Annu. Rev. Biochem. 1997. 66:751–83
Copyright © 1997 by Annual Reviews Inc. All rights reserved

SUBTRACTIVE CLONING: Past, Present, and Future

C. G. Sagerström,[1] *B. I. Sun,*[1] *and H. L. Sive*[2]

Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142

KEY WORDS: differential gene representation, subtraction, positive selection, hybridization, reassociation kinetics

ABSTRACT

Subtractive cloning is a powerful technique for isolating genes expressed or present in one cell population but not in another. This method and a related one termed positive selection have their origins in nucleic acid reassociation techniques. We discuss the history of subtractive techniques, and fundamental information about the nucleic acid composition of cells that came out of reassociation analyses. We then explore current techniques for subtractive cloning and positive selection, discussing the merits of each. These techniques include cDNA library–based techniques and PCR-based techniques. Finally, we briefly discuss the future of subtractive cloning and new approaches that may augment or supersede current methods.

CONTENTS

[1] Equal contribution by first two authors.

[2] Corresponding author; e-mail: sive@wi.mit.edu.

0066-4154/97/0701-0751$08.00

INTRODUCTION

Subtractive cloning is a powerful technique that allows isolation of the differences in the nucleic acid composition of two cell samples (Figure 1). Differences can be at the level of RNA species represented within each sample or within the complement of genomic DNAs. Such differences include genes whose differential expression distinguishes one cell type from another, one growth phase from another, or a normal state from a diseased state. A related procedure, termed positive selection, has been used to isolate differences in cDNA and in genomic DNAs among various genotypes. In this review, we discuss the principle and origins of subtraction techniques that preceded the advent of cloning. We review current subtractive cloning strategies, pointing out advantages and disadvantages of each, and recount several examples of successful subtractive cloning and positive selection analyses. Finally, we discuss limitations of these techniques and prospects for the future.

The Basic Idea

Subtractive cloning uses a process called driver excess hybridization (see Figure 1). Nucleic acid from which one wants to isolate differentially expressed sequences (the tracer) is hybridized to complementary nucleic acid that is believed to lack sequences of interest (the driver). Driver nucleic acid is present at much higher concentration (at least 10-fold) than is tracer, and it dictates the speed of the reannealing reaction. The driver and tracer nucleic acid populations are allowed to hybridize, and only sequences common to the two populations can form hybrids. After hybridization, driver-tracer hybrids and unhybridized driver are removed. This is the subtraction step. The tracer that remains behind is enriched for sequences specific to the tracer tissue source [often called the plus (+) source] and depleted for sequences common to tracer and driver [often called the minus (−) source]. Usually, the process must be performed reiteratively in order to remove all the sequences common to both the driver and the tracer. After subtraction, remaining nucleic acid can be used to prepare a library enriched in tracer-specific clones or to make a probe that can be used to screen a library for tracer-specific clones.

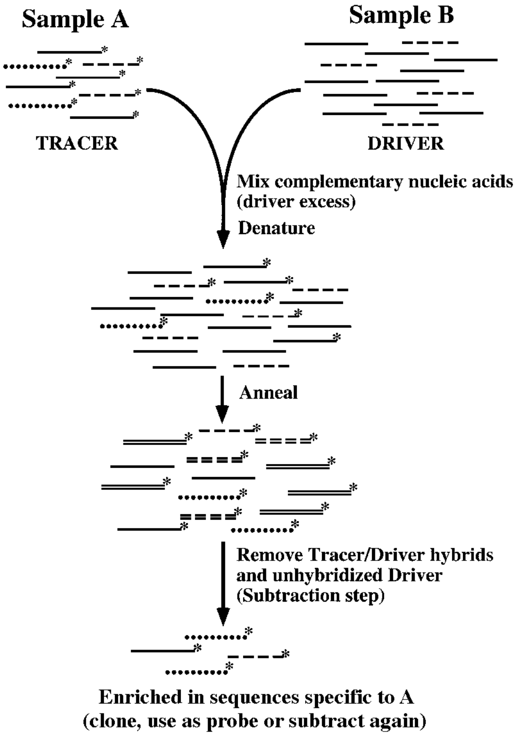

Sample A
Sample B
TRACER
DRIVER
Mix complementary nucleic acids
(driver excess)
Denature
Anneal
Remove Tracer/Driver hybrids
and unhybridized Driver
(Subtraction step)
Enriched in sequences specific to A
(clone, use as probe or subtract again)

Figure 1 General outline of subtractive hybridization. Complementary nucleic acids from two samples are mixed together (driver sequences are present in excess), denatured, and allowed to anneal. Duplexes formed between driver and tracer (asterisks indicate tracer) are then removed, as is unhybridized driver, leaving a population enriched for sequences present in the tracer but absent in the driver. Different sequences are indicated by solid, dashed, and dotted lines; dotted sequences are unique to the tracer.

THE PAST—ORIGINS AND APPLICATIONS OF REASSOCIATION ANALYSIS

Subtractive techniques preceded cloning techniques and have their origins in nucleic acid reassociation technology that was developed after the double-stranded nature of DNA was understood. This technology was not only a powerful tool for characterizing the RNA and DNA composition of cells but also helped develop the methodology required to perform successful subtractive cloning. Reassociation techniques led to an understanding of how many different genes are present in chromosomal DNA, what proportion of the genome is transcribed into mRNA, how many copies of mRNA sequences are present in a cell, and how the mRNA composition differs among cell types. Before methods were developed for analyzing genomic composition by large-scale DNA sequencing, our knowledge of genome organization and activity was dependent on nucleic acid–based hybridization (often termed Cot analysis; 1). Estimates derived from Cot analysis turn out to have been very accurate. In the next section we explore principles of renaturation kinetics and discuss the biologically important information these analyses have provided.

Origins

Watson and Crick's proposed structure for DNA (2) carried with it an inherent implication that the two strands of a DNA molecule could be separated. In the years immediately following Watson and Crick's work, other researchers investigated this possibility further (see e.g. 3, 4). In 1960 researchers demonstrated that separated strands could be reassociated, restoring both the physical and biological properties of native DNA (5, 6). The finding that DNA from different organisms could cross-hybridize (7) was used to analyze genetic relatedness between different species (see e.g. 8). In 1968, Britten & Kohne (1) and Wetmur & Davidson (9) introduced reassociation kinetics to analyze nucleic acid populations. This method was first used to provide estimates for the number of different sequences in the genome (1), but it was soon applied to the analysis of mRNA (see e.g. 10, 11).

Factors Affecting Reassociation Rate

Reassociation techniques measure how quickly complementary nucleic acids reanneal. The rate of reannealing of two complementary nucleic acid strands depends on their concentration—the higher their concentration, the more quickly the two strands will collide and hybridize. As a result, the kinetics of renaturation can be used to determine the number of different sequences and the abundance of each within a particular sample.

In a population containing many different sequences, the relative abundance of each sequence will affect the overall reassociation rate. For example, the overall reassociation rate for a sample containing two sets of sequences in a 1:1 ratio is slower than that of a sample containing two sequences in a 10:1 ratio (Figure 2*a*).

The overall reassociation rate of a sample also depends on the number of different sequences present, since the larger the number of sequences, the lower the concentration of each individual sequence (if total nucleic acid concentration is held constant) and the slower the reassociation rate (Figure 2*b*). Nucleic acid populations often are described in terms of their *sequence complexity*, a measure of the number of different sequences present in a population. Complexity is traditionally expressed as the total length of different sequences present. For

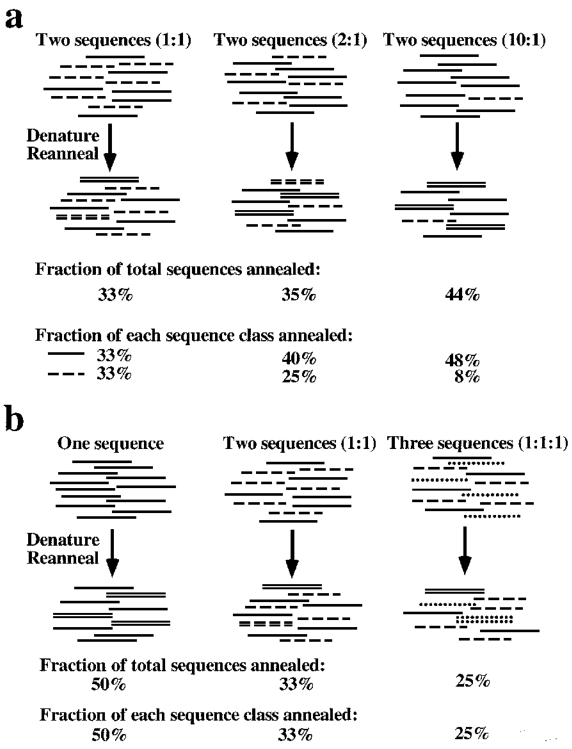


Figure 2 Factors affecting reassociation rate: (*a*) Effect of abundance on reassociation rate: When the overall hybridization of a mixture of two sequences present in a 1:1 ratio (*left*) is 33%, the sequences have annealed to the same extent (33%). When the sequences are present in a 2:1 ratio (*middle*), the overall extent of annealing is 35%, the more abundant sequence is annealed to 40%, and the less abundant to 25%. At a ratio of 10:1 (*right*), the overall extent of annealing is 44%, and the abundant and rare sequences are annealed to 48% and 8%, respectively. (*b*) Effect of complexity on reassociation rate: When a sample containing a single complementary sequence pair (*left*) has annealed to 50%, a sample containing two sequences in a 1:1 ratio (*middle*) will have annealed to 33%, and a sample containing three sequences in a 1:1:1 ratio (*right*) will have annealed to 25%. In this example, each sequence class is present at the same concentration within each sample, so that all sequence classes within each sample will reassociate at the same rate.

example, a cell expressing 14,000 unique mRNAs, each of 2000 nucleotides (nt) in length, would have an mRNA complexity of 2.8×10^7 nt.

