An Australian Government Initiative
National Collaborative Research Infrastructure Strategy

Integrated Biological Systems

Integrated Biological Systems
Phenomics Data & Informatics Workshop 2010

Gungahlin Homestead, Canberra, April 22 – 23, 2010

Presentation Abstracts and Speaker Biographies

ATLAS OF LIVING AUSTRALIA
sharing biodiversity knowledge

Australian Plant Phenomics Facility

apn australian phenomics network

Day 1 Introduction and Keynote: The Why of Informatics

Introduction: Gavin Kennedy, *IBS*

Keynote: Integrating and sharing data for large-scale international phenotyping projects; a mouse's eye view.

Professor Nadia Rosenthal, *EMBL/Australian Regenerative Medicine Institute*

Abstract: We are now in the post-genomic era of biomedical research when many of the challenges faced with generating genomic sequence are solved, and the tools are available to begin assigning normal and pathological functions to genes and their variants using phenotype analysis. In 1981, with his usual foresight, Victor McKusick, the father of modern medical genetics, anticipated the sequencing of the human genome, "perhaps by the year 2000", and noted that the determination of the sequence itself was unlikely to be the main scientific priority. "Even when the anatomy of the human genome is known down to the last nucleotide, we will not know the function of all parts of that DNA..."

The infrastructure required for the support of functional genomics in the mouse is considerable: databases, mouse repositories, informatics tools etc. , but without the acceptance of community standards for semantic coding of phenotype information, and an agreed syntax and data structure for the integration and transfer of data between databases and laboratories it is impossible to combine and analyse data from the multiplicity of projects and investigators. The amount of data already available is huge and is growing daily as the result of hypothesis-driven research. However with the initiation of major systematic mouse phenotyping projects such as EUMODIC (http://www.eumodic.eu/) the volume and complexity of the data being generated is potentially overwhelming. Following the success of the International Knockout Mouse Project (now more than 50% complete) an international project is being developed to systematically phenotype mutant mice lines for every gene in the mouse genome.

In November 2009 NIH announced funding for the KOMP2 program from the NIH Director's Common Fund (http://nihroadmap.nih.gov/KOMP2/). The International Knockout Mouse Consortium (IKMC)[1] has now created almost 14,000 mutant ES cells and 500 knockout mouse lines, and are on track to complete the knockout of every gene by the end of 2011. The new Common Fund KOMP2 program will build upon this resource by expanding the efforts to phenotype the mutant strains. Canadian and European funding agencies are currently considering making their own contributions to this huge effort, to constitute the International Mouse Phenotyping Consortium (IMPC) which will see the generation of an encyclopaedia of gene function including every coding gene in the mouse. A deluge of detailed systematic phenotypic information of this magnitude has never been seen before, and promises to generate

a huge resource for the understanding of normal and pathobiology in mouse and man. This is one of the largest and most challenging projects ever to have been undertaken in the biosciences.

The CASIMIR project (www.casimir.org.uk) was created to examine the informatics requirements, infrastructures, sociological and legal issues required for the success of international efforts in mouse phenotyping, and the exploitation of data and resources emerging from these programmes. The project included partners from across Europe, the USA, Canada, Japan and Australia, and has established consensus approaches and recommendations in all these areas.[2] The first fruits of the harmonisation of informatics approaches to database integration and phenotyping can now be seen in the International Knockout Mouse consortium (http://www.knockoutmouse.org/) and Europhenome (http://www.europhenome.org/) databases, which use a mixture of Biomarts and web services to integrate data from disseminated sources and to provide programmatic access, together with semantic standardisation and adoption of core ontologies and XML schemata. Consideration of the sustainability problems for the database and biorepository infrastructure lead to a study of financial and scientific sustainability models recently published in Database[3] , and the Rome Agenda, a consideration of the community issues for data sharing, was published last year in Nature[4]. Whilst there are specific problems concerned, for example, with the material transfer agreements for mice and the free sharing of mouse models and ES cells, most of the principles of the Rome Agenda apply across the whole of the biological sciences and provide a platform for discussions in many disciplines.

Integration of efforts driven by the community in a bottom-up rather than a top-down way has become the new success story for the biomedical sciences. International agreement on standards and the community resources necessary to realise goals which would not be achievable using only national funding has become a vital new driver pushing the biosciences forward. This represents a challenge to National funding agencies to respond to the vision and energy of the scientific community by cooperating internationally with other agencies and sharing the costs, but more importantly the fruits, of what is now a global endeavour.

1 F. S. Collins, J. Rossant, and W. Wurst, *Cell* **128** (1), 9 (2007).
2 D. Smedley, M. A. Swertz, K. Wolstencroft et al., *Brief Bioinform* **9** (6), 532 (2008).
3 Christina Chandras, Thomas Weaver, Michael Zouberakis et al., *Database* **2009** (0), bap017 (2009).
4 P. N. Schofield, T. Bubela, T. Weaver et al., *Nature* **461** (7261), 171 (2009); E. Birney, T. J. Hudson, E. D. Green et al., *Nature* **461** (7261), 168 (2009).

Bio: Professor Rosenthal is an acclaimed researcher with exceptional scientific credentials, including sixteen years working at Harvard Medical School.

Professor Nadia Rosenthal is the Director of the Australian Regenerative Medicine Institute (ARMI), the headquarters of EMBL Australia. She also directs the influential European Molecular Biology Laboratory (EMBL) Outstation in Monterotondo, Italy, one of five EMBL campuses in Europe.

