Future Directions in Genetic and Genomic Research
Francis S. Collins, M.D., Ph.D.
Director, National Human Genome Research Institute

DR. McCABE: We'll move on, and we'll have a chance to talk with all of our speakers at the roundtable this afternoon. Let's move on now to Dr. Francis Collins, future directions in genetic and genomic research, Director of the National Human Genome Research Institute, and it's great to have you joining us in the mosh pit, Francis.

DR. COLLINS: Nice to be down here.

Good afternoon, everyone. I'm delighted to have a chance to speak to this distinguished group in this inaugural meeting of this new Committee that I believe has a very important mandate and clearly a lot of work to do.

In figuring out what to say to you, I thought I would focus particularly, because I think that was the charge, on some predictions about where genetic and genomic research is going. While you're all ruffling through your papers, let me tell you that I don't have a handout in front of you, but you will find in your briefing books a copy of a paper published in April in Nature under the "Genetic and Genomic Research" tab which I'm going to go over, albeit somewhat lightly because of the time, which will point out a series of areas where we believe the highest priorities now can rest in terms of where genetic and genomic research will be going next. Then I will at the end come up with some suggestions perhaps of particular areas that I think are ripe for further exploration by a group such as this.

So just by way of context to remind you that genetics didn't come around yesterday, we are actually standing on the shoulders of people all the way back to Mendel, and of course many things that followed on after that, the discovery of the DNA double helix, already having been mentioned, exactly 50 years ago. Of course, building upon that was the discovery of recombinant DNA, and then in the 1980s many other additional technologies coming along, leading to the initiation of the Human Genome Project in 1990.

A whole host of things happened that I'm not going to go over in any detail at all, only to remind you that the Genome Project was about a lot more than just getting those 3 billion letters of the human DNA in place. It was also about model organisms, it was about technologies, it was about map development, and it was about ethical, legal, and social issues. The ELSI program, the HGP, founded in 1990, represents the largest investment in bioethics on any topic, and as I said this morning, I hope that research will turn out to be very valuable to this Committee as input into some of the areas that you decide to focus on.

The second component of the genome enterprise over the last 13 years is depicted here, decorated by a number of publications of increasing complexity, including a draft sequence of the human genome in February of 2001, and the finished version of the human genome sequence having been announced just about six weeks ago. But I want to point you also to those three little words down in the right margin there, those three little words that say "To Be Continued," because I think that's what I'd like to now pay some attention to in the rest of this presentation, because I think it would be important for this Committee not only to think about our current situation but what's coming next.

So what is next? Well, in this article which you have, we depicted this future that we're aiming to try to develop as a metaphorical building, a building resting upon the foundation of the Human Genome Project, as you see here, and consisting of three floors of this rather Frank Lloyd Wright-inspired-looking edifice here. One floor is genomics to biology, another floor is genomics to health, and the third is genomics to society. That, by the way, looks an awful lot like the three designations of this Committee, the Secretary's Advisory Committee on Genetics, Health, and Society. What about that?

You will notice also that this building is held together by a series of vertical cross-cutting elements that touch on all the floors and are going to be necessary if this building is going to come to pass and be structurally sound. By the way, notice that the door is open here, which is inviting you to come in and work on any of these floors that you'd like to, and also that the data that will be generated by this enterprise will be accessible to anybody without passwords or other restrictions on access.

But this is a pretty bold notion, that we would try to do this. So what exactly is going to be going on on these various floors? The process by which we develop this set of grand challenges that are described in that particular article involved input from more than 600 scientists and ethicists and public policy experts over the course of almost two years, and out of that, after many iterations, came this series of grand challenges which are aimed to be perhaps a little on the audacious side in that they are, many of them, things which we do not currently see a direct and time-limited pathway to achieving, but they are things which, if they could be accomplished, would make a profound impact on research and on the practice of medicine.

So in that regard -- we have some very interesting biology here in this room. I won't even tell you what's crawling across the rug over here.

