ASSAY and Drug Development Technologies
Volume 1, Number 1-2, 2003

Profile

An Interview with Francis S. Collins, M.D., Ph.D. Director, National Human Genome Research Institute

Francis S. Collins, M.D., Ph.D., is a physician–geneticist and the Director of the National Human Genome Research Institute, of the National Institutes of Health. In that role, he oversees the Human Genome Project, a 15-year project aimed at mapping and sequencing all human DNA. Many consider this the most important scientific undertaking of our time. The project is currently running ahead of schedule and under budget, and all of the original goals will be completed in April, 2003.

Collins was raised on a small farm in Virginia and was home-schooled until the sixth grade. He obtained his undergraduate degree in chemistry at the University of Virginia and went on to obtain a Ph.D. in physical chemistry at Yale University, and a medical degree at the University of North Carolina. After a residency and chief residency in internal medicine in Chapel Hill, he returned to Yale for a fellowship in human genetics, where he worked on methods of crossing large stretches of DNA to identify disease genes. He continued to develop these ideas after joining the faculty at the University of Michigan in 1984. This approach, for which he later coined the term "positional cloning," has developed into a powerful component of modern molecular genetics, as it allows the identification of disease genes for almost any condition, without knowing ahead of time what the functional abnormality might be.

Together with Lap-Chee Tsui and Jack Riordan of the Hospital for Sick Children in Toronto, Canada, his research team identified the gene for cystic fibrosis using this strategy in 1989. That was followed by his group's identification of the neurofibromatosis type 1 gene in 1990, and a successful collaborative effort to identify the gene for Huntington's disease in 1993. That same year, Collins accepted an invitation to become the second director of the National Center for Human Genome Research, following in the footsteps of James Watson. In that role, Collins has overseen the successful completion of all of the Genome Project's initial goals, including the imminent completion of the sequencing phase. In addition, Collins founded a new NIH intramural research program in genome research, which has grown to become one of the premier research units in human genetics in the country. His own research laboratory continues to be vigorously active, exploring the molecular genetics of adult-onset diabetes and other disorders. His accomplishments have been recognized by election to the Institute of Medicine and the National Academy of Sciences, and by numerous national and international awards.

Dr. Collins, please provide our readers with a brief historical journey through your career in the biosciences.

I have had a very nonlinear career pathway. As a young high school student, I was primarily interested in chemistry. That was the first subject that really caught my attention; it has an intellectual foundation that I found very compelling. My undergraduate major was chemistry and my Ph.D. work was in physical chemistry. I avoided biology by every possible means at my disposal because it seemed to be a messy, unsatisfying, descriptive science with no underlying principles. In fact, I was dead wrong, but it took me a long time to discover that. Only as a graduate student at Yale, studying chemistry, did I begin to become dimly aware of the exciting things going on in biology. I began to realize that DNA, RNA, and protein provided a very satisfying, digital underpinning for a science that I had assumed was wholly lacking in such things.

I found that revelation very exciting and realized that my own impressions had been mistaken. I realized, too, that biology was grabbing my fancy more than was quantum mechanics, which is what I was working on at the time. So, to keep my options open, and not knowing what else to do, I went to medical school. That was a terrible reason to go to medical school, but it worked out fine, because I loved it. I loved clinical medicine, I loved taking care of patients, and I loved learning about the intricacies of the body and figuring out how to compensate when the body is not doing what it is supposed to be doing. Ultimately, I also found medicine very frustrating, because the level of ignorance about most conditions was so massive that when I would ask myself, "Am I helping this person by this medical intervention that I'm about to undertake?" The question was often left hanging in the air without a clear answer.

Having gotten excited about the underpinnings of biology—the DNA, RNA, and protein stuff—I began to believe that this was where the future would lie. Early on, I became convinced that I wanted to study medical genetics as a way of bringing medicine and biological research together. That is what I did after completing my clinical training. I pursued a fellowship at Yale in human genetics, mostly working at the bench, but also learning the skills of the clinical geneticist and preparing myself for an academic career in which I would teach, take care of patients, and do research. That is very much what I went on to do at the University of Michigan for nine years, from 1984 to 1993, and I enjoyed that thoroughly. Happily, my laboratory was fairly successful and helped discover the genes for disorders such as cystic fibrosis, neurofibromatosis, and Huntington's disease.