Her research concentrates on embryonic heart development, ageing mechanisms and stem cell-driven regeneration of neuromuscular and cardiac tissue, using the mouse as a model for human response to disease.

She also serves as scientific director of the Heart Science Centre at Imperial College London and is currently working with leading heart transplant surgeon Professor Sir Magdi Yacoub on developing new regenerative ways to treat heart failure.

As Scientific Head of EMBL Australia, Nadia will lead the research directives of EMBL Australia.

Day 1 Session 1: The Biology of Phenomics

An introduction to The Plant Accelerator

Mark Tester, *The Australian Plant Phenomics Facility: The Plant Accelerator*

Abstract: The Plant Accelerator aims to relieve the 'phenotyping bottleneck' which has, until now, limited our ability to capitalise on substantial government and industry investments already made in plant functional genomics and modern breeding technologies. The Plant Accelerator is a national facility, available to all Australian plant scientists, offering access to infrastructure that is not available at this scale or breadth in the public sectors anywhere else in the world. The Plant Accelerator is based around automated image analysis of the phenotypic characteristics of extensive germplasm collections and large breeding, mapping and mutant populations. It exploits recent advances in robotics, imaging and computing to enable sensitive, high throughput analyses to be made of plant growth and function. The Australian Plant Phenomics Facility (APPF) has two nodes, The Plant Accelerator involving the research institutions at the Waite Campus of the University of Adelaide and The High Resolution Plant Phenomics Centre involving CSIRO Plant Industry and the Australian National University in Canberra. New technologies and approaches will be developed particularly at the High Resolution Plant Phenomics Centre to ensure that the APPF remains at the international forefront of plant science. Research networks and established pathways to market will ensure outcomes are delivered for the long-term benefit for Australian scientists and primary producers. The Plant Accelerator is expected to generate in the order of 50 TB of data annually, which needs to be managed and analyzed.

Bio: Mark Tester is Professor of Plant Physiology in the School of Agriculture, Food & Wine, University of Adelaide and a member of the Executive Management Group of the Australian Centre for Plant Functional Genomics Pty Ltd. He is Director of the Australian Plant Phenomics Facility, a $52m organisation established to develop and deliver cutting-edge phenotyping facilities for plant science, including The Plant Accelerator, an innovative 4,500 m^2 plant growth and analysis facility. He also leads a large academic research group ($7m in past 5 years) to understand salinity and drought tolerance, and how to improve this in crops such as wheat. An invited review on breeding technologies to increase crop production was published in 12 February 2010 issue of *Science*.

Phenotyping from the Growth Cabinet to the Field in the High Resolution Plant Phenomics Centre

Bob Furbank, *The Australian Plant Phenomics Facility: High Resolution Plant Phenomics Centre*

R.T. Furbank[1], X.Sirault[1] and M.R. Badger[2]
[1]HRPPC, CSIRO Plant Industry GPO Box 1600 Canberra ACT 2601 AUSTRALIA
[2]HRPPC Research School of Biology ANU Canberra ACT 2001 AUSTRALIA

Abstract: The High Resolution Plant Phenomics Centre (HRPPC) is the Canberra based node of the APPF. The HRPPC provides low to medium throughput, high resolution phenotyping of crop species, high throughput phenotyping of model plants and "reverse Phenomics" for dissection of key traits and the genes underlying these traits. The HRPPC is comprised of 4 modules: the model plant module focussing on A*rabidopsis thaliana Brachypodium distachium and Setaria viridis,* the Crop Shoot Module, for phenotyping cereal and dicot above ground phenotypes, the Crop Root Module for the below ground phenotypes of crop plants, and the field module for high throughput screening of crops in managed field sites of 10m^2 plot arrays.

The model plant module presents challenges for data storage and mining as this system runs at medium to high throughput in collaboration with the ANU. Large datasets of phenotypic characters such as growth rate, plant architecture, photosynthetic performance, yield and stress tolerance must be stored and related to genotype. A large project is currently underway to phenotype the genome wide knock out collection and ecotype collection of the model cereal *Brachypodium distachyon* and make this resource available in a phenotype / genotype database.

A major challenge for the crop and field phenotyping systems is the vast array of data types collected from a variety of sensors, both imaging and radiometric and genotypic information. Comprehensive and uniform metadata schemas will be required to achieve our goal of storing and retrieving this information for the selection of improved agricultural genotypes.

Bio: Dr Robert Furbank is the Director of the Australian Plant Phenomics Facility node: The High Resolution Plant Phenomics Centre, based in Canberra.

Dr Robert Furbank aims to improve crop yield and product quality by researching carbon partitioning and photosynthesis. Dr Furbank currently works on:

- genetic manipulation of carbon partitioning and photosynthesis in crop plants
- plant phenomics and imaging plant performance
- mechanisms of drought tolerance in cereals.

Presentation on the Australian Phenomics Network

Chris Goodnow, *Australian Phenomics Network*

The Australian Phenomics Network (APN) provides Australian and international researchers with mouse models for the study of human and animal disease.

Established in 2007 with funding from the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) and with contributions from state governments, research institutions and the National Health and Medical Research Council (NHMRC), the APN brings together mouse production, strain storage and pathology capabilities. Through this funding the APN has reduced the cost to researchers of accessing mouse models of disease, and provides equipment and expertise to undertake characterization and further research of these models.