(Laughter.)

DR. COLLINS: It's a model organism, I can tell you that.

(Laughter.)

DR. COLLINS: It probably has more genes than I do. The smaller they are, the more genes they have. That seems to be the rule.

(Laughter.)

DR. COLLINS: Well, here we are. Let me try to get back on track here.

(Laughter.)

DR. COLLINS: Are you all getting scared over here?

(Laughter.)

DR. COLLINS: Some of the things that are being focused on with regard to the basic science component of this, the genomics to biology, include the following. We need to understand that 0.1 percent of the genome differs between individuals, because that carries within it the clues to common diseases like diabetes and heart disease and mental illness and hypertension and on down a very long list. It carries within it the clues to differences in drug responsiveness. Many of the things that Wylie and Nick were talking about this morning could be understood in a much more effective way if we had that complete

structure, not only of all the SNPs in the genome, as Nick was talking about, but also how they correlate with their neighbors into something we call haplotypes. I'll come back to that in just a moment.

We need a lot more sequence data. Having the sequence of some organisms only gets us more hungry for more, and now that we have the sequence of the human and an advance draft of the mouse and a pretty good draft of the rat, and increasingly other organisms coming through the pipeline like the chicken and the zebra fish and the chimpanzee and a host of others, the information we learn by looking into this comparison between genomes is profoundly interesting and really does give us in many ways our best handle on understanding function.

In order to achieve that, we believe that we have to drive the cost of sequencing ever downward. It currently would cost me about $50 million to sequence one of your genomes to a high degree of accuracy, and we clearly can't afford to keep doing it at that rate. So the technology not only for sequencing, which is highlighted here, but for many other applications, like genotyping and looking at gene expression, has to come down in cost, and we will only see that happen if we invest in it.

But imagine how things would change if we could today sequence the genome for $1,000. Imagine how that would change the way we practice medicine. You'd be very tempted, with appropriate restrictions on access to who gets to peak at it, to just get the sequence done once and for all and keep it as part of your medical record, and not have to go back and do specific genetic tests on the germline DNA for particular applications. You'd just have it all there, and as new information came along you could quickly in silico determine the consequences and the possible interventions for that individual.

Obviously, comparative genomics, looking at lots of genomes, is giving us a very good window into function, but we need a lot more, other windows, in order to understand that. We now think that about 5 percent of the human genome shows evidence of strong conservation by the evolutionary mechanism, and yet for most of that, about two-thirds of it, we don't know what it does, and we need to figure that out in a robust way that combines experimental and computational approaches.

Clearly, the focus on the proteins that are encoded by the genes is a highly appropriate one, although in many ways the technology is what's rate-limiting at the moment. But we are going to need to push on that and are pushing on that in order to understand how the proteins interact with each other to construct themselves into pathways and networks that carry out function, and how that goes wrong in the case of disease.

And, perhaps, if we do everything right, beginning with simple cells and moving into more complicated ones, we might be able to model a cell in some considerable detail on the computer, making predictions about biology without having to do a wet-bench experiment, or at least making hypotheses that could be confirmed at the bench.

So here are some of the more basic biology things. You may then wonder, well, what's that got to do with this Committee's enterprise? Well, I think a lot, because built upon this will be the clinical and the non-medical implications that this Committee will be wrestling with.

But let me move to that second floor, genomics to health. If we have a good handle on variation --

MR. MARGUS: Can I ask you a question about your previous slide?

DR. COLLINS: Sure.

MR. MARGUS: What is the time frame for those things?

DR. COLLINS: Well, they're different for each one. A thousand-dollar genome is probably the one on here that has perhaps been talked about the most in terms of what is the timetable for that. In order to get there, we really are going to have to jump curbs from our existing approach, which largely depends upon Sanger dideoxy sequencing and/or other enterprises, and obviously your company is engaged in one that people are watching closely. Single-molecule sequencing is a very important new approach that many people are counting on, as well. So perhaps we might get to that in, say, 2015 if all goes well, maybe even sooner if things go really well.

