Linking the EMBL Australia Bioinformatics Resource with the Australian National Data Service

- **JEFF CHRISTIANSEN** – ANDS
- **PIERRE CHAUMEIL** - QFAB
- **DOMINIQUE GORSE** – QFAB
- **MARK RAGAN** – IMB/UQ
EMBL Australia

- Australia has become the first Associate Member of the European Molecular Biology Laboratory EMBL (2009)

- EMBL Australia, when fully established, will consist of:
  - two Groups at EMBL-Heidelberg and/or outstations including the European Bioinformatics Institute (EBI)
  - a Partner Laboratory Network in Australia
  - full participation in EMBL postdoctoral, student & training programs
  - Secretariat (at ARMI/Monash)
  - Additionally a mirror of the EBI database will be established to facilitate biology data access and computation access in Australia

Foundation partners
CSIRO, Monash, UQ, U Sydney, UWA

Also represented on Council
DIISRTE, ARC, NHMRC
EBI Database

- Created and managed by EMBL
- 30+ databases (sequence, expression, pathways, structure, taxonomy, molecular interaction, ontologies)
- 1000+ computational tools
- Based in Cambridge (Hinxton) UK
- One portal @ www.ebi.ac.uk
EMBL Australia Bioinformatics Resource

- A selection of EBI databases and tools mirrored to www.ebi.edu.au

- Increases ease of access to data for Australian-based users

- Will allow access to HPC-based computational analyses of large datasets
Project: Linking EMBL Australia Bioinformatics mirror and RDA
INSTRUCTIONS TO AUTHORS:

GENE SEQUENCES
Authors must deposit all relevant information and data into the appropriate public databases, including but not limited to one of the International Nucleotide Sequence Database Collaboration members. Note that submission into one of the following sequence database ensures that the others will also receive the sequences:

- **GenBank**, the NIH's genetics sequence database
- **EMBL-Bank**, The EMBL Nucleotide Sequence Database
- **DDBJ**, DNA Data Bank of Japan

To prevent an article being delayed in publication provide all accession numbers at final submission.
LOCUS X70800 1821 bp mRNA linear ROD 20-MAR-1996
DEFINITION M.musculus wnt-11 mRNA.
ACCESSION X70800
VERSION X70800.1 GI:312829
KEYWORDS embryogenesis; glycoprotein; wnt gene.
SOURCE Mus musculus (house mouse) Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.

REFERENCE 1 (bases 1 to 1821)
TITLE Murine Wnt-11 and Wnt-12 have temporally and spatially restricted expression patterns during embryonic development
PUBMED 7547479

REFERENCE 2 (bases 1 to 1821)
AUTHORS Wainwright, B.J.
TITLE Direct Submission
JOURNAL Submitted (27-JAN-1993) B.J. Wainwright, Centre for Molecular Biology & Biotechn., University of Queensland, St. Lucia, Brisbane, 4072, AUSTRALIA

ORIGIN
001 gaattcgggc ctaatccgag cctgacgccg gcgggttctcg ggcggttcgg ggagagagcg
061 gactccttcc tcgctcagcc tccccggccc gacccctcct ttgtaatttg aataaaacgc
121 ctcccagccc gcgcgcgccc ttaacccggc gcctggttct cgctgagttg aggcggcgtg
181 cgccgcagaa cacgcagccg gggcctgcagt cggcggagtt cggtgcggct cctgcagggt
cggaccccccc ggacgccggg cccgccgcagc atggagggcgc ggcggcaggt ctgcgcaggt ...
EBI Collections of Australian Interest (1)

- DNA/protein sequences from Australian dwelling species
EBI Collections of Australian Interest (1)

- Species as the lowest collection entry level
  - Fully sequenced genomes
  - Nucleotide sequences
  - Protein sequences

- DNA/protein data for 364,164 species has been deposited in the EBI

- Species taxonomy identifies higher level groupings (EBI utilises NCBI taxonomy)
  - e.g. *Macropus rufus*; Macropodia; Diprodontia; Marsupialia; Mammalia; Vertebrata; Animalia.
EBI Collections of Australian Interest (1)

