Biobanks and beyond—genotyping for the future

As microarray and sequencing technologies evolve to enable faster genotyping, the number of individuals who are being genotyped is accelerating. As of March 2018, the US National Center for Biotechnology Information (NCBI) database of single-nucleotide polymorphisms (dbSNP) contained over 600 million known variants [1]. As other studies continue, that number continues to grow at an astounding rate.

The pace of growth offers exciting potential, but leveraging such vast genotyping resources presents complex experimental challenges. If biomarker discovery is ever to yield routine clinical applications for precision medicine, scientists must build strong statistical evidence of the associations between genetic markers and medical conditions. Large sample sizes are key to identifying both common and rare variants with the statistical rigor that will enable advancements in research as well as clinical and direct-to-consumer applications.

Biobanks provide power in numbers

The genetic information held in large biobanks makes rich datasets of billions of genotypes available to scientists for a wide range of research applications. Each biobank has its own application focus, genotyping strategy, and sources of samples and data from the volunteer contributors.

For example, **UK Biobank** has compiled genotypes of 500,000 consenting individuals from within the UK National Health Service (UK NHS), along with blood and urine samples and complete medical records for each participant. UK Biobank is an exceptionally valuable resource because the UK NHS treats the single largest group of people anywhere in the world, and keeps detailed records on each patient from birth to death. Medical records of the patients are accessible across the entire system. The data from UK Biobank are available to any researcher interested in the wide range of diseases and traits covered by the biobank.

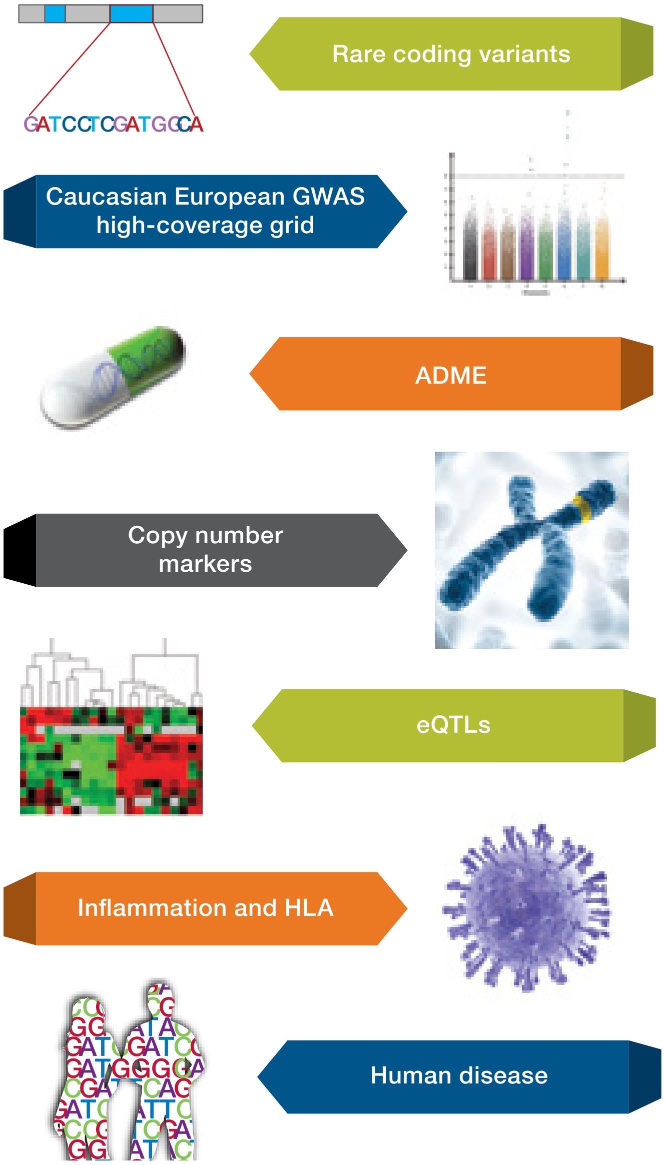


Figure 1. Content on the Applied Biosystems™ UK Biobank Axiom™ Array.

There are other biobanks that are rich resources for investigating causes of diseases in specific populations and worldwide. The number of countries and organizations that are beginning to build their own biobanks is growing. For example:

- The **FinnGen** project launched in October 2018. It is drawing together genetic data and digital health records from a wide network of biobanks across Finland to enable medical innovations such as better diagnostics and new therapies.

- The Tohoku University Tohoku Medical Megabank (TMM) is also based on a unique and specific population. This project is using genetic and biospecimen data from individuals in the Tohoku region of Japan to understand health consequences following the devastating 2011 earthquake, and also to develop a new advanced medical system for the region.

- The China Kadoorie Biobank (CKB) is leveraging China's very large population to investigate the main genetic and environmental causes of common chronic diseases in the Chinese population. The biobank contains genetic data, phenotypic measurements, and blood samples for more than 510,000 adults across ten geographic regions in China.

