JOEL HIRSCHHORN: So the topic is basically comple-- the title is Complex Traits, What to Believe, which I think tries to get the gist of one of the main messages here and fit into the number of characters allotted for a title of a lecture. And it basically has to do with current efforts in human genetics to understand the role that inherited DNA sequence variation plays in regulating what are called complex traits.

And let me just go through all that. So this is basically the plan for the talk, at least as it stands now, and just give you a little bit of introduction into complex traits and common disease and how they relate to each other. So here's a list of-- a somewhat random list of diseases that you might see in a hospital or in a doctor's office and account for a lot of morbidity and mortality.

And it turns out that most of these diseases cluster in families in ways that tell us that genetics plays a role in regulating who will and will not come down with these diseases. Same thing is true for things that aren't diseases but are quantitative traits that you can measure. Some of these, however, are very significant risk factors. I list here just things that are relevant to endocrinology, which is one of my interests. But things like blood pressure, obviously, if you have high blood pressure, that's a huge risk factor for heart disease, stroke, et cetera et cetera.

Increased obesity is a huge risk factor for death and diabetes and heart disease, and many of these other things also have to do with future diseases. And these also cluster in families in ways that tell us that genetics plays a role. But for all of these, whether it's disease susceptibility or the quantitative trait illustrated here, adult height, we know also from the way that they cluster in families that multiple genes are involved and furthermore, that environmental and other non-genetic influences must play a role as well.

So this is not quite so simple a picture as, for example, trying to tract down the gene that's responsible for cystic fibrosis or for Tay-Sachs disease. And so the goal that a lot of people in genetics have is to try to make this connection. There's variation in the human genome, which I'll talk quite extensively about. And there's variation that we observe out in the world. And we know there's a connection here, and it's a question of making that connection.

And if you compare genomes as, again, I'll get to in a little bit, find that people are mostly the same. But there are differences between every person. And again, most of these differences are random spelling differences in the genetic code that have no consequence. But some of them do have biological consequences. Again, the challenge is to figure out which one of these differences matter.

So I said that we know that genetics plays a role, and I just want to briefly get into, how do we know that, actually? So one way is from something called twin studies. I forgot the little-- this is from some country music group that I have no affiliation with and have never heard called the Taylor Brothers, but notable for part of their promotion is that they're identical twins. And you can see things like they look the same, they're about the same height.

I'm not sure whether you can tell whether they have the same preference in clothing or not from this. But they may have some of the same aptitude. So and what you can do is can do twin studies where you compare identical twins to twins who are what are called dizygotic twins, basically where also commonly called fraternal twins. If you compare same sex fraternal twins and identical twins, they're, in theory, raised in about the same amount of shared environment. So they're both raised in the same household, and that's the thing.

But identical twins, or monozygotic twins, share 100% of their DNA, whereas dizygotic twins share only 50% of their DNA. They share the same as siblings. And so if there are things that are more concordant or more similar in monozygotic twins than in dizygotic twins, the usual conclusion is that genetics plays a role in that similarity. So for example, if you look at type 2 diabetes, if you look at monozygotic twins and follow them over their lifetime, there's at least an 80%-- and some people thinks it's actually close to 100%-- concordance rate.

What that means is actually sort of striking, which means that there is a particular set of genotypes, which is incredibly predictive of getting diabetes. Yes, thank you. Sorry. I found a-- I found a-- I found a little cable. So one right one and one wrong one. So I'm all right.

Whereas if you look at dizygotic twins, there's a lot of concordance. And part of that is because they share 50% of their genetic material. And part of it's probably because they share the environment that they shared. But the concordance rate is much, much higher for monozygotic twins. So that says that there's a strong genetic influence.

And you can do other studies that don't depend on twin studies. And you can also do twin studies where there are monozygotic twins reared apart, but those are much less common. But basically family studies where if somebody has diabetes, what's the risk to their sibling for getting diabetes as opposed to the risk of somebody next door, for example? And again, it's much greater for siblings than it is for the general population. That ratio is something called lambda S.

And you can estimate also something called heritability, which purports to be the percentage of variation that's attributable to genetic factors and actually certain types of genetic factors. So that's basically an introduction. So most common diseases are these complex genetic traits where we know genes play a role. And we also know, from those family studies, they don't just segregate in a simple pattern where if the father-- if the father has it, half of the children will have it, and then half of their children will have it, et cetera, and/or in a simple recessive pattern. But it's clearly multifactorial.

So the next section is basically why we think that there might be a connection between common genetic variation and these common diseases. So this is a stretch of a genome. I think I borrowed this from David Altshuler or Eric Lander. I'm not sure. And this is what genetic variation looks like.

So you have a long stretch here. And if you compare two stretches taken from different people, you'll find that there's one place here and one place here where maybe they differ. And most of those variants change a single nucleotide or single DNA letter, and it's called, therefore, Single Nucleotide Polymorphism. And means many forms, or in this case, would be just two forms, or a SNP for short.

And what you might find is there's this stretch of DNA where everybody is the same. And then at this nucleotide, some people have a C and some people have a T. And what we want to know is, of all these different genetic variants, which ones are going to be the ones that are going to affect things that we care about? So here's something that we should all be passionate about in the next few months, so whether you're happy or sad that A-Rod went to the Yankees or something like that.

But no, in seriousness, if this were healthy and diseased, not to say which is which here, then we would be interested to see which variants track along with disease. And that's going to be the general gist of what people are trying to do. So I said that there's a connection between common variants and disease. And so one question is, why might we be interested in common variants?

And if you think about it, before I begin, what's known so far about genetic disease? You might say, well, that's the wrong place to look because if you look at things like Tay-Sachs or cystic fibrosis, those are mostly caused by genetic variants that are quite rare in the population. They're severe deleterious mutations that cause these severe diseases. So if those genetic diseases are caused by rare variants, why wouldn't things like diabetes be caused by rare variants?

And it's an open question, and I'll get into that. But let me just take you through at least some of the background. So I showed you earlier that if you sequence a stretch of genome and imagine this is about 1,000 bases from two different chromosomes-- it could be two chromosomes from the same person, or it could be one chromosome from somebody here and one chromosome from somebody around the world-- you'd get the same answer more or less, which is that, on average, you'd see about 1 every 1,200 bases would be different.

You can then ask the question, well, having identified these two alleles, the C allele and the A allele, if I then look in the same 1,000 bases of lots of other chromosomes, is it that this A allele or maybe this C allele is just a really unique event that's private to one of these chromosomes? Or do I see both alleles around the whole world? And the answer, most of the time, about 90% of the time, is that you see both alleles around the world.

And so what that means is that most of the differences that you can find between two chromosomes are explainable or attributable to the variation, those variants that are common, that are shared throughout the world that are not like the Tay-Sachs mutations. But they're actually old variants that happened a long time ago and are present in multiple populations around the world. So right. And that's basically what that says.

And so this has led to a hypothesis. So the assumptions are as follows-- first of all, that most variation in the genome is evolutionarily neutral. That actually may or may not be true, but actually, whether that-- as long as it's not, most variation is evolutionarily beneficial. We're OK. But let's assume there's some background for the genome.

So then you have a snip that's 22 kb away from the nearest gene and probably doesn't affect the function of that gene at all. So assume that evolution has not particularly cared about that base change. We know that most of this variation, most of the background variation is probably what you would call this, rather than neutral, is due to these common variants. That's what I just showed you, that 90% of that variation is due to common variants.

We know also, what I've been telling you with the Tay-Sachs disease, that traits under negative selection will, in fact, be largely due to rare variants. And there's a nice description of why this should be true in an article by Jonathan Pritchard. But there's a hypothesis that traits that are not under negative selection or at least not under severe negative selection will actually be due to variants that look more like the background, so that are common variants.

