produced simply and rapidly[3]. These developments have led some to comment that the end of the Human Genome Project is in sight. Far from it: in physical mapping, let alone sequencing, the productive period has just begun. Although for the time being the emphasis in mapping has shifted toward the use of YAC clones, whose large inserts allow large chromosomal regions to be analysed much more easily and quickly, considerable progress is also being made using cloning systems that use *Escherichia coli*. While YAC libraries are extremely useful for the localization and isolation of particular areas or genes and can serve as a framework for more detailed analyses, *E. coli* libraries are much more amenable to sequencing whole genomes, the eventual aim of the genome projects.

As regards methodologies for generating mapping and sequencing data, there are three main approaches. The 'classic', entirely gel-based, fragmentation methods are restriction enzyme fingerprinting, extensively used for mapping in *E. coli*, *Caenorhabditis elegans* and *Drosophila melanogaster*[4–6], and enzymatic dideoxy-termination or chemical-cleavage sequencing. Automated versions of these techniques are being applied to several problems, for example, mapping human chromosomes[7] and sequencing the *C. elegans*[8] genome. However despite automation, the practicality of these approaches is limited by the fact that they involve relatively extensive handling of clones. Moreover, mapping of inserts cloned in low-copy-number vector systems such as YACs requires secondary screening using, for example, repeat sequence probes[2].

A second, intermediate, group of techniques depends on the use of a 'tag' of known sequence as a marker sequence. While the sequence-tagged site (STS) approach[9] is widely used and has been successfully applied in complete clone-mapping projects[10,11], its usefulness is restricted by the density of markers generated. Anonymous probes from arbitrary PCR amplification or, for primate DNA, inter-*Alu* PCR could increase the resolution of mapping toward that possible in *E. coli* clone libraries. The usefulness of multiplex sequencing[12], which increases the efficiency of standard gel sequencing by allowing several samples to be separated in the same lane, is also limited by the number of tagged sequences that can be identified.

The third, non-gel-based, type of technique relies entirely on a basic property of nucleic acids: formation of a specific duplex between complementary sequences. This property is exploited in DNA hybridization assays. DNA fragments that range in length from hexamers to megabases can be used both as probe and target, and useful results can be obtained even with complex DNA mixtures. Experimentally, however, this technique can have some drawbacks: for example, hybridization may occur between sequences that contain base-pair mismatches. Nevertheless, since numerous clones can be handled with ease and since many different levels of DNA manipulation – from radiation hybrids to oligonucleotides, from complex inter-*Alu* PCR products to STS markers – can be related to one another directly, hybridization seems an obvious strategy for comprehensive analysis of genomes (Box 1), and for the subsequent, although rarely mentioned, task of comparing individual cases of interest with the existing data.

Among DNA hybridization techniques, oligomer hybridization alone is broad enough to be applied to the entire range of analyses. In contrast to other methods, it provides partial sequence information and, depending on the number of probe oligomers, can

Box 1. Landmarks of the genome project

- Genetic linkage maps
- Physical maps
  - Cytogenetic maps
  - Cell hybrid maps
  - YAC contigs
  - High-resolution contigs
  - Restriction–oligomer maps
- Transcriptional maps
- Gene inventories
- Gene sequences
- Genomic sequences
- Interspecies comparisons
- Gene expression studies

