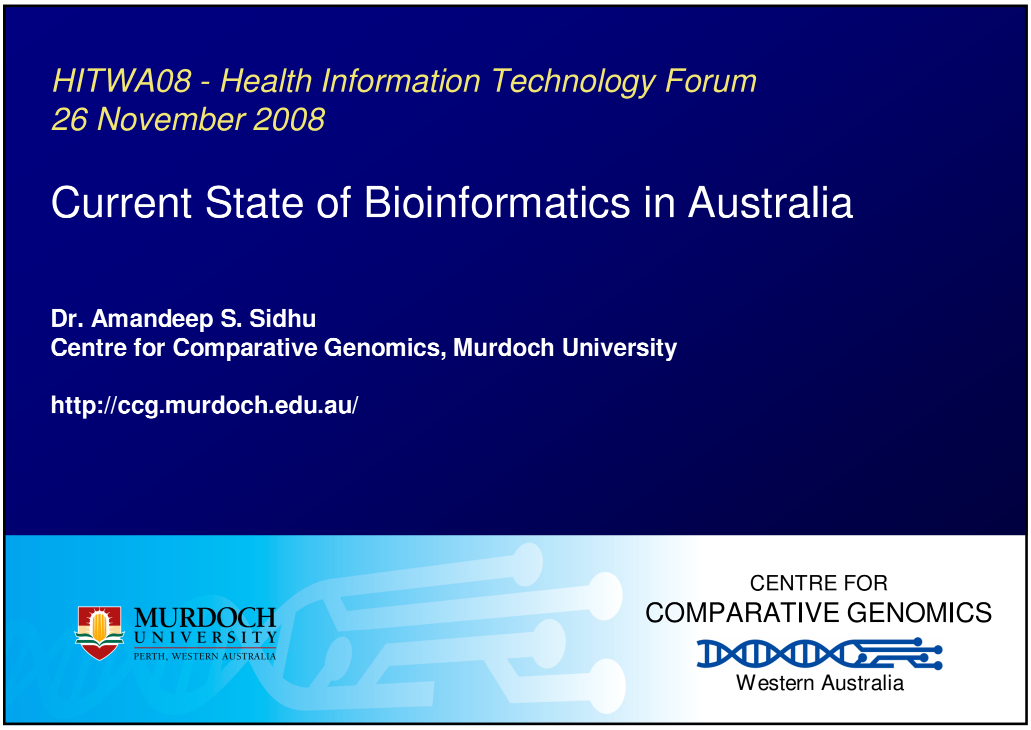
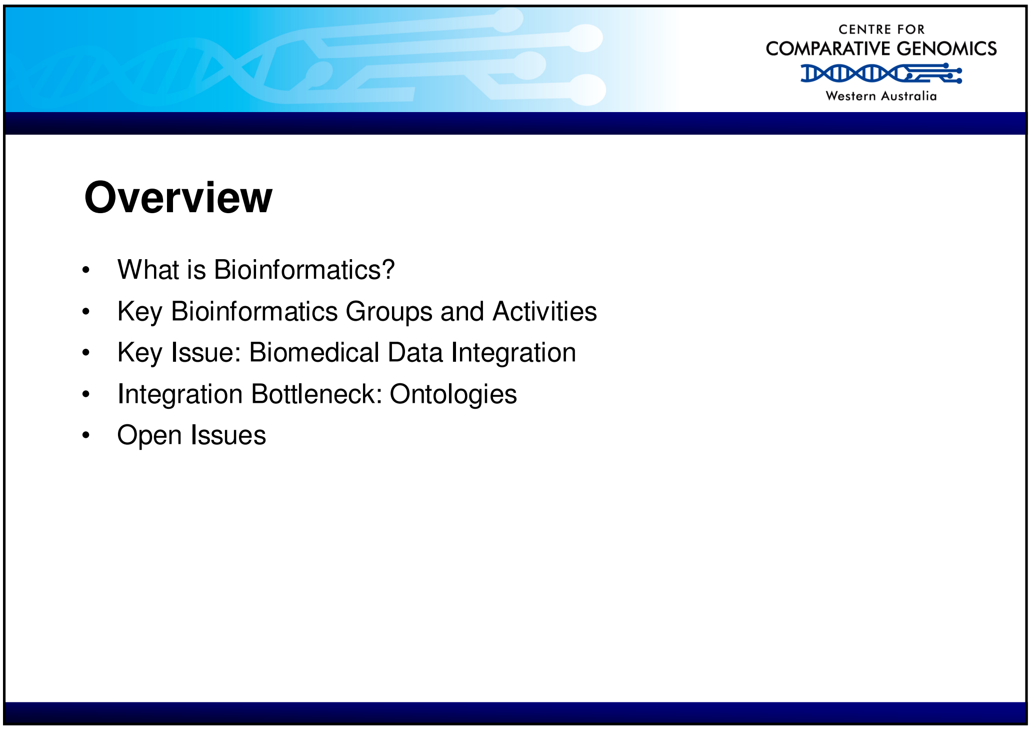


