Biological Datasets

Anna Dearman: OK, hi everybody! I'm going to talk about the biological datasets we use, and go over a few of the same points that Meena did, maybe in a slightly different way.

Just to reiterate "omics": so, all living things are made up of several different classes of molecules so you'll probably be familiar with genes and proteins, for example, and within these classes, there are sort of thousands of individual different molecules and, whereas in the past, people used to take a specific gene or protein that they had a well-worked hypothesis for, that might be related to the thing they were interested in, like disease or something, nowadays with these high throughput laboratory methods we're able to just scan a broad range of individual molecules within a class, and that's what "omics" are, and so it enables people to have a more agnostic approach in their hypotheses. So you can just scan all the proteins or genes that you have data for to see whether they do associate with the thing you're interested in, rather than having to pick a gene or a protein and just look at one at a time. So for Understanding Society we have four biological datasets, three of which we class as "omics", so genomic, epigenomic and then the new proteomics data set, and then for biomarkers that's also - it's a little bit smaller, so we don't really call it an "omic" - but it's also a useful biological data set.

So, I've tried to sort of summarize the basic principles that you apply in "omics". So you start off with your long list of genes or proteins, or whatever the feature is that you're measuring, and then, one by one, you find out whether they seem to be associated with your variable of interest, so educational attainment, for example. And then, as a result of your analyses you'll get your significant hits which will have your low p-values, and you'll get your effect size. So then, using the p-values, you can pick out the genes or proteins - or whatever it is - that's most important, and then you can use this subset of of genes or proteins to investigate what are the biological mechanisms that we know about this subset already from previous research which might be underlying the thing you're researching, so education. You can see whether, if you test a separate population for a sort of aggregated measure of these significant molecules whether you replicate your findings and you can predict outcomes using it. And with certain types of analyses, especially with genetic data, you can sort of start to infer causality and I'll talk about that a bit later.

So one of the great benefits of these biological data when you're researching health is that it's in theory, a better, more objective measure of health because you're literally analyzing something from someone's body, and like Meena said, you can sort of measure biomarkers of

an oncoming disease that someone might not yet have so you can look at people who are at high risk of getting a disease as well, so that's another interesting benefit.

So now i'm just going to give you an overview of how some of this biology works in a kind of simplified form. So whenever we talk about genetics we're talking about our DNA and this, out of all the different biological datasets, your DNA is the one thing that shouldn't be changing over the course of your life or varying with the tissue or organ you take it from. It's the same your whole life and it's the same all over your body, whereas everything else varies with time and sort of your tissue that it comes from. So what this is: is just a very long sequence of these "letters", essentially (you might hear them also called "bases" or "nucleotides") and they're sort of, we know them as being enumerated so this might be base number 58 and that will all be sort of catalogued out there in the world of databases. So they're just big linear molecules of 6.4 million bases, I believe, and they're very tightly wound and organized into the form of chromosomes and these sit within the nucleus, which is a sub-compartment of our cells. And a gene - it's not a separate molecule as such, like the other things like proteins, but it's just a region on this long molecule that codes for a specific protein.

OK, so how does DNA work? It can't just sit there in the cell coding for stuff, it has to actually get used. So here is a zoomed-in picture of a cell with DNA in the nucleus. And then whenever we need to make a protein, we have some complex cell machinery that I won't describe that kind of "reads" the gene and uses it to make little copies of the gene which are called RNA - it's very similar but very slightly different to DNA - and so you'll have multiple copies of a protein-specific gene. So this one is C-reactive protein (this is the inflammatory marker that Meena was talking about earlier) and this is called gene expression. And so from each of these RNA molecules you'll get a protein molecule, and this is also part of gene expression. So once you've made a protein, the RNA molecules can just get destroyed - they don't stick around very long - and then the proteins can get transported to wherever they need to go.

So just to give you an overview, proteins are so incredibly diverse it's almost strange to call them under one name! You can have proteins that form part of the large constituent of your muscles or your hair, for example, but then you can have proteins that sit inside your cells, proteins that sit on the outside of your cells and then some that just do their jobs in liquids like blood - like antibodies and hormones, so they do a whole vast range of different things, and each one is coded for by a gene.

So, as I said, the use of DNA to generate RNA and proteins is gene expression. Now gene expression has to be controlled in order for us to function properly. So this cell that's making C-reactive protein is a liver cell, and it's important that this cell "remembers" that it's a liver cell, so that it makes C-reactive protein instead of toenails, for example. So the control of gene expression is incredibly complicated and there's no time to explain it all here, but one of the

mechanisms of control of gene expression is DNA methylation so this is sometimes equated with epigenetics but it's just one mechanism within epigenetics, but it's quite an easy one to measure and quite widely used, so for the purposes of these talks it kind of gets referred to as epigenetics (DNA methylation). So what DNA methylation is is a long lasting chemical modification of DNA, so it doesn't change the sequence of the bases but it's sort of a chemical modification that serves to promote or suppress the use of the genes. So in this sort of cartoon example, you can see that the methylation has caused a reduction in gene expression of C-reactive protein. DNA methylation is partly inherited and partly a reflection of environmental exposures, so the latter is kind of why it's quite interesting to social scientists.