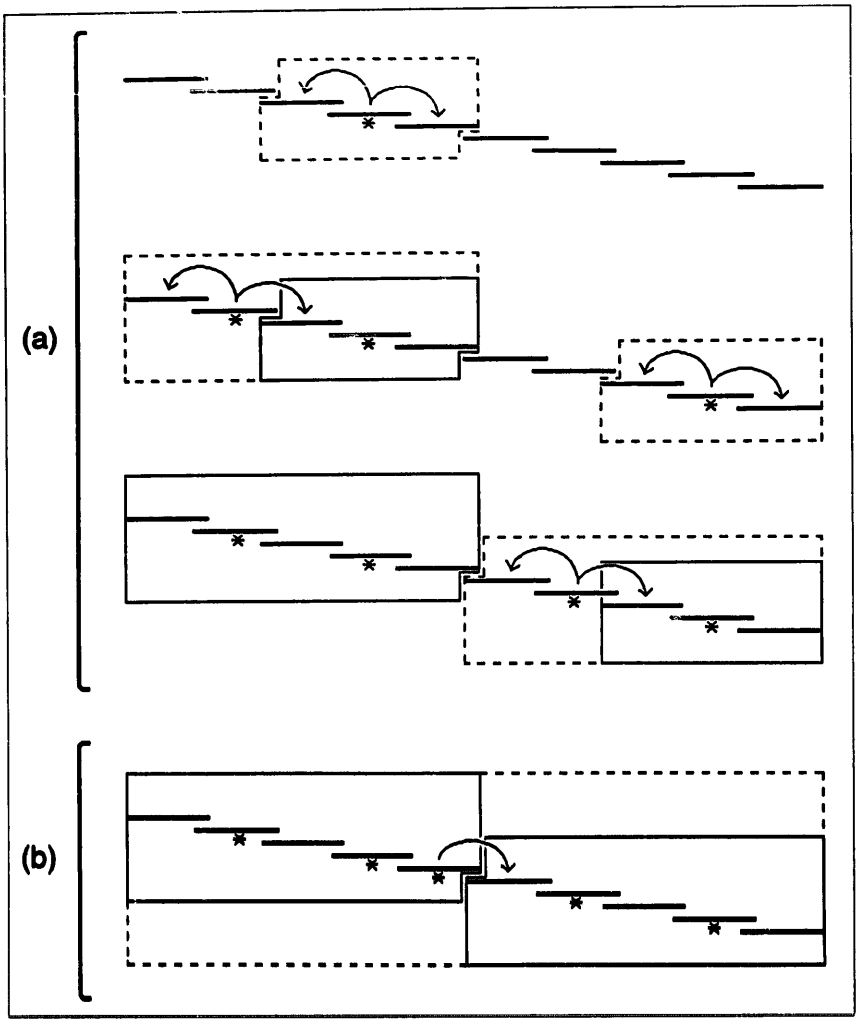

FIGURE 1. The two phases of the *S. pombe* mapping project. (a) Clones for use as probes were randomly picked (*) from a given set of cosmids whose map order was not known. Hybridization identified overlapping clones (arrows). From clones that did not give a positive signal in any earlier hybridization assay (unboxed areas), probes for the next round of experiments were chosen until all clones showed positive hybridization at least once. (b) Gaps in the map caused by the lack of probes for certain overlap regions were closed by using terminal contig clones.

be used, often simultaneously, for map generation, characterization of clones, sequence comparisons and, in principle, the complete determination of sequences[13]. The technique combines the high data output of hybridization techniques with an additional advantage: the amount of information obtained from a given experiment is independent of the genome size of the system being studied. Furthermore, assays that use oligomers as probes are much less subject to artefacts caused by repeat sequences than those using larger, cloned DNA probes.

Genomic and transcriptional mapping

The power of hybridization mapping has recently been illustrated by the completion of maps that span the *Schizosaccharomyces pombe* genome with YACs[14] and, at higher resolution, with P1 phage and cosmid clones[15,16]. Unique DNA probes were the main tool in ordering these clones. There is an important difference between these projects and STS mapping. While STS mapping relies on the presence of well-defined markers, the *S. pombe* mapping work used as probes anonymous clones from the libraries under investigation, selected only on the basis that they had not previously shown positive hybridization (Fig. 1). One group also took advantage of a parallel and related analysis of three different clone libraries[15]. The order of the probes was first established, and the clones were then fitted to that order on the basis of the very same set of hybridization results. Further advantages of the strategy are that the probes, although anonymous, are relatively evenly spaced throughout the genome, and that a redundant analysis of existing contigs is avoided. In an obvious extension of this approach, sequences cloned in cDNA and exonic libraries can be used as probes, combining the ordering of genomic DNA clones with localization of transcribed sequences. Simultaneous hybridization of such probes to genomic and transcriptional libraries would also yield information on sequence homologies between transcribed sequences.

