POPULATION MATTERS: BIOBANKS ACCELERATE GENO-PHENO DISCOVERIES

Sequencing biological samples from people with a specific disease or who belong to a particular ethnic group is improving understanding of how genetic variants influence common diseases,aiding prevention, diagnosis and treatment.

For illumina® by natureresearch CUSTOM MEDIA

POPULATION MATTERS: BIOBANKS ACCELERATE GENO-PHENO DISCOVERIES

Sequencing biological samples from people with a specific disease or who belong to a particular ethnic group is improving understanding of HOW GENETIC VARIANTS INFLUENCE COMMON DISEASES, aiding prevention, diagnosis and treatment.

“Biobanks are crucial for understanding how genetic factors are related to disease outcomes,” says Naomi Allen, an epidemiologist at Nuffield Department of Population Health, University of Oxford, and chief scientist for UK Biobank. “By linking samples to individuals’ electronic medical records, and following them over time, we can correlate genetic profiles with their health outcomes over the course of their lifetime.”

There are more than 120 biobanks worldwide, having evolved over the past 30 years. They range from small, predominantly university-based repositories, to large, government-supported resources. As well as collecting and storing samples, they also provide clinical, pathological, molecular and radiological information for research into personalized medicine.

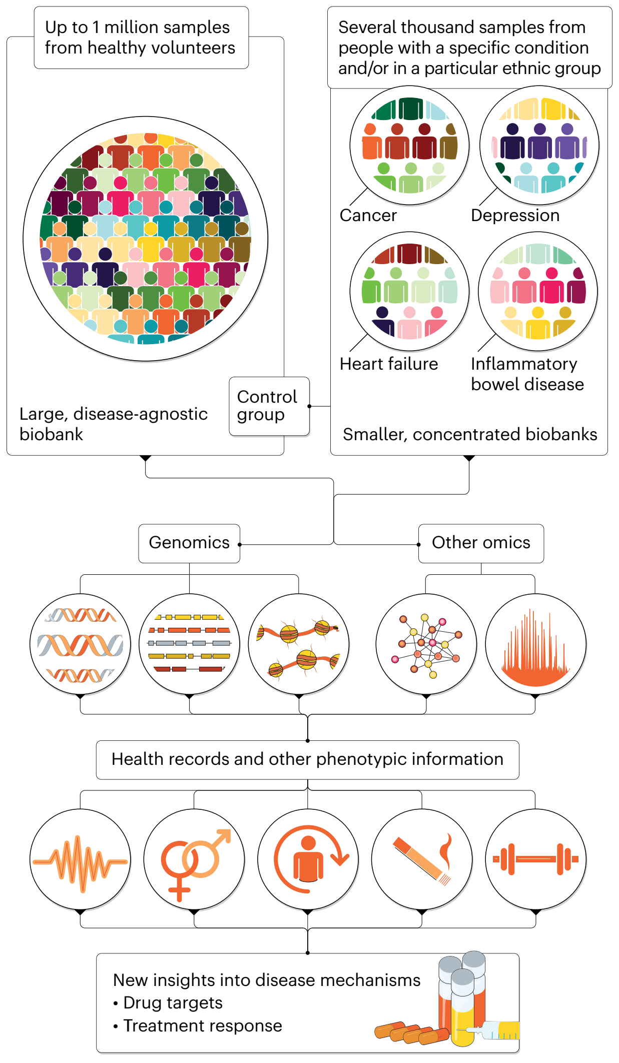
Biobanks allow researchers to explore the causes of disease by helping them link genotypes to phenotypes. This is a process that has been underway for years through genome-wide association studies (GWAS), but it has proved far from straightforward. “What we have learned from GWAS over the last 15 to 20 years is that there are many variants, they have small effect sizes and, even if you total most of the common variants throughout the genome, they account for only a small fraction of phenotypic variance,” says Judy Cho, a translational geneticist at Icahn School of Medicine at Mount Sinai, New York.

Biobanks are speeding up progress. They allow researchers to easily analyse increasing numbers of biological samples and associated clinical data to characterize disease mechanisms, find novel drug targets, and identify patients most likely to benefit from a particular treatment approach. “Embedding genomic information in electronic health-care records, so it can be used through the course of life, is an appealing vision,” says Dan Roden, a clinical pharmacologist and Director of the BioVU Biobank at Vanderbilt University Medical Center, Tennessee. “But it’s one that is hard to realise; there are lots of logistical problems in creating an infrastructure like that.” BioVU’s step towards achieving this vision is to store DNA extracted from discarded blood, collected during routine clinical testing, linked to de-identified medical records. Around 250,000 DNA samples are now available for Vanderbilt investigators.

