2015-02-19_Richard_Myers_converted.mp3

And, you know, the science was not -- the biology was not so good.

[unintelligible] I'm sorry. We are rolling, whenever you want to start.

Okay.

So, just to start out, just tell us who you are, and where you are at this moment in time.

So, I'm Rick Meyers, I'm at the Hudson Alpha Institute for Biotechnology in Huntsville, Alabama. It's a relatively new, six year old non-profit research institute.

Alright, so to start off, tell us about your time at Stanford, and the significance of Stanford for the history of population genetics, The Human Genome Project, and population genomes, which is a very broad term.

Okay, actually I started that at UCSF. So my first faculty position was at UCSF and I started working in the mid 80's, I went there and started working with David Cox, who was a medical geneticist, and he and I sort of joined at the hip immediately when I got there. And our two labs were almost joined as one lab studying genetic diseases. We were really trying to look at Huntington's disease, several others as well. And part of that was David had developed a mapping method called Radiation Hybrid mapping, and I was getting into mutation detection and things like that -- I had done that as a post-doc. And that led us to, you know, be participants in the first grants that were funded for being genome centers. We were one of the first five -- four or five that were funded in the United States in 1990. And it was because of the mapping that we did that.

Now, I should jump backwards a little bit because, even though I was a bio-chemist and I was a molecular biologist, and not really doing anything in genetics or human genetics, during my post-doc years I almost accidently started developing methods for mutation detection, as a spin-off from a mutagenesis method I developed. I was with Tom Maniatis during my post-doc years. And because of that development, Tom got invited to an early human genetics meeting -- or an early meeting in the 80's. I think it might have been '84, at Alta, Utah, and he didn't want to go, so he sent me.

And it was about 20 people, and we were talking about how to measure the mutation rate. The germ-line mutation rate in humans, and it's so low that people were doing all these cell-biology type methods and various things, and I had done these -- developed these methods for mutation detection. Maynard Olson was my roommate at this meeting. And it turned out that people made the statement -- somebody did, not me -- that the rate was so low we're going to have to sequence the human genome in order to figure this out. And that's actually when the Department of Energy, who funded that meeting, went back and I think came up with the notion that we ought to do this in '85, '86 or so, which of course, you know, snowballed into NIH and Sanger Center, you know, -- DOE, too, in leading the way on doing that.

So now jump back to the late 80s when we were writing our first proposal to be part of the Human Genome Project. Started in -- the grants were funded and they started in October, 1990. And so I was still at UCSF until '93, David Cox and I were there, and we started mapping. It's so interesting to think about how crude the technologies were back then because we certainly didn't know how to do much DNA sequencing then and even the mapping methods -- there were genetic maps and physical maps -- we built back and then -- I mean, yak-clone maps and then later back maps of the genome. Physical maps, contig maps, but radiation-hybrid maps and genetic maps all trying to tie that together to get a scaffold for -- and to give the clones, frankly, for sequencing the genome. So we really -- you know, most of the sequencing didn't start until seven or eight years later. We were not even sure that it would work then. By then, capillary sequencing had been invented, so if we had had to continue to do slab-gels even though it was four color fluorescent slab-gels, if we had had to continue doing that, I don't know if we would have ever finished. So the capillaries came along. They didn't work so well at first, but by 1999 or so, the machines got better and were able to do it. So my participation in the Genome Project was partly working with David on the mapping parts but then we started doing sequence finishing for three of the human chromosomes with the Joint Genome Institute, and that's the project I led by then at Stanford. And I guess that started in the late -- you know like 1998, 1999 or so, and then we got the draft in 2001 and the real finished sequence in 2003.

But that's all very -- somewhat different from your work in, say, variation research.

Yeah. Right, so yes. You asked me about how DNA sequence variation fed into this. Well I mean I think the whole idea behind the Genome Project was that so much money and time had been spent, you know, trying to find genes like Duchenne muscular dystrophy and the cystic fibrosis gene, which was reported in 1989. You know, hundreds of millions of dollars, many, many labs going after one disease, and that just was ridiculous. We had thousands of diseases. There was no way we would do that. So I think a big part of the notion, initially, was human genetics. Let's try to figure out how to find these genes. And in order to find genes you need -- you know, back then, almost everything in terms of disease genes was -- were Medelian disease genes. Ones where you had families, you had DNA linkage, you had to use the genetic map to help, you know, hone in on the region, and you know even in the late 90s, you know, and early 2000s we would still have to walk around the regions to find clones that we would then look for transcripts and then look for genes in those regions. And -- so you had to do all the mapping and identification of clones that you would then want to look at DNA sequence variation to find mutations.

And so my early involvement, what actually got me into human genetics in the mid-80s was these mutation detection methods. There were two methods that we developed when I was a post-doc at Harvard with Tom Maniatis that, in retrospect, were pretty crude, but they were not RFLPs, they were not southern blot related, they were not using oligonucleotides. They certainly weren't using DNA sequencing yet, and they were methods that screened DNA fragments as to whether they had variants in them. One of them was a funny kind of gel system called Denaturing Gradient Gel Electrophoresis that separated DNA molecules on the basis of their melting behavior. And a single base change can change the melting behavior. Leonard Lerman, one of our collaborators, had discovered that and invented this gel system and I then developed it and then applied it. And it actually was used by quite a few people for a while.