The reassociation rate is also affected by the ionic concentration and the temperature of the reaction mixture (9). These parameters are usually standardized to 0.18 M Na^+ and 60°C, respectively. Conversion tables are available for converting rate constants obtained under experimental conditions to their values under standard conditions (12). Because the reassociation rate depends on the length of the nucleic acid fragments (9), reassociation experiments were often performed with nucleic acids fragmented to 300–500 nt to control for this effect.

Equations Describing Driver Excess Hybridizations

Information concerning complexity of nucleic acid populations as well as other useful data are derived from mathematical analysis of the renaturation kinetics, often called Cot or Rot analysis. In this review, we concentrate on the kinetics of driver-excess reactions, since these reactions most closely approximate the conditions of a subtractive cloning reaction.

When one strand of the reannealing population (the driver) is present in vast excess over the other strand (the tracer), the reannealing rate can be described by a simple equation. For this equation, we will assume that the driver is single-stranded RNA and the tracer single-stranded complementary DNA, since this is the experimental method that was used historically, and as we discuss later, RNA is still a good driver choice. However, similar considerations also apply for single-stranded DNA drivers. The change in tracer concentration (C) with time (t) can then be defined by the following equation:

$$\mathrm{d}C/\mathrm{d}t = -\mathrm{k}C_0R_0 \qquad 1.$$

where k is a rate constant that depends on sequence complexity, ionic concentration, temperature, and fragment length; C_0 is the starting single-stranded tracer concentration, R_0 is the starting single-stranded driver concentration, and R_0 is complementary to C_0.

The equation is negative because the concentration of the tracer decreases with time. Because the driver is present in such excess that its concentration (R_0) remains essentially unchanged through the course of the reaction, the equation does not need to consider changes in driver concentration.

With integration and rearrangement, one can derive the following equation from Equation 1:

$$C/C_0 = \mathrm{e}^{-\mathrm{k}R_0t} \qquad 2.$$

Equation 2 is useful because it can be used to plot C/C_0 against R_0t in order to obtain the reannealing rate of a specific nucleic acid population (see next section). This rate is independent of the absolute concentration of driver (as long as driver is in large excess over tracer), because the graph plots the extent of reassociation relative to the amount of driver, as well as relative to time.

Measuring Reannealing

Since reannealing assays measure the conversion of single-stranded nucleic acid molecules into a double-stranded form, sensitive assays were needed to distinguish between these forms. This section describes several assays that were used in the original reassociation analyses to exploit physical differences between single- and double-stranded molecules.

ABSORPTION OF UV LIGHT Nucleotides absorb ultraviolet light with an absorption maximum near 260 nm, owing to the heterocyclic rings of the bases. In single-stranded nucleic acids the bases are unordered; however, when two nucleic acid strands hybridize the stacking between their bases becomes ordered. This structural change leads to interference between rings, resulting in a decrease in absorption of about 40% per nucleotide. As hybridization proceeds, the transition from single- to double-stranded nucleic acid can be monitored by recording changes in UV absorption (9).

BINDING TO HYDROXYAPATITE Nucleic acids bind to hydroxyapatite (HAP), a crystalline form of calcium phosphate, through an interaction between the phosphate groups of the nucleic acid and the calcium ions of the HAP. HAP was first used for protein chromatography but was later found to separate single- from double-stranded nucleic acids (13, 14). A double-stranded nucleic acid molecule interacts with HAP more strongly than does a single-stranded molecule, presumably as a result of the greater steric availability of its phosphate groups. Therefore, when phosphate buffer is used to compete the bound nucleic acid away from the resin, single-stranded molecules elute at a lower phosphate concentration than do double-stranded molecules. This finding can be used to quantify how much nucleic acid has reassociated at a particular time. This procedure can also be used on a preparative scale to purify various components of a reassociation reaction.

NUCLEASE SENSITIVITY A nuclease that specifically degrades single-stranded nucleic acids (e.g. S1 nuclease) can be used to distinguish between single- and double-stranded molecules. This assay is often used in experiments in which one component is present in trace quantities (driver-excess reactions), since under these conditions single-stranded driver is present in such large quantities that annealed double-stranded molecules are difficult to detect either by absorption changes or by binding to HAP. By labeling the tracer and measuring the fraction of labeled material that becomes nuclease resistant with time, researchers can estimate the fraction of tracer annealed.

Useful Information Derived from Hybridization Kinetics

Reassociation assays can be employed to quantitatively address fundamental questions in biology. For instance, the utilization of such techniques has led to estimates for how many genes there are in the genome, for what fraction of all genes are expressed in a given cell type, and for how gene expression varies among different cell types. In this section, we give several examples of such experiments and discuss the results.

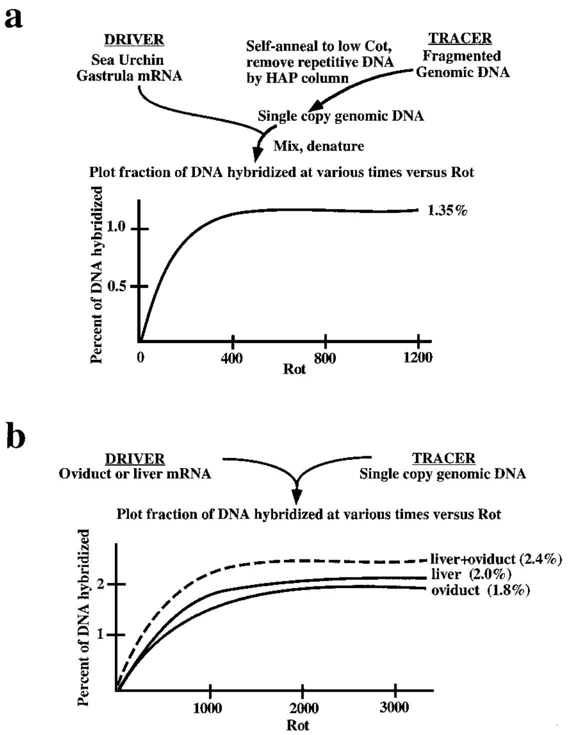


Figure 3 Information about gene expression derived from reassociation experiments. (*a*) Galau et al (20) performed saturation hybridization of sea urchin embryo single-copy genomic DNA tracer with excess gastrula mRNA driver and found that 1.35% of total single-copy genomic DNA hybridized to gastrula stage mRNA, indicating that 14,000 genes are expressed in the sea urchin gastrula. (*b*) Axel et al (17) used the same technique to perform additive saturation hybridization. Tracer from chick single-copy genomic DNA was hybridized to excess liver mRNA, oviduct mRNA, or a mixture of liver and oviduct mRNA. Eighty-three percent (2.0/2.4) of sequences are shared between these two tissues: The difference represents 2000–4000 distinct genes.

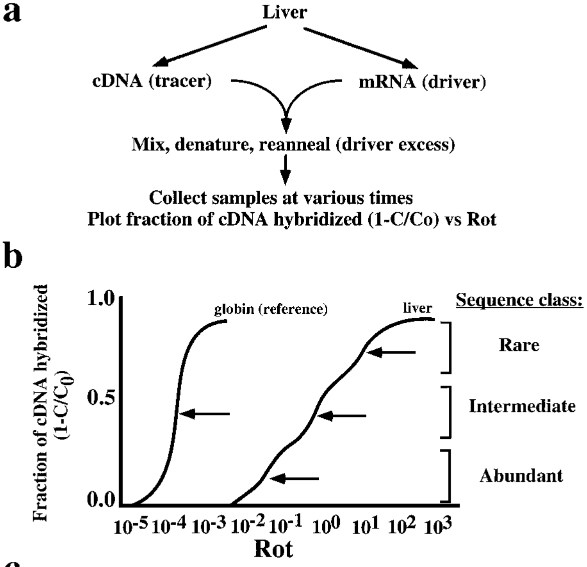
HOW MANY GENES ARE EXPRESSED IN A GIVEN CELL OR TISSUE? Several experimental methods indicate that between 10,000 and 30,000 genes are expressed in various mammalian cell lines (HeLa cells and L-cells; 15, 16) and organs (brain, liver, kidney, and the chick oviduct; 15, 17, 18), though some investigators have suggested a higher number of genes for the brain [closer to 100,000; (19)].

As an example of how these estimates were reached, Figure 3*a* describes an experiment in which Galau and coworkers determined the number of genes expressed in the sea urchin embryo (20). The investigators first determined the proportion of genomic DNA that is transcribed into mRNA in the sea urchin embryo. Given this information, and knowing the size of the genome as well as the average size of a gene, they then calculated the number of genes expressed in the embryo. The experiment involved the following procedure: First, labeled genomic DNA was fragmented and allowed to self-anneal to a low C_0t value, enabling noncoding repetitive DNA to hybridize. Hybridized repetitive DNA was then removed on hydroxyapatite, leaving single-copy genomic DNA corresponding to genes. This DNA served as tracer in a hybridization reaction with excess driver mRNA derived from sea urchin polysomes. Samples were taken during the annealing reaction, and the fraction of single-copy DNA that had hybridized was determined.