And so this is what's called the common disease-common variant hypothesis. And one of the more forceful, I would say, descriptions of it is a paper by Reich and Lander in *Trends in Genetics,* 2002. And the idea for why common diseases might be like-- might be due to alleles that are not under severe negative selection goes like this. First of all, they're actually common. So in some sense, it sounds tautological, but it's actually not.

They can't be that evolutionarily deleterious if they're extremely-- if they're extremely common. So that's one argument. Two is that they're multifactorial. So each gene that contributes to these common diseases actually only contributes a little bit to susceptibility. And that's going to be a theme that'll run through this whole talk. So even if it were a-- even if they were somewhat evolutionarily deleterious, each individual allele would only be contributing a small fraction of that.

Third is they often tend to be late onset or after reproductive age. So again, evolution cares less about stuff like that. Although, we don't know. May be good to have had your grandfather around in Africa 50,000 years ago. So it could be the late onset diseases were bad.

Another thing is that it's been hypothesized that some of these diseases like hypertension or diabetes actually have what's called balancing selection, which is that yeah, they give you diabetes late in life, but the reason that they give you diabetes is because they help make you hold on to calories better. So that gives you obesity now when we have McDonald's around. But 100,000 years ago when food was scarce, maybe that was really good.

There's something called the thrifty gene hypothesis. So for hypertension, maybe it was good to hold onto salt if you lived in a desert, that sort of thing. So if we're looking for a common variation or thinking about that this may be a good place to look for the genetic variants that underlie common disease-- and again, this hypothesis does not say that rare variation won't also contribute to common disease. Just says a common variation will be part of the contributors.

So if we're looking for it, it's been estimated that there are about 10 million common SNPs where both alleles are found at more than 1% in the general population. About 6 million of those are actually already in databases. So it's actually most-- a good fraction that have been found. And as you'll see, there's a lot of redundancy between SNPs. So probably a very high fraction of the variation, common variation in the genome, actually has been cataloged, which is a remarkable thing.

When I started in this field not very long ago, I think six years ago, there was a paper that came out of the Lander lab, which is a massive tour de force, cataloging 2,000 SNPs across the genome. And this quintupled the number of SNPs that were known. And so this is now-- this 3,000 times as many now, just in the last six years. And you can find all of these in websites like dbSNP or the UCSC Genome Browser, and CBI has a genome browser.

And they're all right there. So this is a screenshot from the UCSC Genome Browser. Each one of these little lines is a SNP. And this is probably already out of date, even though I took it not too long ago from there. And this is just in one gene, each of these are-- on each of these two columns is a SNP.

So if we want to use this catalog of common variation to try to find the alleles that are responsible for disease, how can we do that? Well, one way would actually be to try to study every variant or at least every common variant. This would be great. It would be an unbiased search of the entire genome, basically the entire universe of common variation. And this would be great if we could do it, but it's not currently practical.

10 million people-- sorry, 10 million variants. And as I'll show you, you need many thousands of people to do this. So it actually works out to hundreds of billions of genotypes. So that's something like the Pentagon budget or something for a year to do one study. So NIH doesn't like grants like that.

So basically, until this gets much cheaper and/or we apply some of the shortcuts that I'll tell you about, you need to pick which genes and variants we're going to study. So this brings into methods for finding traits. Any questions so far? So we're on the track of trying to find which of these common variants might be contributing to common diseases.

So really, three approaches, two of which have to do with common variants, and one actually which explicitly has to do with rare variants. And I'll get to that at the end. So the traditional approach to finding genes for disease has actually been something called linkage. I don't know. Have guys done linkage?

Is that something that's been done here? Yes? No? So I'll briefly go through linkage. So linkage requires families where you have, in the case of disease, multiple affected relatives in the same family. So those could be brother/sister, or it could be cousins or an uncle and a niece or something like that. And the idea is that there is a single disease-causing allele that's segregating that family.

So it's being passed on to some of the people in the family and not others. And the other assumption is that the disease will more or less-- doesn't have to be perfect, but has to be pretty good-- will more or less track with that allele. Now, for linkage, you're not actually typing the allele itself. What you're doing, you're typing markers that surround the alleles. So they may be a couple million bases away, actually.

And that's a typical thing is about markers spaced every 10 million-- every few to 10 million bases. And the idea is that since you're just looking in a family, it turns out that a million bases is what's called one centimorgan, which means that recombination events happens once every hundred meiosis. so within a few centimorgans, there's generally not going to be any recombination events within a family. So the marker that's a few million bases away will travel with the disease allele in the family.

So you use that marker as a surrogate for the disease allele, and you follow the marker through the family. And you say which of these markers. So if this marker is traveling with disease, you say maybe that's pretty close to the disease allele. But a marker on some other chromosome won't be segregating with the disease. So you'll say that region isn't linked. So that's linkage, basically.

AUDIENCE: [INAUDIBLE]

JOEL HIRSCHHORN: So traditionally, it's been microsatellites, which are markers that are things like repeats of the sequence TATATATATA, and some people have-- DNA polymerase doesn't like to copy those very well. So some people have 50 copies, some people have 44, some people have 53, that sort of thing. And you can type those. Those are a pain in the ass to work with, actually. And so people are shifting over, starting to, to use SNP-based linkage analysis.

Because all you need is basically a marker that lets you distinguish each of the chromosomes that are segregating in the family. So probably, microsatellites are-- it's going to-- there's going to be a shift. But it's been wildly successful for single gene disorders. And the reason it's been successful is because there's always just a single copy in a family of the disease allele, and the segregation with disease and that marker is really very good because if you have the disease-causing allele, you get disease, and if you don't, you don't.

So for common diseases, none of those things are true. First of all, because multiple genes are involved, you might not have the disease-causing allele but get the disease anyway and vice versa. You might have the disease susceptibility allele, really it should be called, and not get the disease. And furthermore, there might be multiple-- because it might be a common variant, there actually might be multiple copies of the susceptibility allele traveling in the same family.

So the markers that are a couple million bases away now no longer perfectly mark the disease susceptibility allele because maybe on this side of the family, it's marked by-- it travels with this marker, this allele. But then on this side of the family, where it's segregating, it travels with a different marker. So all of these things conspire against you for common diseases. Nonetheless, people have found a couple of things through linkage of common diseases and tracked it down. DeCODE genetics in Iceland is probably the most prolific example of that.

So if linkage isn't working, it brings us to the other main approach that's used, which I'll spend a lot of the talk talking about because it is the main method. It's called association studies. And basically, what this means is that rather than searching through the whole genome, you pick candidate genes. And Rich and [INAUDIBLE] have shown that this is much better suited than linkage to this scenario where you have common alleles that have only modest effects on disease, what's called modest penetrance.

And then there's another approach that people are taking, which is they say, OK, I understand linkage might not be the best way to go. But I don't believe this common variants stuff. I want to look for rare variants that contribute to disease. And so there, what you are doing is, again, you're picking the genes to look at. But instead of just taking SNPs off the public map and looking for common variation, they're actually resequencing to try to find those rare variants that are present only in disease.

And the best examples of this are BRCA1 for breast cancer, which actually originally identified by linkage, but that's a rare disease, and also MC4R mutations. That's a melanocortin-4-receptor gene, where there's a lot of rare variants that contribute to early onset, particularly, severe obesity. The difficulty with this approach right now is it's very expensive. It's much more expensive to resequence than it is to do genotyping of polymorphisms.

