This Week in Genomics
9/5/2023

Addressing Diversity in Genomics: Mexican Biobank and BioBank Japan

https://www.insideprecisionmedicine.com/news-and-features/addressing-diversity-in-genomics-mexican-biobank-and-biobank-japan/

Since the completion of the human genome project, it has been increasingly attractive for researchers to ***set up biobanks to collect a range of biological samples and carry out large scale genetic and genomic analyses.***

Relatively rare 20 years ago, there are now estimated to be more than ***120 biobanks around the world***. These range in size from having ***hundreds to millions of participants*** and ***vary in their scope***, with some focused on ***keeping a broad epidemiological record and others targeting certain disease groups.***

Research carried out using these biobank cohorts is becoming ***increasingly important*** as we move more into an ***age of preventive precision medicine***. For example, ***genetic testing to predict risk*** of developing diseases like ***cancer*** is becoming ever ***more common***. However, it is important to consider the ***source and the quality of the data*** before making widespread predictions. Most genetic tests currently being used in clinics are built on ***genetic data that is largely of white-European origin***, which means the ***value of these tests*** for people from other population groups can be ***questionable***.

This is ***beginning to change***, however, and biobanks in different parts of the world are helping to make this happen. At the recent ***European Society of Human Genetics*** conference in Glasgow, Inside Precision Medicine senior editor Helen Albert spoke to ***Mashaal Sohail***, an associate professor and group leader at the National Autonomous University of Mexico, about her experience working on the Mexican Biobank project, and to ***Yukinori Okada***, a professor at the University of Tokyo and Osaka University and a team leader at RIKEN Center for Integrative Medical Sciences, about his experience developing BioBank Japan.

Mashall Sohail – Mexican Biobank. Associate Professor, National Autonomous University of Mexico

How did you become involved in the Mexican Biobank project?

I did my PhD at Harvard University. While there, I got to do work in a lot of exciting projects in human genetics, and ***I trained in population genetics*** in particular. I was involved, for example, in the Genome of the Netherlands project. Towards the end of my PhD, I also focused on the history of science. I was looking to see how science could be done in a different context. I had also been seeing talks and becoming aware of the diversity issue in genomics and starting to think about that.

It turned out that when I was looking for a postdoc that Andres Moreno Estrada was looking for a postdoc for the Mexican Biobank project. And it just seemed like a really exciting opportunity for me that fulfilled all these different fronts. I went to the interview and I just really liked everything that I saw, and the role that I could play in particular, in doing this. And so that's how it started.

What are the goals of the Mexican Biobank project and how did it start?

The project started as a collaboration between the National Institute of Public Health and the National Laboratory of Genomics for Biodiversity (Langebio, Cinvestav) in Mexico, which is where I did my postdoc in Andres Moreno Estrada's group. ***The National Institute of Public Health collected samples as part of the year 2000 survey from across Mexico. They have a biobank of 40,000 individuals.*** Andres Moreno, and the team that he put together, saw the opportunity in that because data was already collected, which is the difficult part.

They wrote a grant to get funding to genotype the samples for the ***kind of analysis*** that I led, which includes ***ancestry and what ancestry tells us about demographic histories and subsequent impact on genetic variation relevant for complex traits.*** That's how it started and got funded, and then I came on as a postdoc.

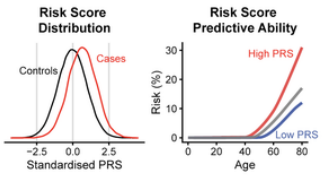
We're at phase one: we have ***genotyped about 6,000 individuals*** and ***carried out whole genome sequencing for 50 individuals***. We hope to ***expand in future***, but it all ***depends on funding***. We don't have ability to recall participants in this case, but they had originally signed consent to allow the kind of analysis that we did.

Why is it important to make sure we have biobanks in lots of different places around the world?

