MAMMALIAN CELL GENETICS

M. Wigler J. Alexander, A. Brady, J. Brodsky, A. Buglino, K. Chang, M. Chi, J. Douglas, D. Esposito, S. Gass, J. Healy, R. Lucito, A. Reiner, M. Riggs, L. Rodgers, S. Rostan, E. Thomas, J. Troge, R. von Roeschlaub, J. West

THE CENTRAL PROBLEM OF CANCER

Cancer is a disease caused by genetic mutations. The particular combinations of genes mutated in the different tissues of origin probably determine the characteristics of individual cancers: their degree of malignancy and their particular therapeutic vulnerabilities. Science knows some of the genes that become mutated in cancer, much about the cellular pathways on which they act, and a little about the way that cancer cells interact with the host. Molecular oncology has finally begun to make very promising contributions to the treatment and diagnosis of cancers. So it is beguiling to think that the "Cancer Problem" has been solved, or will be solved by the routine cranking of the wheel, and thus be tempted to turn one's attention elsewhere, to other biological or biomedical problems that appear to hold greater mystery. This would be a grave mistake. The adequate treatment, diagnosis, and prevention of cancer in patients will require a detailed understanding of the specific molecular mechanisms that go awry in specific cancers, both within the cancerous cell and between the cancer and the host. Without this knowledge, we cannot improve diagnosis, prognosis, and treatment.

Science is far from having a complete list of either oncogenes or tumor suppressor genes that are mutated in even the most common types of cancer. Our work during the past years attests to this, as do the efforts of our direct collaborators and other scientists using similar approaches. Major new oncogenes and tumor suppressor genes abound, and, probably, relevant new cellular pathways. Science does not yet have a significant handle on the ways in which cancer cells interact with their host, or the cooperative ways in which cancer cells within a tumor may help one another. Moreover, medicine does not yet have a practical method for assessing the mutations of individual cancers. During the past years, we have developed new approaches to these problems that we expect will translate into improved diagnosis, prognosis, and treatment. We expect that the same approaches we are developing will lead to solutions for other complex human genetic diseases, problems in genomic mapping, and genome evolution.

REPRESENTATIONAL GENOMIC ANALYSIS

In the early 1990s, we developed a powerful approach to genomic analysis called RDA (representational difference analysis). RDA was designed to solve this problem: Find A when one genome is $X + A$ and the other genome is X (Lisitsyn et al., *Science 258:* 946 [1993]). Many biological searches can be reformulated as this problem, including the search for oncogenes, tumor suppressors, pathogens, or even genetic differences between individuals. Essentially, RDA was a combination of a new method for difference cloning, based on differential DNA hybridization, and a method for reducing the complexity of genomes. Complexity reduction was achieved using a process that we call representation (see Fig. 1), in which a reproducible subset of the genome is derived from the entirety. The value of representations lies in the

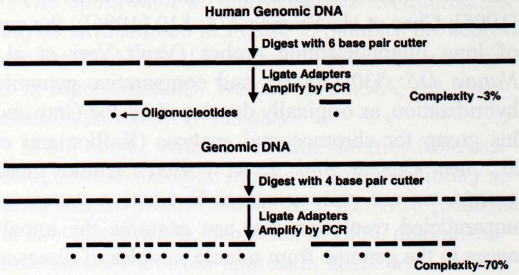


FIGURE 1 Schematic of the preparation of a low-complexity (*upper panel*) and high-complexity representation (*lower panel*). In designing the array, we chose probes from the amplified representational fragments.

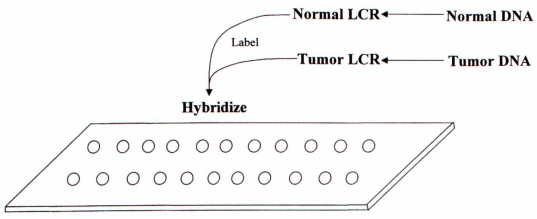


FIGURE 2 A typical format for our gene copy number arrays.

enabling of many protocols dependent on hybridization of genomic DNA. The entire human genome is just too complex in sequence to hybridize efficiently and specifically.

RDA led to the discovery of several new oncogenes and tumor suppressor genes by our collaborators and us. One of these, *DBC-2*, is reported jointly with Hamaguchi and another, *DIGIT,* by Lucito, both in this Annual Report. PTEN, which was discovered jointly with Ramon Parson's group at Columbia (Li et al., *Science 275:* 1943 [1997]), is described here. Additionally, many candidate oncogenes were found with our collaborators at Tularik. Several of these oncogenes belong to novel classes of gene families and have gone into drug discovery screens at Tularik. They will not be discussed here.

