Biobanks, phenotypes, and population variation

Revolution or evolution in complex trait genetics?

Professor Mark McCarthy from the University of Oxford, UK outlines how the intersection of biobank capabilities, new study designs, and technological advances will improve our understanding of complex trait genetics.

Using UK Biobank's genotyping array as an example, Professor McCarthy discusses how new genetic studies can now integrate the power of imputation with our knowledge of population, disease, and biological function to answer some of the remaining questions about the future role of genetics in precision medicine.

Introduction

Over the last few years, the genetics literature has featured a robust debate on the relative success or failure of genome-wide association studies (GWAS) [1–4]. It is clear that, even with over 2,000 significant GWAS loci reported across hundreds of traits [4], there is still much to do to dissect the genetic architecture of many common diseases.

The next chapter of complex genetics looks particularly exciting. Advances in sequencing and genotyping technology are expanding our knowledge of genomic variation in human populations. Visionary initiatives like UK Biobank offer access to new phenotypes, outcomes data, and the opportunity for new and powerful study paradigms.

Recently, Thermo Fisher Scientific talked with Professor Mark McCarthy about the past, present, and future of complex trait genetics.

As a key member of the team that designed UK Biobank's genotyping array, Professor McCarthy also gave his view on how our knowledge of population variation, imputation-based data analysis, disease, and biological function should be integrated into the design of state-of-the-art studies and the genotyping arrays that enable them.

Mark McCarthy is Robert Turner Professor of Diabetic Medicine at the University of Oxford and Consultant Endocrinologist at the Oxford University Hospitals Trust, Oxford, UK. His research team has a long-standing interest in the genetics of complex traits including type 2 diabetes (T2D), obesity and growth. They have played a leading role in international efforts such as the Wellcome Trust Case Control Consortium, that have applied genome-wide association and next-generation sequencing approaches to study these traits. Professor McCarthy's team, along with international collaborators, has identified at least 50 new regions involved in susceptibility to T2D, and a similar number that impact other traits. Their work has provided novel biological insights into the pathogenesis of these conditions, and underpins future efforts to translate genetic findings into clinical practice with which they are closely involved.

Thermo Fisher: Arguably, the GWAS era began in earnest with the 2007 Wellcome Trust Case Control Consortium (WTCCC) paper [5]. What have been the key successes in complex trait genetics since then?

ThermoFisher SCIENTIFIC

McCarthy: There were earlier GWAS successes, for example, in age-related macular degeneration [6] and inflammatory bowel disease [7]; but in terms of scale and scope, WTCCC marked the point where, for many in the field, the potential of GWAS became clear.

The key success has been that the GWAS approach has proved to be so robust. This is, in part, because you are testing a relatively simple hypothesis: specifically, whether or not you see a difference in allele frequencies between cases and controls at a given variant. This has been sufficient to pick up many common variant signals without the need for complex analytical strategies.

Another important point, particularly in retrospect, is that the field applied strong benchmarks early on for defining statistical significance. Few of the signals that reached these thresholds have subsequently proven to be false. This is very welcome when we remember the quite dismal era of candidate gene studies when almost nothing replicated.

These genomic studies also encouraged rapid development of collaborations and the formation of large international consortia. As a result of this data aggregation, the field found a way to move rapidly towards robust signals, avoiding years wasted while individual labs published underpowered studies. This was a major advance in the field and has since had impact on other areas of biological sciences.

All of this showed that high-throughput, large-scale genomics can be done and produces robust results. Out of that came thousands of loci for hundreds of traits.

Thermo Fisher: And, against these successes, what challenges remain that this latest era of complex trait studies will hope to address?

McCarthy: The biggest challenge lies in translating these robust GWAS loci into biological insights about disease pathogenesis. This is slower than some people expected or hoped, and the translational impact of GWAS has been relatively modest to date.

This is principally because most variants identified by GWAS map to regulatory sequence, and most have quite modest effects on risk. Clearly, there's all sorts of activity to address this challenge, not least the Encode Project [8,9], but it will take time to pull it all together.

Thermo Fisher: So, how should genomic strategies evolve to build on these achievements and address the challenges?

McCarthy: Actually, the answer is the same as it ever was: we need more alleles, more phenotypes, more samples, and more analyses. This has been the case in genetics and other fields of science for some time. If you dissect this further, "more alleles" means exploring more populations, including multiethnic and isolate samples. It also requires us to expand our ability to detect different types of allelic variation across a broader frequency spectrum. We will want to look at somatic as well as germline variation.

This is all about identifying as many of these human genetic "accidents of nature" as possible. Each allele that we find and link to a disease phenotype has the potential to improve our understanding of disease mechanisms.

