COPY-NUMBER ANALYSIS AND HUMAN DISEASE

M. Wigler C. Algieri, T. Baslan, M. Bekritsky, S. Chakraborty, K. Cook, Y. Eberling, D. Esposito, V. Grubor, I. Hakker, J. Hicks, J. Huang, R. Kandasamy, J. Katz, J. Kendall, A. Krasnitz, Y.-H. Lee, A. Leotta, D. Levy, S. Marks, J. Meth, N. Navin, M. Oswald, D. Pai, M. Riggs, L. Rodgers, M. Ronemus, J. Troge, Z. Wang, B. Yamrom

Mammalian Genetics

We study variations in the human genome and their association with disease and genetic disorders. The variations we study arise when a large segment of the genome is duplicated or deleted (Lucito et al., *Genome Res 13:* 2291 [2003]; Sebat et al., *Science 305:* 525 [2004]). Such "copy-number" variations, or CNVs, can arise somatically or in the germline. Somatic variations are often seen in cancer and distinguish cancer cells from the normal cells of the body. They provide clues for the origin and behavior of the cancer (Hicks et al., *Genome Res 16:* 1465 [2006]) and possibly its early detection. Analysis of cancer genomes at the single-cell level will enable us to better study how cancers evolve (Navin et al., *Genome Res 20:* 68 [2010]) and provide new tools for clinical evaluations of many sorts. Germline variations distinguish individuals from each other and may be inherited, in which case, they are known as copy-number polymorphisms, or CNPs, or they may arise spontaneously, in which case, they serve as engines of human diversity and can sometimes cause devastating genetic disorders, such as autism (Sebat et al., *Science 316:* 445 [2007]). We have formulated a hypothesis unifying spontaneous and inherited mutation in the etiology of autism (Zhao et al., *Proc Natl Acad Sci 104:* 12831 [2007]). Much of the lab is dedicated to devising methods for data interpretation and building quantitative genetic models.

Cancer and Leukemia

We use copy-number data and DNA-methylation status to study breast cancer (Hicks et al., *Genome Res 16:* 1465 [2006]) and B-cell chronic leukemia (Grubor et al., *Blood 113:* 1294 [2008]). We seek to identify the loci most frequently mutated in cancers and leukemias, and among them to determine which might be causative (Zender et al., *Cell 125:* 1253 [2006]; Xue et al., *Genes Dev 22:* 1439 [2008]; Zender et al., *Cell 135:* 852 [2008]; Bric et al. 2009), and to determine if genomic data can be used to predict the outcome of the disease (Hicks et al., *Genome Res 16:* 1465 [2006]), its response to therapy (McArthur et al. 2009), and the early detection of its recurrence. In addition to assessing the role of copy-number mutation in cancer etiology and outcome, we have developed methods to assess the role of DNA methylation changes (Hodges et al. 2009; Kamalakaran et al. 2009; S Kamalakaran et al., in prep.). A particular emphasis is using copy-number data to assess the population substructure of tumors (Navin et al., *Genome Res 20:* 68 [2010]). We have had success with single-cell genome sequence analysis, a tool that shows extensive promise for applications in clinical and basic cancer biology (N Navin et al., in prep.). All of the above studies are collaborations with scientists at CSHL (Powers, Lowe, McCombie, and Hannon labs) and at other institutions.

Genetic Disorders

After our discovery that copy-number variation is common in the human gene pool (Sebat et al., *Science 305:* 525 [2004]), we studied the role of CNVs in human disease and, in particular, the role of spontaneous (or de novo) germline CNVs. Our findings established that germline mutation is a more significant risk factor for autism spectrum disorders (ASD) than previously recognized (Sebat et al., *Science 316:* 445 [2007]) and established a new approach for the further study of the genetic basis of this and other genetic disorders. We also study the role of spontaneous mutation in congenital heart disease (a collaboration with Dorothy Warburton at Columbia University), rheumatoid arthritis (with Peter Gregersen at North Shore University Hospital), and pediatric cancers (with Ken Offit at Memorial Sloan-Kettering Cancer Center).

One of the de novo events we identified in autism was a deletion on 16p (Sebat et al., *Science 316:* 445 [2007]). This event has now been shown by two other groups to explain perhaps as much as 1% of autism. We assisted Alea Mills at CSHL to engineer mice with the orthologous deletion on mouse chromosome 7, and she has continued to search for phenotypic consequences. We are hopeful that these mice will provide animal models suitable for understanding the underlying neuropathology of the condition and the search for palliative treatments.

