METHODS: A Companion to Methods in Enzymology 9, 136–144 (1996)
Article No. 0018

Parallel Processing in Genome Mapping and Sequencing

Cassandra L. Smith,*,[1] Denan Wang,*,[2] Natalia Broude,* Nikolay Bukanov,* Galina S. Monastyrskaya,† and Eugene Sverdlov†

*Center for Advanced Biotechnology and Departments of Biomedical Engineering, Biology and Pharmacology, 36 Cummington Street, Boston University, Boston, Massachusetts 02215; and †Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, ul Miklukho-Maklaya 16/10, 117871 Moscow, Russia

Conventional genome mapping and sequencing involves the analysis and processing of individual samples and pieces of experimental data. Although these methods work, it is quite clear that more efficient and less expensive methods are needed. Our top down physical mapping experiments have focused on the parallel processing of information from multiple samples at one time. This approach has aided the construction of genomic restriction maps and allowed us to assess the degree of large-scale conservation across wide regions of the human genome. The principles of parallel processing were applied in top down experiments that ordered an overlapping cosmid library from the 14-Mb *Schizosaccharomyces pombe* genome. This approach produced an eight-fold increase in efficiency in clone ordering over similar efforts. Recently, we have developed an enhanced sequencing by hybridization protocol that allows DNA sequence information to be collected on a large number of samples at once. Our current research focuses on applying parallel processing principles to make genome-wide comparisons between pairs of samples for analyzing disease states.

In the past, genomic mapping and DNA sequencing methods focused on analyzing single samples one at a time. The complexity (e.g., single-copy DNA size) of the sample that could be analyzed was limited by the analytical method. In such experiments, information collected in series on different samples was compared after the primary data were obtained. Now, a number of techniques that allow the parallel processing of multiple samples of the same complexity have been developed. Examples of parallel processing are simultaneous analysis of arrays of samples or analysis of complex mixtures of samples. Obviously, these approaches create an increase in experimental efficiency by increasing the speed at which data accumulate. Some of the approaches use comparative information (map or sequence data on mixtures of samples or differences among these samples) to construct the primary information itself (map or sequence data on individual samples or species).

In recognition of the considerable increase in efficiency of parallel processing methods, many funding and scientific organizations have focused on developing genomic resources that can be utilized by multiple scientists simultaneously. Such resources provide access not only to primary material but also to primary information about such material.

This review focuses on describing how parallel processing methods have been applied in our past genomic mapping, library ordering, and DNA sequencing experiments. Also described are several comparative methods that use parallel processes to evaluate and identify DNA and RNA differences between pairs of samples.

DESCRIPTION OF THE METHOD

It is quite obvious that parallel processing of genomic samples potentially greatly increases the efficiency of experiments. What is not obvious is what the best way is to apply these principles to particular experiments, even though, amazingly, there is a limited repertoire of techniques that are used to analyze and manipulate nucleic acids. The available techniques include direct DNA sequencing (e.g., single-base determinations), hybridization analysis (multiple-base determinations), electrophoretic analysis (e.g., size and conformational determinations), and amplification (e.g., using the polymerase chain reaction (PCR)).

[1] To whom correspondence should be addressed. Telephone: (617) 353-8500. Fax: (617) 353-8501. E-mail: clsmith@darwin.bu.edu.

[2] Current address: Howard Hughes Medical Institute, Third and Parnassus Avenue, Room U-426, San Francisco, California 94143-0724.

1046-2023/96 $18.00

In most of these experiments, there is a probe and a target. Parallel processing uses pools of probes, pools of targets, or pools of both. This allows a number of data points to be collected simultaneously. The particular implementation is dependent on technical or analytical limitations or both. Some implementations use a hierarchical approach. The first tier experiments begin with very large pools, which are then subsequently broken down into smaller and smaller pools. If possible, it is also useful to take advantage of efficient methods like binomial sieving pooling to reduce the number of pools and to allow particular traits to be assigned to particular pool members.

RESULTS AND DISCUSSION

Genomic Restriction Mapping

Genomic restriction mapping involves cleaving genomic DNA with a restriction enzyme into a manageable number of megabase fragments. The fragments are fractionated by size using pulsed-field gel electrophoresis (PFG; 1, 2) and ordered by hybridization experiments with cloned DNA sequences. In many cases, other mapping experiments have already located a cloned DNA sequence to a particular chromosomal region with some degree of accuracy. Genomic restriction mapping focuses on linking neighboring restriction fragments in this region. Neighboring fragments are identified using a variety of hybridization experiments. These include the use of linking clones as hybridization probes. These clones span restriction enzyme cleavage sites, thus identifying adjacent fragments (3).

In many genomic mapping experiments, a single-copy sequence is used as a hybridization probe to DNA digested to completion with one or more enzymes or partially digested with a single enzyme. In the former case, the map is constructed by determining the overlaps between fragments from different enzymes. In the latter case, a ladder of fragments is identified; the difference in size between the fragments gives the distance between the restriction sites. We have combined these methods with another that we currently favor. The latter method uses a single complete enzyme digest of multiple DNAs. The analysis takes advantage of the differences and similarities in the samples to determine overlaps (see below).

