Hello and a welcome to this Science/AAAS audio webinar. My name is Sean Sanders and I’m the editor for custom publishing at Science.

We have a very exciting webinar for you today, in which we will be discussing the topic of human population genetics and novel locus discovery in non-European populations. Historically, genome-wide screening has looked at mostly populations of European descent, but more recently, researchers have begun to see the value in looking beyond European cohorts to study common and rare variants in populations around the world.

The two expert panelists I have with me today will be walking us through this topic explaining how the discovery of population-specific rare variants expands our understanding of complex diseases and how population genetics integrates with the genetics of complex disease to reveal novel disease genes.

So, it's my pleasure to introduce these speakers to you now. They are Dr. Charles Rotimi from National Institutes of Health in Bethesda, Maryland and Dr. Carlos Bustamante from Stanford University School of Medicine in Stanford, California. I'm very pleased that you could both be with us today.

Before we get started, I have some important information for our audience. Note that you can resize or hide any of the windows in your viewing console. The widgets at the bottom of the console control what you see. Click on these to see the speaker bios or additional information about technologies related to today's discussion or to download PDFs of the slides.

Each of our speakers will talk briefly about their work. After which we will have a Q&A session during which our guests will address the questions submitted by our live online viewers. So if you're joining us live, start thinking about some questions now and submit them at any time by typing them into the box on the bottom left of your
viewing console and clicking the submit button. If you can't see this box, just click the red Q&A icon at the bottom of the screen. Please remember to keep your questions short and concise, as this will give them the best chance of being put to our panel.

You can also log in to your Facebook, Twitter, or LinkedIn accounts during the webinar to post updates or send tweets about the event, just click the relevant widgets at the bottom of the screen. For tweets, you can add the hash tag, #sciencewebinar.

Finally, thank you to Affymetrix for their sponsorship of today's webinar.

**Slide 3**

Now, I'd like to introduce our first speaker Dr. Charles Rotimi. Dr. Rotimi is a genetic epidemiologist and a biochemist, and currently works as a senior investigator in the Inherited Disease Branch of the National Human Genome Research Institute intramural program at the NIH in Bethesda. He is also the director of the Center for Research on Genomics and Global Health. He did his undergraduate studies at the University of Benin in Nigeria and completed his Ph.D. at the University of Alabama in the US. Dr. Rotimi develops large-scale genetic epidemiology studies that explore the patterns and determinants of common complex diseases in human populations with particular emphasis on populations of the African diaspora. He is the founding and current president of the African Society of Human Genetics. Very warm welcome to you, Dr. Rotimi. Thanks for being with us.

**Slide 4**

Dr. Charles Rotimi: Thank you, Sean. Today, I'm going to talk to you about the excitement brought about by the sequencing of the human genome and also the subsequent spinoffs like the HapMap and the 1000 Genome, and what we are learning from the sequencing of individual genomes, and what kind of challenge that poses to us in terms of understanding human genetic variation in relation to human history and human health.

**Slide 5**

So understanding human genetic variation is the key to our ability to use genomic information across different human populations. So, I think *Science* was right in 2007 to name human genetic variation as the breakthrough of the year because that is really what understanding the sequencing of the genome is all about. It's our
ability to characterize variation at a genome level and genetic variation is indeed the hallmark of our genome. How do we do this across human populations, I think is really the key challenge that we face today.

Slide 6

So, it is of no surprise, I think, to most people on this webinar that human history started somewhere in Africa and that we indeed as humans share a considerable proportion of our genetic variation. Estimates range, you know, anywhere from 95% to 90% of variation that is shared across human populations and these variants are considered common. But there are also rare variants as a result of human beings going to different environments and incorporating different agricultural practices over the past 10,000 years. Our ability to adapt to different environments has indeed brought about some allele that we may consider still restricted to a certain geographical locations. So, those variants are indeed rare and they have not had time to spread across human populations. So again, trying to understand human genetic variation, we need to bring in to bear tools to understand both common and rare endogenous variants.

Slide 7

So one of the very important things that we've learned as a result of sequencing the genome and also projects like the HapMap and 1000 Genomes Project is that human genetic variation is nonrandom. There's pattern to it. There's a structuring. For example, we know that there is greater linkage disequilibrium in European population and Asian population than in African populations, and that human genetic variation is influenced by multiple characteristics. Some are genes-specific factors for example selection, rates of mutation, recombination and also genetic variation is influenced by demographic factors including population size, structure, founder effect, and admixture.

Slide 8

So, by studying human genetic variations, scientists are indeed in a position to shed light on ancient human population migrations, how different human groups are biologically related to one another, and how some disease-causing alleles occur at greater frequency in people from specific geographical regions. So I said, you know, here, borrowing from the work we did with Nature Genetics in 2004, that human genetic variation is our indeed our hallmark and our diversity is not an illusion, we need to study it. But the challenge that we face is we need to do this in such a way that we don't reaffirm old
prejudices. We just need to understand that because we've lived in different environments that certain things are different, certain things are emphasized, certain variations have been deemphasized. It doesn’t mean that we are indeed not members of the same race.

Slide 9

So the major force that has shaped our genome is indeed our environment. So for example when you are around the equator, you do indeed better have dark skin complexion to protect you against the sunburn and skin cancer and some of those biochemical parameters that you do need for survival. So in fact, the major characteristics that we see in our genome have indeed come as the result of our ability to adapt to different environmental stresses. That is why today when you look at the genome of Africans, they indeed have more of that interactions because they have lived the longest or humans have lived the longest on the continent of Africa.

Slide 10

So, some of these forces including climate, also diet has been a tremendous force that has indeed shaped our genome. The example that I have here for you is the relationship of a starch diet and the human amylase gene copy number variation. This was a fantastic study that was done recently that shows very clearly that even populations that are close by, but have different starch content in their diet, have very drastically different copy number variants for the amylase gene. Which shows very clearly that it is sometimes not just geography or proximity to each other, they actually even come down to what you eat and even sometimes how you cook what you eat. So, this is a very clear example of the relationship between the ability of diet in the shaping of our genome. Again, this shows that if we are indeed wanting to study the genetic basis of human amylase gene in relation to diet, we better be studying different human populations.

[0:10:03]

Slide 11

So, to sort of summarize that first session, you know, geographical structuring of genetic variation has implications for understanding human history and health including the design and implementation of genome-wide association.

So now, I want to talk to you a little more about the genome-wide association studies and why it has indeed been concentrated among European populations, and some of the things that we have missed
and we may continue to miss if we do not expand the application of genome-wide association and other genomic approaches to study global populations, not just European populations.

**Slide 12**

So genome-wide association is basically again the agnostic process of searching the genome with thousands of markers and basically asking the question where do I see a signal in relation to a particular trait. You know, for example, you could look in at type 2 diabetes and this pattern that you’re looking at on your screen is called the Manhattan plot, which is basically named after the sort of the skyline that you see in New York in Manhattan. Then we typically are excited when we see the blip, like the Empire State Building, which shows, you know, tremendous success in terms of significant association between the variants that we have on the genome-wide chip and our trait. So that's basically what the genome-wide association study is all about. It’s basically saying that we do not know where the signal is and we are just going to query the whole genome.

**Slide 13**

It is no surprise that we have indeed been very, very successful on using this approach to identify various parts of the genome that are related to various diseases and also non-disease traits. So this graph that you're looking at basically shows sort of the highlights of all of the areas that we have peak signals. Again, if you do it today, this picture you are looking at will be indeed more dense because we continue to find new parts of our genome in relation to the various human diseases and also human traits.

So again, I will say to a very large extent, genome-wise association study has indeed been very successful in improving our knowledge about biology and also showing us pathways that we did not know before were indeed important for certain diseases like Crohn's disease, hypertension, and some forms of cancer.

**Slide 14**

But the problem with the genome-wide association study is that it has not been applied systematically across human populations. As you will see from this paper from the Goldstein group, it basically summarizes, the data published, all of the genome-wide association that has been done so far. It is absolutely, you know, to me very surprising that over 90% of the work that has been done has been done in European populations. There are some in Asian populations and only one, you know, for the malaria project in an African
population. Although, the number in the Asian population is increasing, what has been done in African populations still lags tremendously behind in terms of genome-wide association study.

So one can easily say that there is inequity in the way that we are using genome-wide association study to answer questions across human populations. The consequence of that I think we'll disclose in subsequent slides and also, the reason why I think this has been occurring.

Slide 15

One of the important characteristics that was pretty evident from looking at data from the HapMap in relation to the genome-wide association technology, the chips that were used, are in terms of trying to show the relationship between various variants and traits. It's very, very clear that those chips are indeed less efficient in African populations.

[0:14:57]

For example, if you're looking at a minor allele frequency of 0.05 or greater and \( r^2 \) of about 0.8, you see that you need about a million SNP on the chip to reasonably cover the Yoruba sample from Africa compared to about 50% of that is about 500,000 European population. So clearly, you need more SNP on the chip and I think this was again, you know, as a business it may not be very cost effective to develop those chips for African populations. So there had been issues in terms of how appropriate is the current technology for African populations. Again, so those are some of the things that indeed are important in our ability to do work in different human populations. If the technology is not right then we are indeed sort of working with our hands tied right from the very beginning.

Slide 16

So, some of the consequences of the fact that we have not successfully applied genome-wide approaches to different human populations, an example is on this slide. Again, a genome-wide study was done in the Asian population that picked up this KCNQ1 gene in relation to type 2 diabetes. Despite the numerous GWA studies that was done in European populations, this variant was not picked up. It is no surprise because if you look at the allele frequency between the Eastern Asians and European, you will why this variant was indeed not picked up in the European population. It would have required a tremendous amount of sample size to be able to pick it up in the European population.
So for example, if you look at the power calculation that was done by Mark McCarthy in this slide, he said the power to detect the diabetes allele is over 80% in East Asians and less than 1% in Europeans just given the differences in minor allele frequency. So if we have not done this work in East Asians, we probably will never pick up this variant using the genome-wide approach. So that's again a direct consequence of the fact that if we don't go across human populations, we're maybe missing very important signals for various diseases.

Slide 17

Another example that I have here is the work from our own group looking at the FTO gene, one of the most replicated variants or genes for obesity. Again, the initial finding for FTO was in intron 1 in European populations, but in doing this work in African populations, we were able to use the weaker LD pattern across the African population to indeed better localize and also find other signals that were indeed not picked up in European population, for example in intron 8 in this study. So again, that's another example of why we indeed need to study different human populations.

Slide 18

This next slide here is perhaps what I'm particularly proud of in terms of the contribution from our group to understanding how human genetic variation relates to different diseases across human populations. One of the things that we noticed early in our own work, in our group at NIH is that our ability to replicate what was found in the European population in our African population was quite reduced. Even for something like height, we were only able to replicate about 8% of about 50 something variants if we directly query the same variant that was picked up in European population.

So again, the work from members of our group including Daniel Shriner and Ramos and other members of group, clearly show that perhaps the reason is again we are looking at different LD structure and there are local characteristics that we need to consider in our approaches. So the slide that you're looking at here shows again that if for example your tag SNP and your causal variant is on the same block, LD block, in your discovery and follow-up sample, that you may be successful to query the exact SNP in your follow-up sample. But if the pattern is different and it's similar to what you see on the right side of the slide, that the tag SNP and the causal variant are indeed on different LD blocks, therefore, you will not be successful if
you query or most likely not to be successful if you query the exact SNP.

So one of the strategies that we developed was really to take the reported SNP and to go to about 250 kb in both directions and using an $r^2$ of about 0.3 and doing appropriate multiple comparison controls, and basically query all the SNP in that region. What we find is that we indeed begin to achieve very high level of replication and therefore we are able to show that some of the variants that are picked up in European populations transfer to our African population. Again, this is taking advantage of understanding population genetics and differences in terms of LD pattern across human populations.

One of the interesting things that also came out as a result of this work is our ability to fine-map signals that you find in European populations using an African sample.

So for example, this slide that you have here, it is an example with LDL, a variant that was first reported in European populations. The region again was about 16 kb, but using our African sample in the study that has been submitted for review from our group, we were able to reduce this region using the weaker LD pattern in African population to less than 500 base pair. Again, that's tremendous advantage that you do gain for follow-up studies and we have indeed used this successfully for multiple traits including blood glucose, serum uric acid, and also bilirubin groups. So this is a strategy that I think is absolutely quite successful and I think it's something that needs to be considered to encourage collaboration between people who have African samples and people who are doing studies in European and Asian populations. There are other very good examples in terms of TCL7 work that we did with Decode Genetics.

So again, one of the advantages of making sure that we continue to use genomic approaches in different human population is that again we take advantage of some of the demographic history of human populations, for example admixture. Admixed populations present to us a tremendous opportunity to use the fact that the LD caused by admixture extend beyond what you normally would see. If you take good advantage of this, you can indeed reduce the number of
samples that you need and also markers that you need to find a signal in the genome and this approach has been used successfully.

Also recently, you know, in a work led by Daniel Shriner in our group has shown very clearly that admixture also present us with the opportunity to do both admixture mapping and association, and that if we do this very well that we indeed increase our opportunity to pick up signals in areas that may not be picked up if we just do straight association studies. So this is again very exciting work. This method is available publicly and if you're interested, you can indeed download it and use it if you have an admixed population. So that again is one thing that we do gain by extending genome-wide association approaches to different human populations.

**Slide 21**

The example that I want to illustrate to you here is the work that was done by NIH group and also Johns Hopkins, you know, trying to use admixture mapping to indeed shed light on the very high disparity that we see in terms of end-stage renal disease in African-Americans compared to their European counterparts.

**[0:25:11]**

So this again is a successful study. The hypothesis here is given the fact that we see this huge disparity and if you have the genetic basis then that genetic basis should indeed be seen in the African ancestry part of the admixture of African-American. A genome-wide admixture scan was indeed done in cases and controls by this group.

**Slide 22**

This again was very, very successful. A signal was picked up MYH9, one of the myosin genes, but a subsequent study using the data from the 1000 Genome showed that the actual variant may indeed be the gene that encodes the apolipoprotein L1. It turns out that this particular gene is in strong LD with the MYH9 gene.

What is absolutely fascinating about this finding is that this variant is indeed thought for all intents and purposes to be absent in European and Asian population. It is about 18% in African-American controls and about 40% in Yoruba sample. The question was why would such a deleterious variant rise to such a high frequency in a particular population and I think the answer came from the fact that this variant is protective against sleeping sickness, a devastating condition in the western and some eastern parts of Africa. So you can see how the evolutionary history, a demographic history of the population can indeed help us to pick up variants and again just
given the fact that this is not present in Europeans, you probably
would never find these things if we had not done this work in an
African ancestry population.

Slide 23

Again, the other point here is, you know, the work that was
published recently that shows again that the findings we are getting
from genome-wide association in European populations when we try
to compare them to non-European populations that what we do see
is very weak correlations when we do a pair-wise comparison. This
ranges from a correlation coefficient of about 0.2 to about 0.3 and as
you can see from this graph here, that the correlation is indeed quite
weak. This also shows again that we are indeed missing some
information by just restricting our studies to European populations.

Slide 24

The last point that these authots made is on this slide, which shows
again in a very powerful way that the point estimates when you look
at European and non-European populations that we are in opposite
directions or differ more than twofold, about 57% in European
versus Asians, 79% European versus African, and 89% Asians versus
African comparison. That basically is the conclusion that the modest
correlation, differing risk estimates, and considerable between-
association heterogeneity suggest that different ancestral effects can
be anticipated and genomic risk markers may need separate further
evaluation in different ancestral populations. So again, it is pretty
clear that we are indeed missing something by not studying different
human populations in a comprehensive manner.

Slide 25

This is the work led by Adeyemo in our group, which basically says if
we take all of the variants that we are seeing for various diseases
and we look at four HapMap populations of African ancestry, how do
these variants vary? As you can see if you look at type 2 diabetes and
so as of the publication date of this, we have about 38 loci and the
median minor allele frequency is about 26%, but it varies from about
3% to about over 64%, so tremendous variability. This is again only
African ancestry populations. So you can see that if you are making a
public health decision about the importance of these variants for
example, across these four African populations, you will indeed be
arriving at very different decisions and the importance of these
variants are going to vary tremendously. So, it is indeed important
for us to make sure that we study different human populations.
So, the last slide that I have here basically, this was something really hot off the press. I was reading this as I was traveling for this presentation actually. This is from Eric Lander's group in Boston, which basically shows that the mystery of missing heritability may indeed lie in an overestimation of the denominator of our calculation of heritability, which they called the phantom heritability. Basically, what this paper showed or what they claimed here, I need to actually study it more myself, is that a substantial portion of missing heritability could arise from overestimation of the denominator again creating what they call phantom heritability, but the question is, how is this possible.

