

Christophe Lambert: Well, good morning, everybody. I'd like to welcome you all to our webinar. Should say good afternoon, good evening in some parts of the world. My name is Christophe Lambert, as opposed to Christoph Lange, our presenter. It's always a pleasure to be confused with Christoph Lange because of all the wonderful things he's done. I'd like to first go over some webinar logistics, and should see on your screen a little screenshot of the GoToWebinar pane.

The webinar is listen-only from your end, but if you have any questions, you can type them in the question and answer pane, and we will have time at the end of the presentation to answer your questions. So, also, if you have logistic issues or problems with the webinar, there'll be people who can answer them more quickly, but we'll get to the scientific questions at the end of the webinar. So, I'd like to turn then to introducing our speaker. Dr. Christoph Lange is the associate professor of biostatistics at the Harvard School of Public Health. He's also affiliated with Harvard Medical School and Brigham and Women's Hospital.

Dr. Lange has worked with us over a number of years to take his popular PBAT package and make it user-friendly and suitable for genome-wide analysis on very large data sets, as well as more recently he's enabled copy number genome-wide association studies with his PBAT package. Today's talk is focused on a genome-wide SNP scan of Alzheimer's focusing both on methods as well as results. What's particularly interesting for me about this talk is how he's really maximized the extraction of information of the phenotype, both the affectation status for this family-based study as well as looking at the longitudinal course of the Alzheimer's disease and being able to get more power out of his analysis. So, I think the methods that Christoph's gonna show today are – can be applicable to many other studies, and we're very excited to hear how you've gone about this Alzheimer's study and how we might apply it to our own research. So, with that, I'd like to welcome Christoph Lange, and we look forward to hearing your presentation.

Christoph Lange: Thank you very much, Christophe, and I very much appreciate the opportunity here today to present our research on how to design and how to analyze genome-wide association studies for late-onset diseases. And as a moderating example, we will use a 500K scan for Alzheimer's disease that we very recently published in *The American Journal of Human Genetics*. And of course, this was the work of many, many authors. The first author of the paper was Lars Bertram, who led initiative together with the senior author,

Rudy Tanzi, and then also with Deborah Blacker and David Becker. I should say that the entire project wouldn't have been possible without the funding of the Cure Alzheimer's Fund, for which we are very thankful.

I'm going to go over my talk now, the overview. I'm going to start a little bit with the genome-wide study for Alzheimer's disease and the particular challenge, statistical challenges that we faced there for late-onset disease in family designs and how – which component we can take to maximize the statistical power then in the analysis phase. I'm going to talk a little bit about two analysis decisions that we made. One was, first of all, the test statistic, how we were able to combine all available phenotypes into one overall phenotype here for Alzheimer's. It was affection status and time to onset.

And then, the second analysis decision, the testing strategy that we applied. Here we used the algorithm that's implemented in the PBAT package, which was called using the same dataset for screening and replication and that was published in VanSteen et al and *Nature Genetics* in 2005 and recently modified by Ionita-Laza at *The American Journal* to maximize the power further. And then I'm going to talk a little bit about the results of the 500K analysis, but I have to warn you, I'm a statistician on that, so I don't know – I don't understand much about the loci that we identify. But, anyway, I'm going to focus now on the Alzheimer's study, so as for many other diseases, before the Alzheimer's scan was done, there were only a few – a very small number of genetic loci that were consistently associated with Alzheimer's disease, over three loci for early onset and, of course, the famous APOE loci for late onset. And otherwise, there have been many reports about associations, but none of these associations has really reached the level of replication of the APOE locus.

And the findings have mostly been inconsistent, and a very nice overview of these previous association results can be found at alzgene.org. That is a website that is maintained by Lars Bertram and Rudy Tanzi. And so, the goal was to conduct genome-wide association study on a population of Alzheimer's families, we had a total 1,376 subjects from about 400 families, and the goal was really to identify new DSRs and then replicate them in other family studies.

