

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

Anand Swaroop: Thanks to Emily for giving this opportunity to talk here. When she asked me to talk about the genomic perspective of complement factor H association to MD, I just, sort of, didn't know what it really meant, and it was a particularly difficult task, so given that I was going to come after all these excellent speakers who have already talked to you about the work that has already been either presented, and there have been some new information, so what I thought I will do is, I'll give you a very brief overview of association and things like that, and then move on to show you some of our recent data that we have had for some time, but sort of, the analysis was a lot more difficult than what we had sort of imagined.

All the work that I'm going to talk to you about was done, all the statistical analysis was done by Goncalo Abecasis who has been a major part of HapMap Project as well. So when I was, sort of, thinking about genomic perspective, I could think of genomic perspective at the level of DNA and what that meant to me was detailed investigations of genetic variants that might be involved in disease susceptibility, and, again, I'm keeping all that, sort of, in respect to macular degeneration. Then the second one I could think of was at the level of RNA, that meant, sort of, like expression analysis, global changes that might be happening and, sort of, in response to these different variations. And the third could be the level proteins, and when you think of proteins means no protein exists in the cell alone, and what you see at the functional level is really mediated by different macular molecular complexes. Our knowledge of RNA and protein perspectives at this stage is rather premature and, so, I thought I would rather focus on DNA.

So when you think of DNA, in public databases, there are over eight million genetic variants and mostly SNPs. So what are SNPs? SNPs are really single nucleotide polymorphisms. That means that that particular base pair, _____ nucleotide, say A, you could now have a different nucleotide, say G or C, or whatever. So that's the variation that you can see at the level of DNA. And almost eight to ten million of these exist, sort of, in the human genome, at least that have been identified so far. So our challenge is to find those that directly influence AMD, and that has been very eloquently done earlier, as you heard by Josephine Hoh.

But, again, going a little bit further, so what this genomic approach also means is that you may want to do a very comprehensive survey of genetic variation. When you're looking at susceptible disease, you may want to look at all genetic variations that might be associated at the whole genome level and, again, several speakers earlier already talked about it; you could do a whole genome linkage studies, or association studies, or you can focus on targeted region. And when you're focusing on a specific region, you may need to do only a few hundreds of variants. And this was just a sort of summary that I put in the form of a cartoon of a lot of these linkage scans, whole genome linkage scans, that have been done, and so when you put them together, there was something like sixteen different chromosomes where one can identify these susceptible loci. The two chromosomal regions that came again and again in various studies were chromosome 1Q and also on chromosome 10, and both of these seem to have

been identified at this stage. And so a lot has been talked about already about these variants, along with these four that we also published at the same time in May, this association working with Josephine in a large-scale study, as well.

So when we begin to look at the CFH region, we realize that there's not only this complement factor H there, there are a lot of other related genes within that region. And when you have a situation like that, it's, sort of, very important to know whether a particular variant that you have identified really a causal or there could be other things there happening as well.

So let me just give you another very brief background of genetic association. You've already heard about it, but for the benefit of those who don't do these things often, so basically what we're doing in association is you search for short stretches of chromosome that are shared between distantly related individuals having similar phenotype, and then what you do is compare the frequency of these different variants in these regions, and in our case, the AMD, we can use cases and controls and, basically, it's very critical that the phenotype is very, very well defined because, you know, if you have sort of mixed phenotypes, sort of, your controls are really not good controls you could get into a lot of problems in the genetic analysis, and that has been emphasized earlier, as well. So the beauty of these association studies, compared to the linkage analysis, is that you can identify even simple changes of variations that have relatively very small contribution to disease risk, because you're right in the region, but the problem with that is that you really need to be right in the region of the disease locus rather than – you can't be too far away, just in simple terms to explain that.