Then another method where DNA variation -- and again, this is primarily to look for mutations and disease at the time, at least. And this method was used -- it was RNA-DNA hybrids that you would make and if you had a mismatch you would cleave that measure, the fragment lengths of the cleaved RNA probes. And both of those methods -- and then others came out that were similar, or even more sophisticated or whatnot, it was really only when we started sequencing that, you know, you could really think about this being done on a population scale, for instance. And of course that's made all the difference in the world. Now we find mutations, you know, without linkage, without knowing anything about where on the genome a gene might be, we can find them often just by sequencing deeply since it's so fast and cheap now.

Right. You're all of a sudden touching on about six or seven different things so I'm going to try to unpack it a little bit.

Okay. That's fine.

Not for me. For people who might be watching this. So, one of the things you've said that I found quite interesting that I understood from my research in the archives is the overall general crudeness of early genome sequencing and mapping techniques. For you, from this perspective, in 2015, what are the real, sort of, shifts in sequencing technology and mapping technology development that you have seen? And, in particular, after you answer that I want to hone in something very specific about the GES-TECH grants at the NHGRI and, sort of, their significance for early technology R and D.

Okay. So, I mean, it is interesting to look at the evolution of DNA sequencing. I did lots of DNA sequencing as a graduate student and a post-doc in the late 70s, right after it had been invented. Sanger sequencing and actually Maxam-Gilbert sequencing, radioactive -- very slow, and very error-prone. I sequenced a lot for a time and, you know, it's laughable now, of course, but the error rate was probably, you know, one, two, three percent, and you didn't sequence things over and over again, so it was really hard to get with these manual methods. We thought it was a god-send when four-color fluorescence was invented and it certainly sped things up but you still were running gels and having to track lanes and things like that, and so it was still very slow. So we went from a few hundred base pairs that you could maybe sequence in a day, or if you were, you know, maybe a thousand or so, to, you know, much, much different even by the time the four-color fluorescence was invented. And then the capillaries made a big difference. Because it just spread -- it sped up -- you could just do higher through-put, but it was still remarkably slow.

So here's a good way to think about this. In the peak of the Human Genome Project, after we had really gotten the capillaries to work and all these public laboratories were working together, I think we calculated that in our peak year we sequenced about 125 million pieces of DNA. These were, you know, with the technology where you had to clone the fragments, you had to pick out the bacteria colonies, and you had to grow them up, and you had to make DNA preps, and then you did the chemistry on them. 125 million, and that was huge numbers of machines and our lab was not one of the biggest ones but we had 25 sequencing -- capillary sequencing machines, and we weren't even one of the big centers. And, you know, now we're running, in 2015 with the

fastest technology, we're sequencing six, seven billion pieces of DNA in two days on one machine run. You know, now, these are smaller pieces, but they still -- I mean, the speed of this and the number is just phenomenally different.

Now that came about -- there's a massive change in the way people thought about sequencing when that happened. It went from, you know, cloning and preparing fragments of DNA, which were all that we knew -- and yeah we used robots to help pick the colonies and we did, you know, some robotics for the liquid handling and things like that. So that helped us speed it up, but it still was remarkably slow. And the difference now, with the new technologies, is that you don't address each little piece of DNA, you just throw them onto a surface -- they land on a surface. It's much, much miniaturized, of course, and, you know, because DNA is very sub-microscopically small you can fit a lot of DNA fragments onto a small microscope slide. Something -- you know, billions of them. And so that was fundamentally different, they land in a place, they're held in a place, they're amplified and the sequencing happens, and you read the sequence as its coming off on millions of little spots that required lots of -- very high resolution camera development, and in fact all the improvements we're getting now, many of the improvements, are better and better cameras so that you get higher and higher resolution to detect, and then also to detect a signal from a few hundred molecules that you've amplified in each in situ place. A lot of the image analysis, in fact I think they use some of the same analysis that was used in astronomy to address where you are on this image of, you know, several billions spots that change with each cycle of DNA sequencing.

So that fundamental change is -- and when you calculate that compared to the beginning of the Human Genome Project it's at least a million fold faster, cheaper, better. Huge, huge, huge difference. So you start -- you know, that's why we can imagine -- not just imagine, but now use these sequencing methods in the clinic and really imagine having everybody sequenced as part of their healthcare.

Yeah, and -- but also from my perspective [unintelligible] I'm very interested in, sort of, early granting mechanisms and technology development. The most fascinating little aspects I've found are these so-called GES-TECH, these early sort of proto-center grants. And you were one of the early grantees. And so describe, sort of, what that was and what that contributed to, if you can recall.