- The US Veterans Affairs Office is focusing on US military veterans. They have initiated the Million Veteran Program (MVP), which is building a database of genetic data and health information to study military-related and other diseases.

Imputation for array design

The number of known SNPs and other variants has become so large that conventional SNP selection methodologies using tagging no longer provide enough coverage to address the entire genome. For example, in 2007, Kaiser Permanente and the University of California, San Francisco (UCSF), began a collaboration on a genome-wide association study (GWAS) of 100,000 Kaiser Permanente participants. The team sought to genotype large multiracial and multiethnic cohorts, which had higher genetic diversity and more low-frequency variants than previous population cohorts. Initial studies used conventional SNP tagging to capture the most common variations in most human populations. They also leveraged imputation genotyping algorithms to improve fine mapping of associations and facilitate combining results across studies.

However, to achieve the genome-wide coverage for greater diversity and more rare alleles required a higher density of SNPs than was available on a single array at the time. Even combining SNP tagging and imputation was not a practical or cost-effective approach for designing and producing arrays for a very large-scale project such as this. The array design needed to be made more efficient by enabling as much imputation as possible from fewer SNPs on the array.

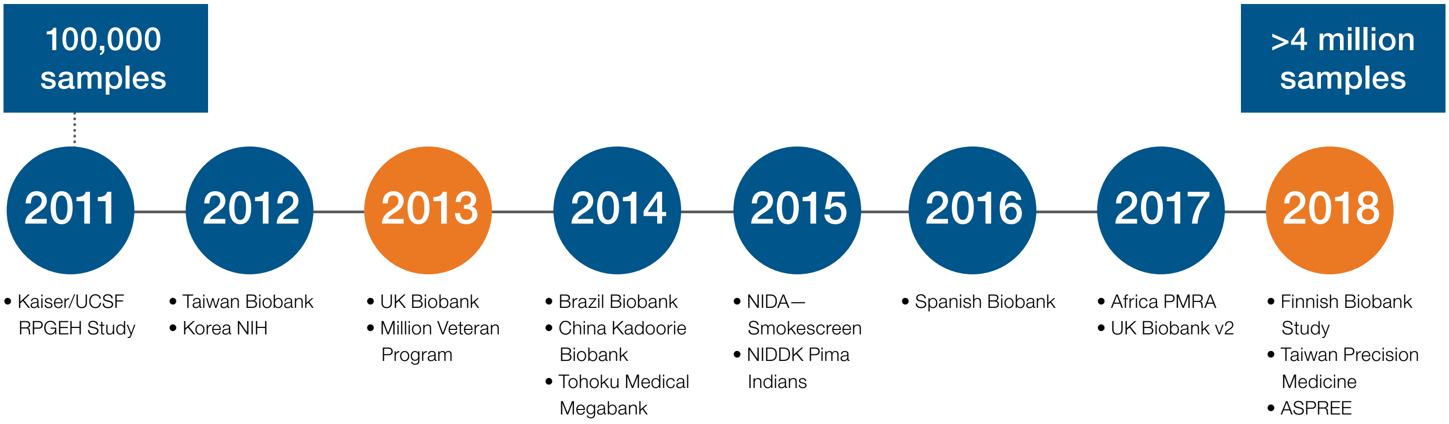


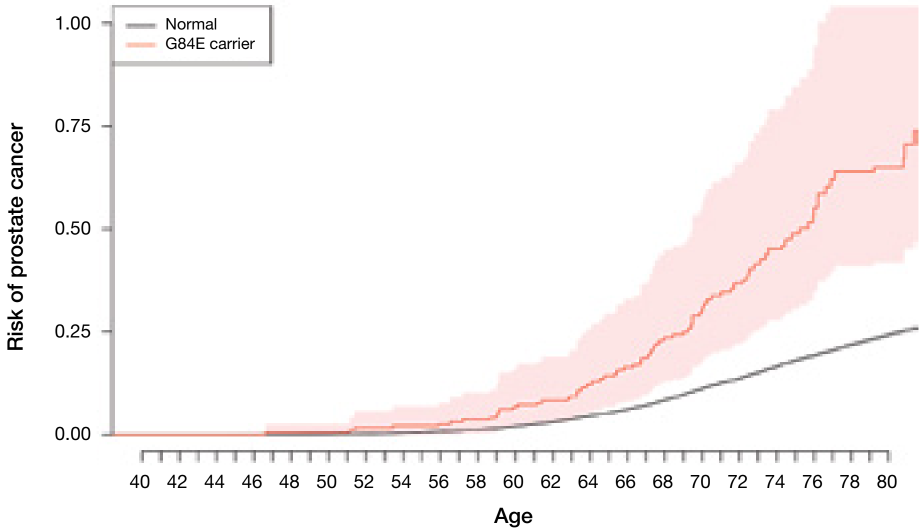
Figure 2. Rapidly expanding Applied Biosystems™ Axiom™ Biobank and precision medicine network.