And because at that time when we finished our replications in the family studies, also replication in publicly available case control scans. The initial cohort that we used for the discovery phase was

the NIMH Genetics Initiative study sample. It's a sample size, the number of families that I described.

It's the largest uniformly ascertained and evaluated Alzheimer's family data, the AD diagnosis of definite, probable or possible AD made through the standard criterias. The ascertainment was two affected siblings per family. All of them had to have an age of onset of at least 50, and the age of onset was available for these programs was an important part of our analysis strategy later on. Generally, we had model affected per family, but also unaffecteds, and parental genotypes for most of these families were missing.

So, the phenotyping, we had 400 and 941 affecteds and around 404 unaffected. Individuals and in terms of the diagnosis, you see the – between definite and probable and possible. You see this split up here, and when we had no AD and no unaffected status, we considered this phenotype as unknown in the analysis. And so, genotyping platform we had available, the Affymetrix 500K scan chip. So, that leaves us with how to analyze, and we had a very limited number of programs, around 1,400.

It was a family dataset, which are not as highly powered usually as a standard case control analysis, and so we tried to minimize the disadvantages of this design in terms of statistical power by good or smart choices in the statistical analysis to take full advantage of this – of the unique features of the family dataset that are here at play.

So, the first decision was really what test statistic to take. The standard, the golden standard at the moment is affection status, and there was, of course, what we sought initially as well, but then another important phenotype that Alzheimer's, but also many other late-onset diseases, are characterized by is age at onset. And so, our idea was to combine both affection status and age of onset as our screening test statistic for association. What – why – the reason for that was essentially one could assume when you have an early onset, you would assume you have more genetic reasons, more genetic causes for that, and the late onset is probably more due to environmental reasons.

The examples for that are, for example, childhood asthma. We have a very early onset at the age range between two to three or something like that. You could probably have a hypothesis that is more genetically determined why late onset can be more to environmental reasons. Here for Alzheimer's, let's draw a similar – if you assume we have an onset say, for example, in the late 60s

or mid-70s, then could have more stronger genetic component than an Alzheimer's onset of in the mid-80s. So, in one way, time to onset or age at onset of the disease is a more refined definition of affection status.

It, by definition, it captures more information, and how to define a test statistic based on that, I have to introduce a little bit of notation. I assume here I have N trios for the late application. This will simply reflect in and is the number of families that we have. The methodologies that are outlined here in the slides is for trios, but the implementation in PBAT and general FBAT approach, as I discuss it here, straightforwardly applies to families with no parental genotypes. Then, x_i denotes a marker score of the offspring in the i th family.

P_{i1} and P_{i2} denote the parental genotypes. T_i is the age at onset or the event time, or for the unaffected problems, the censoring time when we observed them the last time and they were still healthy. C_{ij} denotes the censoring variable, 1 for people who got _____ the event and got affected and 0 for things that _____ study participant. And we assume for the notation that we have K distinct event times. Then, to construct an association test for age of onset, what we can do is we can simply follow a methodology from the clinical trial literature, and there we have treatment groups.

We compare different types of treatments, and this translates straightforwardly to the genetic context where we simply will place as treatment group as shown here by the genotype X . And for each genotype, we compare how many observed number of events we have minus the expected number of events under the null hypothesis, and this will allow us to construct a test statistic for age at onset. The modification of this logrank approach is then to put higher weight to early onset if we believe there are stronger genetic component as we did in our study. And the way this can be achieved is simply here by adding this weight, N_k . N_k denotes the number of offspring at risk at the time point, and that means because at earlier time points we will have more offspring at risk.

We will here put larger weight to this difference in the Wilcoxon test statistic. Then both for the logrank and the Wilcoxon test statistic can then be translated to the standard FBAT notation, which we can see on this slide here, where see here X_i is the observed marker score minus expected marker score conditional on the parental genotypes, or if they are missing, conditional on the sufficient statistic. So, this mendelian residual multiply with T_i ,

which is a coding of the trait, and the coding for the trait, if we want to use the logrank approach, is simply for the i th offspring we have the censoring whereabout c_i minus lambda at T_i . And lambda had T_i . It's a rough estimator for the proportion who has that model. If we wanna use the Wilcoxon aspect, it's a very similar story.