Nine Australian facilities and institutions constitute the APN: The Australian National University, Monash University, The Walter and Eliza Hall Institute of Medical Research, The University of Melbourne, Queensland Institute of Medical Research, the Institute of Medical and Veterinary Science, the Centenary Institute, the Menzies Research Institute and the Animal Resource Centre.

The APN partners contribute their expertise and infrastructure for the production of mouse models, as well as providing cryopreservation and pathology services. In addition, the APN is working with the Atlas of Living Australia to develop a framework for Australia's e-science infrastructure to improve the capture, annotation and dissemination of research data.

The APN's core expertise and infrastructure is also extended by key national and international partnerships. These include the Garvan Institute, the Institute of Molecular Bioscience, the National Institutes of Health (United States), the Wellcome Trust (United Kingdom), and the University of Manitoba (Canada).

Bio: Chris Goodnow has illuminated the mechanism of immunological self-tolerance through innovative integration of mouse molecular genetics with cellular immunology. His discoveries have changed our concepts of how self-tolerance is acquired and autoimmune diseases are prevented, by revealing that self-reactive lymphocytes are controlled by a series of mechanisms serving as checkpoints at each step along the process of antibody formation. He has elucidated how these checkpoints achieve self-nonself discrimination, through an ability of antigen receptors to switch between signalling lymphocyte proliferation or triggering tolerance responses via qualitative changes in the intracellular second messengers elicited.

After a BSc(Vet) and Veterinary Medicine degree at the University of Sydney, Chris trained in molecular and cellular immunology at Stanford University with Mark M Davis, at the Walter and Eliza Hall Institute with Sir Gustav Nossal, and at the University of Sydney with Antony Basten. From 1990-1997, Chris headed a laboratory at Stanford University Medical School as an

Assistant Investigator of the Howard Hughes Medical Institute. Since 1997, he has been Professor of Immunology and Genetics at the John Curtin School of Medical Research at The Australian National University, and is currently Head of the Department of Immunology. Chris was the Founding Director of the Australian Phenomics Facility – a major national research facility for mouse molecular genetics. In translating his scientific expertise, Chris served on the founding scientific advisory board of Illumina Inc – now a leading genetic analysis technology company – and was founder and chief scientific officer for Phenomix Corp, a private biotechnology company with treatments for diabetes and infection in clinical development.

Chris Goodnow has authored many papers in Nature, Science and Cell, and serves on the editorial advisory boards of the Journal of Experimental Medicine, Immunity, Genome Biology, and Mammalian Genome. His honours and awards include the University Medal from Sydney University, Assistant Investigator of the Howard Hughes Medical Institute, Searle Scholar, American Association of Immunologists/Pharmingen Investigator Award, the Gottschalk Medal of the Australian Academy of Science, Commonwealth Centenary Medal, Fellow of the Australian Academy of Science, Federation Fellow of the Australian Research Council, Fellow of the Royal Society London, and the Australian Health Minister's Prize for Excellence in Medical Research.

Data production at the HRPPC

Xavier Sirault, *The Australian Plant Phenomics Facility: High Resolution Plant Phenomics Centre*

Sirault, XRR[1,2], Deery DM[1,2], Furbank RT[1,2]

1: CSIRO PI, Black Mountain, Cnr Clunies Ross St & Barry Dr, Canberra, ACT 2601
2: High Resolution Plant Phenomics Centre, Cnr Clunies Ross St & Barry Dr, Canberra, ACT 2601

The High Resolution Plant Phenomics Centre (HRPPC) is the CSIRO/ANU based node of the Australian Plant Phenomics Facility (APPF). The High Resolution Plant Phenomics Centre is a world class facility exploiting advances in robotics, imaging and computing to enable sensitive, non-destructive, high-throughput analyses to be made of plant growth and function.

One of its core business is the development of new technologies. As such, a range of imaging platforms has been designed and developed by the HRPPC for monitoring/recording the phenotype or growth characteristics of model and crop plants both in the field and in the lab. In this presentation, I will introduce three very different platforms in term of scale and through-put, which are currently being used in the centre:

1- The phenomobile: a custom-built, high through-put, phenotyping buggy designed to record information on crop grown in the field while passing over

the crop at speed of up to 10 km.h^{-1}. It integrates a range of remote sensing technologies: spectral reflectance radiometers (chemical fingerprinting of the crop), infrared radiometric sensors for monitoring canopy temperature and stereo-video rig for real-time, 3D reconstruction of canopy surfaces;

2- A medium through-put, infra-red, thermal, imaging station to screen genotypes for stomatal behaviour under varied controlled conditions; and,

3- A high through-put, digital imaging platform for growth-analysis studies: the PlantScan. The PlantScan is equipped with a conveyor belt system to increase the limited through-put of the commercial system currently used in the centre. In the future, the PlantScan will be equipped with far-infrared imaging capabilities and equipped with an imaging arch for full 3D reconstruction of plants.

I am hoping that through these three specific examples, you will have a better idea at the range and variety of sensors and data formats the HRPPC has to deal with on a daily basis.

Bio: Dr Xavier Sirault is based at the High Resolution Plant Phenomics Facility, the Canberra (Australian Capital Territory) node of the Australian Plant Phenomics Centre.

His work currently involves:

- constructing an automated, multi-sensor 3D (three-dimensional) imaging platform for crop plants,
- mapping environmental variation in breeding trials using wireless sensor technologies, and
- developing processing algorithms to automate analyses of 2D (two-dimensional) images generated by imaging sensors.