Sure, well, I'll recall the best I can. I mean it is interesting that the first grants were actually production grants. But the truth is everyone was doing technology development because we really didn't know how to do it when we first started. So 1990, we had a -- I mean, it's embarrassing to think about how little ambition it seems that we had. We were going to map 400 markers on chromosome four with a giant grant over a five-year period. Now it moves so quickly that all of us did way, way better than that, and it moved to, you know, a different model of getting going and having fewer centers. We had about 20 centers at the beginning.

But it's interesting. Technology development was and continues to be a critical part of all of this. And while you want to do technology development when you're generating large amounts of DNA it actually helps to have -- large amounts of data, excuse me, generating large amounts of data, it's really hard to combine production with development. You have a pipeline that's

going, it's working and if you want to change something in that you mess up your production. So most groups -- and this is when the grants came in, and I don't remember when they started calling them GES-TECH grants, it was early on, it wasn't the very first ones, but -- where you would really need to have a development arm, almost separate from your production arm, but not too far away because you don't want to develop something that doesn't need to be done. The production environment and the people doing that truly understand what the problems are. Where, you know, this step is the right limiting one, for instance.

So that was a hard lesson, and in fact, I think we -- early on tried to learn something from industry. A bunch of us from NHGRI went to the Motorola cellphone factory in Illinois and learned how they did their assembly line. And one of the lessons there is that every step of a 30 or 40 step process, every step had some sort of Q.C., some sort of quality analysis and quality control. And you would stop it if it didn't fit the quality control, so we tried to implement those kinds of things into -- so that's part of technology development too, is making a start-to-finish pathway -- pipeline. It's not very glamorous, it's pretty boring in some ways when you think -- you know, it's not some new, shiny technology, it's putting the pieces together. And we still have that as one -- now, we have that now with one that goes so much faster than it did before but we always want more, faster, better, cheaper, you know, and higher quality along the way.

Rather embarrassingly, I actually -- one of my big research topics is quality control and [unintelligible] project. So I consider that to be -- and the development of the data coordination center, that model for all the subsequent genomics programs. But no, you absolutely make a central point, and I think this leads me to two further questions. I mean, could you speak a little bit more to the idea of -- you have, sort of, the NHGRI and the government, sort of, development of things and the program and the public in government industry collaboration which you haven't had with AVI and Illumin-Ed [spelled phonetically]. How -- is that a consistent feature in genomic science from the beginning?

It is interesting. So there were new developments during the Genome Project, the SNP Consortium being the first one, where it was a partnership between industry -- David Cox actually played a big role in having to set that up, as did others obviously, a lot of it was NHGRI. But -- and that was a new thing for industry to be that open where they would say -- and these were competing pharmaceutical companies, mostly, "Okay, this is going to be good for everybody so we'll get SNPs at least DNA sequence variants out there." And that worked pretty well, I'm sure there were ups and downs. And then in HAT-MAP, you know, they continued to do some of that. I don't think we do enough of this. I think there are too many barriers, there are often barriers because of NIH -- or, sorry, government funded agencies not being able to do that. There are road blocks, I think, that are unnecessary, but not completely. And then industry flirts with us off and on. We just had a project funded by a large bio-tech company that -- they wanted the data to be open and it helps the whole world. It's just wonderful. I think we need more and more of that.

I do think that we suffer and we did throughout the Genome Project from near monopolies in some of the technologies from the companies and while the technologies have gotten better and better and while we quote, collaborate with them the big centers do, especially, our center do, in

the sense of we're still vendors, I mean they're still vendors to us. We buy stuff from them and they -- I think that often the advances are more slowly incremental than they could or should be because of business decisions, and that's where competition would come in and, you know, minimize that. There's a little bit of competition, not enough now.

Nevertheless the technologies have really, really, you know, soared. They are way, way, way better than they used to be. And I think, you know, often, I will say this -- this will not make me popular among some of the companies, but often they use us to figure the bugs in their systems. Every new sequencing machine from the beginning -- the big centers, and actually other users, you know, the machines weren't working well. Every time they would release them they wouldn't work so well and we'd work out a lot of the issues with them. Some of the best analysis technologies came out of the public sector, out of the Genome Institute, for instance, that, you know, the companies reluctantly ultimately adopted because they were -- you know, wanted you to use their own. So that continues.

I think it's just part of the -- and some of that tension I think is fine, I -- you know, it doesn't have to be, you know, altogether for everything. I do think the promise of being able to use this -- and it's not just human health, I mean, this -- we're using this in agricultural genomics, and other things as well, really, really powerful, that there's a huge, huge market size for that so I think there's room for a lot of people to be in it.

Yeah I'm -- you also, and it's typical of this interview, seem to mentioned about six other things.

[laughs]. Sorry about that.

It's fine, it's fine. We're just trying to get through it all. So, one of the things I also wanted to ask you about was -- you discussed SNP consortium and one of the most important, sort of, elements of the pre HAT-MAP period is, sort of, how the consortium becomes the HAT-MAP project. And sort of the early days of various -- sort of, a big push for variation research, you know, in '99 and 2000. What are your thoughts or your reflections on that?