Then we have as a coding here for the phenotype T_i , we simply multiply n_i , the number of people at risk times the censoring variable minus the sum over all events that happened earlier. And we can use either of these codings and then derive standard FBAT test statistic using the notation above here, which has then chi-squared test statistic was 1 degree of freedom.

Just to show you some initial application and how we can benefit from that, this is an application to the Alzheimer's set by Deborah Blacker which was published in '98. We assume here again that we have 143 nuclear families with 2 to 10 siblings, at least 1 sibling with Alzheimer's; parental genotypes are unknown. And what you see here is the distribution of age at onset for the three different genotypes that are defined by the four allele APOE locus.

And here below we have the age range, and here we have the probability of being disease-free. So, what we see here then, for example, at the age of 70, or the genotype 2, the probability of being disease-free is 40 percent. While for the other two genotypes, $x = 1$ and $x = 0$, the probability of being disease-free is about 80 percent, so what we see here very clearly, the huge genetic effect of the APOE locus and all longitudinal or the age range of all study participants. And because this is much more precise information about affection status, we hope by incorporating this in our screening test statistic that our overall analysis approach will be more powerful. This is also reflected when we look at the FBAT test results, when we take the Wilcoxon coding for a time to onset or the logrank coding here or that we take the quantitative coding for time at onset – of age at onset.

When we take the standard age of onset, or the FBAT test statistic, we get a P value of 007, which is not – which is significant, but it does not really reflect these huge differences in the survival curve that we see on the previous slide. However, when we take the time to onset codings, we see here really for the FBAT logrank and Wilcoxon, the P values are in a completely different dimension with minus 8 for the logrank, but even smaller, much smaller for the Wilcoxon where we assign much higher weight to the early-

onset families. So, the idea is then to combine both time, age at onset and affection status into one screening test statistic. But, we could, of course, only use age of onset because we, in order to construct this FBAT logrank or Wilcoxon, we have to use the information on age at onset and affection status. But, of course, it's clear by definition if we condense two variables into one variable we will lose information, so the idea is here construct a multivariate test for the first phenotype is really affection status, and the second phenotype is time or age at onset.

And we will do that by the FBAT-GE approach. And the FBAT-GE approach is an idea that is based on the Generalized Estimating Equation concept that was introduced by Liang and Zeger in '86, and Prentice and Zhao then followed up in '91. And by constructing estimating equations for our model, we can derive then again a multivariate FBAT-GE test statistic that has the very simple, very plain form that is given here. Again, we have mendelian residuals times T_i , and T_i is not a univariate genotype here, but it's a coded vector of phenotypes here, and this, in a similar way, we construct the variants here. And then, by simply taking the product of the vector and the matrices, we get R value for the test statistic and then overall test statistic as in chi-squared of it's M degrees of freedom where M are the numbers of phenotypes that we want to test.

So, in our 500K Alzheimer's scan, we were using the – whoops, sorry – we were using the thrust dimension, otherwise a thrust phenotype PS affection status, and the second one is a coded age at onset as discussed on the previous slide. So, this was – were the considerations we put into place in order to select our test statistic. In terms of now the analysis strategy in terms of the testing strategy, we decided to go with what is implemented or what is the so-called PBAT screening using the same dataset for screening and replication as described in the VanSteen paper. We later on improved the power of the approach quite a bit by using a weighted Bonferroni approach, and this was published by Ionita-Laza in 2007. And I'm going – we used this approach in the actual analysis, and this is what I will discuss here then a little bit as well.