UCSF, Kaiser Permanente, and Affymetrix (now part of Thermo Fisher Scientific) formed a collaboration to solve this problem using an entirely new approach to array design. The rationale for developing these new arrays was to maximize the number of high-quality SNPs and low-frequency variants for genome-wide coverage; to provide complete and redundant coverage of regions known to be associated with diseases, traits, and outcomes; and to leverage data from sources such as the 1000 Genomes Project and the International HapMap project to improve coverage of both common and uncommon variants.

The team developed a unique algorithm to select SNPs for **Applied Biosystems™ Axiom™ arrays**. The algorithm uses imputation to infer millions of markers from far fewer known haplotypes from HapMap reference samples. Imputation-aware array design enables increased genome-wide coverage across a broad allele frequency range from the same fixed number of SNPs, so it is a much more efficient way to achieve coverage than pairwise tagging. Efficient marker selection is particularly important for multiethnic studies, which often need coverage of low-frequency alleles, more markers, and more space on the array.

In the collaboration between UCSF, Kaiser Permanente, and Thermo Fisher Scientific, imputation-aware array designs yielded an overall increase in statistical power of up to 10% [2]. As a result of this success, the team developed four high-density Axiom arrays with complete genome coverage of both common and rare variants, designed specifically for studies of populations of European, African American, Latino, or East Asian ethnicity.

Genotyping biobanks such as the Kaiser Permanente Research Bank, UK Biobank, FinnGen, TMM, MVP, and others now use this method of imputation-aware array design to create Axiom arrays with content that is specifically tailored to their unique needs.



Normal	33177	33177	33176	32966	32268	30195	24228	12972	4013
G84E carrier	300	300	300	299	291	274	223	111	31

Number at risk

Figure 3. Age-specific risk of prostate cancer by *HOXB13* G84E mutation carrier status. This figure highlights both the value of large biobanks such as RPGEH in undertaking important research across a range of diseases and the ability of imputation-based arrays such as Axiom arrays to evaluate both common and rare variants.

Large-scale GWAS enables polygenic risk scoring (PRS) for human medicine

It is clear that the combined power of the vast genotyping data available in biobanks and imputation-based array design is generating valuable GWAS data from populations worldwide. The ultimate goal now is to apply the data to generate clinically actionable results that will advance the promise of precision medicine. How can GWAS be used to determine cause, diagnosis, prognosis, and therapeutic response for human diseases?

The past decade or so of genetic research has revealed that for many common complex diseases such as some types of heart disease, diabetes, and brain disorders, multiple common variants throughout the genome may play a greater role than rare single-gene mutations. A single trait may be affected by thousands of variants. A significant challenge in understanding the role of genetic variants in complex diseases and traits is that the contribution of an individual SNP is likely to be very small, so they are poor predictors of disease both individually and as haplotypes.

Individual SNPs might not be very useful in predicting complex traits; however, quantifying the cumulative effects of very large sets of variants may overcome this challenge. Such polygenic scoring is widely used in crop and livestock genetics to predict traits. Linking variants to specific traits requires both the genotypic and relevant phenotypic data for each sample. Finding statistically significant associations requires data from many samples. In agricultural settings, very large numbers of samples are readily available, and traits of interest are easily measured.

Predicting human traits is a very different situation. However, with access to over 600 million known SNPs and other structural variants, very high-density imputation-aware arrays, and vast amounts of genetic and phenotypic data available in biobanks, researchers are beginning to develop PRS methodologies to stratify patients into disease risk categories based on their genetic mutations.

Summary

The vast amount of genetic, phenotypic, and outcome data that are stored in biobanks offers opportunities for completely new study paradigms. Researchers have tremendous opportunities to assemble cohorts, genetic data, and phenotypic information for the creation of studies. The large number of samples available increases the likelihood of identifying rare variants with enough statistical power to support association. Combined with more cost-effective imputation-aware array designs, the genetic variant data available today are enabling researchers to design PRS-based tests that may yield valuable information for patients and clinicians. The possibility of using GWAS to enable human disease diagnosis and prognosis is on the horizon.

References

1. NCBI dbSNP Build 151. https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?build_id=151
2. Hoffman TJ et al. (2011) *Genomics* 98(6):422–430. doi: 10.1016/j.ygeno.2011.08.007

Have a question for our technical specialists? Contact us on **thermofisher.com/genotyping-microarray-contact**

For information on genotyping strategies adopted by other researchers, visit thermofisher.com/**scientistspotlight**

ThermoFisher
SCIENTIFIC

For Research Use Only. Not for use in diagnostic procedures. © 2019 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified.