RESEARCH PROFILE

W. Richard McCombie

Advances in sequencing technology are revealing the genetic basis of human disease

In a cavernous room within CSHL's Woodbury Genome Center, 16 machines somewhat larger than the average dishwasher are cataloging the sequence in which the four DNA "letters" or chemical bases—A, T, G, C—appear in an organism's genetic code, or genome. At the moment, some machines are sequencing the genomes of prostate cancer patients; others, the genomes of people with bipolar disorder and depression. One is even sequencing the genome of wheat, which at six times the size of the human genome presents quite a decoding challenge.

As the overseer of these projects and a few others, Professor W. Richard McCombie is CSHL's sequencer-in-chief. He is at the leading edge of a technological front, which in the last few years has begun moving so fast that it's "allowing us to think of questions and study them in ways that weren't feasible even a month ago," he enthuses. As a result, his team is forging ahead in two ways: pushing "next-generation" sequencing technology to spell out new genomes faster, cheaper and with greater accuracy; and using this data to advance our understanding of the role of genetic and epigenetic variations in cancer and cognitive disorders such as schizophrenia and bipolar disorder.

These variations—person-to-person differences in DNA and its chemical tags, respectively—play a role in whether an individual has a higher or lower risk of getting a particular disease. When McCombie and his colleagues have crunched through the millions of megabytes of DNA data generated by those 16 machines, their findings will help unlock a universe of information critical for improving human health.

The incredible power of genomics

McCombie's interest in disease-related genetics goes back to the early 1990s when he joined the National Institutes of Health (NIH) after a brief stint in the biotechnology industry. Trying different approaches to understand the genetics of Huntington's disease, "We still failed to get very far, despite being at the cutting edge of sequencing," McCombie recalls.

And cutting edge it certainly was. McCombie's boss at the NIH was J. Craig Venter, who would soon embark on the race that led to the first complete human genome sequence. Initially tasked by Venter to improve biochemical and molecular methods of studying cellular receptors in the brain's neurons, McCombie soon veered toward DNA sequencing. A 1982 paper that he had read as a Ph.D. student at the University of Michigan had etched a deep impression on him.

The paper, by Frederick Sanger, who won his second Nobel Prize for inventing DNA sequencing, laid out the DNA sequence of a virus called lambda. "The paper explained a huge amount of the virus's biology by just referring to its sequence," McCombie remembers. "It struck me that having a genomic sequence gave one an incredible power to understand an organism's biology and link its genetics to any problem." But he also realized that sequencing at the time was "neither fast nor large-scale enough to really make an impact."

By the time CSHL recruited him in 1992, McCombie had made significant headway at the NIH in addressing both of these drawbacks. He led one of the first groups ever to carry out automated sequencing of genomic DNA on a major scale. And he helped Venter organize the first large-scale project involving Expressed Sequence Tags (ESTs)—tools that have since been used to identify thousands of genes and predict their function.

Illumina sequencing machines in Richard McCombie's lab are decoding genomes of patients with cancer and cognitive disorders.

A decade of genomic "firsts"

These triumphs would serve McCombie well at CSHL. At a party for CSHL's Richard Roberts, who had just won a Nobel Prize, a chance conversation between McCombie and plant geneticist Rob Martienssen about sequencing triggered a hugely successful partnership. In 1996, with funding from the National Science Foundation, the duo formed one of three teams in the country that began to sequence the first plant genome —of *Arabidopsis thaliana* (mustard plant), a workhorse in plant genetics labs.

Propelled by a change in technology—from "gel-based" sequencing to "capillary-based" sequencing—the team finished the project in 2000, two years ahead of time. Around the same time, as capillary sequencing "leapt ahead by a couple of generations," a CSHL group led by McCombie joined an international research consortium to sequence the mouse genome. And in 2001, team CSHL made history as one of 20 groups to collectively publish the first draft of the human genome.

McCombie's next challenge was harder: sequencing crops like corn and rice. Not only are their genomes much larger than that of humans, they are stuffed with chunks of so-called repetitive DNA that are difficult sequence and contain few genes. With funding from the U.S. Department of Agriculture, McCombie and Martienssen developed a method—called methylation filtration—to capture only the gene-rich regions. With this shortcut, they rapidly sequenced part of the corn genome in 2003 (its complete genome was published in 2009). They also helped sequence the entire genome of rice in 2005.

Finding the mutations that cause disease

That's when, McCombie says, technology changed yet again, advancing genomics by a factor of 10 annually (in contrast to computing, which according to Moore's law, advances by a factor of 1.5 per year). Invited to sit in on a meeting between then-CSHL President and DNA pioneer Dr. James D. Watson and the founders of 454 Life Sciences, a company that wanted to sequence Watson's genome, McCombie recognized the 454 platform as a potential game-changer.