And if the expense were flipped, people would probably do this resequencing because this gets you the common variants as well. And it's really just-- it's largely an expense issue that we can't do this. So I talked about the study that would take just the Pentagon budget. If you were going to do this, this would take probably the GDP for the history of the United States or something like that. So we're still a ways away from this.

AUDIENCE: That flies in the face of the whole hypothesis, though, [INAUDIBLE]

JOEL HIRSCHHORN: Yes. So I'm just saying that if you choose to-- I don't mean to say that the hypothesis has been proven. I believe that common variants contribute to common disease. There are still people who believe that that's not true at all and that really, this is the only thing that's ever going to yield anything is looking for rare variants. So and there are definitely a couple of cases where this has been true. And this really hasn't been explored to any degree just because it's so expensive.

And I put it there just because if all of a sudden, to sequence somebody's genome for $100 became feasible, you would take 1,000 people with diabetes and sequence their genomes. And that would be the way you do the experiment. So that's why it's on there. So this is a cartoon version of what I just told you. Linkage analysis lets you search the whole genome and points you to a region, although it does not point you to the gene.

So remember, those markers are 10 million bases apart. And in fact, because of some of the vagaries of complex trait genetics and the reasons that linkage isn't that well-suited, usually you get about a 20 centimorgan region or about 20 million bases, which can have something like 200 genes in it. And it'll say somewhere in those 200 genes, there's something that's contributing to disease.

Whereas association, you're looking one gene at a time, so you're not-- but you have to guess correctly where to go. And so you have to-- so there's this haystack here of 30,000 genes or 30,000 straws. And you can pull one out and say, is that a straw or a needle? No, it's a straw. Next one. And look.

And obviously, the success depends on you being able to guess correctly. But we actually-- at least for the first time, we know what the 30,000 straws are. So there's a website now with the entire genome on it. And we're getting pretty good, as I'll tell you, at telling whether things are real contributors to disease or not.

But it's this interpretation of whether this is a straw or a needle that is going to be the focus of a lot of the talk because that's the study that a lot of people do is they pick a gene and they test it, and they claim it either does or does not have a role in disease. And I just want to give you hopefully some sophistication in reading those sorts of studies to try to say should you believe everything that you read?

The obvious answer is no. But so this is what an association study is. You can imagine that you had healthy individuals and people with Alzheimer's disease. And there's a common variant in the APOE gene called APOE4. It encodes apolipoprotein E.

And it is a common variant because if you look in the control population, you'll see it's there about 10% of the time, something like that. But if you look in people with Alzheimer's disease, you'll see those frequencies increased about threefold. And so this is the-- this is the paradigm for associations of common disease and common variants, and this has been seen over and over again. You look at just about every study that's been done, you see this increase in risk. Yeah.

AUDIENCE: Sorry.

JOEL HIRSCHHORN: No, no. Sure.

AUDIENCE: Should I-- You also looked, obviously, [INAUDIBLE]

JOEL HIRSCHHORN: Yeah. So actually, yeah. So there's a little bit of a subtlety to even this association. So it has been seen in other ethnicities. It looks-- although, the last time I looked at this, the data were not strong enough to say that there was definitely a difference.

But it looks like the association with Alzheimer's disease might be a little bit weaker in African-derived populations. There could be a wide variety of reasons why that's true, and I'll go through a bunch of them just in general.

AUDIENCE: Because there's less [INAUDIBLE] like if you just made that for all individuals-- if you made that [INAUDIBLE]

JOEL HIRSCHHORN: No. So the rate, the frequency of APOE4, in this case, doesn't actually vary that much across populations. It's an old allele. Some alleles do, and you're getting into something called ethnic admixture, which I'll definitely get to in a little bit. So as I said, you have to guess right. So there's an issue of which genes do you guess?

And ideally, again, one would like to do all of them. And it's getting closer to be able to do that, but-- yeah?

AUDIENCE: Individuals with the BCRA1 example, when you take the linkage study and find that region that appears to be linked--

JOEL HIRSCHHORN: So that's definitely one thing you would do. So it's actually been hard enough until basically this year to do that. So you have 200 genes. And to thoroughly go through a gene and basically look at all 200 of them has been something that's just basically been out of reach.

So to give you an example, I just got, I think, funded to do exactly that for an obesity linkage peak. And we couldn't have really written to do that as part of a regular R01-type grant. We could not have written an R01 to do that a year ago. It was just too expensive and too hard.

It's just the information about which variants to type, and I'll go through, again, using haplotypes and stuff like that to do it. The information just wasn't there, and the expense was too high. But yes, I agree. Those would be the first candidates. So you would use information from linkage studies.

And you might just say, I don't care about anything else that anybody says. I'm going to take linkage. If your complex trait or your common disease is fortunate enough to have a linkage peak that you believe, which is not true for most of them, actually, but some of them it is, you would say, those are the 200 best genes. I'm going to go for those. That's definitely an approach.

You might say, well, linkage is not very well-powered for common disease. Probably most of the genes are not under linkage peaks. And most of the genes under the linkage peak are not the right one. So I'm going to ignore that. I'm going to go for things that-- I'm going to say 100 years of research has meant something.

And if I'm studying diabetes, I'm going to do the insulin signaling pathway and beta cell function and obesi-- fat cell differentiation genes and stuff like that, and those are the genes I'm going to go for. Or you might say, well, this is just lamp post science because it's the biology that people have been doing. And this is not very powered, but this is a genome-wide tool. We can use expression analysis, so survey all the genes, find things that are expressed at different levels in disease versus healthy people and say that some of those or many of those changes might be secondary to the disease state.

Maybe some of those represent a primary regulatory variant that affects disease. And so I'm going to look for those. And what we're trying to do and other people are trying to do is actually the most interesting genes are probably the ones that happen to fall in here, although there's no saying that they have to fall into this intersection. Maybe they fall in here or there or there or even just one of these or in the universe outside altogether.

But I'll just give you an example for type 2 diabetes just because we are actually doing this sort of thing. So there is hundreds of years of biology or 100 years of biology probably on diabetes, a lot of mouse models, single gene disorders that cause the disease. So that's another thing. If severe variation causes a severe form of the disease, maybe mild variation causes late onset form. So you can-- there's a list of genes there.

So for diabetes, this is one of those things where there really isn't a perfect linkage peak to go under. Graeme Bell's group identified the Calpain-10 gene by going under a peak where they had significant linkage. Although, if you actually put all the linkage evidence together from everybody, you would not have said that there was any significant linkage to diabetes there. That association is probably true, but not-- so you can do this, but it's not the end all.

And then there's been expression studies in diabetes that's pointed to a pathway of oxidative phosphorylation. So some colleagues of mine, [INAUDIBLE], David Altshuler, [INAUDIBLE], Cecilia Lindgren, and others, and also a paper from Nick Joselyn. Mary Elizabeth Patty's pointed to that pathway. And there are genes that are actually in this red area, a few of them. So we're going after those genes, for example.

So that's picking genes, and it's more of a-- at this point, more of a art form than any computerized algorithm to do it. Although, just even getting all the information into one place where you could even do the queries of what genes are under linkage peak, what genes are differentially expressed, what genes are in known pathways, that's actually tools that we're just trying to develop.

So which genes, and then which variants? So ideally, you have the disease-causing allele in your hand and you're genotyping it, and you test it because they're the marker that you're testing, as opposed to being a distance away, is perfectly correlate with the variants. That's the ideal situation. And so that implies that maybe we should actually really concentrate our efforts on finding the variants that are most likely to be functional, so similar to the way you might try to find the genes that are most likely to be responsible.