HITWA08 - Health Information Technology Forum
26 November 2008

Current State of Bioinformatics in Australia

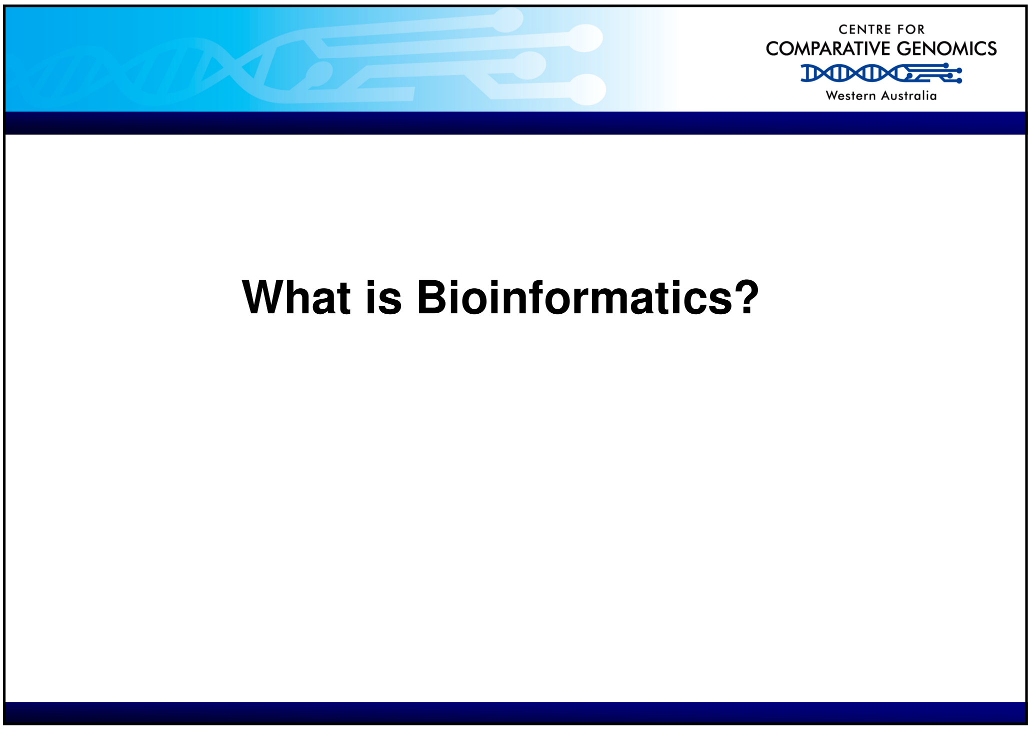
Dr. Amandeep S. Sidhu
Centre for Comparative Genomics, Murdoch University

http://ccg.murdoch.edu.au/



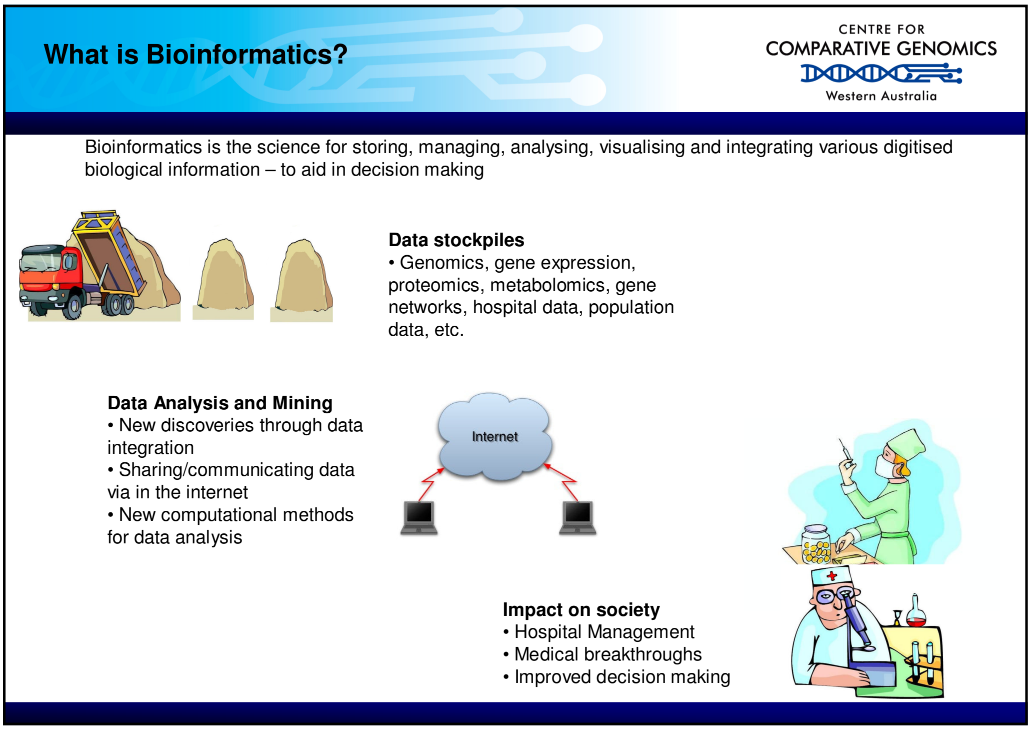
Overview

- What is Bioinformatics?
- Key Bioinformatics Groups and Activities
- Key Issue: Biomedical Data Integration
- Integration Bottleneck: Ontologies
- Open Issues



CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

What is Bioinformatics?



What is Bioinformatics?

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Bioinformatics is the science for storing, managing, analysing, visualising and integrating various digitised biological information – to aid in decision making

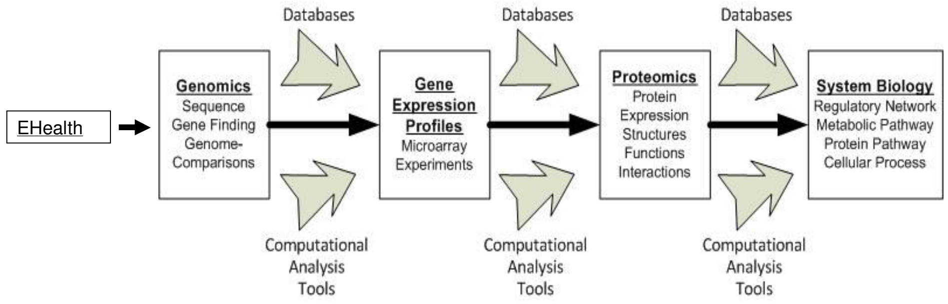
Data stockpiles
- Genomics, gene expression, proteomics, metabolomics, gene networks, hospital data, population data, etc.

Data Analysis and Mining
- New discoveries through data integration
- Sharing/communicating data via in the internet
- New computational methods for data analysis

Impact on society
- Hospital Management
- Medical breakthroughs
- Improved decision making

