Multi-omics

Anna Dearman: "Multi-omics". So you may have been wondering what the overlap is between the Understanding Society biological datasets. So, in terms of number of participants, the two biggest datasets are biomarkers and the genomics. So, as I said, data availability varies a little bit by biomarker but there's normally 12,000 to 13,000 people for each one. So, the overlap for all four of the datasets is over 1,600 people. This is sort of a heat map, so the redder the zone, the more - the higher the number is, So for our new proteomics data, the majority of it, so around 4,700 have genomic data as well. And you can see the numbers that overlap for yourself there. And one interesting thing to bear in mind is that, for each person, the measures of all these things come from the same blood collection. So I mentioned earlier that everything except your genomics will vary depending on the point at which the sample is taken, so that's a nice thing to know that you can make certain assumptions, based on the fact that the these were all collected in the same blood collection.

Right, so how to integrate the different omics. So first I'm going to talk about how integrating genomic data sets with other types of data allows you to infer causality if you design your study well. So genetic variants are not subject to reverse causation like all the other omics and, provided that you choose genetic variants that meet a certain set of criteria (that I won't go into right now), they can be used as instrumental variables in Mendelian Randomization studies. So Mendelian Randomization is used when we want to estimate the effect of and exposure on an outcome but we really can't just because there are too many confounding variables that we can't control for. So instead of using a measure of the exposure itself on the outcome, we make a genetic proxy for that exposure and investigate this pathway here. If you are interested in all the detail around this methodology, I recommend these two papers that talk about how you design a study well and the limitations and things, and the different variations on it.

So I'm now just going to walk you through one example that I read about where they used Mendelian Randomization to investigate the role of blood cholesterol on dementia risk. Now, just to mention, in this world of biosocial research, an exposure, while you might think of an exposure as like pollution or the way you're treated or smoking, your own blood can be an exposure, so in this paradigm here your blood cholesterol is the exposure - just to mention that before I get started. So cholesterol-lowering drugs such as statins are prescribed to reduce blood cholesterol (LDL cholesterol) and when the authors were writing this paper, they were concerned about the potential risk of these drugs on dementia, that there was some evidence for. And they also could rationalize this because cholesterol is part of the protective substance called myelin that protects your brain, so they had theories as to how this could be working. So they wondered whether it was via this pathway that statins were reducing cholesterol, and then that was having a knock-on effect on dementia risk. But unfortunately they couldn't really test this because there are too many confounding variables that, separately, had an impact on blood cholesterol and dementia risk. But so they knew that there were genetic variants that also had a role in blood cholesterol. And so what

they did was they used a few different ways of calculating somebody's genetic propensity towards lower blood cholesterol and tested whether that was associated with the risk of dementia, so they used a few different ways to measure this genetic liability and a few different ways to measure dementia - different outcomes. And what they actually found was no association, except for one finding where one of the polygenic scores for low blood cholesterol actually had the opposite effect, specifically in Alzheimer's disease. So, in order to make sure that the genetic factors constituted a valid instrumental variable the researchers had to rule out any arrows going this way, so rule out any association between the genetics that they were using and factors such as age, sex, hypertension, smoking, physical activity, alcohol consumption, education level and menopause, all of which can affect both blood cholesterol and risk of dementia, and checking that the genetic factors they were using didn't associate with those things, that meant that they could use it as a an instrument in this study.

So while we're on the topic of genetic factors, I just want to talk a little bit about genetics, before moving on to more examples of integrated omics. So, speaking of potential overlapping associations like that, one thing to bear in mind is that DNA indirectly affects our behaviour, and our behaviour affects our exposures. So, for example, we know that there are genetic associations with risk-taking behaviours and smoking and sexual behaviours and things like that, so there is the complication of our genetics being related to our exposures, so it's important when designing Mendelian Randomization to check for those potential, confounding issues. And this is called gene-environment correlation. So here we see a correlation heatmap that shows just the genetic overlaps between things like behavioural traits, psychiatric disorders, cognitive ability and other things. And it's interesting if we zoom in on one thing, so ADHD here, it's interesting that it has some negative correlations with IQ - genetically speaking, of course - and income and things, and has some positive correlations with behavioural traits like risk-taking, sexual and substance behaviours, but also with depression and loneliness so it just raises so many interesting questions, and our challenge as scientists is to work out what these associations mean and what are the causal pathways involved. Now, techniques have been developed to take a list of genetic factors that contribute to a trait such as ADHD which, in this image is the first column here, and, from that list of genetic factors, subtract the ones that overlap with something else, so in this example socioeconomic status. So this axis here is kind of the amount of variance in ADHD status that can be explained by genetics, so when you just take the results of your GWAS, this is how much variance is explained, and when you subtract the genetic factors that overlap with socioeconomic status, the amount that genetics has a role in ADHD drops quite significantly and, interestingly, we see that also for smoking initiation and age at which smoking starts. So since socioeconomic status is arguably something largely imposed on you by society and not necessarily a fully inherited biological trait, we ask ourselves, does this suggests that ADHD is maybe less heritable than it first appears because some of the genetic factors are more, might be acting via socioeconomic status, than directly on their own brain or something. It's also interesting to note - sorry, I've written a note to mention the smoking ones, but I already did that.