At saturation, about 1.35% of the single-copy genomic DNA had hybridized to sea urchin mRNA. Since transcription is likely to involve only one strand of the DNA, this corresponds to 2.7% of total single-copy double-stranded DNA. The size of the sea urchin genome is 8.12×10^8 bp, and 75% of this is single-copy DNA (21). Therefore, the amount of the genome transcribed is 1.6×10^7 bp ($0.027 \times 0.75 \times 8.12 \times 10^8$). Taking the average size of a sea urchin gene to be 1200 nt (22), the investigators estimated the number of different genes expressed in the sea urchin gastrula to be 14,000 ($1.6 \times 10^7/1200$).

HOW MANY GENES ARE DIFFERENTIALLY EXPRESSED BETWEEN TWO CELL TYPES? The question of how many genes are differentially expressed between two cell types was addressed in an experiment similar to that described in the previous section (see Figure 3*b*). Axel and coworkers compared the amount of single-copy chick genomic DNA to which mRNA populations from chick liver and oviduct could hybridize (17). If each organ expressed a distinct complement of genes, then hybridizing a mixture of the two mRNA populations to the genomic DNA should have led to an additive level of hybridization. Conversely, if the sets of genes expressed in the two organs were identical, no increase in the amount of hybridization to genomic DNA would be observed. The result was intermediate: The amount of hybridized genomic DNA increased (from 1.8% hybridized for oviduct and 2.0% for liver mRNA alone, to 2.4% hybridized for the mixed mRNAs). This result indicates that 83% (2.0/2.4) of sequences expressed in liver are also expressed in oviduct. If each tissue expresses 10,000–20,000 genes, this difference represents 2000–4000 distinct genes.

Variations on this type of experiment indicated that in sea urchin embryos 56% of transcripts from the earlier blastula stage were distinct from those found in the gastrula (23). Furthermore, analysis of differences between cell types found that fibroblasts differ from lymphocytes by expression of about 20% of genes (or 2000–4000 genes) (24), whereas two closely related cell types (B and T lymphocytes) differ by expression of only 2% of genes (or 200–400 genes) (25).



c

	Abundant	Intermediate	Rare	Total
Fraction of total sequences:	22%	41%	37%	100%
$Rot_{1/2}$:	0.039	1.64	30.8	
$Rot_{1/2}$(if pure):	0.0086	0.67	11.4	
# different transcripts:	10	840	14,200	15,000
Average # of copies of each transcript/per cell: (Assuming 500,000 transcripts per cell)	11,000	244	13	

Figure 4 (*a*) Hastie & Bishop (18) hybridized excess mRNA from mouse liver with complementary cDNA. They collected samples at various times and measured the extent of hybridization by sensitivity to single-stranded nuclease. (*b*) The concentration of cDNA that remains single stranded ($1-C/C_0$) is plotted against the product of starting concentration and time (R_0t). The reassociation of pure globin RNA with its complementary DNA is included as a reference. (*c*) The $R_0t_{1/2}$ of each abundance class can be read from the graph and used to calculate the number and abundance of transcripts in liver mRNA.

THREE DISTINCT mRNA ABUNDANCE CLASSES Figure 4*b* shows that the reassociation profile for mRNA driver hybridizing to cDNA tracer is tripartite (15–18). In the experiment outlined in Figure 4*a*, Hastie and coworkers (18) examined the composition of mouse liver mRNA. They used labeled first-strand cDNA as a tracer and excess mRNA as a driver. After denaturation and mixing, samples were collected at various times and the extent of hybridization was measured by sensitivity to S1 nuclease. As outlined in Equations 1 and 2, the concentration of cDNA that remained single stranded at time *t* was designated *C*. The fraction of initial cDNA concentration (C_0) that had hybridized at time *t* was then calculated ($1-C/C_0$). This value was plotted against the product of starting driver concentration (R_0) and time (t). The graph in Figure 4*b* has three transitions, indicating three abundance classes in the sample. Also, the curve is shifted to the right relative to the pure globin RNA standard, indicating a high sequence complexity in liver cDNA.

The graph in Figure 4*b* can be used to calculate the total number and abundance of transcripts in the liver mRNA. The $R_0t_{1/2}$ is defined as the R_0t value at which half the nucleic acid in a sample has renatured. The $R_0t_{1/2}$ of each abundance class can be read from the graph (see arrows in Figure 4*b*). This value must be corrected to account for the dilution of the abundance class by the other sequence classes in order to obtain the $R_0t_{1/2}$ (pure). For example, the $R_0t_{1/2}$ of the abundant class is 0.039, but the abundant class makes up only 22% of the sample, so the $R_0t_{1/2}$ (pure) is 0.0086 (0.039×0.22). The number of different transcripts in each class can then be obtained by comparing the $R_0t_{1/2}$ (pure) to the $R_0t_{1/2}$ of a single transcript (globin, $R_0t_{1/2} = 0.0008$). The abundant class contains about 10 different transcripts (0.0086/0.0008). If each cell is taken to contain about 500,000 transcripts (18), the number of copies of each transcript per cell can be estimated: 22% of the 500,000 transcripts would be represented by 10 abundant transcripts, so that there are 11,000 copies [$(500{,}000 \times 0.22)/10$] of each transcript in the abundant class. Similar calculations can be performed for the other abundance classes (Figure 4*c*).

To determine the abundance classes in which differentially expressed transcripts can be found, investigators first purified each of the three abundance classes from kidney cDNA using HAP columns. Each class was then used as the tracer in hybridization reactions to excess liver or brain mRNA driver. Differentially expressed transcripts were found in all abundance classes, although some abundant sequences from kidney were absent from brain and liver, and others were detected at much lower levels. Analysis of the intermediate and rare sequences from kidney showed that about 10% of the sequences in these classes were absent in liver and brain. This observation indicated that differences between tissues are due both to quantitative changes (altered expression levels) and to qualitative changes (absence or presence of particular transcripts) in gene expression (18).

SUMMARY Collectively, these results showed that a given cell or tissue expresses about 20,000 distinct genes in three abundance classes. Although organs differ by expression of about 20%, or about 4,000 genes, two closely related cell types, B and T lymphocytes, differ by expression of only 2%, or about 400 genes (25). Differentially expressed genes may reside in any of the three abundance classes. Note that the examples given analyzed the composition of differentiated cells and tissues. Comparisons between tissues during their development may reveal smaller differences, particularly if they share a common precursor. The expression of about 20,000 genes in a given mammalian cell (mouse or human; 15, 16), as indicated by reassociation experiments, has led to the estimate of 50,000–100,000 genes in the human genome (26, 27). This number is in remarkable agreement with recent estimates of 71,000 human genes, based on direct sequencing of portions of chromosomes followed by extrapolation to the whole genome (27). Similarly, estimates based on the frequency of CpG islands in the human genome have suggested a total of 80,000 genes (28).

Relevance of Renaturation Studies for Subtractive Cloning

Reassociation kinetic studies provide several useful lessons for subtractive cloning. First, one needs to think about the differences between the two tissues to be compared. The greater the number of differences, the more time it will take to sort through the cloned genes. In general, it is wise to choose cell types or populations that are as similar as possible but that still display the differences one wants to define in terms of specific genes.

Second, if the nucleic acid populations to be compared are complex, as in whole organs or whole embryos, there are likely to be several differences to clone, and the large number of different sequences will make clones that are rare on a per cell basis even rarer in the context of a large number of different cell types. As a result, the rarer the sequence, the more difficult it is to remove it by hybridization from a nucleic acid population, since higher Rot values must be obtained. Therefore, it is useful to compare tissues that are as low in complexity as possible. When very complex nucleic acid populations must be compared, multiple rounds of subtraction need to be performed.

Third, it is useful to know the abundance class into which the differences between nucleic acid populations of interest fall. This information indicates how complete the subtraction needs to be in order for these differences to be isolated. In the absence of such information, it is worth assuming that one is trying to clone rare or moderately rare differences. These considerations require that one knows a fair amount about the characteristics of the nucleic acid population being studied. When beginning a subtractive cloning experiment, one either needs to extrapolate from known systems, as is generally done, or to perform Rot measurements oneself.

THE PRESENT—USEFUL TECHNIQUES

We now discuss subtractive cloning methods, including the basic steps in a subtractive cloning scheme and how to choose a particular subtraction protocol. Protocol determination includes selection of tissue sources and preparation of tracer and driver nucleic acids. We also examine parameters important for successful hybridization and methods for hybrid removal. Then, we review positive selection schemes and discuss several prototypical schemes that work efficiently. Finally, we discuss situations in which schemes other than subtraction might be equally or more useful.

Strategies for Subtraction

The original subtraction cloning methods used first-strand cDNA as the tracer, polyA+mRNA as the driver, and HAP to remove hybrids. This method is still useful if one is using single cell types and is able to obtain a large quantity of starting materials. However, when materials are limiting, the available tissue must be converted into a form that can be amplified before subtraction.

When complex tissues are used, multiple rounds of subtraction must be performed to remove rare common sequences more completely. Reiterative subtraction requires that the tracer be regenerated or amplified after subtraction [by the polymerase chain reaction (PCR), by amplification of a cDNA library, or by in vitro transcription]. Some of these reiterative schemes allow the driver to be enriched for rare common sequences, so that it will be better able to remove these sequences from the tracer in subsequent rounds of subtraction (29). A disadvantage of any reiterative subtraction procedure is that biases in the relative representation of clones can occur during amplification. Nonetheless, reiterative subtraction has allowed differentially represented clones to be isolated from extremely complex tissues (29, 30; M Patel, J Kuo, V Apekin & HL Sive, unpublished information; CG Sagerström, HL Sive, unpublished information).