Large biobanks are combining vast numbers of samples with detailed phenotyping data.
Wellcome Images

"More phenotypes" means expanding GWAS into new case-control studies, but especially into large biobanks where the link to health records makes a broader spectrum of phenotypes available. Biobanks allow us to discover how variants of interest from one disease contribute to risk in other diseases.

All of this is, of course, massively enabled by advances in genotyping and sequencing technology that allow us to capture the genetic variation that lies within these large and well-characterized biobanks.

Finally, "more analyses" means developing methods tuned to take advantage of these datasets that are coming on-stream. For example, multivariate methods will allow us to look at the relationship of genetic variants to many traits, not just one. This will improve power and reveal novel relationships.

Thermo Fisher: Most GWAS have been performed with arrays optimized to capture common European alleles. How does this need to change to enable studies of more alleles in more populations?

McCarthy: Clearly, there is a great deal of similarity in the variation, and the patterns of variation, between many ethnic groups, but there are also important differences. At the common variant level, the set of tag SNPs that you might want to select to most efficiently cover the genome will differ from population to population. This will matter even more as we focus more on functional alleles that are more likely to be rare and ethnicity-specific.

"The ideal scenario is to redesign your array every time so that it is optimized for your particular population and research needs."

With a finite amount of money available and limited real estate on a given array, the ideal scenario is to redesign your array every time so that it is optimized for your particular population and research needs. And if you want, as many do, to combine GWAS and functional content—which, with our current level of understanding, most likely means variation in coding regions—you should design the most efficient set of SNPs that addresses both of those questions. That's what we did in relation to UK Biobank. People working on other populations would ideally want to do something analogous, if cost and technology permitted.

Thermo Fisher: Can you explain a little more how you applied this approach in the context of UK Biobank?

McCarthy: Yes, of course. The design group, led by Peter Donnelly in Oxford, used UK sequence data from 1000 Genomes [10,11] and elsewhere to optimize the array content for the most efficient imputation of haplotypes seen in UK samples. This included extending the genome-wide scaffold beyond common variants and adding content that allowed the array to improve imputation in the low-frequency range. Such alleles have not previously been targeted with imputation-ready marker scaffolds.

We also filtered the functional content to select alleles that we knew to be present in the UK population. We wanted to avoid wasting time and money on alleles likely to be monomorphic or present at extremely low frequency in UK Biobank participants.

Overall, having the ability to specify content on a SNP-by-SNP basis is obviously a huge advantage when it comes to efficient coverage of the swath of variants that you're trying to capture, either directly or by imputation. This is something that some manufacturers are better able to do than others. When arrays are made using predefined pools of SNPs, it is much more difficult to be so flexible.

Thermo Fisher: Would this approach of designing population-focused arrays cause difficulties later if different array datasets are combined in meta-analyses?

McCarthy: No. For common variants, it's clear that imputation works extraordinarily well and continues to improve as reference data panels improve. There's not much of a penalty when combining data from studies that used different common variant array content. We rarely found ourselves troubled by that, and it becomes even less of an issue as imputation improves.

Of course, lower-frequency variants are more population-specific anyway. This requires a focused approach to array design; otherwise, your array will include variants that may be useful in other populations but are monomorphic or too rare to be informative in your population. This just wastes money and space on the array.

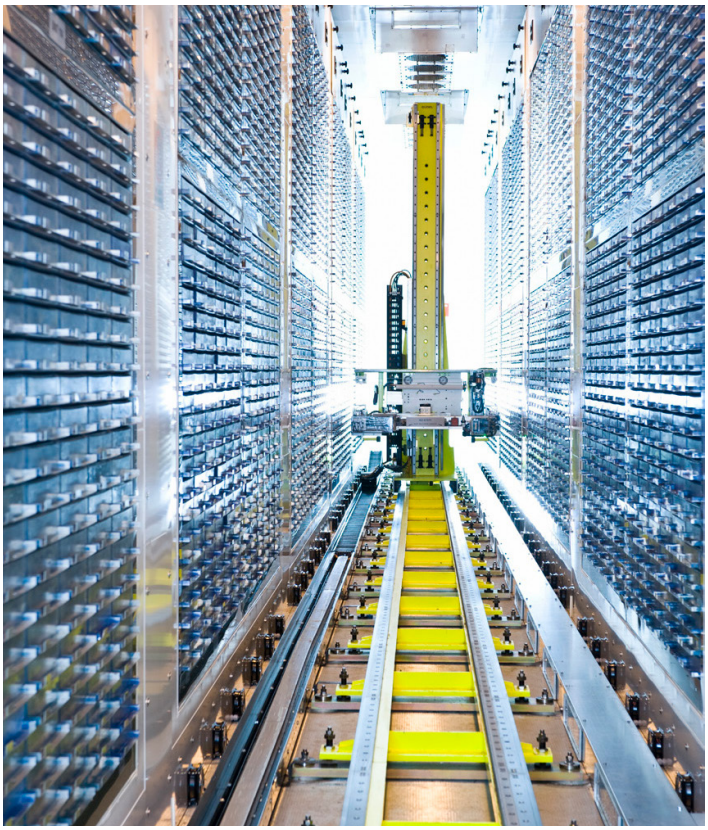
Thermo Fisher: So, what does a good GWAS imputation scaffold look like, and how is it different from the older pairwise tag SNP array designs?