Then the call came to see if I was interested in stepping into those very large shoes previously occupied by Jim Watson and come to lead the Human Genome Project at NIH. After initially turning it down, because I really wasn't interested in moving away from what I was doing, I began to realize that this was an historic opportunity, one that would not come along again, and that I must be absolutely, certifiably crazy to consider turning it down. So, on the second go-round, I decided to accept the position, and I came here in 1993 as the director of the project.

You have been credited with coining the term "positional cloning." Can you explain for our readers the basis for this method? For which diseases has it helped uncover the molecular basis, and how relevant is it today in defining the basis of disease?

Positional cloning means finding or cloning a gene by its position in the genome. That is in contrast to finding a gene by having an idea of what its function is, which is the only way we were able to find disease genes up until the mid-1980s. In the early 1980s, positional cloning seemed to be prohibitively difficult. The goal was to try to find something as subtle as a single base pair that was misspelled in a genome of three billion letters. The likelihood of having tools powerful enough to do that seemed remote indeed.

With the advances of being able to build genetic and physical maps of the genome, and now, ultimately, sequencing the entire genome, those barriers have gradually come down. At first they came down slowly. Certainly in the 1980s, when we were chasing the gene for cystic fibrosis, I had many doubts myself as to whether this was a wise undertaking for a junior assistant professor trying to make some kind of plausible attempt at a career. But with hard work and collaboration with other very talented scientists, especially Lap-Chee Tsui in Toronto, we were able to narrow down the territory, using the tools of genetics, and find a three base pair deletion that was the common cause of cystic fibrosis.

That was a fairly significant event, because it was the first time that a disease gene had been found solely by this strategy of positional cloning, without any other available shortcuts. When you look back on it now, it seems almost like a trivial exercise. A graduate student in my lab could probably do it in about two weeks, given access to DNA samples from affected families, the available databases, a thermal cycler, and a sequencing machine. However, that was not the case in the 1980s. Hundreds of disease genes have now been identified using this strategy, and there is increasing optimism that this will not only work for diseases inherited in a Mendelian fashion, but also for common diseases such as diabetes, asthma, and heart disease, which are much larger contributors to morbidity and mortality.

Estimates for the number of genes encoded by the human genome have varied from 30,000 to 100,000 or more. Is there a more accurate estimate since the initial reports? What factors make this number hard to pin down?

We first have to be clear about the definition of a gene. If we call a gene a stretch of DNA that codes for a protein, then the number 30,000 will probably be about right. Now that we have the mouse sequence and can compare that to the human sequence—and increasingly, we will have other genomes coming along, too—a number of the predictions about genes that were made when we only had the human sequence can be significantly refined. We are finding out, for example, that some genes we thought might be two genes are really only one, or something we thought was a gene is actually a pseudogene and is probably not functional at all.

The number of genes that code for proteins has been shrinking somewhat since February 2001, when we published the initial analysis of the human genome, and that number was already surprisingly low. The other thing we have learned, however, particularly in the past year or so, is that there may be quite a large number of genes that specify short, noncoding RNAs—RNAs that function in some way yet to be determined, but that probably play a significant role in regulation. Some may function as antisense molecules, and some as the real biological equivalent of RNAi—this wonderful way of interfering with gene expression that has become a tool of the molecular biologist probably has a counterpart *in vivo*. If you consider those as genes, and by some definitions they would be, then there might be thousands more genes, perhaps even tens of thousands. Yet those are genes that we have had relatively poor tools to recognize until recently, and so that question is very much up in the air.

Being able to identify them in the genome sequence is proving to be the challenge. The best way to identify them is to catch an RNA that has been made from one of those DNA sequences, prove its existence, and prove that it has a biological function and is not just noise. That is a bit of an issue. It is clear that there is some transcriptional noise in the system. Finding a single copy of an RNA floating around in a cell does not mean that it has biological meaning. The noise in the system has confounded some of the efforts at enumerating the total complement of genes. This is an example of how evolution can help us enormously. If you are going to argue that something is biologically important, then you would expect that it would show evidence of conservation over evolutionary time scales.

Have there been any immediate revelations that have come from having the entire human genome laid out before our eyes?

A long list! One of our greatest fears was that there would not be any, and that our reaction would be, "Oh, that's basically what I thought." What a downer that would have been! But if you look at the February, 2001 *Nature* paper describing the initial analysis of the human sequence, there is a long list of blow-your-socks-off revelations, beginning, of course, with this observation that there are not nearly as many genes as we thought. The surprises continued by revealing glimpses of how we get by with such a short list of genes; alternative splicing, for instance, is much more frequent than people had previously estimated. Additionally, looking at the architecture of mammalian proteins led to the discovery that they are rather complex in the way they are put together, compared to their homologues in simpler species such as yeast or worms. Mammals have cobbled together more domains into one protein than you usually see in simpler species, and that may account for our complexity, by allowing proteins to take on multitasking capabilities. On top of that, the genome sequence has taught us a prodigious amount about the so-called "junk" in the system, the repetitive DNA that constitutes more than half of the DNA in our genome. At least some of that "junk" seems to carry the earmarks of having had a functional contribution; otherwise, it would not have been retained in the gene-rich regions throughout evolution. That was a big surprise and caused all of us to reconsider the use of the word "junk" to describe any part of the genome, since we seemed to have been wrong about the most repetitive elements.

We also learned interesting things about mutation rates. Perhaps the most well cited discovery resulted from studying the sequence of the Y chromosome, which revealed that compared to the rest of the genome, male mutation rates are twice as high as female mutation rates. That is, the mistakes made in passing DNA from parent to child are made twice as often by fathers as they are by mothers. I have yet to meet a woman who is surprised by that!

You have said that the haplotype map of the human genome will be the "personalization" of the genome. Can you please elaborate on this?

The sequence of the genome was derived from anonymous individuals. It did not particularly matter who they were, because we were focusing on the 99.9% that we all have in identical form. Our similarities are quite dramatic when you consider that 99.9% of my sequence is the same as yours. But the 0.1% is of profound interest for medical purposes, because buried within that small number of differences between individuals must lie the clues to hereditary susceptibility to virtually all diseases. We now have the opportunity and responsibility to explore those differences—to be able to pinpoint who is at risk for what disease, to use that information to discover what pathways have gone awry in diseases such as diabetes and heart disease, and to develop the therapies of the future that are based on that molecular understanding.

The principle of studying that set of variations, that 0.1%, and figuring out how they correlate with disease is an extremely compelling one. But the reality of actually doing that is extremely daunting. There are an estimated 10 million sites in the genome where common variations in the sequence exist. Of those 10 million, only a small fraction, perhaps a few hundred thousand, fall in places in the genome where they have some consequence, and only a much smaller fraction of those will be involved in disease. How do we sort this all out? One ap-

proach would be to develop a catalog of the 10 million places that vary and to set up very large studies in which you check each of those sites in perhaps 1,000 people with the disease and 1,000 people without the disease, in order to track down which of those variants are associated with risk. That would be a very powerful approach. The problem, though, is that it would be prohibitively expensive.

The haplotype map is a shortcut. It takes advantage of fundamental features of the human genome. This shortcut is possible because those 10 million variable places in the genome are correlated with their neighbors. If you have a variant in one place—a site, for example, where I might have a C and you might have a T—and you walk down the chromosome and come to another variant a couple thousand base pairs away—where you might have a G and I might have an A—it turns out that those are not independent observations. In fact, in general, there is a rather tight correlation over a distance of 10,000 to 20,000 base pairs, but the exact organization of that correlation varies from place to place in the genome. We will not know exactly what the correlation looks like until we go in and experimentally determine it. That is what the haplotype map is all about—to determine how that correlation works, what the "neighborhoods" look like. If you knew that, you could simply pick a gold standard subset of variants that represent the majority of the variation in the genome and develop a genome search around those. The positives would tell you that you are in the right neighborhood, and you could save yourself about a factor of 20 to 30 in the amount of labor necessary to do such a study. That begins to bring such studies into the realm of possibility in the future.

This is a very powerful argument, and I firmly believe that for medical applications of the genome, having this haplotype pathway to gene discovery will be a major advance. The project aimed at generating a haplotype map is well under way. In October, five countries and nine laboratories began working on this $100 million project, and we will reveal a haplotype map covering 80% to 90% of the genome in as little as three years.

