Muddyman *et al. Genome Medicine* 2013, **5**:100
http://genomemedicine.com/content/5/11/100

Genome Medicine

CORRESPONDENCE

Implementing a successful data-management framework: the UK10K managed access model

Dawn Muddyman[1*], Carol Smee[1], Heather Griffin[2], Jane Kaye[2] and the UK10K Project[3]

Abstract

This paper outlines the history behind open access principles and describes the development of a managed access data-sharing process for the UK10K Project, currently Britain's largest genomic sequencing consortium (2010 to 2013). Funded by the Wellcome Trust, the purpose of UK10K was two-fold: to investigate how low-frequency and rare genetic variants contribute to human disease, and to provide an enduring data resource for future research into human genetics. In this paper, we discuss the challenge of reconciling data-sharing principles with the practicalities of delivering a sequencing project of UK10K's scope and magnitude. We describe the development of a sustainable, easy-to-use managed access system that allowed rapid access to UK10K data, while protecting the interests of participants and data generators alike. Specifically, we focus in depth on the three key issues that emerge in the data pipeline: study recruitment, data release and data access.

Introduction

The principle of open access to sequence data was established in the Human Genome Project (1999 to 2004), whose sequence data were published on the internet as soon as it was available. One of the leading laboratories in this project was the Wellcome Trust Sanger Institute (WTSI), which, with the support of the Wellcome Trust, continues to advocate strongly for data sharing and implements data-sharing policies for most of its sequencing projects [1]. At the time of writing, UK10K (2010 to 2013) was Britain's largest genomic sequencing consortium, having been awarded £10.5 million by the Wellcome Trust. The purpose of the award was to investigate how low-frequency and rare genetic variants contribute to human disease, and to provide the lasting legacy of a research resource that could be used by the wider research community [2].

It was decided at the outset of the project that UK10K data should not be deposited openly on the internet and should only be accessed through a managed access system. This was due to the potentially sensitive nature of some of the sample sets used, the restrictions on data usage imposed by some Research Ethics Committees (RECs) and the nature of the existing consents concerning study participation. The challenge was to develop a managed access system that allowed rapid access to the data, while protecting the interests of participants and data generators. This required putting into place a governance system involving a series of checks and balances that were proportionate, easy to use and sustainable in the long term. This paper outlines the history behind open-access principles and then goes on to describe UK10K's managed data access process. To illustrate some of the challenges of implementing data-sharing principles, we focus on three key issues in the data pipeline - study recruitment, data release and data access. During the project, we considered ethical issues such as whether or not to provide feedback in the form of health-related findings to research participants. Although the discussion of these particular issues remains outside the scope of this paper, the project's procedure for dealing with the feedback of such findings may be found in the UK10K Ethical Governance Framework [2] and in a separate publication by the authors [3].

Open-access principles

The first international document to lay out the principles for open access in the field of genomics was the Bermuda Agreement made in 1996 [4], followed by the Fort Lauderdale Agreement made in 2003 [5] and the Toronto agreement made in 2009 [6]. These documents set out the key principles that now dominate thinking and practice regarding open access to genome

* Correspondence: dm11@sanger.ac.uk
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
Full list of author information is available at the end of the article

sequence data. The Bermuda Agreement stipulated that pre-publication sequence data 'should be freely available and in the public domain in order to encourage research and development and to maximize its benefit to society' [4]. The Fort Lauderdale Agreement set out a plan for the establishment of 'community resources' to achieve rapid and open data release, as 'community resource datasets benefit the users enormously, giving them the opportunity to analyze the data without the need to generate it first. The datasets are, in general, much larger, richer and of higher quality than individual laboratories could normally generate' [5]. Such datasets have been presented as the 'drivers of progress in biomedical research' and therefore they should be 'made immediately available for free and unrestricted use by the scientific community to engage in the full range of opportunities for creative science' [5]. In line with these principles, the Wellcome Trust requires the researchers that it funds to ensure that 'genome-sequence data should be made freely available as rapidly as possible' [7]. Data sharing ensures the maximum use of data that have been generated using a limited number of samples.