Although RDA has been a useful method for finding genetic lesions in cancer, and more effective than any technique before it, it is not sufficiently powerful to satisfy the needs of the cancer problem. We therefore began to fashion a microarray-based version of RDA wherein representations of two samples are compared by array hybridization (see Fig. 2). Our array method builds upon the success of microarrays to analyze gene expression (Schena et al., *Science 270: 467* [1995]; Chee et al., *Science 274:* 510 [1996]), the use of long oligonucleotide probes (Van't Veer et al., *Nature 415:* 530 [2002]), and comparative genomic hybridization, as originally developed by Joe Gray and his group for chromosomal analysis (Kallioniemi et al., *Sem. Cancer Biol. 1:* 41 [1993]). Unlike those groups, we use representations to create a system of unparalleled resolution that can examine the imbalances in the genome from minute samples of diseased (cancerous) tissues. The same methodology can also be used to study other complex human genetic diseases, assemble physical maps of probes from any genome, and enable us to compare the genomes of related species. Such comparative genomic studies will provide maps of rearrangements, segmental duplications, and deletions between related organisms.

OTHER RELATED PROJECTS

Our representational analysis using microarrays has led us, by necessity, to develop a novel genome-informatics algorithm for examining exact sequence matches within a genome. The application of this tool to genome analysis has led to a series of unexplained and interesting observations about genome and gene structure. We will describe our method and observations below, together with some theories. Finally, we describe a new effort to use expression microarray profiling to study the important and largely neglected subject of cancer cell communication: the study of the messages that cancer cells send and receive that may influence their persistence, growth, and spread in the host, and their cooperativity during tumor formation.

PTEN

The *PTEN* tumor suppressor gene was identified as a collaboration between Ramon Parson's laboratory at Columbia University and our own (Li et al., *Science 275:* 1943 [1997]). A deletion locus was first detected using RDA on breast cancer biopsies, which was then recognized as coincident with a locus implicated in a rare genetic disease, Cowden's syndrome, that predisposes affecteds to breast cancers. The *PTEN* gene was found to be mutated in a number of other cancer types, particularly brain cancers (glioblastomas) (Wang et al., *Cancer Res. 19:* 4183 [1997]). The gene has homology with protein phosphatases, and subsequently, Nick Tonks and Mike Myers here at CSHL showed that *PTEN* encodes a protein and phosphatidyl inositol (PIP3) phosphatase (Myers et al., *Proc. Natl. Acad. Sci. 94:* 9052 [1997]; Myers et al., *Proc. Natl. Acad. Sci. 95:* 13513 [1998]). In a collaboration with our group, the PIP3 phosphatase activity of PTEN, which opposes the action of PI3 kinase, a known oncogene and a candidate effector for the *Ras* oncogenes, was determined to be the most likely function that suppresses tumorigenicity.

Last year, in a collaboration with Vivek Mittal here at CSHL, we published a gene expression array analysis of PTEN (Stolarov et al. 2001). Using an efficient ecdysone-inducible promoter derived from the work

of Evans and co-workers (No et al., *Proc. Natl. Acad. Sci. 93:* 3346 [1996]) and a retrovirus-mediated delivery system, we showed that the genes which were under the control of PTEN, both positively and negatively, are also the genes under the control of a small molecule inhibitor of PI3 kinase. These experiments perhaps most clearly demonstrate that PTEN acts by controlling the PI3 kinase pathway. Among the genes that we observed to be negatively controlled by PTEN were those encoding transforming growth factor-β (TGF-β) and genes of the cholesterol biosynthesis pathway, including HMG-coA synthase and reductase, and squalene synthetase. Since PTEN may have a role in opposing the insulin response (Butler et al., *Diabetes 4:* 1028 [2002]), it is possible that these observations may indicate a link between the insulin signaling pathway and elevated cholesterol biosynthesis.

We have pursued these studies by investigating the role of TGF-β in the response of cells to PTEN. TGF-β is a growth factor that has a role in development and differentiation and has been suggested to have a role in tumor development and angiogenesis. When PTEN is induced in a glioblastoma cell line, the cells flatten out and form adherent sheets. In fact, these morphological effects are reversed by the administration of TGF-β. However, expression analysis has so far failed to provide a convincing link between TGF-β and PTEN. The genes controlled by both are a partially overlapping set, but the transcriptional effects of PTEN induction are not reversed by administering TGF-β. In fact, many of the PTEN responses are enhanced. The interaction between TGF-β and PTEN is thus likely to be complex.