Where do you see the best use of genomic information derived from the Human Genome Project?

I see the best use everywhere. I want to see this information used in every academic laboratory to speed up the process of making basic science discoveries, and I want to see it used in the private sector in every imaginable way to accelerate the process of developing better diagnostics and therapeutics, which is what the private sector does best. It is gratifying to see genomics being embraced in the biotechnology and pharmaceutical industries, where it is being applied to transform therapeutic development into a new rational process based on an understanding of how things really work. That is the promise that was put forward 12 years ago by the planners of the project, and it is now coming true.

Although we have the information to begin to make this possible, do we have the necessary tools?

Yes and no. We have the basic sequence, and it is available in an absolutely free and unfettered way in the public domain. Anybody with Internet access can have immediate access to the information. That was a very important outcome of the last few years, making sure that the barriers that might have slowed down the process are not there. With regard to the tools, we have some of them, but we need a lot more. Some of those tools are new technologies that enable us to do very cheap genotyping, so we can use the haplotype map when it becomes available to good advantage. We also need tools to enable very cheap sequencing, so that you and I can have our genomes sequenced for $1,000 or less, which would radically transform the way we do research and practice medicine. We need improved methods for studying gene regulation using microarrays. Additional methods are needed that allow us to explore human proteomics, to understand how pathways work and how proteins touch each other, and that are scalable to the whole proteome. The goal is to accelerate the process of figuring out how the parts all fit together and interact with each other. All of those are critical needs, and many of those tools are still in the development process. For NHGRI, I see that as one of our most pressing mandates, to make sure we focus on moving that process along.

What should the role of the National Human Genome Research Institute be now that the human genome has been sequenced, and has that role already been redefined?

NHGRI is on target to unveil a new plan for genome research this April, which will be bold and ambitious. Based on the opportunities we have had to try it out on people, I think we are well on the way to achieving that goal. The Human Genome Project was originally proposed in 1988 by a National Research Council advisory panel, and we have followed that blueprint closely and carefully. To the credit of all the scientists involved, we have achieved every single one of those original goals, either on time or ahead of time, and within a budget substantially less than was projected in 1988. All of the goals of the Human Genome Project originally envisioned will be realized by this April, when we will have the human

genome sequence available in its essentially completed form on the Internet.

That will be quite a moment of celebration. The completion of the human sequence will fall in the same month as the 50th anniversary of Watson and Crick's discovery of the double helix. That will also be the month when we will unveil NHGRI's plan for the future—a very appropriate time to do so. We will be celebrating all that we have already accomplished and can then ask, "So now what?" We aim to have a very thorough and inspiring answer to "So now what?" It will include a focus on how we bring genomics to biology, particularly how we understand the products of genes, the proteins—what they do and how they interact with each other. It will also emphasize the implications of understanding genetic variation, and the importance of continuing to define sequence from many different organisms, because the comparison of those genomes is incredibly powerful.

This plan will focus much more heavily than was possible in the past on the application of genomics to medicine. The opportunities are now becoming much more real. There will be a focus on how to make it possible for academic laboratories to have access to high throughput screening of small molecule libraries in order to identify compounds that could be very useful probes of biological pathways, something which the pharmaceutical industry does every day as part of identifying lead compounds in drug discovery. I am very excited about trying to promote this transition of the technology, so that it is not limited to the private sector. That is happening in a few places, but most researchers outside the private sector have not had much opportunity to think about how they would use such tools. This ought to be a major priority for the next few years. It offers a wonderful opportunity for collaboration between academia and the private sector, in that the discoveries of small molecules that are useful as probes for exploring biological pathways could then become the first step of a drug discovery program, which could be done collaboratively with a biotechnology or pharmaceutical company, speeding up the process.

Another component of the plan focuses on how to ensure the benevolent applications of genomics to society. This includes a consideration of the consequences of understanding things about ourselves that are not necessarily directly medical in their implications, but might be important for understanding behaviors, the differences between populations, and the complex questions of race and ethnicity. Other issues we will consider are how genetic information will be used in the legal system and in deciding who has access to medical care.

Should the institute support the development of new technologies for genome analysis or concentrate on making existing technologies as widely available as possible?