The principles of data sharing have been widely accepted and endorsed by the research community, but the way that they have been implemented over the past few years has been subject to change. A number of sequencing projects, such as the National Centre for Biotechnology Information (NCBI) database of single nucleotide polymorphisms (SNPs), the HapMap Project, the 1000 Genomes Project and Encyclopedia of DNA Elements (ENCODE), have been based on the principles of open access, which requires that sequence data are deposited online. When these projects were established, it was assumed that there would be no risk of re-identification of research participants who had donated their DNA for sequencing. This assumption was overturned when the article by Homer *et al.* [8] demonstrated that data from individuals could be distinguished in genome-wide association study (GWAS) data using only summary statistics. More recently, it was demonstrated that male participants could be re-identified by linking SNPs on the Y chromosome with data found in publicly available datasets on the internet [9]. This has resulted in a significant policy change, with some existing datasets being removed from the internet. Newer datasets are being put into managed access systems whenever it is considered that managing access helps to protect research participant confidentiality; ensures that data are used within the boundaries of consent and/or other relevant approvals; and enables checks to be made to ensure that data requestors are '*bona fide*' researchers. All of this activity has led commentators to suggest that because of the potential identifiability of genomic information, it is important that participants who are involved in genomic research understand that although their privacy cannot be guaranteed, appropriate legal mechanisms have been put in place to protect participants from exploitation [10].

UK10K

The UK10K project was a collaborative project that brought together researchers working on obesity, autism, schizophrenia, familial hypercholesterolemia, thyroid disorders, learning disabilities, ciliopathies, congenital heart disease, coloboma, neuromuscular disorders, and rare disorders including severe insulin resistance. These conditions are regarded as sensitive because of the social stigma that may be attached to them, the health implications for individuals and their families, and the fact that some were rare conditions only affecting small numbers of the population. DNA samples from close to 5,500 individuals who presented with one of these disease phenotypes were provided by collaborators from their existing collections. These 'disease' samples were whole-exome sequenced. A further 4,000 highly phenotyped 'control' samples were supplied from the TwinsUK [11] (King's College London) registry and the ALSPAC Study (Avon Longitudinal Study of Parents and Children, Bristol University) [12]. The samples from these two studies, referred to as the 'cohorts' group, were whole-genome sequenced at 6x depth - sufficient coverage to detect variants down to a 0.1% allele frequency. This detailed sequence reference database was further enriched with extensive phenotype data collected by the TwinsUK and ALSPAC studies.

Similarly comprehensive data were generated for participants recruited to the 'disease' arm of the project, with 5,500 samples being whole-exome sequenced to 72x depth. The data generated were sufficient to discover novel rare and low-frequency variants associated with the diseases investigated in UK10K. Further information about the project and a list of scientific publications arising from it can be found at the UK10K Project website [2].

To enable the project to develop insights into these conditions and to provide an invaluable, easily used resource from which the whole research community might benefit, a data pipeline had to be developed. There were three key stages in this pipeline: recruitment of studies or clinical collections into the project; release of the sequence data; and access to the dataset. The pipeline and the various steps that were part of the governance framework are illustrated in Figure 1 and are described further below.

Recruitment

Setting up a project such as UK10K involved the collaboration of the clinicians and cohort studies that supplied

