Imperial College Healthcare Tissue Bank
Green Shoots Grant Outcomes Report

A Research Data Management (RDM) "Green Shoots" Pilots

Project Report by Geraldine Thomas, Sarah Butcher and Christopher Tomlinson

Imperial College London

This project was funded as part of Imperial College's RDM "Green Shoots" Programme. In 2014, the Vice-Provost, Research, approved an allocation of £100K for academically-driven projects to identify and generate exemplars of best practice in RDM, specifically frameworks and prototypes that would comply with key funder RDM policies and the College position. The call for projects outlined that frameworks could be based either on original ideas or integrating existing solutions into the research process, improving its efficacy or the breadth of its usage. There was an expectation that solutions would support open access for data; solutions that supported Open Innovation were strongly encouraged.

Six projects were funded, covering different disciplines, faculties and research areas. The projects ran for six months, finishing at the end of 2014. Project reports were made available in 2015.

For more information on the programme and projects please visit:

http://www3.imperial.ac.uk/researchsupport/rdm/policy/greenshoots

Imperial College London

Imperial College Healthcare Tissue Bank Green Shoots Grant Outcomes Report

PI - Prof Geraldine Thomas (Dept. Surgery and Cancer) together with Dr Sarah Butcher and Dr Christopher Tomlinson (Bioinformatics Support Service, Dept. Life Sciences); Imperial College London

Introduction

Tissue banks collect and store biological material in different formats obtained from healthy individuals or from patients. However, this material is only useful for research if associated with data. The richer the dataset that accompanies the sample, the more useful the sample is for research purposes. This data can include information on variables that affect the suitability of the material for different types of analysis, data on the phenotype of the disease, the individual, and increasingly the results of clinical tests, including genotypic information. Obtaining data on treatment and outcome can be challenging with a mobile population. A patient may start their journey in one NHS Trust and end it in another. The majority of NHS Trusts do not have data sharing agreements in place, which means negotiation with each Trust for access to data. One possible solution is to link to national databases that record longitudinal outcome data for individual patients, such as the National Cancer Registry Service.

One of the richest sources of data to enrich annotation of remaining samples from the same individual is often not collated – data generated from research use of samples. Most tissue banks provide analysis-ready samples for use by researchers (extracted nucleic acids, tissue sections) as a way of making a limited resource go further. This results in multiple research groups being provided with samples from the same patient, and even from the same piece of tissue. Tracking research data back to individual samples and individual patients is a challenge, but one that can provide a rich source of varied data for secondary use by bioinformaticians interested in developing algorithms for systems biology approaches to understanding disease processes.

The Imperial College Healthcare Tissue Bank (ICHTB) is a diverse collection of physical tissue specimens that have been collected from operations on patients in the hospitals trust. At the current time, there are approximately 60,000 available samples from around 15,000 different donors and over the past ten years more than 20,000 tissue bank samples have been issued to researchers. Information about the specimens is contained in an online database system. The system contains detailed, anonymised records about donors, operations and samples including pathology reports and allows searches of the collection to take place.

Objectives

In spring 2014, the ICHTB team was awarded a Green Shoots grant to augment the existing ICHTB database with further information that would sit alongside the patient and sample record. In our proposal we focussed on gathering data from recent experiments that have taken place on samples

from the collection and recording this alongside additional information from the patient record. The integration of external information greatly enhances the utility of the samples in the collection.

Process

In the initial stages we carried out a survey to identify the types of experiments that are currently or typically carried out on samples from the ICHTB. This initial survey identified a wide range of different and specialised data types, each with a requirement for different associated formats and metadata. For the scope of this project we narrowed our search to look for data types that are routinely created from tissue bank samples, to act as an exemplar. This approach would have the most impact in the project lifetime, as an import pipeline for a popular data type could be reused every time data of that type is generated.

Within the Imperial Hospitals NHS trust, it is now policy to sequence tumour samples for certain types of cancer to aid diagnosis and treatment decisions. This approach is currently used for lung cancer samples and (we feel) is likely to be extended to other cancers in the short/medium term. Currently, a targeted sequencing gene panel approach is performed on the IonTorrent sequencing platform and after considerable analyses, a standard format report containing information about suspected mutated genes is produced by the laboratory staff. We were able to exploit existing contacts within the NHS trust to meet with the laboratory staff to gain access to sample data and to aid our understanding and interpretation of the current sequencing reports. It is anticipated that this working relationship will extend beyond the end of the RDM grant and into the foreseeable future.