A simple implementation of this was used in the construction of a *Not*I genomic restriction map for the 4.7-Mb *Escherichia coli* genome (4). This enzyme cleaved the *E. coli* genome into approximately 22 fragments that were clearly visible by direct ethidium bromide staining. Initially, a large number of *Not*I restriction fragments were regionally assigned by simply comparing the pattern of *Not*I restriction fragments from *E. coli* strains containing genetically characterized rearrangements. For instance, in one experiment seven *Not*I fragments were assigned by comparing the restriction fragment pattern of *E. coli* strains containing the 50-kb λ bacteriophage genome integrated into different genomic regions.

The *E. coli* effort was quickly followed by the construction of a genomic restriction map for the human histocompatibility locus (5). The publication of this map was shortly followed by the publication of a number of maps for this region in several DNA samples by others. Differences were readily apparent among these maps. Initially, it was not clear whether the differences in the maps were due to actual DNA polymorphisms. Hence, we subsequently examined (6) the physical structure of this region in a larger number of samples. These experiments showed that some megabase polymorphisms existed, but also that some differences that others had attributed to megabase polymorphisms could perhaps be traced to experimental artifacts such as those described in Doggett *et al.* (7) and also to differences in size standards used in particular studies (for a discussion see 8).

Even in these early mapping experiments, our efforts have focused on taking advantage of comparative information. Specifically, most of our work used one enzyme and multiple samples. Many of these experiments used the restriction enzyme *Not*I. This enzyme has been a fairly consistent focus because it produces the largest average-sized fragments from quite a few genomes, including the human genome. The *Not*I sites in mammalian genomes are further distinguished from other restriction enzymes that produce large fragments because most sites are either totally methylated or totally unmethylated. This appears to arise from the fact that most of the *Not*I sites are located in unmethylated CpG-rich islands that are found preferentially at the 5′ end of genes (9).

An approach we have called "polymorphism link-up" was used quite extensively for the construction of a chromosome 21 *Not*I restriction map (8). Here, neighboring *Not*I fragments were identified by analyzing the *Not*I polymorphisms present in a set of DNA samples (Fig. 1). For instance, if a *Not*I site is absent in one cell line and present in another, a hybridization probe (A) located near one end of the fragment will detect a small fragment of size X in the latter cell line and a large fragment of size Z in the former cell line. Another hybridization probe (B) located at the other end of fragment Z will detect a fragment of size Y (which is equal to the size difference between X and Z) in samples that contain fragment Y. Both probes will detect fragment Z wherever the *Not*I site is missing. Such an analysis of the large restriction fragment polymorphisms helps in the map construction by showing that probes A and B are in fact in the same region and do not coinciden-

tally detect a similar-sized fragment Z in some cell lines and unrelated fragments X and Y in other cell lines.

The polymorphism link-up approach was combined with hybridization analysis of complete and partial *Not*I digests with single-copy sequences, *de novo* isolated linking and telomere clones, and human-specific repetitive sequences to construct what must be one of the largest low-resolution genomic mapping datasets. The chromosome 21q *Not*I map was created by ordering 60 distinct *Not*I restriction fragments (totaling 43 Mb), 80 DNA markers, and 11 chromosomal breakpoints in nine different cell lines containing an unselected sample of chromosomes (8). The map revealed a remarkable large-scale conservation of the human chromosome 21q arm. All hybridization probes were present in all samples and no large-scale insertions, deletions, or rearrangements were detected.

Preservation of chromosome structure might be expected in diploid human cells, perhaps by mechanisms that involved chromosome pairing during cell division. However, our study revealed that the large-scale structure of human chromosome 21q was preserved in cell lines that included monosomic rodent human hybrids containing single copies of chromosome 21. It is not clear what forces would preserve large-scale chromosome structure in human or monosomic hybrid cell lines, especially in hybrid cell lines containing single chromosomes copies such as those used in this study.

A single map made using a single DNA source is useful for a number of applications. However, it is also quite clear that the usefulness of maps increases as their structure in a population is known. Such informative comparative information is inherently provided in some of the approaches described above.

Clone Library Ordering

The utility of parallel processing in ordering genomic libraries was recently demonstrated for the *Schizosaccharomyces pombe* genome (10). In these experiments a ~1700-cosmid clone library for this 14-Mb genome was ordered by 61 hybridization experiments. This research represented an eight-fold increase in efficiency over previous efforts to order overlapping cosmid clones by other groups.

The top down hybridization experimental approach that we used on *S. pombe* is schematized in Fig. 2. The first three hybridization experiments used PFG-purified chromosomal DNA to assign cosmids to one of the three *S. pombe* chromosomes. The second tier hybridization experiments assigned the clones to chromosomal regions. In these experiments, the hybridization probes were PFG-purified large genomic restriction fragments. The number of experiments was minimized by pooling one restriction fragment from each of the three chromosomes for each hybridization experiment. The fact that the chromosomal assignment was known in advance from other experiments meant that regional assignments on specific chromosomes could be done in parallel.