They gave several, you know, important points. I think perhaps the critical one is the fact that current estimation of heritability completely ignores genetic interactions, or what we refer to epistasis, among loci, okay, and also the claim and I think it's correct that this assumption is indeed not correct because the current data or observable data is consistent with the interactions. Also, that because of this, that total heritability may indeed be much smaller, and thus the proportion of heritability explained much larger. So we may not be missing as much as we think we are, it's just that we have indeed overestimated our heritability by ignoring genetic interactions.

They gave an example using Crohn's disease, in which they showed that over 80% of the current missing heritability could be due to genetic interactions, if the disease involved interactions among three pathways. So again, this is quite a compelling case and I think it needs to be studied further and it may help also explain the fact that genome-wide association may indeed be more successful than we think it really is.

I think that is the end of my presentation and I thank you for listening.

Great. Thank you so much, Dr. Rotimi. We're going to move on quickly to our second speaker for today and that is Dr. Carlos Bustamante.
Dr. Bustamante received his B.A. and M.A. and Ph.D. from Harvard University and was a postdoctoral fellow at the University of Oxford and affiliated with Cornell University prior to his appointment as a professor of Genetics at the Stanford University School of Medicine. Dr. Bustamante is a population biologist who mines DNA sequence data for insights into the dynamics and migration of populations and the mechanisms of evolution and natural selection. In studies of humans, he analyzes single-nucleotide polymorphisms from many individuals to infer changes in human populations and their relationship to specific gene mutations. Together with research colleagues, he has used DNA markers to assess the impact of shared language and geographic obstacles on migration patterns and genetic composition of human subpopulations in Europe, Africa, and Latin America. Welcome and thanks for being with us, Dr. Bustamante.

Dr. Carlos Bustamante: Great. Thank you, Sean. It's a real pleasure to be here and to be on this webinar with my good friend and colleague Charles Rotimi who I have a tremendous amount of appreciation for and always learn so much from him.

Slide 30

So what I'd like to do is take off from what Charles talked about and dig a little bit into what we're seeing from the new wave of data coming from the 1000 Genomes Project and other efforts. So as Charles really elegantly demonstrated, we've now undertaken thousands of genome-wide association studies. These have been remarkably successful in identifying common genetic variants, those above say 5% frequency in the population that explains some modicum of disease risk largely in populations of Northern European descent.

[0:35:13]

What one fails to get from that first statement is that number one, we're not explaining most of the heritability. So for a trait like height where we have 180 bona fide hits, we're maybe only explaining about 10% and we can argue about why that maybe. One large possibility could be that we've only really queried common genetic variants and common genetic variants are only a small fraction of all genetic variants. So one of the things I want to talk about is how, now that we can sequence thousands of human genomes, we begin to see what that spectrum of rare genetic variation looks like, how many millions if not tens of hundreds of millions of genetic variants maybe there, and their population genetics. Really, the sort of the stark message of my talk is that most of that variation is very
population private. So if we hope to make some inroads into understanding the genetic basis of complex disease in different human populations then we really have no choice, but to increase representation because it is very likely that once we start delving into associations that go down into the 5% and below, most of that will be population private.

Slide 31

So, if we think about for example a very important study such as the Wellcome Trust Case Control Consortium study, this was the work of my former postdoctoral advisor Peter Donnelly, a real sort of watershed moment in complex trait genetics, here they looked at 3000 shared controls in 2000 different cases for -- I'm sorry, 2000 cases for 7 different diseases simultaneously. You can look at this Manhattan plot and really be kind of overwhelmed by the short of sheer success of the enterprise. We find very strong hits for type 1 diabetes and rheumatoid arthritis over HLA. There's the hit for coronary artery disease over 9p21 that's now been followed up. I mean just a really eye-opening and very powerful moment where we're beginning to see the genetic basis of these complex diseases.

Slide 32

As I like to say, this is great if you happen to be British and born between the months of March and April in 1958 then the Wellcome Trust has invested millions and millions of Pounds to figure out why you get a particular disease. But if you happen to be born in Caracas or Karachi or Cairo, how relevant are these variants to your particular disease risk and do you harbor other variants that aren't going to be catalogued by these kinds of studies simply because you're focusing only on a small subset of populations?

We can choose different statistics to represent this. One of the stark cases is perhaps the proportion of participants in these kinds of studies. So if you look, you know, at the proportion of people who've been genotyped in these large-scale studies, the number is like 95% are of European descent. So putting aside the number of studies, the actual just sort of large numbers that you need to get very strong significance, the disparity is even more pronounced.

Slide 33

So, here's for example a counter example of what you would find when you do look at different populations. This is the work of Victor Acuna and Samuel Canizales in Mexico and I have one postdoc Andres Moreno, involved in part of this work. Here, what you're
seeing is the frequency of a mutation in ABCA1 cholesterol transporter, which has a strong impact on your HDL cholesterol. So as many of you know, HDL is one of the types of cholesterol that you measure. You want to have a high HDL or your so-called good cholesterol. If you happen to have this amino acid change in ABCA1, your cholesterol is about half of what it should normally be. Very important mutation, very strong and pronounced and you found it by going in and looking at populations in the Americas and it's at reasonable frequencies in some populations. It reaches somewhere 5% to 25% on which population you look at.

The most important part of the study though is that it's completely absent from the rest of the world. So you could look at 10 million Europeans and you're never going to find this particular ABCA1 transporter mutation, but that mutation could be very relevant to millions of people if and you won't find it unless you really do the study in the right populations.