Well the SNP consortium really was only to find variation, and to catalogue it, and to make them -- you know, make the variants available. Of course back then that meant they were PCR-able markers that you would have and done in that way. And I thought it was set up well because they looked broadly and they -- you know, there was, you know -- but it wasn't really the use of those, so much, it was just finding them.

So the HAT-MAP consortium came about. I was actually not part of it but I was an advisor for Francis during that time because he had an external advisory committee, so I got to see that up close and it ran really well. You know, I'm sure that there was in-fighting probably, just like there is in everything, but they -- it was really a consortium. And, you know, it wasn't -- it was one approach towards trying to find out how you use variation to find diseases. A lot of controversy because a lot of people think -- who don't really understand genetics, I think -- think that GWA studies and looking for, you know, complex associations between variants and complex diseases was a waste of time. That's not true. It is hard. We're only finding a small

amount of the variants that account for these complex phenotypes. But it led to a lot of technology development, and a lot of cohorts being developed, and some of those, are -- and it certainly led to results. I mean, there's no question we found results. There was a lot more going on in addition to the HAT-MAP consortium.

HAT-MAP consortium was meant to be like the Human Genome Project, it was going to provide a resource of the haplotypes which is still very valuable to have those. Now, of course it was only four populations at the beginning and now it expanded out. And that information, even for a rare disease, that kind of information is very helpful. But the world is changed, and the technologies now allow you -- because -- the find these via sequencing, you still might use the variants that you get and the haplotypes that you got from HAT-MAP, but you -- now we sequence and try to find what we can. And some of that's -- you know, for complex disease, a lot of it's for rare disease as well. Simple genetic disease, so.

Yeah, I -- there's a definite school of thought that -- I mean, I can't tell you that everyone in that school of thought understands genetics perfectly, but to say that there is kind of a division between people who think GWAS is sort of the answer to the, sort of, genetic foundation of complex diseases versus genomics and environments of GNE. From what I've read of your research you think that those approaches are extremely complimentary.

Of course. I think whenever anybody says "this is the only way to go," you should immediately become suspicious. It's just never true in that way. Also -- and I think we've learned a lot, and certainly environment. I mean, it's ridiculous to think that this is all genes, but I'm also surprised because we are now able to sequence unexplained disorders especially in children but in adults as well, that many, many, many of these are simple, they're single genes. The reason we didn't find them is the rare -- there are two reasons why I think we didn't find them before. They're relatively rare. When you look at mental disorders -- intellectual delay and decline, there may be 5,000 genes that, when mutated, either alone or even in combinations could end up leading to these fairly severe early phenotypes. And we're finding lots of those by sequencing them. We could never have gotten them before, we would not have gotten them with HAT-MAP, or with linkage because you didn't have multiple family members. And it's partly because they're rare, but actually there's a new, I think it's relatively new knowledge even though people knew that de novo mutations happened. When you have severe phenotypes it looks like, at least in our case where we're looking at a few hundred of these, of children with these unexplained disorders, in our case so far about half of them -- I mean, the hit rate for identifying the genes is much, much higher than I ever expected. Almost 50 percent, and it probably will get better, and half of those are de novo mutations, so they would have looked like complex disease maybe, because you don't -- you know, you don't -- you can't find linkage, you don't find family members, and it doesn't look inherited in that way, and yet, it is, essentially Mendelian because it's one gene that's mutated in the sperm or the egg of the parents that led to this child or this person with this disorder.

So that's not going to explain everything by any stretch, but to me it's a surprise because it's a lot more frequent than expected. I think complex disorders, you know, like rheumatoid arthritis or, you know, things that we know are common, and have a genetic component because they're inherited -- sorry, they show inheritance in families. You can do twin studies all sorts of things

that tell you that genes are involved, but it's not simple genetics. I think all of that information together when you then combine it with some sort of functional analysis, like all of the data we have for histil [spelled phonetically] markers, and transcription factors, and small RNA's and everything that's coming out of the road map projects and ENCODE project and things like that. When you combine it with that and some very clever analysis that tools that people, like Greg Cooper and Jay Shedure [spelled phonetically] and David Goldstein and others are developing for saying "are these functional variants or not?" A lot of people are working on this. I think that's helping to then put the GWAS results into perspective, at least with regard to how genes play a role. And then, you know, the environment part is the hardest part, of course, because you have to -- I think functional essays and mass models and cell models can help to try to tease that out as well.

