MAMMALIAN CELL GENETICS

M. Wigler J. Alexander, J. Allen, K. Chang, M. Chi, J. Douglas, D. Esposito, H. Grasmo-Wendler, I. Hall, J. Hicks, J. Healy, L. Muthuswamy, U. Nath, N. Navin, A. Reiner, M. Riggs, L. Rodgers, J. Sebat, E. Thomas, J. Troge, J. West

I liked the start of my section of last year's annual report so much, that I have decided to use the first two paragraphs again. I apply the adage I first heard from Sol Spiegelman, "any thing worth saying, is worth saying twice." Here it is.

It is a poor and unnecessary gamble to act as though either our theory or our knowledge of cancer is complete. Future progress in detection, prognosis, and treatment of cancer will depend on the accuracy and completeness of our understanding of its specific molecular causes. This knowledge is likely to become increasingly important as cancers, or suspected cancers, are detected at earlier and earlier stages.

There are simple tests for the completeness of our understanding of how cancers survive in and kill their hosts. If our knowledge were complete, we would see a plateau in the number of genes commonly found mutated in cancers. If the principles were few, even advanced cancers with a large number of accumulated genetic lesions would show only a small number of commonly affected pathways. It follows from this that if mutation in a single gene were sufficient to affect a given pathway, then even advanced cancers would show only a small number of commonly affected genes, the remainder of lesions being more-or-less random.

To approach the question of a "complete" understanding of cancer, we have developed a microarray-based method, called ROMA (representational oligonucleotide analysis). ROMA is based on part on our previous technique called RDA (representational difference analysis). Unlike cDNA microarrays, which can describe the "transcriptional" state of the cell, ROMA measures changes in "gene copy number" at loci that undergo amplifications and deletions, hallmarks of oncogenes and tumor suppressor genes, respectively. Although there are many other possible mechanisms that alter critical genes, such as point mutations, many if not most oncogenes and tumor suppressor genes will eventually be found in the types of lesions that we can readily detect. In principle, our method can also detect changes in the methylation of DNA, imbalanced translocations, origins of replication, and long-range features of chromatin structure.

Our basic assumption is that if a locus is recurrently found altered in cancers, that region harbors a candidate cancer gene. Therefore, the application of our method to a large series of cancers, and the comprehensive comparative analysis of such data, should reveal the position and number of candidate cancer genes in cancers. We have progressed reasonably on the task of collecting the data from cancer cells that will lead to the definition of these recurrently abnormal regions.

Using our methodology, we have also determined that there are many copy-number differences in the human gene pool, i.e., large regions of the human genome that are present in individuals in unequal amounts. These variations are germ-line, Mendelian in inheritance, distributed throughout the genome, and rich in genes. Many are common polymorphisms, found in almost equal numbers throughout the human gene pool. It seems likely to us that many of these regions are under selective pressure and will be shown to be associated with disease resistance and sensitivity. In any event, it is necessary to make a database of these variations, so as not to mistake them for cancer lesions, and we are well under way to accomplish this.

THE TECHNOLOGY

The basis of our ROMA technology has been explained over the past years. It involves making complexity-reducing representations of genomic DNA and hybridizing these representations to microarrays of oligonucleotide probes designed informatically, from the published human genome assembly, to be complementary to the representations (Lucito et al. 2003). The probes are chosen from the genome so that they have a minimal overlap with unrelated regions of the genomes. The method for making this computation was published this year (Healy et al. 2003) and has been a tool in the discovery of a new feature of mam-

malian genomes by a graduate student, Elizabeth Thomas (see below). The algorithms allow counts of exact matches of sequences of any length throughout a sequenced genome and are based on a Burrows-Wheeler transform of the genome sequence.

