Conceptual Basis for Large Population Studies of Human Genetic Variation and Common Disease
David Goldstein, Ph.D.

DR. TUCKSON: To begin, let me just thank David for coming, and we are very interested in the next half an hour to hear you talk to us about the conceptual basis for large population studies of human genetic variation and common disease. David, thank you and welcome.

By the way, folks, I think what we'll do, depending on how long the presentations take, I think if they stick to their half-hour allotment, what we may do is if you have an urgent, burning question that you want to ask the individual speaker, we can probably take one or two of those right after, but then we'll also try to query the panel later.

DR. GOLDSTEIN: Well, thank you very much for the invitation to come here and talk about the conceptual basis for large population studies.

What I'd like to do in half an hour is try to cover two things. One is why we might want to undertake such an enterprise, and secondly, how we might go about it in terms of what the technical requirements would be. I'm going to kind of bounce back and forth between those two things.

But kicking it right off, why we would want to set up a powerful framework for studying the genetics of common diseases, the basic motivations are indicated there. We would like to be able to predict risk, but importantly, and I'm going to come back to this a few times, we would like to be able to not only predict risk, but do something about it. It's not really good enough just to predict risk. This is not for insurance companies. It's not good enough just to predict. We have to be able to intervene. So that's something that's going to up, I think, in a few places.

The other motivation is not about prediction and intervention, but it's about identifying genes and pathways that might help us in the drug development process.

inally, the aim would be to identify genetic determinants of treatment response, and that's traditionally thought of in terms of pharmacogenetics, which I will talk a bit about, the genetic determinants of what drugs are safest and work best for a given patient, but you can also think about the genetic determinants of other kinds of treatment responses, such as when there options for surgical procedures and non-surgical procedures and so on. So in general, in the genetics of treatment response.

So the first thing that we need to be clear about is what kind of genetic variation we're talking about, and the first thing that needs to be said is we're not talking about the kind of genetic differences indicated on the slide here, where you've got a mutation that is segregated in a family that causes a disease. So in that simple Mendelian case, there is a 1:1 correspondence often between a genetic difference and the disease that we're interested, and that's actually quite straightforward to work with genetically and the community is now extremely good at finding those kinds of causes of disease.

Now, unfortunately, common diseases aren't like that. The genetic contributors to common disease don't have that kind of 1:1 correspondence.

So the kind of genetic variation that we're talking about here is illustrated with this cartoon. The idea is that our genome is a big place. There are many places in that genome where individuals tend to differ one to the next, and in fact there are now estimated to be more than 10 million common polymorphisms, and that is to say a site where the rare form has a frequency of more than 1 percent. There are more than 10 million of those different places in the human genome, and if you allow for rarer variants, then of course there are many more than that.

These variants, the different forms of many of these sites, we know often have very subtle effects. So they change physiology in some subtle way. That's very difficult to measure.

Then these variants influence the phenotypes that we're interested in -- that is, the kind of diseases that people get -- in some kind of complicated interaction, both with other genetic differences in our genetic makeup and with the environment. That's what really creates the challenge. There are a large number of variable sites in our genetic makeup. They interact with one another, they interact with the environment, and then ultimately they have some kind of influence on what we're interested in looking at, and that is the health of the individual.

I really just want, in walking through this, to emphasize that at the end of the day what we're talking about is the probability of certain conditions being influenced by these variants. The variants do not determine the conditions, and for that reason I think it isn't really appropriate to talk about genes for diseases. We're not doing the same thing as we did with Mendelian disease. We're finding the gene for diabetes and the gene for asthma and so on. We are understanding on how genetic differences influence these conditions. So its a different kind of thing.

So that's what our aim is, is to understand how all those genetic differences that we have influence our health. That's the aim. It looks like it's going to be difficult. There is now really no question about that.

But what I'll now turn to is some of the technical requirements that we're going to need in order to be able to make progress. I'll spend the most time talking about the requirements to efficiently represent genetic variation.

