COPY-NUMBER ANALYSIS AND HUMAN DISEASE

M. Wigler
J. Alexander
M. Andrade
M. Chi
C. Danielski
D. Esposito
H. Grasmo-Wendler
V. Grubor
J. Healy
J. Hicks
L. Hufnagel
J. Kendell
A. Krasnitz
Y.-H. Lee
E. Leibu
A. Leotta
J. Meth
L. Muthuswamy
N. Navin
D. Pai
A. Reiner
M. Riggs
P. Roccanova
L. Rodgers
J. Sebat
J. Troge
H. Wang
B. Yamrom
S. Yoon

Our studies of cancers and leukemia center around a technology developed in this lab called representational oligonucleotide microarray analysis (ROMA). ROMA is a high-resolution, high-throughput method for detecting copy-number changes in the genome, such as amplifications and deletions characteristic of the mutations that drive malignancies. These changes are manifest as alterations in the genome "profile" of a cancer or leukemic cell.

GENOMIC ANALYSIS OF CANCER AND LEUKEMIA

Breast Cancer. In a long-standing collaboration with Anders Zetterberg at the Karolinska Institute in Stockholm, Sweden and Anne-Lise Børresen-Dale at the Norwegian Radium Hospital, University of Oslo, Norway, we have profiled more than 250 breast cancer tumors from separate collections of Norwegian and Swedish populations. We have reached several important conclusions. First, we observed a type of genomic rearrangement that we call "firestorms"—regions of chromosomal instability localized within chromosome arms—and we have determined which chromosomal regions are prone to this destabilization. Even a single firestorm is an especially significant prognostic marker of poor outcome. Second, the number and spacing of genomic lesions correlate very strongly with survival. We derived a mathematical measure that captures the essential geometric features of the profiles that correlate with survival in the population of pseudodiploid cancers ($p = 10^{-7}$). This is a far stronger correlation than is obtained from knowledge of any particular locus and may subsume any locus-based predictive determination. The information in this measure is independent of the location of amplifications and deletions, and of classical clinical parameters such as node status, stage, and expression of hormone receptors. Our measure will thus have clinical utility in patient evaluation.

Our data from the above studies also help to delimit the location of loci of oncogenes and tumor suppressor genes. We have worked out a variety of probabilistic and heuristic methods that sum over the data set that mark what we believe to be the most likely locations for cancer genes. We observe a similar set of loci whether examining pseudodiploid or aneuploid cancer subpopulations, even though the latter have many more lesions, suggesting that these loci confer advantage to the tumor, regardless of the degree and type of genomic instability.

The overall similarity in the location of regions of genome instability in breast cancers in Swedish and Norwegian populations is striking. These data provide a baseline from which we can seek differences in individuals of other ethnic backgrounds and geographic locations. Such comparative studies might reveal the extent of genetic and environmental contributions to the initiation and development of breast cancer or, for that matter, any cancer.

Mouse Models of Human Malignancy. In a collaboration with Robert Lucito here at CSHL, we have used our experience in designing a human ROMA copy-number array to design and test a mouse ROMA chip. This array can be used to study the amplifications and deletions that occur in mouse tumors, and that data can be compared to profiles of similar tumor types in humans. In another collaboration with Scott Powers and Scott Lowe's labs here at CSHL, this approach has been used successfully to define a common gene amplified in mouse and human liver cancers, which encodes an inhibitor of apoptosis. The mouse-human synteny relationships are used to define the common genetic elements in amplified or deleted loci. As a general approach, the profiling of human cancer genomes and the analysis of mouse models of human cancer is a powerful combination for defining the causes of this disease.

In a collaboration with David Botstein and Robert Pelham of Princeton University, New Jersey, we have used mouse ROMA to examine the "normal" stroma that grow in response to implantation of a human cancer cell line in mice. We detect clear evidence that stroma have clonal subpopulations that carry small deletions and duplications. Although preliminary, the implications could be profound: If tumors cultivate or recruit mutant stroma, a vulnerability in the cancer-host communications could be potentially disrupted, leading to novel therapeutic approaches.