Next, the clones were hybridized *en masse* to pools of probes randomly distributed along the genome. These pools were generated by cleaving genomic DNA with a restriction enzyme containing 4- or 6-bp recognition sites. The fragments were separated by size electrophoretically and collected into fractions by cutting the gel lanes into a set of slices. The DNA contained in each piece was used as a hybridization probe.

The hybridization probes used in the first and second tier experiments were considerably larger than the cloned sequences. The hybridization probes used in the third tier of experiments were similar in size or smaller than the cloned sequences. This meant that clones that showed coincident hybridization patterns in the first and second tier experiments need not overlap. Clones that were coincident in some or most third tier hybridization experiments were most likely overlapping, so

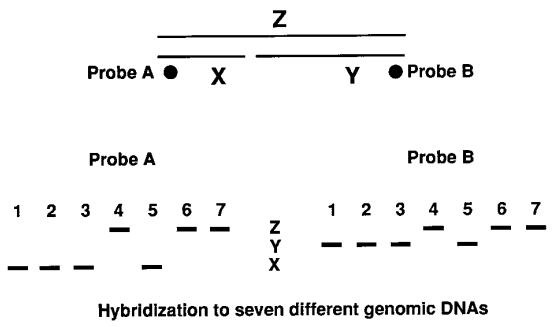

FIG. 1. An example of polymorphism link-up in which a set of seven DNAs from different sources are studied simultaneously. A hypothetical case is shown where a single restriction site is polymorphic so that it is cut in some cell lines, giving fragments X and Y, and not cut in others, giving fragment Z (see text for details).

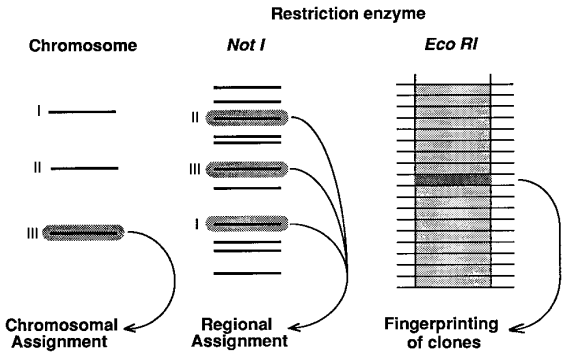

FIG. 2. Efficient ordering of genomic libraries by parallel fingerprinting of regionally assigned, overlapping clones. Three successive sets of probings are used to narrow down the location of large numbers of probes and provide information on probe overlap.

long as they had previously been shown to be contained within the same region.

DNA Sequencing

Conventional gel-based DNA sequencing methods in widespread use today focus on analyzing single samples at one time. Usually, comparative studies are done after the DNA sequence is obtained. These comparative studies have led to a large number of biological insights. More recently, several groups have focused on developed DNA sequencing methods that take advantage of parallel processing principles.

Church and Kieffer-Higgins (11) described a parallel process DNA sequencing method that they called "multiplexing." In this method, 20 different vectors are used for cloning. The vectors are distinguishable because the DNA sequence on both sides of the cloning site in each vector is unique. One clone from each library is pooled; the resulting mixture is connected to a mixture of sequencing ladders that is fractionated on a polyacrylamide gel. Thus, the number of rate-limiting size fractionation steps needed is minimized. The fractionated DNA is transferred to a membrane. Then, 16 consecutive hybridizations are performed on the same membrane. Each hybridization is done with a probe corresponding to one of the unique sequences surrounding one of the cloning sites of one of the vectors.

An alternative method for conservation of effort in DNA sequencing projects involves the use of short modular primers in primer walking approaches (12–15). Here, short probes (primers) can be used in parallel on a large number of targets to build up a set of stacked primers of sufficient length to prime DNA sequencing reactions. The shortness of the individual primers, 6 nt in the initial implementations, precludes their individual use in the DNA sequencing reaction.

Sequencing by hybridization (SBH; 16, 17) is another alternative method that not only analyzes samples in parallel but also uses the information collected in parallel to reconstruct the sequences. There are two formats used in SBH approaches. In one format, format I, different (unknown) target sequences are immobilized in arrays (16). The arrays are interrogated with oligonucleotide probes of known sequence. Thus, each experiment collects bits of sequence information about a number of different samples. For *de novo* sequencing efforts, it is particularly powerful to build up the primary sequence by analyzing similar samples at the same time, e.g., different alleles of the same gene. In a second format, format II, arrays of oligonucleotide probes of known sequences are immobilized (17). The probe arrays are used to interrogate single samples one at a time. In these approaches, the probe sequences are 8–9 nt. The sequence is reconstructed by attempting to determine the minimum tiling path between the positive probe signals. The minimum tiling path is the shortest continuous sequence that contains all the sequences detected and does not contain any specific sequences that were tested for and not detected.