So now I'm just going to go through the different biological datasets that we have and kind of reiterate the ways in which individuals vary.

So here we have two people and, like all of us their DNA only differs by 0.1% So I've zoomed-in here on a small stretch of the DNA and you can see that it's the same sequence for everybody, except for this one position here and this can either have an A or a G, and because we all have two copies of every chromosome (one from Mum and one from Dad) your genotype at this SNP (single nucleotide polymorphism) can either be AA, AG like this person or GG. So, the genetic dataset: we have half a million SNPs tested in nearly 10,000 people. Meena mentioned 14 million, but you can use half a million SNPs to estimate the remaining 13 and a half million by a process called imputation - I won't go into that now, but you tend to co-inherit little blocks of DNA, so you can infer what the nearby variants will be without directly measuring them. So what the data file will be like is: you'll have your SNP ID, so if you look up an rs number out there in the world of biological databases, it will be pointing you to a specific chromosome, a specific position on the chromosome - I actually made this all up, so it probably won't relate to anything in real life. And then, for each of these SNPs, you will have the two variants or alleles as they're sometimes called, that you can commonly expect to find there, and then for each person, they will have two copies, and we will have found out which two they have. But often when we are thinking about SNPs, we actually instead of having two letters will have a count of one of the letters. So we kind of won't really think too much about this column here we'll think about one of the variants and we'll have a count of 0, 1 or 2 for this variant, and when we run our analyses the effect size that we get will correspond to one of the two variants. So, as Meena mentioned, we use data like this to do genome-wide association studies (GWAS). So we test all the half a million / millions of variants for association with whatever the thing is that we're interested in - the outcome - and get our most significant ones, and we can construct a polygenic score. So what that will be is we'll have our effect size or our "weight" for one of the variants and for each person we'll multiply the number of copies they have by the effect size and then aggregate that into a sum that will be a single number that will reflect any of the traits that we've done a GWAS for.

So Understanding Society's genetic data has been used by researchers to generate polygenic scores for a whole range of traits and we are about to start depositing these. So for now we've got polygenic scores for BMI and testosterone and in future we will deposit more. So, for this you just get a single number for each person for each trait. So when it's a disease phenotype (where you can either have it or not, and it's like a dichotomous situation like that), your polygenic score kind of reflects a "liability", like a risk for having that disease, whereas when the trait is a continuous thing like BMI, there's just a sort of rough correlation - so if you have a high polygenic score for body mass index you're more likely to have a higher body mass index, but like Meena said you can't always necessarily predict very accurately from the polygenic score, but there's a sort of rough correlation.

So for our epigenetic data set - this is DNA methylation - so from each individual, you have a sample of blood from which we get white blood cells, and this is where the DNA comes from in a blood sample, and you'll have sort of hundreds, thousands of white blood cells and each of them will contain your full genome, so in the DNA sample you'll have thousands of copies of each person's DNA. And so, for each site along the DNA that is subject to methylation, it either will be or won't be methylated, but because you've got so many copies, you can still get a quantitative measure of methylation because, for example here, this person's got methylation at this first base on every copy of their genome whereas this person's only got methylation at half of the copies of their genome, so this can be expressed as a proportion. And so, in the data file so we've got data for over three and a half thousand participants, 850,000 sites on the DNA, and each of these sites - it's called a CpG - has an ID like the SNPs do, and for each person you just get a proportion of methylation, and 1.0 would represent 100%. So, unlike with genetic studies, pretty much all the other omics, you can test for associations with both exposures and with outcomes. Sometimes we call these "MWAS" - M for methylation instead of E for epigenetics. So, in a similar way to polygenic scores, the outputs of EWAS studies can be used to generate DNA methylation signatures, or scores, so basically the same idea: you multiply for each person their value for the the site you've typed and the effect size from the EWAS and sort of calculate their total burden of, for example, exposure to smoking. So basically a DNA methylation signature for smoking might be: these specific sites in red in your average smoker would be more methylated than your average non-smoker, and the sites in blue they might be less methylated in your average smoker than in your average non-smoker. So some hypothesize that for a given exposure, like smoking, DNA methylation scores might be a more effective and precise measure of exposure, but as Meena mentioned, it can vary - sometimes people find that self-report of how much you've smoked is actually better and more accurate than measuring methylation and some people find it the other way around. So the jury's still out on that one. So one type of methylation signature is the epigenetic clock, which is a score to represent biological aging, or wear-and-tear, and so these broadly correlate with your chronological age so, for example, you can see here we've got five different methods of