I actually think that it's a nuanced question that I got to think very hard about over this project. In general, what we're understanding from ***genome wide association studies (GWAS)*** is that possibly the ***genetic architecture is actually very shared across different ancestries, in terms of what drug targets or what genes may be involved in a trait.***

But when it comes to thinking about ***polygenic risk***, being able to predict polygenic predisposition to a disease, that's where it becomes really relevant—what ***cohort*** you use for ***discovery*** and then what ***cohort*** you use to be able to ***make those polygenic scores***.

Polygenic risk scores can provide a measure of your disease risk due to your genes. Combining polygenic risk scores with other factors that affect disease risk can give a better idea of how likely you are to get a specific disease than considering either alone. (**from: Center for Disease Control and Prevention**)



It's very relevant to get ***cohorts that are as similar as possible to the cohort you want to check predisposition*** in. It needs to be similar, not only in genetic ancestry, but also in various other environmental factors. For example, in our association studies, we ***didn't find any novel genetic variants*** associated with any of the diseases or complex traits we looked at ***compared to GWAS's already done in Europe,*** but we did see ***better prediction*** for some traits with our GWAS compared to the U.K. Biobank GWAS, which is ***orders of magnitude larger***.

Also, for the Mexican Biobank, all the genotyping and all the analyses were done in Mexico. It's important to build that local capacity … ***it is really key to make this kind of project sustainable in different economies in different regions.***

What are the key findings or outcomes so far from the Mexican Biobank?

From this first phase, we are able to replicate some older findings, in particular, ***create a genetic map of the ancestries that are found in Mexico.*** We were able to replicate the fourth ancestry that's relevant for Mexico, which is ***Asian ancestry on top of the European, Indigenous, and African ancestries***. We were able to really look at that in ***every state and show the fine scale variation that exists***, then we were able to infer population size histories in different Mesoamerican regions. This is coming on top of previous work done by other groups as well, but it is the ***first time*** we're able to do it at a ***nationwide scale***.

We were able to show ***founder effects*** that vary in ***different regions of the country*** based on different ***civilizational histories*** in different regions of Mexico and highlight the ***heterogeneity that exists within a single country***. Then we were able to show how that ***impacts genetic variation in terms of runs of homozygosity***.

Founder effect **-**the reduced genetic diversity which results when a population is descended from a small number of colonizing ancestors. Effect strongest in small populations ***due to genetic drift***.

We also showed that the number of rare genetic variants that an individual carries is ***correlated with their ancestry profile***, as well. Then we were able to replicate various GWAS findings in the biobank, which is 6,000 individuals so far, and make polygenic risk scores, which I think is very relevant for future work. We were also able to ***compute polygenic scores across a range of 10 traits and diseases.***

Lastly, we did an analysis to jointly ***model genetic and environmental variables*** to look at which variables are ***most relevant for a given trait—it could be height, or triglyceride level***, for example. We were able to infer what we are calling a ***"trait profile"*** for a range of traits, showing if it's ***most relevant*** whether you live ***in an urban or rural environment***, or if your ***genetic ancestry is more relevant***. For example, for ***glucose and triglycerides***, we find that it is really relevant ***what proportion of your genome has indigenous ancestries from the Americas***.

What is next for the Mexican Biobank and your research?

We're looking at the role of ***mitochondrial genetic variants***, which have never been looked at, at large scale, in Mexico. We're looking at archaic introgression in the Mexican biobank, such as ***Neanderthal and Denisovan*** ancestry, and making a map of that and looking at how this may be relevant for complex trait variation.

We're really looking carefully at how to do polygenic prediction best by comparing different methods that assume different levels of sharedness of genetic architecture among the different ancestries that are found in the Mexican biobank.

There's also an initiative to create a ***meta-analysis with existing cohorts of Mexican individuals in the U.S.*** and to do a larger, ***more well powered GWAS, especially for the metabolic traits*** that we have. Those are just some of the projects that we're working on.