But going big isn’t the only solution: insights into common diseases can also be found by analysing smaller disease- and/or ethnic-specific cohorts, which concentrate on important genomic signals. Advances in genetic technologies and increased efforts to capture genetic diversity in biobanks are helping researchers to make robust geno-pheno associations in a cost-effective manner (see ‘Biobanks and geno-pheno associations’).

BIOBANKS AND GENO-PHENO ASSOCIATIONS

Small, ethnically diverse, disease-specific biobanks can reveal how genetic factors are related to disease using fewer people than big, disease-agnostic biobanks. The big biobanks can also provide a control cohort for these smaller ones.


Up to 1 million samples from healthy volunteers
Several thousand samples from people with a specific condition and/or in a particular ethnic group
Cancer
Depression
Heart failure
Inflammatory bowel disease
Large, disease-agnostic biobank
Control group
Smaller, concentrated biobanks
Genomics
Other omics
Health records and other phenotypic information
New insights into disease mechanisms
• Drug targets
• Treatment response

“FURTHER UNDERSTANDING GENETIC DIVERSITY IS GOING TO BE REALLY ILLUMINATING.”

Improving biobanking efforts

The UK Biobank is one of the world’s largest, storing samples from 500,000 people, recruited between 2006 and 2010, linked to a range of electronic health records. In 2019 it received funding to perform high-coverage (X30) whole-genome sequencing for all samples, with the first 200,000 available in 2021. “Because you need large sample sizes to see small genetic effects, many studies have had to piece together data from different places,” Allen says. “The beauty about UK Biobank is that it provides standardized, high-quality genetic and health outcome data for half a million people in one dataset.”

Samples from the UK Biobank have helped to establish the genetic and socio-demographic risk factors for COVID-19[1] and are helping to identify variants in immune system genes that might help predict the outcome of hospitalized patients with the disease.[2]

However, there are drawbacks to the large biobanks. Volunteers tend to be healthier and wealthier than the national average, and don’t represent the diversity of the population. In the UK Biobank, 95% of samples are from white people. “It is not an especially good cohort for studying ethnic differences,” Allen concedes. “There is a need for complementary studies to be carried out in Black and minority ethnic groups to compare genetic differences and how they affect health outcomes.”

Cho is part of the US National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Genetics Consortium, working to define the genetic factors underlying susceptibility to inflammatory bowel disease (IBD) in African Americans and Ashkenazi Jews.[3,4] “Current genetic risk scores do not predict well across populations,” she says. “Further understanding genetic diversity is going to be really illuminating.”

To improve diversity, biobanks are engaging with underrepresented populations. Cho, who also manages the BioMe Biobank, with more than 50,000 samples, says she is making it a priority to increase samples from African-Americans. “It is right for social justice reasons, and it is right for scientific reasons.”

Towards precision medicine for everyone

Erin Dunn, a psychiatric epidemiologist with expertise in genetics at Harvard Medical School, says that diversifying biobanks is also a good way to engage minority populations in health research. “Without including all population groups in genetic research, we lack an understanding of whether and how genetic variants in one population generalize to another.”

Using data from the Hispanic Community Health Study/Study of Latinos, Dunn and colleagues conducted the largest GWAS of depressive symptoms in Hispanics/Latinos to date.[5] They found different genetic variants associated with depressive symptoms from those reported in studies of people with European ancestry. "This is a start, but there is still a lot of work to do to piece together what's similar and what's different," Dunn says.

In collaboration with researchers at Penn University, Cho is also finding health implications in genetic differences among people of different ancestries. Using Illumina's Global Sequencing Array to screen more than 9,000 individuals of African and Latin ancestry, enrolled in either the Penn Medicine or the BioMe biobank, they found a significant association between a variant in the transthyretin gene (TTR) and a treatable form of heart failure.[6] These findings suggest that TTR testing could identify at-risk people from these ethnic groups, leading to earlier diagnosis and, importantly, treatment.

A cost-effective way to assess selective genetic variants is to use biobank data as the 'control' group, and do pooled analyses using samples that have undergone similar genetic analyses. "Researchers often use the UK Biobank data as a general population control group to compare against their own 'case' cohort of patients," explains Allen.