We use two forms of oligonucleotide microarrays: the printed form that we make ourselves, and a form in which oligonucleotides are synthesized in situ on the array surface using laser-directed photochemistry. A company called NimbleGen Systems makes the latter, and their technology has given us substantially greater flexibility in the design of arrays and the selection of representations. Pictures of these microarrays are shown in Figure 1, with the printed array on the left and the Nimblegen array on the right. We have shown that each format yields very similar measures, probe for probe. This work was conducted in collaboration with Robert Lucito here at CSHL, and a report of our findings was published this year (Lucito et al. 2003). With NimbleGen, we typically array 85,000 probes (85K format), averaging one probe per 30kb, but even greater densities can now be achieved, nearly 200,000 probes per chip. To facilitate our close relationship with NimbleGen, which fabricate their chips in Reykjavik, CSHL has set up an Icelandic subsidiary with personnel trained by our laboratory.

CANCER LESIONS

We have applied our method to both tumor biopsies and cancer cell lines and have observed gross chromosomal copy-number alterations, and highly localized amplification, imbalanced chromosome breaks, and deletions. In the latter case, we expect that we have observed both hemizygous and homozygous deletions. In our data analysis, we have used algorithms for statistical segmentation. The first version of this was designed by Adam B. Olshen and E.S. Venkatraman of the Memorial Sloan-Kettering Cancer Center (Olshen et al. 2003). Subsequent versions were designed in collaboration with Kenny Ye of the Department of Applied Mathematics and Statistics at Stony Brook University, and are based on minimization of variance and an assumption of log normal distribution of ratio data.

We observe a large number of lesions, of varying sizes, per cancer. Breast cancers appear to divide into two types: those with large numbers of genomic changes and those with very few. The former also further divide into two types: those that appear to be evolving lesions slowly and those that appear to be evolving rapidly. Virtually all of the known lesions have been seen, as well as many new ones. We are now in the stage of accumulating data and performing comparative analysis. We expect this study to give us a good estimate of the number of pathways involved in the development of breast cancer, while identifying those major pathways.

NORMAL VARIATION

We have applied our method to the comparison of normal genomes and discovered that there are a large number of extensive regions of copy-number variation

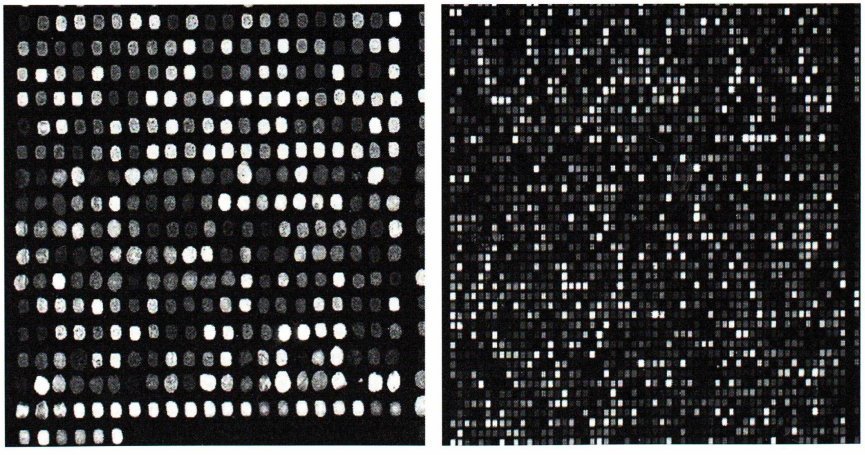

FIGURE 1 Two forms of oligonucleotide microarrays.

between any two humans. We have made 24 comparisons and have noted on the order of 240 differences at about 80 specific loci. Of these, nearly half of differences are recurrent, indicating that they arise from common polymorphisms in the human gene pool. The regions showing copy number are rich in genes and are distributed fairly uniformly across the human genome, with the exception of the X chromosome. We have confirmed these variations by polymerase chain reaction (PCR), interphase fluorescence in situ hybridizaton (FISH) (a collaboration with Anders Zetterberg of the Karolinska Institute, Sweden, and Barbara Trask, University of Seattle, Washington), and independent ROMA using *Hin*dIII (as opposed to *Bgl*II) representations. The regions arise by both polymorphic gene duplications and deletions. These regions must be categorized, as otherwise they will be mistaken for recurring cancer lesions in our cancer surveys. Furthermore, these normal variations may be associated with inherited disease susceptibility or resistance. In a collaboration with Conrad Gilliam of Columbia University College of Physicians & Surgeons, we are conducting a large survey of normal genomes and genomes from families with children with autism syndromes.