I mean, but you -- would you say that there was a specific point in time when genomics and epidemiology really started to play along, or has that always been the case? But -- you know, the epidemiology was harder to measure and it wasn't as mechanistic as the GWAS stuff and--

Yeah. I think that some people -- David Cox again was a very close colleague and we worked closely together for many years. I remember him pushing -- he might not have used the word epidemiology at the time, but pushing, you know, the population studies need to be done, Aravinda Chakravarti, a bunch of people got this early on, I think. Getting the epidemiologist, who were from a very different culture to then work with the genomic side and the human genetics side was, you know, took some time, but, you know, people, some people started doing that earlier. Clearly, that's a major part of it, and the epi -- what the epi does for you is helps to, at least, get some of the history of this so that you also are at least thinking about environmental exposure. We know, for Parkinson's disease and things like that, you can sometimes even definitively show that, you know, some environmental factors play a role, at least in some of the cases. I think that is our -- it's hard because -- and it's not even a cultural thing now that's the hard part, it's the science is hard, but it's clear -- and tools are getting better, I think, but clearly that combination of those needs to be put together for, you know, for trying to crack these problems.

Yeah. Because historically, I mean, you have -- you had, you know, a fully throttled, at least theoretically, GWAS, you know, understanding being put forward by David Altshuler right along with HAT-MAP, and kind of the broad way --

Yeah, yeah.

-- of looking at population genomics. And then, you know, you have GWAS, GWAS, GWAS, and 2007 is the, you know, the year of GWAS and everybody's going to Stockholm. And then to -- but then around 2009, you finally have, you know, genes-in-environment initiative, and so is that just because the -- you know, the science is just that much harder to --

Well, I think one of the things that happens in all fields of science is that you get paradigms that -- and strong personalities that -- and results that end up pushing them forward, and I've always thought that it's never all or none for anything. So I think that sometimes those are over played. It sometimes hurts the field because when you draw a line in the sand that it's my way or the

highway then you usually that means you disenfranchise a whole group and then you end up instead of working together end up fighting about it.

So clearly, I certainly don't -- never felt that GWAS was useless. I do think it was overemphasized that we should still be doing it, but not at the expense of -- not avoiding some of the other things. And so clearly they've come back, and it's -- so there is this phenomena in human behavior that phenomenon where people, you know, it's almost like camps, and you're on my side or you're the enemy, which is, of course, just silly. Science doesn't work very well that way. You'll have these paradigms going and sometimes going so long that they really wear themselves out, and I'm not saying that's happened here but you can see that. And then some new thing gets discovered or some new personality comes along and then it ends up shifting. It's kind of fun to study that. I think sometimes it's not always to the -- for the best science to have it happen that way.

And GWAS had some really fantastic early results, I mean --

Yeah, yeah.

There were issues --

And there's still, you know --

And still great results from all that silliness.

Absolutely. Some of the results -- I mean part of it now is interpreting some of those. We're finding where some of those hits are. We're finding, yes there is strong evidence that something is going on in this part of the genome. And you know, even if it's only a few percent of people with that particular disorder that can be explained by that or it's only a part of their disorder, sometimes these are really useful insights into pathways. I'm -- even think maybe even treatments. You know, it's sort of like saying the rare forms of Alzheimer's, and there are some really, really rare forms of a disease like that -- three different Mendelian disease genes. They don't explain -- they explain a tiny percent of people with Alzheimer's, and yet we know a whole lot more about the pathways. So I think that some of that comes out of GWAS studies, too. I don't -- I think we should never think it's all or none, and I think that was -- there certainly was a trend for a while where that was the only way that -- and, you know, granting -- grant reviews and even grant opportunities that come along often follow those trends, and sometimes you have to break away from them.

Actually, I mean, looking at the history of GWAS and then looking at some of the new stuff that's coming out of genes and environment is really fascinating --

Yeah.

-- from a historical perspective. So one of the things that we've touched on, and I know you had an office with Luca Cavalli-Sforza and, as you know, you related to people and --

Yeah.

And, you know, people have been interested in ancestry and, you know, the forces of -- you know, the forces of evolution and, you know, selection, drift, migration. How have -- these conversations have going on for a really long time. How have programs like HAT-MAP and how have deep sequencing and better genotyping and things like that, how has that improved our understanding of the historical causes of human variation and, you know, it's just --

[affirmative.]

-- who we are as individuals, and nations, and people?

Well, so -- but Luca Cavalli-Sforza was in the office next to me for sixteen years, and I actually knew him long before that. And even though I wasn't even a geneticist when I first met him and I certainly didn't do population genetics, I couldn't help but let this stuff rub off. He was -- is -- a remarkable man. Alan Wilson also, who unfortunately died quite young, was one of the real leaders in this. He was one of my teachers when I was in graduate school at Berkeley. So I got some exposure to this, and I actually had the good fortune with Luca, Mark Feldman, Greg Barsh, a woman named Wya Tang [spelled phonetically] at Stanford to do a population-based genetic study on the Human Genome Diversity Project samples that Luca had collected over the years, about a thousand -- well, many more than a thousand, but a thousand individuals we looked at from 50 populations and it's really interesting.

