biobank uk
Improving the health of future generations

22nd March 2017

Q&A: Exome sequencing of UK Biobank samples

UK-based pharmaceutical company GSK and Regeneron, based in the US, have successfully applied to undertake genetic sequencing of UK Biobank samples, that focusses specifically on 1-2% of the human genome, known as the exome. Scientists believe this is the area of most relevance for new drug therapies, and the work could herald a new era of drug development.

What does UK Biobank hope will come out of this approved use of sample?

The ground-breaking project by GSK & Regeneron to sequence the exomes of all 500,000 UK Biobank participants provides an extraordinary opportunity to make the UK Biobank resource even more useful for health-related research that is in the public interest, and to bring about advances in preventing and treating a wide range of disorders more rapidly.

UK Biobank was established to make such data available on a very large scale to academic and commercial researchers all over the world so that they are able to study many different conditions reliably. Without this commercial investment, these exome sequence data would not become available so soon because the costs of such assays are too high to be covered by most charity or government research funding schemes.

How will the UK Biobank resource benefit from this collaboration? What value will the GSK/ Regeneron collaboration add to the resource itself?

UK Biobank will have the benefit of exome sequence data on each of its participants, which will be invaluable for future medical research – much of which is likely to have a genetic component. Exome sequencing typically costs about $300 per sample. This exome sequencing project represents a big investment in the UK Biobank resource: a value of about US$150M for all 500,000 samples (scheduled to be completed during the next 3 years). Having such valuable data provided at no cost to UK Biobank, so that they can then be made available to researchers worldwide will be an enormous boost to medical research and will help bring about advances in diagnosis and treatment more quickly.

How big – and how important – is this development?

UK Biobank is of national and international strategic importance in the global effort to improve health. Exome sequencing builds on UK Biobank's existing enhancements (genotyping all 500,000 participants, detailed MRI imaging of 100,000 participants, key biochemistry on 500,000, week-long activity data on 100,000).

This investment by GSK/Regeneron will allow the resource to be used for even more detailed and pioneering research, and could encourage further investment from industry to

do more analyses that are beyond the pocket of government and charity funding. A sequencing initiative on such a huge scale has never been done before. It is a world first, and an endorsement of UK Biobank and the vision of the funders (MRC & Wellcome Trust) who established the resource 10 years ago. Exome sequencing in UK Biobank seems likely to produce novel findings which could be important for many different conditions, especially since the first phase of exome sequencing will be for the most deeply characterised UK Biobank participants (including those who have been imaged).

What is the benefit for UK Biobank participants and the wider public?

On joining UK Biobank, participants were told that their participation may be of no direct benefit to them, but that it was an investment for future generations. However, the potential benefits for the wider public (including UK Biobank participants) of having these exome sequence data available for research are the possibilities for significant advances in understanding and treating diseases.

What is the exome? What is exome sequencing? How does it compare to genotyping?

The human genome is made up of about 3 billion "letters" of DNA in each cell. The order of the vast majority of letters is the same in every individual, but changes in the order can account for different human traits, including diseases. There are three main ways to study the genome:

1. Whole genome sequencing – where every letter in an individual's genome is measured;
2. Exome sequencing – where every letter is measured, but only in those parts of the genome (the "exome") that are directly used to produce proteins; and
3. Genotyping – where several hundred thousand letters are measured throughout the genome and associations of regions in the genome found with various conditions (for large studies, this has been the most common approach to date);

UK Biobank has already undertaken genotyping of its 500,000 participants and this information is already contributing to health research. More than 800,000 letters ("genotypes") were measured and this information has allowed more than 70 million other genotypes to be estimated ("imputed") since areas of the genome tend to be passed on together from one generation to the next.

By contrast with only measuring a small proportion of the letters and the uncertainty of estimating other letters with genotyping strategies, exome sequencing involves the direct measurement of the letters at each location throughout the exome. Exome sequencing is several times cheaper than is whole genome sequencing (although information in the non-exome part of the genome that may well be important for health is not measured), and it has smaller data storage costs and better developed analytical approaches for data on this very large scale.

Why is the project being done in phases?

The work is being done in three phases (50K samples in 2017; 100K samples in 2018; and remaining 350K samples in 2019) so that, based on emerging information, UK Biobank and GSK/Regeneron can decide whether (or not) to continue after the first and second phases.

Why is exome sequencing being proposed on all 500,000 participants? Would whole genome sequencing be more valuable in the long run?

Scientists make applications to use UK Biobank based on their needs. GSK/Regeneron applied to do exome sequencing and this will add value to the UK Biobank resource. If others wish to undertake whole genome sequencing of UK Biobank samples, then they can apply to do so in the same way; there is plenty of stored DNA for subsequent sequencing. The GSK/Regeneron exome sequencing project is being undertaken in three phases. So, it could change to whole genome sequencing if that was considered to be appropriate in the future, such as in light of new information from this study (e.g. early findings) or perhaps a big drop in the cost of whole genome sequencing.