So I mentioned to you earlier that there are eight to ten million of these SNPs, or single nuclear-type polymorphisms, that exist in our genome, but looking at all those polymorphisms is a rather daunting task, but it has been, sort of, recognized that in our genome, there is something called "linkage disequilibrium." What that means is that, basically, that our modern chromosomes that we have are really mosaics of these ancient chromosomes; what that means is that if we had the same ancestor and you had this particular kind of, sort of, stretch of DNA, now in the present day, you have all kinds of mosaics that have been created by the recombination events that have happened, but still, depending on how recent the population is, you can have a small or large, sort of, regions of the DNA; like, in this case, if you are putting together all these different chromosome, there's a small region of this size here, which is sort of common in all of these chromosomes. So these are called, sort of, LD blocks, if you want to call it, and all the markers in that will behave in a very similar manner. So if you recognize that, then what one can do is we can use – we don't need to have eight or ten million SNPs to do all the mapping studies, or analyses, genome switching analyses, we can just use maybe about five hundred thousand SNPs and it can allow us to classify almost all mosaic fragments, okay. And that is true, really, for Caucasians and Asian populations, but for people of African descent, you need a lot more SNPs. So if e-chips, which are 500K right now, and _____ 300K chips, each of these cover about 80 percent of these SNPs, so they are really highly, highly useful; and this is what I was trying to say, that in the early analysis, you really don't need to look at all of these genetic variants.

So what we did was, we followed up on what Greg Hageman was talking about in their paper, and Michael Dean alluded to, what we did was, we thought we will

take it a step further, so we thought we'll do a very detailed analysis of the CFH region. So, in about a 125 KB region, we could identify about 200+ SNPs in the database, and we tried to optimize all of them, but 84 of these we picked up at the end, and the remaining either did not follow Hardy-Weinberg tests, assays or had very low frequency, or the PCR's were not very successful, or they were other problems; so we dropped that. But we had 84 SNPs that were assayed in all of our population. And this is particularly critical because – one of the reasons that we didn't get many of these SNPs, sort of, successfully assayed by Hardy-Weinberg, also, at least that there were quite a few of these, and we think that could be because of these homologous sequences that might be there, and so our primers may not be, sort of, giving the right kind of answers. So this is just a slide to show you that in this region that we tested, each of these triangles represents a SNP, basically, in the region and we tested all of these SNPs, and this is the frequency of that, and what this shows you is the degree of LD base with these red regions show you a bit with these SNPs.

So this is the data. This is really the most important, sort of, figure that I have in my presentation, at least; so I'll have to explain that to you a little bit more. Goncalo had to explain that to me, so if I don't explain that to you clearly, you can blame Goncalo. Anyway, so basically – so this is the single SNP association analysis, so all these 84 SNPs were analyzed throughout this region, and you can see there are a few different – these dots are there. Each dot represents a single variant. There are a few different colors there, you can see; there are these green dots – I'm colored blind, so he told me it's green, so I'm telling you it's green, but I think it's green. Anyway, so there are these green dots there and then you have these purple dots, and you have these black dots, and you have these square kind of things, which are red. So all the dots in the same color, particularly the green ones and the purple ones, are in the same LD group, linkage is being blocked; they represent one group, basically. And this is another group, these purple dots are another group. The line that you see – Oh, on this side, what we have plotted are the log values of the – the log of P-values, basically. And this line indicates that number for this famous Y402H SNP. So, as you can see here there's a fantastic haplotype here with the purple, which includes the Y402H also – not a haplotype; I'm talking about single association, here, so just sort of bear with me – so this is that, the large separator markers there, and then, but you have these green ones which are in a different LD group, but they even show higher association, so when you talk of chi-square values, based on the single association SNP analysis basically, when Y402H in our study shows about 110 chi-square value, some of these go as high as 160, so you can see that there are SNPs there which are even more strongly associated with macular degeneration. And these SNPs that I'm talking about, these are not one or two, there are about 20 SNPs that show even stronger association compared to Y402H; and these 20 SNPs, 18 of these are in the non-coding region.

