The UK Biobank was set up more than 10 years ago to collect information about more than half a million people across the UK, following them through time to see what happens to them as they lives their lives, creating an unparalleled resource for biomedical research. October 2018 sees the arrival of two new papers based on UK Biobank's work, published in the journal Nature – the first announcing the release of genetic data from all 500,000 participants, and the second describing new insights into the relationship between genes and the structure of the brain. Here's one of the lead researchers, Professor Jonathan Marchini from the Wellcome Centre for Human Genetics at Oxford University, speaking to Kat Arney about what these new data represent, and what researchers could do with them.

Jonathan: It's genetic data on 500,000 people from the UK, and for each of these people we ran an assay – we call it a chip – which measures approximately 800,000 positions in the genome. So for each of those positions in the genome we're measuring which type that individual has – it can either have zero, one or two copies of the mutation we know exists at that site. We've got 800,000 positions in half a million individuals - this really encapsulates the genetic variation these people have and allows us to build on it and link that data to health outcomes.

Kat: So this isn't looking at the entire genome - all the As, Ts, Cs and Gs, the spelling of everyone's DNA - it's looking at those 800,000 specific locations and saying what has each person got here then trying to map that back on what the person is like and what happens to them in their life?

Jonathan: Exactly. The human genome is actually 3.2 billion positions - we've measured just a fraction of that - but they are positions that have been chosen very carefully so that we can capture the majority of the genetic variation these people have. Also in this project we have increased that number up to around 96 million, using statistical methods known as genome imputation. These methods build on other external reference datasets that are much denser in the number of sites they have assayed, and the idea is that we are leveraging relatedness between people. So, for example, my genetic code will be very similar to the genetic code of members of my family - my close family, cousins, second cousins and so on - so we are leveraging that shared genetic information between people to make predictions into the UK Biobank dataset.

Kat: So if I get this right, the way that imputation works is if your data says that someone has this particular variation here, and this particular variation here, and you look at another dataset and also find this variation and this variation, but that dataset has a bunch more data around it, you can infer that your person is likely to have the same?

Jonathan: That's exactly right. We're matching the DNA from people in the UK Biobank with other people in these external databases and doing that exact prediction process. It's more complicated than that and we have to use statistical methods to make those predictions accurately.

Kat: It's a huge challenge, the scale of this data. How have you managed to deal with this much data? You have half a million people and now you've expanded out to information about 96 million locations in the genome – how do you manage that and keep a handle on it?

Jonathan: You're right, the dataset ends up being over 2 terabytes of data, and that's actually in a very compressed form. There were substantial challenges over the last 3 to 5 years, refining our methods so that we can handle that. We've been used to dealing with datasets of one, two three or four thousand people, so lots of the datasets and the tools were using to deal with this data really wouldn't scale. There was quite a lot of effort from people in my group and others in Oxford and around the world to develop new methods that would do the same things that we were used to, the same quality of the data processing, but just much faster.

Kat: When you look at all this data and then the people that are in the UK Biobank, were there interesting things that you noticed? For example, I noticed that you described in the paper that there is a surprising number of families that seem to have all gone in for the study together?

Jonathan: Yes, that was surprising initially - we found that about 150,000 people in the dataset have a relative in the dataset of the degree or closer. A third degree relative is the equivalent of having an aunt or uncle in the dataset. Roughly a third of the people in the dataset have someone who is that closely related, and so we were initially surprised at that, but when you think a bit deeper about it and you think about how the dataset was collected, by asking people at their GP's surgery whether they'd like to be enrolled in the study. You can imagine that someone might say yes to that, they might go and do that half day assessment process for the project, and they might tell their friends and their family about the project - if they had a positive experience that might encourage their family members to take part.

Kat: And can that actually be useful? Why is it useful to have related people in this kind of study?

Jonathan: One key thing that's very useful about the relatedness in the dataset is that it allows us to make measurements about how genetic a given trait is. If we take diabetes, for example, having a whole range of relatives in that dataset allows us to infer how much of diabetes is genetic versus non-genetic, and by nongenetic I mean due to our environment. And because UK Biobank has measured many traits for all of these individuals, we can make statements about how genetic all of those traits are.

Kat: So you've got all this data – you've got data from half a million people, you've got loads of data about what they're like, loads of scanning data, health data, you're going to be following them up through their medical records – what happens next? What can people do with this data and how do researchers get their hands on it?

Jonathan: Researchers all over the world are using the data already for a wide variety of different scientific questions and projects. I guess maybe the primary one is if a particular researcher is interested in a disease they can take each of those 96 million positions in the genome and look to see whether the genetic data at that position is correlated or linked to the disease information on all of those people.

Basically, you're scanning through the genome and you're looking for each position and asking the question does it look like this position is related to whether a person has the disease or not? And that's very useful because if you do find a position of the genome it might indicate that that genetic variant has a causal relationship with that trait of interest, and then when you look at that region in more depth and you look at which of the specific genes in that region it might give you information about how you start to design a new drug to help treat that disease.

