Exploring along a Crooked Path*

Patrick O. Brown

Thank you, Evan and the Awards Committee, for choosing me to receive this honor.

I confess that before I learned of this award, I hadn't known much about Curt Stern, but when I finally read some of his work in preparation for this lecture, it really struck a chord. In an essay he published in 1944, entitled "The Journey, Not the Goal,"[1] Stern addressed the dilemma we often face between the desire to direct our research toward an altruistic goal, like understanding human disease, and the pure joy of exploration and discovery, which sometimes diverts from the goals that we set out after. Stern's essay makes a strong case for the value and importance of exploration for the sheer joy of it—that the little crooked paths taken by scientists who stray from the main road can unexpectedly turn into major highways.

And how appropriate, because the research that brought me here today *began* by wandering, and the tendency to wander off from my original goals has been a recurring and vaguely embarrassing feature of my scientific personality, but now I guess I can own up to it.

I wandered onto this particular crooked path in the fall of 1989. Everyone in my lab was then working on a great research problem—trying to understand the molecular events between entry of a retrovirus into a host cell and integration of the viral genome into the host's genome.[2] It was fun and challenging, and we had a very clear and unquestionably important practical goal—to try to understand enough about the basic molecular mechanisms so that we could eventually find ways to block HIV from infecting human cells.

But, as we all do, I liked to read about interesting problems *far* from my own research and to daydream about what I might do if I were working on them. One day, I happened upon an article about imprinting and its potential role in human genetic diseases.[3] At the time, there was only a handful of imprinted genes known, so it started me thinking about how I might find them all systematically. As a DNA enzymologist, the rather cockamamie scheme I dreamed up involved reciprocal crosses of two divergent strains of mice, a series of hybridizations of cDNAs, and the use of mismatch-repair enzymes to selectively isolate perfectly base-paired duplexes, eventually yielding a pool of sequences representing just the imprinted genes.

It was a pretty far-fetched idea, and I might have just chalked it up as the usual scientific daydreaming and forgotten about it, but, once I had a basic biochemical strategy worked out in principle, I encountered a set of great papers by Neil Risch, last year's Stern award recipient, about mapping complex traits by using affected relative pairs and mapping where in the genome they shared identity by descent.[4–6]

It occurred to me that the same basic biochemical methods could be used as an efficient way to get a high-resolution genomewide map of all the sequences identical by descent between two genomes, which, 16 years ago, seemed a useful contribution to linkage or association mapping. The final step in such a procedure, as I envisioned it, would involve a hybridization step analogous to FISH, but using a more organized version of a metaphase spread, in which cloned segments of the genome were laid out in a grid on a glass slide, so that we could read out a high-resolution map of identity by descent.

The challenge of making this wild scheme work and of taking on some really great mysteries in human genetics was irresistible. I was soon joined on this project by Stan Nelson, and, in a couple of years, the method, which we called "genomic mismatch scanning," was working, as Stan demonstrated using yeast as a model.[7]

The experiment was to map, for each of the haploid offspring of a cross between these two yeast strains, which genetic intervals were identical to each of the two parents, using our biochemical procedure to separately purify the sequences identical to the each parent and then hybridize each selected pool of DNA to an array of clones on a filter to produce the map.

With the basic procedure worked out, our next major tasks would be to get the same method to work on the human genome; to miniaturize the array, so that the reagent and production costs and handling properties of the arrays representing a large genome like the human genome would be manageable; and to use two-color comparative fluorescence to make the results robust and quantitative.

I had a 2-year pilot grant from National Center for Human Genome Research (NCHGR) to fund the initial work, so I wrote a renewal application in which I proposed to adapt it to the human genome and to develop the DNA microarray technology. Since we were fresh from our success in developing this new mapping method, I thought this grant would be a sure thing. When I got the reviews,

From the Department of Biochemistry, Stanford University School of Medicine and Howard Hughes Medical Institute, Stanford

Address for correspondence and reprints: Dr. Patrick O. Brown, Department of Biochemistry, Stanford University School of Medicine and Howard Hughes Medical Institute, Stanford, CA 94350. E-mail: pbrown@cmgm.stanford.edu

* Previously presented at the annual meeting of The American Society of Human Genetics, in Salt Lake City, on October 29, 2005.

Am. J. Hum. Genet. 2006;79:429–433. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7903-0005$15.00

I was devastated—the grant got the worst priority score I had ever seen. The reviewers liked the idea of adapting our biochemical procedure to the human genome, but they singled out the microarray proposal as unnecessary, premature, and very impractical.

