

Complement Factor H
29-Apr-06 - 1:00 pm - 4:30 pm

Jonathan Haines: Ok well, I would also like to thank the organizers for inviting me to come down, and for putting together such a nice symposium today. I'm going to be talking a little bit about the methodologies, the genetic methodologies behind some of the studies that you have heard about, and that you will be hearing about for the rest of the session. I think it's important to start off by understanding the different types of human genetic disease that are out there. And as was pointed out by Dr. Sieving early on, there are diseases of what we call a simple genetic architecture. These are the Mendelian diseases. The eighteen hundred or so that have been mapped, and a lot of the different genes for these have been identified. We call them simple, because you can tell how the trait has passed in a family. It follows a recognizable pattern. In this particular case you've got a dominant type pattern, this vertical transmission through families. You usually have only one gene that's responsible for the risk of the disease in a single family. These are often called Mendelian, because they follow simple Mendelian rules of inheritance. And they're often called causative genes, because the variations or the mutations in those genes tend to be very severe, and tend to be the only that's necessary to cause the disease. Some examples of that would be congenital glaucoma, juvenile glaucoma, and lots of different types of retina depigmentosus. Of more interest and of more relevance to macular degeneration is the second category, the diseases of a complex genetic architecture. This is what has been teasing and frustrating geneticists for about the last five to ten years. The problem with these is that there's no clear pattern of inheritance. In most cases you have either a single affected individual with no family history, or maybe there's a sibling, or maybe there's a cousin who's affected. Through various epidemiologic studies you may have moderate to strong evidence that the disease is in fact inherited that there is a genetic component. But it's not as simple – a simple one in any one family. They tend to be much more common in the population. Obviously macular degeneration is one. Glaucoma would be another, and some of these others. Typically they are thought to involve many genes, and possibly gene and environment interactions and gene, gene interactions as you just heard. And we tend to call these susceptibility genes, or the variance in these genes tend to be called susceptibility variance, because they are not sufficient within themselves to cause the disease. They are simply elevating, or in some cases protecting from the risk of developing the disease, and primary open angle glaucoma is one example, and macular degeneration is obviously another example. So it's the complex genetic architecture that we are trying to tease apart when we start doing these genetic studies. So if you want to do a genetic study in complex disease, here's how you do it. Write it down. The point of the slide, and I'm not going to spend the next hour walking through this, is that it's a fairly complicated process. And it's a very iterative process, you'll notice that there are not a lot of different things one has to consider, and these arrows here indicate the recursive and reiterative nature of what you do. You're constantly reevaluating where you are, and what techniques you want to use. And based on the information you get, you may want to modify the approaches that you use. So although this looks somewhat linear, it really isn't. It's a much more complicated process than that. And there is no standard paradigm for how to do this. For Mendelian disease there are fairly straightforward ways to attack the problem. For complex disease

