A guide to accessing the multi omic date in Understanding Society

Yanchun Bao: So I think I just use about 10 minutes up to 15 minutes quickly guide you about if you want to apply our data. So this is about how to apply this multi-omic data Meena and Anna just talked about of Understanding Society and, yes, just overview what data are available to you and where to get the data, and if you have question around that I can quickly answer that. So I mean Meena and Anna quite often use "genomic" and sometimes we also call it "genetic" but basically what we mentioned here is just SNP data. And for Understanding Society are sample size for SNP data is 9,920, so you know just a little bit below 10,000, and we have this genotyped, which means physically measured about 500,000 SNPs and we got some 14 million SNPs imputed and the SNP data measured by this Illumina HumanCoreExome chip and performed at Sanger Institute. And it's imputed using combined UK 10K and 1000 Genomes reference panel and the population, I want to mention is white European because sometimes we got some inquiry ask if you have you know minority group of genetic data and, unfortunately, our data based on White European sample.

OK, the second omic data we have is you know DNA methylation, as Anna mentioned, we often refer it as epigenetic data, and we have two batches. The first one is measured in 2017 at that time we got some funding to measure a relatively small subset so we have 1,174 individuals, about 857,000 sites after we have some quality control. For second one we did it later at 2020 and for this one, we have relatively larger sample size so 2,480 and again, we have similar CpG sites after QC and this one also the number of CpG sites is slightly different, but the majority are overlapped so you don't need to worry too much the difference and basically we use Illumina Infinium Methylation Epic Arrays performed by Professor Jon Mill in Exeter and the data normalized and cleaned in Biological Sciences department in University of Essex and, again, the population is White European and as Meena, sorry, as Anna also mentioned, we have five epigenetic clocks based on this methylation, some subset of the methylation, and this related to age. And the third one, is our proteomic data, so we have two panel: cardiometabolic panel, so we have 6,180 individuals, 92 proteins; similar for neurology panel same number of people and the same number of proteins. And it made by some company called Olink and they uses Proximity Extension as a method and Anna did some quality check for that and we created some kind of User Guide. And for this data we have 92% of White people, and we do have some other minority group have this data so that's that.

So that's basically three omic data we have and there's two way to use that data, so depends on what you want to do. The first one is that if you just want to have genetic data or epigenetic data as your control data for some studies and you basically just wants those information, then you can directly go to the European Genome-phenome Archive to download the data and the only thing I need to mention is that because it is deposited by different partners, so the genetic data and epigenetic data do not link to each other, and we are going to re-deposit those data later, and in that case you should be able to link these two datasets of course the epigenetic data is a subset you know the sample is a subset of the samples of the genetic data and also, you cannot use the IDs to link to any of the main survey data. On the other hand, if you just want to use proteomic data or

epigenetic clocks then we're going to make it available publicly so, which means we're going to deposit in the data archive with our nurse visit data so we're going to update that very soon. The second type of way of using the data is that if you want to link those genetic or epigenetic data to any of the survey data, then you need to apply for that.

Of course you can also link those data with protein data or epigenetic so I give here the link of our website to do that. So basically if you click this link, you will find that you need to fill two things. The first one is your application form, so you briefly introduce what you want to do, and you know give some summary and why you think it's important to your study and list the principal applicants and anybody else that are going to use the data. And also in the form, you are going to find something like you know do you want to use genetic data, epigenetic data or do you want to have the imputation and so that's information. Additional to that you need to fill in list of variables you want to use, which means survey data. So basically, you will see somehow like a form, like the one I list here, so you should give something like what's the files or what variable comes from and your variable name and which wave it comes from so from B1 to B18 it's just BHPS data and from 1 to 10 is our Understanding Society one until wave 10. Sometimes, some people just give very raw description so, for example, I want any variable related to smoking. And that would cause trouble because we don't know what exactly you want, so do you want just to know, you know, who is a current smoker, you know, previous smoker or non-smoker, or do you want to know more about their history: when they stop smoking, when they start to smoke. Or you know, other thing. So, in that case you know if you gave information like that it's very likely we're going to contact you again to ask you to clarify which variables you need. So to save the time of your process it's best to you first go to the Understanding Society website to look at those available variables and find out the name and a list of them. Of course, if you want to know, for example, you know, what information available, and so on, you can contact us before you start to make application.

OK, so briefly give you some idea about this data, how to say or size of the data you're going to receive if you apply the data and get it approved and finally, you know, receive data. So we have basically we're going to use a way called Zendto which is just an email sent to you with some link is password protected way to send your data but that's only send some data under 2 gigabyte. And the for the large data, we are going to send by FTP website with password protected, and if you just send it as measured genetic data, then it's just up to 3 gigabyte so you know, once we zip it we probably can send you by Zendto, but if you require any imputation data, then it were, you know, up to 95 gigabytes so quite could be quite large and it can be open in R or Plink or some other software and for epigenetic data we have two .gds files. And it up to 23.4 gigabytes and it can be opened in R and for survey data normally it's in Stata or text format we're more likely send you by Zendto and if you also additionally apply for proteomics data or epigenetic clocks those were sent by Stata or text format and basically, all of those data will be linked by a unique ID, but you will not be able to link to other survey data, except for the one you know you applied, and you should not try to do that and I think I already mentioned, but I want to emphasize that if you just want proteomic data or epigenetic clock alone, not with any genetic data or epigenetic data you just go to the data archive to download it so you don't need to go through this application process.

And last to see, I want to mention, we have some User Guides for biomarker which is online and we're going to put the proteomic User Guide online very soon, and also, we have the epigenetic clock User Guide online very soon, but for genetic and epigenetic data we will send you some readme file, with more information about quality control, how we imputed data and so on this type of information, but we will not put it online because this data, you know are available only through application so we try not to confuse people by just to put some readme file, but without the data.

Okay, so last bit would be if you have some general question, you can contact Meena with her email address I list here. But if you really start to make application and want to know the progress and so on, or you have particular question about some variables and so on, you can contact the genetic team which I give the information here.