So the most obvious things are missense variants. And these are easy to recognize. It's been shown, if you look at their-- they're much rarer than they should be, which means evolution has cared about them to suppress them. So many of them are actually mildly deleterious. Maybe they're not severe deleterious, but maybe this is just the right balance of things to cause common disease. And even if you're doing rare variants, you can recognize them and group them together and study them as a class, if you want to say-- if you have the model that rare variants cause disease.

So but we know, and there are plenty of cases now, where the variant that's been finally identified is actually for a complex trait has not been a missense variant. It's been a regulatory variant that affects the level of the gene but not the structure of the protein it encodes. There's a great example from type 1 diabetes from John Todd's group and others where there was a missense variant that was thought to be the right variant, but then they did a huge study and showed that in fact, all of the association that they saw with the missense variant was actually explained by something in the 3 prime UTR, and it affected the level of the gene.

So the question is, well, these are hard to recognize. There's lots of variation out there we have no idea how to reach. So the reason we can recognize missense variants is we know the genetic code. They cracked that soon after DNA. But we don't know the regulatory code.

So there's some effort now to try to crack, essentially, the regulatory code by using multiple species and doing comparisons. So this is from Eddie Rubin's group. This is a great website. It's called the ECR browser. And you can see, what you see here is regions that are actually evolutionarily conserved across species and therefore, may be important.

Evolutionists has cared about them again. So they may be important regulatory regions. And there's experimental backup for that hypothesis. So basically, that would suggest that we should-- and [INAUDIBLE] have suggested this approach, at least for missense variants, that we should resequence targeted regions, try to find every variant that's in there, and make sure we type all of them. And again, the issue is that resequencing is expensive.

But we might not have the causal variant in hand if we don't go through this approach or if we're bad at guessing what the regulatory variants are. So we should try to understand also the correlation between variants. I alluded to the fact that the SNPs are redundant with each other. So if you have the causal variant that hasn't been genotyped, you have to infer its effect by genotyping neighboring SNPs. And so they have to be correlated or what's called in linkage disequilibrium with the causal variant.

It turns out that a lot of the times, that's true. So this is actually what the genome looks like most of the time or most of the genome, where that slide that I showed you before, imagine this is now zoomed out. We're looking now 30,000 bases of a bunch of different chromosomes. What you see is if you have a C here and a C here, you can predict that you're going to have a G here, a C here, an A here, and a C there.

And these patterns of alleles are called haplotypes. And in most of the genome, there are these blocks of correlation or blocks of linkage disequilibrium where there are only a few common haplotypes. In this case, there's the red, the blue, and the yellow. And occasionally, there's a shuffling by recombination.

These were first identified by Mark Daly and colleagues in genetics, and then this shown to be true on a genome-wide scale with Stacey Gabriel, et al. And the reason this is useful is that if there's a causal polymorphism but not one that you've typed, you can predict it pretty well by the neighboring markers. So that if you have a C here and an A here, then you probably are carrying this causal polymorphism.

I'll take you through briefly. But I'm going to blow through this a little bit, just the range of what these patterns look like. There's a measure of correlation, which is not important to understand, called D prime. Basically, red means that the SNPs are correlated with each other, and any other color means that you're not sure if they are. And I'll show you a little plot like this, where this is the data for markers 1 and 2, and this is the data for markers 3 and 5.

And this is what blocks look like, where you see a whole bunch of red triangle here, which means that all the markers 1 to 6 are correlated with each other. And then here, 8 to 13 are all correlated with each other. And that corresponds to there being two blocks of linkage disequilibrium.

And again, from the Gabriel, et al. paper, if you look, just concentrate on the figure on the right, in, for example, a European population or Asian population, over almost 40% of the genome is found in blocks of over 50 kb, and the average block size is about 22 kb. So on average, if you look in the genome, there's this correlation that extends for tens of thousands of bases on the order of the size of a gene, although it doesn't necessarily correlate perfectly with genes.

And within those blocks, as I showed you with that red, blue, and yellow pattern, there's only a few common haplotypes present, so in the order of three to four for non-African populations and one or two more for African populations. And they explain the vast majority of the chromosomes in that population. And furthermore, the haplotypes themselves are generally shared, and Africans have a little bit more. And this is because when humans emerged from Africa, only a subset of the geniculate diversity went with them. So Africa is actually a more diverse continent than the rest of the world is.

And I won't go through the reasons that blocks exist, but basically, it has to do with the nature of recombination, the fact that there are hotspots of recombination and this emergence-- this out of Africa pattern. And so that's led to a proposal that basically, if you can recognize these red, blue, and yellow patterns, you can then identify a few tagged SNPs, in this case, two, which distinguish those patterns.

So if have the red variant, you're-- if you have the variant for tag SNP 1, you're red, if you have the variant for tag SNP two, you're yellow, and if you don't have either variant, you must have been blue. And so this means that if you can type enough markers to distinguish these patterns, you can then identify these tag SNPs. And that greatly reduces the effort you need to cover a region.

And so that's the goal of the human haplotype map, which is ongoing. It's a $100 million international project. It's basically to type about a million or more markers in three different reference populations-- European-American, West African, and then East Asian, which is Japanese and Chinese. And you can go to the hap map website or see this nature description. But its goal is to identify all these tagged SNPs.

And so the approach would be basically take the SNPs from the database, genotype these SNPs in reference panels, measure the linkage disequilibrium, and then identify the tag SNPs, and then select them and then take those SNPs and type them in the population. So that's how one would set up, in 2004, to do an association study, where you pick genes. You would understand the haplotype structure. You might even do some resequencing to make sure you had all the missense variants.

And then you would take some set, some subset of variants, and type them in populations and do your association study. The problem is that after you've done all that work to set up to do this, you do an association, you publish it, and you find that it's very difficult for anybody else to reproduce it. So that was a lot of work for nothing, right? So we wanted to understand, if we were going to try to use this as a tool, well, why are association studies not reproduced?

And there's really three possible explanations. One is could be false positives. So when you did your association, it was just wrong. Maybe you've tested a lot of genes and tested a lot of variants and a lot of genes and by chance, you got a p value of 0.01. So you go ahead and you publish it.

And it was just because you had flipped a coin a bunch of times, and you got a row of-- you got a run of heads. Doesn't mean the coin is two-headed. It could mean that, in fact, you were right, but then the people who came after you had little, small studies, and they were looking for a modest effect, and they just didn't have enough power. So it actually was-- the coin was unfair.

So it does come up heads more often than tails, but they only flipped it five times, and it came up tails-- or they only flipped it six times, and it came up heads three times and tails three times. This is equivalent doing a study with not very many people. So that's just a lack of power. And then there could be true differences between populations so that the marker really is associated with disease in your population but not in the population that somebody else studied.

So we set out to overview the association study literature and say, well, what explanations are actually going on here? And we identified, as of 2000 or 2001, just over 600 associations between common variants and disease. And there are way more than this now. There were 166 that have been studied at least three times, so we get to get a handle on their reproducibility.

And the first thing we noticed, of those 166, only 6 were actually highly consistently reproducible. So only 6 were seen 3/4 of the time. So you could say, well, this is just terrible. Almost all associations just are not really reproducible. Or you could say, well, actually, you've only tested 166 things. You found 6 real ones. That's pretty good. There are 10 million things to test out there.

So these are the 6. And one of them is this paradigm that I showed you, APOE4. Then there's things like CCR5-delta32, the CTLA-4. Interesting, this is that missense variant that I told you about that actually now this should be corrected to the 3 prime UTR. So even though they didn't have the right variant, it was still pretty reproducible.

Factor V in deep venous thrombosis and a couple of others. And there are more on this list now, although not many, that are that highly consistently reproducible. So what about the other 160? Are they just complete nonsense, or what? Why are they-- why are they not reproducible?