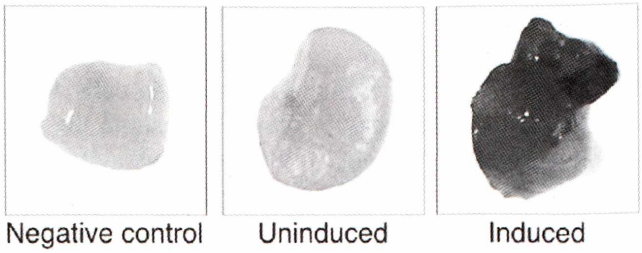
Last year, we reported the development of an effective ecdysone-inducible viral vector system, the system that was used in the above studies. Initially, we were unable to determine if the system would function in vivo. That difficulty was illusory and has been overcome (Fig. 3). Tumors that carry an ecdysone-inducible β-*GAL* gene can be induced to express that gene upon injection of an ecdysone homolog into the peritoneum of tumor-bearing mice (Fig. 3, top). This has enabled us to test the expression of PTEN in the glioblastoma cell line U87MG that has deleted its endogenous *PTEN* gene. The growth curves shown in Figure 3 (middle) suggest that expression of PTEN can arrest tumor growth. Considering that U87MG has been maintained in cell culture for many years after its removal from a patient, this is quite remarkable. Our observations support one of the fundamental hypotheses of cancer genetics: that pharmacological correction of the critical genetic defects of a cancer can lead to a cure. Figure 3 (bottom) shows the reverse result, as a control: When an oncogene (H-*ras*) is induced in U87MG, the cells form tumors more rapidly. These studies illustrate the utility of the ecdysone-inducible system for assaying both tumor suppressor genes and oncogenes in vivo.

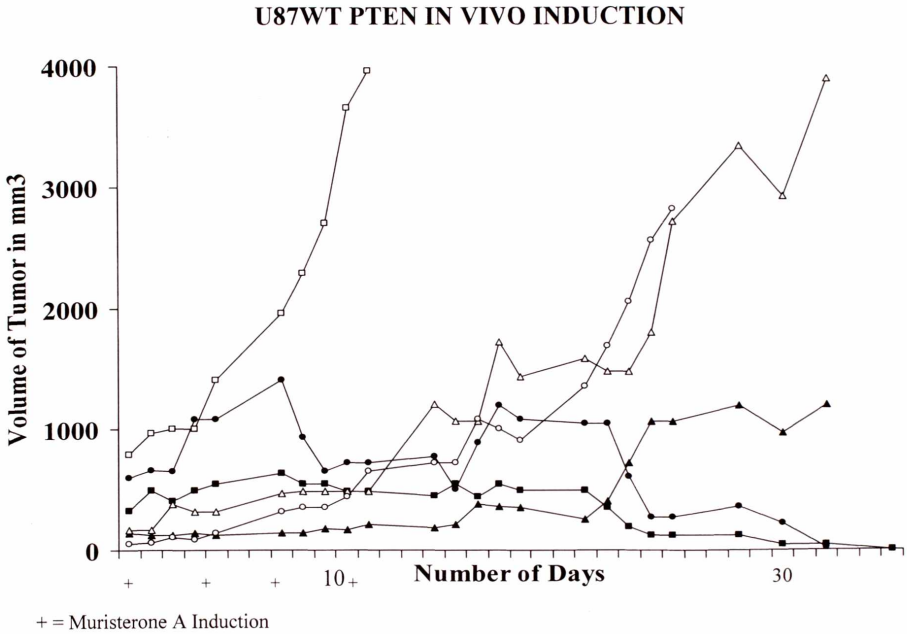
REPRESENTATIONAL MICROARRAY ANALYSIS OF THE GENETIC LESIONS IN CANCERS

The majority of our research effort is, in fact, directed to the use of microarray analysis to determine the genetic defects of cancer cells. As mentioned in the introduction, this effort has evolved from our early RDA studies, which clearly indicated that there were many oncogenes and tumor suppressor genes that had not yet been discovered. The essential core of our approach is the representation, which simplifies a complex genome so that its analysis by hybridization becomes more feasible. As illustrated in Figure 1, and described in past Annual Reports, we make representations by cleaving genomic DNA with a restriction endonuclease and then select only small DNA fragments by polymerase chain reaction (PCR) amplification. This enables us to reduce complexity in a computationally predictable manner and also analyze genomic DNA from minute amounts of specimen. A long history of the application of representations supports its reproducibility and reliability (Lisitsyn et al., *Proc. Natl. Acad. Sci. 92:* 151 [1995]; Lucito et al., *Proc. Natl. Acad. Sci. 95:* 4487 [1998]). If a sequence is present in a representation made with a particular restriction endonuclease, it derives from a small restriction fragment, and if a sequence is in a small restriction fragment, it is in the representation made with that enzyme.

The application of representations to microarray analysis is illustrated in Figure 2. Samples are represented and hybridized to arrays of probes that are present in representations. We call this RMA, for representational microarray analysis.

In our earlier studies, we made libraries from representations and arrayed fragments from the library as probes (Lucito et al., *Genome Res. 10:* 1726 [2000]). Our first experiments, published in 2000, demonstrated the feasibility of our approach. There were several problems with this approach, however. First, it was necessary to derive map positions for our probes. We accomplished this by sequencing (a collaboration with Richard McCombie here at CSHL). Second, these