We need to do both. While much technology development goes on very effectively in the biotechnology industry, and we are delighted about that, the earlier steps in technology development, particularly when it is not clear how long it will take to develop a financially viable product, have traditionally gone on in the academic sector. This needs to be supported even more vigorously in the future given the importance of these advances in achieving the outcomes that we have been talking about. If you look back at the last 10 years, you will find examples of how NIH support has been critical in getting such technologies off the ground. The company Affymetrix, for example, was essentially founded on work supported by an NIH grant in the early 1990s. At the same time, NIH has to be realistic about where our investments can best be made, and we have to recognize that this is high risk and that a lot of the money one puts into technology development does not immediately pay off.

What role will sequencing of nonhuman genomes play in NHGRI, and how does this serve the institute's mission of "Improving human health through genetic research"?

Every new genome sequence gives us a chance to compare that with the growing database of other genomes and, invariably, we learn something. We learn how those genomes are all similar, which points you to central functions that living organisms need to be able to carry out, and also how they are different. We already have a vigorous program to allow us to evaluate proposals about which genomes to sequence next. This program involves the submission of a White Paper by an investigator or community of investigators who are interested in seeing that genome sequenced. Those get reviewed by a peer review committee and are placed into high-, medium-, or low-priority bins. As sequencing capacity becomes available in our centers, an organism is picked from the high-priority bin. We have started sequencing the chicken and the chimpanzee and relatively soon will begin the cow, the honeybee, the dog, and a group of fungi.

I find this to be an extremely compelling part of what NHGRI should be supporting. We are basically able to look into evolution's lab notebook and see what has worked. That is a powerful way to understand gene function. The challenges are considerable in terms of the computational aspects of learning everything you would like to once you have six or ten mammalian genomes and you can line them all up alongside one another. How do you do that, and do it in a fashion that digs out the most important information? I think the appetite for sequence will

grow, and as the cost of sequencing continues to drop, we should be able to support a lot of this activity without breaking the bank.

Do you believe there are any drugs directly attributable to genomic sequencing, and what are the next diseases you believe will yield to therapy by drugs as a result of the genome project?

There are drugs that have been profoundly influenced by the availability of gene sequencing, but given the recentness of the human genome and the length of the drug development pipeline, perhaps not yet by genome sequencing. Consider a drug such as Gleevec; we would not have that drug if people had not spent decades understanding how the translocation between chromosome 9 and 22 results in a fusion protein partly coded by the *bcr* gene on chromosome 22 and partly by the *abl* gene on chromosome 9. That whole pathway of discovery was directly based on understanding genetics and involved gene sequencing as part of the enterprise to figure out what the target should be. We now have a drug that, by everyone's estimation, is considered to be a real "home run" in the treatment of chronic myeloid leukemia.

Another example is Herceptin. We would not even have had the idea for that drug had the role of the HER2 receptor not been realized, particularly the fact that it is amplified in some types of breast cancer. That whole pathway is also a genomic success, although it came along prior to our having much of a grasp of the genome sequence *per se*. Consider also the work being done at Human Genome Sciences and other companies that are focusing drug discovery on cDNA sequences.

How much should the institute focus on low-prevalence, single-gene disorders versus high-prevalence, multigenic disorders?

That is an interesting question. The NIH as a whole has a strong interest in all diseases, and the disease-specific institutes have generally paid attention to those disorders that are the most high prevalence, because they are often the major source of morbidity and mortality and economic losses. But I would not say that rare diseases have been neglected. The National Human Genome Research Institute has a particularly strong role to play in the study of rare diseases, and we have recently partnered up with the Office of Rare Diseases of the NIH to make this a very robust area of focus. The focus will include an effort to identify specific rare disorders that are ripe for a therapeutic advance. Those conditions would not be likely to get much attention in the private sector, because they are rare enough that it is unlikely there would be a viable financial market. But approaching those diseases as models could be extremely useful, just as approaching the disease familial hypercholesterolemia led to the discovery of the statins. In this next phase of research, in which we will be paying specific attention to clinical applications of genomics, you will see us focusing on rare diseases that represent paradigms for areas in which therapeutic approaches might now be possible.