Yukinori Okada – BioBank Japan; professor at the University of Tokyo and Osaka University and team leader at RIKEN Center for Integrative Medical Sciences

BioBank Japan—Yukinori Okada

How did you become involved in the BioBank Japan project?

I have been closely working with BioBank Japan for around 15 years. When I was a resident physician, I decided to become a researcher instead and I was looking for which scientific fields were going to be interesting. I saw a paper on the International HapMap Project, a completely new concept. I was confident when I saw the paper that human genome projects, especially human genome projects that focused on the difference of sequence across different people, should be the next wave of science. I looked across Japan and at this time, BioBank Japan was the place that had a certain amount of human genomics going on, so I decided to join.

What are the goals of BioBank Japan and how has it developed over time?

At this time, ***BioBank Japan has 20 years of history and now has around 260,000 participants***. I think it's probably the largest non-European biobank. It started in the early stages of human genomics and has been leading the human genetic resource in studies in East Asia for the last 20 years. For example, 15 years ago the methodology of GWAS had not been validated, the number of samples was much smaller, and there were not so many people with relevant computational knowledge.

BioBank Japan itself is (now) a typical design, a ***hospital-based, disease- and patient-oriented biobank***, which ***collects DNA,*** but also other ***clinical information*** … it was not so common 20 years ago. One of the people who designed and launched the biobank, Professor Yusuke Nakamura, knew what was necessary for the future of human biobanks and he made a very nice design. BioBank Japan has ***focused on around 50 target diseases***, including ***cancers, immune diseases, cardiovascular disease, and diabetes***, and it now has high statistical power to analyze genetics of such diseases.

Why is it important to make sure we have biobanks in lots of different places around the world?

Diversity sometimes means ***ethnicity***, but diversity also refers to ***human sequence variations***. Having biobank resources across many regions worldwide can ***maximize the power*** to assess differences in mutations and phenotype. Disease itself is also more population specific than we usually think.

Of course, ***common diseases*** occur in all populations, but some ***disease biology is different***. For example, some types of diabetes in East Asian people have different biology and different outcomes to in ***European people***. It is not a different disease, it is the ***same disease***, but the proportion of the ***subtypes is different***. There are also some cultural and environment factors. For example, the ***South Asian population has more consanguinity***. If consanguinity increases, then ***genetic diversity decreases***, but the people who are homozygous for ***loss of function variants increases***. So, many things differentiate environmental phenotype so it is especially important to provide such resources across the world.

Polygenic risk scores are an important piece to implement personalized medicine into society. I think they are informative to predict disease, but of course different types of harmonization or diversity coverage is required … when you return to genetics you do the GWAS and find genetic markers linked to disease risk. It is simple, but including genetic backgrounds from across global populations is very important.

What is next for BioBank Japan and your research?

We've had a recent focus on the ***Global Biobank Meta-analysis*** Initiative. It's a nice project. Even though there are so ***many biobanks***, which in total ***covers global diversity***, unless they are unified, or ***coordinated***, it means then the ***diversity is not finely deflected***. The point is how biobanks, locally or globally, coordinate in the same direction and make good collaborations. I think it's a nice approach to coordinate the global biobanks together for some specific project and research efforts.

We've also been helping people in other projects to analyze data. Now, ***most analytics is open source***, so actually, people can easily do it by themselves using knowledge on the Web, but 15 years before, there was no software, no guidance on how to do that analysis. Each of us needed to write our own scripts to do a GWAS. It's funny that many people in Japan, even people I never knew, use my scripts.

Now things have changed, because there's so many people in the biobank field. I am shifting to more human omics integration. Increasing our sample size is somehow saturated. To ***get much more exciting results***, just based on sample size increases, the ***s***. Our strategy is how to expand the multi-modality with a finite resource. It means probably using the ***proteome, or metabolome***, somehow to expand the longitudinal follow up the clinical information.

My feeling for this area is that it continues to change rapidly. I think no one imagined the situation now 50 years before. Just following such dramatic changes is really fun. I still have some years before my retirement so I want to keep going and following this interesting field.