So, going back to integrating the different omics (sorry, I'm just trying to move my Zoom toolbar) sometimes your dataset doesn't include all the variables you're interested in, so Meena mentioned earlier, that in Understanding Society, the nurses took bloods in people's households and then

posted it to the lab to process, so by the time the blood gets processed in the lab, the RNA in the blood will just be destroyed so there's no feasible way for us to have RNA data, if you remember the diagram from earlier, we have DNA to RNA to protein. But, in these situations it can be useful to include summary-level data from previous studies in your Mendelian Randomization study. So there are databases out there of SNPs (so the genetic variants) that associate with levels of RNA, levels of DNA methylation, protein levels and a range of other quantitative outcomes. So these SNPs that correlate with something quantitative are called Quantitative Trait Loci (QTLs). So, here are some of the databases where you can look up these associations. I had a go at looking up this SNP here and three of the top hits I got were that this SNP correlates with platelet volume so that's something in your blood that helps you clot and form scabs and things, so this SNP correlates with how big those are, it also correlates with inflammatory bowel disease and income. So, yeah, there's so many interesting findings out there, that you can integrate into your research.

So now I'm going to give you an overview of a few examples of integrated omics to describe how we can use the data and hopefully to encourage you, as viewers, to come up with some new interesting questions using these as like a framework, maybe. So this study explored how our DNA affects the levels of 92 different proteins in the blood, so DNA and methylated DNA, and they also use these quantitative trait loci databases, and Mendelian Randomization, to explore the causal relationships between DNA methylation, RNA levels and proteins. One of the 92 proteins was already suspected to have a role in Alzheimer's disease, so they kind of zoomed in on this one a bit and explored it in a bit more detail to try and understand the pathway. I'll go over it all in a couple of slides. There were lots of analyses, so I won't describe all of them. So the main findings - they were primarily looking at the genetic and epigenetic factors associated with blood levels of these proteins. So that's already three "omics": genomics, epigenomics and proteomics. So they found 41 SNPs were linked to concentrations of 33 proteins, 26 methylation sites were linked to concentrations of nine proteins. And they looked up the genes that were linked to these methylation sites and explored the biological pathways that those genes were involved in, to sort of get a feel for what processes might be going on to do with these proteins. They also performed Mendelian Randomization to look at the causal relationships here, so they, so the RNA omics is called transcriptomics, so they were able to incorporate that fourth omic, and they found that, for some of the proteins, RNA levels were known to cause the protein levels; for other proteins it worked the other way around, and that there was a reciprocal relationship between methylation levels and protein levels for for these proteins. But that's all very biological and how does it help us understand outcomes that we're interested in? So, as I said, they went into more detail for one of the proteins. So the poliovirus receptor gene was previously known to be related to Alzheimer's disease, so I completely made up this cartoon version of the poliovirus receptor gene and I made up some SNPs - I don't know whereabouts the SNPs really are in relation to the gene, but you'll get SNPs that are close to a gene before and after it, and in the middle of it - that's just a rough example. So, as I said, they knew that this gene was implicated in Alzheimer's, and they found that a SNP near the gene was linked to levels of the protein itself, so they wanted to see whether the same SNP was driving the protein levels and Alzheimer's disease, because if so it might indicate that it was the same - that the protein was involved in a pathway to the disease. They found that it's actually two different SNPs, but then they also explored whether protein levels seem to be causing Alzheimer's disease risk or whether perhaps Alzheimer's disease resulted in higher levels of the protein by some other mechanism, and

they found that it was more in the direction of the protein being a certain level contributes to Alzheimer's disease.

So, using the previous example as a model, can you think of any other outcomes, where we already know there's a genetic component, but we might want to explore whether there is a pathway from genes, to protein levels, to outcome. This might be especially important if the protein level can be modified by an intervention like medicine or exercise. So that's just one example for you think about.

And then, what about the earlier example of cholesterol and dementia, where confounding variables were making it impossible to look at causality, so that a genetic proxy was used to represent the exposure. Is there any other examples that we can think of that would fit into a model like this?

And so here's an example of integrating behavioural and biological data to explore disease. Now, there's a lot of arrows going on here, so we'll just go through them slowly. So it was already known that cardiovascular disease, could be affected by genetics, blood lipids and smoking behaviour. We also know that genetics could influence smoking behaviour. blood lipids and health outcomes, and we also knew that smoking could influence blood lipids and it could influence cardiovascular disease. But what this study revealed was that some of the genetic associations here, between - so the associations between genetics and blood lipids - were only present in smokers, or were only present in non-smokers, but not both. So it teased apart genetic associations that were lost or gained when you look at smokers and non-smokers separately, and this shows how our behaviours can interact with our genetics, to increase our risk factors for disease.

And then there's just one more example - it has a slightly less confusing flowchart. So another way to integrate biological and sociological variables is to use a machine learning approach, so in this example the researcher was interested in looking at the ability to predict limiting long term illness after one or five years, and he fed in genetic factors, a list of biomarkers and some other variables. In this example it just turned out that age was quite good at predictin,g and not much else, but you can always adapt this at every level, so you can choose different inputs, different machine learning approaches and you can measure different outcomes, and hopefully this kind of approach will yield good predictive models for a range of outcomes.