Tracer and Driver Preparation

Both RNA and DNA can serve as either tracer or driver; however, RNA generally makes a poor tracer since it is easily degraded. Conversely, RNA makes a good driver since driver molecules not removed during the hybrid removal step can easily be degraded enzymatically or by using alkali.

poly(A)+ DRIVER AND cDNA TRACER cDNA is prepared by oligo(dT) or random priming (Figure 5). The problem with this method is that large amounts of starting material must be obtained, and only two rounds of subtraction can be performed before the amount of remaining tracer becomes too small. With complex starting tissues, subtraction is never complete at this point.

AMPLIFICATION METHODS cDNA libraries can provide an excellent source of full-length clones, and currently this is the best method to isolate full-length clones after subtraction (31–34). Full-length cDNA libraries can be subtracted as single-stranded phagemids or as RNA derived from inserts in the library by in

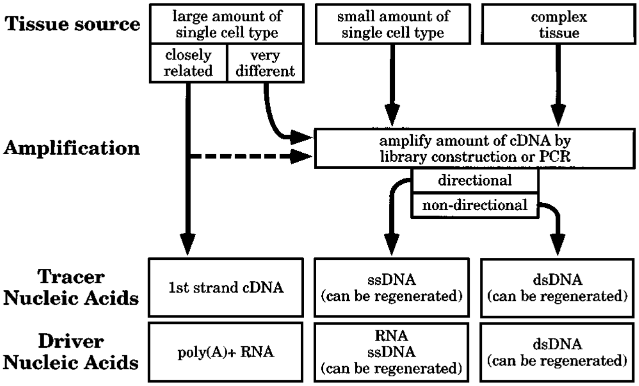

Tissue source
large amount of single cell type
closely related
very different
small amount of single cell type
complex tissue
Amplification
amplify amount of cDNA by library construction or PCR
directional
non-directional
Tracer Nucleic Acids
1st strand cDNA
ssDNA (can be regenerated)
dsDNA (can be regenerated)
Driver Nucleic Acids
poly(A)+ RNA
RNA ssDNA (can be regenerated)
dsDNA (can be regenerated)

Figure 5 General outline of tracer and driver preparation, showing the sources of driver and tracer nucleic acid, an indication of whether they should be amplified, and the forms of tracer and driver nucleic acids for hybridization. Broken arrow indicates an option to amplify the nucleic acids.

vitro transcription (Figure 5). To preserve full-length clones after subtraction, some selection for the original ends of the cDNA must be performed. For example, one can select for restriction enzyme sites engineered at the 5′ and 3′ ends of the cDNA during initial library construction, or for intact phagemids that can transform bacteria after subtraction. It is difficult to perform reiterative subtraction with this method, however, since the library must be regrown before each step.

Another disadvantage of this method relates to the hybridization rates of different length cDNAs. As discussed previously, the reassociation rate is inversely proportional to the length of the nucleic acid strands analyzed. Since full-length cDNAs vary widely in size, from a few hundred to tens of thousands of nucleotides, the reassociation kinetics of different length cDNAs will vary, suggesting that successful hybridization and subtraction require higher C_0t values than are necessary for small fragments of cDNA.

An alternative method uses PCR to amplify cDNA. This method allows one to work with tiny amounts of starting material and to perform multiple rounds of subtraction easily (29). A major drawback of this procedure, however, is that small pieces of cDNA are the end product. Small clones are isolated because DNA is fragmented, or synthesized initially as short fragments, to prevent introducing bias during PCR where small fragments are preferentially amplified (35). As a result, full-length clones must be isolated after subtraction, which prevents, for example, direct functional testing of the subtracted products.

The Hybridization Step

An important parameter controlling the success of the hybridization is the tracer:driver ratio, which should be at least 1:10 to allow the driver to govern the subtraction. Other factors include the absolute concentration of driver and the time allowed for hybridization, both of which should be as large as practical because both factors affect the Rot or Cot (for DNA driver) that is achieved and therefore determine the extent of hybridization. In general, a Rot of at least 1000 should be obtained. The hybridization time is limited by the degradation of driver and tracer during hybridization, particularly when the driver is RNA.

It is also important to consider whether the driver and tracer will be single or double stranded. Single-stranded driver is the most efficient choice, since the concentration of driver decreases only slightly as driver hybridizes to tracer, which allows high Rot or Cot values to be reached. In contrast, when double-stranded driver is used, driver-driver and driver-tracer duplexes form during hybridization. Driver-driver duplex formation competes with the desired reaction, and over time, decreases the concentration of driver available, reducing the efficiency of subtraction. As a result, more rounds of subtraction must be performed with a double-stranded driver than with a single-stranded driver to obtain equivalent subtraction. Despite these disadvantages, subtractions can be performed effectively with double-stranded driver (29, 30). This method allows the driver to be made by exponential PCR and is a good way to use small amounts of starting material.

The Subtraction Step: Removing Driver-Tracer Hybrids and Excess Driver

Several methods exist for removing driver-tracer hybrids and excess driver. We describe here the principles of various methods, including their variations, advantages, and disadvantages (see Table 1 for a summary).

HYDROXYAPATITE HAP binds double-stranded nucleic acid more tightly than it does single-stranded molecules, enabling separation of driver-tracer and driver-driver duplexes from unhybridized tracer. The advantages of this method are that it is proven and the separation is efficient. One disadvantage is that this method is cumbersome since HAP columns need to be run at high temperature (65°C) in water-jacketed columns. Another disadvantage is that unhybridized driver molecules cannot be removed. This is not a problem if RNA is used as the driver, since the RNA can later be degraded by RNase or alkali treatment; however, this method cannot be used with a DNA driver.

Table 1 Methods for subtractive enrichment

Method	Principle	Variations	Advantages	Disadvantages	References
Hydroxyapatite (HAP)	Differential adsorption of ssNA and dsNA to HAP ($Ca_{10}(PO_4)_6(OH)_2$) at different temperature and phosphate ion concentration	Column or batch adsorption	Efficient	Running column is cumbersome (needs high-temperature water jacket); batch adsorption is not as efficient as column	12, 55, 85, 86
Biotinylation and streptavidin	Biotinylated driver nucleic acids can be removed after streptavidin (or avidin) treatment	Biotinylation:			
		1. Photo-biotinylation	Simple to prepare biotinylated driver	Low biotin density; hydrophobic reagents and products often insoluble	36, 41, 43, 72
		2. Biotinylated-nucleotide incorporation	Simple to prepare biotinylated driver; high biotin density	Could over-biotinylate; relatively expensive	30, 37, 38, 87, 88
		Driver-hybrid removal:			
		1. Cupric-iminodiacetic acid agarose beads	Efficient	No prominent disadvantage	42, 43
		2. Streptavidin binding followed by phenol extraction	Very simple	No prominent disadvantage	41, 72
		3. Streptavidin-(or avidin)-coupled beads	Easy to use; driver recycling	No prominent disadvantage	31, 38, 40, 44, 45

Chemical cross-linking	2,5 diaziridinyl-1,4-benzoquinone cross-links GC hybrids, rendering them inaccessible to polymerases		Easy for subtracted probe preparation	Largely unproven	46, 76
Driver immobilization	Immobilized driver hybridizes to and traps common tracer sequences	DNA-cellulose, oligo d(T)-coupled cellulose, latex, or magnetic beads; nitrocellulose membrane	Driver recycling	Kinetically unfavorable to immobilize driver on solid surface; largely unproven	47–51
Restriction enzyme digestion	Hybrids of tracer and driver DNA cleaved by frequent cutting restriction enzymes			Largely unproven	53
RNase H	RNA tracer that hybridizes to ssDNA driver is digested by RNase H			RNA tracer easily degraded; largely unproven	54

BIOTINYLATION AND STREPTAVIDIN Biotinylated driver nucleic acid has been used efficiently to remove driver-tracer hybrids and unhybridized driver molecules. The driver can be either RNA or DNA. Biotinylated driver-tracer hybrids and unhybridized driver can be removed by exploiting affinity of biotin for the proteins avidin or streptavidin (see below). Several methods are available for generating biotinylated nucleic acid. For example, RNA can be photobiotinylated (36), where photobiotin acetate and nucleic acid are irradiated with a sun lamp (absorbance between 261 and 473 nm causes photolysis). Another method involves incorporation of biotinylated nucleotides during driver synthesis using thermostable DNA polymerases during PCR (37), RNA polymerases (38), or Klenow fragment (39). Finally, biotinylated primers can be used during a PCR reaction (40).

The above-mentioned methods are simple; however, photobiotinylation has two drawbacks: First, the biotin density on the nucleic acid is low (about one per three hundred nucleotides), potentially resulting in submaximal driver removal. Second, products of photobiotinylation are often insoluble in aqueous solutions because of the hydrophobicity of the side arm to which the photoreactive biotin is attached (41). Incorporation of biotinylated nucleotides overcomes these disadvantages. One potential problem of biotinylation by incorporation is that too much biotin may interfere with hybrid formation; however, titration for the optimal biotin density can be performed (30).

After the hybridization reaction, biotinylated driver-tracer hybrids and excess driver are removed by exploiting the high-affinity biotin-binding protein streptavidin (Kd $= 10^{-12}$). Originally this step used cupric-iminodiacetic acid agarose beads (42, 43) that fractionate protein-bound nucleic acid (i.e. the streptavidin-biotinylated nucleic acid complex) from protein-free nucleic acids. Thus, single-stranded biotinylated driver and driver-containing duplexes were separated from unhybridized tracer. In a more commonly used method, biotinylated nucleic acid can also be removed after streptavidin treatment by phenol extraction (41). The streptavidin–nucleic acid complexes partition to the aqueous-organic interface during phenol extraction, while unhybridized tracer remains in the aqueous phase. Biotinylated driver can also be removed by reaction with streptavidin-coupled beads. For example, vectrex-avidin (38), avidin-sephacryl S-1000 (31), streptavidin-magnetic beads (44), and streptavidin-agarose (45) are easy to use. This method also allows the driver-containing beads to be recovered, thus enabling the driver to be recycled.