McCarthy: A good scaffold is a set of markers that demonstrates, for the population of interest, the most efficient ability to impute genomic coverage across a broad allele frequency range.

There's no doubt that designing a scaffold using an imputation-aware, multimarker tagging approach is a much more efficient way to achieve coverage than using pairwise tagging. It requires fewer markers, and this is particularly important if you are designing cosmopolitan arrays for multiethnic studies or need coverage of low-frequency alleles. These types of designs will inevitably need more markers and more space on the array, so selection efficiency, during design, acquires even greater importance.

"There's no doubt that designing a scaffold using an imputation-aware, multimarker tagging approach is a much more efficient way to achieve coverage."

With the UK Biobank array, we didn't want to ignore common variation, of course, because many new findings will come from studies in the 500,000 biobank participants. However, common variants have been well covered in European populations, so there was the desire to push imputation down into the low-frequency space. This has proved to be feasible, to some extent because UK Biobank provided a relatively constrained focus to the study.



UK Biobank is being developed as a global research resource.
Wellcome Images

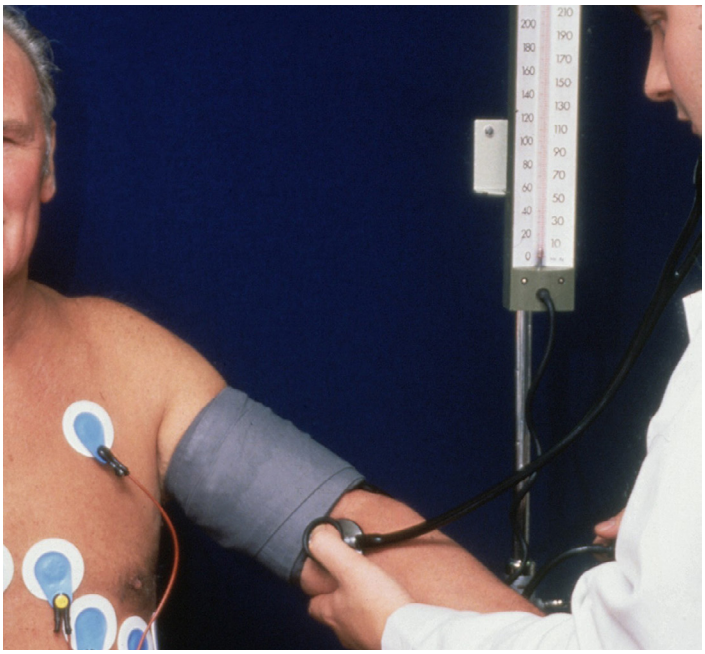
Thermo Fisher: As you noted, advances in sequencing and genotyping technology are providing access to millions of new variants. How should investigators prioritize what goes onto their genotyping array?

McCarthy: Well, clearly, there's a set of things that should be on every array: the known GWAS signals because you wouldn't want to miss out on those; ancestry informative markers (AIMs); fingerprinting markers; X/Y markers; human leukocyte antigen (HLA); and so on. They don't take up a lot of real estate and they are tremendously useful for sanity checking, but also because they invite testing of specific hypotheses. Beyond that, it depends on your objective and on the hypotheses you have about where the strongest biological effects lie.

We had exactly these discussions in the context of the UK Biobank array. For example, you have to make a decision on how much real estate you give to your GWAS scaffold and how much you give to functional or coding SNPs. This depends on your expectation of what you will discover with each.

It also depends on the extent to which you up-weight coding variants over noncoding variants on the basis of their biological importance. Coding variants, for now at least, are that much easier to take forward into biological and physiological studies. That doesn't mean that I have any prior belief about their relative contributions to genetic architecture. There's still little empirical data on this, particularly for low-frequency variants. But, all else being equal, and given the pressure to show that these approaches produce biological insights, I'd generally favor discovering a coding variant over a noncoding variant.