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

A Biomedical Informatics Perspective of Life Science Processes



EHealth

Databases

Computational
Analysis
Tools

Genomics
Sequence
Gene Finding
Genome-
Comparisons

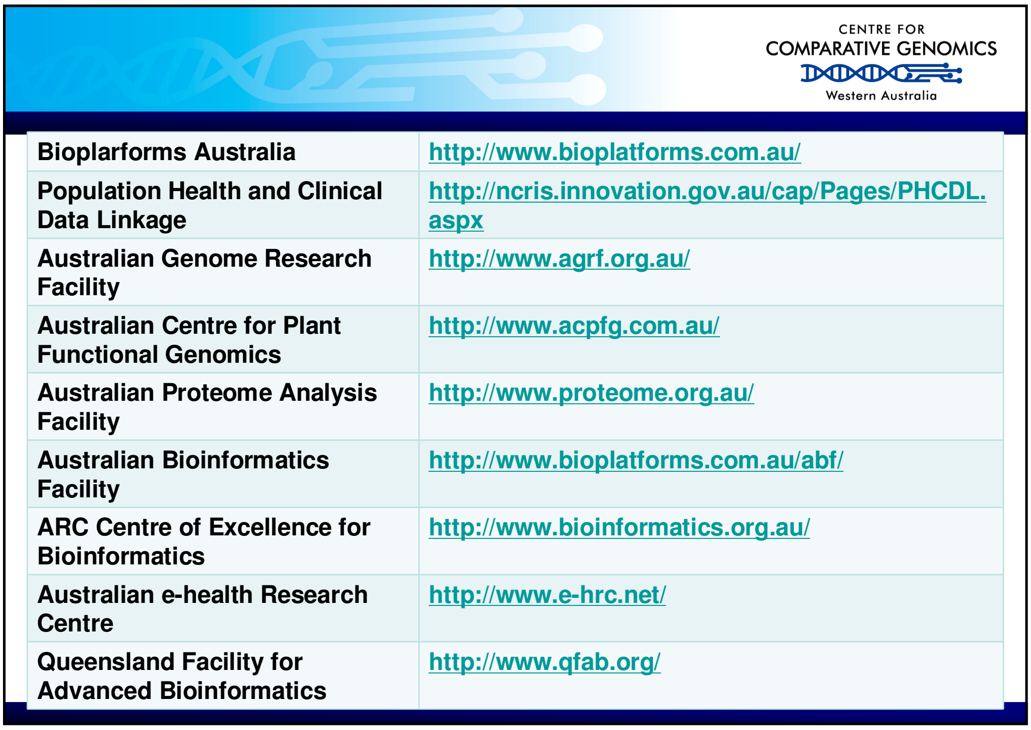
Gene Expression Profiles
Microarray
Experiments

Proteomics
Protein
Expression
Structures
Functions
Interactions

System Biology
Regulatory Network
Metabolic Pathway
Protein Pathway
Cellular Process

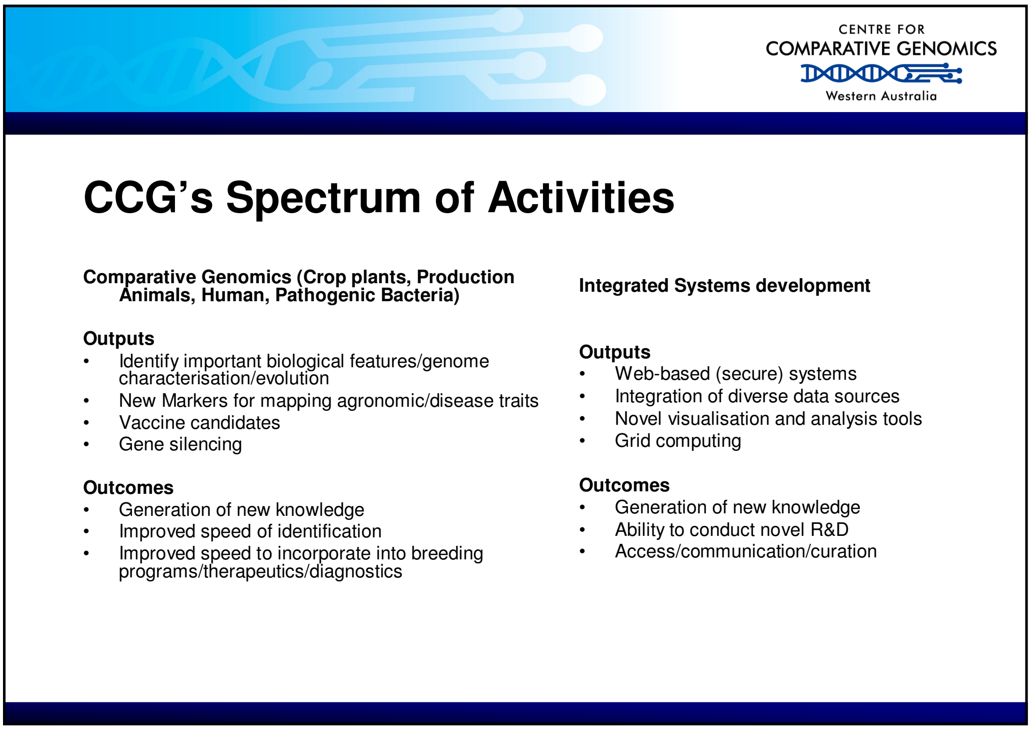
CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Key Groups and Activities



CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Bioplarforms Australia	http://www.bioplatforms.com.au/
Population Health and Clinical Data Linkage	http://ncris.innovation.gov.au/cap/Pages/PHCDL.aspx
Australian Genome Research Facility	http://www.agrf.org.au/
Australian Centre for Plant Functional Genomics	http://www.acpfg.com.au/
Australian Proteome Analysis Facility	http://www.proteome.org.au/
Australian Bioinformatics Facility	http://www.bioplatforms.com.au/abf/
ARC Centre of Excellence for Bioinformatics	http://www.bioinformatics.org.au/
Australian e-health Research Centre	http://www.e-hrc.net/
Queensland Facility for Advanced Bioinformatics	http://www.qfab.org/



CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

CCG's Spectrum of Activities

Comparative Genomics (Crop plants, Production Animals, Human, Pathogenic Bacteria)

Outputs

- Identify important biological features/genome characterisation/evolution
- New Markers for mapping agronomic/disease traits
- Vaccine candidates
- Gene silencing

Outcomes

- Generation of new knowledge
- Improved speed of identification
- Improved speed to incorporate into breeding programs/therapeutics/diagnostics

Integrated Systems development

Outputs

- Web-based (secure) systems
- Integration of diverse data sources
- Novel visualisation and analysis tools
- Grid computing

Outcomes

- Generation of new knowledge
- Ability to conduct novel R&D
- Access/communication/curation

CCG Team consists of

- Research Scientists
- Bioinformaticians
- Bioinformatics Officers
- Software Developers
- System Administrators