Different 8- to 9-nt sequences anneal to their complementary sequences at significantly different temperatures. Thus, hybridization experiments must be done over a range of temperatures to distinguish matched from mismatched sequences. Even so, discrimination of matched from some mismatched sequences is difficult, especially when end mismatches occur. In many cases the difference in hybridization signal intensity between matched and end-mismatched 8- to 9-nt sequences is only a factor of two, and in some cases there is no discrimination. Hence, these experiments cannot be done under a single set of experimental conditions.

Recently, an enhanced method of SBH was described by us (18). This method, called positional SBH, reads sequence only at the end of a duplex (Fig. 3). In this method probe sequences consist of a duplex region and a 5- to 6-nt single-stranded overhang. Stacking interactions between the perfectly matched duplex probe sequences and the hybridized target sequences provide for enhanced discrimination between matched and mismatched sequences. Additional, enzymatic enhancements increased the discrimination between matched and mismatched sequences. In one format, DNA ligation was used to covalently link the matched target to the probe sequence. In another format, the 5-nt probe sequence was used to prime a DNA polymerase reaction. In a model system, the hybridization intensity difference between matched and end-mismatched sequences varied between 20 and 100 under a single set of experimental conditions. This method also provides a powerful sequence-specific capture protocol for input into other sequencing or DNA analysis procedures (see below).

Differential Display/Comparative Genome Hybridization

Traditional experiments focus on analyzing the expression of single genes one at a time. An alternative approach focuses on developing expression profiles of cells. For instance, it is possible to use conventional sequencing methods to sequence large numbers of cDNAs from a single cell type. A comparison of the expression profiles of appropriate pairs of samples highlights differences between them. This approach identifies known and unknown mRNAs (e.g., cDNAs) and provides comparative information about their relative levels in different cells. This method does not efficiently sample genes that are expressed at low levels. This cDNA profiling method evaluates the behavior of several thousand genes in a typical sample. Note that it is estimated that there are 50,000–100,000 genes in the human genome. Current sequencing costs do not allow this type of information to be collected in parallel on a large number of samples. Instead a number of

techniques have been developed to highlight the differences between pairs of samples.

One alternative to direct DNA sequencing is a method that has been called differential display. This method depends on randomly primed PCR to amplify unknown DNA sequences in pairs of samples. The PCR reaction serves to test for the presence of multiple, unknown sequences and to reduce the complexity of the genome to a level that can be analyzed. The multiple PCR products produced from different samples are compared after size fractionation by electrophoresis. This method has been applied to analyzing genomic DNA and mRNA.

The differential display method uses arbitrary PCR primers to amplify DNAs (or RNAs) of unknown sequence. Williams *et al.* (19) used a pool of primers of variable composition to amplify the unknown sequences. Welsh and McClelland (20) used a two-step PCR amplification method. In the first step, a single PCR primer was used in two low-stringency (low temperature) PCR cycles. The low stringency allows imperfectly matched primers to initiate DNA synthesis. The primers, now located on the ends of the products of the first two PCR cycles, serve as tags for subsequent high-stringency PCR cycles.

In practice, these methods have not been very reproducible even among researchers in the same laboratory. Thus, a number of improved protocols have been developed. Some protocols simply ligate tags onto restriction fragments (21, 22). Other approaches use partially or completely degenerate primers for PCR amplification (23, 24). These methods often produce uneven

