Thank you, Peter. Don't go anywhere. If I can encourage our speakers to come back to the stage, we'll have a slightly shortened Q and A session, try and get us back on time. There should be a couple of microphones there available. Okay, Ewan?

Hello, it's Ewan again. I'll try not to ask too many questions. For both Peter and [?Sec 0:00:42.1] - lovely, lovely talks. I get a tingling worry about fine grain geographical structures in these PRS' and GRS' and sophisticated uses of them. In your case, Peter, for looking at this very old question and Sec, in your case, for this risk projection. I know I've talked to both of you about this, but I'd just like to - two things. How paranoid are you about it and how do you reassure the people who are paranoid? I suspect there are more people in this room who are paranoid about it than I am.

Thank you, Ewan. I think you're asking basically, about the external generalisability of the scores derived from UK Biobank, and for any predictor, the key is just take it to other samples and see how well it does. We've taken this predictor to three other samples, separate studies in the United States and it works almost identically - remarkably to what we've found in the 300,000 people in phase two and one of those is a multi-ethnic study of several thousand people who are hospitalised with early heart attack in the United States, and the predictor works just beautifully.

It's a great question, Ewan. I think to translate it to my talk. The question is essentially I think, if people in Cornwall were slightly taller than people in Scotland and people in Cornwall tend to find their partners there and similarly, in Scotland, then you might also see a pattern of assortative mating which just reflects mating according to ancestry. Well, we spent 95 per cent of the time on this project in trying to look at how to deal with population stratification. It's a very good question, we're very well aware of it and it's the Achilles' heel for these studies. We're pretty confident that we're not suffering from that in the conclusions that I've drawn from the traits that I have. But if you're not careful, you'll find evidence for assortative mating which is just, pure population stratification. I agree with that.

Okay, thank you, Ewan. Next question over here.

Nick Cross, University of Southampton. As part of the exome sequencing and genome sequencing studies, you're going to pick up individuals with clonal haematopoiesis driven by somatic mutations, and these individuals with an increased risk of developing haematological malignancies. What's going to be done with that information? Is that going to be fed back to individuals, or what?

Yes, so my understanding of it and certainly, the UK Bio team can say more. We're aware of those types of

variants and some are ones that others have recognised and we already see that in other exome sequence data sets and can see that here, and there's certainly many important research questions that are going on from that, but this is not something - nor will any other genetic information, I think at the time, will be returned back to these participants, but I think that's beyond our comments.

Maybe I could speak to that. So, everybody who participated in UK Biobank did so on the basis of clear consent, there will no feedback of individual results to the participants. As we've gone through the imaging assessment as well, we've reiterated that point to participants. During the baseline assessment visit, if an incidental finding was noticed then that would be mentioned to them. Similarly, in the imaging, if during the actual imaging assessment, the technician notices something of concern, then that's checked by a radiologist and about two per cent of participants then get feedback from something spotted during the assessment visit. But the consent is to take part in the study originally and reiterated for the imaging, as anything found subsequently will not be fed back.

Thanks.

I think if I could just add one more thing on this. I think it's an important area of research that is really just starting to be looked at in terms of the consequences of those, and this data set will be great to look at that. We've had some experience looking at our [unclear word 0:05:15.1] data base, which has a much wider age range, so from 30 to centenarians and so, you can detect a lot of these simply by looking at different age strata and their presence there. We also have - like UK Biobank, a very broad phenotypic catalogue, if you will there, and so, just some caution, you've got to be quite careful about the associations we make there, because there's a lot of age-associated phenotypes and causality we ascribe to some of these variants of certain phenotypes versus just a by-product of age is going to be a very important question.

Okay, thank you. Do we have another question over here?

Hi, Dan [unclear surname 0:05:54.1] of the Institute of Cancer Research. All of the really great advantages of the exome sequencing project are also entirely captured by the whole genome sequencing, with the added of advantage of evenness of coverage due to capture inefficiency in exome sequencing. Why do both or why continue with the exome sequencing when the whole genome is going to do it all, anyway?

Well, I think we've all talked about this and we're all complementary approaches and I think that's the short answer, they're all complementary. I think everyone wants to have the value of 50,000 exomes now, 500,000 exomes by the end of next year. Does anyone want to pass on that opportunity and wait for five years to have everyone sequenced? That challenge is an enormous challenge in terms of scale and infrastructure and it will

add some additional valuable information that Nicole talked about. I think they're extremely important initiatives. I don't think anyone here would probably disagree with that, that we ought to carry on with them full steam ahead.

F: Yes, and to add to that, clonal haematopoiesis is a clear example of, you'll have a better resolution from a 50x exome that you would do from a 30x genome, so I think there is complementarity, not only in timeliness of the essence but also, in terms of the biological information that we can get.

Yes, one thing I can say, in ten minutes you can't say much, but thank you to Mark for an extra thank you to John Overton and Geoffrey Reid who's here, and they have done a really exquisite job in terms of the sequencing and the informatics. The coverage in these exomes is extremely high, so mean coverage is not the metric when you're talking about exomes, so we'll look at the proportion of targeted bases that have, you know, let's say 20x coverage or so, right? On average now, in this first QC report of the 50k, we're talking about 94 to 95 per cent of targeted bases at 20x or higher - that's high, so this is high-quality data.

Thank you. I think we probably have time for one more question. There's a question just here.

F: Hello, I'm [unclear phrase 0:08:13.3] - Istanbul, Turkey. I want to ask, there is current genomic data in the UK Biobank and the exome sequences are in the current study; what is the additional information when we compare the genomic data currently for the whole data set and the exome sequencing data?

I tried to give a little bit of a teaser on that but essentially, and you also heard about it in the whole genome talk, that imputation is a fantastic tool and it will continue to improve with more sequencing and better reference panels, but it can only go so far and as you get to a certain allele frequency, imputation accuracy is quite poor. One way to think about genetics is it's all rare variation, certainly when you get to this population scale, the common variation plateaus quite quickly and so, the overwhelming majority of all genetic variation you're going to see is going to be rare variation, so the sequencing efforts are filling in an enormous amount by counts and other metrics of variation that doesn't exist, period, or isn't very accurately imputed in the data. That's what I would say.

F: Yes, so if we think of a rare variant tool, it would be [unclear word 0:09:43.5] about ten per cent of the time in an imputation data set, but we can recover the completeness of genetic variation down to variants that are covered by a single person, or like some allele genome sequencing data. I think the other point that I mentioned as part of the BB project, it was really an important one is that, having a wide distribution of variation really allows you to model extremely accurately pressures on the genome. For instance, selective constraint or mutational constraints, which we know very well are not constant over the genome. Our ability

to really understand what is a known-normal pattern in the genomic region, for instance, this disease associated variant, really critically depends on our ability to model genetic variation through the entire genome. Having that background information that tells you for instance, in that region, mutation rates are higher or lower than average, also informs how you interpret definitive [unclear words 0:10:46.7] variant. For this reason, these data sets are all highly complementary. It will you give a different part of the picture.

Okay, thank you. I think we need to bring the Q and A session to a close. What I would suggest is, if you have a burning question, then please try to collar our speakers after this session. We'll draw this to a close and I'd like to thank our speakers for their presentations.

[Round of applause]

[END OF TRANSCRIPT]