GENOMIC ANALYSIS OF CANCER AND HUMAN GENETIC DISORDERS

M. Wigler
J. Hicks
J. Alexander
P. Andrews
T. Auletta
T. Baslan
M. Bekritsky
H. Cox
D. Esposito
E. Grabowska
I. Hakker
Q. Huan
J. Huang
R. Kandasamy
J. Kendall
Y.-H. Lee
A. Leotta
D. Levy
B. Ma
S. Marks
J. McIndoo
N. Navin
M. Riggs
L. Rodgers
M. Ronemus

The Wigler and Hicks group studies human cancer and genetic disorders from a population genomics perspective. The cancer effort focuses on breast and prostate cancer (the latter jointly with Lloyd Trotman) and involves collaborative clinical studies (with local, national, and international collaborators) to discover mutational patterns predicting treatment response and outcome. We also develop methodology for single-cell genomic and RNA analyses to detect cancer cells in bodily fluids such as blood and urine. This last has major potential applications to the early detection of cancer and monitoring its recurrence and response to therapy during and after treatment. The single-cell analysis has also led to insights about the clonal evolution and heterogeneity of cancers (Navin et al. 2011). This work may lead to a better understanding of initiation, progression, and metastasis and shed light on the stem cell hypothesis of cancer and host responses.

Our genetic efforts are largely focused on determining the role of new (de novo) mutation in pediatric disorders with a strong genetic contribution. We study autism as part of a large study of simplex families organized by the Simons Foundation (Fischbach and Lord 2010), as well as congenital heart disease with Dorothy Warburton of Columbia University and pediatric cancers with Ken Offit of Memorial Sloan-Kettering Cancer Center (MSKCC). In collaboration with Alea Mills, we helped create a mouse model for one of the most common genetic abnormalities contributing to autism (Horev et al. 2011). Recent work has confirmed and extended our previous observations on the role of de novo copy-number variation (large-scale mutation) in autism (Levy et al. 2011), with similar results in the other disorders. In a large-scale exome sequencing project in collaboration with Dick McCombie (at the Genome Center at CSHL) and the Genome Sequencing Center at Washington University in St. Louis, we have proven the role of small-scale de novo mutations that disrupt genes in autism (Iossifov et al. 2012). Overall, our results succeed in confirming our previous genetic models for autism incidence and identify strong candidate mutational targets, from which biological models of autism can be made and tested.

Breaking with tradition, the remainder of this report contains two position statements on the direction in which we are taking our science, and why.

Applications of Single-Cell Analysis to Biological and Medical Problems

For several years, we have explored methods for gathering and analyzing nucleic acid sequences from single cells. There are at least five broad areas in which single-cell methods can be applied: (1) cancer, (2) neurobiology, (3) disorders of stem cell renewal, (4) detailed dissection of cell-state transitions, and (5) genome assembly. We are working on each. All of these applications require improvements in the gathering of single-cell data, both its quality and cost, and developing new processing algorithms. Some of the mathematical challenges are new. To improve data generation, we make use of DNA barcoding to facilitate massively parallel sample processing, increase the uniformity of data quality, and reduce costs.

The applications of single-cell analysis in cancer are nearly endless. Single-cell genome analysis can be used for detecting cancer cells in samples such as blood or urine obtained by noninvasive or minimally invasive procedures. A more thorough analysis of standard biopsy material, such as needle biopsies, can be achieved by single-cell genome profiling, revealing details of cancer heterogeneity, subpopulation structure, host reaction, and the presence of cancer cells in surgical margins and lymph nodes. Through single-cell RNA expression, the host response at the sites of cancer growth can be monitored: the presence of capillary endothelium, immunocytes of all types, and reactive stroma. The identity of soluble factors to which these cells are responding may be inferred by observing their response signatures. By means of statistical methods correlating clinical outcomes with single-cell observation, we shall be able to improve detection, prognosis, therapeutic choice, and monitoring response to therapy. The most transformative application would be a blood test for early detection of cancer onset, where the

DNA analysis provides evidence for the presence of cancer cells and disease stage, and the RNA analysis indicates the cancer's tissue origin and subtype. Both together can aid in prognosis and choice of therapeutic intervention. Finally, many fundamental questions of cancer biology can be addressed by single-cell analysis. Among these: How do cancer populations evolve? Is there in fact a stem cell population that is genetically distinct from the majority of the tumor population? From what cell or combination of cells do metastases arise? To what extent and through what processes do cancer cells cooperate?