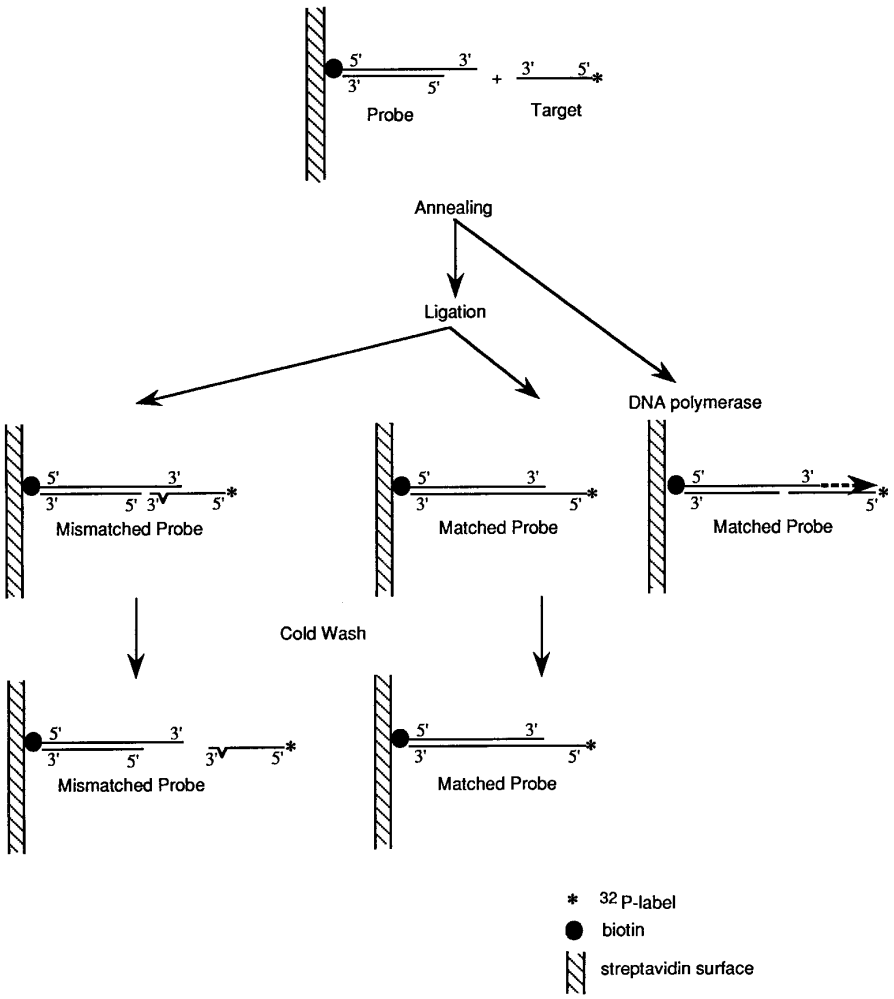

FIG. 3. Enzyme-enhanced sequencing by hybridization. A probe with a five-base 3′ single-stranded overhang is used to capture a target. The fidelity of the capture can be enhanced in two ways. DNA ligase can be used to check that correct base pairing has occurred at the 3′ end of the target. A cold wash removes targets that have not been ligated. DNA polymerase I extension can be used to check that correct base pairing has occurred at the 3′ end of the probe. The polymerase will incorporate label only if it is presented with a template–primer complex with correct terminal base pairing. In practice, both proofreading methods can be combined in a single protocol.

DNA amplifications and products unrelated to the target sequence, by the formation of primer dimers. One method, tagged PCR (T-PCR; 25) developed by us uses a two-step PCR protocol. The first two rounds of PCR are done with a chimeric primer consisting of a 5′-constant and a 3′-variable region. The products of these two rounds of PCR have the constant ("tagged") sequence at their ends. The chimeric primer is removed, and subsequent PCR cycles are done with primers complementary only to the constant or tag sequence.

The principles of randomly primed PCR have been applied to analyzing cellular RNA transcripts *en masse* (26–29). The PCR products are fractionated by size so that they can be displayed simultaneously. This method has been called differential display because the transcripts from many samples can be analyzed in parallel on a single gel. Ito *et al.* (29) have published two protocols that appear to be quite robust for differential display. In these protocols, randomly primed PCR is used to add the fluorescently labeled PCR primer. The patterns of PCR products are compared by fractionation on a high-resolution polyacrylamide gel using an automatic DNA sequencer. The results are recorded and analyzed electronically. This use of an automatic DNA sequencer allowed for high-throughput automatic analysis of multiple samples.

Comparative genome hybridization (CGH; 30) allows whole genome comparisons. However, CGH is currently a rather low (10–30 Mb)-resolution method. In CGH, DNA from two different cell types is amplified as uniformly as possible, using randomly primed PCR methods, so that the products are differentially labeled. One DNA probe is labeled with biotin and the other DNA probe is labeled with digoxigenin. Metaphase chromosomes from normal cell lines are used as targets for fluorescence *in situ* hybridization (FISH) experiments, employing equal amounts of the differentially labeled DNA probes simultaneously. After hybridization, the two DNA probes are differentially stained, the first with fluorescein-conjugated avidin (or streptavidin) and the second with rhodamine-conjugated anti-digoxigenin. Unlabeled, highly repetitive human DNA, i.e., $C_0t = 1$ DNA, is used as a competitor to prevent the FISH images from being overwhelmed by the more efficient hybridization of high-copy-number repeats. The resulting pattern of simultaneous hybridization of the two probe sets is examined by quantitative fluorescence microscopy. Simultaneous hybridization of DNA from the same cell labeled with two different fluorophores serves as a control to compensate for inherent experimental variation in different regions of the genome.

A two-color assay allows sequences differentially present, or absent, in a pair of samples to be highlighted. For example, fluorescein florescence is green to the human eye, rhodamine is red, and the combination is yellow. Thus, regions of the genome that are the same in the two different probe sets will appear yellow, while regions deleted in one probe will take on the color of the other; regions amplified in only one probe set will show that color, predominantly because of the more efficient kinetics of hybridization at higher probe concentrations. CGH has proven to be extraordinarily useful in providing a rapid overview of DNA rearrangements in various types of tumor cells (30, 31).

CGH provides positional information about genomic sequences, but it does not provide the sequences themselves. However, once regions containing known candidate genes are targeted, these candidate sequences are available for further testing. When no such candidate genes are available, further laborious conventional positional cloning experiments are necessary. The great power of CGH is that the entire genome is scanned at once, and attention is focused on just those regions where significant differences in DNA content occur.