So, if you look at a genome-wide association study, this is a statistic that – the view that a statistician would probably have on it. It's essentially we have many, many marker loci here, the human genome, and we want to test hundreds of thousands of loci. And the idea of the PBAT screening algorithm is essentially rather than facing the matter with testing problem that is posed by these hundreds of thousands of tests, we want to come up with a

screening statistic that allows us to select a very small number of these SNPs and then would test only then this small number here in a second stage where we have to adjust for multiple comparison. And the idea is essentially to come up with a test – with a screening statistic that allows us to look at all 500,000 loci and then select, based on that value, a very small subset of test statistics and compute only for those the association test statistic. Here in this case, instead of looking at 500,000, or it's probably 50 SNPs, we look here at the test statistic and maybe it was 4 loci.

And this approach sounds kind of too good to be true, and it has only one major constraint in order to work. The screening statistic has to be statistically independent of the testing statistic, and by that is how we can construct this for family-based association studies. This is something I'm going to talk about on the next slide. If then we look at the joint likelihood of the family-based study, we have the phenotypes and the genotypes, and as described in Laird and Lange in 2006, we can derive or split up the slide – this joint likelihood in the following way where we look at the phenotype, the genotype and then condition on a function of the screening statistic that is also a functional piece, quantities times here the phenotype and the genotype as a function of S . And then, by default, these two parts become independent.

It's a simply-based rule, and we can use this first part as a screening step here and then later on as a testing step, and by definition here, these two steps are independent. The only trick will be to construct a screening statistic S here so that we do not extract all the information about the association from the data and leave information for the association test here, so that's how information is maintained. So, how can we do that? So, in the first step here, how does it work? It's piece by piece, and the screening statistic here is applied to all 500,000 SNPs, and based on that small subset of, say, 10 or 200 markers are selected, and then the testing step is then – we compute the actual association test.

This only apply to these 10 or 200 markers, and instead of adjusting for 500,000 comparisons, we then just have to adjust for 10 or 200, which means a substantial reduction of the multiple testing problem. And, of course, both steps have to be statistical independent for that. So, in the family-based side, this partitioning of the likelihood into a screening statistic and into a testing statistic happens in a very natural way. We look at here at the joint likelihood. What we can do is we simply condition here on the parental genotypes and the phenotype as well, and what we get

then is the sufficient statistic defined by the genotype and parental – the phenotypes and the parental genotypes here.

And this is the so-called conditional mean model that we introduced in 2003, or if you are like the notation of the between or within component of the likelihood approach, here we have the between-family component that's what S reflects. And then, here in the second part of the likelihood, we have essentially the information, the genotype here conditional on the phenotype and the parental genotype. That is the distribution that is used for the FBAT test statistic. S would use S introduced by Speerman et al for the classical TDT and later on by Laird et al in 2000 for the FBAT approach. For the properties of these testing strategies when we implemented and compared to the standard FBAT and adjusting from a multiple comparison in terms of statistical power, we outperformed the approach by many magnitudes.

And we have been able to apply it now in several genome-wide association studies successfully, including the association studies that I'm going to present today. So, just to visualize this approach a little bit more – in more detail, here we assume we have trait and simplification of genome-wide association studies with six marker loci and the – for – in the first step we do the statistical trick. In the conditional mean model that [inaudible] that's the observed marker scores here on all six SNPs are missing, and we impute them based on the parental information. So, we have a place to observe marker score that's the expected marker score, and then use a trait value to estimate for each combination a genetic effect size, and based on this genetic effect size, we estimate the conditional power. And so, for of the FBAT test statistics, so we're gonna – for the actual study, we are able to estimate in a predictive way the statistical power of the FBAT test statistic.

We do this for all six SNPs here, and then, in this case, we select the SNP with the highest statistical power based on an actual effect size estimate that is locus specific and specific to the area frequency at this locus. Try and push only this locus and forward to the testing step in which we assume that we have found the data again. The real data is available, and then we compute the FBAT test statistic only for this SNP, and the beauty of this approach is here if we just selected one SNP, we don't need to adjust the model of comparison at all, and this P value is significant by definition.