Leukemia. In a collaboration with Nick Chiorazzi at North Shore University Hospital in Manhasset, New York, we have initiated a study of chronic lymphocytic leukemia (CLL), the major form of lymphoma/leukemia in adults. We use ROMA to examine copy-number changes in leukemic cells, using normal blood neutrophils from the same patient as a control. We have completed a series of 21 patient samples. In summary, 10/21 samples have a 13q deletion, and five of these deletions are homozygous. The epicenter spans DLEU7 but not mir-15a or mir-16-1, the micro RNAs (miRNAs) that others have proposed to be in the epicenter, throwing into question the validity of the hypothesis that expression of these miRNAs are important in this form of leukemia. In addition, 3/21 samples show multiple narrow deletion events and define a new class of chromosomal instability in CLL, which is not, to our knowledge, previously reported. Two new loci are the recurrent target of small-sized events, one at a gene proposed to be involved in membrane trafficking and one at a locus containing no known genes.

ANALYSIS OF HUMAN GENETIC DISEASE

Our studies of human genetic disease likewise center around ROMA. In general, with ROMA we can observe two types of phenomena: (1) spontaneous mutation that results in a copy-number change in an afflicted human and (2) an inherited polymorphism of copy number that affects the likelihood or severity of a disease. Neither of these genetic events were readily detectable by prior methodology, so ROMA, or an equivalent copy-number measurement technique, opens new possibilities for understanding forms of human disease with strong genetic components. Our particular interests lie in autism and other neuropsychiatric disorders of early onset, in a collaboration with Jonathan Sebat of CSHL; congenital heart disease with Dorothy Warburton and Wendy Chung of the Columbia UniversityCollege of Physicians and Surgeons, New York; and familial prostate cancer, in a collaboration with Bill Isaacs of Johns Hopkins University School of Medicine, Baltimore, Maryland.

As a baseline for all of these studies, we must analyze the "normal" variation of the human genome. This baseline catalog of copy-number polymorphisms (CNPs) is necessary to distinguish novel events that might be associated with disease from preexisting events that are part of normal human copy-number variations. Currently, we have identified approximately 510 unique CNPs from 500 individuals.

Distinction between Sporadic and Inherited Autism. We have established a distinction between sporadic and familial autism, in particular, we established the role of de novo mutation as a cause of autism in the former, as we originally postulated. In 90 samples from families with one autistic child but with no other history of the disease, we have observed six confirmed examples of spontaneous deletion or duplications. There may be additional examples of these, much smaller in size, that await confirmation by higher-resolution ROMA. In contrast, in the AGRE (Autism Genetic Research Exchange) set of 170 patients who derive from families with two or more affected children, we see only one possible example of a spontaneous mutation. This is on the X chromosome and is found in both affected children, but not in the parents.

We have limited data on the rate of spontaneous amplifications/deletions in normal family trios, but in 30 family trios we have found no clear examples of new mutations. We need to solidify these observations with more rigorous standards and more data, but our findings point to the conclusion that sporadic autism can result from spontaneous mutation.

This observation has an important implication for how the community of genetic researchers proceed in the discovery of mutations that cause autism. Trios of mother, father, and child, where there is no other affected child or history of autism in the family, may be the preferred population in which to search for spontaneous mutations. Unfortunately, most of the community's effort has been directed to the collection of families with more than one affected child, where transmission genetics has a greater role.