Single-cell expression analysis can also be used to develop a deeper understanding of neural and nonneural subtypes in the nervous system (brain, chord, and gut) where tissue heterogeneity obscures bulk analysis. The result can be a better modeling of neuronal networks through identification of neuronal subtypes and a better understanding of the response of neuronal cells to stimuli. The role of somatic mutation in nervous system disease can be explored. Finally, these methods can help refine our knowledge of the fates of pluripotent stem cells induced to differentiate into neuronal cells. If successful, we could explore the functional consequences of candidate gene mutations.

The identity of stem cells and the role of aberrant stem cells in medical disorders are still highly unexplored. By an iterative process coupling cell separation and regenerative assays, single-cell RNA analysis can help identify which cells in a tissue or in distant tissues are in fact the cells with regenerative capacity. Several medical disorders, particularly those of aging, may be the result of somatic mutation in stem cells causing the replacement of critical cell types over time with dysfunctional versions. The types of disorders that might be caused by this mechanism include a wide variety of neurodegenerative conditions (e.g., Alzheimer's), disorders of autoimmunity (e.g., lupus), cardiomyopathies, and a variety of cutaneous (e.g., psoriasis) and connective tissue (e.g., scleroderma) diseases. The mechanisms underlying these disease processes might become apparent by combining DNA and RNA analysis of single cells, looking for aberrant genomes or expression patterns in stem cells. The latter may be identified at the single-cell level in complex organs, thereby overcoming the problems that beset the analysis of mixed-cell populations.

Examples of cell-state transitions abound in all living systems: the progression through the cell cycle, response to nutrient opportunity, differentiation, responses to growth factors and other signaling molecules, and even responses to contact with other cells. Often, these transitions involve a causal sequence of changes in the expressive state of the cell, the proteins and RNAs that they make. Because these changes typically occur in cell populations that are not in temporal synchrony, the study of the sequence of changes often cannot be determined with precision. Single-cell analysis offers one way around this problem, as each cell represents a snapshot in time at the point when it is destroyed and converted to an ensemble of macromolecules. In principle, the overall series of changes can then be assembled into a coherent whole provided that (1) there is some overlapping signature in the macromolecular composition between time-adjacent states and (2) the temporal series is relatively similar from cell to cell.

We can make sequence libraries from a fraction of a cell genome. Genome analysis itself can be improved by doing this. If the DNA of a cell is diluted sufficiently into isolated "packets," any contiguous region we sequence from a given packet is likely to represent only one of the two parental haplotypes. If the DNA is not broken into small pieces during dilution, but diluted in large blocks—even if we subsequently break the DNA into small pieces when we make libraries—we can reassemble the short-read sequences into large haplotype blocks. The result, if conducted efficiently using many cells as a starting point, would ultimately be a haplotypic assembly of the genome of any organism at roughly the cost of sequencing the entire organism in the conventional manner. As a practical matter, this converts the most efficient sequencing apparatus currently available, which is a short-read apparatus, into a long-read single-molecule apparatus.

Genetic Models of Autism

A large proportion of our lab works on autism genetics. The purpose of these notes is to give some scientific perspective to what the group has achieved and the future direction it will take. The rewards of studying autism from a genetic point of view are great. The disorder affects some of the most profoundly human traits. A complete analysis may reveal new sources of phenotypic variation. A good set of targets provides guides to therapies. And there are unmet needs for early detection and counseling families and individuals coping with the condition.

It was evident to some by as early as the mid 1980s that autism was likely to be a disorder caused by new mutation. The disorder was variable, hinting at multiple underlying causes; the incidence was high; the her-

itable component was appreciable but low; but the concordance in identical twins was higher than in any other cognitive or behavioral impairment. Human geneticists had failed repeatedly to produce evidence of heritable underlying events, except in the limited cases of X-linked disease. In both of these cases, Fragile-X syndrome (FXS) and Rett syndrome, new or recent de novo mutations were clearly the root cause. Geneticists using transmission genetics would explain their failure under the rubric of a "complex" disorder, caused by subtle interactions of multiple genes that would therefore be hard to crack. But what we knew in fact pointed to the existence of singular events of strong penetrance. Cytologic analysis had showed that many rare large-scale chromosomal events could result in developmental anomalies. Cytology could only scratch the surface of the richness of the consequences of new mutation.