What was the process for taking the decision to allow GSK and Regeneron to be provided with samples? What was the role of the MRC and the Wellcome Trust, which fund UK Biobank?

Scientists from academia or from industry can apply for access to UK Biobank samples and data provided they are to be used for health-related research that is in the public interest and the data generated by the research are put back into the resource so that other researchers can use them. GSK and Regeneron submitted an application for this project in the normal way. The UK Biobank Access Sub-Committee – which is a sub-committee of UK Biobank's Board responsible for overseeing application to use the resource – considered this request carefully before recommending it to the Board. The MRC and Wellcome Trust are represented on the Board, and were made aware of the proposal from an early stage and also agreed to the project.

Why is this being paid for by the private sector? Couldn't the MRC or government and or the Wellcome Trust have paid for it?

GSK and Regeneron want to do the work now and will undertake it. The MRC and Wellcome Trust have already committed many millions of pounds to UK Biobank and their initial investment is engendering this additional funding, which will enhance the resource for other researchers. Exome sequencing costs about $300 per sample (compared with about $50 for genotyping and $1000 for whole genome sequencing). Assuming that all three phases are completed, then this will represent an investment of $150M in making UK Biobank even more useful for health research.

Why are the companies being granted a 9-month exclusivity deal? Isn't that unfair? Is this standard practice?

Academic and commercial organisations undertake assays at their cost, so there needs to be some incentive for them to do so. UK Biobank's practice is to grant exclusivity of between 6 and 12 months (depending on the cost, complexity and utility of the assay data) before the data are made available to other researchers. The policy is the same for academic and commercial researchers.

Is this the first application of this type? Is this the first time the UK Biobank cohort has been sequenced?

The whole UK Biobank cohort has been genotyped (measuring several hundred thousand of the most common differences in DNA), which was done with grants from the MRC, Department of Health and British Heart Foundation. There are on-going MRC-funded assays of the length of the chromosome "telomeres" (the ends of the chromosomes) which is thought to be a marker of aging. There has also been a previous approval for exome sequencing by an academic group in participants with coronary disease. All applications to use the UK Biobank resource are handled in the same way.

Have participants given their consent for their data/DNA to be used by commercial companies?

Yes.

Will UK Biobank participants be identifiable through the data?

No, identifiable data is not provided to researchers.

Is there a requirement that results must be published within a particular time? If so, when and where?

All researchers using the UK Biobank resource are required to report annually on the progress of their work and, where appropriate, will be expected to publish their findings and tell us what, if anything, they have patented or otherwise found out that is significant for health. Researchers are also obliged to return their results to UK Biobank so that other researchers can benefit from them.

Why can't GSK/Regeneron collect their own samples? Why should they have access to a publically-funded cohort?

UK Biobank was established to make research easier, quicker and more cost-effective by building one very detailed resource and then making it widely available. Collecting and storing samples from large numbers of people and then following their health is expensive. The Wellcome Trust and MRC have funded UK Biobank to do this, freeing researchers to get on with the science. The UK Biobank resource is available for health-related research to all bona fide researchers, whether they are in academic or commercial organisations, from the

UK or elsewhere. By investing in these assays for their own research, GSK and Regeneron are also supporting the research of many others because these data will be made freely available after the agreed period of exclusive use.

What Open Access and Open Data requirements are in place? Are they consistent with MRC's standard requirements?

Yes. The agreement is subject to the UK Biobank's published Access Procedures. Researchers are required to use their best endeavours to publish the findings of any research deriving from the resource in an academic journal or on an open source publication site within 6 months after the date when it was agreed that the research would be completed. Within 6 months of publication or 12 months of when the research project was to be completed, they must provide to UK Biobank the results of the research, and the raw data behind them, for inclusion in the resource in such detail and format as is reasonably required.

What happens if the companies breach the terms of the MTA, and what happens to the data?

If an applicant breaches the terms of the MTA, they potentially lose their right to use the UK Biobank resource, which may apply to the individual researcher/research group or the institution or company they work for. UK Biobank understands the importance of maintaining participant confidentiality and would take whatever steps are open to it to ensure this is upheld and respected.

Would all of the Intellectual Property (IP) derived from the research be owned by the companies? Are there any restrictions on any IP?

As is standard in UK Biobank, any IP that is derived from research using the resource is owned by the researcher (in this case, GSK and Regeneron) for the approved research that they do. Researchers are not allowed to patent genotype-phenotype data because UK Biobank considers that it is not reasonable for any researcher to try to assert exclusive rights over genetic or phenotypic data.