In 2007 Dr Sirault completed his doctorate on leaf rolling in wheat looking at understanding its role on:

- soil water depletion
- water relation characteristics
- light interception
- leaf-area.

Prior to completing his doctoral studies, Dr Sirault worked on identifying quantitative trait loci (QTLs) for carbon isotope discrimination in barley (INRA Montpellier) and fine QTL mapping for Type-I diabetes in mice (Institut Pasteur, Paris).

Data Production at the APN University of Melbourne Histopathology and Organ Pathology Laboratory

John Furness, *University of Melbourne, Pathology and Histopathology Service*

The Histopathology and Organ Pathology facility of APN provides a comprehensive service to researchers across Australia who require histopathological analysis of specific mouse lines. It provides comprehensive reports and histological images and data on specific mice or mice at certain development stages. Data, reports and images can be accessed and interrogated by researchers on line.

The Histopathology and Organ Pathology service also generates histological and organ pathology data on genetically modified mice available through the APN.

Staff have extensive histology, diagnostic and electronic imaging experience and the facility has engaged the service of medical and veterinary pathology consultants to aid in the interpretation of histological features.

The facility offers a Digital Slide Scanner Service capable of producing high quality digitally accessed, interrogable images of stained slides. It has access to a modern Histology Facility, confocal microscopes, a multi-colour fluorescence microscopy and a high resolution white-light microscopy all with electronic image capture.

The Australian Phenomics Network Histopathology and Organ Pathology facilities are based at the Department of Anatomy and Cell Biology at The University of Melbourne and the Institute of Medical and Veterinary Science in Adelaide.

The facility offers several levels of service:

- The First Line Phenotyping Service surveys 25 organs, using conventional staining and includes both Necropsy & Histopathology, Pathologist consultation, full pathology report and virtual digital images. Second line phenotyping involves specialist stains, including immunohistochemistry. Investigation of embryo tissues is also offered, as is examination of other species, including human. We have also expertise in investigation of tissue/ materials interfaces and have been involved with investigation of tissue interactions with prostheses.
- Large amounts of data are being produced, and there is a strong demand for access to complex data sets on-line. Moreover, it is an expectation that images and meta-data will be archived and kept for several decades, at least. The establishment of first-class data archiving, integrity, security and access regimes remains a significant challenge for APN Histopathology and Organ Pathology.

Bio: John Furness is Professor of Anatomy and Cell Biology at the University of Melbourne and Director of the Autonomic Neuroscience laboratories (http://www.anatomy.unimelb.edu.au/researchlabs/furness/index.html)

He is best known for his work in unravelling the intrinsic circuits in the digestive tract (the enteric nervous system), for the chemical coding hypothesis and for the discovery and identification of sensory neurons intrinsic to the digestive tract. The major focuses of his current work are on the cellular basis for control of enteric neurons and the identification of therapeutic targets, visceral sensory neurons (particularly those responsible for visceral pain) and the investigation of drugs that can reduce visceral pain, the control of ion channels that determine the excitabilities of neurons, and the investigation of therapeutic approaches to solving autonomic dysregulation that is consequent on spinal cord injury. He is the leader of the Histopathology and Organ Pathology section of the Australian Phenomics Network, which utilises virtual slide technology and the handling and analysis of large image files and their integration into a national phenomics database. He is the leader of the Autonomic Spinal Cord Injury Program of the Victorian Neurotrauma Initiative.

He spends his days mucking around in the lab.

Honours and Awards include:

Fellow of the Australian Academy of Science
Fellow, Academy of Science of Bologna (L'accademia delle scienze dell'istituto di Bologna), the second oldest scientific academy in existence, 2005
Centenary Medal, Govt of Australia
Honorary Life Member, Australian Neuroscience Society, for distinguished neuroscientists who have rendered notable service to the society, July 2007
Janssen International Research Award, 1993
Distinguished Research Prize, Gastroenterological Society of Australia, 1994
Australian Physiological and Pharmacological Society Lecture and Medal, 1995
Grossman Lecturer, Cambridge, U.K., 1995
Davenport Medal of the American Physiological Society, 1997
Distinguished Achievement Award, Australian Neuroscience Society, 2003
Walsh Memorial Lecture, University of California, Los Angeles, March 2004.
Honorary Vice-President, International Society for Autonomic Neuroscience, 2000.
Fellowship of the American Gastroenterological Association (AGAF) February 2008.
Gold Medal, L'accademia delle scienze dell'istituto di Bologna, for service to scientific development in Italy, June 2008

Day 1 Session 2: The Data of Phenomics

Researcher Practice in data management

Margaret Henty, *ANDS*

Abstract: Data management is a bit like motherhood: everyone is in favour of it, but actually doing it is a different matter. This talk will look at researcher attitudes to data management, researcher practice and some ideas about how to improve it.

Bio: Margaret Henty is Manager of the Capabilities Program for the Australian National Data Service (ANDS). Prior to joining ANDS, she worked with the Australian Partnership for Sustainable Repositories (APSR) where she managed outreach program and conducted small research projects related to repositories and data management.

Providing annotations for the Atlas of Living Australia and other services.

Ron Chernich, University of Queensland/*Atlas of Living Australia*

Abstract: Although Web publishing has become ubiquitous in modern research as a means of presenting and sharing information, this media has been essentially one-way. To address this limitation, numerous systems have proposed and demonstrated mechanisms by which readers can create and share mark-up of on-line documents. Although studies have demonstrated the value to collaboration of shared annotations on Web resources, there is still no wide-spread, universally-accepted, standardized approach to annotation services. Moreover there are a number of distinct, intrinsic limitations within the World Wide Web's underlying technologies that prevent the wide scale implementation of interoperable, shared annotations. This presentation will examine these limitations and show how our *Danno* research project has managed to overcome many of the inherent technological barriers and deliver a light-weight, web browser-based service for creating and displaying annotations for whole web pages, or selected regions of text, or images.