Slide 34

So, the other part of this is that as we're beginning to sequence more and more genomes, we could begin to get a sense of the distribution of functional genetic variation. Several years ago, we had looked at 39 exomes if you just sequenced the coding part to the genome and made predictions about how strong the selection need to be to maintain mutations at given frequencies. The upshot of this figure is if you look at new mutations that are entering the human genome based on these 39 exomes, you can estimate that about half of those mutations are quite deleterious. That is, they're going to impact fitness by more than about 1%, okay, or 0.1%.

[0:40:10]

That means that those are mutations that evolutionarily aren't going anywhere. They'll be a very strong impact on your evolutionary fitness and those are those in red and orange. But if you follow that out to say the 2% or 5% class, now you're talking about things that we've queried for association, well most of the polymorphism we've queried for association is basically bang on neutral. So if we believe there's any link between the functional impact of the mutation and the impact that that may have on fitness, then everything we've been querying has been neutral, and there may be a tremendous amount of rare functionally important genetic variation that is now going to be catalogued and we need to understand the population genetics of it.

Slide 35
So really, it comes down to understanding how often are allele shared, whether they are common or rare across different populations. We'd like to use this to predict the power and discovery rates for future experiments and we need to be able to understand and analyze admixed populations such as African-Americans and Hispanic-Latinos, two populations we devote a lot of our time working on. So, what I want to focus on really is just the first parts of the pilots from 1000 Genomes and tell you a little bit about what we've learned now that we've sequenced several hundred human genomes.

Slide 36

So this is part of the project that was published in November of last year. This is the 1000 Genomes pilot project data, which sequenced 180 individuals to 2 to 4x from three continental populations, CEU from Utah so it's representative of Northern European ancestry, the Yoruba from Ibadan, Nigeria, and a population sample from Chinese and Japanese. That sequencing alone found 20 million SNPs. So if we think about the kind of variants that were queried for associations in the studies that Charles talked about, I mean those were maybe half a million to a million SNPs. You're already talking here about 20 times more SNPs that we know now exist in the human genome.

Furthermore, there was a subset of the project that looked at exon sequencing data and those were a subset of genes sequenced in 700 individuals where we developed our techniques that allowed us to merge the high coverage data with the low coverage data. And also give you a little bit of insight into what the NHLBI Exome sequencing project has learned, that's part of a consortia that we're involved with have sequenced 7000 exomes including 2000 of African-Americans.

Slide 37

So one way of thinking about this problem is to calculate what's known as the join site frequency spectrum. So what we're seeing here is the frequency of variants going from 0% to 10% and 0% to 10% in the second population. So, if you had drawn two samples from the exact same population, you'd get the two-dimensional histogram that you have on the left. Basically, it looks like Pascal's triangle and you can think about it in the following way. If I were to flip a coin twice, right, a quarter of the time I'd get two heads, a quarter of the time I'd get two tails, and half the time, I'd get one head and one tail, okay. Now if on the other hand what you had were two populations that had split a long time ago, then every time
you'd sample, you'd really be only getting polymorphisms that were private to those populations and that's what you see on the right-hand side.

Slide 38

So, this is what theory predicts and this is what the data looks like from the actual sequencing. This is the work of Simon Gravel, a postdoc in my lab and we published this in PNAS last year. The amazing thing is that here is the frequency in the Yoruba, here is the frequency in the Chinese and Japanese. The vast majority of genetic variants that we find are number one private to populations and they are rare. So the modal class are singletons, mutations you only see once. You only see them in one population and not the other and there's a real possibility paucity of shared rare genetic variation. What does that mean? That means if I'm looking for an association of mutations that are at 10% frequency in one population, I'm very unlikely to see it in a different continental population.

Slide 39

We can now look at different pairs of populations. So if we now take this and put it on its side, here's comparing the CEU in the Chinese, here's that joint frequency distribution. In the middle here is what it would look like if what you had were just sampling, re-sampling. On the right-hand side are the residuals so if we subtract the middle column from the left most column. The scorching red is telling you what you're missing, which are basically shared rare variants. So you might say fine, Carlos, the Chinese and CEU those are quite different populations, now let's look at the Luhya and the Yoruba, these are two populations from Africa. They are about 0.5 FST, which is less than northern and southern Europe. Even Luhya versus Yoruba, you see a scorching red along the diagonal meaning that you're missing some degree of shared variants. Chinese and Japanese, the same so there are going to be associations that are private to the Chinese or the Japanese populations that you don't expect to replicate and that goes up as you look at rarer and rarer variants.

[0:45:31] 

Slide 40

There's a way to quantify this using a statistic that Simon developed and what we're showing here in this graph. So as a function of minor allele frequency, you might imagine comparing pairs of populations and asking how often do you share across a population versus within a population.
So if you're at 100% or 1 here, then the sample were to come from one single population. What you see here for a pair of European populations, the CEU and the Tuscans, that at 1% and 2%, you're seeing some population specificity, but about 3% or 4% you can't really distinguish those European populations, which is good. That means that when you're looking at those rare associations, you know, maybe at about 4% or 5% they'll replicate within Europe. The one percenters may in fact only replicate within one of the populations.

But now look down here when you're comparing an African population to either a Chinese or a European population. In that sub 10%, you're seeing extra sharing of 5, 10, maybe even 50-fold, right, which means that there are many, many, many of these associations in European populations that you're not going to expect to replicate in other populations and vice versa. So the only way to get these associations if we believe these rare alleles are important and again, rare alleles are the vast majority of genetic variants in the human genome, right? If we believe those are important then we really need to do the population studies at a very fine scale.

Slide 41

One thing we can do is take these kinds of data and build demographic models and I don't want the belabor the mathematics of it. We can basically take the multi-dimensional data and fit a model of human expansions and migrations. This is the work of again Simon Gravel and Ryan Gutenkunst, a former postdoc in my lab.

