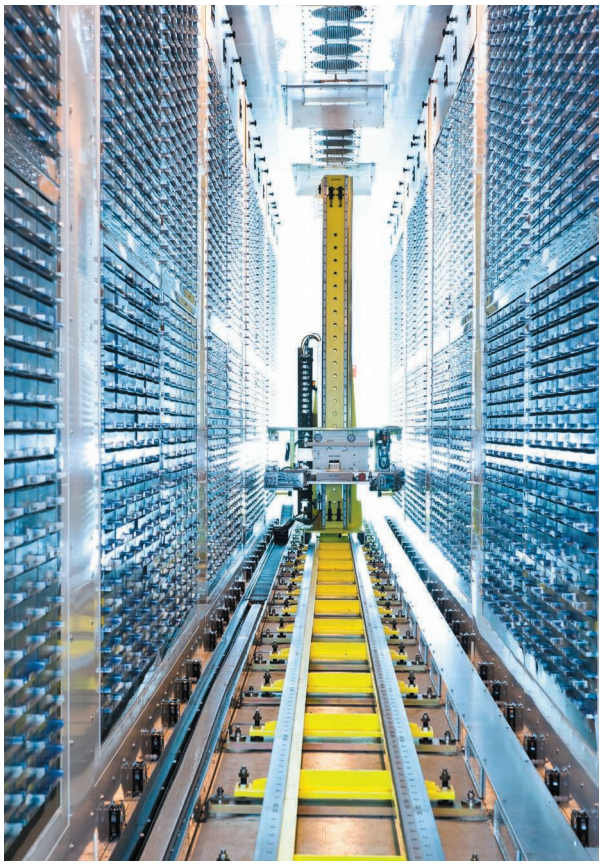
HUMAN GENOMICS

Biobank for the masses

UK Biobank contains a wealth of data on genetics, health and more from 500,000 participants. A detailed overview of the biobank and an analysis of its brain-imaging data show the value of this resource. SEE ARTICLES P.203 & P.210

NANCY COX

Huge sample sizes are often needed to discover the genetic variants that contribute to disease. Meta-analyses of many genome-wide association studies (GWAS), which test for such links, are now beginning to search for associations between DNA variants and common diseases in more than one million individuals[1]. But perhaps equally important is detailed clinical and biological information about the participants, which enables researchers to better test for more associations — including those that give insight into disease mechanisms. Writing in *Nature*, Bycroft *et al.*[2] and Elliott *et al.*[3] describe a huge resource called UK Biobank that marries large-scale genomic and detailed clinical data for 500,000 people. The biobank promises to aid the discovery of relationships between genome variation and common human diseases, and to improve our understanding of the mechanisms that underlie those associations.



WELLCOME IMAGES

Figure 1 | Biological samples in a storage freezer at UK Biobank. Two papers[2,3] describe the set-up of the biobank and analyse some of its data.

As Bycroft *et al.* describe, UK Biobank's 500,000 participants donated urine, saliva and blood samples (Fig. 1), which were used for genetic analysis and evaluated for known biomarkers of disease. The participants were aged between 40 and 69 when they were recruited to the study between 2006 and 2010. This age range meant that participants would be at risk of developing common diseases of adulthood. The volunteers filled out thorough questionnaires about a wide range of factors, including family disease history, demographic background and lifestyle. They also gave consent for researchers to access electronic health-record data. Subsets of participants underwent more-comprehensive examinations, including extensive imaging and lung-function studies.

The sample size of this resource, combined with the breadth of data that have been collected — and that will continue to accrue as the participants age — is unprecedented, as is the generosity of the project's data-sharing plan. From the beginning, the intent has been to share the data in their entirety with any health researcher. As a consequence, thousands of scientists from all over the world have been doing research on these data since July 2017.

In 2007, the Wellcome Trust Case Control Consortium published a landmark study[4] that set the standard for how GWAS should be performed and the resulting data shared, greatly influencing how GWAS were conducted. Similarly, Bycroft *et al.* provide a wealth of detail on how they designed their study and analysed the resulting genetic data. As such, their paper promises to influence a new generation of data scientists.

