Last, but not least is young Adam Butterworth who has done some beautiful work, and Adam has been tasked with telling us what other [unclear word 0:00:06.0] UK Biobank should consider, and Adam told me before he's 37, but I think you look like me at Adam, 32 actually.

That's chronological age, not biology age, so I'm not going to tell you about polygenic risks cores. I'm not going to tell you about all of the [unclear word 0:00:25.4]. I'm going to tell you about proteomics, and I don't speak for a particular company, I don't speak for a particular community, so I'm a genetic epidemiologist rather than a proteomicist, but I'll hopefully show you some data, which is why I think proteomics would be a good thing to do in UK Biobank, much as the previous things we just heard about, I think would be also great to do.

You've seen various flavours of the slide in the last couple of representations, but I think by overlaying some of these molecular layers in something like UK Biobank, there is lots of possibilities for understanding fundamental biology, disease, aetiology, risk prediction, drug target prioritisation et cetera, and I'll hopefully show you some examples of that.

Can you speak close enough?

Sure, so we have a mini Biobank by UK Biobank standards based in Cambridge of 50,000 people and that's been, if you like, a good testbed for like the [?Didi Asset 0:01:14.5] Scale. Do they work well in subsets of these individuals? We haven't measured all these things in all 50,000 people, but you can see here how we're trying to bridge the gap between DNA and disease. There is the heritable proportion of disease, but how does it act? What are the intermediate pathways that we might try and target to prevent diseases? As I say, ultimately, I'd like to see all these things in UK Biobank. Some of them we have already like the blood cell traits that Nicole mentioned earlier. We've just heard about the Nightingale Platform that's about to be done, but I think each of these layers could offer distinct and complementary advantages.

I'm going to talk particularly about the work we just published in the last couple of weeks using our particular platform called SomaLogic, which is one of a number of different proteomic platforms we tried, but, you know, I'm not saying we should necessarily do this particular one, but I'll tell you a bit about what we find and why we think it was useful.

First of all, why look at proteins full stop? Well, they're the drivers of biology. The DNA might carry the recipe, but it's really the proteins that are actually acting in biology, and often dysregulated in disease, and, of course, many of the drugs that we have, the medicines that we use target at the protein level, so it's a very relevant domain of biology I think. As we heard, many different assays are now maturing and becoming high throughput, enough to use, but also very wide angle, and as we'll see examples like the SomaLogic or the

Olink Platform now, give the possibility not to measure tens of biomarkers, as we heard, or about to be released, but maybe hundreds or even thousands of biomarkers. Again, when we think about the limited precious UK Biobank samples, what we want to do with those, perhaps things that give you a lot of data points and a lot of potential answers might be the way to go. Of course, protein is very relevant, therefore, to a wide set of disease and phenotypes. As we've heard in UK Biobank there is almost any kind of disease one could imagine, so there is potentially a very wide relevance here.

People sometimes well, in studies where you have RNA and DNA, why do you need proteins as well? Well, we think there is an imperfect correlation, actually this gives you another layer of detail, and we find things that you wouldn't see in those levels.

Here is the assay we used. It's used SOMAscan. It's from a tech company out in Boulder, Colorado. We tested five or six different assays at the time in a small pilot study as we heard about with the whole genome sequencing, trying to find the best thing to do. Ultimately, probably this won out because it was the broadest, and as you can see here the company now does, in fact, I think more than 5,000 different proteins that they can measure in human plasma in a relatively scalable way, and that's using this aptamer-based approach, which I won't go into the details off for time. On the right here, you can see it a very diverse set of proteins, so it's not focusing on a particular biological domain, although it's not a completely unbiased approach, so they have been selected to be ones that might be of particular interest.

Before you run these assays in a large scale, you often want to know, well, do they work? Are they reliable? In some ways we viewed our experiment, which was essentially doing Duas of 3500 proteins in about 3000 of our blood donors as a generic validation, so what you see here across the bottom is the position of a generic variant on the human genome, chromosome is one to 22 and then we utilised the fact that each of the proteins that we're measuring is coded for by a gene somewhere in the genome, so on the y axis here you have the where is that gene that codes for the protein? You can see these diagonal red lines that runs through. These are what we call the SiS association, so often you do your genome-wide screen, you look at ten million imputed variants across the whole human genome, and the only region that lights up is right where the gene is that goes with that protein. That already tells me that that assay for that particular protein is picking up the right thing, so in some ways we can use the genomics to tell us that we're measuring the right thing.

The blue dots that you see, they are sort of pale blue, so what we call trans. That's a region somewhere else in the genome that seems to be influencing this protein, and often these make good sense, so we might have that the gene is a receptor for the protein or that the genetic signal is near the gene that is the receptor for the protein, but much of the time we don't know what these are, and these are starting points for novel biological understanding, why does this particular region of the genome influence levels of this particular protein?