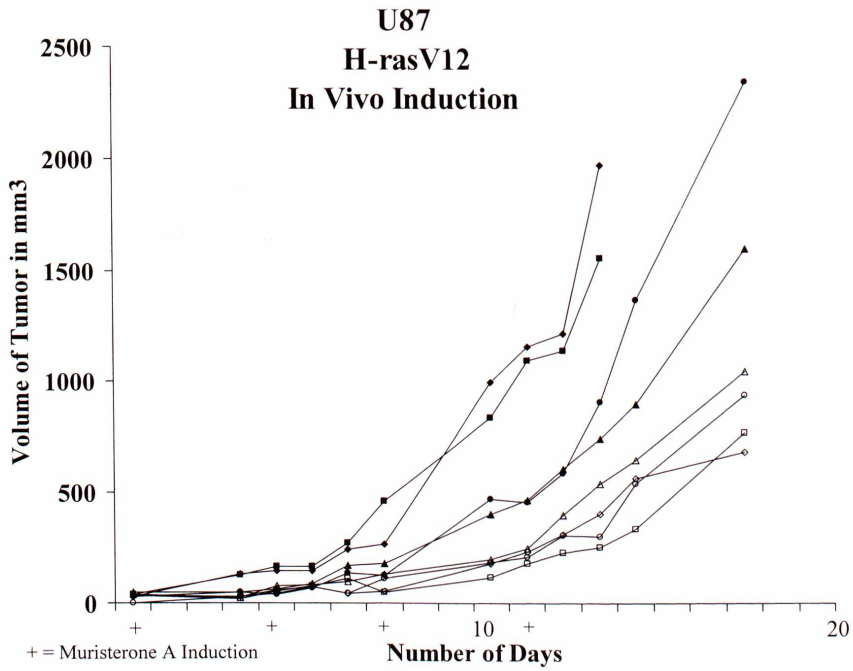


FIGURE 3 (*Top*) Induction of β-galactosidase expression under the control of the ecdysone promoter in a tumor. (*Middle*) Inhibition of tumor growth resulting from induction of PTEN expression under the control of the ecdysone promoter. (*Bottom*) Increased tumor growth in tumors due to the induction of oncogenic *ras* under the control of the ecdysone reporter. (*Open symbols*) Uninduced; (*closed symbols*) induced.

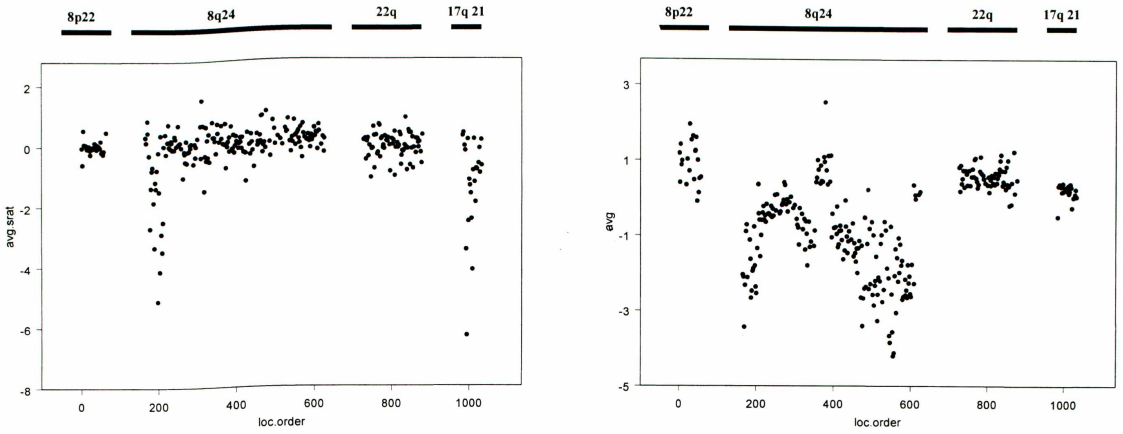


FIGURE 4 Two breast cancer samples are examined by probes distributed among four loci, on 8p11, 8q24, 22q, and 17q21. Each point in the graphs is the average of four measurements at each probe. Probes have been arranged according to their chromosomal position. Gene amplification is seen on the vertical axis as values in the negative range.

studies indicated that repeats were more widely distributed in the genome than we had expected, and probes with repeats were not useful for measuring gene copy number in a sequence-specific way. As a consequence, fewer probes were usable than we had expected. A third disadvantage of our first approach was that our probe set was randomly distributed in the genome, and hence we could not focus probes in regions of interest. Fourth, the infrastructure of maintaining large physical collections of probes was daunting. Finally, the arrays made from fragments are generally of a variable quality.

We were encouraged by the success of arrays of long oligonucleotide probes, which had been used by other investigators for expression analysis. Our first test of oligo arrays was successful (see Fig. 4). Importantly, we can design oligonucleotide probes directly from the published human genome assembly inferred to be in representations (see Fig. 1). In this way, we can avoid repeats and have many more probes that are usable. We can therefore make arrays with higher resolving power. In addition, we can place oligonucleotide probes at will near genes of interest, and thus increase our resolution in important regions of the genome. We call this procedure ROMA, for representational oligonucleotide microarray analysis.