Subtractive Hybridization

Subtractive hybridization methods select DNAs that are present in one sample and absent in another sample. Subtractive methods not only provide a means of isolating sequences present in only one of a pair of samples, but they also provide a means of analyzing sequences that are present at very low levels. Subtractive hybridization was used some time ago to isolate the gene involved in Duchenne muscular dystrophy (32). The original method, as well as a number of derivative methods developed since then, were quite difficult to use. A very efficient, genomic subtraction scheme has recently been described by Sverdlov and several of his former co-workers (33–40). The DNA sample that is missing the sequence of interest, i.e., the target DNA, is called the driver DNA. The sample that contains the target sequence is called the tracer DNA. In addition to the target DNA, the tracer DNA contains all sequences that are present in the driver DNA. One round of subtraction will increase the relative amount of sequence unique to the target (i.e., absent in the driver) by the ratio of the starting concentrations of target and driver. A second round of subtraction will give a purification proportional to the cube of this ratio (33). The improved procedure described below addresses problems encountered in the original protocols.

The goal of subtraction is to enrich the amount of the target DNA in a tagged tracer sample. This is done in several ways. The first step of subtraction involves hybridization of the tagged tracer DNA to a large excess of biotinylated, differentially tagged driver DNA. The two DNA samples are denatured, and the single strands are mixed together, usually at a driver:tracer ratio of 100:1 (mole:mole). The presence of excess driver

means that a tracer sequence is much more likely to find its driver complement than its tracer complement.

Both the double-stranded hybrid molecules (driver:-tracer) and the double-stranded driver molecules will have biotin located at one or two ends, respectively. The biotin is used to capture these molecules on streptavidin-coated magnetic microbeads. Then beads are removed along with the bound biotinylated DNA molecules. The tracer DNA remaining in the supernatant is amplified using primers specific for the tags of the tracer DNA. The presence of specific tracer tags allows the tracer DNA to be amplified in the absence of amplification of the driver sequences. The PCR-amplified product is then subjected to additional rounds of subtraction until material of the desired purity is obtained.

The time needed for hybridization depends on the concentration of the DNA samples. A high concentration of driver DNA allows hybridization to occur within a reasonable period of time (e.g., usually overnight). The concentration of specific DNA sequences is also increased by reducing the complexity of the sample. This can be done by focusing on cDNA libraries and hence only on expressed sequences (estimated to represent about 5–10% of the entire haploid human genome, of 3×10^9 bp). The complexity is also reduced by typical PCR reaction conditions. For instance, conventional PCR preferentially amplifies small fragments. Multiple tag sequences added to the ends of the genomic DNA sequences allow for PCR amplification and subtraction of different subsets of the genome.

One subtraction protocol has the potential to be applied to the entire genome in a single set of experiments. In this procedure, two genomic DNA samples, cut with the same restriction enzyme, are mixed and fractioned by size electrophoretically (41, 42). This enhances subtraction because each subtraction is performed on DNA contained within one gel slice. This is very similar to coincidence cloning methods that have been described that clone only those sequences that are present in two samples (43).

CONCLUDING REMARKS

Some of our recent experiments have focused on further improving the efficiency of comparative genomic procedures. Recent genomic mapping experiments on chromosome 20 have focused on using PCR or PCR-enhanced hybridization methods to analyze gel slices containing electrophoretically fractionated *Not*I restriction fragments (44; Bukanov *et al.*, unpublished results). In this approach, gel slices containing DNA are analyzed directly using PCR primers specific for single-copy sequences such as those used in genetic mapping experiments. Use of the single set of genetic markers developed by Weissenbach and colleagues (45) facilitates mapping, since all experiments are done using a single set of PCR conditions. Not all markers need to be tested against all slices. Instead, intelligent pooling of the slices is used to minimize the PCR experiments. A similar approach has been used to analyze YAC libraries (46). Alternatively, if the DNA source is a monosomic hybrid cell line, then inter-*Alu* PCR (47) can be used to amplify and fingerprint the human DNA contained in each gel slice.

We are developing the use of such slices as targets in a modified CGH approach with higher resolution and simpler image analysis. Digestion of total human DNA with *Not*I and fractionation by PFG slices can divide the entire human genome into 150 fractions. The DNA in each fraction could be cleaved into smaller pieces using a second restriction enzyme and subjected to a second size fractionation. If the second fractionation were divided into 10 slices, the entire human genome would then be divided into 1500 fractions having on average about 2-Mb resolution (obviously, the genomic DNA could theoretically be divided into any number of fractions). These samples could be arrayed and used as targets and/or probes in two-color CGH experiments. The 2-Mb complexity samples are equivalent to large genomic clones. In contrast to large clones, such arrays represent unrearranged genomic samples, and they can be relatively easily generated from multiple samples.