What you can see as well is that there are these kind of stacks, and these are some of their kind of very highly pleiotropic, some of their fundamental regulators of biology or perhaps areas which affect big protein cascades like complement factor, for example, so we think there is lots to learn about biology. In our paper,

which you can read about, just came out in *Nature* that we found about 2000 gene-type protein associations, so something, there is about four times more than had been published from studies previously. I was going to try and create a cartoon of what it might look if you did this in 100,000 or 500,000, but I think the blue would be probably a bit overwhelming to see.

What can you do with this data? Well, actually, we have a relatively healthy cohort of blood donors, so unlike UK Biobank we don't have many incident disease events at the moment, but others have used this platform in other studies to show that you could start with 1000 proteins and boil then down with statistical reduction techniques to a nine-protein score, and they showed that you could actually predict future cardiovascular events by dividing people into deciles of this score. You can see people here in the tenth decile have a much higher chance of having a future event. One of the proteins that was particularly interesting to us, the second strongest one was this protein called Matrix Metallopeptidas Protein, it's 12 Mmp12, and they showed that people who had higher levels of this had a higher risk of events.

What we did in our study, because we were looking at genetic data, we tried to answer the question of it might be predicted by a marker, predicted protein, is that a casual protein for heart disease? We looked at a series of about 14 different generic variants in the Mmp12 gene region that all independently affects your level of Mmp12 protein, and surprisingly we found that the [unclear word 0:07:18.4] here that increased plasma in Mmp12 actually give people a decreased risk of heart disease. We have an interesting situation where you might have been quite excited, particularly if you are one of the companies who has an Mmp12 inhibitor that people have thought about for preventing or treating lung disease, you might have thought this would be good for also preventing heart attacks, actually our data suggests that - and this is perhaps protected by a marker for heart disease perhaps produced in response to initial damage, so there is definitely ways we can find causal insights into disease, sometimes in surprising directions with relevance for therapeutic targets.

The other example I'd like to show you is about this idea of bridging the gap between DNA and disease. The Duas era for diseases has been phenomenally successful at finding statistical associations, very small P values, regions of the genome that relate to diseases, but often they look messy like this. This is a region where there is a signal within inflammatory bowel disease, and there is what, I don't know, 30/40 genes in the region? How do we know which one it is? Probably many biologists could make a good case for any number of these genes.

One of our protein signals happens to look almost exactly the same, statistically co-localises, and actually it was a cyst signal for this protein with the gene highlighted here called MST1. This suggests that actually maybe the causal actor in this region is MST1. Actually, when we looked at our other proteins, we also found other proteins that shared almost exactly the same signal. If I flick backwards and forwards here you can see the disease signal and the protein signals looking very similar in the same region, and this was for another protein that's coded somewhere else in the genome called BLIMP-1, and actually, when we looked at the inflammatory bowel disease, the Duas papers, there is also a hit in the gene region for PRDM1, which encodes BLIMP-1. Again, previous studies have made good cases for both of the two genes in the

neighbourhood here as being the biological actor. We actually think our data suggests that this is PRDM1 in this region, and MST1 in the other region, so we can start to fill in the layers between DNA and disease and start to try and pinpoint, go from beyond the statistical associations for disease to some meaningful biology.

Now, I could stand here and tell you proteomics is going to be a wonderful [unclear word 0:09:29.9], we should go out and do it immediately and it's all going to be fun. Of course, there are challenges along the way. Some of them are specific to proteomics, but, of course, many of them are germane to many mix assays or even smaller scale biomarker assays. We obviously want to validate these assays before we go out and do them at scale. We obviously used assays that have been used in somewhat smaller studies than our study, and, again, maybe studies like ours to act as the pilots, if you like, for a large-scale biomarker in UK Biobank. There are always challenges of interpretation which, whatever domain you look at in proteins, we perhaps have challenge about whether a protein is free in the blood or is in a complex and how we interpret those measurements and results. There is great potential for non-biological variation batch effects et cetera to influence things that need to be carefully controlled for, and we know that the UK Biobank team has great expertise now in doing this, and, of course, by the time we get to going beyond one protein, one measurement at a time, how do we go from the 35 biomarkers that are about to be released to 3500 proteins? That brings different challenges in terms of statistics, computation, biothematics, et cetera, so it may not be easy, but I think there is enormous value of doing these kinds of things at scale.

Just to conclude, I think certainly, and many omex assays, and particularly some of the multiplex proteomics assays are now becoming sufficiently high throughput and reliable for use in very large population studies, and they are also particularly wide-angled now. They are really growing in the diversity. A few years ago, we were talking about panels of ten or 13, now we are talking about panels of hundreds or thousands, and that just enables us to answer a really greater range of questions and, of course, we shouldn't forget these challenges.

I'd really just thank Ben Son, who is the PhD student who led all the work on the proteomics, and also the diverse range of funders, both pharma, tech companies as well as national and charity funders who enabled all of that study to come about, both with measurements, creating the study, blood samples, et cetera, so really a big team effort and perhaps some hints about how these things might be funded going forward. Thank you.

[END OF TRANSCRIPT]