Well, the next thing we noticed was that of the 160, 91 of them were actually seen multiple times, so not just by the person who reported them, but by somebody else as well. And so that suggested that maybe there was something real going on, but we wanted to look at that in a more formal way. And so to do that, we basically picked 25 of these 160 associations for some diseases we cared about and some others at random.

And we got rid of the first person to report it. And we said, of all the other people who tried to replicate it, how well did they do? So there were 301 studies for these 25 associations. And if all of those 25 associations were incorrect, we would expect that the follow-up studies, in theory or naively, should have-- 5% of them should have, again, been statistically significant with a p value of 0.05. 1% should have had a p value of 0.01, et cetera, et cetera.

So what did we see? We actually found that about a fifth of them had P values of less than 0.05. So this is way, way more than you would expect. This is not just by chance. And encouragingly, most of the associations were in the same direction. So it was the same allele that was associated with disease as was in the original report.

So it's actually-- that would only happen not 1 in 20 times but 1 in 40 times. And furthermore, it wasn't that each of the 25 had a few replications. It was that there was a subfraction of them that had a lot of replications, and then the rest of them were never seen again. So you might say, could this be publication bias?

So we obviously could only look at the reports that were published. So it might be that there were lots of people doing these studies out there, and when they got a P value of less than 0.05, they broke out the champagne, celebrated, and sent the journal-- sent the report off to the journal. But maybe if they did it and the p value wasn't over 0.05, it lingered in the desk drawer. And that's actually the formal name for this thing. It's the desk drawer phenomenon.

And so what we asked was, could there be a universe of unpublished studies lingering in desk drawers around the world that would explain why we saw so many of these studies? How many unpublished studies would we have to hypothesize per association to explain the fraction that we saw? Was it one or two or three, or was it just a ludicrous number of studies out there? And it turned out, actually, you would have to postulate a ludicrous number of studies to explain the positives we saw.

You would have to explain possibly about 40 to 80 unpublished negative studies lingering in desk drawers around the world. And there just aren't even that many people doing association studies for any of these diseases. So we think this is-- and we did some other work to show this is not publication bias. So you asked about ethnic admixture stratification. So I want to talk about that for a little bit.

AUDIENCE: I have one question.

JOEL HIRSCHHORN: Yeah, sure. Go right ahead. Yeah.

AUDIENCE: [INAUDIBLE] a rare disease.

JOEL HIRSCHHORN: Yeah. Yeah. So yeah, it's not a common disease. You're right. So it's a very environmental disease. So what it is is actually, it's probably a common disease, but just the environmental exposure that's required is rare.

So it's like-- and actually, another one was HIV, which unfortunately is not a rare disease now because the environmental exposure is so common. But you could imagine, if the world were a different place, would be a rare disease and the same sort of thing. And it may be the same thing as diabetes. Type 2 diabetes may be a common disease in part because the environmental exposures.

AUDIENCE: Like I said, I'm not sure how they can-- how can you get how you're looking--

JOEL HIRSCHHORN: Yeah, so it turns out it's actually a huge risk factor. It's a huge genetic risk factor. There's an interesting article in *Science* looking at this variant, showing that there's been waves of selection in favor of the resistance variant in tribes where prion diseases are endemic.

AUDIENCE: It's just a pretty common sequence.

JOEL HIRSCHHORN: Yeah. It's pretty common, and it turns out that it's a big effect. So that's how you get power so that even if-- and there are enough sporadic-- people collect a couple hundred sporadic cases here and there. And it's just a big effect, and that's why they were able to see it. So even rare diseases can have common variants that contribute to them.

So ethnic admixture. What is that? So this is an ideal epidemiology study. So when people collect cases and controls, they often will ask about ethnicity. And they have what's called self-described ethnicity.

And the ideal thing is that you would-- clearly, everybody matches on what people put down. But people will self-describe differently. And there may not be an accurate self-description for what people's actual ethnic and genetic backgrounds are. So you hope that somehow that this self-described ethnicity is good enough that you actually have a good match. But you can imagine that within some ethnic group, there might be a couple of subgroups.

And this is obviously the distinctions. I've drawn this as there's the blue subgroup and the green subgroup. But in reality, would probably some shading from blue to blue-green to green-blue to green. So you can imagine that there might be some different ethnic subgroups. And you could think about this, for example. European might be Northern European and Southern European or something like that.

And there might be diseases where the disease actually is more prevalent, let's say, in northern Europe. So you might have more cases from the Northern European ancestry and more controls than Southern European ancestry. And for example, type 1 diabetes might fit this pattern. Much more common in Finland than in South of France, for example.

So if you then had a marker that was tracked genetically with Northern European ancestry but not with Southern European ancestry-- and there are not a lot of markers like that, but there are at least one that I know of-- then you might find that that marker would be overrepresented in the cases relative to the controls just because that ethnic group was overrepresented.

And this is thought to be more of a problem also in populations that have had recent admixture, like African-American populations or Latino populations, where instead of this being Northern European/Southern European, this might be degree of ancestry, so what fraction of your chromosomes have alleles that are more common in Europe versus alleles that are more common in Africa. And you can end up with this same setup.

So you can get, in theory, false positive associations from that. So there are ways around it now. So the first way that was proposed by Rich Spielman and then a whole bunch of other people is to use what are called family test-based-- family-based tests of association. And I'm not going to go through the details of those other than to tell you that they are immune to this problem.

And the other way is to basically, if you're doing-- if you can't do the family-based test, and there are a lot of reasons you might not be able to, you can use something called genomic control, where you type a lot of random markers, and you see, well, do those random markers give you these spurious associations? And you can use those to correct your association results or even to rematch your cases and controls to make sure that their genetic-- regardless of what the self-described ethnicity is, you can match them by their genetic background as assessed by these 100 markers.

So we actually looked-- at least for our little study, most of the things that had replicated are actually seen in family-based studies or in multiple ethnic groups, which makes admixture less likely. But doesn't mean that this isn't going on in a rampant way. And for example, the ones that were never replicated, could well be that some of those false positives are due to admixture.

Then there's true population differences that could be coming into play. There's a lot of different things. It could be different diets in different populations. It could be that there are other genes that are more common in one population than in another, and there's some interaction between what you're looking at and that.

And I want to talk about one very specific type of difference that at least we could address, which is that there might be different patterns of correlation in different populations. And the marker that's being used might not be the causal marker. It might be something which is correlated. So I showed you this diagram, where you could use this CA haplot-- if you look at this two-marker haplotype, you use this CA haplotype to try to predict whether the causal polymorphism is there.

And it does pretty well because there wasn't very much of this recombinant haplotype here. So here, actually, the prediction-- I haven't drawn the arrows. But the prediction would actually not be right for this one chromosome. But most of the time, the correlation is pretty good. But what if there was a lot of that recombinant chromosome in some other population?

Now, all of a sudden, you have a common CA haplotype without the causal SNP. And so in addition to being-- the correlation is now no longer so good, so you might not see an association. So that's at least one possible thing to keep in mind is that the marker that you're looking at might not be the causal thing, so basically understanding the LD patterns around an association. So once an association has been found, typing lots of other markers in there and seeing, is there another marker that explains it better will become important, and almost nobody does that right now.

And then finally, the last thing that we looked at was, could it be that some of the associations that we were seeing in our meta analysis are actually true, but the studies that said that there was no association were actually incorrect? And so to try to get a sense for this, what that requires is that the sample sizes were too small to detect the association, which means, in turn, that the association must have been a weak associations or modest effect.