Another area that the UK Biobank will be useful is to make links between our genetics and our environment. All the people in the UK Biobank have undergone a large number of measures in addition to their genetic data - many of the individuals have information about their diet and they've been asked lots of questions about their lifestyle. We think that for many diseases there is likely to be a link between someone's genetics and their lifestyle, their environment, and this dataset will allow us to dig deeper and answer those questions.

Kat: And are there particular regions of the genome that you've been really interested in?

Jonathan: Yes, there is one particular region of the genome called the HLA, which is where all of our immune genes are located, so in the beginning of the project we designed the array to particularly focus on that region. We've got very detailed information and this allows us to then look at that particular region of the genome when investigating diseases of the immune system.

Kat: Another paper that's come out of UK Biobank is looking at the connection between genetic variation in the people in UK Biobank and their brains. What was the idea here?

Jonathan: Another exciting extra part of UK Biobank is that the project is going to brain image 100,000 of the half a million people in the study. Now each of these brain images is able to measure the function of someone's brain and the structure of their brain in various different ways. It's a very high dimensional data set that you end up with when you image someone's brain, and we know from other brain imaging experiments that these types of images are relevant to particular diseases of the brain – neurodegenerative diseases like Parkinson's, Alzheimer's, multiple sclerosis, stroke, motor neurone disease and things like depression and schizophrenia. Because we've made these measures that are related to these diseases we can then ask the question, are there any genetic influences on these measurements?

One thing we do when we take these brain images is we measure the volumes of particular structures in the brain. We've got these volume measurements of things like the hippocampus in the brain, so we can ask the question does our genome influence the size of the hippocampus in someone's brain? We can measure properties of the white matter in somebody's brain and ask whether there are genetic influences on the white matter.

Kat: And when you started to look at all the people's brains and mapping that back to their genetic data and their health, were there any interesting things that you noticed? What stood out from those data?

Jonathan: Well, we made a lot of findings. We found about 150 different regions of the genome that linked to at least one of the different measures that were carried out on individuals' brains. That was quite striking because if we're finding 150 regions of the brain when we've only looked at 10,000 people, when we actually have the full 100,000 we are really looking forward to making some big discoveries. Some things stand out – we found one particular region of the genome had a very strong relationship to the structure of our white matter, and this region of the genome is very related in other studies to diseases like multiple sclerosis, stroke and motor neurone disease.

Kat: And the white matter, that's sort of the connecting stuff between the brain cells?

Jonathan: Yes, that's right, the cortex of our brain is on the outside – the grey matter is where the majority of the neurons are - and you have these links between various parts of the cortex through the white matter, the more central part of the brain.

Kat: What can we do with that information? That sounds like something that's really interesting.

Jonathan: It is! I think it focuses people very specifically on the genes. It allows people to make more detailed experiments of how these genes affect white matter, understand the biology in a deeper way, and then do follow-up studies specifically linking genes to these particular diseases. It's a very important step in the process of understanding the biology and then potentially building a new treatment for these diseases.

Kat: What's next for the brain imaging aspect of UK Biobank?

Jonathan: Roughly 20,000 people are being imaged every year up to 100,000 people, so I think researchers will continue to do these kinds of genetic studies as that dataset increases in size. And as the dataset increases in size we'll make more discoveries. The hope is that the biological story, the biological basis of these brain traits and the diseases that they may impact will become clearer. And the more the biology becomes clearer, the more we know how we might make interventions and build new treatments for people.

Kat: What's next for UK Biobank? You've announced this dataset looking at 800,000 positions in half a million people, but where are you going next with this project?

Jonathan: One of the most exciting things that is going to happen over the next year or so is the gradual release of what we call exome sequencing data. A consortium of pharmaceutical companies has come together and offered to pay and carry out a very detailed genetic sequencing in the exome – these are particular regions of each gene around the genome, approximately 2 to 3% of the genome. And the genes are the main coding functional part of our genome.

The dataset we currently have, while it is amazing it doesn't interrogate those regions of the genome in quite the level of detail that we like. Specifically, the current dataset we have isn't able to look at very rare variations or very rare mutations that might exist in our genetic code. By using a different technology – genome sequencing specifically in the gene regions - we can go deeper and we can look very specifically at these very rare variants that might, we hope, enable us to make new discoveries that are much easier to take forward to the next step. If we know very specifically that a particular gene has a very rare mutation in it that links to our disease risk, then by knocking out that gene or modifying the gene we might be able to make a drug that impacts the treatment for that disease. Beyond that, what we'd ideally like is to have whole genome sequencing but that is much larger and much more expensive project.

Kat: Finally, the announcement of the release of this dataset does mark a milestone in the project. How do you feel to have got to this point?

Jonathan: Well, I'm very excited! I feel very fortunate to have been part of the project and it's really nice to see now that the dataset is being widely used and we're able to play a small part in that process. UK Biobank is a visionary project - I think it will take a long time for other countries to collect data similar to Biobank at this scale.

Professor Jonathan Marchini from the University of Oxford's Wellcome Centre for Human Genetics and UK Biobank. To find out more about UK Biobank or apply to use the data, simply visit ukbiobank.ac.uk