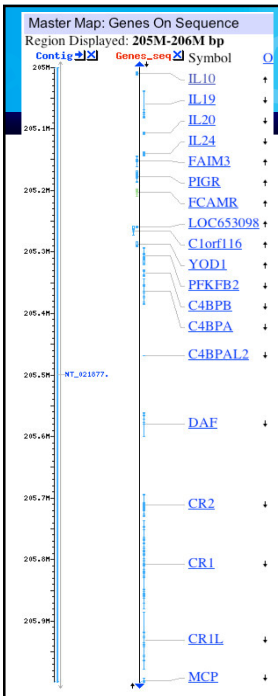
Key Issue: Biomedical Data Integration

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Example: Integrating publicly available Asthma genomic data

- Internet resource 1:
  - NCBI 58 asthma associated (genome regions) QTLs
- Internet resource 2:
  - NCBI 144 asthma annotated genes
- Internet resource 3:
  - NCBI many GEO datasets asthma gene expression studies
- Internet resource 4:
  - NIH Genetic Association DB 46 positively associated asthma genes
- Internet resource 5:
  - PubMed 50 microarray and asthma related papers

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia



H. sapiens Genome (Build 36.1) #Chromosome: 1 Region: 205000000.00..206000000.00

Symbol	Description
IL10	interleukin 10
IL19	interleukin 19
IL20	interleukin 20
IL24	interleukin 24
FAIM3	Fas apoptotic inhibitory molecule 3
PIGR	polymeric immunoglobulin receptor
FCAMR	Fc receptor, IgA, IgM, high affinity
LOC653098	similar to chromosome 1 open reading frame 116
C1orf116	chromosome 1 open reading frame 116
YOD1	YOD1 OTU deubiquinating enzyme 1 homolog (yeast)
PFKFB2	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2
C4BPB	complement component 4 binding protein, beta
C4BPA	complement component 4 binding protein, alpha
C4BPAL2	complement component 4 binding protein, alpha-like 2
DAF	decay accelerating factor for complement (CD55, Cromer blood group)
CR2	complement component (3d/Epstein Barr virus) receptor 2
CR1	complement component (3b/4b) receptor 1 (Knops blood group)
CR1L	complement component (3b/4b) receptor 1-like
MCP	membrane cofactor protein (CD46, trophoblast-lymphocyte cross-reactive antigen

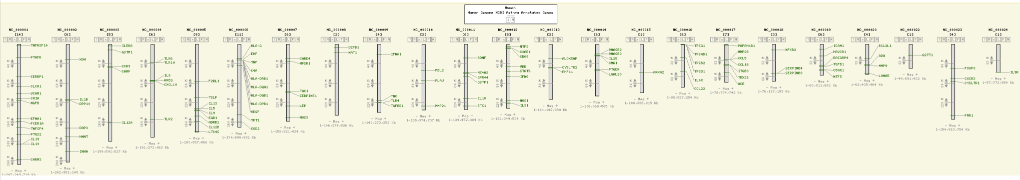
CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Initial questions

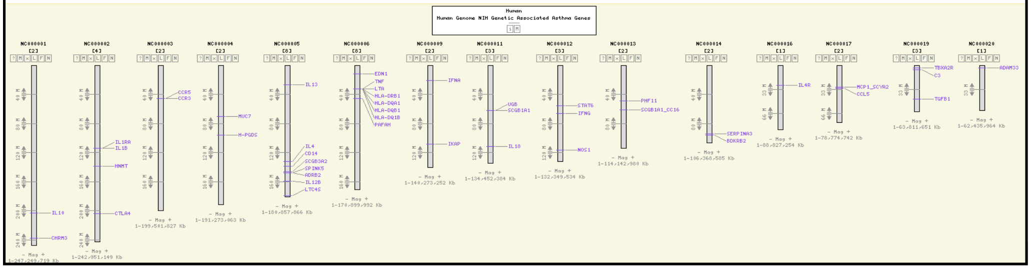
- Do the ‘NCBI’ annotated genes co-locate with the Asthma QTLs?
  - What about the NIH Genetic Association DB 46 positively associated asthma genes?
- Do any of the 58 microarray gene expression studies identify common genes that are upregulated/downregulated?
  - Are the non-randomly clustered?
- Are there any available bioinformatics resources that integrate all this information or allow a researcher to do so? How to go about this?

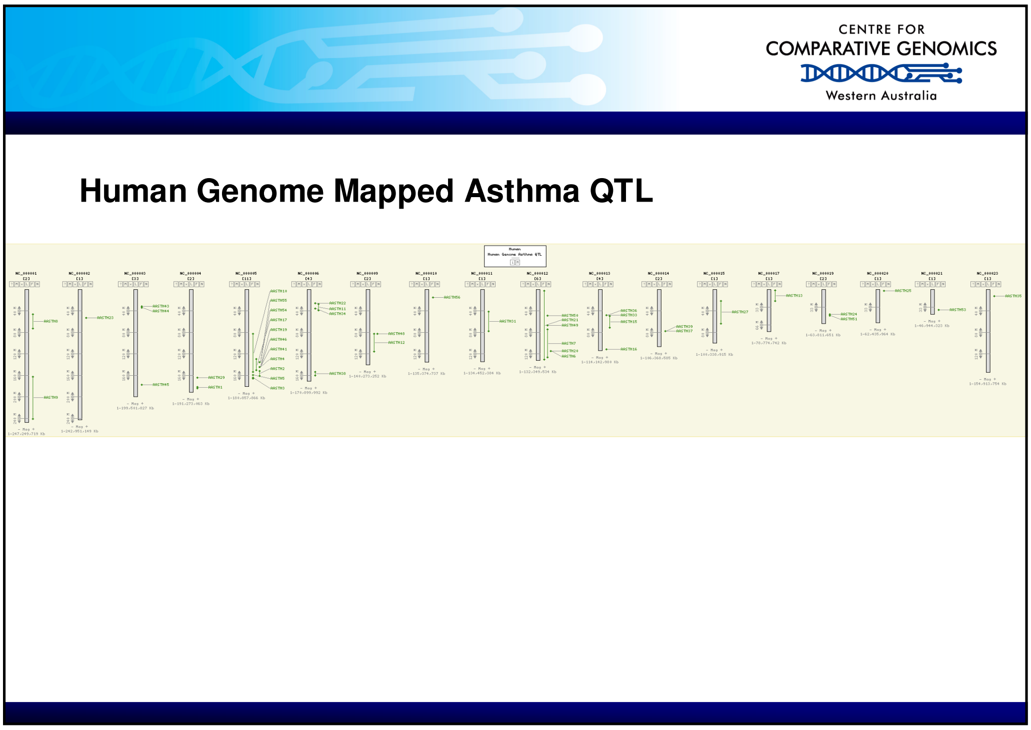
Human Genome Mapped NCBI Asthma Annotated Genes

