



**Report from the NCRI Informatics Initiative
Genetic Variation Workshop
Royal Society
3 February 2006**

The NCRI Informatics Initiative was formed to address the problems of data sharing and integration in the cancer domain. Several areas were identified for action including the area of Genetic Variation data standardization, access and integration in the cancer domain. This workshop addressed these areas and posed several questions for discussion by participants. Participants from across the cancer domain were selected to ensure that the full range of NCRI partner activity was represented and speakers were selected in consultation with the NCRI task force mandated by the NCRI board to advise the Informatics Initiative.

Summary of Presentations

All workshop presentations are available at:

<http://www.cancerinformatics.org.uk/workshops.htm#hgv>

Richard Begent Director of the NCRI Informatics Initiative (NII) provided an introduction to the activities and mission of the NII.

Richard Wooster (Wellcome Trust Sanger Institute) provided an overview of the current data resources, standards and emerging data in the Genetic Variation domain; he also provided detail on the cancer specific COSMIC mutation database held at the Wellcome Trust Sanger Institute. <http://www.sanger.ac.uk/genetics/CGP/cosmic/>

Steve Walker (UK Biobank) provided an overview of Biobanks activities in acquiring samples and data and the IT provision required.

Rakesh Narajan of University of Washington (St Louis) and a participant in the caBIG caTissue project provided an overview of a tool 'Mutation Viewer Pipeline' to support mutation analysis data acquisition, analysis, and reporting and visualisation.

Chris Mattocks of the National Genetic Reference Laboratories provided an introduction of the needs of the diagnostic genetic labs including clinical reporting, and local data storage in and between the various NGRL labs.



Anthony Brooks (Univeristy of Leicester) presented a summary of genotype-phenotype challenges and solutions and introduced the PML – polymorphism markup language a new standard for representing polymorphisms. It was noted that this standard lacked any semantic content, and that a new version PML would be developed to remedy this. ([PML publication link:](#))

Tina Boussard (Stanford Univeristy) provided a presentation on the development of data standards for emerging technologies based around the needs of the PharmGKB database which incorporates data on mutations, pharmacokinetics and integrates these with HTP genotyping information generated on arrays. She highlighted the need for semantic annotation of SNPs.

Adrian Moody from Astra Zeneca provided an industry perspective on the generation and use of genetic variation data within industry, their need for access to public data in the drug development process.

Breakout Sessions

Breakout groups were convened so that each group had a diverse mix of expertise (clinical, scientific, technical, etc). Three groups were chaired by selected workshop attendees (Paul Lewis, Anthony Brookes, Rakesh Nagarajan and facilitated by members of the NCRI Informatics Initiative Coordination Unit. Some points for discussion were provided in advance for the working groups to discuss, but questions and suggestions were also taken from the floor at the end of the presentations and groups were able to add their own topics for discussion during the workshop session time permitting.

Discussion topics

Discussion topics are shown in bold, and the working group conclusions are shown in context in italics

1. Standards

Q. What is the awareness of current standards, who is using these, and what are the limitations?

- *Standards are important but must come from the community*
- *Semantic standards are a good place to start*



- *To get infrastructure and data models funding is key*
- *Stakeholders and journals play a key role in encouraging and enforcing use of standards e.g. adherence to MIAME for microarray experiments and sequence deposition*
- *Phenotype is a particularly difficult thing to describe in a standard way and the current work is incomplete*
- *Lessons can be learned from the rare disease world as through necessity they are already sharing data.*
- *Standardization should be targeted to areas of high impact such as improving the power of meta analysis.*

2. Resources

Q. Identify useful resources and their utility, what are the current limitations?

- *Just finding and book-marking data is not enough. We need a second generation of tools that allows layering of public/private data so that the totality of data can be used across SNP (genotype) and phenotype data.*
- *Ontologies are needed and must be used correctly*
- *There is a changing concept of what a database is – journals and database line is becoming fuzzy*
- *Managing and using high dimensionality data is problematic, cancer is a good model, but any work should not be limited just to cancer*
- *caBIG paradigm works, is semantically typed, feeds into semantic standards, but these need to be available early.*
- *PML (Polymorphism Mark-up Language) – is not semantically typed, for data exchange need both an OM (as PML has) and an ontology in provided via infrastructure like EVS.*
- *In the clinical space the HL7 SIG is working on genotyping/pedigree modelling. This model could be leveraged.*
- *Data provenance is important so that the data can be trusted, and raw data must be available*
- *caBIG software is certified, some NCRI certification might be considered and could extend the utility of the NCRI [Planning Matrix](#)*
- *Cutting edge technology and data generation should not be forced into existing standards*
- *Both open source and proprietary models can work, support should be provided for both, core models etc should be open source, applications that support these do not have to be*
- *Raw data is needed especially for HTP experiments*



- Training sets of data are useful for clinicians in the areas of standardization of collection methods, describing phenotypes.
- *Not all data is worth preserving, the 80 percent solution is OK, data quality is important in sample processing, data processing*
- *caBIG is leveraging simple models, this is a good thing to learn from*

3. Connecting for Health

- Q. What's needed for interaction with CFH
- Q. Between CFH and Research/Between Research and CFH
- Q. Diagnostics and development of new assays/technologies

- *In the CFH programme HL7 and Snomed are being promoted for genetic data*
- *CFH/research connection; predominantly there is a desire for clinical data from the research community. However, the format of research data is not immediately is not always appropriate in a clinical setting and appropriate transfer of knowledge is needed*
-

4. Funding?

- Q. Who should pay for infrastructure/services/data integration
- Q. Are the peer reviewed funding streams meeting the need
- Q. Who is providing current funding
- Q. What are the consequences of losing funding for research, for healthcare
- *In the US model NCI funding requires that all grantees demonstrate how they interact with the caBIG process. This allows grantees access to resources, e.g. caCORE and the GRID initiative. This model is not easily translatable to the UK which has a collection of funders (NCRI stakeholders)*
- *The funders therefore need to a common approach*

5. What should the NCRI's role be?

- *Provide fora and facilitation, help build consensus*
- *A series of workshops could develop these ideas*
- *Communicate the needs to the stakeholders*
- *Help us build international collaboration*
- *Coordinate working groups*
- *Assist in the development of a plan and interact with other communities, get involved with other initiatives - EU genetic variation, interact on level of international standards and share application development internationally*
- *Patient advocate groups are powerful and should be involved, focus on communicating with this community is important*