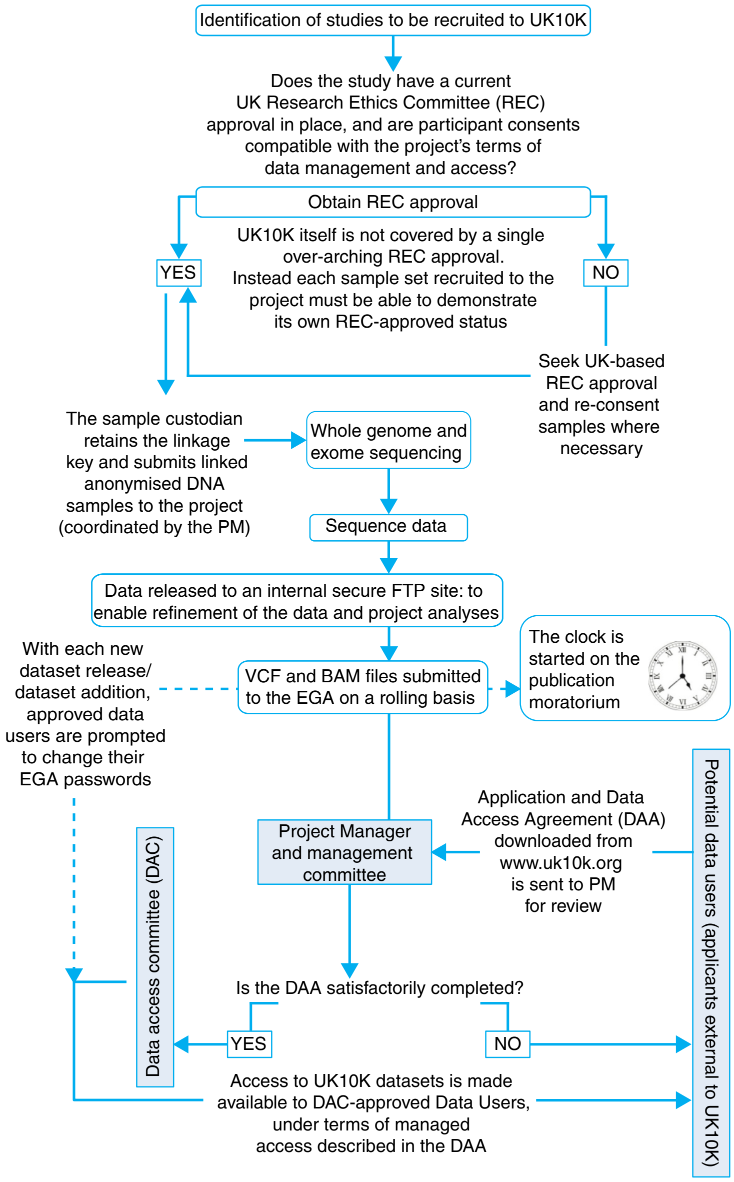
Identification of studies to be recruited to UK10K
Does the study have a current UK Research Ethics Committee (REC) approval in place, and are participant consents compatible with the project's terms of data management and access?
Obtain REC approval
YES
NO
UK10K itself is not covered by a single over-arching REC approval. Instead each sample set recruited to the project must be able to demonstrate its own REC-approved status
Seek UK-based REC approval and re-consent samples where necessary
The sample custodian retains the linkage key and submits linked anonymised DNA samples to the project (coordinated by the PM)
Whole genome and exome sequencing
Sequence data
Data released to an internal secure FTP site: to enable refinement of the data and project analyses
With each new dataset release/ dataset addition, approved data users are prompted to change their EGA passwords
VCF and BAM files submitted to the EGA on a rolling basis
The clock is started on the publication moratorium
Potential data users (applicants external to UK10K)
Application and Data Access Agreement (DAA) downloaded from www.uk10k.org is sent to PM for review
Project Manager and management committee
Data access committee (DAC)
Is the DAA satisfactorily completed?
YES
NO
Access to UK10K datasets is made available to DAC-approved Data Users, under terms of managed access described in the DAA

Figure 1 Flow diagram to illustrate the process of data flow, access requests and approvals. BAM, binary alignment map; PM, project manager; VCF, variant call file.

the samples, as well as WTSI support in terms of staff, resources, expertise and project management. Before the project could commence, appropriate research participant consent and/or REC approvals had to be in place, as did policies and processes to ensure that the storage and use of the samples complied with relevant regulations. In addition, an Ethical Governance Framework document [2] had to be agreed upon that detailed the data management framework. A suitable operational governance structure for the project was also developed to allow decision making and to ensure accountability to the WTSI and the Wellcome Trust. Without these structures in place, there could have been significant delays to the project or sequencing output could have been jeopardized. The development and implementation of UK10K's data management framework involved identifying likely challenges. In the initial stages, these included reaching agreement on the terms of data access and the recruitment of suitable (ethically compatible) sample sets, the release of data using an appropriately secure and managed mechanism, and the establishment of a system for sustainable data sharing once the project had ended. The recruitment stage of the project also involved a number of challenges that had to be resolved.