Most published clone maps are based on unique probes (herein the term unique probes includes inter-repeat PCR products that characterize, for example, a particular YAC clone) or on repeat elements that define unique features of the DNA, such as restriction fragments. However, for effective fine-mapping in preparation for large-scale sequence analyses, the relative inefficiency of discrete single-copy probes means that it is almost always necessary to use non-unique probes. This can be achieved either by mixing unique markers in pooling schemes that allow a hybridization signal to be assigned to a particular marker in the pool[17], or by using sequences that occur very frequently in genomic DNA. The most widely applicable approach is to use short oligomeric sequences[13] (Fig. 2), which, unlike repeat elements, are represented in every type of genome at high frequency. As a bonus, the generation of an oligomer-based map can obviate the need to construct a restriction map. The concurrence of the sets of data obtained using either oligomers or unique probes to map *S. pombe* cosmids[15]

demonstrates the potential of the approach. However, while oligomer hybridization mapping is in theory a most efficient technology, some technical problems remain. One is the possibility of DNA contamination, a problem that is inherent, for example, with the *in situ* filters used in the *S. pombe* experiments. This can be overcome by using specifically designed oligomers whose sequences are represented only rarely in the genome of the vector host, but are frequently found in the cloned DNA of interest. In practice, only about one-third of oligomers designed actually gave this desired pattern of hybridization. Additionally, since the amount of target DNA on such filters varies greatly, the intensity of the hybridization signals also varies widely, so normalization and automated scoring of signals is essential. The occurrence of mismatch hybridization is not as problematic, so long as it occurs reproducibly with all targets. Analysis of fingerprint information for ordering the clones is relatively straightforward: even with average hybridization frequencies higher than the reported 3.8% (roughly equivalent to one binding event per megabase of sequence) most signals can be treated locally as being unique. Nevertheless, several algorithms for multilocus analysis have been described[13].

Data obtained from oligomer hybridization with selected (motif) and randomly generated oligomers not only provide fingerprint profiles for mapping purposes, but also yield structural information, such as the nature and position of repeat and regulatory elements. Thus, the suitability of a DNA segment for further investigation could be assessed before work intensive and costly sequence analyses are carried out. For the localization of transcriptional units, for example, Melmer and Buchwald[18] have shown that in more than 50% of cases at least one splice site of a given gene could be identified in cosmids by hybridizing two degenerate oligonucleotides corresponding to splice-site consensus sequences.

An extended form of the oligomer hybridization technique can produce high-resolution maps that could be used directly as an ordered template for sequence analyses[13,15]. Very short and pure DNA fragments, such