So what we did – so a lot of analysis that has gone into that and all that is being put together by Goncalo – I think he was supposed to – I've been traveling so Goncalo was supposed to send the paper yesterday or today; hopefully he has done that. But anyway, the bottom line is that he has been trying to do a lot of these analyses and I'm just going to show you the next stage analysis, which is – so what he did was, he selected five SNPs to build these haplotypes, okay. And he did the association analysis of these. So, I don't know if it's clear. Maybe I

should have used different colors. I was just using the bright colors that I can see, so pardon me if you can't see, because I can't see from here, anyway. So what I can show you here, which is, I think, very, very clear, there are five different markers that were used, and these are different haplotypes on this side, and look at these numbers, basically, which tells you the frequency of these. None of these markers were Y402H, so we put a conditional _____ whether any of these haplotypes have Y402H. And now, the major risk haplotype, like Hageman et al, had identified, we also found a different haplotype that we made here, with the ninety-four percent chance that it includes Y402H; okay, very high chi-square for that. But in addition to that, we have multiple other haplotypes; these are rare haplotypes, and this is a relatively common haplotype, seven or so percent; but these never has Y402H there. And then, there are two protective, instead of one. We have found two protective haplotypes, with very, very high chi-square value.

So a detailed examination of this region, what it has told us is, something really which we were not expecting. And what we found was that there were a number of these non-coding SNPs that showed very, very high association, a very strong association, compared to the coding SNPs, actually. And, basically, we found four common haplotypes, two associated and two protective, and multiple rare haplotypes.

But something else, which was very interesting, was that no single SNP, including Y402H, would explain the observed effect and, basically, none of the SNPs could, sort of, explain fully the effect of _____ SNPs that we had. And this is in contrast to what has been reported in some other cases, like ApoE and all where you can see association can be explained by just a simple few alleles here.

But I have to say still one more thing, and that is, that Y402H still is probably the most important of all of the allele system, is part of the most common haplotype, as well; so it is very important. But you cannot, sort of, ignore the significance of other variants, and that I think we need to look at. So what we think is – and Greg has pointed out in his talk at that time, actually, that in addition to the structure of CFH, what I think will be important is the expression of CFH, as well. And maybe some of these non-coding SNPs are really, in some way, affecting the expression levels of either CFH or the neighboring genes, because, you know, you have this whole cluster of genes. So I'm going to stop there. The graduate student who was working with Goncalo did all the analyses _____, and in my lab various people did different aspects of the work, and Dr. Ken Lively at Broad Institute, people who closely collaborated with us and did a lot of the genotyping was the group of Stacey Gabrielle. Questions?

Q: You mentioned the 18 out of the 20 in that second linkage disequilibrium block were non-coding. What about the two that were coding? Were they conserved changes or conservative, or what?

Anand Swoop: One of those was a highly conserved change, and I don't remember exactly – one was highly conserved and one did not change the amino acid action. So, basically, I would say, that the variants that we have found are highly associated really should not affect the structure directly.

Q: Anand, as you probably know, we also found, and we published that in the *Science* paper, SNPs that had a higher P-value or higher chi-square score

associated with AMD, than the 402H SNP, and when we looked at the haplotypes – the haplotypes are estimated haplotypes, because they're not family data, they're estimated by an algorithm called the EM algorithm and others. But when we looked at those estimated haplotypes, we found that all of the risk haplotypes contain the Y402H SNP. That is to say, that there were no – although there were several haplotypes that were associated with AMD – there were no risk haplotypes that did not contain that specific variant. I wondered if you had risk haplotypes that did not contain the histidine allele?

Anand Swoop: Yea, this is what I was showing. That's what the slide was, that we had haplotypes. We have at least one risk haplotype, which is fairly common. It's not as common as the haplotype which includes Y402H, but the other, number two in that list, which did not have Y402H at all, and then there were rare haplotypes and when I say rare, we discarded the variants which were less than .05 or whatever the criteria is being used; but there were rare haplotypes as well, which did not include Y402H. Now that means the H, histidine allele. So we have several haplotypes that do not, but they're not as common as the one risk haplotype that includes Y402H.

Q: Do you recall how common they are in aggregate?

Anand Swoop: I thought the numbers were there. I think – we can go back or we can talk later.

Q: We'll talk later.

Anand Swoop: Okay.