Protective Variation in Autism. In familial autism, we have preliminary evidence for the possible involvement of relatively common alleles in the penetrance of the disorder. In particular, our data suggest the involvement of the *CHRNA7* locus encoding the nicotinic acetylcholine receptor $\alpha 7$. This locus is of interest because of its suspected involvement in schizophrenia and attention deficit disorders. Our data indicate that this region is hypervariable in the human population and may thus be under strong selective pressure. Duplication at this locus is observed in individuals from all populations, including sporadic cases of autism, at rates about 15%, but in the AGRE set, the frequency of duplication is much lower, around 5%. These results suggest that the presence of duplication at this locus might protect against autism of the inherited variety.

This hypothesis is important in several respects. If correct, it suggests that in some cases, pharmacological approaches might alleviate development of the disorder. Second, it may provide mechanistic insight into the nature of the disease. Third, the hypothesis raises hope that other genetically linked disorders could yield to similar approaches. Important hypotheses require extremely careful analysis, and far more work than what we have completed thus far is needed to confirm this observation.

ADVANCES IN ROMA TECHNOLOGY

ROMA offers a fresh perspective on human disease, but its potential is still largely untapped. Improvements to the methodology in terms of resolution, reliability, data access, and cost are a continuing and challenging activity within our group.

New Designs and Applications. We have successfully developed and tested a high-resolution 390,000 probe ROMA chip, with more than four times the probe density of our previous 85,000 probe chip. With this microarray, we can see cancer lesions with far greater clarity than before, which should aid greatly in determining the location of genes causing cancers and leukemias. Work remains to modify the processing software that we have developed for the 85,000 array, including normalization and segmentation protocols. (Normalization is a method for standardizing interpretation of raw hybridization signals, and segmentation is a method for interpreting the hybridization data as alterations of chromosomal loci.) We have also developed methods that should enable the calibration of each probe on the array, which we predict will dramatically improve our segmentation methods and also help in selecting probes as we design new arrays in the future.

Additional statistical methods now allow us to perform hybridizations in single color, minus a reference sample. The ability to perform hybridization in one color, the omission of a reference sample, and the ability to reuse arrays after high-stringency washing reduces the consumption of arrays and other reagents and reduces costs almost twofold. Finally, we have established that we can obtain reliable data from tissue freshly fixed in formalin, which is the way in which breast and most other cancer biopsies are routinely prepared for clinical-pathological appraisal. All of these developments should facilitate the adoption of our methods in a clinical setting.

Data Access. ROMA generates raw data files on the order of megabytes per sample, and each sample is associated with critical biological information. The "accessibility" of our data has been a nagging problem. We have assays on over 2000 samples in our collective group (cancer and genetics), and each sample may have several dozen important associated clinical facts. Each assay may yield a table 40 x (85 x 10^3), and four times that for the high-resolution arrays. The raw data undergo a variety of processing steps and are stored in a variety of workspaces and formats distributed throughout the dispersed computers in four different geographic centers and two continents. Accessing all of this information in a seamless manner has been a trying experience. Our first implementation as a large centralized relational database of raw data was only a partial success, because it did not reference properly either sample information or the processed data files.

We have therefore taken two steps. The first is to implement an "Ark." The Ark contains all of the processed data files in a logical folder architecture on a central server as ASCII text, with specially designed "headers" in each file that identify and describe the internal structure of each file. Each folder contains an index to its contents in a standardized "language." We have developed tools for the rapid integration of this data back into the working directories of our numerical processing environment for comparative analysis, for graphical display, or as a jumping-off point to the Internet.

The second step has been to develop a customizable database language. It is a hybrid between object-oriented and relational database languages. Called "Pentuple," it uses something akin to natural language—a flexible linguistic structure with hierarchical permissions to alter that structure, allowing a user group to enter, organize, integrate, and search very large amounts of data. We hope that Pentuple will lead to a general purpose multiuser archival language system.

PUBLICATIONS

Jobanputra V., Sebat J., Troge J., Chung W., Anyane-Yeboa K., Wigler M., and Warburton D. 2005. Application of ROMA (representational oligonucleotide microarray analysis) to patients with known cytogenetic rearrangements. *Genet. Med.* **7:** 111–118.