CHEMICAL CROSS-LINKING After hybridization, the two strands of driver-tracer duplexes are cross-linked by 2,5 diaziridinyl-1,4-benzoquinone while unhybridized tracer remains unaffected (46). Subtracted probes can then be prepared from single-stranded tracer while cross-linked driver-tracer (or driver-

driver) hybrids will not be accessible to polymerases. The advantage of this method is the ease of subtracted probe synthesis, which can take place immediately after the cross-linking reaction, without physical separation of driver-containing hybrids from tracer.

IMMOBILIZATION METHODS Tracer sequences can be hybridized to driver immobilized on solid phase. The unbound tracer is enriched for sequences not represented in the driver. Variations include immobilizing the driver on cellulose (47), oligo(dT)-cellulose (48), oligo(dT)-latex (49), Dynabeads oligo(dT) (50), or on a nitrocellulose membrane (51). These methods are useful because driver can be recycled. The disadvantage is that the kinetics of hybridization to solid phase are unfavorable relative to solution hybridization (52).

ENZYMATIC HYBRID REMOVAL Specific digestion of driver-tracer hybrids by restriction enzymes has been used. In the single reported use of this method, both tracer and driver are single stranded: Tracer is prepared from a library as single-stranded phagemids, and driver is first-strand cDNA (53). After hybridization, the driver-containing hybrids are digested with restriction endonucleases, and the remaining DNA is introduced into bacteria. The phagemid tracer is capable of transforming bacteria, but digested driver-tracer hybrids and excess driver are not. Because this method relies on the ability of phagemid DNA to transform bacteria, it cannot be adapted easily to other forms of tracer.

In another method, RNaseH is used to remove hybrids (54). Labeled RNA tracer is hybridized with single-stranded DNA driver, and the reaction mix is treated first with RNaseH to remove hybridized tracer and subsequently with DNaseI to remove excess driver. The remaining RNA can be used directly as subtracted probe. This method is simple, but the use of RNA as the tracer presents a problem because it is labile and may be degraded significantly during the hybridization reaction. Another disadvantage is that the subtracted products cannot be cloned directly.

Positive Selection

In contrast to subtractive cloning, which actively removes unwanted nucleic acid, positive selection actively isolates the desired nucleic acid and leaves behind the undesired nucleic acid (Figure 6). Most positive selection methods use double-stranded nucleic acid for both tracer and driver, and specifically isolate tracer-tracer hybrids, passively removing tracer-driver and driver-driver hybrids as well as unhybridized tracer and driver (Table 2). Thus positive selection is a dual subtraction–active selection method. The earliest positive selection methods subtracted common sequences first and then in a separate step selected for tracer-tracer hybrids (55, 56). More recently discovered methods are more efficient and simultaneously select for tracer-tracer hybrids

Table 2 Methods for positive selection

Method	Principle	Variations	Advantages	Disadvantages	References
Cohesive restriction sites	Only tracer-tracer hybrids regenerate clonable cohesive ends	In-gel reassociation	Great enrichment	Rare sequences may not anneal	58, 60, 63, 64, 66
Specific primer binding sites	Select for tracer-tracer hybrids that can be amplified exponentially by PCR	Mung bean nuclease treatment to remove ssDNA before PCR	Small amount of starting material needed; great enrichment	Rare sequences may not anneal	67, 68, 89–91
		Suppression PCR enhances normalization		Largely unproven; rare sequences may not anneal	69
Protection from enzymatic degradation	Exonuclease digestion of dsDNA except for tracer-tracer hybrids, which have thio- nucleotide incorporated at their ends			Largely unproven; rare sequences may not anneal	70

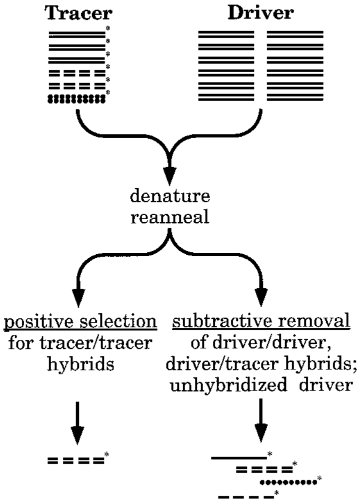

Tracer
Driver
denature
reanneal
positive selection for tracer/tracer hybrids
subtractive removal of driver/driver, driver/tracer hybrids; unhybridized driver

Figure 6 Illustration of the two different outcomes of positive selection or subtractive enrichment. Solid lines represent common sequences, dashed and dotted lines represent tracer-specific sequences, and asterisks indicate sequences of tracer origin. Note that positive selection may not isolate rare clones (*dotted line*, see text), whereas subtraction leads to isolation of all differentially represented clones (*dotted* and *dashed lines*) but may retain some common sequences (*solid line*).

while removing common sequences (see Table 2 for a summary of these techniques).

Positive selection can be more efficient than subtraction for isolating differentially represented genes, since subtraction rarely goes to completion and a background of common, unsubtracted clones may persist. However, a major disadvantage of positive selection is that since the desired tracer-tracer hybrids take longer to form than do driver-tracer hybrids (because tracer concentration is so low), positive selection is likely to miss rare clones that have not reannealed. One solution to this problem is the phenol-enhanced reassociation technique (PERT; 57–59), which increases the rate of hybridization by decreasing the aqueous volume. Polyethylene glycol (PEG8000) serves the same purpose (60). Another way to make positive selection more effective is to use a tracer population whose sequences are all present at similar abundance. This is the case with genomic DNA or with cDNAs that have been normalized or self-subtracted to low C_0t to decrease the representation of abundant sequences and increase the representation of rare sequences (61, 62).

COHESIVE RESTRICTION SITES This method selects for the ability of tracer-tracer hybrids to be cloned because both ends of the hybrids contain a particular restriction site (58, 63, 64; see Figure 7*D*). Driver-tracer and driver-driver hybrids do not have the correct ends to be cloned. Tracer is prepared by restriction endonuclease digestion that generates cohesive or “sticky” ends, and driver is prepared by sonication, which leaves “ragged” ends. After hybridization the entire nucleic acid mix is combined with DNA ligase and vector with ends compatible to those of the tracer-tracer hybrids, which will be the only molecules cloned efficiently.

In a variation of this method, called in-gel competitive reassociation (60, 65, 66), the hybridization mix is run on an agarose gel before denaturation and reannealing are carried out within the gel. Running the hybridization mix on a gel facilitates isolation of restriction fragment length polymorphisms (RFLPs). In this case, both tracer and driver are digested with the same restriction endonuclease, and the driver is dephosphorylated after digestion.

SPECIFIC PRIMER BINDING SITES This positive selection technique uses selective amplification of tracer-tracer hybrids by PCR. In a variation of the compatible restriction site method, specific adapters are selectively ligated to tracer-tracer ends, followed by PCR (67). Only tracer-tracer hybrids have adapters at both ends of the duplex DNA and are amplified exponentially by PCR. Tracer-driver hybrids are amplified with linear kinetics, whereas driver-driver duplex and single-stranded tracer are not amplified at all.

In another method (68, 69), tracer DNA fragments are ligated to adapters that have both strands dephosphorylated and, therefore, become covalently

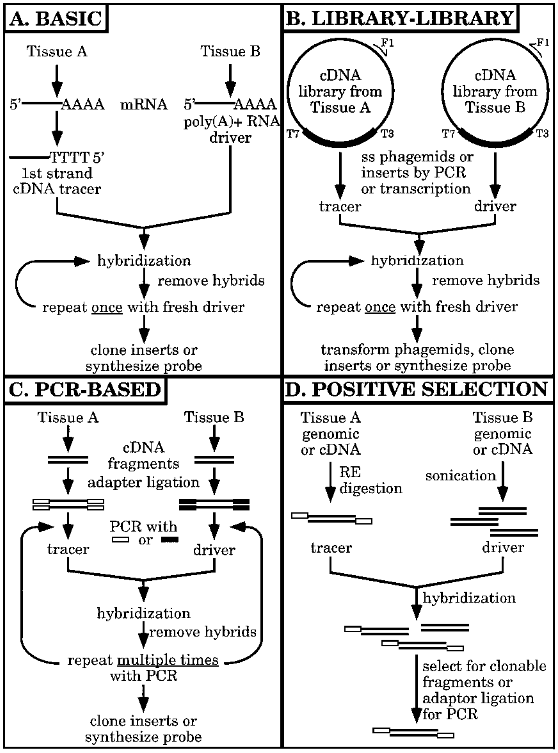


Figure 7 Comparison of subtractive enrichment and positive selection. (*A*) Basic scheme: First-strand cDNA serves as tracer and mRNA as driver. A large amount of tissue is necessary for this scheme, and subtractive hybridization can be performed only twice. (*B*) Library-library subtraction: Solid bars represent cDNA inserts. In this scheme, the final subtracted products represent full-length cDNA. (*C*) PCR-based subtraction: Open and solid boxes represent different adapters. This scheme allows multiple rounds of subtraction to be performed easily. (*D*) Positive selection based on cohesive restriction sites method: Open boxes represent regenerated cohesive ends, which can be ligated to vector or adapter.

attached only to the two 5′-ends of double-stranded tracer molecules. After hybridization, the ends of DNA molecules are filled in and subjected to PCR, and the adapter again serves as primer. Again, only tracer-tracer hybrids are amplified exponentially.