The architecture of most common diseases looks very similar at the common variant level, but this may not be true at lower frequency. As empirical data appears over the next couple of years, how you might more carefully weight these design decisions may become clearer. Some diseases may be more dominated by coding variants, for example, and this information will help to make the decision on how best to use the real estate on the array.



Harmonized phenotyping in biobanks will provide a boost in study power.
Wellcome Images

Thermo Fisher: You mentioned noncoding variants are harder to take forward. What role can eQTLs play?

McCarthy: Expression quantitative trait locus, or eQTL, analysis [12] is one tool we have that allows us to link regulatory variants to the transcripts whose expression they may influence. It's extremely valuable when you find a strong eQTL that's coincident with a GWAS signal.

It's best if that eQTL is detected in a pathologically relevant tissue, but that's not always possible, and many eQTLs act across multiple tissues. Our current strategy is that a combined eQTL discovery approach is best. First, you can use eQTL data from whole blood or lymphoblastoid cell lines. These data, available from many thousands of samples, allow you to create an inventory of cis, and possibly trans, eQTLs. It won't necessarily be very specific for the eQTLs in a different tissue, but it will be very sensitive for a subset of disease-relevant eQTLs because of the very large sample sizes.

Second, you can explore your tissue of interest, usually involving much smaller samples. One of our interests is diabetes so, with colleagues, we have completed expression studies in subcutaneous fat, muscle, and pancreatic islets, for example. The smaller sample size means you will only pick up large eQTL effects specific to the tissue, but you can at least filter the larger lymphocyte and blood-based datasets for their relevance to your disease. I would argue you probably want to combine both approaches to get the optimal balance of sensitivity and specificity.

Thermo Fisher: Large prospective biobanks are being created in many countries. Some, like UK Biobank, are making data freely accessible to the global research community. How will these resources be used by the global community?

McCarthy: In several ways. The prospective component of UK Biobank is particularly important because you have biosamples, data, and exposures at baseline. By linking these to electronic medical records, you have the potential to track what happens to those people over time. This is a powerful design for biomarker identification and validation. The long periods of prospective follow-up enable you to identify nested case–control groups based on incident disease. That, in turn, means you can hopefully identify genetic and nongenetic biomarkers that are prospectively and causally related to onset of disease. With the exposure data available, you can also do much more comprehensive studies of gene–environment interactions than have been possible before.

The scale, harmonization, and standardization of phenotype and genotype that we will see in large biobanks will almost certainly pay dividends even for traits where we already have studies involving several hundreds of thousands of people. The combination of harmonized genotyping and phenotyping, and the availability of individual-level data on all participants, will provide a boost in power above and beyond that which is possible from the sample size alone.

When you compare a biobank such as this to conventional case-control samples, one real advantage is the ability to perform these analyses across a wide range of phenotypes. Understanding pleiotropy may enable us to pick up signals that were not quite detectable before and to better understand the relationships between traits. A good existing example of this is the alleles in the *HNF1B* gene that increase risk of T2D but which are also protective for prostate cancer [13,14]. These two diseases don't occupy the same chapter of a medical textbook, but the shared genetics point to unexpected mechanistic connections.

Prospective biobanks are also going to be very valuable for the evaluation of human therapeutic targets. Imagine that you have already found an interesting association between a coding variant and your disease of interest. The variant is in a gene that looks eminently druggable. By looking across diverse phenotypes, you can test whether the same genetic variant has other trait associations. You can then assess the effect of perturbation of the protein, in the direction appropriate for the disease of interest, on those secondary traits. Your promising drug target becomes a lot less attractive if it's likely that any drug, developed to mimic the beneficial effect on the disease of interest, will be likely to cause undesirable effects on those secondary traits.

Thermo Fisher: Recently there has been some debate on using controls from prospective biobanks for studies of cases from other countries. What are your thoughts?

McCarthy: I would recommend caution. We've seen several examples where researchers have over-interpreted the differences they found between cases and controls when the two have different origins or have been typed on different arrays. There are some major technical challenges to such analyses, and it's not trivial to overcome them. I'm not a great fan of doing formal analyses of this type.

However, access to large biobank datasets can be useful for evaluating apparent case associations, particularly for rare variants. For example, you may have a rare variant that looks interesting and which you have seen in only three out of 10,000 individuals, all of them cases. Going to any single source of controls and seeing the same variant in zero out of 10,000 individuals won't tell you very much. But, if you look across many different biobanks, the representation of that allele, or of other alleles in the same gene, becomes meaningful. Has that variant ever been seen in those biobanks? If so, what is its relationship to your trait of interest? This can be really helpful as you try to evaluate the significance of those three instances where disease cases carry the rare allele.