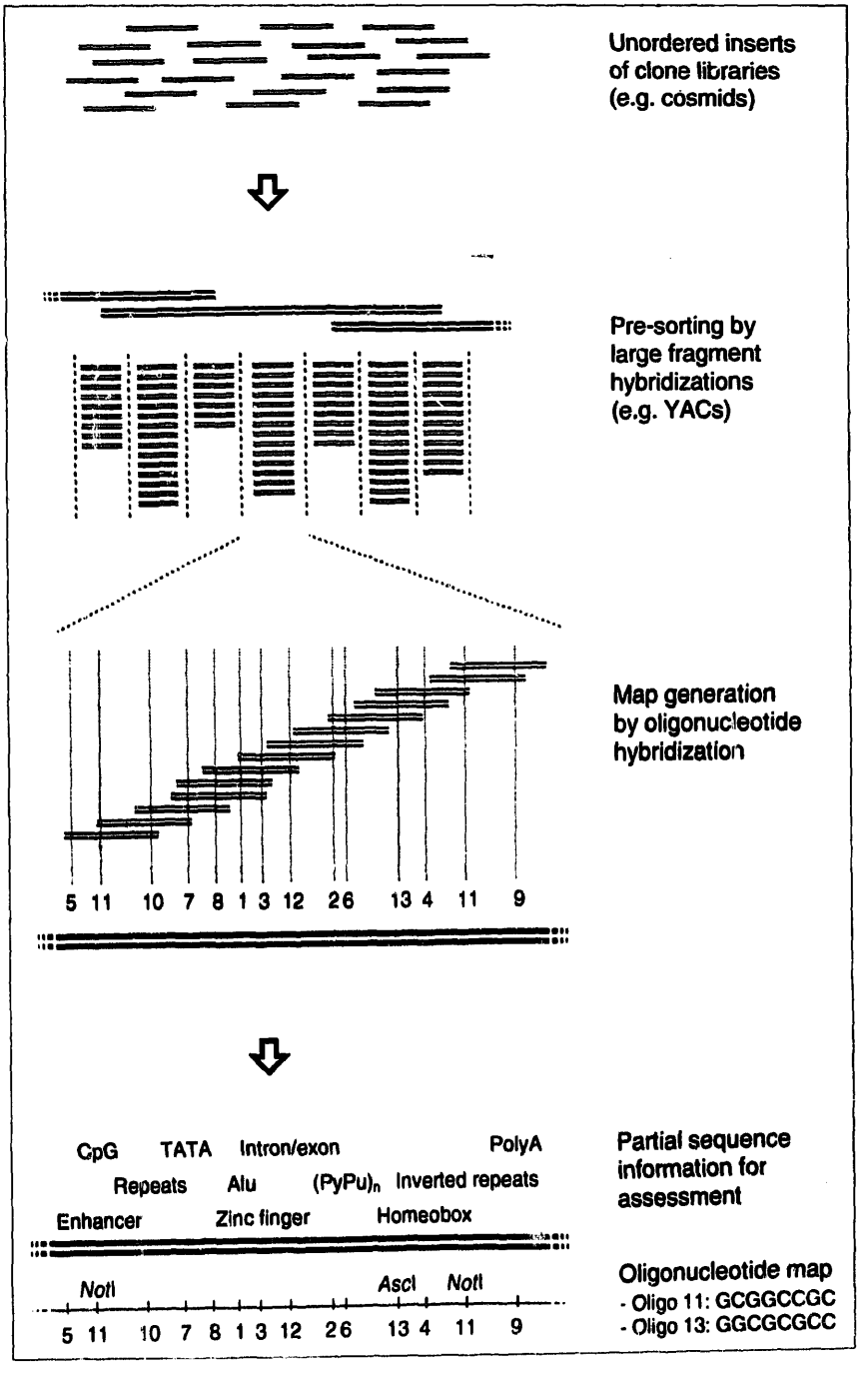

FIGURE 2. Oligomer mapping. A high-resolution library (e.g. cosmids) is subdivided by hybridization of low-resolution DNA fragments (e.g. YAC clones). Fingerprint data for establishing the order of clones are produced by hybridization with short oligonucleotides. Locally (i.e. within the intervals defined by the low-resolution fragments), most oligomers bind only once. Besides providing mapping data, this method yields an oligonucleotide map and partial sequence information concurrently.

as the widely used phage M13 DNA or PCR products[19], are the ideal substrate for very specific hybridization of short oligonucleotides (that contain between eight and ten nucleotides) under conditions that do not permit any mismatched base pairs[20], an approach that eliminates most of the technical problems described above. Since the number of different oligomer probes needed is independent of the size of the genome being screened, this procedure is particularly well-suited to analysis of large genomes. Assembly of a 'template' map would reduce the redundancy of such sequencing to a defined minimum, which could be varied locally according to the quality of the results obtained. Moreover, a minimal amount of sequence overlap would then be sufficient to allow contigs of sequences to be constructed.