There are two reasons for that. One is that I was explicitly asked to do that, but the other reason is that's where we're farthest along. When you actually hear people talking about the genetics of common disease, nine times out of 10, people are talking about how good we're getting at sequencing and genotyping and how much we know about genetic variation. We actually have gotten quite good at that side of it.

That's the easiest side of it by far. The difficult side is things that we actually haven't made much progress on, which is knowing exactly how to measure in patients what we need to measure and knowing how to relate that to the genetic variation. That's the harder bit. So I'll spend more time talking about what we're better yet, and then just sort of telegraph what we're not so good at and some ideas about how we might improve on that.

So first, kicking off, the genome is a big place and it's got a lot of genetic variation and, as good as we are now at sequencing and genotyping, we can't simply get very, very large numbers of individuals that suffer from a certain condition and individuals that don't and exhaustively compare them genetically. We're not capable of doing that right now. We might at some point, but that kind of capacity has always been promised to be right around the corner and it never quite arrives. So what people have been thinking a lot about is more efficient ways to make these comparisons and more economical ways.

Something that's getting a lot of attention right now is called "haplotype tagging," which I'll now spend a few minutes talking about. The basic idea here is to find a framework for efficiently representing the genetic variation either in a region of our genetic makeup that you're interested in or in the entire genome.

I don't know how well you can see this, but what's shown here is a cartoon representing a stretch of the genome. You could consider that a gene, and indicated are each of the sites in that stretch of the genome that differ, where there's a polymorphism.

So there are 12 sites indicated there, and I'd just point here to this green group. Those are four polymorphisms that are indicated in the gene, and so you the first row is one chromosome you might sample from the population, and in that chromosome, that find site has a T allele and then the fifth chromosome you might sample from the population has an A allele there. Then you've got the next polymorphic site which has the alleles that it has and so on.

The point here is that members of the green group are all associated with one another. So in this case, if you know the allele that's present at the first sites, it tells you the allele that's present at the second site in the green group, and the third, and the fourth.

Now, those associations among variable sites in our genome are due to a whole raft of population genetic forces which I won't go into, but they do exist. There are these associations. They're usually not perfect. I'll say something about that in a minute.

But they do exist, and because of that, if you were interested in looking to see if any of those sites associated with a trait you were interested in, you would not have to directly assay all of them. You could assay one member of the green group and it would tell you about the others. You could assay one member of the pink group or whatever color it is and it would tell you about the others and so on.

These associations are called "linkage disequilibrium," and so another name for this is linkage disequilibrium mapping, but the point is these associations do exist and if you understand the nature of these associations, then you know how to select out a subset of the variable sites that tell you about the others.

In this particular case, obviously the subset that you can use is one member of each color group, and there is no loss of information at all because each member is telling you about the others. So if one of the ones that you did not assay was influencing the phenotype, you would still see it through the one that you did look at.

So that is, at its conceptual core, the entirety of haplotype mapping or linkage disequilibrium mapping, and it is in fact the primary motivation, I think as far as I'm concerned and most people are concerned, for the HapMap Project, which is an effort to characterize these patterns of association among variable sites, so that you can select out a subset that efficiently represents the variation in our genetic makeup. So that is an extremely important tool currently because we can't look at variation comprehensively, and that's the conceptual core.

Now, in fact the association, because we're doing biology here and this is not physics, these associations, of course, are never perfect. So you actually have to use a whole bunch of messy statistics to go through this step of choosing one member of each color group, but that really is a technical detail. This is the basic aim.

What I'd now like to do is just take a couple of minutes addressing the issue of how well we expect this work. So can we feel comfortable that we really do have a good framework in hand for efficiently representing variation? I'm going to try to give a yes or no answer to that question.

I'll illustrate that with some work that we did on a data set that we collected together with GlaxoSmithKline, where we looked at these patterns of association among 55 genes that encode major drug metabolizing enzymes. There were a bunch of these variable sites or polymorphisms that were assayed in a number of individuals, both of European ancestry and Japanese ancestry, throughout all of these genes. So that's the data set.

This just indicates the way that this sort of analysis is carried out. This is the stretch of sequence indicated and there are genes indicates, and there are all the polymorphisms indicated that we looked at as thin lines. Those are about 60-plus of them spread through four genes that are contiguous.