So we estimated the effects for all of these associations and by pooling all of the data. And so eight of them replicated. And the question was, what was the effect size? And basically all eight increased the risk of disease by less than twofold. And some of them was just a 10% increased risk of disease associated with having the allele.

And for 10%, for example, you need thousands of cases and controls. And almost none of the studies we looked at had that. So almost all the studies were underpowered for almost all of the associations we looked at. So lack of power is rampant in this field. And a negative study by itself, unless it's got many thousands of people, really can't be taken to be worth much other than by itself. It can be worth a lot in context of other studies.

One last thing that I want to just take you through again because it's an important issue thinking about interpreting the literature, when we did this estimation of what was the actual genetic effect, and then we compared those less than twofold risks to what the first report had claimed, almost always, the first positive report had overestimated the actual genetic effect. And we wanted to know if it's consistent with a phenomenon called the winner's curse.

So winner's curse is best described for auction theory, and it goes like this. Imagine that there are a bunch of people who are bidding on an item. And they all have an accurate-- I'm sorry, yes, an unbiased estimate of the value of that item. But it's not that precise. So people are just as likely to overestimate its value as underestimate it.

And they all place bids. And those unbiased bids will therefore-- and let's say they bid what they think that the item is actually worth. Those unbiased bids will fluctuate around the true value. And then one of those will have fluctuated up the most, and that will be the winning bid. So the winning bid, conditional on it being the winning bid, will almost always have overestimated the value.

And the best description of this is a Samuelson article called, "I Won the Auction but Don't Want the Prize." So beware on eBay. But so in association studies, basically the winning bid is equivalent to the first person who gets an exciting enough result that they can publish it quickly and in a prominent journal and that sort of thing. And conditional on them having found something interesting, it's likely that they overestimated the true value.

And the true value might either be a weak effect, a weaker effect, or actually no effect at all. And only time can tell. And so we found that it was consistent-- that the degree by which they overestimated was generally consistent with the winner's curse phenomenon. So this is the take home messages for reading the association study literature.

So if you pick up a paper and it says that something is associated with a disease, there is a possibility that it will be true. But it's I would say well under 50%. It's not 0% either, but it's well under 50%. The genetic effects for associations are likely to be quite modest. So if somebody is claiming a fivefold effect and their confidence intervals go from 1.1 to 40, it's probably closer to 1.1, if not 1.

And then because of these modest effects, you need large study sites to detect these reliably and also to narrow those confidence intervals down. Now, I'll just give you an example, our favorite example, for association of a missense polymorphism, or we think that's the causal polymorphism. In type 2 diabetes, there is a gene called PPAR, which is important for fat cell differentiation and is a nuclear hormone receptor. Probably binds fatty acids.

So the first study that was published is right here. So each line is a study. The point is the point estimate for the effect on diabetes. And actually, in this case, anything to the left of this line is an association with diabetes. And the line around the point represents the confidence interval around that estimate.

So for example, this study up here, even though it shows the-- it trends towards association was considered negative because this line crosses-- the 95% confidence will cross as 1. So this is the first study which you can see the winner's curse in action here. Overestimates is the strongest estimate of risk. And it was followed by, I believe, these three and a bunch of other studies, all of which were considered to be negative because their 95% confidence intervals crossed 1.

We then did a pretty big study where we showed an association. We actually pointed out that all of the literature to date was consistent with an association somewhere in this range. Much weaker than originally described, but modest enough that these studies didn't have power to pick it up. And in fact, as you get larger and larger samples, what you can see is that the confidence intervals-- is that things tend to really focus in.

And you put all the data together, you get a pretty narrow confidence interval. And now, the p value for association is 10 to the minus 9. So it's real, but it took 20,000 alleles from people with diabetes and controls to get there.

AUDIENCE: Are all the studies in the same [INAUDIBLE]. Can you really put them all together and do the overall?

JOEL HIRSCHHORN: So for diabetes, the definition of diabetes is pretty clear. There's a WHO consensus. And they're more or less the same definition. I would say close enough. The allele itself is clearly the same variant.

And assuming that the quality of genotyping was good, probably they were looking at the same thing there. Obviously, the populations are different. So I talked about population differences. So it turns out, also one of the best examples of heterogeneity between populations is for this polymorphism, not so much in diabetes, although it hasn't really been looked at well, but actually, obesity. So this has been looked at a lot for obesity, and wildly divergent results.

It turns out that how much trans fatty acids you eat, or saturated to unsaturated fatty acids, affects whether this polymorphism has an effect on obesity. I can't remember right now which way it goes. Basically, on one diet, which is much more common in Western countries, there's I think no effect. And on another diet, which is much more common in Asian countries, there is an effect on obesity.

So you can imagine that there- and actually, if you look, it turns out that these three are all in Asian populations. So it could be that there is some heterogeneity going on there, although it's statistically marginally significant whether any of these studies are different than any of the others. So these are mostly consistent with if you took darts and threw them all at the same odds ratio, you might end up with points that looked more or less like that.

So but it could well be that there are differences. For things like-- for less well-- so for Alzheimer's disease, there's a little more debate sometimes of whether it's pathology-confirmed or not and that sort of thing. So all right. So what to believe is the question. So should you believe it?

I would say initial skepticism is warranted. Replication, especially with low p values-- and maybe we can discuss what that means-- is encouraging. Large sample sizes are crucial. So if you're going to take anything home, there you go.

So I want to quickly-- so this is a quantitatively-oriented class-- talk about applying Bayes' theorem to interpreting association studies. So Bayes' theorem is one of the fundamental theorems of probability. And in this case, what I've done is I've substituted in the words causal for probability of observing association. And causal is the probability that the variant that you're looking at actually is causal.

So if you go through just plugging it into Bayes' theorem, this is what you get. And what you're interested in is basically given that I've seen an association, what's the likelihood that it's true? That's what you want to know when you pick up the paper. So this has to do with three terms.

One is the probability that something is associated given that it was causal. So that's the power of the study. So would you see an association if it were causal? Would I have seen this association data if it were not causal? Well, that's the p value, the probability of observing the data by chance.

And then the probability that it's causal to begin with, which is your prior probability. So this is the key to Bayes' theorem, which is that you have to specify the priors, which is, of course, always the hardest part. It's very easy to write down Bayes' theorem, but then specifying your prior distribution can sometimes be quite tricky. All right. So I'm going to take you through what I believe prior distributions look like within a couple of orders of magnitude.

So what are the prior probabilities? So I'm going to say there are about 600,000 independent common variants, and it has to do with the degree of redundancy that's been determined by the hap map and stuff like that. There are 10 million in total. Some of them aren't that common, so I'm going to shove those under the rug for a little bit. And of the rest, there's enough correlation that I'm going to say there's about 600,000 variants that you can type.

The other assumption is that at least a few of those are going to be causal. So I'm assuming the common variants play some role in disease. So I'm going to say somewhere between 6 and 60. It may be more, but I would say causal to a degree where you have any power to detect them at all because obviously, if they have no power to detect them, they may as well not exist from the point of view of this study because if you plug in 0 for your power, then everything starts canceling out.

So your prior probability is 1 in 10,000 to 1 in 100,000 more or less. How about candidate genes? Let's say, well, I'm not studying-- you say this paper is not studying some stinking random variant. This is the greatest candidate gene in the whole world.

Well, so we've gone through the exercise of writing down all the candidate genes that we think are great candidate genes for diseases. And you usually come up with somewhere between, say, 100 and 500 genes, so let's say 300 that you come up with. And I would argue that you could tell as equally beautiful a story about almost all 300 of those genes once you knew post hoc that there was an association.