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia



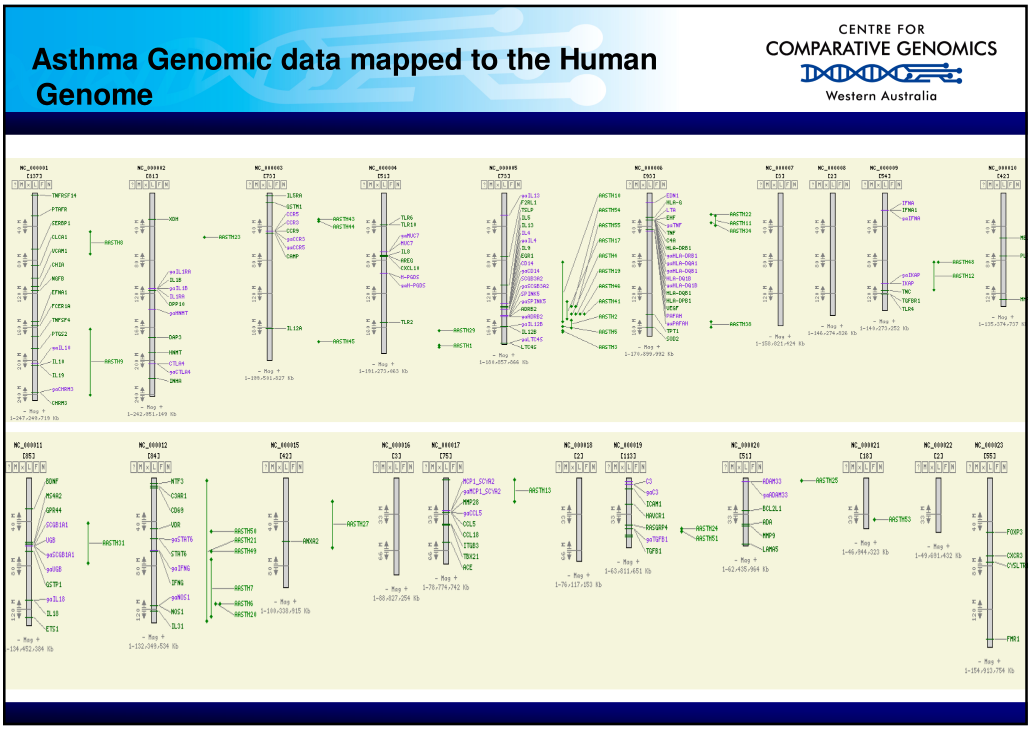
Human Genome Mapped NIH Genetic Associated Asthma Genes





CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

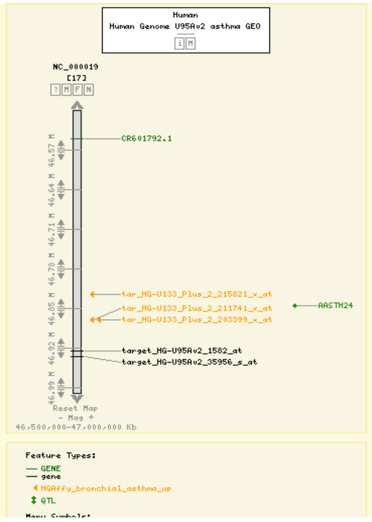
Human Genome Mapped Asthma QTL

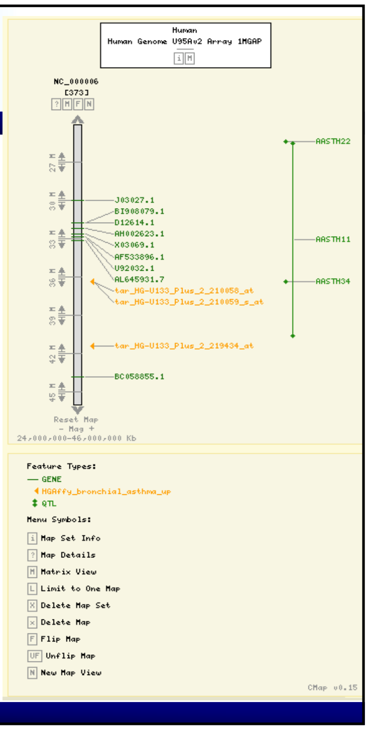


Asthma Genomic data mapped to the Human Genome

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Mapped Asthma annotated QTL's to the genome





CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Nature of Biomedical Data

- The biological data sets are organized in loose hierarchies that reflect our understanding of complex living systems.
- They range from plain text of laboratory records, nucleic acids and proteins, three-dimensional structures of molecules, biomedical images, to various experimental outputs from technologies in laboratories and hospitals.
- Data from these systems is incomplete and often inconsistent which presents challenges in challenges in informatics, modelling, and simulation.
- Furthermore, biomedical research is an expanding phase, and many fields of biology are still in the developing stages.

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Inconsistency in Biomedical Data

- Creators of biological data sources use data descriptors with specific intended meanings in their own contexts.
- Heterogeneity arises because different biological data sources have different vocabularies and semantics.
- The creator of a data source structures the information in the source such that the interests of the creator and the immediate community are best satisfied.
- The creator is frequently less concerned about the use of the data source by a secondary system or community.

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Problem of Biomedical Data Integration

- Typically, researchers require integrated access to data from multiple domains, which requires resolving terms that have slightly different meanings across the communities.
- This is further complicated by the observations that terminology is being used either is not explicitly identified or the terminology evolves over time.
- Data access methods change regularly. These changes are an attempt to keep up with the scientific evolution occurring in the community at large.

Data Integration Solutions

Integrated Data Retrieval

- Data Integration Solutions can be divided into three major categories:
  - Link-driven approaches using hyperlinks
  - View integration approaches using standardized schema
  - Semantic integration using common vocabulary

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Need for Common Languages

- Mechanisms of intercommunication between databases of different structure and format is the need for common semantic standards.
- As noted by Hafner (Hafner and Fridman, 1996), general biological data resources are databases rather than knowledge bases: they describe miscellaneous objects according to the database schema, but no representation of general concepts and their relationships is given.
- (Schulze-Kremer, 1998) addressed this problem by developing ontologies for knowledge sharing in molecular biology.