What you do is do a statistical version of selecting one member of each color group and you identify nine out of those 60-plus polymorphisms that you assess are able to represent the other variation that's there. Then the question that you want to answer is, well, how well is that really going to work in representing variation that, A, you don't yet know about, and B, variation that's in a somewhat different population from the one that you looked at originally?

That's important because you have to remember that the way this works -- for example, the way that we're all going to use the HapMap data, is the HapMap looks at a number of individuals, for example, from the SETH repository -- so these are individuals of North European ancestry -- selects these special tagging SNPs, and then goes and applies them in a different group. For example, our case, patients with epilepsy and so on. So you have to ask the question how well will they represent variants that you may not know about initially and in a somewhat different population? So you need an answer to that.

So in this case, we find these nine SNPs to represent all these others, but what you want to know about is how well they represent SNPs that you actually don't yet know about and in a somewhat different population. So you think of statistical ways to do that, which I'm not going to talk about, and evaluate how well they do.

We went through a few of those exercises, which, as I said, I'll skip, but what I'll do instead is show a direct evaluation of whether or not they work, and that is taking these SNPs that you identify out to a brand new population sample and assessing whether or not they predict variable sites that we know are functional. So there are in these particular genes lots of sites that we know change the activities of the enzymes, for example. Those are exactly the kind of differences that we're looking for and we can ask do these tagging SNPs work?

This shows the result. Shown here is the minor allele frequency of the SNPs that we're trying to predict, that we're proposing not to type, and here is a measure of how well we can predict them. It doesn't really matter how that measure works, but what does matter is that if you're up here at the top in this performance measure, that is exactly the situation, and you can show this formally, of the cartoon. If you're up here at 1 in this performance measure, it's exactly like taking one member of each color group that exactly predicts the others with no loss of power whatsoever.

If you're in this range, you do very well, and if you're down here you do very badly, which is to say that if there was a SNP down here that you did not type and it was influencing the condition, you wouldn't see it.

So how do you? Here's the minor allele frequency of what you're trying to predict, here's the performance, and once you're above about 5 percent, you do great. So it's fair to say the short, non-technical version is that out here, if any of this stuff was influencing the phenotype and we only typed our tagging SNPs, not these things directly, we would still see it. So that is really encouraging.

This is the very discouraging note. It's a small sample size so far, but the very discouraging note that these rare things may not be predicted at all. Sometimes you predict them and sometimes you don't.

Now, we've gone on and done a bit more of that kind of thing, and our impression is that this is a fairly general outcome, that in this framework you just can't reliably pick up the variants that are rare in the population, where rare is something between 3 and 5 percent as a cutoff. More work needs to be done, but that's how it looks to us at the moment.

So what's the conclusion from that? What I'd like to emphasize is that we are talking about a truly dramatic economy. In the 55 genes that we looked at, we estimated that there 4,000 common polymorphisms, and what we show is that about 200 of these specially selected SNPs can represent the other 4,000.

Now, you can select these in different ways and some people would use methods that would result in a number slightly larger than 200, but it is really dramatic economy that you can achieve this way, and I would assert that it is now not controversial whether or not you can represent common variation in this framework. It's still discussed a little bit in the literature, but I think that debate really now has gone out of date. I think it should be viewed as demonstrated that this framework can officially represent common variation.

I should say that I have no association with the HapMap Project, so I don't feel any need to support the necessity of the HapMap Project. It's just a technical evaluation. That framework really does seem -- not seem. Has been demonstrated to work well in representing common variations. So I think that's really encouraging, and of course, these data that we have are by no means the only data that make this case.

So common variation can be efficiently represented. We should view that as non-controversial.

It seems unlikely that rare variation can efficiently represented. So for that, we don't have an economical approach. If we want to also identify the rare variants that influence both common diseases and responses to treatment, we're going to have to do more difficult, more expensive things, and we should because, without a doubt, rare variants will also contribute -- I'm not going to go into that whole debate, but I think it's quite clear to most people that both common variants and rare variants contribute to common disease. The relative importance of those two things, we don't know, but they're both going to make some contribution.