Another nice feature that we hear in the first step to a population-based analysis, which has all the pros of a population-based analysis, and here at the second level, we do this family-based

analysis, which has all the pros of the family-based approach. But, by combining it, we really minimize the caveats of both the population analysis of a pure population-based analysis and a pure family-based analysis.

But, here we have this robustness against population of mixture and stratification, and here in the first step, because it's a family – it's a population-based analysis, we have really this robustness against genotyping error. If you have genotyping error here, it will only reduce our power, so it's unlikely to be pushed forward and get assigned high power values. But, in this step here, we do the proper family-based test. Of course, one criticism of this approach could be that we select only a very, very small number of SNPs that are actually tested for association. And we have been able to modify this approach and maximize the power further by allowing really all SNPs to be tested by using a weighted Bonferroni approach that was introduced by Ionita-Laza et al.

And the idea here is simply that we take the power estimates for each SNP, and rather than selecting only a handful of SNPs, we test all the SNPs, but the significant level that we assign to each SNP is a function of the power estimate. And here, for example, in this case, you have assigned to the two highest power estimates a significant level of 2 percent, and for the remaining four SNPs, we assign a significant level of just a quarter percent. And how you do this writing is a little complicated, and we discussed it in the theoretical paper in detail how you can maximize this writing scheme. But, essentially, it allows you to test all six or five markers in the testing step and don't have to do this selection of just a few and for a number of SNPs. Let me apply this to the genome-wide association study for Alzheimer's disease.

Of course, the first step was the cleaning step, and I'm not going to bother you with all the data cleaning. I'm just going to show you here the Q plot as shown in the paper, and what you can see here that we are maintaining about the diagonal line after we did the data cleaning. This is for the FBAT-GE test statistic using affection status and time to onset here – oops, sorry – and then also, we restricted it to all – to families had to have at least 20 informative that – only families who allowed to contribute to the test statistic was more than 20 informative families. The results that we found are shown on this slide. Four SNPs with significance here using the weighted Bonferroni approach, by Ionita-Laza et al using the FBAT-GE approach.

Of course, the F – APOE locus reached significance as well, but it

was eradicated, so I'm showing you only here the new loci S in the paper. And here you see always the FBAT-GE, then affection status of P value here, and then also for age at onset here. And what we did then was Rudy Tanzi had in his lab three additional Alzheimer's studies, family-based Alzheimer's studies, so these four SNPs were genotyped in these three additional replication studies. And based on the direction of the signal, we computed one-tier P values for these replication sets, and here you see in this column the replication P values and then the total P values. And what we observe is that for the – this – the first SNP and the third SNP give good replication and replicate nicely across the three studies when we combine the replication P values, but two and four is not that great.

Then, we compared it, our findings, to previously published case-control findings of genome-wide association studies that were publicly available on the Web. And so, there's a TGEN sample that has in total here 1,300, 1,340 probands and all through the sample by GSK there's a similar sample size. And essentially, for the first SNP and for SNP No. 4, we see replication. For the other SNPs, it is not that good, and for the third SNP here, that was actually missing in both studies. That SNP was clean, but in our study, the genotyping quality seemed to be reasonable.

So, just as a conclusion, we identified four new loci that looked interesting for Alzheimer's disease, and we were partly able to do that by analysis strategy that took good advantage of all the phenotypic data and tried to minimize the model or testing problem, the family context, the impact of the model testing problem as much as possible. And this – the type of analysis that we did here we want to include age of onset together with affection status, or another phenotype is straightforward, and the other phenotype could be incorporated in this FBAT-GE test statistic. We could use sliding windows-haplotypes, etc. And, but, one conclusion that remains no matter how well you design the statistical analysis replication at the end as the gold standard, and you can't do enough.

That was pretty much everything I had. Now, we'd be more than happy to take questions.