Bio: Ron is a Principal Research Fellow working in the field of metadata management and annotation services for the eResearch Lab within the School of Information Technology and Electrical Engineering (ITEE) at the University of Queensland. Ron has over 30 years of experience in IT working in Australia and overseas in commercial software development. For the past 11 years, he has specialized in the areas of metadata management and associated software tools. Currently, Ron is leading the development of a metadata annotation service for mutli-media digital collections that will become part of the Atlas of Living Australia.

The yin and yang of meta-phenomics.

Hendrik Poorter, *Forschungszentrum Jülich*

Hendrik Poorter, Frank Gilmer & Uli Schurr,
JPPC (Jülich Phenotyping Centre),
Research Center Jülich, Germany.
H.Poorter@fz-juelich.de

Abstract: In this talk we discuss two aspects related to high-throughput plant phenotyping. The first relates to the various aspects of quality control in the process

of phenotyping, with emphasis not only on the procedures to be followed, but also on the resulting population of plants one gets.

In the second part we discuss challenges at the other side of the spectrum: how do you combine information from various experiments in a meaningful way, which provides a stronger and more coherent picture than the data of the individual experiments separately? This forms part of an effort at the Research Centre Jülich to set up a database on a wide variety of plant traits, in a systematic way. Aim of this database is to unlock the information that is present in the literature, where a large amount of data for a wide number of species are presented in a non-systematic and non-unified way, but it can also be used in more dedicated experiments in a phenotyping centre. We use this database to describe the response of plants to a wide range of environmental factors by means of dose-response curves. Furthermore, our insights can be increased by analyzing the (biological) variation behind these response curves. Final goal is to achieve a quantitative picture of the phenome of the plant.

This approach - that we refer to as 'meta-phenomics' - not only will serve as a benchmark for future and comprehensive phenotyping efforts, but it will also represent a very valuable tool *per se* in understanding the integrated response of plants to their environment.

Bio: Hendrik is an ecophysiologist currently working in Jülich

---

Day 1 Session 3: The Management of Data

Managing Big Scientific Data: Capturing, Integrating and Presenting Mouse Data at MGI

Cynthia Smith, *Mouse Genome Informatics*

Cynthia L. Smith, Janan T. Eppig and the Mouse Genome Informatics Group.

In the scientific community, there has been an exponential increase in traditional publishing over the past ten years, and additional sources of new information are only being provided in digital form (i.e. mouse mutant repositories and sequencing centers). In order for humans to make sense of this data deluge, computational data warehousing and data organization is necessary. Metadata tools such as controlled vocabularies and ontologies are required by curators to organize and integrate information, and for accurate data retrieval by data miners. Using these tools ultimately allows for computational methodologies to fully exploit and realize the potential of these data, and to allow new trends and insights to emerge that would previously have remained hidden behind the obvious.

The Mouse Genome Informatics resource (MGI, www.informatics.jax.org) provides free access to integrated data on the genetics, genomics and biology of the laboratory mouse, facilitating navigation through sequence, polymorphism, spatiotemporal expression, mapping, biochemical function and process, sub-cellular topology, mammalian homology, phenotype and disease model data. MGI users can

explore these data using a suite of navigation tools, including an enhanced Quick Search Tool, Mouse GBrowse, and web-based vocabulary browsers and data type-specific query forms. Robust querying parameters include standardized terms from the Mammalian Phenotype Ontology, a dynamic DAG-structured vocabulary that supports phenotype annotations to background-specified allelic genotypes at varying degrees of granularity. Parallel use of OMIM and other MGI bio-ontologies, including the Mouse Anatomical Dictionary and Gene Ontology (GO), fosters complementary routes to examine anatomy-based gene expression profiles, map functional features of gene products to specific disease states, and establish associations between observed mouse phenotypes and orthologous human gene mutations for which defined mouse genotypes model the human pathological state.

We will review in detail the data sourcing and integration processes at MGI; the problems of capturing, managing, analyzing, integrating and distributing heterogeneous data sets and provide examples of enhanced data retrieval using controlled vocabularies.

Supported by NIH grant HG000330.

Bio: Dr. Smith received her PhD in Cell Biology and Anatomy from Weill Cornell Medical College in 1995 and did postgraduate work in developmental neurobiology at Harvard Medical School and at The Skirball Institute at the New York University School of Medicine before joining the Mouse Genome Informatics project in the fall of 2000. Her initial responsibilities at MGI were to begin to catalog and describe mouse mutations from the literature and later projects include the construction and maintenance of the Mammalian Phenotype Ontology (MP), as well as use of this ontology in classifying mouse phenotypes at MGI.

PODD: An Ontology-driven Data Repository for Collaborative Phenomics Research

Yuan-Fang Li, *University of Queensland*

Abstract: Phenomics is becoming an increasingly critical tool to understand phenomena such as plant morphology and human diseases. Phenomics studies make use of both high- and low-throughput devices for data capture and measurement. As a result, high volumes of data are generated on a regular basis, making storage, management, annotation and distribution a challenging task. Sufficient contextual information, the metadata, must also be maintained to facilitate the dissemination of these data. The challenge is further complicated by the need to support emerging technologies and processes in phenomics research. This talk presents our on-going effort in designing and developing an ontology-driven, open, extensible data repository to support collaborative phenomics research in Australia.