So you can say, oh, of course. That's why protein kinase C theta is the key to diabetes. And you write your story. So I'm going to say that there are 300 candidate genes, all of them about equally likely to be associated. Again, using some things about there are about three or four blocks of linkage disequilibrium per gene, about three or four haplotypes per gene.

I'm going to say there are about 12 independent variants per gene. So there's about 3,600 candidate variants on your list of 300 genes that you've just made. Let's assume-- and this is I think somewhat generous, but let's assume that all that biology that went into making that list is worthwhile and that half of all the causal variants are in the candidate genes. Well, then if you do all the multiplication, the prior probability is about 10 times higher for a candidate gene as it is for a random gene, 1 in 100 to 1 in 1,000.

And then what are the probabilities you suggested looking under a linkage peak? Let's go do all the genes under a linkage peak. Well, what are the prior probabilities there? Well, let's say actually not 200 genes. Let's say there are only 100 genes under a linkage peak.

Again, the same 12 variants per gene. But this time, there's only one-- the assumption is there's only one gene or one variant that's under that linkage peak that's actually contributing. So again, the prior probability works out to be about the same as a candidate gene, 1 in 1,000. So I would say you should probably, at this point, be agnostic as to whether you're focusing all your effort under a linkage peak or on a list of candidate genes that you've come up with through biology.

So let's assume that you now do your study, and you achieve some magical p value of 0.05. What's the likelihood that there was an association that was correct? So the greatest candidate gene in the world. So this is actually-- I'm going to give you 10 times more likelihood that this was correct than your typical candidate gene. Prior probability of 0.01.

P value, 0.05. Posterior probability. This is assuming also perfect power to detect your association, which means you did this in a big study. If you do it in a small study, this number goes down because this thing scales with power. So the posterior probably, chance that you're right, about 15%. So that's a little bit disheartening, maybe.

How about a typical candidate gene? This is the typical association study that's done. There's about a 1% chance that a p value of 0.05 represents a true association. Candidate or linkage peak, so Calpain-10, 1% chance with a p value of 0.05. Now, their p value was lower, but still, that's not much lower.

Random gene in the genome. Basically guaranteed to be wrong with a p value of 0.05. So what kind of p values do you need to actually have pretty good posterior problem? Because again, this is assuming good power. So basically, what you need is p values that are somewhere in the range of 10 to the minus 4 to 10 to the minus 6. And this could be a single study or all studies put together.

So basically, most report associations are probably incorrect but not all of them. Some will turn out to be correct. And I would say that if you have low p values replicated or there is a really, really, really, really good reason for plausibility, then that makes it much more likely to be right. And then just this last point, genes under linkage peaks are about the same as any other kind of candidate. Obviously, if you have a candidate that's in a good pathway and it's under a linkage peak, probably raises its prior probability. Oh, sorry.

AUDIENCE: [INAUDIBLE] you're not just looking at the state that's [INAUDIBLE] independent SNPs being present, contributing to [INAUDIBLE].

JOEL HIRSCHHORN: Yeah. So I haven't talked at all-- I only alluded briefly to gene environment interaction with the PPARgamma and the diet thing. And haven't talked at all about gene-gene interaction. So there's a hot debate going on in the field, I would say, about whether one should look for gene-gene interaction without any evidence that either gene is involved, which basically involves doing all the pairwise at least, if not three-way or four-way or n-way combinations.

And the challenge for that becomes apparent when you say that there are actually 600,000 independent variants across the genome. So even if you're just doing pairwise, that's 6 times 10 to the fifth. So it ends up being 3.6 times 10 the 11th variants-- sorry, pairwise combinations that you have to look for. And so you know that you're actually going to get a p value of 10 to the minus 11th for interaction at some point just because that's how many pairs you've tested.

In fact, you can get 3.6 of them. So it becomes a issue of power again in that to correct for all the different hypotheses that you'd be testing, you need an enormous sample size, or you need to hope that the interaction effect is just so enormous that you actually get a p value that survives that correction for multiple hypothesis testing, which might be the case. But then the question is, if that interaction is so enormous, you have to draw very strange models that you wouldn't also pick up at least one of the two as being causal.

Now, what should be done but isn't is that once you have something that you really believe-- so for example, we work in the diabetes field, and we really believe people are gamma Pro12Ala. But yet, we do not yet-- and we should-- test for pairwise interactions for every single thing we genotype in Pro12Ala because you can draw plenty of reasonable models where you would only detect one of the two things that are interacting. And the other one would be completely masked or largely masked.

So we don't do that, and we should. And but I would say that if you're-- again, if you're picking up a paper which says we didn't see anything with either one of these variants, but then we did the pairwise test and we see an interaction, think about your prior probability that that pairwise test was actually associated. The number of pairs in the genome versus the number that are actually going to be causal, your prior probabilities are probably several orders of magnitude lower.

So unless there's, again, some really great reason to know that these genetic variants interact. So that's my incredibly biased answer to that or not biased but opinionated answer to that question. I won't go through why you get the same sort of winner's curse and things like that in linkage. But just to take you through rare variant association studies, again, just in case you pick up literature on this.

So this is the typical rare variant association study as it is practiced in 2004. Genes are resequenced in affected individuals to try to find, say, missense variants. And then those missense variants are then taken and typed in unaffected individuals. So it's sequencing here, genotyping here. And what's often observed is that they're not there.

So I'll take you through a possible resequencing association study. Let's resequence your favorite gene, gene X. 200 diabetic individuals, and, lo and behold, you actually find 10 rare missense variants, each seen in one person with diabetes. You then type 200 healthy individuals. Those variants are not seen at all.

So you conclude that rare missense variants in gene X cause diabetes. So now I'm going to take you through an equally possible resequencing study. Resequenced gene X in 200 Red Sox fans. You will identify-- if this is a gene that happens to tolerate rare missense variation, you will identify 10 rare missense variants. You might end finding some common variants, but you might not.

Many genes don't have any common missense variants. If you then take those 10 missense variants and type them in 200 other people, say Yankees fans, you will not see those variants. So this is very plausible. And you will conclude that rare variants in gene X make you root for the Red Sox, which of course is not the correct answer.

And the reason for this has to do with the fact that in identifying these rare variants, you have to go to very great depth in doing the resequencing. So you get common variants if you sequence two chromosomes. So that's what I showed you at the very beginning of the talk. So if you just compare two chromosomes and then look, are those variants common in the rest of the population, the answer is almost always yes.

And the average allele frequency is about 1 in 5. So average is 20% for the minor allele and 80% for the other allele. But if you then sequence a large group of people, you obviously find all these common variants. You find those in the first four chromosomes you do. You basically run through all the common variants.

But you actually keep finding variants as you go every once in a while, and this has to do with population genetics and what the frequency distributions are and stuff like that. But most of the variants you find, if you do deep resequencing, end up being these rare variants. They don't account for most of the heterozygosity because they're so rare. And in fact, their average frequency is something like 1 in 10,000. So if you type them in 200 people, you're never going to see any of them again.

So what needs to get done is basically, you need to resequence the controls as well. There are other ways of doing it. For example, if you have families, can see that the rare variants actually segregate with disease. That's sort of equivalent to resequencing control. So that's one way around it.

But that's basically what you need to do. And furthermore, what I said there in that previous slide was you identified 10 missense variants. Well, what if you identified 11, but one of them was common? And you said, well, that's a common one, and it's present in cases and controls, so that's probably not functional. That's probably-- and I'll just concentrate on the ones that are only present in cases.

Well, that's obviously rigging the results beforehand. So if you limit yourself to just those variants that are present in cases and say those are the only ones that matter, obviously they're only going to be present in cases. So you have to have some way of grouping these rare variants for analysis before you know the results of the association study of which ones are in cases and which ones are in controls, so something like either I'll only look at frame shifts or nonsense mutations, or I'll only look at things that have this particular in vitro functional property or something like that.