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Ontologies in Biomedicine & E-health

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Ontologies

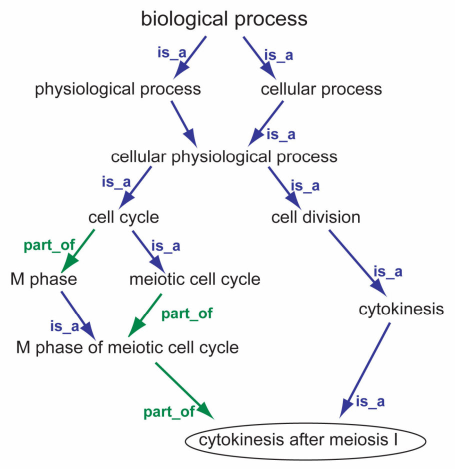
- The term ontology is originally a philosophical term referred to as “the object of existence”. The computer science community borrowed the term ontology to refer to a “specification of conceptualisation” for knowledge sharing in artificial intelligence (Gruber, 1993).
- Ontologies provide a conceptual framework for a structured representation of the meaning, through a common vocabulary.
- Vocabulary usually includes concepts, relationships between concepts, and possibility of defining ontology rules and axioms.
- Ontologies in biomedicine have emerged because of the need for a common language for effective communication across diverse sources of biological data and knowledge.

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

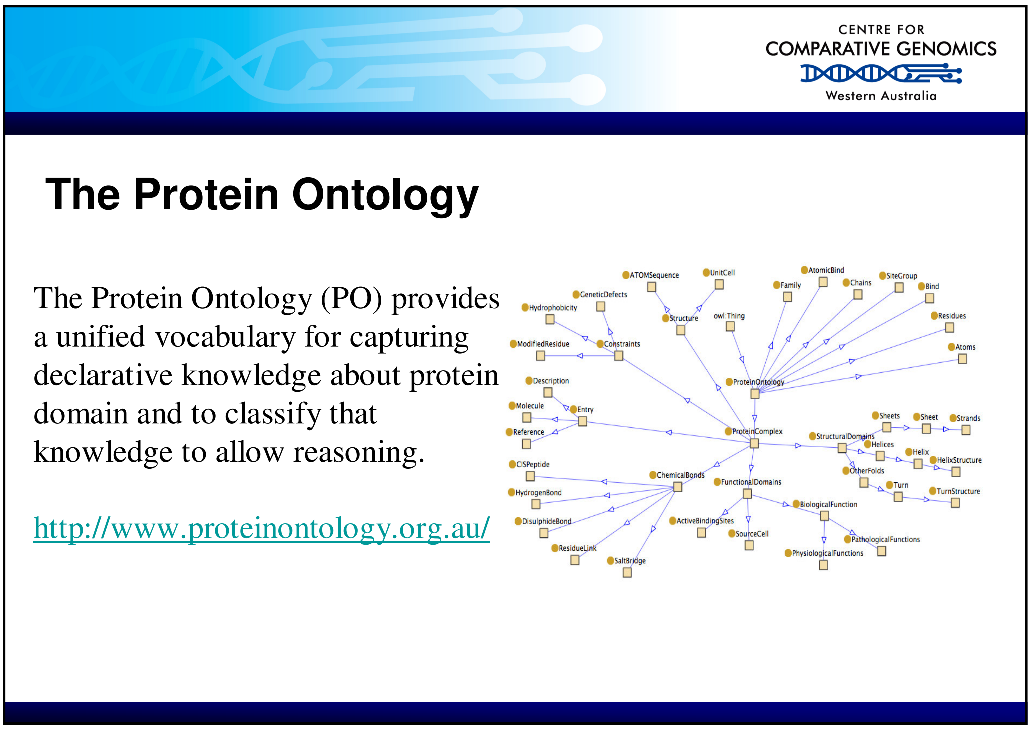
The Gene Ontology

The Gene Ontology (GO) is a collaborative effort to create a controlled vocabulary of gene and protein roles in cells, addressing the need for consistent descriptions of gene products in different databases.

http://www.geneontology.org/



biological process
is_a
is_a
physiological process
cellular process
is_a
cellular physiological process
is_a
is_a
cell cycle
cell division
part_of
is_a
M phase
meiotic cell cycle
is_a
part_of
cytokinesis
is_a
M phase of meiotic cell cycle
part_of
is_a
cytokinesis after meiosis I

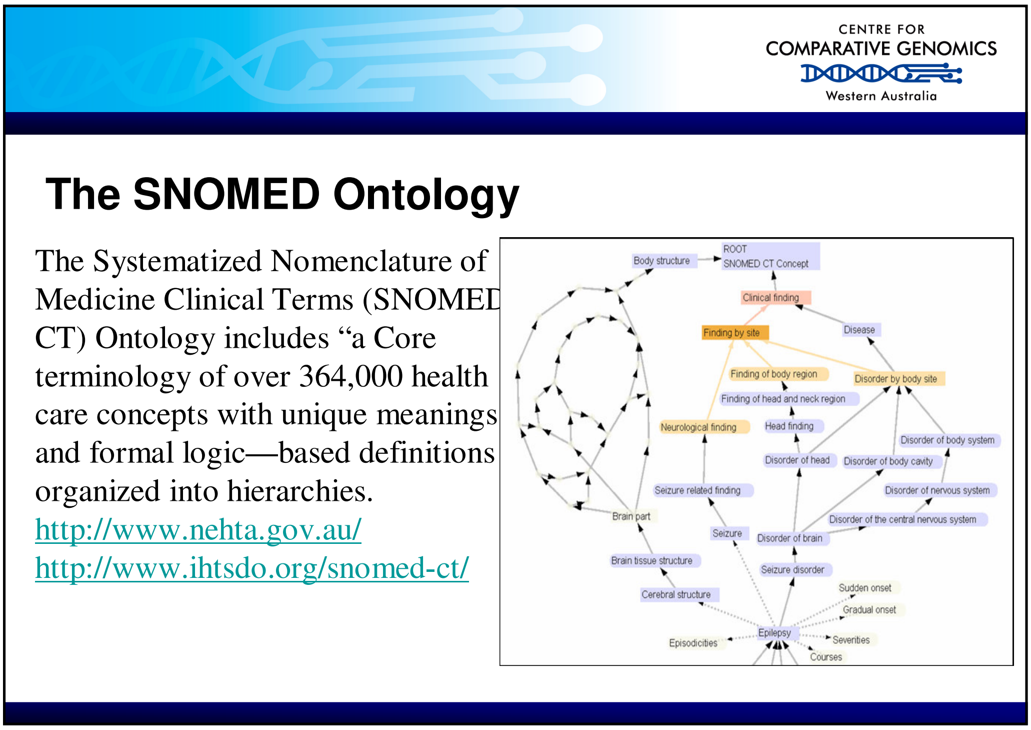


CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

The Protein Ontology

The Protein Ontology (PO) provides a unified vocabulary for capturing declarative knowledge about protein domain and to classify that knowledge to allow reasoning.

