



# NCRI INFORMATICS INITIATIVE

**NCRI Informatics Initiative  
Integrating Clinical Trials and Functional Genomics Workshop  
Royal Society of Medicine, London  
31st January – 1st February 2005**



Functional genomics studies are providing new information about cancer at an ever increasing rate. While there is outstanding potential for application in cancer prevention and care it is becoming increasingly clear that interpretation of data is a complex problem which needs to be approached in a systematic way if valid conclusions are to be drawn. Clinical trials in which there are defined populations, and careful quality control, provide an ideal situation in which to investigate the limitations and power of functional genomics data. There are many potential sources of error in bringing the diverse molecular and clinical data together and standards are required for data acquisition, and the management of data at the molecular level (Bioinformatics) and the clinical level (clinical informatics). The workshop brought together a diverse group of people including functional genomics and clinical investigators, bioinformatics and clinical informatics experts and computer scientists in order to define the issues and determine a strategy for progress towards optimal exploitation of functional genomics in cancer research.

## Summary of Presentations

The workshop began with a review of the current state of functional genomics in the clinical domain from James Brenton. The types of genomics platforms being used in clinical cancer research include expression microarrays, array comparative genomic hybridisation (array CGH), tissue microarrays and arrays for mutation analysis. It is expected that this technology will lead to better prognostic and predictive markers and so enable the goal of individualised cancer treatment.

A number of factors currently affect how this goal is reached. Difficulties in combining current clinical studies in a systematic way include methodological inconsistencies and lack of important clinical variables in the studies. Microarray technology itself has a number of inherent sources of variability including the probe preparation and labelling, the platform used, the reference sample chosen and effects of choice of segmentation method during scanning. Study design is key to a meaningful array-based clinical project and problems affecting the field include cohort selection, statistical analysis, validation of results and reporting of raw data.

Dipak Kalra's review of clinical information standards and an electronic health record (EHR) then posed a number of questions about how the emerging genomics fields would impact on routine clinical care. In terms of an EHR: how does functional genomics research fit in with patient care: what do we need to capture in the lifetime of a patient; and what clinical information does genetic research need to know about its human data subjects? Furthermore, which standards are relevant for the EHR, for terminologies, ontologies and ethico-legal standards? Activities such as openEHR ([www.openehr.org](http://www.openehr.org)) and the CLEF project ([www.clinical-escience.org](http://www.clinical-escience.org)) are addressing some of these issues and learning from their experiences will be critical. Challenges remaining for integrating research data with an EHR include separation of knowledge, reference data and salient health record data, avoidance of simple models of phenotypes and a recognition that clinical data needs to meet ethical and legal requirements while being interoperable.

The workshop continued with a description, by Sue Dubman, of the work of the NCI Centre for Bioinformatics and how it fits with the greater NCI vision of integrated, targeted cancer prevention and care. The challenges are similar to those faced outside the US – lack of consensus on common standards and terms, fragmented locally used systems that are rarely interoperable and systems built to variable standards of quality. The NCICB aims to provide foundational biomedical informatics infrastructure, tools, applications and data to NCI research initiatives and to the cancer research community. The Cancer Biomedical Informatics Grid (caBIG) initiative will facilitate the sharing of infrastructure, applications and data, the key goal being interoperability (the ability of multiple systems to exchange information and to be able to use the information that has been exchanged). The challenges faced by caBIG and its related developments include setting realistic expectations, understanding user requirements, being secure and private while providing open access and 'eating the elephant one bite at a time'.

Phil Quirke explained some work at Leeds University where pathology images are captured in a high-resolution digital format and stored in a tissue microarray database in a standardised way and can be linked to other genomic and clinical outcome data. In collaboration with a number of other centres a project is underway to link pathological image data with other clinical images such as radiological and MRI scans.