Recruiting participant studies

Agreement had to be reached at the outset regarding the content of the participant information sheets and consent forms, so that the use of samples in the project and the terms of data sharing were explicit. For example, there is often an assumption that if there is mention in the information sheets of research results being 'published', that this may infer that data can be stored indefinitely in an electronic archive and shared with other researchers, even without specific mention that the data will be stored in this way. Some studies had to be re-consented before they could be included in UK10K, and this provided an opportunity to design new information sheets that described precisely how the data generated by UK10K would be kept and shared with only *bona fide* researchers, thus removing any ambiguity or assumptions. It was also important to explain that although measures would be put in place to prevent the re-identification of research participants from the data, there would still be a small potential risk that this could happen.

Use of pre-collected samples within existing consents

Patients enrolled in the studies that were included in UK10K had consented to their materials being used in the study for which they were originally recruited. Nevertheless, there were differences between studies as to how 'broad' the existing consents were regarding the subsequent use of the samples in other projects. Some consents, for example, were study-specific and the samples could not be used in other projects without first gaining participant re-consent (as was the case for the ALSPAC samples used in UK10K). Other studies required that results could be shared but only via a managed data-access process. To address this, WTSI (as lead institution) sought signed assurance from sample custodians that the following common components of consent and/or REC approval were in place prior to accepting samples: i) appropriate consent and/or REC approval had been sought to allow samples to be used in UK10K; and ii) the linked anonymized data generated could be archived indefinitely and shared with researchers outside the project via the European Genome-phenome Archive (EGA) at the European Molecular Biology Laboratories (EMBL) Bioinformatics Institute, Hinxton, UK.

If sample custodians were unsure as to whether the consent provisions or REC approval covered the use of the samples or the sharing of data, they were required to amend their REC approval to do so. All collaborators in the project were keen to ensure that their REC approvals were updated, understanding and agreeing with the importance of this procedure and the need to make sure that the participants' trust was not jeopardized.

The need to review existing consents and REC approvals with potential collaborators and sample custodians as early as possible in the project rapidly became apparent, as this could have become a significant 'rate-limiting' step. The fixed-length project run time meant it was necessary to exclude some studies from UK10K as there was simply not enough time to re-consent participants, or to request that a local REC approve the use of sample sets collected outside of the UK. In both instances, time was lost in establishing the REC approval status of these samples, and then in having to source alternative available studies whose ethical approvals were compatible with inclusion in UK10K.

Dovetailing with existing systems

Different sets of samples had been collected by collaborators as part of distinct projects prior to any plan that they may be used in UK10K. The recruitment of research participants into the original studies had therefore taken place with no or little consideration that data arising from the use of these samples may be shared with other researchers unconnected with the immediate project into which research participants were being recruited. Ensuring that anonymized samples could be legitimately used in UK10K without compromising the original consent was a challenge for the RECs, who were approached to review and approve amendments to the original research proposals in order that sample sets could be included in UK10K. Nevertheless, every REC approached concluded that although many of the consents did not specifically state that the results of the research would be shared with other researchers, the potential 'harm' to the research participant did not outweigh the benefits of the research.

For the well-established and widely used TwinsUK and ALSPAC resources, approval and consent mechanisms were already in place.