Bio: Yuan-Fang Li is a research fellow at the School of ITEE, the University of Queensland, Australia. He received both his Bachelor of Computing and PhD's degrees from National University of Singapore. His main research interests include

the Semantic Web, ontology languages, semantic querying and inference, large-scale information processing, information visualization and formal methods.

Genomics Data Management - a case study

Joe Thurbon, *Intersect*

Abstract: Next-generation sequencing generates data that presents two data-management challenges. The volumes of data generated are very large, and the distinct scientific uses of the data are manifold. This talk presents a system, developed by Intersect and about to go live, which directly addresses both problems. The goals of the system are to provide collaborative access to the metadata and data associated with next-generation sequencing, and to provide support for researchers to keep track of their analyses of that data.

Bio: Joe Thurbon is an eResearch analyst at Intersect. In his role, he helps academics without IT backgrounds make best use of IT in their research. His own research background is in mathematical logic, diagrammatic reasoning, and machine learning. He has ten years R&D experience in academic and industry settings, and fifteen as a software engineer.

The iPlant Collaborative: A Cyberinfrastructure-Centered Community of Plant and Computing Scientists

Matthew Vaughn, *iPlant*

A dominant theme in 21st century science is the emergence of biology as a data-driven science. Advances in laboratory automation, microfluidics, and imaging, in combination with accessibility of high performance computing, have converged to allow rapid production of large, sophisticated biological data sets. Most familiar is ultra-high throughput sequencing (UHTS), which generates extremely sizable, but relatively simple data. New phenotyping technologies produce data of similar size, but of much higher complexity. Consequently, massive collections of images, morphological measurements, or biochemical/metabolomic profiles are becoming commonplace and require sophisticated storage, archiving, integration, and analysis capabilities. To date, solutions to large and/or complex data problems have been developed somewhat *ad hoc* in the form of (often high quality) standalone databases, services, and tools that are usually not mutually communicative.
The overarching mandate of the iPlant Collaborative (iPlant) is to build a robust, accessible, extensible, and sustainable cyberinfrastructure that ties together powerful, user-centered analytical tools capable of meeting the analytical challenges of modern biology. We solicited proposals from the plant science community to identify grand challenges that could be addressed given a comprehensive cyberinfrastructure. Gathering an unprecedented number of biologists and computer

scientists in six community workshops, we further crystallized these proposals into two specific grand challenges to focus our initial round of development: 1) building a tree of life to incorporate 500,000 plant species (*iPTOL* project), and 2) elucidating the genotype to phenotype continuum in plants (*iPG2P* project).

Major initiatives underway for the iPG2P project include: 1) pipelines to provide storage, cycles, best-of-breed analysis routines, and interpretive tools for UHTS applications; 2) infrastructure-grade general linear model-based association mapping capable of supporting analysis of data tables in the 1E+12 size range; 3) a component-based interactive visual programming environment for scientific inquiry, with development focused initially on supporting analyses common to G2P analyses, including cross-species inference, prediction and interpretation of networks, and integration of metabolite and expression data; 4) database and analysis systems capable of acquiring, processing, and storing complex phenotype data.
In designing these applications, we have identified a number of issues, including paucity of common formats and access APIs (even for genomic data), lack of support for storage and retrieval of large, complex measurement data, a requirement for robust metadata and provenance management, conflicting models for computational services, and an abundance of storage formats that hamper data integration. Our Common Semantic Model for data integration, the development of semantic web technologies based on SSWAP, the iPlant metadata model, and efforts to foster innovation in phenomic database technologies will be discussed, as well as the concept of the iPlant CI as a platform for new analytical services, collaborative analysis, and data sharing.

Bio: Ph.D. University of Illinois under Dan Bush 2001; Physiological and Molecular Plant Biology; Worked on regulation of photoassimilate partitioning and sucrose transporter proteins in Arabidopsis.

Postdoc Cold Spring Harbor Lab with Rob Martienssen; Worked on functional redundancy in plant genetic architectures, transitioned into bioinformatics and worked on miRNA target prediction, antisense RNAs, and epigenetic regulation of transposable elements.

Research Assistant Professor, Cold Spring Harbor Laboratory (2007): Devoting half time to serving as lead for Genotype to Phenotype in iPlant and the rest to research on computational epigenomics, sequence assembly in polyploid genomes, and cloud computing.

Day 2 Session 1: The Language of Phenomics

Using semantic web ontologies in molecular genetics

Melissa Davis, *Institute of Molecular Biology/Queensland Facility for Advanced Bioinformatics*

Abstract:
Semantic web technologies promise improvements in the way we capture data, data provenance, and deal with complex issues of knowledge representation and inference. The use of ontologies has already provided demonstrable benefits in molecular biology and genetics, and use of Semantic Web ontologies is expected to enhance these benefits. Here I will describe our research in the development of Semantic Web ontologies in the application domains of biomolecular interaction, small molecule research, as well as our technical research in ontology integration and ontology engineering.