In the early 2000s, methods for examining copy-number variation (CNV) developed in our lab (Lucito et al. 2003) became very powerful when combined with the data of the human genome sequence assembly (Healy et al. 2003). This led to the discovery of widespread CNV in the human gene pool (Sebat et al. 2004) and the hope that we could demonstrate autism was associated with new mutations likely to disrupt the dosage of functioning genes. Supported by the Simons Foundation, we succeeded in showing this (Sebat et al. 2007), and our work fed the acceptance of the idea that disorders associated with drastically reduced fecundity would be often caused by new mutation through the action of gene dosage effects. Altering the functional dosage of certain genes, whether by duplication, deletion, or disruption, would manifest upon transmission as dominant traits, and we sought evidence for this in the AGRE data of multiplex families. We found the evidence (Zhao et al. 2007), namely, that boys born to parents with two previous offspring with autism spectrum disorder (ASD) had a 50% chance of being on the spectrum. These observations were confirmed in independent studies by ourselves and others.

Mathematical modeling of the family risk function using the AGRE data, and data from other sources, led to several predictions (Zhao et al. 2007). Up to half of autism might be explained by new mutation, and a large part of the remainder would be due to transmission of strongly penetrant variants carried by asymptomatic parents. We speculated that most often the carrier parent would be the mother.

Subsequent work has focused on defining the list of autism target genes and developing a more quantitative genetic model. We seek a list of target genes in the expectation that such a list would provide medical geneticists with improved tools for diagnosis, especially early diagnosis, and also yield insights into physiologic mechanisms and thus ideas for intervention. Our most recent studies of autism were based on the Simons Simplex Collection, an unparalleled sample set, and are published in *Neuron* (Gilman et al. 2011; Iossifov et al. 2012; Levy et al. 2011). These papers provide a leap forward in assembling the list of gene targets (Gilman et al. 2011), validated the role of de novo CNV in autism (Levy et al. 2011), and provided evidence of a roughly equal role of transmission of rare CNV. Levy et al. (2011) provided evidence that the autism candidate loci contained a functionally convergent network of genes, and they extracted a list of the most likely autism candidate loci from the CNV data. Iossifov compared sequence data within families to find compelling evidence that de novo mutations which disrupt genes contribute to autism. Many of the findings of Iossifov are found in three smaller studies published essentially concurrently (Neale et al. 2012; O'Roak et al. 2012; Sanders et al. 2012). Recurrence analysis confirmed previous estimates that there are on the order of 200–800 (most likely, 350–400) dosage-sensitive genes that when disrupted can contribute significantly to autism. Most significantly, the list of genes showed a strong overlap with genes that encode proteins whose translation may be under the control of FMRP, the product of the gene responsible for fragile-X syndrome (FXS). Nearly half of autism dosage-sensitive genes may be so controlled. FMRP is one of the central regulators of synaptic plasticity, the physiologic mechanism underlying the response of neural networks to repetitive stimuli.

In addition to producing an extensive list of well-vetted autism gene targets, we have focused on obtaining a better, more detailed quantitative genetic model for autism incidence. What answers would such a model provide, and why would they be useful? A detailed quantitative model would determine with greater precision the number of dosage-sensitive gene targets and determine more precisely the overall role of these in incidence. Without such information, it will be hard to know how much of causation we are missing, and whether we need to consider extragenic mechanisms to explain phenotypic variation. Without such a model, we would be unaware of whether we were dealing competently with the problem. An adequate model would explain the role of gender bias in transmission, and whether certain targets are gender-specific. Properly understood, gender bias might guide thinking about ther-

apy. At the present time, there are gaps in the evidence supporting any model or mixture of models. Filling those gaps either with further evidence or by correcting the models will be valuable contributions to our understandings of genetics and cognition.

LITERATURE CITED

Fischbach GD, Lord C. 2010. The Simons Simplex Collection: A resource for identification of autism genetic risk factors. *Neuron* **68:** 192–195.

Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. 2011. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70:** 898–907.