Despite this terrible review, the folks at NCHGR were kind enough to give me a little bridge funding to keep the project alive and advised me to resubmit a proposal focusing on *just* the biochemistry and to get *rid* of the specific aim to develop DNA microarrays. So I grudgingly submitted that eviscerated proposal, and, this time, I got the grant.

But, meanwhile, we "blasted ahead" with the work on DNA microarrays, thanks to a smart, ambitious engineering student, Dari Shalon, and in a year—by the time that grant was funded—we had a simple system working for making and using DNA microarrays.[8]

At first, I thought that using DNA microarrays to look at global gene expression was just going to be a fun side project and that the more important and exciting application was for large-scale genotyping. So, when Joe Derisi joined my lab and printed the first microarrays representing the complete yeast genome, his very first hybridization was to get a genomewide genotype using genomic-mismatch scanning (fig. 1). It was a really beautiful result, one of the most exciting pictures I had ever seen. The red spots represent the genes inherited from one parent, the green spots genes inherited from the other parent.

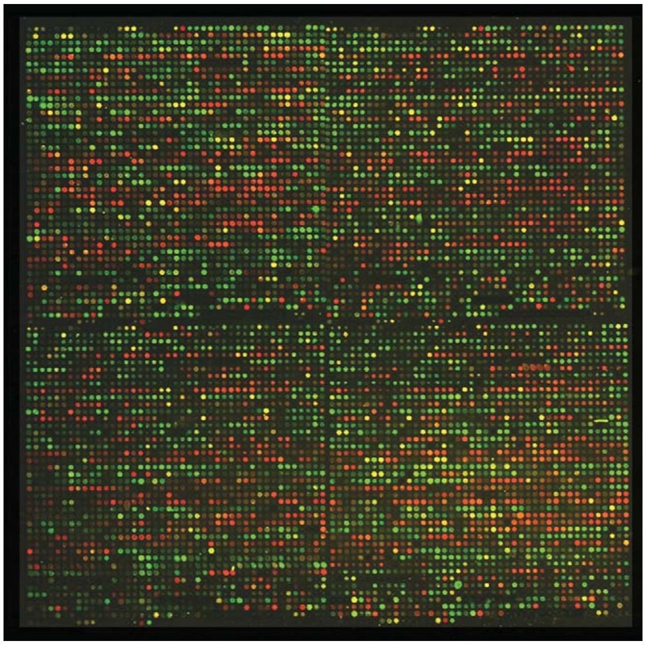

Figure 1. First result with a whole-genome DNA microarray: genotyping a *Saccharomyces cerevisiae* isolate by genomic mismatch scanning (experiment performed by J. DeRisi).

With four microarray hybridizations, we were able to determine the complete high-resolution genotypes of all four haploid products of a single yeast meiosis at single-gene resolution. It looked as if we were well on our way toward our goal; yet, almost 10 years later, we still haven't published this work. So why did we wander off the trail again?

The core idea of genetics is that variation in traits can be linked to variation in genes. We're used to thinking of this principle as it applies to the links between variation in gene sequences and variation between whole individuals. But there's a much richer kind of variation in both genes and traits inside each of our bodies, and, in many ways, it's more accessible to study. Although the cells that make up our bodies each have an identical set of genes, they are astonishingly diverse in shape and size and behavior. That's possible because they are equally diverse in the patterns in which their genes are expressed—just as we can use the 30,000 or so words in our vocabulary to write a million different stories by combining them in different patterns. Our cells can use the 30,000 or so genes in their genetic vocabulary in different and dynamically changing patterns in each cell to specify the unique and dynamic characteristics of that cell.

What made us wander again off the path we had started on was that some of the earliest experiments carried out by Joe DeRisi, Mike Eisen, and Vishy Iyer showed that, by using DNA microarrays to look systematically at the patterns of mRNA expression in diverse cells and tissues, we could make a map that shows relationships between expression of specific sets of genes and the distinctive characteristics of each human cell or tissue.[9–11]

We could apply the fundamental logic of genetics to this new kind of data, to make a new kind of map of the genome that shows connections between traits—of cells, tissues, or individuals—and genes—in this case, not their sequence but their expression—on a genomewide scale.[10] In this new kind of map, the genes could be ordered, not by their chromosomal location, but by their pattern of expression, and, thus, indirectly by the biological role that pattern represents. And cells, tissues, physiological or developmental processes, or individuals could also be clustered or grouped on the basis of the similarities in their global expression patterns, which are closely tied to their phenotypic characteristics.

