M: Thank you, Alice, and also just to recognise I think there's a very strong collaborative working relationship that's been developed between Biobank and Regeneron through the course of this project. Certainly to recognise the contributions, particular individuals, so John Overton, Jeffrey Reid, John Penn and the many others in Regeneron who are really contributing to the success of what is quite a complex delivery programme to achieve. We've talked about genotyping, we've talked about exome sequencing and so I'd like to welcome to the stage Professor Nicole Soranzo, who is the head of the Human Complex Trades Group at Wellcome Sanger Institute, and a member of UK Biobank's whole genome sequencing expert working group, to share some of the perspectives on the whole genome sequencing programme that's about to start.

Thank you. Good morning everybody, so mine will be a very data-poor talk, so I'd like to give you some perspective about how we're thinking about whole genome sequencing in UK Biobank. It seems really hard to think that less than a year ago we met with a group of investigators at a Wellcome Trust to discuss this prospective project and the project indeed already start. This is really a testament to the ability of UK Biobank and our funders to really buy into the vision and get things done. The genome sequencing project is part of the UK Government's Industrial Strategy Challenge Fund for - really has an ambitious aim to produce a deep sequence characterisation of the whole genome sequencing of the entire UK Biobank. Many logistics are still to be, if you want details, still to be worked out in terms of what scientific technological and organisational data, but really the ambition is to complete the sequencing of the entire cohort within the next five years.

The work has been divided into two main stages, so a pilot or Vanguard phase, which will sequence the first 50,000 individuals and then the main phase which will sequence the remaining 450,000 participants. Indeed as I mentioned, the Vanguard phase has now officially started, we'll sequence a 50,000 participants using Illumina short-read technology, for those of you who like sequencing, this will be done using 150 base per [unclear word 0:02:36.8] PCR free libraries. We aim to achieve 30x average coverage of the human genome with 85 gigabytes minimum per genome. The sequencing will be carried out at the Sanger Institute and we'll start officially the 31st of July of this year, and the plan is to have the entire 50,000 individuals turnaround for sequencing in the next 18 months. Just to give you an idea of the scale of the challenge, so this is predicted to generate about 4.5 petabases of sequence, which is pretty close to the entire sequencing output of the Sanger Institute until now, so it's really a major scale-up for us as well. In addition to this project we will also be running a series of innovation projects as Mark, or pilots, have already alluded to, which will really inform technology for the main phase of the project.

I really want to stress that this is in no way a competitive effort to the exome but rather builds on and really is complimentary to the genotyping and [unclear word 0:03:45.4] sequencing datasets. I thought it would be really worth it to think about what kind of information or data we will get out of the genome sequencing project. This plot is depict in a proportion of single or [unclear word 0:04:02.1] variants, so that it

is indeed informed by an empirical evaluation study, where we are the same individual sequence with exome sequencing at 50x, genome sequencing and imputation. This is really illustrative cartoon which represents a lower bound estimate, but the main point here is that using a standard traditional exome capture array, you are only able to see between two and two and a half percent of the total variation that we see from whole genomes. We know that much of this variation is actually not accessible to imputation, even with later, more advanced imputation reference [?planner 0:04:41.2], we know that both of the percentage of [?snips 0:04:44.2] that can be recovered, and also the accuracy of the imputation falls quite dramatically below 1 per cent [unclear word 0:04:51.4] frequency, so that indeed the vast majority of variants below this threshold are not imputable.

What the genome sequence will give us is access to the remaining 98 per cent or so of the human genome and also access to the hard-to-impute fraction of human genetic variants. Obviously the first thing we are interested in understanding is what is the contribution of this rare variant to human phenotype. You heard very well about the value of coding variation, but the question is is there equivalent scope and value to look outside protein coding regions. We think this is true, this is the case and one of the study that informed this thinking was actually based on phase one of the UK Biobank imputation, so this was published about 18 months ago by my group with many collaborators. Here we carried out a whole genome base scan in about 170,000 individuals, looking at 36 phenotypes and the capture variation in full blood cell counts. What we found is the largest to date number of independent variants with 2.7 thousand statistically independent associated variants. What was quite remarkable is that for the first time we discovered a large number of what we call low frequency between one and five per cent, or rare genetic variants, and so this really allows us to study the properties of these variants in bulk.

What you see here is this the allelic spectrum, is that indeed the first observation is that there was really a [unclear word 0:06:31.9] of genetic effects between the common low effect variant, which we discovered in [unclear word 0:06:36.9] Human Association Studies, and indeed variants that are low in frequency but begin to approach, effect size, they are comparable to clinically actionable variants. [Unclear word 0:06:49.8] what was even more remarkable, if we divide these variants based on where in the genome they fall, we see that for common variants as we would predict, the vast majority sit outside protein coding region, as highlighted from this yellow, blue and green colour. Indeed nearly 70 per cent of the rare variant associations also map outside protein coding genes, and what is even more striking is that if you plotted the average effects size for the association, this rare predicted non-coding variant have similar effect sizes to rare highly damaging coding mutation and much greater than the common counterparts. It's really suggested that there is potentially a pool of clinically very informative genetic variants that are outside protein coding regions that we want to discover.