Generation of gene inventories

The technology of zero-mismatch hybridization of very short oligomers is also being used to make gene inventories[19,21]. This system is based on the statistical probability of clones that share a given number of hybridization events having identical sequences. The degree of homology between sequences is determined from representative libraries of 10^4–10^5 transcribed DNAs, and clones are classified and catalogued this way (Fig. 3). Comparison with previously determined sequences entered in databases can reveal similarities to genes of known function, and even previously unknown sequences can often be grouped into a particular functional or structural category on the basis of consensus sequences. In contrast to the 'tag sequencing' technique[3], information is obtained from the entire length of a cDNA rather than from a single sequencing reaction at one end of the molecule. A complementary characterization by hybridizations with pooled total cDNA libraries can identify and compare gene expression patterns in tissue- or stage-specific libraries[22]. In this way, the experimental scope of 'tagging' experiments can be widened to allow the study of many different organisms; applications might include epidemiological investigations or basic research on, for example, the evolution of exons.

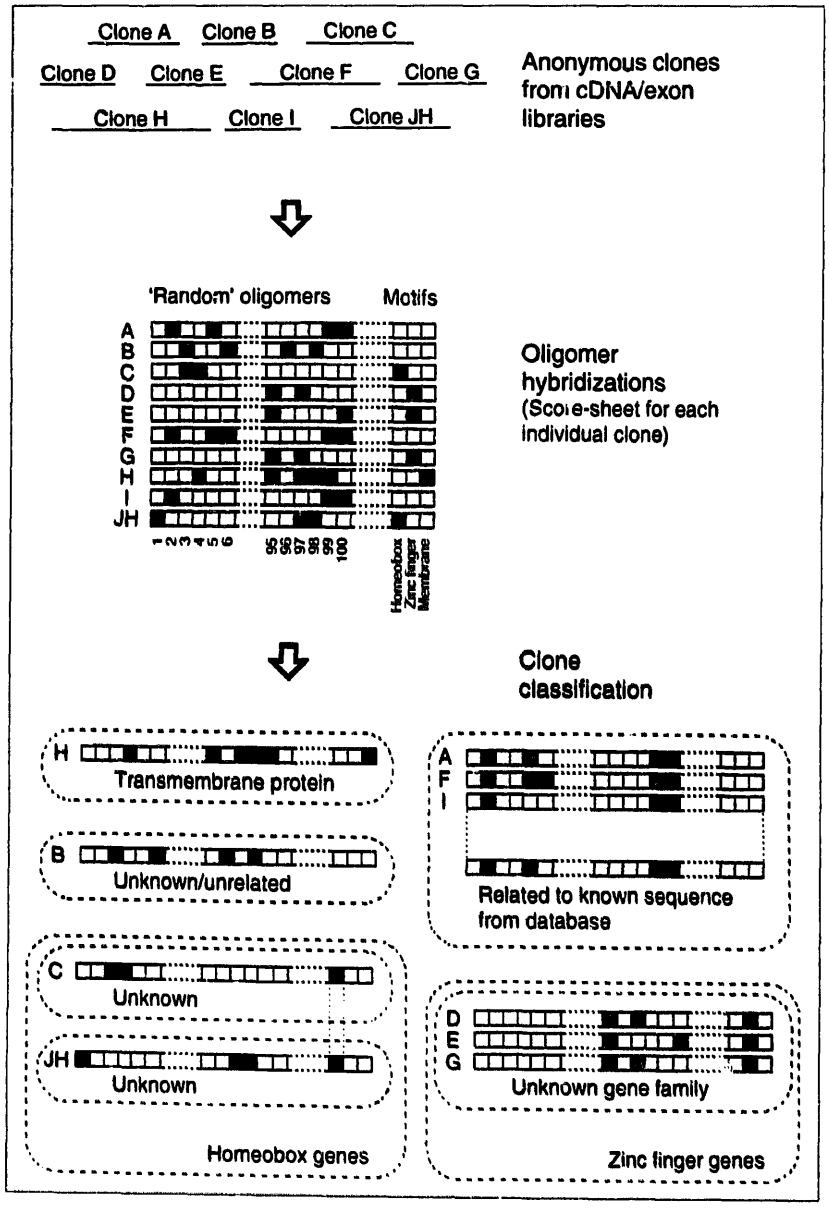

FIGURE 3. Oligomer profiling of representative (cDNA/exon) libraries. Hybridization with a set of short oligonucleotides produces partial sequence information on a clone library. On the basis of their oligomer signatures, clones can be grouped, or characterized by homology comparisons with known sequences.

Sequencing by hybridization