Large datasets allow these kinds of "informal" analyses. That's why the Exome Variant Server (EVS) database [15] arising out of the Exome Sequencing Project (ESP) has been so useful. Researchers generally don't do an association study comparing their cases against ESP controls. Instead, when they see a variant of interest in their own case-control analyses, they go and look in EVS and get a sense of how frequently others have seen that variant. They also look at associated phenotypes because that allows them to prioritize particular variants for further study.

Thermo Fisher: This all suggests that exciting times are ahead for complex trait genetics. Looking ahead, what will we learn in the next five years, and what clinical impact would you hope to see beyond that?

McCarthy: Well, over the next four to five years we'll have sequenced and genotyped enough people to have a pretty good idea of the genetic architecture of complex traits. We'll have a better understanding of the contribution of rare alleles and low-frequency alleles for many traits, and we'll know if all diseases behave fairly similarly or whether they have rather different architectures. For example, maybe we'll find more rare coding variants in one disease than another. There's a lot of opinion in this area; we really need empirical data to establish which hypotheses are sustainable. This will have a major impact on how we plan our strategies for further discoveries.

These answers will also help to define how we might be able to use genetic data to drive prediction and stratification of disease. From what we know today, DNA sequence alone will provide rather imprecise estimates of individual risk for most common complex traits. However, by gathering omics data at intervals during a person's lifetime, we could derive "molecular signatures" of disease. In combination with sequence and exposure data, we can hope to refine and update an individual's disease risk profile in real time. It would, of course, be transformative if we can use genetics to stratify disease subtype, make mechanistic inferences, and suggest specific treatment options for individual patients. It remains to be seen to what extent this is possible: this is another area characterized by a lot of opinion and very little data. We're already seeing the potential of this approach in various types of cancers, but I suspect the power of this approach will vary between diseases.

applied
biosystems

As I've noted already, when we get deeper into this through sequencing backed up with custom arrays, we will learn more about functional variants in drug targets. Where we have good human validation data, we will be able to parse those targets for their potential adverse effects, and that will clearly benefit pharma. The industry is increasingly disappointed by targets that have been validated in pre-clinical models but perform poorly as soon as they hit first-in-man trials. I would hope, as we have seen with *PCSK9* [16,17], that human genetics will be a very powerful tool for putting better-quality targets forward for drug development.

Thermo Fisher: Thank you very much.

Further reading:

Applied Biosystems™ Axiom™ UK Biobank Genotyping Array

The array designed by the UK Biobank team is now available to all investigators. To learn more about this design and options to modify it for other populations, go to **assets.thermofisher.com/TFS-Assets/GSD/Technical-Notes/uk-axiom-biobank-genotyping-arrays-datasheet.pdf**.

UK Biobank

To find out more about UK Biobank and how you can access data for your own research, go to **www.ukbiobank.ac.uk**.

References:

1. McCarthy MI et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356–369.
2. McClellan J, King MC (2010) Genetic heterogeneity in human disease. *Cell* 141:210–217.
3. Sullivan P et al. (2011) Don't give up on GWAS. *Mol Psychiatry* 17:2–3.
4. Visscher PM et al. (2012) Five Years of GWAS Discovery. *American Journal of Human Genetics* 90:7–24.
5. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678.
6. Klein RJ et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–389.
7. Duerr RH et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314:1461–1463.
8. Division of Genome Sciences, NHGRI, NIH. The ENCODE Project: ENCyclopedia Of DNA Elements. www.genome.gov/10005107 [Online].
9. The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306:636–640.
10. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
11. The 1000 Genomes Project Consortium. 1000 Genomes: A Deep Catalog of Human Genetic Variation. www.1000genomes.org [Online].
12. Cookson W et al. (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10:184–194.
13. Gudmundsson J et al. (2007) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* 39:977–983.
14. Frayling TM et al. (2008) A genetic link between type 2 diabetes and prostate cancer. *Diabetologia* 51:1757–1760.
15. NHLBI GO Exome Sequencing Project (ESP), Seattle. Exome Variant Server. http://evs.gs.washington.edu/EVS [Online].
16. Hall SS (2013) A gene of rare effect. *Nature* 406:152–155.
17. Willrich MA Baudhuin LM (2013) PCSK9 and the road less traveled: how an unconventional approach led to a major discovery. *Clin Chem* 59:1283–1284.

Read more interviews at
thermofisher.com/scientistspotlight

ThermoFisher SCIENTIFIC