Figure 4 illustrates results of the microarray analysis of two breast cancer specimens. The microarrays were designed with probes distributed over a region containing the c-*Myc* oncogene on chromosome 8 and the *ErbB-2* gene on chromosome 17, as well as concentrations of probes from elsewhere on the genome. As can be readily seen from the figure, amplification at these loci, and at another locus distant from c-*Myc*, can be readily detected. These arrays contain about 2000 probes, arranged in chromosomal order of the horizontal axis. The vertical axis displays the concentration in the cancer sample relative to their concentration in a normal sample for each probe, measured as a ratio of intensity of cancer to normal. From the peak of Figure 4 (left), we can see this patient has an amplification at *ErbB-2* and one at the c-*Myc* locus. Figure 4 (right) shows the analysis of a tumor from a patient with amplification around c-*Myc* and another locus at the other end of that segment, but no amplification about *ErbB-2*. These experiments demonstrate the ability of our microarray system to characterize the cancers of patients, and at the same time, point to new loci that are likely to be relevant to our understanding and treatment of breast cancer.

Although amplification is a hallmark of oncogenes, deletion is a hallmark of tumor suppressor genes. In Figure 5 (top), we demonstrate the ability to detect (the equivalent of) homozygous deletions. We prepared two representations, a *Bgl*II rep and a *Bgl*II rep "killed" with *Eco*R1. The latter is a *Bgl*II representation from which fragments containing *Eco*R1 sites are mainly destroyed. On the left of the figure are all of the probes derived from *Bgl*II fragments containing *Eco*RI sites, and on the right are those derived from *Bgl*II fragments that do not contain *Eco*R1 sites. The vertical axis is the ratio of hybridization signal, one representation versus

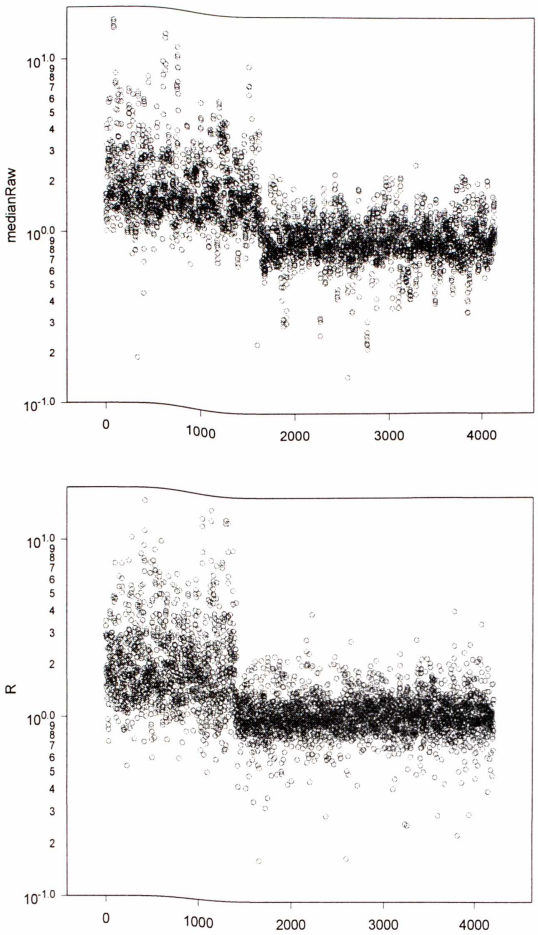


FIGURE 5 (*Top*) Actual data from a control experiment. In this control experiment, probes are not in chromosomal order. Rather, probes that are expected to show increased ratios, because they derive from *Bgl*II fragments with *Eco*RI cleavage sites, are on the left. (*Bottom*) Simulation of the top panel, based on sampled probe parameters.

the other. It is clear that the probes set on the right "report" differently from those on the left.

SIMULATION OF ARRAY RESULTS: PREDICTIONS OF PERFORMANCE

It is clear from Figure 5 (top) that not all probes report equally well. Some probes fail to perform. There are a variety of possible reasons for this: Some oligonucleotides are not properly synthesized, some *Bgl*II fragments do not amplify well, the genome sequence from which we derive our probes is inaccurate, some probes recognize repeat sequences, and some probes hybridize poorly due to subtleties of their secondary structure. However, microarray hybridizations such as that shown in Figure 5 (top) enable us to characterize our probes, and future results can be corrected for the performance character of each probe.

Experiments such as Figure 5 (top) also allow us to simulate overall array performance, using statistical sampling protocols. We assume that the basic equation for probe hybridization is given by

$$I = \alpha(Ax + B) + \beta$$

where, for each channel, I is hybridization intensity, α is a multiplicative system noise, A is the specific hybridization parameter of the probe, x is the concentration of complementary sequence to the probe in the sample, B is the nonspecific hybridization parameter, and β is the additive system noise. We can then simulate array performance for any number of probes, assuming that the distribution of probe parameters A and B of the probes we have sampled will be representative. Figure 5 (bottom) illustrates a simulation of the data shown in the top panel.