Christophe Lambert: Thank you very much, Christoph, and if anyone has questions, if you would type them in the question pane, I'll read them out loud to Christoph, and we'll answer them. If you have some follow-up, you can also type it in the question pane. So, while we wait for people to type their questions, Christoph, what's your thoughts on

this whole issue of how so many studies don't replicate? Are we seeing, you know, just sort of fluke occurrences, or you know, what's the right way to think about replication in genome-wide studies?

Christoph Lange: I think that's the \$100.00 question. It's a moment – I think it's – it could be a false positive. Of course, another issue, it could be that there's confounding in the replication study in terms of that the discovery dataset and the dataset that you try to replicate it in is really – was designed different slightly – the affection status definition is slightly different. The study was collected at different time. I think in general I would summarize it, because we need large samples for both really for the discovery phase, but also for the replication phase, taking one study and splitting it up or something like that is not something that doesn't work really.

It wouldn't be true replication anyway, so you're trying to replicate findings in different studies, and this simply introduces heterogeneity, and I believe probably part of the problem is that we, at the moment, still can't handle this heterogeneity.

Christophe Lambert: Mm-hmm. As we wait for people to type their questions, I guess I'll ask another. So, I think I misspoke at the beginning when I characterized the study as longitudinal, but in a sense, it's sort of longitudinal with respect to the whole population, perhaps, when you look at time to onset. But, how – can these same approaches also be applied to longitudinal data where you look at the time course of a disease, progression, or for instance, you know, the – some measure of how the phenotype is changing over time, and what's been your experience on how that might add power to studies?

Christoph Lange: I mean, if you have really repeated measurements, you can use it in this FBAT-PC approach that we developed, and that is included in the PBAT software, and what it means essentially is as long as these phenotypes have positive environmental correlation, by any additional repeated measurement that you add, you add the same amount of – to your sample size. So, if you have, say, four repeated measurements and you have positive environmental correlation, you have four times the original sample size. And that was why, I think, we were partly successful when we applied the strategy to one of the first genome-wide association studies in the Framingham Heart Study where we just had a sample size that was below 1,000. But, looking at BMI, they are having six repeated time points that we could use. We got a much larger effective sample size in our testing strategy.

Christophe Lambert: Could other parameters besides P values help quantify the degree of association, such as relative risk or odds ratios?

Christoph Lange: I mean, if you want to characterize, I mean, these are the important quantities to go for, I absolutely agree, and this is why we do take – that's why we like, as a selection criteria, of prioritizing criteria conditional power, because the conditional power will depend on effect size that is specified as relative risk or odds ratio and on allele frequency. And these are the two factors that are important for – that – for – that make a finding important if both criterias are met, and also these other criterias that you need in order to replicate.

Christophe Lambert: Can I use these methods for migraine headache where the age of onset is not clear-cut?

Christoph Lange: I'm not sure what clear-cut means in this context. Is it –

Christophe Lambert: So, I guess, in that case, it's not clear when migraines first started for a person.

Christoph Lange: It – I mean, the approach, as any FBAT approach, is robust again in the specification of the phenotype, so you will lose power if that's a major issue. And if you have some age of onset that you can rely to some degree on, it's definitely worthwhile giving it a go, yes.

Christophe Lambert: Okay. Next question, how do you test or really define independence of the two steps, screening versus testing?

Christoph Lange: By another null hypothesis, whether they are statistically independent.

Christophe Lambert: And I think for some people who really question the independence, it – in addition to sort of proving it in your papers, you've done extensive simulation studies as well to demonstrate this, right?

Christoph Lange: Yeah. We made extensive simulation studies as well, but the core piece of it is just really the equation as it's shown here. Your – this is a typo here, so it's just a genotype of the offspring conditional on this guy times this probability here, what we use in the screening step. And by default, the motive here is this independence of the two steps.

Christophe Lambert: So, here's another question on the screening/testing. The partition of likelihood into two parts, screen and test, may not be independent if there are lurking environmental factors, isn't it?