Defining the structure of human variation by this haplotype map, we aim to have that done in the course of the next couple of years; and sequencing lots of additional genomes, it depends on what you define as lots, but we'd expect to have another 30 to 50 gigabases of DNA sequence in public databases in the course of the next three or four years, or we're not doing a very good job with the possibilities that are there.

All functional elements of the genome. Well, you know, it depends on how rigorous you want to be about all, and how much you have to know about their function. That's a bit of a squishy definition. This blueprint that you can read about in the Nature article is not as focused on timetables as our prior five-year plans were for the Genome Project, because in many ways we're in a different circumstance. The Genome Project had a set of deliverables which were supposed to be produced by 2005. They all got delivered, and now we are in a circumstance of looking at a much broader array of opportunities going well beyond what was contemplated for the Genome Project, per se.

I think as we move along, some of these goals are going to need to be tied to more specific timetables. Right now, they're sort of put out there as challenges to the scientific community, saying can you do this? If you can, the consequences will be substantial. There is more information in the article than I have time to put on the slides, and some of it does get more specific about the timetables. It's not as squishy as it may sound from what I just answered in response to your question. But I do think this is a circumstance where we're trying not to be overly restrictive in terms of only putting forward things that could be done, say, in the next five years. There are certainly things here that reach well beyond that.

Genomics to health. Here is perhaps the one that I want to dwell on the most because it is perhaps the most relevant to where the future of genetics is going in terms of its medical applications, to actually identify the genetic and the environmental risk factors for all common diseases, and to do so with those things being studied in concert as opposed to separately, because obviously a lot of the important revelations about common disease are going to come from an understanding of how heredity and environmental triggers interact with each. I'll come back to that in just a moment.

We also need to push forward on some of the things that were discussed in some ways this morning, sentinel systems that would allow you to detect disease before symptoms have appeared, and also ways, as Nick described, using things like microarrays, to take a disease where currently you lump what is probably several different conditions together under one label and you distinguish them by a careful understanding of the differences in their molecular taxonomy.

An area that I think many of us are quite excited about but which is really quite a paradigm shift for academic researchers would be to put into their hands the kind of capabilities which are the mainstream of the pharmaceutical industry by allowing academic investigators access to high-throughput screening of small molecule libraries to identify compounds that act as agonists or antagonists for particular pathways of biological interest. Those, of course, also could become the first steps towards drug development, and

that kind of greater partnership between academia and the private sector, in many people's view, would be a very valuable direction to go in.

One that I think we ought to think carefully about here this afternoon is the need, if all of this is going to come to pass, to have really large human cohorts in order to try to understand genotype/phenotype and environment correlations. I think we've learned over the course of the past several years that those kinds of studies really need to be set up in a fashion that's unbiased. They need to be large, otherwise one tends to draw conclusions that don't end up getting replicated because you're looking at relatively modest contributions from any particular gene variant that's involved in a disease.

And yet, many people perceive there to be barriers to this kind of large-scale cohort study, and I want to come back to that in a moment.

Also in this genomics to health floor, clearly we need not to limit our studies on genomics to any particular population, and certainly it would be a mistake to focus, for instance, on the majority population in a particular area. We need to understand whether health disparities have a contribution from heredity, as well as from other areas which are probably more likely to be involved in most instances, such as socioeconomic status, access to health care, and cultural and dietary practices. But we'll never really know unless we in fact carry out those studies in a rigorous way.

And perhaps you will say this is a bit idealistic, but I don't think so. The ability to use genomics to unravel the causes and potential treatments for conditions such as malaria excite many of us, that here's a science that does have an opportunity not only to touch upon people in the developed world but also to go after diseases that have been largely neglected with a new focus built upon the field of genomics.