So we have a very economical method for representing common variation. We don't for representing rarer variation. I don't expect that tagging will actually serve the purpose, but you may find more clever methods to do it perhaps, and we probably need to think about alternatives.

So I think in terms of representing common variation, the genetic side, we really are now in pretty good shape. Even though we've got a challenge for rarer variation, it's terrific that we can now start asking questions about those 10 million genetic differences among us all. That's terrific. That's a real tool that will no doubt lead to advances.

But what is much, much more complicated is deciding about how to look at individuals that are being studied genetically, both individuals that have diseases and individuals that don't have diseases.

So for example, if you're thinking about prospective studies, and many people have been making arguments for the advantages of prospective studies, and that is where you enroll people that are random samples from the population, for example, in one design and monitor them over time, and as they become affected by different common diseases, you can then carry out genetic studies knowing about the background of the individual because they've been in your study for awhile.

So as we move to carrying out those kinds of studies, which do have a lot of advantages, we need to think about exactly what information we need about individuals at the time of enrollment, and I don't have time to go into details here, but I would say that that's something that we really don't have a very good idea about.

For example, if you're interested in cardiovascular disease, exactly how much information do you need at the time of enrollment for a large population sample in order to understand the state of the person when they're 50 well enough that it really tells you extra things about why they had a heart attack when they were 66? And we don't know exactly what we should be looking at when we enroll individuals for cardiovascular disease or for other things. We really don't know.

So if we move towards very large prospective population studies, that's something that we're going to have to figure out. Obviously, lots of people have ideas about it, but it's not like the genetic side where we really know what we're doing. It's definitely an area of active work.

The other thing I'd like to raise as an issue is the question of what types of information are the most important. For example, we've been carrying out a variety of studies in epilepsy, and a common way that people have been thinking about doing epilepsy work is the sort of thing that people usually do, which is you get a lot of individuals that have epilepsy and you compare them to a lot of individuals that don't have epilepsy.

Yet epilepsy has quite a striking potential, in that in cases where patients don't respond to pharmacological treatment, surgery is carried out and the actual affected tissue is then available for study, so that you can look at the seizure-focus tissue in those patients that have to undergo surgery.

That is basically not being done in epilepsy research, and you can actually write out a long list of striking opportunities like that if we look at the right place and interface correctly with the actual care, clinical care, of patients where we might really figure new things out if we actually look at the right kind of information, and sometimes that right kind of information doesn't come from simply enrolling a million people in a study.

I'm not disparaging that. I'm saying there are other kinds of data that are available that emerge from clinical care that we are not making systematic use of. In the area that I'm familiar with, it's certainly the case, and in a variety of other areas. So I think we have to think very carefully about how we interface genetics work with health care to make sure that we really do capitalize on the most important types of information as, for example, we most certainly are not doing in epilepsy, although, of course, we're trying to change that now.

Another point that I would like to raise in that context is the overwhelming importance of having detailed information about how patients respond to treatment. I'm not going to have a lot of time

to talk about this, though I'm going to talk a little bit about it, but I think that it is now very, very clear that genetics plays a major role in influencing treatment response -- in particular, responses to medicines -- but in order to make progress in identifying the genetic differences among patients that influence how they respond to medicines, it is essential to have very detailed information about what medicines they were given, in what doses, in what combinations, and exactly how they responded. So we're not going to be able to make progress unless we have that available and that's very, very difficult to get.

In that context, I'd like to mention that one opportunity for getting that kind of information may in fact be through managed health care providers. Where the patient records have been electronic, that may be a framework for getting exactly the kind of information about drug response that you need. But in thinking about very large population studies, I would say that it is absolutely essential to make sure that you do the best job that you can do in representing how patients respond to medicines.

So I'd like to just end in the last four or five minutes with a couple of thoughts, A, about what we're trying to do, and then B, about the case for more serious attention to pharmacogenetics.