Christoph Lange: It depends whether your – I'm not sure. I don't think – I mean, it depends what you do here on the conditioning here, but as long as you condition properly here on these two variables only and you only use these two variables here, you will always be fine.

Christophe Lambert: It is a concern maybe about confounding environmental factors.

Christoph Lange: I mean, the first step is always a population-based analysis, so it can be confounded as any population-based study as well. And then, that's essentially here, so – but, then by having then two family-based association tests in the second step, you don't have to worry about that. So, I would see it probably even as a strength of the approach. If you believe that you have confounding here that you can control for at a population-based level, it uses information to prioritize it, but here the final call, whether the signals for real are not the standards the end of the day by the FBAT test statistic. And unless you have reasons to believe that the FBAT test statistic is incorrect, you are all set. And what you essentially need for that is [inaudible]. That's the only concept, really.

Christophe Lambert: So, there's another question, someone just asking you to explain again why there – the screen statistic is independent of the testing statistic. Is kinda the intuition, Christoph, that basically you don't actually look at the child genotypes in the first stage? You only –

Christoph Lange: That is correct. I mean, this is what I try to visualize here in the first step here. When you take your, in quotes here, your genome-wide association study with six SNPs, and in the first step you pretend that they are missing, and then you just impute based on the parental genotypes and the missing genotypes in the offspring, and because you condition on the parental genotypes later on the FBAT test statistic, you maintain the independence of the two steps.

Christophe Lambert: So next question, what do you think about the value of doing GWAS using familial cases versus using sporadic cases?

Christoph Lange: I mean, if you believe in the sporadic case, I think there's not much point in doing a family genome-wide association studies. If you have cases with – that show strong familiarity, then the family-based assigned is more likely to be successful because you have

the evidence of genetic heridability in there already. Probably you can use the same study. I think that would be a key advantage.

Christophe Lambert: All right, next question. I knew someone would ask this on you about the biology, Christoph. How do these four new loci relate to other susceptibility genes, proteins and Alzheimer's disease?

Christoph Lange: I have to be frank. I don't know much about that. I don't understand much about it. They are sitting in predicted genes that do make sense to my collaborators. Why that is exactly the case, I don't know. I have to admit, I am just a statistician on this project.

Christophe Lambert: Just a significant signal to you.

Christoph Lange: Yeah, exactly.

Christophe Lambert: Next question, has independent validation been done other than conditioning on the selected SNPs from screening phase? Is this presentation published citation?

Christophe Lambert: I think they're basically asking – you mentioned earlier that the whole screening thing is published in *Nature Genetics*, right?

Christoph Lange: Yes, and, I mean, the replication, I mean, when you go back to the slides here, these are our primary findings, and here you see the independent studies, the three independent studies that were used for replication. And, of course, they are, by design, a completely independent – and also the case-controlled studies by TGEN and the GSK that we used.

Christophe Lambert: Next question, someone asked if these statistical methods – I think perhaps they're referring to either the screening step or the combination of phenotypes – I'm not sure which. Can these statistical methods be applied to non-family-based genome-wide association studies?

Christoph Lange: We have something in the making that's pending that we can do a very similar approach for case control studies. We are still working on the software and the methodology paper. It's still in the making for that, but we hope to get that out relatively soon.

Christophe Lambert: That's great. Someone asked, can we split extended pedigrees into many nuclear families?

Christoph Lange: We've tried that extensively, and so when you go into the PBAT package, you have those options. Actually, you can split it up into

nuclear families, or you can analyze it [inaudible], and in particular, if you have these extended pedigrees and you have quite a bit of missing genotypes, they tend to take computation device quite a long time. And when we compared the power gain of analyzing the extended pedigree versus splitting them up and treating them as nuclear families, the power gain was relatively small. So, my recommendation these days is technically you get more power out of it by analyzing the entire pedigree as a whole. Computation-wise, it might be a huge burden, and power-wise, you don't get that – gain that much usually anyway.