We can judge from our simulations how many probes with what probe parameters enable us to detect which kinds of lesions. For example, a 60K probe set with the characteristics of our current probe set will be able to detect reliable fivefold amplification of a 400-kb region, similar-sized homozygous deletions, and 3-Mb hemizygous deletions. However, if we use 120K probes with twice the performance characteristics of our current set, we will be able to detect even 150-kb hemizygous deletions. If probes are arranged six to a gene, we believe we can detect hemizygous deletion of a single gene. These predictions have important applications to the analysis of the genome in individuals with inherited genetic disease.

APPLICATIONS OF ROMA TO GENETIC DISEASE

The applications of ROMA for the analysis of cancer have been illustrated. The method provides a rapid way to define the regions of the chromosome of cancer cells where amplifications and deletions take place. This analysis can proceed in a highly parallel manner, at high resolution, from small amounts of cancer biopsies. The benefits from this study would be a catalog of lesions that contain oncogenes and tumor suppressors and could quickly narrow the search for those genes involved in human cancer. My estimate is

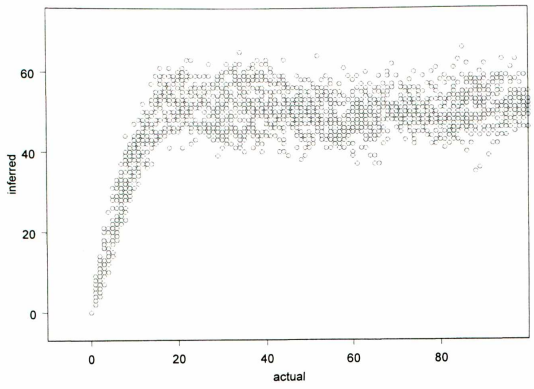


FIGURE 6 Hamming distance. Horizontal is Euclidean distance between probes. Vertical is Hamming distance. The Hamming distance provides a local metric.

that we know fewer than a quarter of the genes that are frequently mutated in cancers. When these data are correlated with clinical data, then we also derive knowledge about which genes are most commonly mutated in which kinds of cancer, and which mutations or combinations of mutations have the greatest value in predicting malignancy and outcome to particular therapies. Knowledge of the new oncogenes and tumor suppressor genes is likely to lead to new therapeutic approaches to cancer and new forms of diagnosis and detection.

The applications of ROMA, however, are likely to extend beyond cancer to gene identification in other complex human genetic diseases. Many of the mutant genes that cause genetic disease are mutated by deletion. I estimate that 5–10% of loci associated with disease are lost through deletion rather than point mutation. With our method, we can detect hemizygous deletion. We compared a patient with velo-facial cardio-syndrome with a normal. This patient had a 3-Mb deletion of chromosome 22, which was readily seen by array hybridization. Our simulations predict that with a dense enough microarray, we will be able to detect hemizygosity at individual genes. Thus, we could be in a position to scan the entire genome of affected and unaffected humans for such deletion events. This may be an efficient way to analyze humans with spontaneous deletions in the germ line, or with complex genetic diseases, as it is postulated that the aggregation of nonlethal mutations (e.g., hemizygous loss) of genes with a common function can lead to disease susceptibility. For example, a cohort of unrelated schizophrenics might show a statistically high level of hemizygous deletion of certain genes critical in brain function.

PHYSICAL MAPPING OF GENOMES AND COMPARATIVE GENOMICS

There are major applications of ROMA to the physical mapping of the genome. It is widely recognized that the publicly available sequence assembly of the human genome is quite flawed. Many regions of the genome are misassembled, there are large gaps, and in particular, the duplicated regions are poorly assembled. The extent of the defects of assembly is largely unknown but could adversely affect the techniques we have developed, in particular since our oligo-based representational methods are strongly dependent on the colocalization of probes by map position.

Last year, in a collaboration with Bhubaneswar Mishra and Will Casey of the Courant Institute of Applied Mathematics at New York University, we described algorithms whereby the order of arrayed probes could be inferred from a series of hybridizations of arrays with random pools of BACs (bacterial artificial chromosomes). The central concept is that probe proximity will be inferred by a similar pattern of hybridization signals from nearby probes throughout the series of hybridizations. Probes less than a BAC's length away will tend to show related patterns, and the closer the probes, the closer their patterns will match. The Hamming distance gives a quantitative measure of pattern matching, and for nearby probes, the Hamming distance is proportional to Euclidean distance (see Fig. 6) over the short range. The Hamming distance provides a local metric from which large stretches of ordered linear map information may be inferred (Casey et al. 2001).

We estimate that on the order of 100 hybridizations are sufficient to create correctly ordered contigs of about 75 probes, assuming an average arrayed probe density of five probes per BAC. The boundaries of regions of segmental duplications should also be evident from the analysis. The relative order of the contigs can be established by a variety of other methods. We can also map regions of the genome that have not yet been entered into the public databases, as we have a set of sequences from representations that are not yet in the published databases (a collaboration with McCombie, CSHL).