The sharing of microarray data is an established principle now and Alvis Brazma described some of the standards and infrastructure for doing this. The Microarray Gene Expression Data (MGED) Society have successfully implemented MIAME with support from the publishing journals and further development has led to the markup language MAGE-ML and the MGED ontology MO supporting deposition of the data in databases such as ArrayExpress. Health Level 7 (HL7) is a healthcare messaging standard organisation enabling disparate healthcare applications to exchange key sets of data. Amnon Shabo explained that a Clinical Genomics special interest group of HL7 is working on how to represent core genomic data with objects of the Reference Information Model in HL7 – the Genotype Shared Model. Part of this work will investigate how to integrate with established bioinformatics models such as MAGE-ML.

The second day focused on the solutions to the obstacles to sharing and integrating clinical trials and functional genomics data. The NCRI Informatics Planning Matrix and progress of data sharing pilots being undertaken by NCRI Partners were described by Peter Kerr and the concept of a Platform Reference Model for the cancer informatics domain was introduced. Anthony Finkelstein exposed some of the dangers and challenges of undertaking a large-scale software project and stressed the importance of a sound architecture in such projects.

The CancerGrid E-Science project, which is due to start in May 2005, was presented by Jim Davies, Sylvia Nagl and Steve Harris. This project will endeavour to address some of the issues described in the workshop through provision of open informatics standards for clinical trials and development of GRID-enabled tools to manage clinical trial data. Part of this work will include the development of ontologies for clinical trials in collaboration with the NCICB.

Subha Madhavan presented the Rembrandt and ISPY Trial studies which are combining molecular and clinical data in the context of brain neoplasia and breast cancers respectively. Both leverage NCICB and caBIG infrastructure components and it is envisaged that these will be two of the studies that will be built on the caIntegrator framework which will provide a mechanism for analysing and aggregating biomedical research data in an integrated fashion. Hani Gabra and Tim Aitman explained how Imperial College and the West London Cancer Network are using a number of genomic profiling techniques to develop better treatment for ovarian cancer through both cohorts from clinical practice and clinical trials. The NEAT project aims to define the fields that represent ovarian cancer and define these using controlled vocabularies and ontologies. The CSC/Imperial Microarray Centre is developing clinical microarray data warehouse to enable meaningful queries and generate new knowledge through this data integration.

A number of clinical studies in breast cancer are underway to investigate whether genetic or genomic tests can improve treatment. James Mackay described how patients with an inherited BRCA2 mutation will be tested for sensitivity to platinum in a clinical trial developed in conjunction with the Breakthrough Breast Cancer Centre. London Genetics is a collaboration between several London medical schools and academic institutes which will offer genetic services within clinical trial activity and will align itself with national initiative such as the National Cancer Tissue Resource. The TRANSBIG project is a large breast cancer clinical study that is attempting to validate a 70-gene microarray signature to help avoid unnecessary chemotherapy following surgery. Thierry Sengstag talked about issues such projects faced including variability of the technology across different centres and the lack of standardisation of clinical information.

Monica Jones presented the National Cancer Research Network (NCRN) work in developing electronic remote data capture (eRDC) for clinical trials. A proof of concept pilot project has been completed successfully where a number of NHS trusts and clinical trials units assessed functionality and usability. The second stage will involve approximately 30 centres and will also involve a non-oncology based trial to demonstrate generic applicability. Any national system will obviously have to tie-in with the NHS National Programme for IT and the principles and processes of eRDC across the NCRN will be established independently of any particular application solution. This project will examine integration with the CancerGrid project in 2006.

## **Barriers and solutions to integrating clinical and genomic data**

Breakout groups were convened so that each group had a diverse mixture of expertise (clinical, bioinformatics, computer science etc). Three distinct groups addressed the barriers and solutions to integrating clinical and genomic data. Their findings fell into a number of categories.

To address infrastructure issues the question 'why not have one central database?' was posed to the groups. Issues identified included security, ownership, curation, control (technical, access and quality), risk, cost, feasibility, and compatibility with recent legacy, flexibility, extensibility, and evolution. The physical location also brings similar issues. It was thought that centralised and federated approaches should coexist and architecture should support both. It was pointed out that assets are not just technical but also people and expertise. We also need to know the necessary level of interoperability and so requirements and values need to be determined. One model is that of the search engine Google where web pages are distributed but the web index is centralised.