So, we typically take the split-up or what we have now a little bit in the program, this new option that does a hybrid thing. If you have these small in quotes extended pedigrees in – if you find in the pedigrees – in the large pedigrees, we have a lot of missing data, smaller pedigrees where the genetic information is more or less complete. PBAT then would analyze them as one pedigree out of this big pedigree. And what take then these pedigrees that aren't as complete within this huge pedigree as independent.

Christophe Lambert: Another question, although the between and within tests are independent, is it possible for population stratification to mask an association leading to Type 2 error in the screening test? Have you any comments regarding robustness of the two-stage approach to this problem?

Christoph Lange: The – I mean, the first step is, of course, population based and can be inferenced by confounding there, but the second step is a true family-based test. So, any confounding that happens in the first stage can lead to reduced power, but the alpha level will always be maintained. If you believe strongly that confounding is an issue for the first stage in your dataset that you want to analyze, what you can always do is because it is a population-based analysis that you run, you can apply corrections at that stage if you want as well to maximize the overall power of the approach. So, based on the imputed genotypes, you could use one of the standard population-based methods to control for population of mixture and stratification.

Christophe Lambert: Gettin' to the end of the questions. Couple more. Here you've used – they're clarifying if you have used the genotypes of sibships to reconstruct the parental genotypes. And they're just clarifying that you did not have the genotypes of the parents.

Christoph Lange: We did not have the – no, and in the FBAT approach, technically speaking, we don't reconstruct the parental genotype. What we

construct is for all these configurations is sufficient statistic and computes and the offspring distribution conditional on that. In the end – and at the end of the day, from an analysis point of view, it's a very similar thing. Technically speaking, slightly different.

Christophe Lambert: Okay, last question. The late age of onset Alzheimer's disease, whether unaffected family members can be considered as good internal controls 'cause they may be too young, be affected at their early ages, example, 30- or 40-year-olds?

Christoph Lange: I'm not sure that I understand the question correctly, but this is –

Christophe Lambert: So, I guess they're probably asking, are the unaffected siblings old enough that we can be sure that they're good internal controls?

Christoph Lange: I mean, that was precisely the reason why we wanted to incorporate age of onset in the analysis, because by this – particularly the logrank rating or the Wilcoxon rating, if you are unaffected, that happens at a relatively early time point and you [inaudible] that you have relatively different weight in the test statistic. But, if you are 90, 95 and you're still unaffected, the FBAT logrank or Wilcoxon would give you a relatively high rate. So, the beauty of a time to onset or age at onset using that as a phenotype, additionally our analysis was exactly taking this into account.

Christophe Lambert: Well, one more question came in. I have multigenerational extended pedigrees with intergenerational – I think they're saying consanguineous marriages – how can I deal with such a family?

Christoph Lange: Our software is not, by default – if you have genotypes for other parental genotypes, then you can relatively easily split them up and doing this conditional analysis, because the way – this design – this analysis design, it's always conditional on the previous generation. And that doesn't really matter whether the genotype – whether the two parents were related, because from one generation to the next, the transmissions under the null are always mendelian. You have to do the split-up however manually. If you have these scenarios, PBAT is not able to do it automatically for you.

Christophe Lambert: Okay. Well, that's all the questions we had. It was quite a rich set of questions, and we'd like to, again, thank you, Christoph, for a very interesting and engaging presentation. I'd like to thank everybody for attending. For those of you who have asked if this presentation will be available, yes, we will put it, this recording, up on our website within a couple days. And so, if you would like to

watch it again or direct any of your colleagues to it, we'll have it on the webinar section of the Golden Helix website. So, again, thank you so much, Christoph, and look forward to many more exciting discoveries in the years to come.

Christoph Lange: Thank you very much, Christophe. I appreciate it. Thanks a lot. Bye-bye.

Christophe Lambert: Okay, bye-bye, everyone.

[End of Audio]