CANCER GENES

In the previous year, we have described the discovery of new oncogenes (Mu et al. 2003) and tumor suppressors (Hamaguchi et al., *Proc. Natl. Acad. Sci. 99:* 13467 [2002]) using RDA or array-based methods. Although this work is continuing by our collaborators, and we are continuing research on PTEN, we are focusing presently on the accumulation of massive amounts of copy-number data from cancer cell lines and tumors, and we have demonstrated our ability to obtain this information from clinical material archived as either frozen or formalin-fixed. We have seen promising cancer genes in tumor amplifications, encoding proteins such as kinases, transcription factors, receptors, and antiapoptotic factors and have observed interesting candidate tumor suppressor genes in regions of loss, encoding cellular components such as checkpoint control proteins and ubiquitin ligase subunits. But our general approach has been to hold off on the difficult task of functional validation until we have collected a sufficient amount of copy-number profiles to winnow the candidates and determine priorities. In future work, it will also be a priorty to correlate patterns of gene loss and gain with clinical outcomes.

GENOMES AND EVOLUTION

In collaboration with Bud Mishra and Will Casey at the Courant Institute for Applied Mathematics at New York University, we developed the algorithmic basis for the use of microarray hybridizations to map genomes (Casey et al., *Lect. Notes Comput. Sci. 2149:* 52 [2001]). Joe West and John Healy in my lab collected a full set of data in a model organism, *Schizosaccharomyces pombe*, which has a complete sequence assembly and is putting these ideas to test. We are finding that the empirical data fit the mathematical model closely, and apart from the difficulty of assembling centromeric and telomeric regions, and additional hybridization data needed because of "noise" in the system, the probes by and large map into long linear or only slightly branched structures. We predict from this work that array hybridization is a feasible way to validate a sequence assembly or to obtain a rough "local" probe map of a new organism.

Elizabeth Thomas, a Watson school graduate student, in collaboration with John Healy here at CSHL, Nathan Srebro at the Massachusetts Institute of Technology, and Bud Mishra of the Courant, has used our exact matching algorithms to discover a new and fundamental feature of genomes. Mammalian genomes are densely populated with "doublets," short duplications between 25 and 100 bp, distinct from previously described repeats. Each doublet is a pair of exact matches, separated by some distance. The distribution of these intermatch distances is strikingly nonrandom. One interesting characteristic of nearby doublets is that both exact matches tend to occur in the same orientation. By comparing doublets shared in human and chimp or mouse and rat, we can see that at least nearby doublets seem to arise by an insertion event that does not affect the neighboring sequence. Most doublets in humans are shared with the chimpanzee, but many new pairs, especially adjacent ones, arose after the divergence of the species. New doublets are most likely to be adjacent, whereas older doublets are almost equally likely to be nearby or adjacent, indicating that adjacent doublets may be unstable, disappearing over time. A genomic mechanism that generates short, local duplications while conserving polarity could have a profound impact on the evolution of regulatory and protein-coding sequences.

PUBLICATIONS

Healy J., Thomas E., Schwartz J.T., and Wigler M. 2003. Annotating large genomes with exact word matches. *Genome Res.* **13:** 2306–2315.

Lucito R., Healy J., Alexander J., Reiner A., Esposito D., Chi M., Rodgers L., Brady A., Sebat J., Troge J., West J., Rostan S., Nguyen K.C.Q., Powers S., Ye K.Q., Olshen A., Venkatraman E., Norton L., and Wigler M. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13:** 2291–2305.

Mu D., Chen L., Zhang X., See L.-H., Koch C.M., Yen C., Tong J.J., Spiegel L., Nguyen K.C.Q., Servoss A., Peng Y., Pei L., Marks J.R., Lowe S., Hoey T., Jan L.Y., McCombie W.R., Wigler M.H., and Powers S. 2003. Genomic amplification and oncogenic properties of the *KCNK9* potassium channel gene. *Cancer Cell* **3:** 297–302.

In Press

Olshen A., Venkatraman E., Lucito R., and Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* (in press).

Ira Hall