ALSPAC permits the sharing of data generated from the analysis of ALSPAC samples via managed access, but only after approval has been sought by the ALSPAC Executive Committee (which is responsible for coordinating all requests for new and existing data collection). The 2,000 ALSPAC samples included in UK10K had to be re-consented so that the participants agreed to managed access of the data via the EGA. It was agreed that the ALSPAC Executive Committee would not require separate reports regarding requests to access ALSPAC data in UK10K (as would usually be the case), provided that they had a representative sitting on the UK10K Management Committee (MC) who would see all requests, and be in a position to report this back to the ALSPAC Executive Committee as required.

By contrast, TwinsUK operates a REC-approved opt-out system; research participants are informed about forthcoming research studies in which their samples may be used, including the data access processes linked to these prospective studies. Research participants in the TwinsUK registry automatically agree to be included in new studies, unless they specifically opt out.

Another example of reconciling existing mechanisms for managed data access with that of the project is provided by the inclusion of the Generation Scotland sample set. Generation Scotland is a multi-institution, population-based resource [13], and a condition of agreeing to UK10K's management of access to the data derived from their samples was that any requests to use these data would be reported back to a Generation Scotland sample resource representative, thus satisfying their own metrics for monitoring data usage.

Although TwinsUK, ALSPAC and Generation Scotland operated different approval systems, it was clear to us that a good, informed relationship between research participants and these projects was key. In addition, sample custodians and associated steering committees understood the benefit of sharing data with other researchers, despite their own investment in setting-up the initial projects and recruiting research participants.

Data release

Once the necessary approvals had been confirmed, samples were submitted to WTSI for genome or exome sequencing. It was decided at the outset that the EGA would house both all sequence data generated by UK10K and phenotype data for selected traits, such as height and body mass index (BMI), that were collected by both the TwinsUK and ALSPAC studies. Data were released to the EGA as a series of cumulative 'rolling releases' - so that as newer, more complete datasets for each sample set were generated, they were made available to approved data users as soon as possible. It was initially hoped to release data on a quarterly basis, but making releases at such specific intervals ultimately proved impossible because of the increasing length of time required to prepare progressively larger and more complex datasets for EGA submission. Instead, datasets were released as soon as it was possible to do so. So as not to hinder project progress, the collaborators (sample custodians) who had provided samples and were collaborating on primary analyses with WTSI-based researchers were able to access the data internally via a secure FTP site rather than waiting to access data through the EGA.

Each time a new data release was made, the EGA would send an automated email alert to all approved data users with existing access rights, informing them of the need to update their password. Initially this caused some confusion in the data management process, as it was not clear to data users why they were being prompted to change their login details. This was quickly resolved when the EGA amended the alert to explain that UK10K studies had been updated with additional dataset(s) and that a new password was required in order to maintain security, prompting data users to access the latest version of the data.

To protect study participants, a great deal of effort was invested in protecting the privacy of the data. Some of the disease phenotypes included in UK10K were so rare or at such extreme ends of a disease spectrum, and the number of documented cases were so few, that these patients could potentially have been at risk of re-identification. It was agreed at the start of the project that sequence data would only be made available to third-party researchers outside the project in a linked anonymized (coded or pseudonymized) form, and that samples would be provided to UK10K by collaborators (sample custodians) in a linked anonymized form with the linkage key retained by the sample custodian. This ensured that the collaborating principal investigator (sample custodian) would be the only party able to link data back to the research participants' identity. This included, but was not exclusive to, situations in which the collaborator also held a clinical duty of care to the research participants. In this way, patient identity was protected both within and outside of the project when the data were released to the EGA. This system also enabled clinically significant findings to be returned to participants, after a number of checks had been carried out according to the UK10K policy [2].

Data access

The commitment to data-sharing principles in research requires a considerable amount of time and resources to manage requests for access. There were two bodies that were responsible for oversight of UK10K data-access requests from researchers outside of the project (a model based on the Wellcome Trust Case–control Consortium (WTCCC) Data Access process); the MC and the Wellcome Trust Data Access Committee (DAC). Composed of representatives from each of the cohort and disease groups, the MC acted as the first formal stage of review for applications to use project data. To ensure the most efficient use of MC time, applications were first checked by the UK10K project manager (PM). This informal review at the point of submission ensured that all requests brought before the MC were valid, correctly executed and respected the constraints of dataset usage. The DAC was set up by the Wellcome Trust to be independent from the project and, once approved by the MC, applications were forwarded onto the DAC for final, formal approval. The DAC were then responsible for requesting that the EGA create the appropriate user accounts for the data requestors.