So just in the last couple of minutes-- or I'm actually way ahead.

AUDIENCE: Seven minutes.

JOEL HIRSCHHORN: Oh, no. Not way ahead. Good. 12:30. Good. For a minute, I thought I had 37 minutes left. I was like, I've been talking longer than that.

OK. So what could we learn? And I think this is also important to think about in what are the goals, and what are the likely outcomes? There's a lot of hype around the whole field of genomics and SNPs and common disease and just to try to take a look at what is and is not likely. So I don't know if anybody's seen the movie *Gattaca.*

So *Gattaca,* I think people actually have *Gattaca* in the back of their mind when they are thinking about this, which is that we should be able to tell by genotyping, if we could sequence somebody's genome, we will be able to predict-- if you look at the opening scene in that movie, it's like, probability of coronary artery disease, 98%. Probability of stroke, 100%. This sort of thing.

So we already know that genetic variation, actually, even if we had perfect information, only explains 30% to 50% of common disease. But what's even the likelihood that we'll get that good a prediction? I think it's probably not that great because it's probably a lot of alleles that are going to interact in complex ways that we may not be able to figure out for a very long time. But nonetheless, there is some thought that we may be able to identify people who are at high risk or particularly, people who might be at low risk.

If that were the case-- and I've shown here probably a very optimistic scenario, which is that you could enrich this much for the high-risk people-- you might say that these people would get a more aggressive intervention preventive measure, and these people might get the standard of care or might even be able to get slightly less aggressive interventions depending on how the risk-benefit played out. This obviously requires that there be effective preventive measures and things like that.

But you could imagine, for things like diabetes, you could say, you really are at high risk for getting diabetes. You really need to exercise and lose weight, and maybe that there would be some-- you could hook that person up with a nutritionist and all the personal trainer and all that sort of stuff. And it might not be something that we, as a society, could afford for everybody but maybe at least for the people that are the highest risk, we could afford it if we could get there.

But I don't think that this is necessarily really where the true promise lies. There's another possibility. This is called pharmacogenomics, which basically reclassification of individuals to guide therapies. So you have this mix of people who have some disease, like type 2 diabetes, but it turns out it's actually type 2A, 2B, 2C, and 2D. And maybe we could classify them by DNA sequence or expression profile or some genomic method.

And we could say, well, these people will do best with treatment A, these people with B, these with C, and these with D. And again, that may or may not come to pass, and certainly, people are looking at it. And I think it's going to partially come to pass. I'll just give you one real-life example that actually is not in the clinic yet, but I'm not sure why not.

So there's a gene called CYP2C9, which actually metabolizes warfarin, which is used commonly to treat conditions involved with thrombosis or blood clotting. This is an anti-clotting agent, actually, vitamin K. And this didn't quite-- I'm sorry. The white on yellow is not a good color scheme. But basically, most of the population has zero low-activity alleles.

20% of the population has one low-activity allele, and 5% of the population have two low-activity alleles. And if you look at the average dosage that people end up on by their genotype-- so this dosage is not based on their genotype. It's like there's trial and error by the physician, and they find, oh, all of a sudden, you have a nosebleed that won't stop. Maybe your warfarin dose is too high or your PT value is too high. We have to cut the dose.

You see that there's a very nice allele versus dose correlation here so that if you have two low-activity alleles and you get put on the typical 5 milligram dose, you're going to have bleeding problems. And in fact, you're going to be at six times the risk of serious complications. So personally, if I were going on warfarin, I would probably want to know what my CYP2C9 genotype was. But it's not yet standard practice.

One of the only cases that I know that is is actually if you have cancer and you're getting six [INAUDIBLE] purine or related drugs, there's a gene called TPMT, which again metabolizes that into a toxic intermediate. And that has now become pretty much the standard of care because again, the risk of death is higher if you have the wrong genotype. So death gets people's attention quite--

AUDIENCE: And that's--

JOEL HIRSCHHORN: --vigorously.

AUDIENCE: Average dose, right? So--

[INTERPOSING VOICES]

JOEL HIRSCHHORN: Yes. So there's a range.

AUDIENCE: --histogram on the--

JOEL HIRSCHHORN: Oh, absolutely. Yes. So I have to say, I should have put error bars on here. I was looking at that and just noticing I did not have them.

[INTERPOSING VOICES]

AUDIENCE: --because even if you did a histogram--

JOEL HIRSCHHORN: Right. But--

AUDIENCE: --it doesn't necessarily change--

[INTERPOSING VOICES]

JOEL HIRSCHHORN: Ah. So right. So if this were-- if you could say, well, physicians can do this without knowing ahead of time, then this should not happen, right?

AUDIENCE: Oh, yeah.

JOEL HIRSCHHORN: Yeah. So yeah. So this is the argument is oh, well, physicians are going to be good at this anyway. They're not. They suck at it.

So but I think the other thing that is potentially useful, so if you look at least the two most well-established risk alleles for type 2 diabetes, turns out, they're both in drug targets. Now, this is admittedly a little bit circular because the reason they were looked at was because it was known ahead of time that they were drug targets. So they made sense as being good candidate genes. But it is interesting that two of the major drugs that are used to treat type 2 diabetes, there's variants in their targets that affect the genetic susceptibility to diabetes.

And of course, the promise here is that if you could find all of the other genetic pathways, even if they only had a 1.1-fold effect on disease-- so PPARgamma is only about a 25% risk effect on risk of disease. But you could easily imagine that you come up with a pharmacologic intervention that does a lot more than the genetic polymorphism does. So there's only so much tolerance for how much genetic variation evolution will let happen.

But it might be that giving thiazolidinediones might be like having 10 times as much of an effect as having an alanine allele does. So the idea is that hopefully by identifying the genes that are responsible-- and this is true for severe forms of the disease as well-- you might identify drug targets. And finally, of course, this helps just understand human biology, and that's always a worthy goal.

So should talk about potential difficulties from all these-- from all these advances. There are obviously concerns about privacy with genetic data. And don't really have time to go into it, but obviously, the big fear has to do with insurance discrimination and employers knowing about genetic diseases. And there's bills that keep making it almost all the way through Congress and not quite to deal with this. So lobby your congressman.

There's the *Gattaca* mindset about improper interpretation of predictive information, how predictive genetic information is. There are companies that will try to sell you that you should be on diet A or diet B based on SNP gene types that are in those irreproducible associations. There can be psychological impacts of if you say you have the high-risk allele for diabetes, that might-- without an understanding that this is a correlative rather than a perfectly predictive measure, that could have significant harm as well as impact on reproductive choices, again, depending on how predictive people think things are.

There are huge interactions with concepts of race and ethnicity. We've alluded to this a little bit with different allele frequencies and different populations. Most alleles are present around the world, and you can't use generally any one variant to tell different ethnic groups apart, but there are definitely differences in allele frequencies between populations. And that has significant implications.

And then finally, I certainly don't study the genetics of performance, but you could imagine that if people started getting into that, that could be quite dicey. And even if the science were perfect from a technical standpoint, it might create a lot of trouble.

So David Altshuler is a collaborator with me on the diabetes work and also very involved in the hap map with the work that was done by Stacey Gabriel, Mark Daly, Steve Schaffner on haplotype blocks. The diabetes group also is Noelle Lake, Cecilia [INAUDIBLE] did some expression stuff. And Kirk and Lee and Eric were on the meta analysis association study. And there are a lot of genome project-type things that are obviously giving us all the tools to do this. So thanks.