The economics of data integration raised further questions and answers. Do we need to know what we currently spend on informatics? The success of networks such as NCRN means that larger trials are being conducted which demands greater informatics infrastructure – this is combined with the increase in genetics/genomics data which trials units aren't necessarily equipped to handle. What do we need to do now to ensure support for projects that develop infrastructure? The NCICB model was discussed where members of the community tender to provide outputs subject to conditions. Engagement with NPfIT, HL7, Snomed involves a FTE commitment in a project ie resourcing/training for connecting with other initiatives. The goal is to develop an infrastructure for improving biomedical research – is it interesting for computer scientists or lab or clinical scientists? Challenges are interesting but it's not research. Need funding specifically for this work and appropriate acknowledgments must be made. The research councils and RAE should give better recognition for multi-disciplinary research. There is a training need – specifically training scientists to understand the language of other disciplines.

The cultural and sociological issues in this initiative are at least as important as the infrastructure and economic ones. The requirement to establish methods of -de-identification of patient data was given a high priority by participants and strategies for moving forward collectively were suggested including dealing with ethics committees in a unified way and addressing issues together as part of NCRI. It is clear that different levels of access agreements may be required for different data sets and involvement of patient representatives from the beginning is crucial. A research community consensus on data release needs to be established where possible across these domains and there is also a need to engage journals and funders in enforcing this. The difference between access to data and samples should be clarified – data persists, samples deplete and this needs to be recognised.

In summary, there is an increasing requirement to meaningfully integrate genomics and clinical data. There are a number of technical, cultural and infrastructural issues to be addressed so that this can happen. However, a number of projects are beginning to address these issues and solutions are beginning to emerge. The NCI Centre for Bioinformatics and NCRI Informatics Initiative are national organisations that are addressing some of the issues with the research communities. A number of participants have established collaborations through the workshop and participants agreed that that the NCRI Informatics Unit should continue to provide leadership for coordination of data standards through other workshops and the Informatics Planning Matrix.

## Workshop Participants

<b>Name</b>	<b>Organisation</b>
Alan Rector	Computer Science, Manchester
Alexander Voss	Informatics, Edinburgh
Alvis Brazma	EBI
Amnon Shabo	IBM Research Lab in Haifa
Ankur Agrawal	MRC/IC Microarray Centre
Anthony Finkelstein	Computer Science - UCL
Carlos Caldas	Oncology, Cambridge
Christopher Wroe	Computer Science, Manchester University
Dipak Kalra	Health Informatics, UCL
Euan Stronach	Oncology, Imperial College
Fiona Reddington	NCRI Informatics Initiative
Gavin Kelly	Cancer Research UK, London
Hani Gabra	Oncology, Imperial College
Helen Parkinson	NCRI Informatics Initiative & EBI
Ian Tomlinson	Cancer Research UK, London
James Brenton	Oncology, Cambridge
James Mackay	Clinical Genetics, London
Jane Cope	NCRI
Jim Davies	Oxford University Computing Lab
Jonathan Ledermann	CR-UK & UCL Cancer Clinical Trials Centre
Julie Hearn	Cancer Research UK, London
Justin Hinshelwood	MRC
Malcolm Mason	Wales Cancer Bank
Matthew South	CR-UK & UCL Cancer Clinical Trials Centre
Max Parmar	MRC Clinical Trials Unit
Max Wilkinson	NCRI Informatics Initiative
Monica Jones	NCRN
Pat Soutter	Oncology, Imperial College
Paul Mason	Cancer Clinical Trials Unit, Birmingham
Phil Quirke	Pathology, Leeds
Richard Begent	NCRI Informatics Initiative
Stephen Turner	NCRI Informatics Initiative
Steve Canham	Institute of Cancer Research
Steve Harris	Oxford University Computing Lab
Subha Madhavan	NCI Center for Bioinformatics
Sue Dubman	NCI Center for Bioinformatics
Sylvia Nagl	Oncology, UCL
Tim Aitman	MRC Microarray Centre, Hammersmith Hospital
Thierry Sengstag	Swiss Institute for Bioinformatics
Tom Freeman	MRC RFCGR
Ugis Sarkans	EBI