First, on the point of what we're trying to do, I would like to just raise the issue that in academic genetics research there's been a real focus on a final and accurate determination of whether a given polymorphism really is a risk factor for a given disease. In some contexts, that's something that you would like to know. For example, in prediction, you would like to know whether a polymorphism really is a risk factor, but one thing I think that's not so well appreciated is that there are contexts where you don't need to know with certainty whether a polymorphism really is a risk factor. It's good enough to have an educated guess.

Now, I'd like to make that point by a reference to a project that GlaxoSmithKline has carried out, which I have not been involved in, but I report this with permission, and what they've done is done a genetic study comparing individuals with and without Type 2 diabetes, and they've tried to identify polymorphisms that are associated with diabetes. What they did is they looked at 400 individuals with diabetes first and 400 individuals without, and then they had a follow-up.

The size of those studies, and we know this already from calculations you can do in advance, are not sufficiently powered to reach a final determination with any degree of statistical confidence that a given polymorphism really is a risk factor for diabetes. In fact, reaching that final point of confidence is hugely expensive in diabetes because we know that the effect sizes are small.

However, what they did come up with is a set, when they went through that exercise, of 21 gene variants, genetic differences, that appear to be associated. None of those 21 clearly, with statistical confidence, is in fact a risk factor, but you can ask the question in a somewhat different way. You can say I don't care about any single one of those. I care about the set of 21. What is the probability that at least five or six out of the 21, even though I don't know which one it is, really are disease-associated? That's a completely different calculation, and in fact, in this case, what you find is that probably, with fairly good confidence, five out of the 21 are real, but you don't know which.

Now, that's actually still very useful, because in the context of drug development, that means you can take all 21 and start working on them. You don't have to know exactly which one it is, and you could ask the question if it's going to cost you another $250 million to get really precise assessments for each of those 21, maybe it's actually better to spend $100 million and start screening some of them.

So what I'd like to point out is that when we're thinking about drug development, it is not necessarily always just a matter of reaching a final conclusion, no matter what the cost is, of whether a given polymorphism is in fact a risk factor.

And the ending, two minutes, is the case for pharmacogenetics. I think that in academic research, as far as I'm concerned, there is a slightly inappropriate overemphasis of studying predisposition directly as opposed to treatment response. It's starting to change, but I think it hasn't changed enough, and I just want to make the case that variable responses to medicines is, A, hugely important, and B, easier to do than directly studying disease predisposition.

So these numbers, the study that they're based on has many methodological issues and they are highly debated, but nonetheless, however you look it, it's quite clear that variable responses to medicines is hugely important. It has been estimated that adverse reactions to medicines cause over 100,000 deaths in the U.S. alone, ranking as the fourth or fifth leading cause of death.

In terms of variable efficacy, as in fact a senior vice president for GlaxoSmithKline pointed out, medicines typically don't work. So the average rate at which a given medicine does what it's supposed to do is about 50 percent. It varies across therapeutic areas, and a lot of this variation is genetic. We know that, but we haven't found it.

So I'd just close by saying that when you actually start looking in detail at the genetic determinants of drug response, what you find out is that it's usually quite a bit simpler than the genetic basis of common disease.

That has two components. One is that you often know where in the genome to look for possible genetic determinants of drug response, and two, the genetic determinants of variable drug response often are common. So they are not the rare things that are hard to find.

The final point is that when you find a genetic determinant of variable drug response, there is often the possibility of doing something about it clinically. The possibility. It's not immediate, but you often, for example, have the possibility of suggesting that you use Drug A instead of Drug B or that you change the dose.

That, as a final point, is in sharp contrast to predisposition studies of common disease, where sometimes you find things that really are risk factors and there's nothing whatsoever that you can do about it. For example, ApoE4 is the classic example of that. Certainly, that doesn't mean we shouldn't do common disease predisposition, but it does certainly mean that in thinking about these large population-based studies, we've got to take the drug response side and treatment response side more generally very, very seriously.

I'd like to end there, and I should mention the people that worked on some of the stuff that talked about. Thanks.

DR. TUCKSON: Thank you very much. Very well done.

Is there one hot, burning question? If not, we'll come back and do it at the panel.

(No response.)

DR. TUCKSON: David, thank you very much.