http://www.proteinontology.org.au/



CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

The SNOMED Ontology

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) Ontology includes “a Core terminology of over 364,000 health care concepts with unique meanings and formal logic—based definitions organized into hierarchies.

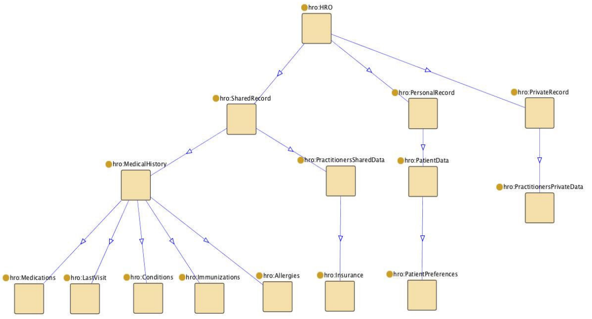
http://www.nehta.gov.au/

http://www.ihtsdo.org/snomed-ct/

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

The Health Records Ontology

Health Records Ontology is a effort to utilize SNOMED terminology as a basis & extend it to define a minimal set of concepts to integrate diversity in various formats to achieve a standardization in communication for healthcare delivery.



CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Open Issues

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

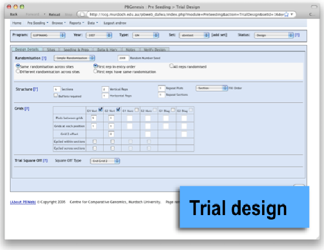
CCG Projects

- Plant Breeding Web (PB Web)
- Wheat Genome Sequencing and Annotation
- Protein Ontology (PO) Project
- Truth in Labelling of Fish
- Beef CRC Project
- Health Records Ontology (HRO) Project
- Yet Another Bioinformatics Interface (YABI)
- .. and more

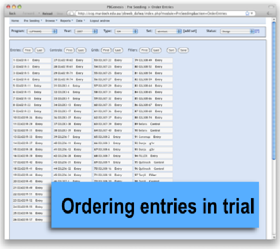
Plant Breeding Web (PB Web)