Alongside the ongoing work on sequencing data, we were able to use the funds from the Green Shoots grant to fully implement a link with the National Cancer Registry. The purpose of this strand was to augment the existing donor information in the tissue bank with outcome information about the donors. Outcome information greatly enhances the utility of the samples for a given donor, particularly if the cause of death is recorded as it is in the NCR records. Collecting this information will enable researchers to examine genetic markers for survival rates in particular cancer types as those patients that have died of a condition can be separated from those that have survived, or have died from an unrelated cause.

Putting the two areas of work together will, at some future point, give the ICHTB a subset of donors whose outcome is known as well as sequencing information on samples taken from them. We believe that this will be of great use to the medical research community at (and beyond) Imperial College.

Results

We were able to first build a tool for automatically exchanging information with the National Cancer Registry and then to register 1884 of our donors with them. The tool automatically sends messages about available samples to the NCRS and receives and interprets the reply. The ICHTB database is updated with the outcome data where it is known about a patient. Primary and secondary causes of death are recorded where they are present and this information is presented back to the user alongside the donor record in the ICHTB user interface. Of the 1884 patients registered with the NCR we found that 572 were identified in their records. Of these 408 were still alive at the time of writing and 164 were dead – the rest of the set are currently unidentified by the NCR. This system is now actively in use and will continue to be used to update these links between ICHTB and NCR data.

Once the NCR data exchange tool was built and deployed, we turned our attention to the more complex problem of handling our prototype direct sequencing datasets. Each analysis of this type yields multiple levels of data, with differing degrees of usefulness for different audiences, also differing perceived levels of reuseability for future research. We discussed how we could best represent data arising from these analyses in a way that was useful in the ICHTB context. We received a representative data report type from the NHS trust laboratory and set about building a pipeline to import this information directly into the ICHTB, together with all relevant metadata required to track their provenance. A pipeline has been built and a prototype user interface to view these derived sequencing data has been constructed. As an adjunct, we have explored how raw data pertaining to the sequencing reports are currently stored (the aligned reads and raw non-filtered variant calls). These files can be large (multiple gigabytes) and should not be copied unnecessarily. In future, we expect to extend the ICTB interface to allow tracking of the raw data files themselves as they are stored and archived separately by the generating laboratory. This will increase the potential for later re-use. This work is still at the prototype stage and is not yet being used with real data. Work on this aspect of the ICHTB will continue beyond the end of the RDM funding period.

To assist with tracking data outcomes arising from ICHTB-associated projects, we also started to investigate ways of improving the linkage between the ICHTB and the research papers arising from ICHTB tissues used in approved research projects. It is a requirement that all approved projects report back to ICHTB and presently, publications are reported back periodically and made available by the ICHTB as a static list. We started to consider how these data could be more effectively linked into the data holdings associated with the ICHTB. Whilst publications can be easily linked to specific studies, each study may utilise many different samples from many different patients. Moreover, there is no trivial link between reported results and any single sample. We considered linking these papers into the university's research data repository Spiral. This proved to be complex for a number of reasons; including formulating the most appropriate mapping level between the publication and the groups of samples used; the fact that some of the papers are linked to external authorship and may not be present within Spiral and furthermore that Spiral has no publically accessible API that could be queried to make the link. Within the scope of the project, this line of work was not pursued further but will be a point for further work.

Impact

This project has established a working mechanism that enables linkage with the National Cancer Registry for all cancer patients who have donated material to the Imperial College Healthcare Tissue Bank (ICHTB www.imperial.ac.uk/tissuebank). The mechanism preserves patient confidentiality, and has been approved by the NHS Trust Caldicott Guardian. In addition, we have developed a pilot system to add data obtained from existing routine next generation sequencing to individual tissue samples held in the bank. This type of linking is becoming ever more important in light of a number of recent developments in this area, (such as Genomics England and other large initiatives) where samples and routine local screening and QC data arising from them may need to be routinely tracked with respect to results returned from externally run analyses.