Slide 42

Here's sort of our best-fitting human demographic model, which begins with humans expansions in Africa about 150,000 years ago, the out of Africa bottleneck about 50,000 years ago, which is a very tight bottleneck, subsequent migrations among the continental groups. Then I'll draw your attention to this right-hand side, which is the dramatic growth that has happened in populations of Eurasian ancestry and particularly populations post agriculture, right. The Neolithic revolution, the fact that since 10,000 years ago, human populations have begun growing not only exponentially but super exponentially is going to have a tremendous impact on the distribution of rare genetic variants. Meaning we'll have even more rare genetic variants than we would predict and that much of that polymorphism is going to be population private.

Slide 43
So what we can now do is take these data and predict given that we've sequenced 100 genomes here, how many polymorphisms do we expect to impact amino acids, those so-called non-synonymous changes versus synonymous changes and draw out predictions for a sample size of 100, 1000, or 10,000 genomes.

One of the very neat results that we see here if you focus in on the right-hand side of this curve is that the top category now once we've sequenced 10,000 genomes are polymorphisms that impact amino acids in Chinese and Japanese populations. So if we were to sequence 10,000 Chinese and Japanese, 10,000 CEU, and 10,000 Yoruba, we would actually find more genetic variants in the Chinese genomes, right. You might say, well how does that make sense? Well it makes sense because we've now got that demographic impact, that huge exponential growth kicking in. So a country like China with a billion people is going to have more of that rare genetic variation than countries with a couple, tens of millions of people and that begins to now jive with one's intuitions. So there's a tremendous amount of rare genetic variation that we've yet to catalogue and understand, but we can now begin to model it and understand it and make predictions based on what we've sequenced.

**Slide 44**

We can now do the same thing with higher numbers. So if we look at the exome sequencing data, this is data produced by Debbie Nickerson and Stacey Gabriel. I'm showing you now the joint allele frequency spectrum for Europeans and African-Americans.

**Slide 45**

What you see here is here is the joint allele frequency spectrum for the CEU and YRI so these are the 1000 Genomes populations. Here are now European-Americans and African-Americans and you see very clearly that the African-American frequency spectrum looks very different from the Yoruba side frequency spectrum because African-Americans aren't just a transplanted African population, rather they have their own demographic history, they have an influx in ancestry from Europeans, about 20%. So in fact, there are some variants in African-American populations, as Charles talked about, that come in from that European ancestry and in fact, you're getting some of those rare alleles contributing now in African-American populations and you begin to see that playing out.

18
Okay. I'm getting some feedback there. So the last bit here is that we can now take these results and also power if we've now gone beyond 1000 sequenced exomes to 10,000 or 100,000 exomes, how many variants we're going to find. We see that there are likely to be millions of coding polymorphisms in the human genome, millions in just the gene coding regions that are likely to be impacting some function and maybe important.

Slide 47

So the sort of the conclusions here is that, number one, population genetic theory coupled with data on exome sequencing is really beginning to support this many deleterious alleles model of disease. There's a plethora of genetic variation out there. Correlating rare genetic variants with disease is going to be much harder than common genetic variants so we need new statistical method, but there's plenty of variation out here. We're now finding the numbers maybe 50 million, 100 million polymorphisms meaning that these are things that you're seeing in tens of thousands of people.

Slide 48

It implies a switch point in efficiency between cataloguing variants and doing tests for association and so we believe the important thing is to try to focus and figure out where might you have some of these globally rare but locally common polymorphisms. So to give you a flavor of what we're doing, we're now working to find a collaborator throughout the world and we've love to hear from other people on the webinar who are interested in collaborating and putting together a really good set of samples from throughout the world with the proper consent to do this kind of sequencing and you get a sense here. In yellow is what my former colleagues at Stanford had done, the Human Genome Diversity Panel. In red are some of the samples that we and our collaborators have put together including samples from West Africa, the Americas, North Africa. In green is what the current best efforts are, which are the 1000 Genomes Project and we're very proud to be a part of 1000 Genomes, but we think there is an opportunity to move and increase representation as well.

Slide 49

I just want to end with this slide. This is now some sequencing we've done in the lab, 53 human genomes from different human populations. The Yakut are a trans-Siberian population, Maya, Cambodia, Mozabite, Mozabite are of North African population, Pathan are population from Pakistan, and the Mbut Pygmy and San. Here, you see the percent of novel variants that we find once we
sequence these populations relative to 1000 Genomes. So some of these African populations 20% of the SNPs you've seen you'd never seen before even though you'd already sequenced 1000 genome. So there's just a tremendous amount of human genetic variation out there left to be catalogued and correlated.

**Slide 50**

With that, I'd like to acknowledge our partners in crime, folks both involved with 1000 Genomes project in particular and with the Exome pilot part of the project, and happy to take any questions folks might have.

**Slide 51**

Sean Sanders: Fantastic. Thank you so much, Dr. Bustamante, and many thanks to both of our speakers for the very enlightening and fascinating talks. We covered a lot of ground there and we're a little bit behind time, but we are going to go over by a few minutes so we can get some questions in.

Just a quick reminder to those watching us live, you can still submit your questions, just type them in the text box and click submit.

The first question I'm going to put out to you both and maybe I'll start with Dr. Rotimi is why do you think GWA studies have currently almost excluded non-European populations the past few years that these studies have been done? Dr. Rotimi?

Dr. Charles Rotimi: Yes. I think perhaps the most important reason is the availability of large cohorts in European population. I think this has made it very attractive for scientists in that part of the world to come together to do genome-wide association studies. So that's the number one factor again the availability of well-characterized, large, human genetic epidemiologic cohorts.

[0:55:05]

I think number two is also that there very good and well-trained scientists who are Europeans who are leading this effort, which is not the case when you look at other ethnic populations with the exception of maybe Asia, which you can begin to see that Asia is also coming up now in terms of genome-wide association study.