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

Horev G, Ellegood J, Lerch JP, Son YE, Muthuswamy L, Vogel H, Krieger AM, Buja A, Henkelman RM, Wigler M, et al. 2011. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci* **108:** 17076–17081.

Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, et al. 2012. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74:** 285–299.

Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, et al. 2011. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70:** 886– 897.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, et al. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res* **13:** 2291–2305.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* **472:** 90–94.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, et al. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485:** 242–245.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, et al. 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485:** 246–250.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, et al. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485:** 237–241.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445– 449.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K, Law P, Qiu S, Lord C, Sebat J, et al. 2007. A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci* **104:** 12831–12836.

PUBLICATIONS

Floer M, Wang X, Prabhu V, Berrozpe G, Narayan S, Spagna D, Alvarez D, Kendall J, Krasnitz A, Stepansky A, et al. 2010. A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell* **141:** 407–418.

Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. 2011. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70:** 898–907. PMID: 21658583.

Horev G, Ellegood J, Lerch JP, Son YE, Muthuswamy L, Vogel H, Krieger AM, Buja A, Henkelman RM, Wigler M., et al. 2011. Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci* **108:** 17076–17081. PMCID: 3193230.

Kamalakaran S, Vinay Varadan V, Giercksky Russnes HE, Levy D, Kendall J, Janevski A, Riggs M, Banerjee N, Synnestvedt M, Schlichting E, et al. 2011. DNA methylation patterns in luminal breast cancer differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Mole Oncol* **5:** 77–92. PMCID: 21169070.

Lee Y-H, Ronemus M, Kendall J, Lakshmi B, Leotta A, Levy D, Esposito D, Grubor V, Ye K, Wigler M, et al. 2011. Reducing system noise in copy number data using principal components of self-self hybridizations. *Proc Natl Acad Sci* **109:** E103–E110.

Levy D, Ronemus M, Yamrom B, Lee Y-H, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, et al. 2011. Rare de nova and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70:** 886–897. PMID: 21658582.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. 2011. Tumor evolution inferred by single cell sequencing. *Nature* **472:** 90–94. PMID: 21399628.

Navin N, Krasnitz A, Rodgers R, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge J, Grubor V, et al. 2010. Inferring tumor progression from genomic heterogeneity. *Genome Res* **20:** 68–80. PMCID: PMC2798832.

Png KJ, Yoshida M, Zhang XH, Shu W, Lee H, Rimner A, Chan TA, Comen E, Andrad VP, Kim SW, et al. 2011. MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer. *Genes Dev* **25:** 226–231.

Russnes HG, Navin N, Hicks J, Borressen-Dale A-L. 2011. Insight into the heterogeneity of breast cancer through next generation sequencing. *J Clin Inv* **121:** 3810–3818.

Russnes HG, Moen Vollan HK, Lingjærde OC, Krasnitz A, Lundin P, Naume B, Sørlie T, Borgen E, Rye IH, Langerød A, et al. 2010. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Science: Transl Med* **2:** 38ra47. PMCID: 20592421.

Shakya R, Reid LJ, Reczek CR, Cole F, Egli D, Lin CS, DeRooij DG, Hirsch S, Ravi K, Hicks JB, et al. 2011. BRAC1 tumor suppression depends on BRCT phosphoprotein binding, but not its E3 ligase activity. *Science* **334:** 525–528.

Tafe LJ, Janjigian YY, Barbashina VV, Kelsen DP, Ilson DH, Tang LH, Hicks JB, Shah MA. 2011. Human epidermal growth factor receptor 2 testing in gastro esophageal cancer: Correlation between immunohistochemistry and fluorescence *in-situ* hybridization. *Arch Pathol Lab Med* **135:** 1460–1465.

In Press

Baslan T, Kendall K, Rodgers L, Cox H, Riggs R, Stepansky A, Troge J, Kandasamy R, Esposito D, Lakshmi B, et al. 2012. Genome wide copy number analysis of single cells. *Nat Protocols* **7:** 1024–1041.

Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee Y-H, Narzisi G, Leotta A, et al. 2012. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74:** 285–299.

Xue W, Kitzing T, Roessler S, Zuber J, Krasnitz A, Schultz N, Revill K, Weissmueller S, Rappaport AR, Simon J, et al. 2012. A cluster of cooperating tumor suppressor gene candidates in chromosomal deletions. *Proc Natl Acad Sci* (in press).