As already mentioned briefly, researchers wishing to use UK10K data could contact the UK10K PM as the designated, visible point of contact with any queries regarding data usage, before submitting their request in the form of a completed UK10K data-access application (designed specifically for the project). The preliminary (informal) review of all data access applications included a number of considerations, which are summarized in Box 1.

Having confirmed that an application was satisfactorily completed (and if necessary contacting the applicant to resolve any issues), the PM would then inform the MC of the application. Provided that no queries or issues were raised by the MC, the PM would then send the request onto the DAC who, acting independently of the project, would consider the request in more detail and approved it, request that the PM contact the applicant to address any concerns raised or rejected it. Once approved, the DAC would contact successful applicants to confirm this outcome and instruct the EGA to create an account with appropriate access permissions for all approved data users. It was proposed that there should be a maximum two-month turnaround time from receipt of a data-access application to the issuing of EGA login details, though, in practice, the procedure sometimes took longer than this, in part because the DAC met less frequently to review applications than was initially anticipated.

The publication moratorium

A publication moratorium was put in place to protect the first publication rights of the project's researchers, sample custodians and data producers. The terms of the moratorium were explained as part of the UK10K Publication Policy (Appendix B of the data access application), and state that 'All data will no longer be subject to the publication moratorium once the data have been published, or until one year has passed since the full dataset required for analysis was released'. Once data began being submitted to the EGA, it quickly became apparent to the MC that a more precise definition was required for what constituted a dataset that was 'full' enough to enable meaningful analyses. It was agreed that the moratorium 'clock' would start from when the majority of samples (≥90%) had sequence data available in the EGA. For the cohort studies (TwinsUK and ALSPAC), the publication moratorium on the whole-genome datasets expired on 2 July 2013, and for all other exome studies, the moratorium expires on 2 January 2014. To ensure complete clarity regarding the expiration of the publication moratorium, these dates were included in the project's data access application document, were posted on the project website and were also added to the dataset descriptions in the EGA.

The first consequence of violating the moratorium, or indeed breaching any term of the UK10K data-access agreement, is that the DAC and/or the MC would instruct the EGA to terminate data access immediately. Once a breach had been confirmed, the DAC and/or MC would contact the appropriate journal requesting that any manuscripts using UK10K data be withdrawn. The project took its obligation to protect the publication rights of the researchers, sample custodians and data producers very seriously. As part of the managed-access process, the DAC was asked to stress the requirement for strict adherence to the publication moratorium, particularly to those data users approved for access during the early stages of data availability. Periodic checks of the literature were made by the PM to ensure that UK10K data had not been published prematurely (pre-moratorium expiry), or used for purposes other than those on which the basis for data access was granted.

Succession planning

As a project with a fixed duration of 42 months, it was essential that UK10K also developed a plan for the

Box 1. The UK10K project manager's review

- Was the document completed satisfactorily?
- Were contact details provided for all listed prospective data users?
- Was the application co-signed by an authorized representative of the data user's institute?
- Were the recent publications listed by the lead applicant authentic?
- Was there an adequate description of the proposed research?
- Were the datasets listed in the application appropriate for that research?
- Would use of those datasets in that way conflict with or violate any constraints on their terms of use? (For instance, using datasets as controls where the data access agreement specified that to do so was not permissible according to the consents attached to that dataset.)

Table 1 Setting up a managed access data resource: project checklist

Pre-project

✓ Have all pre-collected sample collections been identified for inclusion in the project?

✓ For these studies, do the sample custodians agree in principle to the potential inclusion of their sample sets in the project?

✓ For those studies willing to participate in the project, have the sample custodians confirmed total sample numbers available within the required timeframe for the project?