Analysis of autism incidence in families, a collaboration with Kenny Ye at the Albert Einstein School of Medicine, provided evidence for a unified theory of the genetic basis for the disorder (Zhao et al., *Proc Natl Acad Sci 104:* 12831 [2007]). Autism families are divided into simplex (only one affected child) and multiplex (multiply affected children). By inspecting the records from the AGRE consortium, we found that the risk to a male newborn in an established multiplex family is nearly 50%, the frequency expected of a dominant disorder. Autism incidence and sibling concurrence rates are consistent with a model in which new or recent mutations with strong penetrance explain the majority of autism in males and are consistent with a one-hit event.

We are now in the midst of a larger study of spontaneous mutation in autism, based on a population of simplex families collected by the Simons Foundation. This collection is of high-functioning children, and it has a 7:1 bias of males to females. Early initial results confirm our previous findings, and we observe de novo (copy-number) mutation more frequently in children with autism than in their unaffected siblings. The statistical evidence is strong for deletion events, but much weaker for amplifications, an assessment that was not possible before because of lack of statistical power. Because our new studies are performed with higher-resolution microarrays, we also see many more examples of narrow new mutations (altering only a few genes), thus expanding our list of good candidate genes involved in the disorder. There is a male bias to the detection of narrow mutations, but we see little gender bias for broad mutations (altering many genes). Because there should be no gender bias in the incidence of new mutation, the detection biases suggest to us that at least two contributory genes are targeted in the broad lesions, equalizing susceptibility in males and females.

The data deepen the mystery of the male bias in autism. We are developing testable theories, including the possibility that the pattern of monoallelic expression is different in males and females during early development. Regardless of the model, if females require two-hit events, it enables us to estimate the proportion of autism that is one-hit in males. It is (the incidence in males minus the incidence in females) divided by the incidence in males.

Our study based on copy number does not pinpoint the genes that cause autism, because even the narrow events typically contain multiple genes. Pathway analysis, performed in collaboration with Ivan Iossifov at CSHL, does suggest a plausible set of interrelated genes. We are now pursuing our leads by sequence analysis of trios (mother, father, and child) from the Simons collection (a collaboration with the McCombie Lab at CSHL). We are conducting a search for de novo point mutations that disrupt function in our candidate genes. From our unified hypothesis, knowledge of the rate of de novo mutation in the germline, and the rate of autism in males, we estimate that there are on the order of 300–500 autism genes. In the 1000+ trios we expect to sequence, we predict to see a signal in the form of recurrent mutations only in the actual autism genes.

Data Generation, Analysis, and Quantitative Modeling of Genetic Process

A major part of our group's effort centers on the generation, analysis, and interpretation of high-volume data. This includes developing protocols for handling microarray copy-number data, determining quality control, probe evaluation, signal extraction, and segmentation (the method of "observing" copy-number variation); comparisons of sets of experiments, including new statistical measures, data reduction, and data summary; as well as construction of databases so that we can communicate our results to others. Our novel contributions include methods to attenuate system noise in array hybridizations, parameterize hybridization performance, detect and correctly call regions of genetic polymorphism, detect de novo events, classify cancers for outcome, and define the epicenters of genetic change in cancers and leukemias. We are now also using high-throughput DNA sequence data to measure mutation rates. This entails becoming familiar with and developing new methods for data analysis. Our first "product" in this line is a method to correctly call copy number from the DNA sequence read density of single-cell genomes.

PUBLICATIONS

Bric A, Miething C, Bialucha CU, Scuoppo C, Zender L, Krasnitz A, Xuan Z, Zuber J, Wigler M, Hicks J, et al. 2009. Functional Identification of tumor-suppressor genes through an in vivo RNA interference screen in a muse lymphoma mode. *Cancer Cell* **16:** 324–335.

Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, et al. 2009. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* **19:** 1593–1605.

Kamalakaran S, Kendall J, Zhao X, Tang C, Khan S, Ravi K, Auletta T, Riggs M, Wang Y, Helland A, et al. 2009. Methylation detection oligonucleotide microarray analysis: A high-resolution method for CpG island methylation. *Nucleic Acids Res* **37:** e89.

McArthur HL, Tan LK, Patil S, Wigler M, Hudis CA, Hicks J, Norton L. 2009. High resolution representational oligonucleotide microarray analysis (ROMA) suggests that TOPO2 and HER2 coamplification is uncommon in human breast cancer. *Cancer Res* **69:** 2023.