Realizing that "we were about to enter a different world in which sequencing could help find the causes of disease," he began to think about ways to extract and sequence small stretches of the human genome; for example, regions that actually code for protein (now known to be about 2% of the genome) and so are likely to harbor mutations that contribute to disease. The result was a revolutionary method that McCombie and CSHL molecular biologist Gregory Hannon developed in 2007 called targeted resequencing.

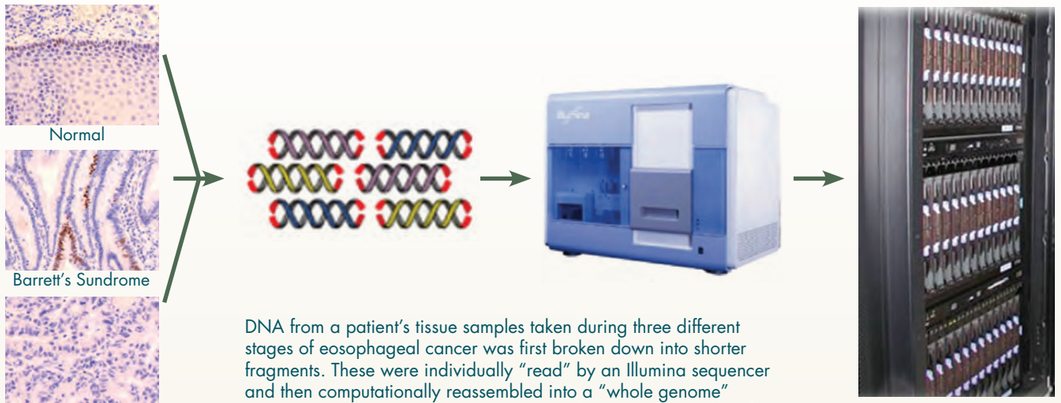
This method has made it possible to sequence and compare genomes, at low cost, of large groups of people, which is key to unearthing insights about disease-causing mutations. CSHL scientists have used it to home in on cancer-related genes as well as study entire "exomes," the totality of coding regions within a genome.

With support from the Starr Foundation, McCombie's group and their collaborators at Memorial Sloan-Kettering Research Center in New York are now examining tumor cells in the blood of prostate cancer patients on chemotherapy to find genetic markers that determine response to therapy. They are also studying, sequencing and analyzing genomic DNA from a patient with eosophageal cancer from samples taken before, during and after tumors grew [image above]. "Studying the genetics of cancer progression this way wasn't feasible even a couple of years ago," says McCombie.

These technological innovations have dovetailed with a dramatic surge in sequencing power at CSHL. The acquisition of next-generation sequencing platforms has boosted output from 78 million DNA bases per month in 2000 to 2.5 trillion bases per month in 2011. The costs of whole genome sequencing are plummeting too. Between 2008 and 2011, there were two 50% drops each year. In February 2011, it cost CSHL $25,000 to sequence a genome. By summer 2011, McCombie expects it to cost no more than $ 6,000.

Such advances have spurred scientists elsewhere to sequence their own and others' genomes, but McCombie isn't tempted to follow suit. "In the absence of a diagnosis of a disease like diabetes or cancer, I'm not sure personal genome sequencing has a cost benefit," he reasons, but admits to thinking about it. "All studies need positive controls," he says. "I wouldn't rule out including myself as one someday." **Hema Bashyam**

Tracking cancer's evolution with whole genome sequencing



DNA from a patient's tissue samples taken during three different stages of eosophageal cancer was first broken down into shorter fragments. These were individually "read" by an Illumina sequencer and then computationally reassembled into a "whole genome" sequence.Comparing this sequence to a reference genome identified genetic alterations that occurred as the cancer progressed.

Genomics unravels cognitive illnesses

Encouraged by Dr. James Watson, philanthropists Vada and Ted Stanley donated $25 million to launch the Stanley Institute for Cognitive Genomics at CSHL in 2007. Under McCombie's Directorship, the Institute is poised to translate findings about the genetics of cognitive disorders into DNA-based diagnostic tests by 2017.

It's now apparent that the mutations responsible for these disorders are rare variations in DNA caused by losses and/or gains of large chunks of DNA sequence (called copy number variations or CNVs) or by changes in single DNA nucleotides. CSHL's advances in next-generation sequencing "have made it possible to drill these disorders down to the single- nucleotide changes that lead to disease symptoms," says McCombie.

With scientists at the University of Edinburgh, McCombie's group recently sequenced the complete genomes of five members of a large Scottish family (135 members) in which many suffer from bipolar disorder or depression. "After identifying all variations among them, we believe we will be able to link these to the presence or absence of disease," McCombie explains. The scientists have also sequenced DISC1—a gene disrupted in another large family with psychiatric disorder—in about 2000 individuals, half of whom have schizophrenia, bipolar disorder or depression. Analysis is underway of the thousands of variants found to pinpoint the rare disease-causing ones.