So let me just say another word here about how we're going to get to this point, because that obviously would change many of the issues that we would be deliberating about around this table. Frankly, most of us are carrying risks for future illness somewhere in our DNA, and at the moment we don't have the ability to know very precisely what those are except in instances, some of which Wylie described very eloquently, where we already are beginning to get a handle on those conditions. But for the most part, we don't. How are we going to get there, and how soon will we get there?

Well, what do we need here? We need this catalog of human variation, and yes, we have 4 million SNPs, and yes, by August we'll have 6 million SNPs, and we need to put those into a map that organizes that variation across the chromosome, so-called haplotypes, because that will be a wonderful shortcut to using that catalog to identify the variants that are associated with common diseases. We need better technology, as Nick pointed out, in order to apply that in a cost-effective way to make associations of particular variants in a particular gene with a disease risk or with drug responsiveness.

And, I would argue, we also need advanced methods for collecting environmental exposure data. If we're really going to understand how those genetic susceptibilities interact with the environment, we need to measure the environment, and that is at the present time something which I think there is a fair amount of expertise, but it's not widely shared with geneticists, and vice-versa. We've got to get these fields together in terms of those who understand heredity and those who are focused on the environment and convince them that they're not actually working at two different purposes. We're working at the same purpose.

Then, if that's going to happen, and here is a case where I think many of us are looking in some optimistic way for perhaps a really new and bold enterprise to emerge, we really need in this country a large cohort study of perhaps half a million individuals who are carefully followed over the course of several years for

whom a consent has been obtained in a fashion that is able to stand up to all possible standards. They will be involved in an ongoing way in such a study, if it could be mounted, so that it's not a one-time analysis.

The incidence of various diseases would be noted over the course of that timetable, very careful records of diet and other environmental influences could be kept, and extensive DNA genotyping as well. If you go through the expectation, that would finally give you an opportunity in an unbiased way to determine what the effect is of a particular variant on disease risk and how that interacts with the environment.

It is fine to do a lot of disease-specific case/control studies, and we're all doing lots of those as well, but they're often chosen in a fashion that they emphasize the more severe end of the spectrum of the disease, and therefore they may tend to overestimate the genetic contribution. If we're ever going to sort that out for common diseases, this kind of a large-scale cohort, as is currently being contemplated in the U.K. with their BioBank program, as Iceland is doing in terms of the whole country in collaboration with a company called Decode, as the Japanese are just beginning to mount with their own BioBank program which is about to get underway, but here in the U.S. we do not have such a plan.

If we're serious about health disparities, for instance, we need to have a plan that involves adequate sampling of some of the minority populations in this country. Otherwise, we will end up again not quite clear on what's happening there. So that's a need, I think, that we need to address very soon.

If we do this all right, both in terms of understanding how to measure genotypes, how to collect environmental data, and how to carry out large-scale studies, there's no reason we can't identify the major contributing genes for the common diseases that fill up our clinics and hospitals in the next five to ten years. That really would, then, position us to be able to offer people the opportunity of a multiplex kind of test to discover what one's individual susceptibilities for future illness might be, focusing of course on those conditions, as Wylie made the point very clearly, for which some intervention is available, because I think those are the ones that people are going to be most interested in. Of course, that will be a subset of the total for which such testing can be accomplished.

That will then put us finally in the circumstance of being able to move into the lower part of this diagram, which is a time description which we hope to traverse from top to bottom for disease after disease over the course of the next couple of decades, and ultimately, of course, get us down here to the point of being able to offer therapies for conditions that we currently don't have very good solutions for. Again, I think Wylie did a wonderful job using the example of hemophilia, showing how many of these various arrows can be traversed, but they do take time and we can't expect that they're going to happen overnight, and the relative speed with which an effective therapy arrives is probably the least predictable of all.

So, I've now touched upon those first two floors. What about genomics to society? What did we put in that particular part of the building? Well, several bullets are here described, and they're ones that have already been brought up, most of them during the course of just this first half-day of discussion of this Committee.