As the amount of data held in biobanks increases to petabyte levels, mostly driven by genomics, it becomes harder to sustain a data download model. Several biobanks are developing cloud-based platforms to reduce need for computer resources. "We want to democratize access to the data, particularly for low- and middle-income countries, so that any researcher in the world can use it," says Allen.

The first large biorepository to implement a cloud-based model is the US All of Us initiative that is developing a cohort of one million participants of diverse ancestries. "The All of Us dataset now includes 250,000 subjects whose data are available to researchers through Researcher Workbench," says Roden, who is co-director of the All of Us Data and Research Centre.

"THE FIELD IS SLOWLY PIVOTING FROM A DISCOVERY PHASE TO AN IMPLEMENTATION PHASE."

Powering GWAS using well-characterized samples

Because the large biobanks don't specifically recruit people with certain illnesses, they are often underpowered for GWAS analysis of diseases. One way to identify significant genotype-phenotype associations is by using smaller cohorts with specific characteristics. "By analysing samples from patients who have a particular disease, have undergone the same type of surgery or had a similar response to a drug, you decrease the number of confounding factors and increase the probability of finding something important," says Anita Grigoriadis, non-clinical deputy for the Breast Cancer Now unit and a cancer bioinformatician at King's College London.

Her team is studying the biological characteristics of triple-negative breast cancer patients to identify distinctive features and guide the development of targeted treatments. They use the breast cancer biobank at Guy's and St Thomas' Hospital, which has been collecting tumour, blood and other tissue samples from patients since the 1970s. "We were able to identify potential drug targets and biomarkers for a high-risk group of breast cancers patients through exploring a rich resource of biobanked triple-negative breast cancers with extensive long-term clinical and pathological follow-up information," she says.

Furthering genotype-phenotype associations

Linking variants with disease is just a first step. "Once these genetic variants are identified, we can then start to determine whether they are truly the causal ones and what types of biological processes they might be linked to," says Dunn.

To disentangle the role of genetics, researchers are analysing gene expression patterns (transcriptomics) and how they are regulated (epigenomics). Further understanding the proteins that these genes encode (proteomics) and their activity (metabolomics) helps piece together complex disease processes and the influence of life experiences and environmental factors. "Delving more deeply into individuals' proteomic and metabolomic profiles is likely to tell us a lot about the mechanisms through which things like diet influence health," Allen points out.

These multi-omic analyses help disentangle the flow of information that underlies disease. "Multi-omics can help us understand every step of the process converting genotype to phenotype," says Cho. She also highlights the importance of conducting single cell sequencing (sc-Seq) to measure transcriptome-wide gene expression at single-cell resolution. Such analysis is crucial to determining the cell types and states in which causal variants operate, and could lead to patient-tailored therapies.[7] "Using sc-Seq we found a treatment-refractory signature in patients with Crohn's disease that will help determine who will and won't respond to anti-TNF therapy," she adds.

Thanks to the expansion of biobanks, capturing diversity in both populations and linked data types, and to technological advances allowing detailed and robust analyses, it is becoming possible to study the genetics underlying any common disease or trait. Many diseases that are currently underdiagnosed could be prevented, or treated earlier or more effectively, by integrating whole-genome sequencing and deep phenotyping into clinical assessments.

"The field is slowly pivoting from a discovery phase to an implementation phase," says Roden. "Translating the massive amounts of information we can produce at a feasible cost into the ordinary course of healthcare is not easy."

But it is a challenge that he and others are keen to tackle. There is little doubt that the delivery of actionable results will allow the optimization of prevention and treatment strategies. "Even little steps," says Cho, "can make a big difference to patients' lives." ■

REFERENCES

1. Niedzwiedz, C.L., *et al. BMC Med* **18**, 160 (2020).
2. https://www.ukbiobank.ac.uk/2020/05/genome-wide-association-studies-for-covid-19-and-its-outcomes/
3. Brant, S.R., *et al. Gastroenterology* **152**(1):206-217.e2 (2017).
4. Rivas, M.A. *et al. PLoS Genet.* **14**(5):e1007329 (2018).
5. Dunn, E.C., *et al. J Psych Res* **99**: 167-176 (2018).
6. Damrauer, S.M. *et al. JAMA* **322**(22):2191-2202 (2019).
7. Martin, J.C. *et al. Cell* **178**(6): 1493-1508.e20 (2019).

illumina®