The DNA contained in each of the gel slice fractions will most likely have to be analyzed after PCR amplification, since current high-resolution electrophoretic fractionations require the use of small amounts of sample. For instance, the DNA contained in a gel slice could be amplified by T-PCR (25). The T-PCR products can be used as targets for hybridization or as templates for additional PCR reactions with different proteins. Hybridization analysis of such samples is simplified because of the possibility of using higher concentrations of DNA than in conventional directly blotted PFG-fractionated genomic DNA.

Other ongoing experiments are using sequence-specific capture methods to compare different subsets of the genome. These build on methods developed by us and others to purify single-stranded and double-stranded sequences containing specific sequences from complex samples (48–53). The first method we developed (48–50) captured homopurine–homopyrimidine stretches by taking advantage of the fact that these sequences from triplex structures at low pH. Hence, an immobilized single-stranded probe can be used to capture simple repeat sequences in duplex DNA. Recently, we have applied these methods to making libraries enriched in tandemly repeating trinucleotide sequences (triplet repeats).

Expansion in triplet repeat sequences has been asso-

ciated with an increasing number of human genetic diseases (54). Conventional positional cloning experiments that have uncovered diseases caused by triplet repeat mutations focused on the testing of individual known repeats found in candidate regions. In contrast, our method captures restriction fragments that contain specific repeats and also differentially displays the captured restriction fragments from multiple samples in protocols that are similar to those by Ito *et al.* (29), described above. Other experiments in progress are on developing the use of arrays of samples to replace the size-fractionation step.

ACKNOWLEDGMENTS

The work was supported by grants from the USDOA (A1BS2154) and NATO (HTECH.LG 94057) TO CLS.