Relating the patterns of expression of specific genes to phenotypic variation gives us a way to learn about the functions of specific genes. Conversely, the gene-expression profile of a human cell or tissue can be a distinctive identifying signature—analogous to the genotype—of that tissue.[12] And that has obvious medical implications.[13]

Soon after we had the microarray technology in place, I began a very rewarding collaboration with David Botstein to profile gene-expression patterns in cancer.[14–18] Figure 2 shows some of our earliest data from several different kinds of common human cancers.

You can certainly tell apart cancers that we already knew were different—for example, the gene-expression patterns in the breast cancers (black box) are clearly very different

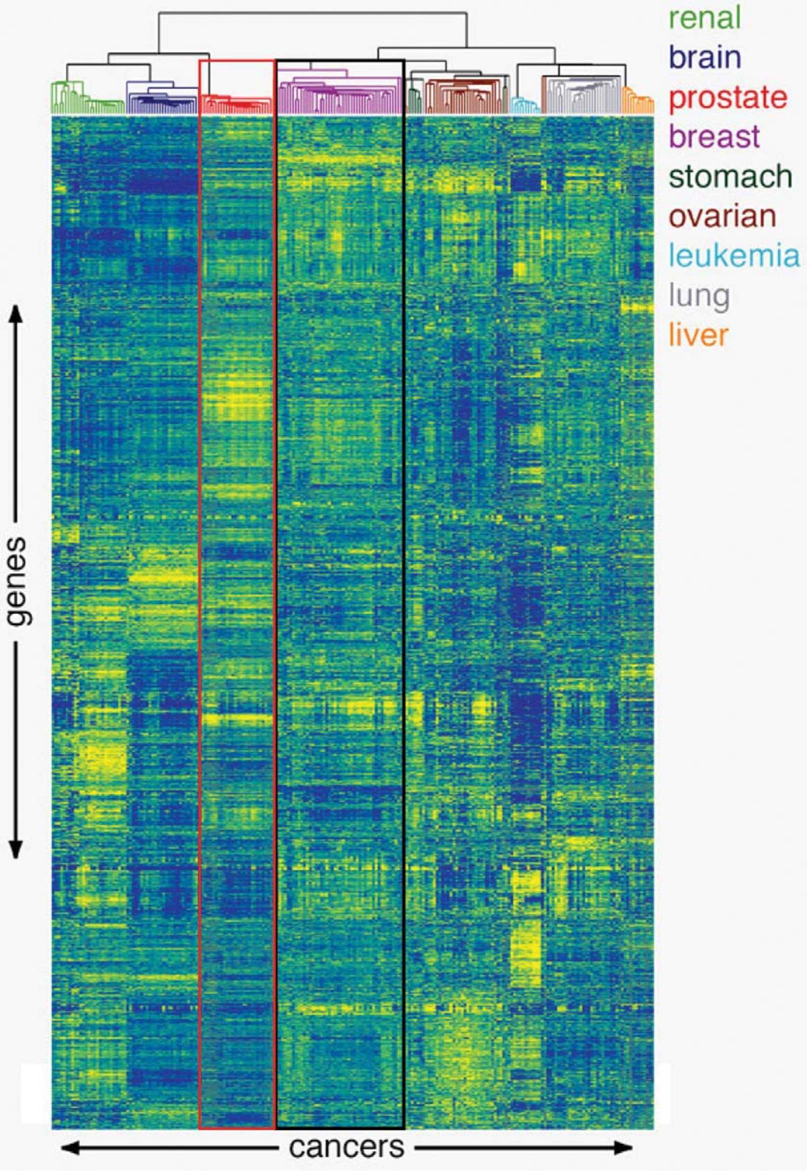

Figure 2. Analysis of gene-expression patterns in human cancer. This "map" shows the results of quantitative profiling of gene-expression patterns in a diverse panel of human cancers in a table format in which each row represents a specific gene, each column represents a specific tumor sample, and the measured abundance of each gene's transcript in each tumor is represented by a color scale in which bright yellow represents the highest expression levels and dark blue represents the lowest expression levels. The genes and tumor samples are organized by hierarchical clustering on the basis of similarity in their overall patterns of expression. Note that tumors of similar histological type—for example, breast cancer (*black box*) and prostate cancer (*red box*)—almost always cluster together, yet tumors of the same histological type can vary significantly in their expression patterns.

from the patterns in the prostate cancers (red box). And the fact that these molecular portraits give us a richly detailed picture of what proteins are present and which regulatory systems appear abnormally active or missing in each of these tumors becomes more and more important as drugs that act on specific molecular targets become available.