✓ Have they also confirmed the REC approval status for these samples?

Specifically has the sample custodian provided signed assurance that:

- appropriate consent and/or REC approvals are in place for the use of samples in the project, and are compatible with the terms of data sharing proposed by the project;
- where appropriate consents and/or REC approvals are not in place, there is sufficient time to correct this within the timeframe of the project (if not, these studies should be excluded at the outset);
- a mechanism is in place for sample custodians to withdraw the use of their samples;
- the potential risks of participant identification (however minimal) have been explained;
- all of these points have been documented in, for example, an Ethical Governance Framework, and is this document available to data users both inside and outside the consortium.

Data collection stage

✓ Has it been made clear to the sample custodian that DNA samples must be submitted in a linked (coded) anonymized form, with the sample custodian retaining the linkage key?

✓ Do the sample custodians agree with the timeframe for submitting their samples, and the amount and quality of material required for sequencing?

Preparing for managing data access

✓ Has a comprehensive, project management committee-approved data-access document been prepared that includes:

- a description of all available datasets and any constraints on the use of the data;
- the project's publication policy;
- a clear explanation of the publication moratoria (if applicable) and its expiry dates;
- a named point-of-contact to whom completed applications (and any queries) should be submitted; in the UK10K example this was the project manager (PM).

✓ Has the data access document been made available for download at the site of released data (that is, the resource; in this case the EGA), and/or on the project website?

✓ Has a person or group independent from the project been identified and appointed to function as a Data Access Committee (DAC)? Have they been briefed on the terms of data access?

✓ Is there a mechanism in place to arrange for approved data users to access the datasets in the resource once the DAC has approved the data request? If not, this needs to be put in place.

Data release stage

✓ Have the data been deposited into the Resource in a linked anonymized (pseudonymized) form, so that third-party data users are unable to identify study participants?

✓ Is the project tracking applications that are made to use project data? (For UK10K, applications were monitored by the PM from the point of receiving an application through to the DAC notifying the PM that the application had been formally approved.)

✓ Providing information as to how datasets are being used may be a condition of some studies' permitting the inclusion of their samples in the project; and failure to do so may result in the dataset being withdrawn.

✓ Are approved data users notified as and when additional datasets are added to the resource?

Post-project

✓ If the project is of a fixed duration, has a mechanism been put into place for managing data access once the project's management structure dissolves?

sustainable management of applications to use project data once the MC dissolved at the end of the project. To this end, it was agreed that in the absence of a PM and MC, all data access applications would be submitted instead to the WTSI's newly designated Data Access Officer. This officer would review proposals, seek DAC approval, and liaise with the EGA to create and maintain data-user accounts. To ensure a smooth transition, this handover process was initiated several months before the close of the project. Additionally, contact details for the sample providers (sample custodians) were added to each of the UK10K studies in the EGA and were also listed with the sample set descriptions on the project website, so that data users could direct queries to an appropriate contact after the MC had disbanded.

Conclusion

As advances in technology further reduce the cost of genomic sequencing, the more accessible and widely used this technology will become, generating ever larger and more complex datasets. The UK10K managed-access system is one example of how it is possible to release data widely while still respecting the contributions of all those involved in the project and the regulatory requirements. While funders advocate data sharing, there are no hard rules governing how groups should manage and share the data generated. Endeavors such as UK10K are, by their very nature, unique - driven by the particular participating institutes and their local policies, influenced by the varying demands of funders, and required to adhere to the specific ethical constraints of consent and/or REC approval attached to the use of participant materials (be that biological material and/or data). As such, the UK10K model is not intended to be taken as an example of recommended practice, but rather a collection of lessons learned that could prove useful to others undertaking a similar approach to data sharing. Table 1 is a checklist that captures the crucial questions that, from experience, are useful to raise and address when establishing a managed data-access system.