The ultimate objectives of the genome projects are the large-scale sequence analysis of genomes and the application of this information to individual case studies in research and medicine. Sequencing by oligomer hybridization has the potential to make an important contribution to both these aims. The technique has two basic formats, in which either the DNA fragment templates[20] or the oligomeric probes[23] are fixed to a solid support (Fig. 4). Many studies have addressed the technical problems involved, such as fixation of clones or oligomers and the stability of the duplexes formed. Since, in principle, an unlimited

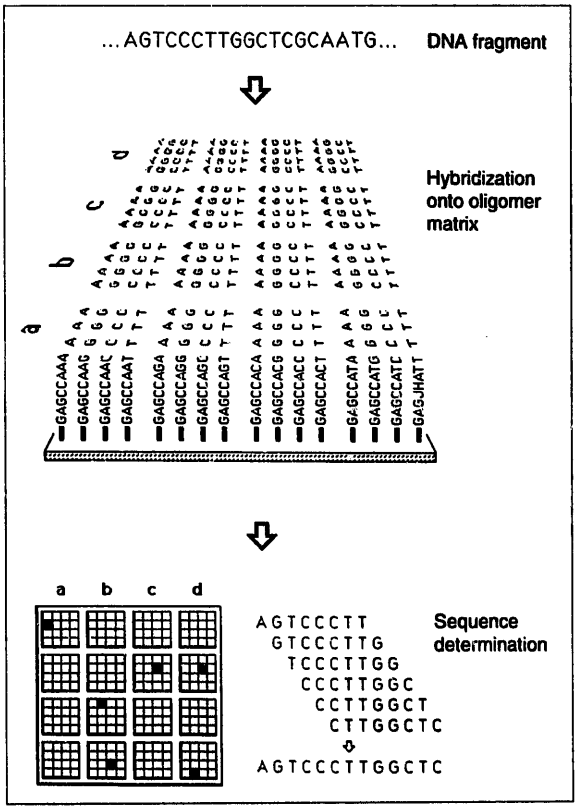

FIGURE 4. Schematic representation of one format of sequencing by hybridization. A given DNA fragment is hybridized to an oligomer matrix and binds to oligonucleotides that contain a complementary sequence. The continuous sequence is determined from collation of these oligomer sequences. (Modified, with permission, from Ref. 27.)

number of different oligomers can be synthesized directly on the surface of a hybridization matrix[24,25], an important clinical application of the technique has already become feasible. Once an association has been established between a particular DNA mutation and a human disorder, sequence analysis with a subset of relevant oligomers can characterize even single-base variations[26]. The potential of oligomer hybridization technology is illustrated by the fact that it has been used to determine the complete sequence of entire, although relatively small, DNA segments[20]. Since many of the biophysical problems involved in this methodology are at least partially understood, further refinements are now directed at improving the more technical aspects of the approach, such as developing rapid detection methods. With its high potential for automation, this technology brings sequence analysis of mammalian genomes well within the range of experimentation.