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

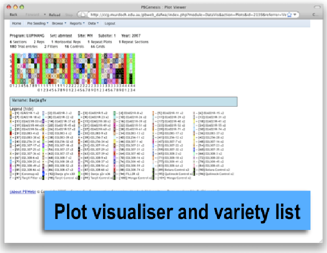
PBWeb splashscreen



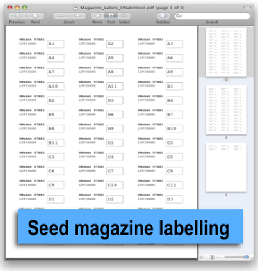
Trial design



Ordering entries in trial

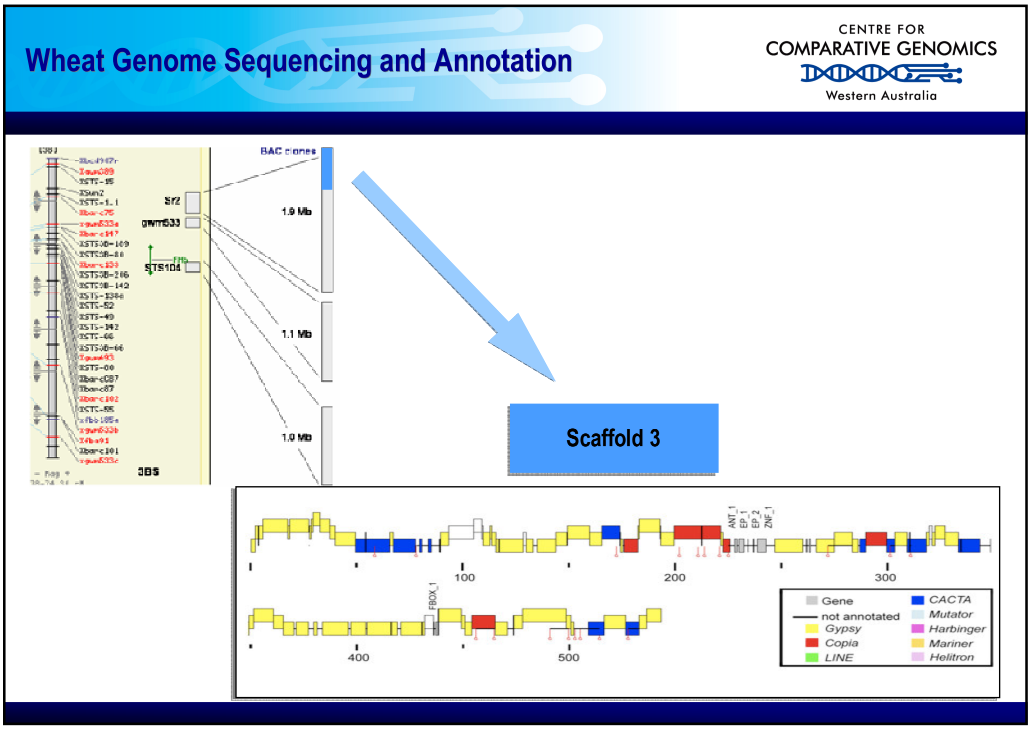


Plot visualiser and variety list



Seed magazine labelling

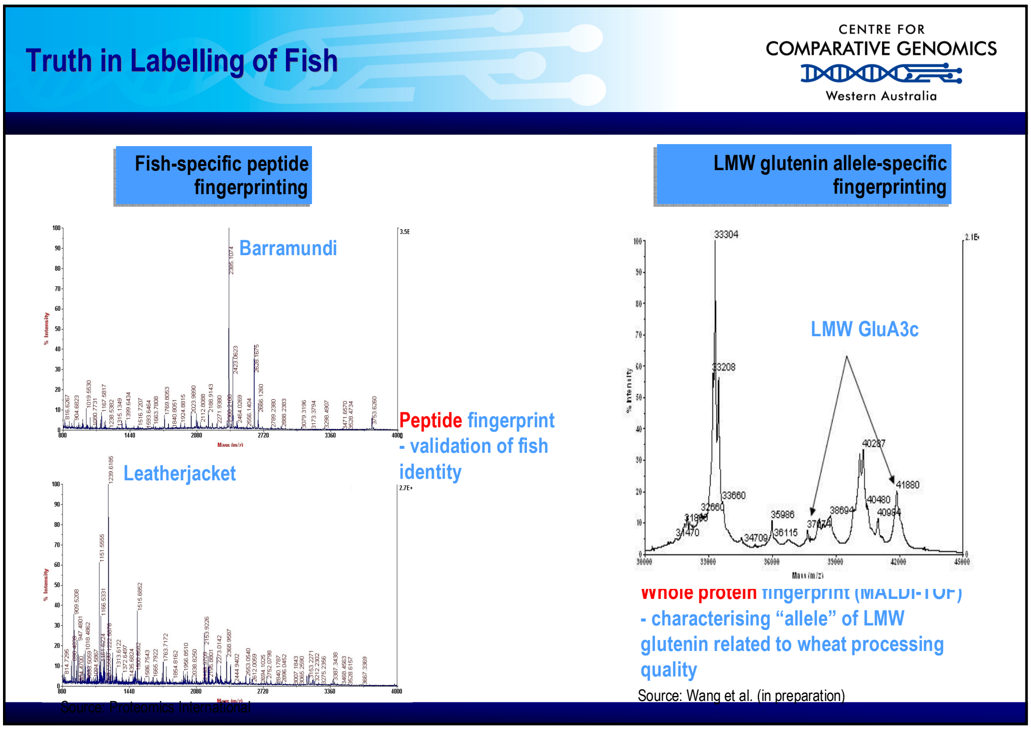
Harvest bag labelling



Wheat Genome Sequencing and Annotation

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Scaffold 3



Truth in Labelling of Fish

CENTRE FOR
COMPARATIVE GENOMICS
Western Australia

Fish-specific peptide fingerprinting

LMW glutenin allele-specific fingerprinting

Peptide fingerprint - validation of fish identity

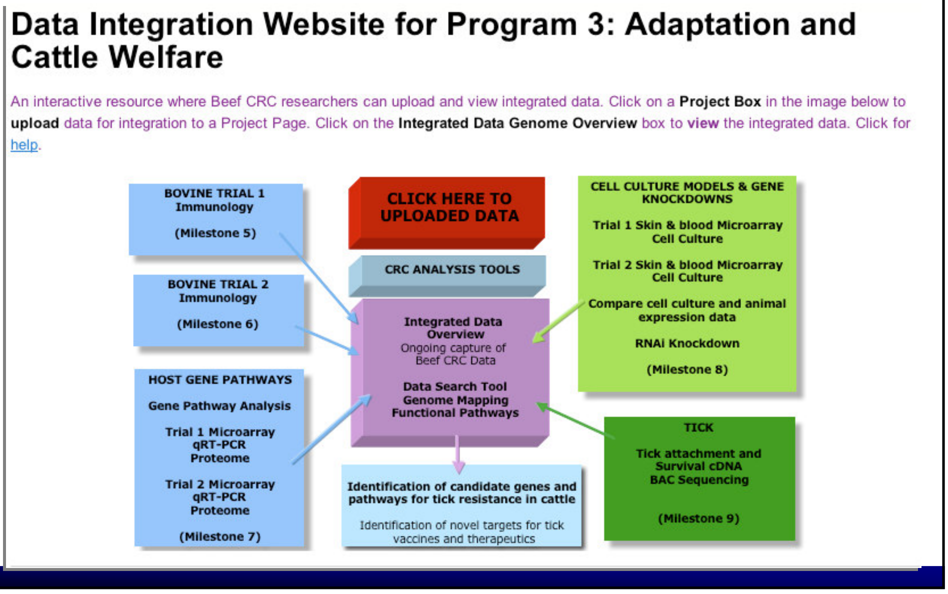
Whole protein fingerprint (MALDI-TOF) - characterising “allele” of LMW glutenin related to wheat processing quality

Source: Wang et al. (in preparation)

Source: Proteomics International

Beef CRC Project

CENTRE FOR COMPARATIVE GENOMICS
Western Australia



Data Integration Website for Program 3: Adaptation and Cattle Welfare

An interactive resource where Beef CRC researchers can upload and view integrated data. Click on a **Project Box** in the image below to **upload** data for integration to a Project Page. Click on the **Integrated Data Genome Overview** box to **view** the integrated data. Click for help.

BOVINE TRIAL 1
Immunology
(Milestone 5)

BOVINE TRIAL 2
Immunology
(Milestone 6)

HOST GENE PATHWAYS
Gene Pathway Analysis
Trial 1 Microarray
qRT-PCR
Proteome
Trial 2 Microarray
qRT-PCR
Proteome
(Milestone 7)

CLICK HERE TO UPLOADED DATA

CRC ANALYSIS TOOLS

Integrated Data Overview
Ongoing capture of Beef CRC Data
Data Search Tool
Genome Mapping
Functional Pathways

Identification of candidate genes and pathways for tick resistance in cattle
Identification of novel targets for tick vaccines and therapeutics

CELL CULTURE MODELS & GENE KNOCKDOWNS
Trial 1 Skin & blood Microarray
Cell Culture
Trial 2 Skin & blood Microarray
Cell Culture
Compare cell culture and animal expression data
RNAi Knockdown
(Milestone 8)

TICK
Tick attachment and
Survival cDNA
BAC Sequencing
(Milestone 9)

YABI

CENTRE FOR COMPARATIVE GENOMICS
Western Australia

Yet **A**nother **B**ioinformatics **I**nterface
- A Web 2.0 Internet application
- Simple design: drag & drop concept
- Simplifying
  - *-omics* workflow creation
  - Customisation
  - Data analysis
- Design your own workflow