PROTECTION FROM ENZYMATIC DEGRADATION This method protects tracer-tracer hybrids from degradation by exonuclease (70). Double-stranded tracer DNA is treated with Klenow fragment of DNA polymerase I before hybridization to replace the nucleotides at both 3′-ends with α-S-nucleotides. After hybridization the reaction mix is treated with exonuclease III, which degrades unmodified DNA molecules but leaves intact those containing phosphorothioate-nucleotides (i.e. tracer-tracer hybrids). The reaction mix is then treated with exonuclease VII to digest any single-stranded molecules before cloning.

Monitoring Subtraction and Isolating Differentially Represented Genes

We now discuss how to evaluate whether a subtraction cloning experiment was successful, and how to isolate differentially represented clones using the subtracted pool of nucleic acid.

ANALYZING ENRICHMENT AFTER SUBTRACTION The amount of tracer should decline with successive rounds of subtraction, and radioactively labeled tracer can be used to determine whether this decrease has occurred. If one is using a protocol that allows successive rounds of subtraction, the counts removed should plateau as the subtraction nears completion. This type of monitoring does not indicate what genes are represented in the enriched population; therefore, it is useful to assay for some known genes that should be present at equal levels in starting tracer and driver populations, as well as genes expressed only in the tracer that is of the abundance class one would like to isolate. The remaining tracer population can then be monitored for representation of the common genes, which should decrease, and for representation of the tracer-specific genes, which should increase. The more genes that behave as expected, the more likely the subtraction has been successful.

A more quantitative method used to evaluate subtraction efficiency is the enrichment (i.e. the increase in representation) of a tracer-specific clone per unit number of total clones, or per mass of total nucleic acid. For example, an increase in representation of a clone from 1:10,000 in the unsubtracted pool to 1:100 in the subtracted pool indicates a 100-fold enrichment. Often, a single number is used to summarize the enrichment of a subtraction process. This number may reflect the overall enrichment of clones of a particular abundance class being studied.

There are several ways to assess the enrichment of a subtraction process, including comparing signal strengths of a single probe to a Southern blot of pre- and postsubtraction nucleic acid, hybridization of single probes to libraries made from pre- and postsubtraction nucleic acid, and dot blot analysis of single probes to pre- and postsubtraction nucleic acid. Table 3 compiles data from some experiments that estimated enrichment after subtraction or positive selection.

Estimated enrichment varies widely, possibly as a result of assaying clones from different abundance classes or using different assay methods. Enrichment also depends on complexity: The lower the complexity and the fewer the differences between two nucleic acid populations, the greater the enrichment

Table 3 Enrichment reported for various subtraction processes

Method for subtraction	Rounds of subtraction	Estimated enrichment	Method for estimation	Reference
Subtractive enrichment				
HAP	1	20	Tracer recovery	92
	3	60	Tracer recovery	93
	3	100–700	Southern blot	39
	3	450	Southern blot	94
Biotinylation	1	10	Not reported	67
	2	50	Colony hybridization	72
	2	50	Colony hybridization	75
	2	100	cfu pre- and post-subtraction	31
	multiple	2000	Southern blot	29
	1	5000	Transformation assay	34
Chemical cross-linking	1	100	Not reported	76
	1	240–300	Southern blot	46
Driver immobilization	1	275	Dot blot	51
	4	300	Dot blot	49
Positive selection				
Restriction enzyme site	2	3000	Southern	60
Selective PCR	1	1000–5000	PAGE of labeled tracer	69
	2	4×10^5	Southern	68

will be per cycle of subtraction, particularly if the differentially represented clones are abundant. A conservative expectation for any subtraction process is an enrichment of 50- to 100-fold for the first one or two cycles of subtraction. In general, unsuccessful subtractions result from insufficient subtraction, leading to an inability to find differentially represented genes among the remaining common clones.

ISOLATING DIFFERENTIALLY REPRESENTED CLONES FROM A SUBTRACTED POOL

There are several ways to isolate differentially represented clones from a pool of subtracted nucleic acid. For example, a subtracted library can be constructed and clones picked randomly. This method can be successful if the subtraction has been very effective, but in general, additional levels of screening are necessary.

Another option is to probe duplicate lifts of a subtracted or unsubtracted library with a subtracted probe versus an unsubtracted probe. However, since subtraction enriches for all rare clones, clones that hybridize preferentially to the subtracted over the unsubtracted probe may correspond to rare, but commonly represented, clones that just become more abundant after subtraction.

A better option is to screen with two subtracted probes, one from the tracer minus driver and another from the driver minus driver (71) [or driver minus tracer (72)]. Both probes will detect rare clones that are enriched during the subtraction, but the former (tracer minus driver) will also detect clones that are specific to the tracer. This approach distinguishes between differentially represented clones and those that are just rare.

Choosing a Subtraction Method

The decision to use a particular method should be based on the amount of starting materials that one can obtain, the complexity of those materials, and the goal of cloning (for example, whether one wants to make a subtracted probe or a subtracted library, or whether one wants full-length clones or would be content with partial fragments).

Figure 7 illustrates some useful cloning schemes and summarizes distinct features of each. The basic scheme (see Figure 7*A*) is perhaps the most straightforward when one is trying to compare similar tissues and it is easy to obtain large amounts of such tissues. When the starting tissue is difficult to obtain or when complex tissue is to be compared, a library-library or PCR-based scheme must be adopted. Library-library subtraction (Figure 7*B*) is the easiest way to isolate full-length clones, but it is difficult to perform this method reiteratively.

PCR-based subtraction (Figure 7*C*) is the scheme that most easily allows multiple rounds of subtraction, but representation of tracer population may be biased because of multiple rounds of PCR required and because the average size of tracer nucleic acids is small. The scheme shown for positive selection (Figure 7*D*) is based on cohesive restriction sites. As discussed previously, positive selection is so sensitive that one would be likely to isolate some clones of interest but unlikely to obtain a full spectrum of clones.

Success Stories

We have compiled lists of representative success stories for which different methods were used for subtractive enrichment (Table 4) and positive selection (Table 5). In Table 4, we categorize the examples into three groups: (*a*) those in which subtracted libraries were constructed and clones were picked and examined randomly, (*b*) those in which subtracted probes were prepared to screen unsubtracted libraries, and (*c*) those in which subtracted probes were used to screen subtracted libraries. (In both cases presented under the second category, differential screening was performed using the subtracted probe compared to an unsubtracted probe or probe from another source.) In cases for which multiple techniques were used to isolate the differentially represented clones, the primary technique used is listed.

Table 4 Representative examples of subtractive cloning

Result	Method used and remarks	References
Subtracted libraries, randomly picked clones		
Subtracted library enriched for gastrula-specific clones of *Xenopus laevis* embryos	Subtracted library constructed using HAP	95
Subtracted library enriched for scrapie-modulated clones	Biotinylated driver removal by affinity resin (This is the early library-library subtraction.)	31
Subtracted libraries enriched for up- or down-regulated genes in *X. laevis* tail after thyroid hormone treatment	Photobiotinylated driver removal by phenol extraction	29
Isolation of F-spondin	Differential screening of a subtracted library constructed using biotinylated driver	75
Construction of a mouse subtracted library enriched for two-cell stage transcripts	Biotinylated RNA driver removal by phenol	96
Construction of a mouse endoderm-minus-mesoderm subtracted library	Phagemid tracer–biotinylated RNA driver removal by phenol	97
Subtracted probes on unsubtracted libraries		
Isolation of myoblast-specific clones	Differential screening using subtracted (HAP) versus unsubtracted probes	56
Isolation of Waf-1	Differential screening using subtracted (chemical cross-linking) versus unsubtracted probes	76
Subtracted probes on subtracted libraries		
Isolation of a helper T cell receptor clone	Both subtractions used HAP	92
Isolation of three cytotoxic T cell receptor clones	Differential screening of a subtracted library constructed using HAP	74, 98
Isolation of growth-arrest-specific clones	Both subtractions used HAP	71
Isolation of *X. laevis* cement gland marker clones	Biotinylated driver removal by phenol	72
Isolation of early neural markers as well as other novel clones	dsDNA tracer/biotinylated driver hybridization	M Patel, HL Sive, unpublished information

Comparing Subtraction, Positive Selection, and Other Methods for Isolating Differentially Represented Genes

Subtractive cloning and positive selection are not the only methods available for isolating differentially represented sequences. Table 6 presents a comparison of subtractive cloning to other methods. In this table, *random sampling* refers to selection of clones for analysis from an unsubtracted library on a random basis. In *differential display* (73), PCR is performed on first-strand cDNA with an arbitrary 5′-primer and a 3′-anchor primer consisting of oligo(dT) with two fixed 3′-bases. The amplified products are resolved on a sequencing gel, and differences can be identified between patterns of fragments from the tissues being compared. *Differential screening* uses unsubtracted probes from different tissue sources to screen duplicate filters of an unsubtracted library.