Sequences between humans and primates are sufficiently close that the same arrays we make for

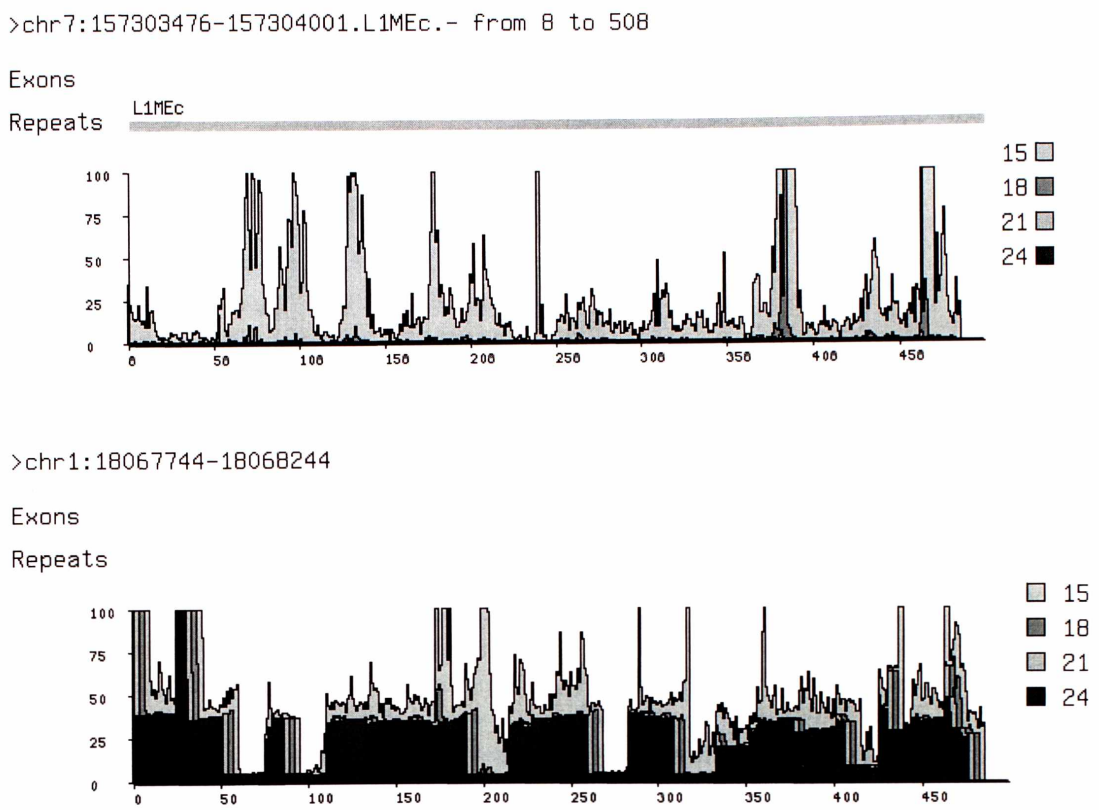


FIGURE 7 (*Top*) Region annotated as repeat, but with low mer frequencies. (*Bottom*) Reverse of the top panel, a sequence of chromosome 1 that is not annotated as a repeat, but that has high mer frequencies. The sequence in the bottom panel is found repeated about 40 times on chromosome 1 only. Frequencies of mers of lengths 15, 18, 21, and 24 are the vertical axis (see Key at right of figure). Each horizontal pixel is a single base pair.