And, if you look more closely, you can see that, even within a single diagnostic classification—for example, among the breast cancers—there is actually tremendous molecular diversity: each individual patient's cancer has a distinctive expression pattern.[15,16] These differences in gene-expression patterns among cancers that we can't tell apart just by looking under the microscope turn out to provide valuable new information about their potential to progress and metastasize—in many cases, much better than the classic pathological predictors.

Of course, we don't want just diagnostic markers. We really want to understand the underlying mechanisms of pathogenesis, and, in fact, our current work using DNA microarrays to investigate cancer focuses almost entirely on working out the pathogenetic mechanisms that underlie specific features of the gene-expression patterns we see in specific cancers.

Maps like these show us a huge number of possible connections between the distinctive characteristics of each cancer and the specific genes it expresses, but making sense of the connections depends on synthesis of what we know about this cancer on the one hand—bits of information contained in thousands of published articles—and what we know about these genes on the other, again information locked away in thousands of published articles. When my colleagues and I started trying to make sense of maps like these, we immediately discovered that we couldn't read all these thousands of articles fast enough to realize the potential emergent value of large, systematic data sets like this—we would need to develop software tools to collect and synthesize this information. But there was a terrible, fundamental problem that completely blocked the way to this goal—the way that published scientific information is managed. The published knowledge that we wanted to integrate was not available to be used in that way for a simple and perverse reason: the traditional scientific publishers consider the work we scientists have carried out and published in their journals to be their private property and won't allow the information to leave their control so that it can be used creatively for the good of science and the public.

The fact that the world's treasury of published DNA sequences is a shared public resource, available to be downloaded in their entirety into individual researcher's computers and used and analyzed without restriction, was the secret to the success of the Human Genome Project and much of modern biology, is so obviously the right thing for science that we take it for granted.[19] And that's the way it can be and ought to be, not just for sequences but for all the work that we publish.

That realization has led me to another winding trail, on which I've been joined by my former mentor, Harold Varmus, and former post-doc, Mike Eisen: the Public Library of Science, which we started as a grass-roots advocacy movement and which has turned into a nonprofit, open-access publisher of great scientific journals, including *PLoS Genetics, PLoS Biology,* and *PLoS Medicine*.[20]

Our goal is not just to publish our own open-access journals but to catalyze a change in the whole system, so that every published scientific article and all its contents become a public resource, freely available to be read or used in any way by anyone in the world, just as published DNA sequences are today.

Sixteen years ago, I was determined to figure out HIV replication, but I wandered far off the trail. My graduate advisor, Nick Cozzarelli, whose insight and wisdom I appreciate more with each passing year, once chided me about a paper I was neglecting to write by saying that I got too much joy from initiating projects and not enough joy from completing them. It was a pretty apt criticism. But I console myself with the closing words of this essay by my kindred spirit, Curt Stern: "the joy of the journey is never ending, that of reaching a goal always passing."[1(p100)]

The accomplishments that the Curt Stern Award recognizes depend on ideas and discoveries and contributions from a wonderful, brilliant group of students and post-docs and colleagues. It's been my great privilege and good fortune to know them and work with them.

Web Resource

The URL for data presented herein is as follows:

Public Library of Science, http://www.plos.org/

References

1. Stern C (1944) The journey, not the goal. Scientific Monthly 58:96–100
2. Brown PO (1997) Integration. In: Coffin JM, Hughes SH, Varmus HE (eds) Retroviruses. Cold Spring Harbor Press, Plainview, NY, pp 161–203
3. Reik W (1989) Genomic imprinting and genetic disorders in man. Trends Genet 5:331–336
4. Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241
5. Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253
6. Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228
7. Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO (1993) Genomic mismatch scanning: a new approach to genetic linkage mapping. Nat Genet 4:11–18
8. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
9. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686
10. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868
11. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO (1999) The transcriptional program in the response of human fibroblasts to serum. Science 283:83–87
12. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33–37
13. Brown PO, Hartwell L (1998) Genomics and human disease—variations on variation. Nat Genet 18:91–93
14. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 96:9212–9217
15. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, et al (2000) Distinct types of diffuse large B-

cell lymphoma identified by gene expression profiling. Nature 403:503–511

16. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. Nature 406:747–752
17. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227–235
18. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98:10869–10874
19. Roberts RJ, Varmus HE, Ashburner M, Brown PO, Eisen MB, Khosla C, Kirschner M, Nusse R, Scott M, Wold B (2001) Information access: building a “GenBank” of the published literature. Science 291:2318–2319
20. Brown PO, Eisen MB, Varmus HE (2003) Why PLoS became a publisher. PLoS Biol 1:E36