Table 5 Representative examples of positive selection

Result	Method used and remarks	References
Subtracted library enriched for mouse Y-specific DNA sequences	Restriction enzyme method for positive selection	63
Cloning of DNA fragments from the Duchenne muscular dystrophy loci on the X chromosome	Restriction enzyme method for positive selection with phenol-enhanced reassociation	58
Isolation of 20 human restriction fragment length polymorphisms (RFLPs)	Selective PCR for positive selection	68
Subtracted library that contains RAG-1, RAG-2 and other novel clones	Selective PCR for positive selection on cDNA	89
Subtracted library enriched for mouse RFLPs	Combination of subtractive hybridization, restriction enzyme method, and selective PCR	90
Isolation of DAZ (deleted in Azoospermia) by YAC subtraction	Combination of subtractive hybridization, restriction enzyme method, and selective PCR	99

In Table 6, the detection limit is expressed in the form of abundance percentage of the rarest class of genes that can be detected by a given method. As an example, random sampling would allow one to isolate rare genes represented at as low as 0.001% or rarer because rare genes comprise a significant portion of a well-represented library. The likelihood of obtaining differentially expressed genes of particular interest by random sampling, however, is extremely low, unless a large number of clones is examined.

Subtractive cloning is not always the best method for a particular application. For example, differential display may be faster than subtraction for obtaining new markers for a particular tissue, but it will not yield a complete spectrum of differentially expressed genes. A combination of methods is often useful. For example, differential screening is routinely performed in combination with subtraction (56, 72, 74–76).

THE FUTURE—EASIER TECHNIQUES?

We come now to the question of what subtractive cloning cannot do at present, and how this technique may change in the future in order to overcome current limitations.

Problems to Overcome

The two major problems with current subtractive or positive selection techniques are (*a*) an inability to easily isolate full-length clones after subtraction and

Table 6 Comparison of various techniques for gene isolation

Technique	Detection limit (%)	Advantages	Disadvantages	References
Random sampling	<0.001	Simple; clones obtained are full length	Very labor intensive; unlikely to obtain complete spectrum of differentially expressed genes	100, 101
Differential display	0.01	Allows simultaneous comparison of multiple samples; very fast; small amount of starting material required; good for obtaining markers	Difficult to optimize PCR conditions; large number of false positives; cDNA isolated is not full length	73, 102
Differential screening	0.05–0.2	Few false positives; clones obtained are full length	Labor intensive	103
Screening-subtracted library with unsubtracted probe	0.01	Likely to obtain clones of particular interest	Labor intensive	56, 76, 104
Screening-subtracted library with subtracted probe	0.001	Very likely to obtain clones of particular interest	Labor intensive	71, 72, 92

(*b*) the difficulty of isolating a complete spectrum of differentially represented clones. Although full-length library-based subtraction can yield full-length clones, it is difficult to perform enough rounds of subtraction to isolate differences from complex tissues. Conversely, PCR-based techniques allow multiple rounds of subtraction and isolation of differences from complex tissues but yield only small cDNA or genomic fragments. It is time-consuming to determine how many different clones are represented by the fragments derived from a PCR-based subtraction and to subsequently isolate corresponding full-length clones. Further, although subtractive cloning is better than any other current technique for isolating a large spectrum of genes, it is still difficult to determine when a full complement of differentially represented genes has been isolated.

Solutions

Three solutions are available for addressing problems associated with current subtraction and positive selection techniques: long and accurate PCR, direct sequencing of cDNA clones, and analysis of a whole genome on glass "chips."

LONG AND ACCURATE PCR One way to obtain full-length clones after subtraction is to perform the entire procedure with full-length cDNA, even when PCR amplification is used. This procedure may be possible using a recently discovered technique for long and accurate PCR (77, 78). Conventional PCR is limited by the inability of the polymerase used (Taq) to replace erroneously introduced nucleotides. The presence of a misincorporated nucleotide at the 3′-terminus is thought to lead to termination of synthesis (77, 79). Taq polymerase is more likely to make a mistake during the synthesis of long templates than during the synthesis of short ones, which may be why short templates are amplified more efficiently than are longer ones.

By using a mixture of Taq polymerase and a thermostable polymerase that has proofreading ability, investigators can use PCR to efficiently copy long (>35 kb) DNA fragments (77, 78). As a result, it may be possible to efficiently copy complex mixtures of full-length cDNAs under carefully controlled conditions through the multiple rounds of PCR that subtractions require. The products of a PCR-based subtraction or positive selection could correspond to the original large fragments of input nucleic acid and obviate the need for subsequent sorting and full-length clone isolation.

DIRECT SEQUENCING OF cDNA CLONES High-throughput DNA sequencing has made it feasible to use random sequencing of cDNA clones to screen for differentially represented genes. By randomly sequencing a large number of clones from each of two libraries derived from different nucleic acid populations, and then comparing the frequency at which different clones are encountered, it is possible to identify genes that are differentially expressed between the populations (see e.g. 80). With this approach, abundant transcripts are sequenced several times, and although this method requires sequencing of only a small portion of each clone (usually a few hundred nucleotides from the 3′-end), a large number of sequencing reactions is required.

Velculescu and coworkers (81) have devised a technique that defines short tags corresponding to the 3′-end of each mRNA and that can analyze many more transcripts simultaneously than can conventional sequencing. Another method for decreasing the number of clones to be sequenced uses normalized, or self-subtracted, libraries (63, 64) for the comparison. The problem with this approach is that small quantitative differences in representation are lost, and only all-or-none expression can be monitored.

ANALYZING A WHOLE GENOME ON "CHIPS" As the nucleotide sequence of the entire genome of several organisms is uncovered, it may become unnecessary to isolate differentially expressed genes de novo and instead may be possible to screen all known genes for differential expression. A method for such an analysis has been proposed recently (82). This technique relies on the synthesis of large arrays (up to 400,000 different 20-mer oligos) of oligonucleotides on small (1.6 cm^2) glass "chips" (82–84). Given 100,000 genes in the human genome, this method would allow representation of all human genes (with 10–20 different oligos representing each gene) on a few chips. The chips are then probed with labeled cDNAs from different tissue sources, and a comparison between the hybridization of each gene to the different sources is made (see 82 for a recent feasibility analysis). This analysis is performed efficiently by computer, and a report of the genes differentially represented in the two or more tissue sources under comparison can be generated. Since the complete set of genes is represented on the chip array, the complete set of differentially expressed genes could theoretically be obtained. Although sensitivity of the method appears impressive (82), it is unclear whether the method will be as sensitive as current subtraction techniques.

It also may be possible to use this method to score relative levels of gene expression in different tissues (82). After identifying differentially expressed genes, investigators could obtain full-length cDNAs from a clone repository. This approach could supersede subtractive cloning or positive selection and would allow identification of differences between many tissue sources far more easily than can be done at present.

The human genome will be sequenced within the next few years, and as DNA sequencing technologies improve, the genomes of other organisms will be analyzed, although it may take many years until all the useful model organisms are ready for chip analysis. Until then, subtractive cloning and positive selection techniques will remain the most powerful methods for analyzing the multitude of biological processes that involve differential gene representation.

ACKNOWLEDGMENTS

We thank our colleagues in the Sive laboratory for their comments on the manuscript and David Bartel, Tom Tuschl, and Peter Unrau for their patient and extensive discussion of renaturation kinetics. This article began as a talk at Genetics Institute in Massachusetts, and HLS appreciates that impetus from which this review grew. Space constraints prevented us from citing all relevant references, and we apologize for omissions. While writing this review, we were supported by grants from the National Science Foundation. HLS is the recipient of a NSF Young Investigator's Award and is the Latham Family Career Development Professor at MIT; CGS was supported by a postdoctoral fellowship from the American Cancer Society, Massachusetts Division; and BIS is an NIH postdoctoral fellow.

Visit the *Annual Reviews home page* at http://www.annurev.org.