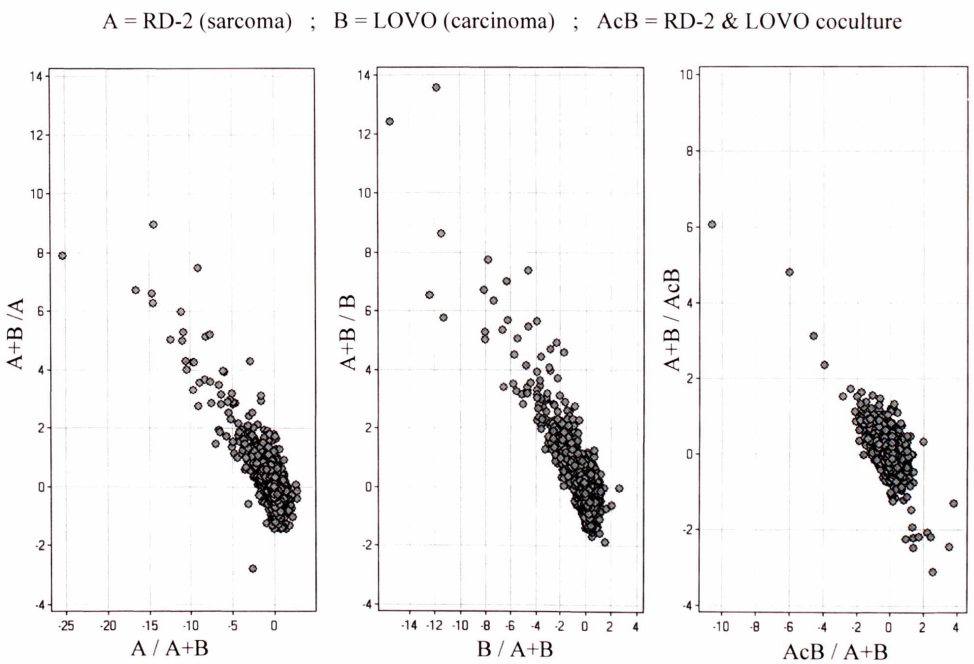


FIGURE 8 Probes that are complementary to genes that are expressed relatively abundantly in the carcinoma are in the upper left of the left panel; those complementary to genes that are abundant in the sarcoma are in the upper left of the middle panel; those complementary to genes that are induced upon coculture of the two cell lines are in the lower right of the right panel; and those complementary to genes that are repressed upon coculture are in the upper left of the same panel.

humans can be used to map primates. All that is needed are BAC libraries of the primate genomes. From such studies, the points of rearrangements between the genomes can be inferred. Perhaps of greater importance, and more accessible, the comparison of patterns resulting from array hybridization of primate and human genomes can indicate which regions of the genomes have been duplicated or deleted in the respective species recently in evolution.

GENOME ANNOTATION BY EXACT MATCHING

The oligonucleotide probes we use in ROMA are 70 mers. To maximize probe performance (i.e., specific to nonspecific hybridization), we decided to minimize the coincident exact matches between smaller "mers" of the 70 mers and other regions of the genome. We accomplished this by creating algorithms that can compute the number of exact matches for mers of any length in a given genome. We then select the best 70 mers for each member of a representation that minimize the number of exact 15-mer matches elsewhere. We demonstrated that 70 mers with frequent 15-mer matches do indeed have poor specific to nonspecific hybridization characteristics.

When we compared our method to repeat masker—a standard method for determining homology of sequences to known repeats—we observed that some mers with matches of high frequency are not annotated as repeats and, conversely, that some sequences which are annotated as repeats have low frequencies of extraneous matches in the genome (Fig. 7). We are now exploring the use of our algorithm for automated annotation of the genome and are finding new families of repeats, many of which are chromosome-specific. Many other features of gene structure and genome evolution become apparent when the genome is annotated this way, and we are continuing genomic analysis with this tool.

CELL-CELL COMMUNICATION

One of the most neglected areas of cancer biology is the study of host–cancer cell and cancer cell–cancer cell interactions. Although angiogenesis has received a lot of attention, many other host-cancer interactions will be important for tumor growth and spread. Furthermore, although most cancers are monoclonal, they are not clonal monoliths. They are genetically heterogeneous, and perhaps functionally heterogeneous and interdependent. We therefore have sought to devise a method to examine cell-cell communication, starting initially with a model system based on carcinoma-sarcoma interactions.

Through the use of expression microarrays, we may detect whether cocultivation of cells alters their joint transcriptional state. To test this idea, we profiled the transcriptional state of two colon carcinomas and one fibrosarcoma cell line, each grown independently. We then cocultivated them in all pairs and examined the joint transcriptional state. The results (see, e.g., Fig. 8) indicate that the colon carcinoma cell lines do not alter each other's transcriptional state, whereas cocultivation of the sarcoma with either carcinoma results in new transcriptional states. We observe both gene repression and induction and are in the process of determining which changes occur in which cells. The effects we observe may be mediated by soluble factors, or interaction between molecules on the surfaces of the cells, or even be the consequence of differential modulation of the culture medium. These experiments suggest that we can at least observe retention of the capacity for cancer cell communication, and they encourage us to believe that we can use this technique to catalog and correlate interactions.

PUBLICATIONS

Casey W., Mishra B., and Wigler M. 2001. Placing probes along the genome using pairwise distance data. In *Algorithms in bioinformatics* (ed. O. Gasscuel and B.M.E. Moret), vol. 2149, pp. 52–68. First International Workshops, Arhus, Denmark. Springer-Verlag, Berlin, Heidelberg, WABI, LNCS.

Scheffzek K., Grünewald P., Wohlgemuth S., Kabsch W., Tu H., Wigler M., Wittinghofer A., and Herrmann C. 2001. The Ras-Byr2RBD complex: Structural basis for Ras effector recognition in yeast. *Structure* **9:** 1043–1050.

Shenghua W., Stolarov J., Myers M.P., Su J.D., Wigler M.H., Tonks N.K., and Durden D.L. 2001. PTEN controls tumor-induced angiogenesis. *Proc. Natl. Acad. Sci.* **98:** 4622–4627.

Stolarov J., Chang K., Reiner A., Rodgers L., Hannon G.J., Wigler M., and Mittal V. 2001. Design of a retroviral-mediated ecdysone inducible system and its application to the expression profiling of the PTEN tumor suppressor. *Proc. Natl. Acad. Sci.* **98:** 13043–13048.