there are lots of different ways, all of which have advantages and disadvantages, and none of them are perfect, so there are lots and lots of different ways of attacking the problem. So I'm going to simplify that a little bit, and talk about a couple, a subset of some of those approaches that have been used fairly successfully now in macular degeneration. So there are multiple different approaches that can be successful, and this is – I want to reiterate this that there's not a single way to get to the right answer. There are multiple different ways that you can get to the same answer. And there are four that I'm going to talk about here. One is the genome screen using association analysis, using association approach. And in that approach one tests large numbers of markers typically they're single nucleotide polymorphisms or SNPs, single base pair changes across the genome, and you test those directly for association with the disease. Does this allele at this SNP associate increase the risk of disease? That's one approach. Another approach is to do a genome screen, but to use what I call a location approach, but basically to use genetic linkage analysis to get at the answer. And in this you sequentially narrow down from the entire genome to smaller and smaller chromosomal region or regions that are identified by genetic linkage, and you test the markers that you're doing pretty much solely on location across the genome. A somewhat related approach is the genome screen using what I call a functional approach, which is essentially sequentially narrow the chromosomal region identified by genetic linkage. So that first step is essentially the same. But once you're into a narrowed region, you test markers, and in this case again usually SNPs, preferentially chosen in the coding sequences of the genes. So you're preferentially looking for coding variations, and focusing in on known variance, and known genes. The advantage of this approach is that you can more quickly get to the susceptibility variant if it is one of those known – in one of those known genes, and one of those known coding variance. The advantage of the purely location approach is if it turns out that it's not that simple that you've got a splice site change, or regulatory change, or it's in a novel gene that's only predicted, and isn't known, you can find it by this approach, but you're not as likely to find it by that approach. And then the fourth approach is a strictly candidate gene approach in which you focus in specifically on certain candidate genes, certain genes, because of what you know about the biological function of that gene, and the biological dysfunction of the disease that you're looking at, and trying to pull those two things together. So I'll talk about each one of those in a little bit more detail. The association approach, there is no prior hypothesis about a specific function of the gene. So you don't make a biological hypothesis. You can do this across the entire – across the entire genome, or you can in fact do this in a somewhat more focused region. It's an un-brought biased approach toward looking at the region of the genome. You can use multiplex, simplex, or case control data sets so basically any kind of data set that's out there can be used. The multiplex data sets would be families where you have multiple affected individuals in them. You sequentially localize the susceptibility allele. You identify initial association signals. That signal will span a very, very small region. And you will narrow the signal using additional association studies, and additional markers in that region to try to narrow that signal as much as possible. And you may be able to get that down to a very, very small region just a few kilobases of DNA, maybe even smaller than that. There's several advantages to this approach. One can discover association to previously unknown genes, predicted genes if it's a splice site variance something like that. You can pick that up. It's a very comprehensive initial

examination of the genome, more comprehensive than for example linkage analysis is. And therefore you more quickly, more rapidly focus in on certain regions. But there are some disadvantages to this approach. The massive number of SNPs needs to be tested across the entire genome. That's getting easier and easier to do with the new chips, and the new technologies that are out there, but it's still not an expensive experiment to do. Statistical methods for the analysis of these data are really still in their infancy. They are very, very different. There are lots of different ways of trying to do this. Nobody really has a great way of doing it yet. So there's no paradigm for the analysis of these data. So you know, we somewhat have an embarrassment of riches in which we are trying to figure out what's going on. And ultimately there tend to be a huge number of false positive results in whatever analysis you do do. So it's sorting through the false positive results is a bit of a problem. All that being said of course, it was very successful, and the paper from Josephine Ho in Science demonstrated that this can work with the identification of the CFH variant. So the locational approach has – the genome screen locational approach has the same advantage that there's no prior hypothesis about the specific function of a gene or a pathway. It's also unbiased. It can however, only use the multiplex families. You have to have multiple affected's in a family effectively to be able to use this approach. You sequentially narrow the region of interest using additional linkage studies if you have the ability to do so, or then you can move into sequential association studies looking at only small regions of the genome. The advantage to this is that you can discover association again to previously unknown genes, or genes of unknown or unexpected function. The disadvantage is you have to screen the entire genome. It requires using multiplex families, which can be pretty difficult to collect. It may take a long time to do. And the signals ultimately from the linkage may not be that strong although the signals once you've focused in and look at the association may be very, very strong. And the paper that we published uses this approach to identify the CFH variant. A variant of that approach, the genome screen functional approach uses – starts off in the same way, an initially unbiased approach towards the entire genome. There's not a specific hypothesis about gene function. There is however, an additional assumption that the susceptibility allele that you're looking for is a coding variant, and is something that's in the coding region of one of the genes. You can use only multiplex families at the first stage, but then you can use case control or family based association methods later. Again you sequentially narrow the region of interest by using the linkage studies or sequential association studies. The advantage is if the coding assumption is correct. If that assumption that you made is correct, you're going to more quickly get to where you want to go. More quickly identify the variant. And even if that assumption isn't right, the variance that you're using work just like any other, any other anonymous variant, and may still serve as proxies for what you're really looking for. So you haven't really lost anything by using this approach. This advantage of course, is you have to screen the entire genome. You have to use the multiplex families to begin with. The signals may not be that strong, and the coding markers may not be evenly spaced. So that if you're – if - they may not be served as perfect, a perfect set of proxies for the region you're looking at. All that being said of course, it worked as well, and now _____ published that as well in identification of the CFH variant. And the final approach that one can use is the strictly candidate gene approach in which you test a specific biological hypothesis based, and you're looking at the genes based on that biological hypothesis and what you know