calculating your biological age and they're all broadly in the early 40s - they differ slightly - but so this person probably will be 41 or something. So they correlate with your actual age, but some people will have accelerated biological aging, and this can sometimes be predictive of age-related disease and mortality, so it's an important biomarker to sort of see life "getting under the skin" and accelerating your age. So, as I said, there are different methods for calculating biological age, called epigenetic clocks, and we've used five of the best-known to calculate this in all the participants we've got epigenetic data for, and these will soon be deposited for download. Now, each method is based on completely different biology and has its own limitations so I've included the titles of some useful papers that review the performance of these different clocks to help people interpret what they might find.

So as Meena mentioned, we also have the biomarker data set. So all the things we've talked about here can be considered "biomarkers" but the more common use of the word would be things that your doctor might test your blood for that are just well-known to correlate with health outcomes. Because there are some things that you can measure that might not necessarily have a direct relevance to health, that we're still trying to work out, but these are things that are very well-defined, so if you have a high or low value the doctor might be concerned. So we've got 21 molecules coming from plasma, serum or whole blood (the difference between plasma and serum is that serum comes from clotted blood, and plasma comes from unclotted blood). So it includes various measures of things that are known to be relevant to health, and this is - in terms of participants - our biggest biological data set. The number of participants varies a little bit between each biomarker but it's normally 12,000 to 13,000 participants. From the nurse visit we have what can also be classed as "biomarkers" that are not from the blood, so things like blood pressure, lung function etc that tell you something about health. So, just like the other "omics", biomarkers can be tested for their associations with exposures and outcomes and can be sort of aggregated together to represent a phenomenon like inflammation, for example, so this is a list of things - mostly from the blood, but also systolic blood pressure - that have been used in the past to indicate somebody's allostatic load index. Now allostatic load is a sort of important biosocial measure: it's kind of, what it basically means is your stress "getting under the skin" and taking its toll on the body, and it's closely related to the idea of biological age / epigenetic age, and also is linked to inflammation so there's overlapping concepts here that are linked together.

So now, with the new proteomics data set which is coming soon. It's a lot more simple to explain how people vary in the proteins - it's just someone might have more than someone else in their blood, and so the data set is pretty simple - it's just the name of the protein and then a sort of relative measure of the abundance of it in the blood. And so we've got this data for 184 proteins for 6,180 participants. Now proteomics is a slightly younger field than genomics so we're still getting an idea of using sort of protein signatures in the same way as polygenic

scores, and things. But this is happening: someone's developed a protein biomarker for ovarian cancer which comes from 11 different proteins, so we're hoping that there can be more things like this that are useful.

So why might a sociologist be interested in proteins? It's quite a new idea at the moment. So a lot of findings from genome-wide association studies, actually, the SNPs that are relevant are often associated with the level of a protein, so why not measure those directly? And then, if there are any mediating forces that actually affect the level beyond what the genetics do, then you're measuring it directly and not inferring it from genetics. So I've put some little examples of papers that are quite interesting. So in this paper they explored the blood levels of a range of proteins in welders and compared them with non-welders. So, in welding, people are exposed to sort of harmful fumes, and they did find that some proteins were found at different levels in the blood in welders. And, based on what those proteins are known to do, they theorized that some of them signified that the brain was furiously trying to repair itself from the insult of these fumes that were attacking the brain, essentially, so this is kind of obviously a sociological thing because it's about your occupational hazards, so it will be interesting to see if any more occupations are found to be hazardous in a way that you can see in the blood. So also, there has been a study looking at protein biomarkers that associate, in middle and older age groups, with cognitive ability and brain volume, so this is very intriguing, and if these findings can be replicated in younger people, then it could be really important for educational attainment because, obviously, the majority of older people are less likely to be going through education that's going to impact their lives the way younger people are. So there was also a small study done, that would be less well powered, but still I found very interesting. So we know that stress can impact our sort of cardiovascular health, ultimately, and there was a little study that tried to break down some of the pathways in more detail and they looked at things like religious struggles, which they defined as like believing God is punishing you, and they measured levels of proteins in the blood just to build up a sort of pathway from these sort of high level concepts in our minds, to stress, to the blood, to cardiac outcomes - so this was really interesting. And then this last example, it's actually not about blood levels of proteins, but CD38 is one of the proteins we have data for at Understanding Society, and when I was researching it for our User Guide I saw that it was related to oxytocin, which you may or may not know is the sort of "social hormone" that's involved with love, and sort of behavioral things, so it'd be interesting to see if blood levels of that have anything to do with social behaviours and outcomes.