Clearly, there continues to be concern about genetic discrimination and genetic privacy. We have major issues, and Lawrence Sung will be talking about them later on this afternoon, about intellectual property, and we should I think move beyond the debates about genes and whether they should be patented because, frankly, that horse is very far out of the barn, and in fact pay more attention to some of the other entities for which that has not yet been settled, such as haplotypes, such as expression data, such as protein crystal structures, all of which are also contemplated as being intellectual properties in a way that may or may not be good for the public in the long term.

Very much I think on many people's screen and high on the list of things that we need to pay a lot of attention to as we focus on the study of variation is how that study might reflect usefully on the topic of race and ethnicity, recognizing that this is not a simple issue and that it is not so straightforward to simply say that race has absolutely no biological basis, as has been I think occasionally said in too strong a fashion. Race is basically a reflection in a very fuzzy way of ancestral geographic origin. Ancestral geographic origin is a reflection in a fuzzy way of genetic variation. They're not completely disconnected, but they're very fuzzily related to each other.

Now, how do we formulate that message in a fashion that is benevolent and actually provides a useful commentary on a dialogue which has often been contentious?

As we understand this variable part of the genome and apply it for medical purposes, it will clearly also be applied for non-medical purposes. We are going to uncover in the next decade or so variations in the genome that play a role in such things as intelligence and sexual orientation, and there will be such discoveries that actually, after people test them, are validated, unlike the ones that have been reported in those areas for the most part up until now. How are we going to fold that into our social discourse and our understanding of ourselves and our fellow human beings?

In that regard, are there boundaries that we don't want to cross in terms of the applications of genomics in the non-medical arena? And if so, who establishes them and who enforces them?

So those are some of the things that we think are most deserving of intense attention in the coming years. Again, this is not my list so far. This is basically the list that this group of 600 advisors came up with and which we formulated into this prospectus for the future. But now, in a somewhat more directed way, and again without being able to completely defend this, because I think there are so many different topics that might have been proposed, let me just mention a few areas that arise from this list that might be appropriate for focus by this Committee, recognizing that tomorrow is largely going to be the point at which that discussion goes forward.

I can't help but put genetic discrimination first. I celebrate the accomplishments of the Senate HELP Committee and I'm delighted by comments from Dr. Rowe that the industry is supporting this bill, and my hope is that the House of Representatives will act quickly on the same kind of bipartisan basis to achieve what we've waited for now for seven years; that is, effective federal legislation that will outlaw the use of predictive genetic information in health insurance and in the workplace. So we need to achieve that. We're not quite there. It's a great moment that it's finally made it through a Senate Committee in a bipartisan fashion, but there are several steps still to go.

Let me then be bold enough to say that while the complications in terms of adverse selection issues are much more complicated than these other types of insurance, and I think that has kept people from wanting to even engage on them up until now because of the need to focus on health insurance in the workplace, perhaps it is time to think are there options, at least for some floor, some minimal level of care that could be, in fact, considered in terms of life insurance disability and long-term care insurance where, again, somebody with a high risk based upon genetic information would not be completely screened out or to the degree that it became unaffordable.

Basically, there hasn't been a lot of discussion about that in this country. There's been a fair amount of discussion about life insurance issues in the United Kingdom, in part because they're not as worried about health insurance because that's covered. But perhaps it might become time to begin to consider that.

Of course, there are other areas of genetic discrimination that have not, I think, received as much attention which are not part of the insurance issues, but the notion that your genetic information might be used against you in other considerations, such as an adoption proceeding or in ability to gain an education or the military, and there are dots here because there are others you can think about as well. So that was bullet number 1.

Bullet number 2, and it has already also come up, we do not at the present time have an effective system for overseeing genetic tests to ensure that clinical validity, and hopefully clinical utility, but at least clinical validity has been established prior to marketing. The current system, as we all were part of the SACGT's discussions, does not allow confidence that a test has achieved that kind of status before it becomes marketed to practitioners, and sometimes even to the public. I think that is an issue which we continually are concerned about and need to return to as part of this Committee's discussions.