The third factor I think is funding. I think where the funding has gone, it has gone towards people who have these very large cohorts and they're the scientists who also can write these very excellent grants, you know. So funding has in a sense been directed towards --
it has been following the people with the large cohorts and people who have been in a sense successful in the past. So, the money keeps going to those who already have in a sense.

The fourth reason I think is also technology. As you know, Carlos and myself showed in the presentation, that the old technology or the past technology for genome-wide association hasn’t been very efficient for African and the other ethnic minority populations again because you require more SNP on this chip to tag these causal variants in an African population specifically.

So, those are four factors that I think have indeed driven, you know, why genome-wide association has indeed been predominantly in European populations who you have large cohorts, scientists, well-trained scientists, and funding, and technology.

Sean Sanders: Excellent. Dr. Bustamante?

Dr. Carlos Bustamante: I can't give a better -- I think those are spot on. I would say it's an issue of funding priorities and I think in study sections when you go in with a small cohort particularly from an isolated population, you get dinged. People feel you need thousands and thousands of individuals to do these studies, which I don’t think is necessarily true. I think if you have the right set of isolated populations, you could certainly make inroads.

I think people don’t necessarily prioritize it because they don’t necessarily view it as exciting. They feel, oh, GWAS, you know, we've done a lot of GWAS, I'm tired of hearing about GWAS, let's move on to the next step, which are rare variants. But of course you can't do the rare variants if you haven’t done the common variants and no one is going to give you money to look at ten million people if you haven’t already looked at thousand or something like that. So I think it's a very big issue and I think the cohort issue is very important and getting proper partnerships in place between investigators in developing countries and developed countries is a huge issue.

There's a tremendous amount of mistrust. People have samples, they don’t want to give up their samples rightfully and so they feel look, you know, if you had the resources in country, I could a tremendously good GWAS so why don’t you just give me the money to do it because my country can’t give me the money. You know, it's a very tough issue. But the end result is that the countries that are already well organized are the ones that are going to continue to
benefit and unless this is viewed as a top priority, you know, the genome sequencing is just going to get worse, right. If you talk about the number of sequenced human genomes, the disparity is even bigger than the GWAS.

Sean Sanders: So as far as the heritability of complex diseases, do you see that these differ in non-Caucasian samples? Dr. Bustamante?

Dr. Carlos Bustamante: In terms of the heritability, you mean --

Sean Sanders: Heritability of complex disease, yeah.

Dr. Carlos Bustamante: You know, I think it's -- what I would put forth is that for any phenotype, human phenotype you want to study, there are going to be environmental and genetic factors and those genetic factors may be different in different populations because they maybe of different frequencies.

Sean Sanders: Right. A question for you, Dr. Rotimi, which diseases do you think are going to benefit most from studying more diverse ancestries?

Dr. Charles Rotimi: You know, in my opinion, all human diseases will benefit from that experience. As, you know, Carlos showed very clearly, if indeed some of these variants that we are picking up are going to be population specific to some extent or have some geographical restrictions, then we need to study them in these various diseases. So we may have something called type 2 diabetes that at the clinical level looks like the same across all human populations, but the underlining genetic associability may vary ever so slightly. If we don’t study different human populations, we may indeed never know others. So right now, I think we need to indeed study all of these conditions in different human populations.

[1:00:13]

But specifically though, there are some diseases that are indeed local that you will never gain insight from them by looking at European populations. For example, we have a study that we are working on now called podoconiosis, which is again a condition that looks like elephantiasis, but it is not. It is actually result of exposure to fine grain sand in the volcanic area in places like Ethiopia and Cameroon. Those conditions don’t exist in Europe anymore so you will never be able to study those conditions if you do not fund an African population or an African scientist or their collaborators to look at those kind of conditions.
To just highlight a point that Carlos made, for that particular study we were quite successful by looking at just 200 cases and 200 controls. So you don't need 20,000 or 100,000, it just depends on the architecture of the disease.

Sean Sanders: So a number of variants have now been identified. How do you address the challenge of validating whether these are actually causal mutations or not? Dr. Bustamante, maybe you can start us off.

Dr. Carlos Bustamante: Well, I mean, you know, the question of causality is a philosophical one, it depends on what burden of proof you require. I think of the GWAS results that have first come out, I would say that there have been some where you've been able to link it to a coding polymorphism, right. I mean the classic example of complement factor H that was sort of a beautiful causal link. There are some where in fact it we'd had to do more work, but it looks like the 9p21 association at least with cardiovascular disease does seem to be a non-coding polymorphism where you could demonstrate that the factors are interacting, you know, chromatin accessibility. So it's a question of digging in deeper and doing the right set of experiments that might include a mouse model, it might include doing more functional genomics work, right, the end code consortium sort of data are going to bear in on that.

I would say from an epidemiological point of view, you know, we're okay with correlation. I think that the first step is trying to really establish the set of variants that are at least correlated with disease outcome and then you can begin to tease apart whether that's due to a compendium of variants that are all in linkage disequilibrium and the functional mechanism. Of course, even though you establish the functional mechanism, you don't necessarily know how to go from there to alleviating or a therapeutic so it's a long road.

Sean Sanders: Dr. Rotimi?

Dr. Charles Rotimi: Yes. I think that if you look at the publications of genome-wide association studies so far, I think for the most part it's been dominated by statisticians, geneticists, and population geneticists. So again, that has led us to a point where we've been able to pick up these variants. To answer your question in terms of how do we know the function, I think we now need to make sure that we fully engaged the bench scientist, the biochemist, the physiologist, the molecular biologist so that they can begin to shed light on, provide insight into what is the variants that we are picking up, what they
actually do. So I think we need to now broaden and that’s a point that I actually made in my presentation. I'm in keystone meeting here in Sta. Fe and the other point that we made yesterday that we need to engage the bench scientists to come, those who know how to do the functional assays to now take these variants and really show to us how they influence human health.