Bio: Melissa is currently a Data Modeller with the Queensland Facility for Advanced Bioinformatics, where she is developing systems biology and bioinformatics research projects in pathway and network analysis of molecular interaction networks and signal transduction pathways. This work is particularly interesting due to the challenges of integrating experimental results with biomolecular networks to add value to both bioinformatic and experimental analyses. Previously, Melissa held a post-doctoral research appointment in ARC Center of Excellence in Bioinformatics at the Institute for Molecular Bioscience, where she developed a number of projects both in ontological engineering and network analysis. Prior to her research appointments, Melissa worked as a knowledge management consultant for a Brisbane-based management company and as a consultant for a Canberra based systems engineering company where she provided specialist consulting services in the areas of knowledge management, ontological engineering, systems biology and biotechnology. Her research interests include, but are not limited to: Systems biology, specifically of the nuclear receptor super-family of transcription factors, and the application of bioinformatics to experimental and translational research. Melissa holds a PhD in Computational and Molecular Biology, and a BSc majoring in genetics, both from the University of Queensland.

Linked Open Data: a New Resource for eResearch

Anne Cregan, *Intersect/ANDS*

Abstract: The Open Data Movement aims at making data freely available to everyone. A Data Commons is rapidly emerging, and the World Wide Web is fast becoming a space not only for linking documents and web pages, but for interlinking data sets. Since inception in 2007, the W3C's Linking Open Data Project, based on the Resource Description Framework (RDF) standard, has grown into a data cloud now containing billions of items. This data cloud provides a useful repository of data for use in research and is also an important place to publish open research data sets to be shared with other researchers and the community. A large component of research data is suitable for linking into the open data cloud, and international researchers have commenced the process of publishing their data sets online as a collaborative research initiative, as it is an excellent way to expose, share, and connect pieces of research data. Greater visibility and ability to process data with a common theme generated by different research groups enables new research studies and insights to emerge.

Biography: Dr Anne Cregan is an eResearch Analyst at Intersect Australia and ANDS. She has a diverse background in a range of academic disciplines and her commercial IT background encompasses programming, business analysis, data mining and senior level management, as well as project management and co-ordination. Anne has a doctorate in Computer Science from UNSW, sponsored by NICTA and specialising in semantic web technologies. She is on the Program Committee for several semantic Web-related conferences and has a BSc (Hons) in Psychology from the University of Sydney.

Biodiversity data federation - The Atlas of Living Australia

Donald Hobern, *Atlas of Living Australia*

Abstract: The Atlas of Living Australia (ALA) has been funded under the National Collaborative Research Infrastructure Strategy (NCRIS) to integrate data on Australia's biodiversity and to make the data available in forms which can support research, policy and education, in particular in the areas of conservation, biosecurity and taxonomic research. Relevant data sets exist in many forms and include structured records of the occurrence of species at different locations, databases of names and classifications, sequences, characteristics and traits, and text information on each species. The ALA aims to catalogue all of these resources using linked open data technologies and to offer a range of interfaces to discover and explore the resulting map of connected information.

Bio: Donald Hobern spent 16 years working in software development for IBM before joining the staff of the Global Biodiversity Information Facility (GBIF) to develop its technical infrastructure for integrating biodiversity information at the global level. Since 2002 he has been very active in standards development for biodiversity data and in promoting the use of shared open architectures to maximise the usefulness of such data. He is currently Director of the Atlas of Living Australia project and Chair of the Taxonomic Databases Working Group (TDWG).

Day 2 Session 2: The Informatics of Phenomics

Accurate estimation of plant biomass from two-dimensional images

Mahmood Golzarian, *PBRC – University of South Australia*

Mahmood R. Golzarian[1], Karthika Rajendran[2], Stuart Roy[2], Mark Tester[2], Desmond S. Lun[1]

1- Phenomics and Bioinformatics Research Centre, School of Mathematics and Statistics, and Australian Centre for Plant Functional Genomics, University of South Australia

2- Australian Centre for Plant Functional Genomics, University of Adelaide

Abstract: Accurate estimation of plant biomass from two-dimensional images is a core problem in high throughput plant phenotyping. In this project, we found that the current linear regression modelling of biomass solely as a function of plant area achieves a large bias for salt-stressed and non-salt-stressed plants. If not addressed, this problem prohibits accurate biomass estimation of plants under stress from two dimensional plant images. We showed that this bias is correlated with plant developmental stage. Based on this observation, we developed alternative methods to estimate plant biomass that significantly reduce the bias. The predictive estimators are based on the colour information obtained from the images of the wheat plants grown in a greenhouse environment. Modelling plant growth as a function of plant area and plant age, we demonstrated that most of the observed variance can be explained and, moreover, a small bias for salt-stressed and non-salt-stressed plants is obtained.

Bio: "I did my Bachelor and Master's degrees in Agricultural engineering in Iran. I did my PhD in the school of mechanical engineering at the University of South Australia. The title of my PhD project was 'computer vision system for monitoring crop establishment'. I graduated in July 2009 and joined Phenomics and Bioinformatics Research Centre at the University of South Australia in August 2009".

Phenomics and Biomedical Imaging

Jurgen Fripp, CSIRO/Australian e-Health Research Centre

Abstract: Biomedical imaging involves the development of novel imaging and image post-processing techniques to detect, diagnose, and monitor the progression of diseases and to evaluate different treatment therapies on disease. These techniques are most commonly utilized in clinical practice and large scale studies, for instance our lab is involved in large scale (3D Magnetic Resonance Imaging) studies into Osteoarthritis, Alzheimer's disease, Stroke, (Prostate and Brain) Cancer. From these studies, we'll discuss imaging and image analysis relevent to plant and mouse phenomics as well as our labs data management and analysis practices. Several examples of these techniques applied to plant data may also be illustrated.