Literature Cited

1. Britten RJ, Kohne DE. 1968. *Science* 161: 529–40
2. Watson JD, Crick FHC. 1953. *Nature* 171: 737–38
3. Meselson M, Stahl FW. 1958. *Proc. Natl. Acad. Sci. USA* 44:671–82
4. Marmur J, Doty P. 1959. *Nature* 183: 1427–29
5. Marmur J, Lane D. 1960. *Proc. Natl. Acad. Sci. USA* 46:453–61
6. Doty P, Marmur J, Eigner J, Schildkraut C. 1960. *Proc. Natl. Acad. Sci. USA* 46: 461–76
7. Schildkraut CL, Marmur J, Doty P. 1961. *J. Mol. Biol.* 3:595–617
8. Hoyer BH, McCarthy BJ, Bolton ET. 1964. *Science* 144:959–67
9. Wetmur JG, Davidson N. 1968. *J. Mol. Biol.* 31:349–70
10. Hahn WE, Laird CD. 1971. *Science* 173: 158–61
11. Davidson EH. 1971. *J. Mol. Biol.* 56:491–506
12. Britten RJ, Graham DE, Neufeld BR. 1974. *Methods Enzymol.* 29:363–418
13. Walker PMB, McLaren A. 1965. *Nature* 208:1175–79
14. Bernardi G. 1965. *Nature* 206:779–83
15. Ryffel GU, McCarthy BJ. 1975. *Biochemistry* 14:1379–85
16. Bishop JO, Morton JG, Rosbash M, Richardson M. 1974. *Nature* 250:199–204
17. Axel R, Feigelson P, Schutz G. 1976. *Cell* 7:247–54
18. Hastie ND, Bishop JO. 1976. *Cell* 9:761–74
19. Bantle JA, Hahn WE. 1976. *Cell* 8:139–50
20. Galau GA, Britten RJ, Davidson EH. 1974. *Cell* 2:9–20
21. Hinegardner RT. 1968. *Am. Nat.* 102: 517–23
22. Davidson EH, Britten RJ. 1973. *Q. Rev. Biol.* 48:565–613
23. Galau GA, Klein WH, Davis MM, Wold BJ, Britten RJ, Davidson EH. 1976. *Cell* 7:487–505
24. Crampton J, Humphries S, Woods D, Williamson R. 1980. *Nucleic Acids Res.* 8:6007–17
25. Davis MM, Cohen DI, Nielsen EA, DeFranco AD, Paul WE. 1982. In *The Isolation of B and T Cell Specific Genes,* ed. E Vitetta, CF Fox, pp. 215. New York: Academic
26. Lewin B. 1994. *Genes V.* Oxford: Oxford Univ. Press
27. Fields C, Adams MD, White O, Ventner JC. 1994. *Nat. Genet.* 7:345–46
28. Antequera F, Bird A. 1993. *Proc. Natl. Acad. Sci. USA* 90:11995–99
29. Wang Z, Brown DD. 1991. *Proc. Natl. Acad. Sci. USA* 88:11505–9
30. Patel M, Sive HL. 1996. In *Current Protocols in Molecular Biology,* ed. FM Ausubel, R Brent, RE Kingston, DD Moore, JG Seidman, et al. New York: Wiley
31. Duguid JR, Rohwer RG, Seed B. 1988. *Proc. Natl. Acad. Sci. USA* 85:5738–42
32. Duguid JR, Bohmont CW, Liu NG, Tourtellotte WW. 1989. *Proc. Natl. Acad. Sci. USA* 86:7260–64
33. Duguid JR, Dinauer MC. 1990. *Nucleic Acids Res.* 18:2789–92
34. Rubenstein JL, Brice AE, Ciaranello RD, Denney D, Porteus MH, Usdin TB. 1990. *Nucleic Acids Res.* 18:4833–42
35. Jeffreys AJ, Wilson V, Neumann R, Keyte J. 1988. *Nucleic Acids Res.* 16:10953–71
36. Forster AC, McInnes JL, Skingle DC, Symons RH. 1985. *Nucleic Acids Res.* 13:745–61
37. Lebeau MC, Alvarez-Bolado G, Wahli W, Catsicas S. 1991. *Nucleic Acids Res.* 19:4778
38. Swaroop A, Xu JZ, Agarwal N, Weissman SM. 1991. *Nucleic Acids Res.* 19:1954
39. Wieland I, Bolger G, Asouline G, Wigler M. 1990. *Proc. Natl. Acad. Sci. USA* 87:2720–24
40. Syvänen AC, Bengtström M, Tenhunen J, Söderlund H. 1988. *Nucleic Acids Res.* 16:11327–38
41. Sive HL, St. John T. 1988. *Nucleic Acids Res.* 16:10937
42. Porath J, Carlsson J, Olsson I, Belfrage G. 1975. *Nature* 258:598–99
43. Welcher AA, Torres AR, Ward DC. 1986. *Nucleic Acids Res.* 14:10027–44
44. López-Fernández LA, del Mazo J. 1993. *BioTechniques* 15:654–59
45. Syvänen AC, Laaksonen M, Söderlund

H. 1986. *Nucleic Acids Res.* 14:5037–48

46. Hampson IN, Pope L, Cowling GJ, Dexter TM. 1992. *Nucleic Acids Res.* 20:2899
47. Scott MR, Westphal KH, Rigby PW. 1983. *Cell* 34:557–67
48. Vitek MP, Kreissman SG, Gross RH. 1981. *Nucleic Acids Res.* 9:1191–1202
49. Hara E, Kato T, Nakada S, Sekiya S, Oda K. 1991. *Nucleic Acids Res.* 19:7097–104
50. Rodriguez IR, Chader GJ. 1992. *Nucleic Acids Res.* 20:3528
51. Maréchal D, Forceille C, Breyer D, Delapierre D, Dresse A. 1993. *Anal. Biochem.* 208:330–33
52. Wetmur JG. 1976. *Annu. Rev. Biophys. Bioeng.* 5:337–61
53. Rivolta MN, Wilcox ER. 1995. *Nucleic Acids Res.* 23:2565–66
54. Kuze K, Shimizu A, Honjo T. 1989. *Nucleic Acids Res.* 17:807
55. Timberlake WE. 1980. *Dev. Biol.* 78:497–510
56. Davis RL, Weintraub H, Lassar AB. 1987. *Cell* 51:987–1000
57. Kohne DE, Levison SA, Byers MJ. 1977. *Biochemistry* 16:5329–41
58. Kunkel LM, Monaco AP, Middlesworth W, Ochs HD, Latt SA. 1985. *Proc. Natl. Acad. Sci. USA* 82:4778–82
59. Sutcliffe JG. 1988. *Annu. Rev. Neurosci.* 11:157–98
60. Yokota H, Oishi M. 1990. *Proc. Natl. Acad. Sci. USA* 87:6398–402
61. Patanjali SR, Parimoo S, Weissman SM. 1991. *Proc. Natl. Acad. Sci. USA* 88:1943–47
62. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. 1994. *Proc. Natl. Acad. Sci. USA* 91:9228–32
63. Lamar EE, Palmer E. 1984. *Cell* 37:171–77
64. Nussbaum RL, Lesko JG, Lewis RA, Ledbetter SA, Ledbetter DH. 1987. *Proc. Natl. Acad. Sci. USA* 84:6521–25
65. Roninson IB. 1983. *Nucleic Acids Res.* 11:5413–31
66. Sasaki H, Nomura S, Akiyama N, Takahashi A, Sugimura T, et al. 1994. *Cancer Res.* 54:5821–23
67. Straus D, Ausubel FM. 1990. *Proc. Natl. Acad. Sci. USA* 87:1889–93
68. Lisitsyn N, Lisitsyn N, Wigler M. 1993. *Science* 259:946–51
69. Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, et al. 1996. *Proc. Natl. Acad. Sci. USA* 93:6025–30
70. Zeng J, Gorski RA, Hamer D. 1994. *Nucleic Acids Res.* 22:4381–85
71. Schneider C, King RM, Philipson L. 1988. *Cell* 54:787–93
72. Sive HL, Hattori K, Weintraub H. 1989. *Cell* 58:171–80
73. Liang P, Pardee AB. 1992. *Science* 257:967–71
74. Saito H, Kranz DM, Takagaki Y, Hayday AC, Eisen HN, Tonegawa S. 1984. *Nature* 312:36–40
75. Klar A, Baldassare M, Jessell TM. 1992. *Cell* 69:95–110
76. el-Deiry WS, Tokino T, Velculescu VE, Levy DB, Parsons R, et al. 1993. *Cell* 75:817–25
77. Barnes WM. 1994. *Proc. Natl. Acad. Sci. USA* 91:2216–20
78. Cheng S, Fockler C, Barnes WM, Higuchi R. 1994. *Proc. Natl. Acad. Sci. USA* 91:5695–99
79. Huang MM, Arnheim N, Goodman MF. 1992. *Nucleic Acids Res.* 20:4567–73
80. Okubo K, Itoh K, Fukushima A, Yoshii J, Matsubara K. 1995. *Genomics* 30:178–86
81. Velculescu V, Zhang L, Vogelstein B, Kinzler KW. 1995. *Science* 270:484–87
82. Lockhardt DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. 1996. *Nat. Biotechnol.* 14:1675–80
83. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. 1994. *Proc. Natl. Acad. Sci. USA* 91:5022–26
84. Fodor SPA, Read LJ, Pirrung MC, Stryer L, Lu AT, Solas D. 1991. *Science* 251:767–73
85. Bernardi G. 1971. *Methods Enzymol.* 21:95
86. Davis MM, Cohen DI, Nielsen EA, Steinmetz M, Paul WE, Hood L. 1984. *Proc. Natl. Acad. Sci. USA* 81:2194–98
87. Langer PR, Waldrop AA, Ward DC. 1981. *Proc. Natl. Acad. Sci. USA* 78:6633–37
88. Brigati DJ, Myerson D, Leary JJ, Spalholz B, Travis SZ, et al. 1983. *Virology* 126:32–50
89. Hubank M, Schatz DG. 1994. *Nucleic Acids Res.* 22:5640–48
90. Rosenberg M, Przybylska M, Straus D. 1994. *Proc. Natl. Acad. Sci. USA* 91:6113–17
91. Lisitsyn NA. 1995. *Trends Genet.* 11:303–7
92. Hedrick SM, Cohen DI, Nielsen EA, Davis MM. 1984. *Nature* 308:149–53
93. Timblin C, Battey J, Kuehl WM. 1990. *Nucleic Acids Res.* 18:1587–93
94. Wieland I, Böhm M, Bogatz S. 1992. *Proc. Natl. Acad. Sci. USA* 89:9705–9

95. Sargent TD, Dawid IB. 1983. *Science* 222:135–39

96. Rothstein JL, Johnson D, De Loia JA, Skowronski J, Solter D, Knowles B. 1992. *Genes Dev.* 6:1190–201

97. Harrison SM, Dunwoodie SL, Arkell RM, Lehrach H, Beddington RS. 1995. *Development* 121:2479–89

98. Saito H, Kranz DM, Takagaki Y, Hayday AC, Eisen HN, Tonegawa S. 1984. *Nature* 309:757–62

99. Reijo R, Lee TY, Salo P, Alagappan R, Brown LG, et al. 1995. *Nat. Genet.* 10:383–93

100. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. 1991. *Science* 252:1651–56

101. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. *Science* 270:484–87

102. Liang P, Bauer D, Averboukh L, Warthoe P, Rohrwild M, et al. 1995. *Methods Enzymol.* 254:304–21

103. St. John TP, Davis RW. 1979. *Cell* 16:443–52

104. Mather EL, Alt FW, Bothwell AL, Baltimore D, Koshland ME. 1981. *Cell* 23:369–78