REFERENCES

1. Schwartz, D. S., Saffran, W., Welsh, J., Haas, R., Goldenberg, M., and Cantor, C. R. (1983) *Cold Spring Harbor Symp. Quant. Biol.* 47, 189–192.
2. Schwartz, D. C., and Cantor, C. R. (1984) *Cell* 36, 67–64.
3. Smith, C. L., Warburton, P., Gaal, A., and Cantor, C. R. (1986) *in* Genetic Engineering (Setlow, J. K., and Hollaender, A., Eds.), Vol. 8, pp. 45–70, Plenum, New York.
4. Smith, C. L., Econome, J., Schutt, A., Klco, S., and Cantor, C. R. (1987) *Science* 236, 1448–1453.
5. Lawrence, S. K., Smith, C. L., Weissman, S. M., and Cantor, C. R. (1987) *Science* 235, 387–1390.
6. Lawrence, S. K., and Smith, C. L. (1990) *Genomics* 8, 394–399.
7. Doggett, N. A., Smith, C. L., and Cantor, C. R. (1992) *Nucleic Acids Res.* 89, 859–864.
8. Wang, D., and Smith, C. L. (1994) *Genomics* 20, 441–451.
9. Bird, A. P. (1987) *Trends Genet.* 3, 342–347.
10. Grothues, D., Cantor, C. R., and Smith, C. L. (1994) *Proc. Natl. Acad. Sci. USA* 91, 4461–4465.
11. Church, G. M., and Kieffer-Higgins, S. (1988) *Science* 240, 185–188.
12. Kotler, L. E., Zevin-Sonkin, D. I., Sobolev, A., Beskin, A. D., and Ulanovsky, L. E. (1993) *Proc. Natl. Acad. Sci. USA* 90, 4241–4245.
13. Szybalski, W. (1990) *Gene* 90, 177–178.
14. Kieleczawa, J., Dunn, J. J., and Studier, F. W. (1992) *Science* 258, 1787–1791.
15. Azhikina, T., Veselovskaya, S., Myasnikov, V., Potapov, V., Ermolayeva, O., and Sverdlov, E. (1993) *Proc. Natl. Acad. Sci. USA* 90, 11460–11462.
16. Drmanac, R., Drmanac, S., Strezoska, Z., Paunesku, T., Labat, I., Zeremski, M., Snoddy, J., Funkhouser, W. K., Koop, B., Hood, L., and Crkvenjakov, R. (1993) *Science* 260, 1649–1652.
17. Khrapko, K. R., Lysov, Y. P., Khorlin, A. A., Ivanov, I. B., Yershov, G. M., Vasilenko, S. K., Florentiev, V. L., and Mirzabekov, A. D. (1991) *J. DNA Sequencing Mapping* 1, 375–388.
18. Broude, N., Sano, T., Smith, C. L., and Cantor, C. R. (1994) *Proc. Natl. Acad. Sci. USA* 91, 3072–3076.
19. Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. (1990) *Nucleic Acids Res.* 18, 6531–6535.
20. Welsh, J., and McClelland, M. (1990) *Nucleic Acids Res.* 18, 7213–7218.
21. Ludecke, H., Senger, G., Claussen, U., and Horsthemke, B. (1989) *Nature* 338, 348–350.
22. Kinzler, K. W., and Vogelstein, B. (1989) *Nucleic Acids Res.* 17, 3645–3653.
23. Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. (1992) *Proc. Natl. Acad. Sci. USA* 89, 5847–5851.
24. Telenius, H., Carfter, N. P., Bebb, C. E., Nordenskjold, M., Ponder, B. A., and Tunnacliffe, A. (1992) *Genomics* 13, 718–725.
25. Grothues, D., Smith, C. L., and Cantor, C. R. (1993) *Nucleic Acids Res.* 21, 1321–1322.
26. Liang, P., and Pardee, A. B. (1992) *Science* 257, 967–971.
27. Liang, P., Averboukh, L., and Pardee, A. B. (1993) *Nucleic Acids Res.* 21, 3269–3275.
28. Wong, K. K., and McClelland, M. (1994) *Proc. Natl. Acad. Sci. USA* 91, 639–643.
29. Ito, T., Kito, K., Adati, N., Mitsui, Y., Hagiwara, H., and Sakaki, Y. (1994) *FEBS Lett.* 351, 231–236.
30. Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, R., and Pinkel, D. (1992) *Science* 258, 818–821.
31. Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W. L., Waldman, F. M., Isola, J. J., Grey, J. W., and Kallioniemi, O. P. (1994) *Cancer Res.* 54, 4257–4256.
32. Hoffman, E. P., Brown, R. J., and Kunkel, M. L. (1987) *Cell* 51, 919–928.
33. Sverdlov, E. C. (1993) *Crit. Rev. Mol. Genet. Microbiol. Virusol.* 3, 26–29.
34. Sverdlov, E. D., and Ermolayeva, O. V. (1993) *Bioorg. Khim.* 19, 1081–1088.
35. Sverdlov, E. D., and Ermolayeva, O. V. (1994) *Bioorg. Khim.* 20, 506–514.
36. Lisitsyn, N. A., Lisitsyn, N., and Wigler, M. (1993) *Science* 259, 946–951.
37. Lisitsyn, N. A., Launer, G. A., Wagner, L. L., Akopyanz, N. S., Martynov, V. I., Levlikova, G. P., Limborska, S. A., Polukarova, L. G., and Sverdlov, E. D. (1993) *Biomed. Sci.* 1, 513–516.
38. Lisitsyn, N., Rosenberg, M., Launer, G., Wagner, L., Potapov, V., Kolesnik, T., and Sverdlov, E. D. (1993) *Mol. Genet. Mikrobiol. Virusol.* 3, 26–37.
39. Lisitsyn, N. A., Segr, J. A., Kusumi, K., Lisitsyn, N. M., Nadeau, J. H., Frankel, W. N., Wigler, M. H., and Lander, E. S. (1994) *Nature Genet.* 6, 57–63.
40. Zaraisky, A. G., Lukyanov, S. A., Vasiliev, O. L., Smirnov, Y. Y., Beliavski, A. V., and Kazanskaya, O. V. (1992) *Dev. Biol.* 152, 373–382.
41. Yokota, H., and Oishi, M. (1990) *Proc. Natl. Acad. Sci. USA* 87, 6398–6402.
42. Yokota, H., Amano, S., Yamune, T., Ataka, K., Kikuya, E., and Oishi, M. (1994) *Anal. Biochem.* 219, 131–138.
43. Brookes, A. J., and Porteous, D. J. (1991) *Nucleic Acids Res.* 19, 2609–2613.
44. Wang, D., Zhu, Y., and Smith, C. L. (1995) *Genomics* 26, 318–326.
45. Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Mil-

lasseau, P., Marc, S., Bernardi, G., Lathrop, M., and Weissenbach, J. (1994) *Nature Genet.* 7, 246–339.

46. Green, E. D., and Olson, M. V. (1990) *Proc. Natl. Acad. Sci. USA* 87, 1213–1217.
47. Nelson, D. L., Ledbetter, S. A., Corbo, L., Victoria, M. F., Ramirez-Solis, R., Webster, T. D., Ledbetter, D. H., and Caskey, C. T. (1989) *Proc. Natl. Acad. Sci. USA* 86, 6686–6690.
48. Ito, T., Smith, C. L., and Cantor, C. R. (1992) *Proc. Natl. Acad. Sci. USA* 89, 495–498.
49. Ito, T., Smith, C. L., and Cantor, C. R. (1992) *Genet. Anal. Tech. Appl.* 9, 96–99.
50. Ito, T., Smith, C. L., and Cantor, C. R. (1992) *Nucleic Acids Res.* 20, 3624.
51. Kandpal, R. P., Kandpal, G., and Weissman, S. M. (1994) *Proc. Natl. Acad. Sci. USA* 91, 88–92.
52. Kijas, J. M. H., Fowle, J. S. C., Garbett, C. A., and Thomas, M. R. (1994) *BioTechniques* 16, 657–662.
53. Tagle, D. S., Swaroop, M., Lovett, M., and Collins, F. S. (1993) *Nature* 361, 751–753.
54. Williams, P. J. (1994) *Nature Genet.* 8, 213–215.