Also this information comes from other studies and this is one by [unclear word 0:07:47.5] based on a DDD study which looked at under, children with developmental disorders that are undiagnosed based on all exome sequencing studies. What they were able to find is that there is indeed a high rate of the novo mutation,

they are mappable to ultra-conserved elements of in this case a relevant tissue type, which is foetal brain. This really suggests that a deep understanding of the genome structure, including mutation rates, demography and natural selection coupled with really deep regulatory annotation and also with the growing number of large scale of experimental study, will really allow us to identify for each disease, a non-coding region, particular irregularly regions that are a particular interest for specific traits and disease. We can think of then looking, association of rare variants under this region in a manner similar to what we are doing now for coding variants in exomes. We done a few [unclear word 0:09:00.1] collation just to understand what spectrum of effect sizes we think we can capture, so in this case we, a [?test 0:09:07.3] association with aggregated rare variants where by modelling different accumulatively effect sizes for combination of these genetic variants and their accumulatively allele frequency.

In the 50,000 of [unclear word 0:09:22.5] study, we have strong power to detect association in this top right part of the corner. Just to give you an example, for instance, we have a full nearly perfect power to the cover association with accumulative effect size of greater than point six standard deviation for variants that are found in six people out of 1000. In the full dataset we obviously have more power and so for the same effect sizes we will be able to detect effect when there is at least four carriers out of 10,000 individuals. We really think this is going to be a very worthwhile effort as I understand, the non-coding part of the genome improves to really discover many more associations. The other thing I wanted to point is that indeed sequence variation is only a very small part of the overall picture, what makes us biologically interesting creatures. Indeed the genome sequence will really allow us to look at very diverse classes of genomic variation. The perhaps most important one are different classes of structural variation, so there's now available software, they use properties of [unclear word 0:10:34.7] sequencing to drive information of many classes of variations, such as Deletion, Duplication, Inversion and Mobile Element Insertion.

It is very clear the genome sequencing will foster the development of even more accurate ways to predict and study these variants, and particularly to evaluate their contribution to human phenotype. The other interesting aspect of biology is telomere length that's being show by [unclear words 0:11:02.5] colleagues, that indeed you can use now computational methods to drive quantitative measurement of the length of telomeres, which is indeed very well correlated to traditional methods such as Southern blot. This really opens up a possibility of beginning to really starting to understand the role of telomere length variation, but in basic biological processes such as aging, but also in a complex diseases as shown in this case, in this example for types of cancer. A third aspect which is also I think going to be really important is our ability to study mitochondrial DNA variation and its impact on human traits. Mitochondrial really [unclear word 0:11:47.0] small molecules with 17 kilobases in humans, there are 13 protein coding genes, they are mostly implicated in electron transport, chain which in turn control OXPHOS function, as well as a few number of tRNA aided control, [unclear word 0:12:05.1] or the translation of MT and coded proteins.

We know already that mitochondrial function is governed by both mitochondrial variants, nuclear variants and environmental factors. Indeed this function, which is also typically associated with aging, leads a

progressive biogenetic decline that has been correlated with many diseases, including immunological, metabolic and degenerative diseases but also with cancers and aging. We will be able to use genome sequencing to study this, so we know already that indeed from array data we can call very accurately mitochondrial haplogroups, which show in Europe this structure of function. What the sequence will give us is the ability to resolve cause and effect, but it's all been effectively variants that define sub-haplogroups with different functional properties. Another interest aspect of mitochondrial biology, indeed copy number, copy number variation on mitochondrial has been associated with tissue specific variation, which has been linked to stress and indeed it's being associated now with clinical endpoints, such as major depression. The mitochondrial has about 100 full coverage compared to autosomes, and so we'll be able to very accurately call mitochondrial a variation from our genome sequencing data and then correlate it to a host of phenotypes and disease outcomes.

This is really to conclude by saying that indeed from the sequencing data w will be able to achieve a broad exploration of different classes of genetic variation, including rare and coding regulatory variants, structural variation, telomere length, mitochondrial DNA. It's undoubtably the case that they sort of produce many useful tool for genome analysis and not unimportantly also an [unclear word 0:14:08.3] imputation reference panel that can be extended to other disease datasets. This will allow us to really address many very fundamental and broader questions about biology, and these are just some of the examples. Around the contribution of rare variants with complex traits and application, for instance, prediction, as you heard before, understanding of the role of regulatory variation and really deciphering the regularity [?grammar 0:14:36.1] of the human genome, ability to really map with high certainty in the landscape of evolutionally mutational constraints in the human genome, and this really provides us with background information onto which we can [unclear word 0:14:49.2] disease causing variants. Finally, importantly processes that are associated with not only this level of variation but human evolutionary processes at large. Really I joined a chorus of people saying that this really will empower our UK and worldwide science, and thanks very much again to the UK Biobank for making this possible.

[END OF TRANSCRIPT]