Conclusions

Analysis of extensive DNA regions using hybridization techniques, and in particular oligonucleotide hybridization, has developed from being an intriguing idea to become an established method. This is mainly as a result of two factors: first, there is now a core of scientists actively pursuing the advancement of the processes involved; and second, their work has already yielded important results that clearly illustrate its huge potential in many types of application. This technique has the capacity to relate directly information from many different types of nucleic acid analyses, and to process large numbers of samples in parallel and repeatedly, and provides an opportunity to optimize the effectiveness of experiments by means of interchanging the probe and target sequences and combining data from various types of analysis. It should therefore make a valuable contribution to the ongoing molecular exploration of genomes.

Acknowledgements

I am grateful to P. Lichter whose goodwill and computers were essential for the completion of this manuscript. I thank many colleagues, particularly those in H. Lehrach's laboratory, for valuable discussions of the ideas presented here.

References

1 Huntington's Disease Collaborative Research Group (1993) *Cell* 72, 971–983
2 Cohen, D. Chumakov, I. and Weissenbach, J. (1993) *Nature* 366, 698–701
3 Adams, M.D. *et al.* (1991) *Science* 252, 1651–1656
4 Kohara, Y., Akiyama, K. and Isono, K. (1987) *Cell* 50, 495–508
5 Coulson, A. *et al.* (1991) *BioEssays* 13, 413–417
6 Sidén-Kiamos, I. *et al.* (1990) *Nucleic Acids Res.* 18, 6261–6270
7 Trask, D. *et al.* (1992) *Genomics* 14, 162–167
8 Sulston, J. *et al.* (1992) *Nature* 356, 37–41
9 Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science* 245, 1434–1435
10 Chumakov, I. *et al.* (1992) *Nature* 359, 380–387
11 Foote, S., Vollrath, D., Hilton, A. and Page, D.C. (1992) *Science* 258, 60–66
12 Church, G.M. and Kieffer-Higgins, S. (1988) *Science* 240, 185–188
13 Lehrach, H. *et al.* (1990) in *Genome Analysis Volume 1: Genetic and Physical Mapping* (Davies, K.E. and Tilghman, S., eds), pp. 39–81, Cold Spring Harbor Laboratory Press
14 Maier, E. *et al.* (1992) *Nature Genet.* 1, 273–277
15 Hoheisel, J.D. *et al.* (1993) *Cell* 73, 109–120
16 Mizukami, T. *et al.* (1993) *Cell* 73, 121–132
17 Evans, G.A. and Lewis, K.A. (1989) *Proc. Natl Acad. Sci. USA* 86, 5030–5034
18 Melmer, G. and Buchwald, M. (1992) *Hum. Mol. Genet.* 1, 433–438
19 Meier-Ewert, S. *et al.* (1993) *Nature* 361, 375–376
20 Drmanac, R. *et al.* (1993) *Science* 260, 1649–1652
21 Drmanac, R. *et al.* (1992) *Electrophoresis* 13, 566–573
22 Gress, T.M. *et al.* (1992) *Mamm. Genome* 3, 609–619
23 Khrapko, K.R. *et al.* (1991) *DNA Sequence* 1, 375–388
24 Southern, E.M., Maskos, U. and Elder, J.K. (1992) *Genomics* 13, 1008–1017
25 Fodor, S.P. *et al.* (1991) *Science* 251, 767–773
26 Maskos, U. and Southern, E.M. (1993) *Nucleic Acids Res.* 21, 2269–2270
27 Hoheisel, J.D., Maier, E., Meier-Ewert, S. and Lehrach, H. (1993) *Ann. Biol. Clin.* 50, 827–829

J.D. HOHEISEL IS IN THE MOLECULAR GENETIC GENOME ANALYSIS LABORATORY, GERMAN CANCER RESEARCH CENTRE, IM NEUENHEIMER FELD 280, D-69120 HEIDELBERG, GERMANY.