What technologies do you think will provide the best route to "triaging" the genome; that is, identifying and validating drug targets and weeding out those genes that are not appropriate for therapeutic intervention?

I am not sure that anybody knows the answer to that question. Traditionally, a certain subset of gene products have been considered to be attractive drug targets because that "territory" has been well worked out. Whether those targets are G protein-coupled receptors or kinases, they are the focus of much attention because assay systems have been developed and they represent strategies that have succeeded and for which many of the details have already been enumerated.

But I think that it would be unfortunate if we limited ourselves, now with the 30,000 genes of the human genome and hundreds of thousands of possible protein products in front of us, to the gene families with which we are already familiar. We should not simply discard the genes and proteins that do not fit into our prior description of good drug targets. After all, our ignorance is pretty substantial, and there may be ways of approaching novel gene products using innovative strategies, such as RNAi and others. Obviously, priorities will be set based on targets about which we know the most, but I would be reluctant to say that there is a certain set of protein products that are "off the table" anytime in the future. We have a lot of exploration to do, and the biotechnology and pharmaceutical companies that are pushing the envelope will reveal ways to get beyond our 500 traditional drug targets into broader territory, with much greater potential for medical benefit.

How successful do you think the Human Genome Project has been in providing unfettered access to newly sequenced genes for the purposes of drug discovery? Will the availability of the sequence itself make it easier for more institutes/companies to discover new drugs, or will that need to wait until functions have been ascribed?

Public accessibility of the DNA sequence, which is updated on the Internet every 24 hours, has been a defin-

ing feature of the public international consortium and a critical part of its success. There are no barriers to access, and there has been no limitation to the immediate use of this information. I strongly defend that decision, made largely by the scientists involved, as having been the right thing to do.

In terms of gene function, I think there are a number of genome-wide approaches to function that are very appropriately following that tradition of making the data freely accessible. Some of those approaches involve trying to understand which proteins interact with each other, which is presently being done quite successfully in a scaled up fashion in yeast. Many of the results of functional studies should also be made immediately accessible. However, some of the data will be proprietary, of course, since a company that is pursuing a particular gene as a possible druggable target will want to protect the competitive edge of that pathway to drug discovery. We all understand that is how the system works.

We have talked mainly about the role of government. Should the activities of industry change in this new genomic era, and if so, how? Should industry focus more on development of validated targets, given the increasing expense of drug development and decreasing reimbursements? If so, is NIH prepared to focus more on target validation that is relevant to industry than it has in the past?

The opportunities for partnerships between academia and industry are on the way up, not down, and in several ways. In one very obvious and gratifying way, we are seeing more and more of these public–private partnerships that have as their focus the derivation of large data sets that everyone is anxious to use. The SNP Consortium was an early and very successful example. Another example would be industry's help in sequencing the mouse genome. The haplotype map effort will be a partnership with the private sector, with everyone agreeing that this is data we should have as quickly as possible and that should be accessible to all.

This is a new paradigm, and the success of these enterprises bodes well for additional partnerships in the future. We need to identify the precompetitive data sets that are most appropriate for these types of partnerships. We cannot expect this type of scenario to work when the data to be produced have competitive aspects; in such cases, industry is going to be justifiably uncomfortable with an insistence that the data be placed in the public domain. Many aspects of genomics are upstream of the competitive phase, however, such as haplotypes, SNPs, or sequences. Much microarray data fall into this category as well. We can envision data that represent the level of expression of tens of thousands of genes in thousands of different human tissues—a data set that would be carefully validated and made widely accessible. Such an effort has been proposed by the International Genomics Consortium and is under active discussion. Those types of opportunities will continue to arise, and we just have to be thoughtful about the nature of the data being proposed.

The other area I want to reemphasize is the need for greater collaboration between academia and industry on high-throughput screening of small-molecule libraries. At present, this is largely a private sector activity. The notion of making available to academics that very powerful toolkit is an appealing one. Researchers in academia could carry out the first steps and identify compounds that have interesting biological activity, but which could also be quite useful for drug development. Academia could then identify an appropriate company that could license the discovery. That paradigm could offer multiple advantages, including increasing the breadth of private sector pipelines and, perhaps, making it more likely that rare diseases would get more attention.

Thank you, Dr. Collins.

—Interview by Vicki Glaser