So that -- we published that in 2007. At the time it was by far the biggest study done like that, and it was 300,000 SNPs that we looked at across the genome, but what you'd learn from that, and -- I just am totally fascinated by this, you certainly learn about ancestry and human migration and human populations. Luca has been studying that -- and others had been studying that -- for decades, but they had very crude tools for doing this. I mean, including Alan -- I remember an Alan Wilson's during the '70s when I was in graduate school there, they were using antibodies in migration of polymorphic proteins and in starch gels to see, you know, if -- what the variation was. They weren't even looking at DNA yet. So that the fact we could do that with genotyping by the mid-2000's, and of course this was greatly helped by having the SNPs to begin with and having HAT-MAP and having other information, even some of the genes figured out -- the genes figured out from sequencing. That taught an enormous amount by looking at 50 different populations, an average of 20 individuals from each one of those.

Now I think those samples are probably fully sequenced, and we know a whole lot more about the variation, and what you learn is migration patterns. That's really interesting, it's actually really important to know this in disease when you're looking, especially for complex disease, to know ancestry. You literally can say, you know, this person, who is French, looks like her genetic distance moves her a little bit further east. Turns out she had Russian ancestry two or three generation before -- that she had not told us about. You can learn even which parts of the genome came from different regions. And this is not -- I mean I think it's redefined the way people think about race, that it's a continuum. You have very, very few variants that distinguish the outward differences from things like skin color and hair shape, and body shape, that those are

very, very recent, superficial variations compared -- because we have -- when you look at the variation that causes disease, for instance, that makes us who we are, you find that in every population in the country -- in the world, rather, that much of that is ancient variation, and it's the relative amounts of that -- of those variants that will that will at least contribute to, not cause, contribute to a higher instance of disease in this population versus that population. That's been really, really helpful and I think some of the population differences are ones that help us understand disease a lot better, as well as normal, basic biology. It's been extremely valuable.

So there is a -- some body of scholarly opinion that basically tries to say that the Stanford Human Variation Project was unsuccessful and the HAT-MAP project was successful. Why did --

You mean the one that Luca lead -- the Human Genome Diversity Project?

So from your perspective, why was the Stanford Human Diversity Project so controversial? I mean, you've talked something about, you know, methods --

Yeah.

[unintelligible] science --

Well, so --

-- difference in outcome --

Well I don't think it -- I disagree with the -- first of all, it wasn't the Stanford, it was the Human Genome Diversity Project. So it was Luca leading it, but there was people from all over the world that he, that -- that's collecting them and doing other studies on them. We took it late in the game, when some of the controversy had already died down. And the controversy had to do with disagreements about how -- I mean they were political, primarily. I think some of those were disagreements maybe about how they were collected. Certainly was not sensitive in the way -- the same sensitivity that NHGRI did when collecting the four population groups for the HAT-MAP Project, where a lot of effort went into doing the right thing here.

I don't think Luca or any of them were doing the wrong thing. They just didn't have the -- those kind of safeguards. Some of these were collected in the '50s and '60s. These were a long, long time ago. And that's not necessarily an excuse for it. So that probably is where the controversy -- I never quite understood it. It certainly didn't fail. People have used those samples -- it was just a set of samples. Now, it did not lead to a huge amount of knowledge about variation, about population variation, partly because a lot of that was done before there were good markers. You know, it was not coordinated in the same way. It did not have nearly the funding that HAT-MAP and others had. So that -- all of those probably contributed to it. It was good science though. I mean it was not, you know, it just, but it was slow science compared to, you know, what happened a little bit later.

I think -- I don't -- I haven't really followed the samples much, but I think a lot of -- I think they have been sequenced because their well characterized with regard to ancestry. When they

collected those, they would go to a place. They did get informed consent. They didn't call it that then, but they got consent. And they were the people who contributed their DNA, I think, knew what they were getting into. It was a genetics -- population genetics study. What was really valuable about it is all of the individuals from those collections truly were indigenous to those regions, which is really important. That's hard to do. If you do that in the United States, with people who migrated here, you're going to get people who barely know where they're from. They might think they know where they're from, whereas, when you go to some of these other populations and you'd ask -- the requirement was where were all four of your grandparent from? And if they were from there, generally, in most of these populations, they truly had been there for many generations. That was true for the Native American population. There are only a few of them in that set of cohorts, but if you took non-native Americans from any of the Americas, they're going to be, you know, add-mixed like crazy, so.

So, I suppose just to cap this one discussion off. I mean, how, in your opinion, do you talk about things like race and ancestry, and the significance of variation for disease, without being reductive?

Right. So first of all, I don't -- I live in the Southeast, I live in Alabama, I grew up in Selma, Alabama, and well, race was a big thing going on through my childhood and certainly now. It is at least good to see how things have changed dramatically. Laws do make a difference, actually. We still have a long way to go, but I don't use the word race, and it's not for political correctness. And the reason, I guess population geneticists might still do this, but we talk about geographical origins because, I mean, say you're African, north African versus sub-Saharan African, it's quite a different physically.