A great deal of anticipation, planning and time were required to set up and then successfully deliver a genomic sequencing project of UK10K's scope and magnitude. Having a designated PM role was crucial to coordinating the implementation of data-management processes as well as to the day-to-day running of the project. A wider group of people (other than researchers and sample providers) were involved in the project than might have been expected; these included administrative staff, regulatory and policy advisers, database and IT support, patient representatives, an independent DAC, an Ethics Advisory Group, the MC and also a project publications committee.

Finally, it should be noted that UK10K has successfully achieved its two core goals: rare and low-frequency variants associated with disease have been detected (primary papers describing these results are currently in preparation; with secondary consortium papers already published to the project website [2]); and a managed access data resource has been produced that will benefit the genetics research community around the world. At the time of writing, over 75 applications have been received and approved, allowing successful applicants to download UK10K data via the EGA from researchers spread across the global research community: in Taiwan, Canada, China, USA, Australia and Europe (specifically the UK, Finland, the Netherlands, France, Switzerland and Austria). Once the project's primary results are published and UK10K is increasingly cited in the literature, we anticipate a rapid increase in the demand for use of the data as awareness of the data resource spreads.

Maximum use of the samples donated has been ensured by sharing the data with researchers outside of the consortium. Furthermore, in making the data available on a rolling basis as they are generated, rather than at the project's end, this benefit has been realized immediately.

Abbreviations

ALSPAC: Avon Longitudinal Study of Parents and Children, Bristol University; DAC: Data Access Committee; EGA: European Genome-phenome Archive; MC: UK10K Management Committee; PM: Project manager; REC: Research Ethics Committee; SNP: Single nucleotide polymorphism; WTSI: Wellcome Trust Sanger Institute.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Funding for the UK10K Project was provided by the Wellcome Trust under award WT091310. JK is funded under Wellcome Trust Award 096599/2/11/Z. A full list of consortium members can be found at the UK10K Project website [2].

Author details

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]Centre for Health, Law and Emerging Technologies (HeLEX), Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK. [3]The UK10K Project.

Published: 15 November 2013

References

1. Dyke S, Hubbard T: **Developing and implementing an institute-wide data sharing policy.** *Genome Med* 2011, **3:**60.
2. **The UK10K Project** [http://www.uk10k.org/]
3. Kaye J, Hurles M, Griffin H, Grewal J, Bobrow M, Timpson N, Smee C, Bolton P, Durbin R, Dyke S, Fitzpatrick S, Kennedy K, Kent A, Muddyman D, Muntoni F, Raymond LF, Semple R, Spector T: **The UK10K management pathway for the return of clinically significant findings.** *Eur J Hum Genet*. in press.
4. **Summary of Principles Agreed Upon at the First International Strategy Meeting on Human Genome Sequencing (Bermuda, 25–28 February 1996) as reported by HUGO** [http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1]
5. **Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA** [http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf]
6. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, Austin CP, Berglund L, Bobrow M, Bountra C, Brookes AJ, Cambon-Thomsen A, Carter NP, Chisholm RL, Contreras JL, Cooke RM, Crosby WL, Dewar K, Durbin R, Dyke SO, Ecker JR, El Emam K, Feuk L, Gabriel SB, Gallacher J, Gelbart WM, *et al*: **Prepublication data sharing.** *Nature*, **461:**168–170.
7. **Wellcome Trust Statement on Genome Data Release** [http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002751.htm]
8. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.** *PLoS Genet*, **4:**e1000167.
9. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: **Identifying personal genomes by surname inference.** *Science*, **339:**321–324.

10. Hayden EC: Privacy loophole found in genetic databases. *Nature News* 2013. 17 January.
11. The TwinsUK Registry [http://www.twinsuk.ac.uk/]
12. The Avon Longitudinal Study of Parents and Children [http://www.bristol.ac.uk/alspac/]
13. The Generation Scotland Resource [http://www.generationscotland.org/]

doi:10.1186/gm504
Cite this article as: Muddyman *et al.*: Implementing a successful data-management framework: the UK10K managed access model. *Genome Medicine* 2013 **5**:100.