The work is a vivid reminder that data generation is perhaps the least challenging aspect of big-data science. The researchers used an array-based approach to determine nucleotide variation at more than 800,000 genomic sites, and then imputed variation at millions more sites. But the scale of the data meant that both the design of this 'genotyping' and the subsequent quality-control analysis needed to be wholly reconceived from methods used for smaller studies. Moreover, much of the software used needed to be substantially revised to achieve reasonable computing times. Software is being made available to scientists, along with the full results of the authors' preliminary GWAS and phenome-wide association studies, the latter of which analyse associations between the entire range of traits logged in the biobank and a single genetic variant.

Bycroft *et al.* conducted several analyses to demonstrate that the data they collected would yield appropriate results in association studies. For example, they analysed a genomic region that harbours several human leukocyte antigen (HLA) genes, which have a role in distinguishing foreign cells and particles from those of our own bodies. It is well established that many variants in these genes are associated with common diseases[5]. The authors confirmed that the HLA types imputed from their genotype data have the expected associations with disease, validating both the genotype and disease data used in the study. The group also performed GWAS to identify genetic variants associated with differences in height — again, their results matched those from GWAS meta-analyses that used independent samples.

Whereas Bycroft and colleagues detail how the biobank's genome data were generated and highlight the quality of the data, Elliott *et al.*[3] give us a preview of how these data can be used to drive discovery and to probe the mechanisms underlying genetic associations with disease.

The authors focused on brain-imaging data from more than 8,400 UK Biobank participants. These data were processed to generate a list of thousands of image-derived phenotypes (IDPs) — traits related to brain structure or function that can be identified through images.

Elliott and colleagues investigated associations between IDPs and genetic variants.

The authors' analysis provides new data on the heritability of IDPs, for instance demonstrating that the volume of a given brain region is more heritable than are measurable aspects of its function. Reassuringly, these results generally replicate those from previous studies that analysed a small subset of the IDPs in a greater number of individuals[6,7].

Elliott *et al.* also demonstrated how GWAS on IDPs can be combined with the results of GWAS on neurological and psychiatric disorders as a way to gain insight into possible mechanisms of disease. For instance, they showed that variation at a particular genomic region that has previously been associated with risk of schizophrenia is also associated with certain aspects of brain volume, pointing to a possible mechanism for how and why variants in this region might be associated with disease risk. This work is just a tantalizing teaser of how much more we will learn once 100,000 UK Biobank participants have undergone brain imaging — a project that should be completed by 2020.

The excitement about the opportunities to advance human genetics using UK Biobank is palpable. Most of the variants incorporated in the biobank's database are common, but sequence data being generated to interrogate rare variants will soon be available to investigators. The size and breadth of the resource, coupled with the many related individuals who have donated their samples to this huge database, should enhance our ability to study the consequences of rare variation on a scale we could not have imagined just a few years ago.

The generosity of the United Kingdom in sharing this resource with the rest of the world is a shining example of the value of investing in the greater good. It can be challenging to make large-scale clinical data publicly available, because of privacy concerns and the difficulties inherent in removing all potentially identifying information from electronic health records. Nevertheless, scientists benefit hugely from the broad availability of all of these data sets. The US National Institutes of Health initiative All of Us is being designed to be broadly available to the scientific community. We can celebrate the United Kingdom's generosity best by emulating it. ■

Nancy Cox *is at the Vanderbilt Genetics Institute, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, Tennessee 37232–0700, USA.*
e-mail: nancy.j.cox@vanderbilt.edu

1. Evangelou, E. *et al. Nature Genet.* **50**, 1412–1425 (2018).
2. Bycroft, C. *et al. Nature* **562**, 203–209 (2018).
3. Elliott, L. T. *et al. Nature* **562**, 210–216 (2018).
4. The Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
5. Moutsianas, L. & Gutierrez-Achury, J. *Methods Mol. Biol.* **1793**, 111–134 (2018).
6. Hibar, D. P. *et al. Nature* **520**, 224–229 (2015).
7. Fornage, M. *et al. Ann. Neurol.* **69**, 928–939 (2011).