Now again, I'll repeat that these superficial differences like the way our faces are shaped, and our skin color, and our eyes, and et cetera, are a relatively small number of genes. We still -- we have the same, you see the same variants in every population, with a very, very rare exception, with a few that look like they truly might be population specific. So the problem with the word race is: Where do you draw the line? Because it's a continuum. It really has to do with our migrations out of Africa. You know, there are probably two, well there may have been more, but there are two major migrations. One going north to Europe, one going east and then populating the world in that direction. And those were ancient by human history standards but really, really recent in terms of human evolution. I mean, you know, we're hundreds of thousands of years old as a species and really more; you know, whereas those migrations happened 60, 70, 80 million years ago.

And so I think the way that you think about this is first of all, there's a huge amount of add-mixture -- I'm doing my part. My children are half-Asian, half-Chinese, so this huge amount of add-mixture. So that confuses, as well as, you know, makes -- it's a different part of our history now. The migration then led to add-mixture. And there are different times in those. I think it's important to -- it's interesting to know this stuff. I mean, most people are quite interested in their ancestry. I know I am a real mixture, mostly European, but a mixture of different places -- just because even knowing some family history but then looking at my DNA. It's useful to actually know that when you, especially for complex disease, to know this region of my chromosome six is, you know, southern European and this region is, you know, northern European in terms of

frequency of diseases and things like that. We might -- I think the more we learn, the more valuable that could be. But I also think that it's not the -- this notion that you have, you know, genes that are specific to Africa, or specific to, you know, France, specific to this part -- other part of Europe or this part of Asia, is really a misconception, and I think that it's a little -- that part is dangerous, not because of the social stuff but dangerous because scientifically it's not very solid.

Yeah, I still do a little research on a little bit of conversations with groups and behavior genetics --

[affirmative].

-- and, well, we'll keep that off camera --

Yeah, yeah, yeah, yeah.

So switching gears a little bit, I just want you to describe what were the most -- I mean, you're really involved, you know, ENCODE and [unintelligible] code, and what were the -- really the thorniest issues that were sort of overcome, or you had to deal with in constructing that program? Sort of the pilot phase and just give two or three I'm thinking pseudo genes or things like that.

Well, so the scientific issues were that, especially when we started, is that we didn't have great tools. And it was actually in the middle of ENCODE where sequencing based methods for measuring RNA and DNA and protein interactions and methylation and things were developed. So that made a huge difference in terms of getting a large amount. I think the most positive thing to come out of ENCODE is that it's a consortium that agreed, after lots of hand wringing, has generally agreed to a set a standard that, "This is the way we'll do it, and the data are freely available for everybody to use." So it is an encyclopedia in that sense.

I think that there has been both confusion and hand wringing and up and down about what role generating the encyclopedia, you know, almost keeps biology out of it. Certainly, ENCODE should not try and claim that they know all of biology. These are reference, you know, datasets from a few cell lines -- some primary cell lines -- and so I think all of that has been very valuable with some over, you know -- I think the biggest key for ENCODE is to stay away from press conferences because that's where the harm has come. The hardest scientific issues -- I mean, those evolve -- once sequencing-based methods got -- we could go, you know, much, much faster. The biggest issue right now for transcription factors because there are 1,800 of these or so, is getting good immune reagents. We're relying on antibodies to do those kinds of measurements, and we have hundreds and hundreds of these to go that -- that are just really, really hard to get good antibodies.

One of the other parts -- and this is not just ENCODE, this is -- I mean realize there are thousands of researchers outside of ENCODE who have been doing this for decades, okay, but on a different scale usually, and not to generate an encyclopedia. So one of the difficult things is integrating all of that -- having a set of standards for how things are done, at least from the encyclopedia, at least allows people to compare whatever they have against that standard, so that

helps a lot. The hardest issue that we have had other than antibodies, is which cells do you do this on? We, by necessity, can only do it with a modest number of cell types. Many of them are cancer lines. Some of them are primary cell lines. Some of them are tissues. So we're not doing entirely wrong -- we're not entirely limited there -- but, you know, in all research, clearly working with organs and tissues, you're dealing with multiple cell types and heterogeneous cell types whenever you do an experiment. And so trying to get this down to the single-call level is pretty critical because clearly you -- we do a lot of work on brain, on post-mortem brain, human and mouse brain, no matter how much -- careful you are about dissecting, you're going to have multiple cell types in a measurement that you make. And so that dilutes your signal out. If you have -- let's say in a disorder you have particular genes changing in particular types of neurons and you've got all these other cells that you're looking at. You have to see that signal over the mixture of all the other cells. So that's probably the biggest challenge. Not to ENCODE only, but to our whole field and that's why some of these methods are getting developed for looking at single cells.

Yeah, I am -- actually you're discussion right now is leading me into because I am also very interested in GTEX is that-- are the same issues percolating in GTEX as well or have you -- in terms of getting the right samples and this and that and the other thing [unintelligible]

So, yes, so I tell people now that our hardest problem is not DNA sequencing, it's not making libraries, it's not even analyzing the data, although that still is a hard one. It's getting the right samples. And sample doesn't mean just a DNA sample or cell sample. It means the clinical information. It means it came from a person; it came from physicians, physician scientists, you know, working in that -- working with that individual and their disease, let's say. It involves their family members, so these are complex things. And the other sort of rate-limiting difficult stuff is deciding what questions -- the question you want to ask -- to answer and then what do you need to get it? How many samples? It's not just power calculations and more complicated than that. And so that's I think any place where people are addressing those problems, the genomics can follow very, very readily if you do the genomics well. You got to do that well, but we know how to do that part well, we don't know how to do the upstream part as well. That's been rate limiting for everything.