A special concern in that regard is the proliferation, mostly on the World Wide Web, of direct-to-consumer marketing of genetic tests, some of which, I must say, are of deeply questionable validity and for which at the present time there seems to be no particular oversight whatsoever. I show you as an example this one from the Web. This is a company that is offering to concerned parents genetic testing for the millennium, as it says here. I'll quote from their Website: "Are you concerned about your children's future? Does your child have the genetic trait that leads to disruptive and addictive personalities?" I'm not quite sure how the parent was supposed to know if the child had that genetic trait. Maybe they had a bad day in school. "DNA testing can help you understand and manage a child's behavior before it gets out of control."

You go down here and it tells you how to take a foam-tipped applicator and rub the inside of the left cheek 25 times, send your DNA sample off, have it tested in some way, and then notice if, in fact, the test comes back indicating some alarm. They will then sell you some neutriceuticals at a considerable price that will perhaps protect your child from a terrible outcome.

This is junk science, and it is not the only example that one can find out there on the Internet of similar such things that are happening in greater and greater profusion, and they run the risk, I think, of perhaps fouling the nest here in terms of convincing the public that genetics is junk science in general. If we don't have the ability to restrict in some way the marketing of such information, we may later find out that the public has concluded that this whole field is not something to be trusted.

Just two more areas that I might suggest based upon the predictions of the future enterprises that the genome enterprise might be engaging in. I must say, when I speak to researchers who are most interested in seeing the medical advances occur in terms of connecting up genetic variation and environmental exposure with disease risk, they are deeply concerned that our current system, with a very uneven focus on protection of human subjects, is making it increasingly difficult for clinical investigators to do research.

Now, I grant you, I think those protections need to be there, and they need to be very rigorous and extremely well thought through, and we have representation here from Dr. Carome from the Office of Human Research Protections. But I think the conclusion of many clinical investigators is that somehow we've built a network that is so complex and so restrictive and so difficult to deal with that it's beginning to get very hard to do research. The public has an interest in the research getting done as well, and perhaps we need to reconsider whether we've got the balance right here or whether there are actions that could be taken that would make it more feasible to undertake large-scale studies of the sort that we really need if we're ever going to sort all of this out.

I don't know that this is an easy question to deal with, but I think it is number one on the minds of many investigators. I just came back from the Cold Spring Harbor genome meeting, where most of the world's major investigators in genomics were gathered, and this was much the discussion at the meeting, both during the meeting and in the hallways afterwards.

My fourth bullet here -- and again, I could go on with a much longer list -- relates very much to other things that have been brought up. Again, I was very impressed with Dr. Rowe's presentation about what Aetna is doing here in terms of making sure that genetic services are offered and are connected up with adequate counseling. That's a wonderful step in the right direction. I'm a little less optimistic that that's going to be happening in quite such a broad way as one would like. There are workforce issues here in terms of who is going to be providing the expertise that's needed in order to interpret all of this information and provide the kind of counseling that the public is going to need.

There are chronic access issues about who actually is able to get the information, and related to that are cross-cultural issues. Are we really prepared to deal with the very different ways in which different people may assess the information and need to have it explained? And reimbursement. Who is actually going to pay for all of this? Where is that going to come from? Those are clearly issues that have to be solved in the next few years or this revolution in availability of genetic information, which I think everybody agrees is coming, may encounter a major problem in health care economics.

So those are a few ideas of areas that might be attended to. Again, I look forward to this Committee's deliberations, and I count on them turning out well, and for that particular optimism I refer to a particular verse from Proverbs, which says "Plans fail for lack of counsel, but with many advisors, they succeed." We seem to have many advisors and expert ones around this table, so we shall count on success.

Thank you very much.

(Applause.)