Sean Sanders: Great. So I'm going to try to sneak in a couple more questions. I hope our audience will stay with this for those last two, just a couple more minutes. There's been a number of questions about studying different populations using GWAS including in South America, which I know Dr. Bustamante showed some data on and also in India. But maybe I can broaden the question and ask, which populations do you see are going to be the ones that are going to be studied in the near future and that will be the most useful for this type of work? Dr. Rotimi?

[1:05:03]
Dr. Charles Rotimi: I will say the African population because the African population has these unique characteristics giving again the evolutionary history of humans that whatever we’ll learn from the African population to a very large extent is going to inform other human populations. So by investing in doing more of these genomic studies in Africa, I think we are not only trying to bridge the inequity that exists right now, but it's also very sound scientific thinking.

Sean Sanders: Right. Dr. Bustamante?

Dr. Carlos Bustamante: So I would place my chips on isolated populations that have here to for not been that well studied and that includes many populations in Latin America. So even admixed populations in Latin America tend to show very fine scale substructure and I believe because the Americas were the last continent to be colonized and that many of the indigenous populations descend from small founding populations that it's a very unique opportunity to study. But furthermore, because of the last 500 years of admixture, you have these populations that have sort of taken off from different starting points and they're sort of very natural laboratories for comparing the genetic basis of complex disease and many on island populations as well. So we're putting a lot of emphasis in the Americas and trying to work on very well-defined populations.

I also believe the Pacific Islands are just an amazing opportunity with what we've been learning from ancient DNA and the great
complexities that occurred there during the last half a million years of human evolution. I think it'd be very interesting to see how these different populations may carry variants that are going to predispose the disease. So, you know, obviously Africa is awesome and I can't emphasize that we study that, but I think Melanesia and the Americas are going to also yield really cool insights.

Sean Sanders: Great. So I'm going to put out a final question for you, which is kind of a two-part question. Do current GWAS platforms cover genetic variation in non-CEU populations and in admixed populations as well? In addition to that, where do you see the GWAS studies going in the next ten years or what would you like to see? Some maybe I'll start with you, Dr. Bustamante.

Dr. Carlos Bustamante: I think it depends on what population you're talking about. We've certainly found that there are private variants that aren't well tagged and that's why I think it is really important that we still have efforts like the 1000 Genomes Project that try to enroll people from throughout the world to build better catalogues. I'd love to see it do a 10,000 or 100,000 genome project where we now broaden representation particularly in the Melanesia, which is now not represented. Africa, we only have a couple of pockets of diversity in Africa. We need many better, many more samples from Africa and the Americas and say 10,000 to 100,000 genome project to really capture variation.

I think the cost of doing this kind of work now is largely phenotype driven. So I think we should really be looking towards building large systematic worldwide phenotyped cohorts that can be used by scientists throughout the world to really tackle these problems and move beyond the, you know, here are my little set of samples that I have, here are your set of samples, but rather view it as an opportunity to build a resource that will provide a genomic medicine for the world.

Sean Sanders: Dr. Rotimi, you get the last word.

Dr. Charles Rotimi: Yes. Again, I think there needs to be a very conscious redirecting of funding to make sure that more global populations are given the opportunity to be studied using genome-wide approaches. I think the reviewers have a role to play here, the funding agencies have a role to play not just to discourage individuals who are starting to collect large cohorts but are not quite there yet but continue to give money to those who have 100,000 or 200,000. So there have to be
rethinking of the funding and how we encourage young and not so young scientists to develop these large cohorts in different populations.

For example, we recently funded, you know, a joint project between the NIH and the Wellcome Trust Call at H3 Africa. It's a good example of that where, you know, there's now an opportunity for African scientists to come together to develop large cohorts that will in the future enable large genomic studies to be done in that part of the world.

[1:10:22]

Dr. Carlos Bustamante: If I can just add one point, which is that, you know, when NIH grants get scored, one of the big categories is innovativeness, is this something innovative. When they look at GWAS, they immediately say, oh, there are already 2000 GWA studies how can this be innovative and they don't realize, they don't give value to the fact that you're looking at a very diverse population. That has to be taken into account and given some large amount of bonus points in terms of going out and collecting these populations and making these links with developing countries, right?

It's amazing to me that an experiment a year ago was highly innovative and super interesting and now you try to do it in a different population that takes a tremendous amount of effort to collect the samples, oh, that's not so innovative because somebody else has done it. But you have no idea whether the results you're going to find are the same, right? But that kind of myopia that NIH reviewers have is really something that I think from the review committee really needs to come down and say look, we really value the increase in diversity that these kinds of studies are going to provide and that has to be given priority.

Sean Sanders: Uh-hum. Great. Well, I think that's a great place to end and many thanks to our speakers for providing us with such interesting talks and great discussion, Dr. Charles Rotimi from NIH and Dr. Carlos Bustamante from Stanford University School of Medicine.

Many thanks to our online audience for the questions you submitted to the panel. Sorry, we didn't manage to get to all of them.

Slide 52

Please go to the URL now at the bottom of your slide viewer to learn more about resources related to today's discussion, and look out for
more webinars available at www.sciencemag.org/webinar. This particular webinar will be made available as an on-demand presentation within approximately 48 hours from now.

We’d love to hear what you thought of the webinar, just send us an email at the address now up in your slide viewer, webinar@aaas.org.

Again, thank you to our wonderful panel and to Affymetrix for their kind sponsorship of today’s educational seminar. Goodbye.