So GTEX is one of those projects that is post-mortem, now larger numbers. A lot of hand wringing about getting that and doing it right, getting the informed consent properly done. One of the things that NHGRI is adamant about is that data get made as freely available as possible. You truly want, when you spend all this money and time and effort, doing -- you know, getting results out, you don't want it to be just behind a firewall where a thousand flowers cannot bloom. You really want everybody to be able to look at it. That is complicated because of the privacy issues, but I still think we can overcome that. We should overcome it because science advances so much better when it's not, you know, clubby and where only some people can get to it. So I think GTEX is -- ENCODE is now working with GTEX on doing some of the -- working on some of the samples that way. They are post-mortem. There, you know, there are all sorts of issues about the quality of tissues and things like that, but they, you know -- and how you dissect them and how you what tissues -- how heterogeneous they are when you're looking at them.

So I don't want to keep you too long over what I'm supposed to keep you, but you keep mentioning, you know, basically free access to data and I have to absolutely ask you about Bermuda, Fort Lauderdale --

Okay.

-- and things like GWAS and [unintelligible] --

Okay.

-- and so on and so forth.

So I had the good fortune to be part of the Human Genome Project from the beginning. And we had three Bermuda meetings. I attended two of them. I think I missed one because my daughter was being born. I can't remember why I missed the middle one. But they were remarkable meetings because it was basically -- we knew before we started, in 1990, that we were going to release the data for everybody to use. That was an absolute requirement -- for one human genome sequence. There were some countries that dug in their heels and didn't want to do this. And that's -- it was very interesting to see this happen at the Bermuda meetings where the decision was made, "Well, then you can't participate." And so they ended up coming through. And at least and then the data were released almost immediately. Then there were other versions. Fort Lauderdale and then the Toronto versions where there were some modifications to that, but the general feature, and I think the Genome Institute and the Department of Energy and others deserve a lot of credit for this, for -- and Sanger Center for that matter, in England, for saying this has to be for the world.

And truly to let everybody to look at it and use it, not for a touch-feely reason, but because the science advances so much faster. I mean so many things happened while we were generating the sequence and the clones and everything. And so I think that's a critical part. I hope we don't ever go back on that. Of course people who do the work need to get recognition for doing it. There are safe -- enough safeguards. Sometimes you do give up and people will compete, you know, sometimes unfairly, not even recognizing that you've, you know, participated, or that you've generated the data, but I think it's the way of science -- this kind of science really has to be. Too many resources going into this to let anybody own it.

Yeah, I mean that also brings me to the kind of factual question, I mean: How different would genomics research be if Bermuda hadn't been reached, and if everyone was sort of in a, you know, we're all -- I don't know, where you basically [unintelligible]

Every country for themselves, yeah, yeah--

[unintelligible] proprietary or, you know, the war of all against all, or something like --

I think it would have been a disaster. I think we'd be 10 years behind what we are now and, you know, maybe ultimately coming out. But the idea of -- I mean we knew from the beginning that this science was going to be different. It was partly to help people who were so concerned about

the Human Genome Project, you know, "You're going to take up all the resources." Well, first of all, it didn't. It's been a -- from an economic development point of view, the Genome Project has paid off 150 fold. I mean it's a -- literally, I mean literally has made that much difference. But by having all of the data go very, very rapidly out there we learn I mean -- so many people, companies as well as research laboratories and universities and institutes, you know, make big, big, big advances that just wouldn't happen if you can get at the data.

And you don't see that retreating at all?

No. I think there's a -- there's always been -- I think -- I was appalled when I first got into the human genetics that a lot of people in human genetics held their samples, held their information. I thought it was appalling actually. I fought against that in the early -- in the mid 1980s, just because it was part of the culture. I think that has changed a lot since then.

But you still have it. You still have people. Of course people are competitive and probably even more so. But I think even with competition on a race, let's say, for something people are looking for, once that gets out there, then you can get other minds and other, you know, graduate students and post-docs who are thinking about -- and faculty members who are thinking about this get them on it and there are many, many examples of where advances are made that were not expected because, you know, other people got to work on it. So, I mean, I think that's kind of the norm in genomics now.

All right. So I don't want to keep you over your time, but thank you very much.

You're welcome. Thank you. All right.

All right, so Albra [spelled phonetically] we're done. So let's update [unintelligible]. I'm actually working on pretty close [unintelligible].

I really enjoyed having that because I didn't have any -- I didn't have to produce anything for it, and so, and you know, that meant that I was -- let's see am I doing this --

Yeah. No, I think that's fine [laughs].

:
Here let me do it around here so it's not tight around your chair --

[end of transcript]