DEVELOPMENTAL BIOLOGY

A dual origin for blood vessels

Contrary to previous assumptions, it seems the cells that line blood vessels are derived from more than one source. In addition to their known developmental path, they can arise from progenitors of embryonic blood cells. SEE ARTICLE P.223

M. LUISA IRUELA-ARISPE

Blood-cell lineages and the endothelial cells that line the interior of blood vessels have an intertwined biology and interrelated embryonic origins. Our current knowledge indicates that endothelial cells differentiate directly from one of the three main cell layers of the early embryo (the mesoderm), and that a subset of endothelial cells subsequently gives rise to haematopoietic stem cells (HSCs)[1,2], from which adult blood cells derive. On page 223, Plein *et al.*[3] reveal a second origin for endothelial cells, and refine our understanding of the relationship between the endothelial and blood lineages.

Transient embryonic populations of red blood and immune cells arise early in development, before the emergence of HSCs, from precursor cells called erythro-myeloid progenitors (EMPs). In line with the model that mesoderm gives rises to endothelium, which in turn gives rise to blood, EMPs originate from endothelial cells located in a structure called the yolk sac that surrounds the embryo. Using a genetic-engineering approach to produce mouse embryos in which yolk-sac-derived EMPs and all their descendants were labelled with a fluorescent protein, Plein and colleagues unexpectedly found that these cells also contribute to the walls of blood vessels.

Analysis of the labelled cells revealed that

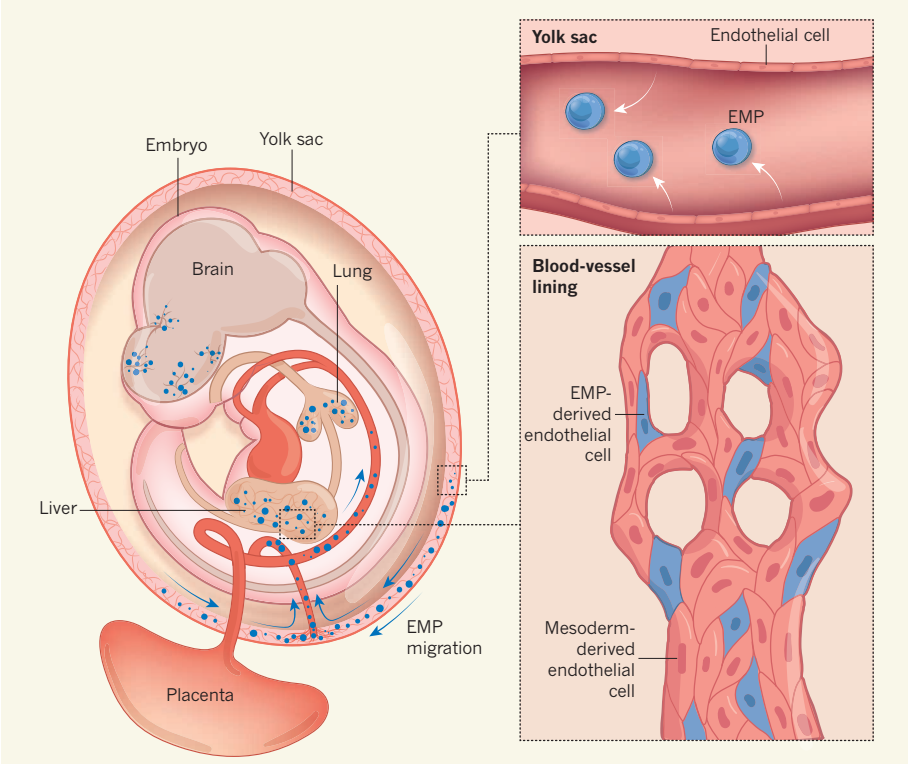


Figure 1 | Two contributors to the blood-vessel lining. An embryonic tissue called mesoderm (not shown) gives rise to endothelial cells, which proliferate to form both the inner lining of blood vessels and the lining of a structure called the yolk sac that surrounds developing embryos. Endothelial cells of the yolk sac in turn give rise (white arrows) to cells called erythro-myeloid progenitors (EMPs), which migrate into the embryo and are known to differentiate into embryonic blood-cell lineages. Plein *et al.*[3] demonstrate in mice that migrating EMPs can also revert to an endothelial-cell type. EMP-derived endothelial cells are incorporated into mesoderm-derived blood vessels in developing organs such as the brain, liver and lung, forming a mosaic pattern across the vessel lining.