Genetically modified animal (like mice), has become commonplace in biomedical research. The study of these animals has opened new opportunities for understanding the biology of development and disease processes, however, it also introduces a need for the development of efficient and reliable (objective) characterization. Our lab has been involved in the development of techniques to allow the quantitative 3D anatomical phenotyping (using 3D micro-CT images) and quantitative functional phenotyping (using PET) of mice and rats. In this talk we'll discuss both the imaging and image analysis used in this research.

Bio: Jurgen Fripp obtained his PhD in the field of Medical Imaging at the University of Queensland. He is currently a research scientist at the Australian e-Health Research Centre at the CSIRO ICT Centre. His main research interests are in

registration and segmentation particularly using techniques from shape and appearance analysis.

Day 2 Session 3: Phenomics Information Infrastructure

The Australian National Data Service: making it easier to manage and reuse data

Andrew Treloar, *Australian National Data Service(ANDS)*

Abstract: The Australian National Data Service has been established with a vision of "more researchers reusing data more often". This presentation will discuss how ANDS is working towards this vision, and the services that ANDS is putting in place to make this possible.

Bio: Dr Andrew Treloar ndrew Treloaris the Director of Technology for the Australian National Data Service (ANDS) (http://ands.org.au/), and in 2008 led the project to establish ANDS. Prior to that he was associated with a number of e-research projects as Director or Technical Architect: ARCHER (http://archer.edu.au/) – an e-Research support environment, DART (http://dart.edu.au) – data acquisition and analysis, and ARROW (http://arrow.edu.au/) – institutional repositories, as well as the development of an Information Management Strategy for Monash University (http://www.monash.edu.au/staff/information-management/). He has a B.A. hons. (first class), majoring in Germanic Languages and Linguistics, a Grad. Dip. in computer science, and a research Masters in English Literature, all from Melbourne University. In 1999 he received his Ph. D. from Monash University with the topic Hypermedia Online Publishing - Transformation of the Scholarly Journal. His research interests include data management, institutional repositories and scholarly communication. He never seems to be able to make enough time for practising his cello, reading, talking to his “chooks” (chicken whispering), or working in his vegetable garden and orchard. Further details at http://andrew.treloar.net/.

The Australian Research Collaboration Service (ARCS)

Paul Coddington, *Australian Research Collaboration Service (ARCS)*

Abstract: Founded in 2007, the Australian Research Collaboration Service (ARCS) provides technology services that enable Australia’s researchers to operate at the forefront of their fields. Offering interoperable and collaborative solutions, ARCS is an integral component of the national research infrastructure. This talk will provide an overview of ARCS and the services that it offers to the research community.

Bio: Dr Paul Coddington is ARCS Projects Manager, working on the initiation and management of projects to develop eResearch services for a wide range of research communities. He is also Deputy Director of eResearch SA, the South Australian

member of ARCS. He has a PhD in computational physics and over 20 years experience in eResearch, having worked on many research and development projects concerning the application of high-performance and distributed computing to scientific problems and the development of online research data repositories.

The Australian Access Federation: An Overview

Bradley Beddoes, *Australian Access Federation (AAF)*

Abstract: The Australian Access Federation (AAF) provides a production framework and support infrastructure to facilitate trusted electronic communications and collaboration within and between universities and research institutions in Australia and overseas. This talk will give an introductory overview of the AAF and benefits it brings to activities in the higher education and research sectors.

---

eResearch in CSIRO

Dr Darrell Williamson, *CSIRO*

Abstract: The research process is rapidly moving to one that is computing intensive and data-intensive based on access to shared data sources, web services and scientific workflow. CSIRO is now in the first year of a 4 year funded program to develop eResearch across the enterprise. An overview will be provided of the progress to date, and the strategy going forward which builds upon and relates to the NCRIS and Super Science initiatives.

Bio: In February this year, Dr. Darrell Williamson was appointed to his current position as CSIRO eResearch Director with responsibility for the strategic development of eResearch services across CSIRO having joined the CSIRO in July 2007 as Deputy Director of the ICT Centre. Darrell served for two years as Chief Executive Officer for the Cooperative Research Centre in Advanced Computational Systems, and six years as Chief Executive Officer and Research Director for the Smart Internet Technology Cooperative Research Centre. Previously, Darrell had a number of senior academic appointments including Senior Research Fellow in the Research School of Physical Sciences & Engineering and Professor and Dean of the Faculty of Engineering & Information Technology at the Australian National University. He has undergraduate degrees in Science and Electrical Engineering, a Masters degree in Electrical Engineering from the University of Newcastle and a Doctoral degree in Applied Mathematics from Harvard University.

EMBL Australia - an opportunity for all Australian Science

Silvio Tiziani, *EMBL Australia*

Abstract: Australia is the first Associate Member of the European Molecular Biology Laboratory (EMBL), Europe's preeminent molecular biology research organisation funded by 20 member states. EMBL Australia has been formed to capitalise on the potential of this membership with plans to develop an Australian Partner Laboratory to the EMBL to include a locally based mirror of EMBL's EBI and a national network of Bioinformaticians to support australian scientist to access the EBI resources.

Bio: Silvio is Executive Director of EMBL Australia, also Chief Operating Officer of the Australian Regenerative Medicine Institute at Monash University. He has worked in management positions at both public and private sector organisations and has extensive experience in business development, strategic planning, financial analysis and budget management and corporate governance.