about the disease and the gene. This can be expanded to include not just the single gene, but a set of genes, a pathway of, a common pathway, genes that interact or known to interact with each other. You can use multiple family types, which is good. The advantage is that you have a specific hypothesis, the statistics actually are, can be fairly straightforward. You have more comprehensive analysis. The ability to do a more comprehensive analysis of the region, the gene, or genes that you're looking at. And if you're right, you may lead very directly, and very quickly to the correct answer. This advantage of this approach is that the biology of most genes is really, is pretty much unknown or poorly understood. So you're basing your information perhaps on a pretty sparse set of information, your decisions on a sparse set of information. And even genes with known function may have as yet unknown pleiotropic effects. That means that the gene may be known to affect cardiovascular function, but it may have a totally different effect on auditory function, and no one's just looked at that yet so they don't know that. So it's possible the genes that have a known function that may not be the complete repertoire of what that gene's responsible for. And there are several examples of this. The CFH again was identified looking specifically at the complement factors, the recent publication of the BF, the other complement factors, VFC2. And a gene that we looked at last year, the _____ have all been looked at identified effects in this way.

So I think I will end there with just this last slide, and I think this sort of sums up where we are in terms of looking at complex diseases. We've got a lot of data out there. We have the human genome sequence at least most of it. And we need to figure out now how do we put that puzzle together, and how do we make sense of all the information that we've got. So I'll stop there and take any questions that you might have. Thank you. It must be very clear.

Q: Could I ask a question? Could you comment on to what extent you can localize a fine map using association down to the individual SNP or nucleotide given the - in other words, is the strongest association necessarily imply causality, because you have to have matching or causal and marker, allele frequencies, you've got LD patterns across the regions. To what extent do you think you can really fine map the causal SNP?

A: Well I think that's a very good question, and the answer that - there's sort of two questions in there. And one is I don't think that any of the methodologies that we have will really be able to get us down strictly through this methodology to a single causal or susceptibility SNP. Get down to the single base pair, or maybe couple base pairs that we're really interested in. And partly because of the other aspects that you talked about, there's linkage, there's equilibrium, there are these patterns of a lot of different SNPs, a lot of different variance in a small region being highly correlated to each other. So you pick up a signal and it's the signal's going to be coming out of an entire region. In the case of CFH, there's a two hundred and fifty kilobase region that has high - a lot of linkage to this equilibrium, a lot of high correlation across it. So you have to end up doing something else in that case. Now you may be - depending on the samples size you have and where, what kind of population you pulled that from, you may have a better or a lesser ability to narrow it, but you're not going to be able to get down to a single, a single variant that way. Yeah?

Q: My question is how will the advancing field of genomic research, first the sequencing of the genome, and now the HapMap Project influence the choice of method for finding candidate – for finding genes for diseases?

A: Well it has tremendously influenced the study designs. We are – all of – most of these study designs that I just outlined would not have been possible even five years ago, because of the technologies have advanced so much. Having the genome sequence out there just makes us able to design experiments and to do studies, and set them up in hours rather than having to go and discover all that information, and it could have taken weeks or months or perhaps even years to get to the same point. We can do that almost instantaneously now by looking at the databases of information that's out there. And as the HapMap data becomes more complete, and more available it will also allow us to perhaps be a bit more judicious in what SNPs, what variance random anonymous variance we choose to search through the genome. So yeah, it's making a huge difference now and as the new technologies come up, it will continue to make a huge difference.

Okay thank you.