- Identification of Australian dwelling species (native, feral, agriculturally important)
  - Australian Plant Census (info from State Herbaria via ALA)
  - Australian Faunal Directory (DSEWPC - info via ALA)

- 70,703 plant species listed (APC)
- 101,048 animal species listed (AFD)
EBI Collections of Australian Interest (1)

- DNA/protein sequences from Australian dwelling species

DNA/protein seq for 364,164 worldwide dwelling species in EBI
EBI Collections of Australian Interest (1)

- DNA/protein sequences from Australian dwelling species

DNA/protein seq for 364,164 worldwide dwelling species in EBI

171,751 Australian dwelling plants and animals
EBI Collections of Australian Interest (1)

- DNA/protein sequences from Australian dwelling species

- DNA/protein seq for 364,164 worldwide dwelling species in EBI

- 171,751 Australian dwelling plants and animals

DNA/protein seq collections for 12,825 Australian dwelling plants and animals
EBI Collections of Australian Interest (1)

- DNA/protein sequences from Australian dwelling species

  - 12,825 species and 697 assoc higher-level (e.g. genus, family, phyla) groupings.
  - Scientific and common names included to increase discoverability
  - 13,522 collection descriptions published to RDA

DNA/protein seq collections for 12,825 Australian dwelling plants and animals
### Australian Organism Collection

<table>
<thead>
<tr>
<th>Name</th>
<th>Phascolarctos cinereus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonyms</td>
<td>Koala</td>
</tr>
<tr>
<td>Rank</td>
<td>species</td>
</tr>
</tbody>
</table>

This data collection contains all currently published nucleotide (DNA/RNA) and protein sequences from the Australian dwelling organism Phascolarctos cinereus, commonly known as Koala. The nucleotide (DNA/RNA) and protein sequences have been sourced through the European Nucleotide Archive (ENA) and Universal Protein Resource (UniProt), databases that contain comprehensive sets of nucleotide (DNA/RNA) and protein sequences from all organisms that have been approved by the scientific and research community. The identification of the Koala as an Australian dwelling organism has been published in the Australian Plant Census (APC) or Australian Faunal Assessment (AFA) of Living Australia.

### EBI Data for Species

- **DNA Records**: The first 100 DNA records for this species are displayed. Other records (total: 148) can be viewed at EBI.
- **Protein Records**: All available protein records are displayed.

172 items found, displaying 1 to 20.

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Sequence Description</th>
<th>EBI Database</th>
<th>View in EBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB011223</td>
<td>Phascolarctos cinereus mitochondrial ND1 gene for NADH dehydrogenase subunit 1, complete cds.</td>
<td>ENA</td>
<td>Go to EBI</td>
</tr>
<tr>
<td>AB241053</td>
<td>Phascolarctos cinereus mitochondrial DNA, nearly complete genome.</td>
<td>ENA</td>
<td>Go to EBI</td>
</tr>
<tr>
<td>AF252263</td>
<td>Phascolarctos cinereus cytochrome P450-4A15 (CYP4A15) mRNA, partial cds.</td>
<td>ENA</td>
<td>Go to EBI</td>
</tr>
<tr>
<td>AF252696</td>
<td>Phascolarctos cinereus olfactory receptor koa2 gene, partial cds.</td>
<td>ENA</td>
<td>Go to EBI</td>
</tr>
<tr>
<td>AF252697</td>
<td>Phascolarctos cinereus olfactory receptor koa3 gene, partial cds.</td>
<td>ENA</td>
<td>Go to EBI</td>
</tr>
</tbody>
</table>

[First/Prev] 1, 2, 3, 4, 5, 6, 7, 8 [Next/Last]

The search term: **Phascolarctos cinereus**
EBI Collections of Australian Interest (2)

- DNA/protein sequences submitted by Australian Research Organisations
EBI Collections of Australian Interest (2)

- Identification of Australian Research-Associated Organisations
  - ARC (incl. linkage) grants (2002 - 2012)
  - NHMRC grants (1990 – 2011)

- 201 Australian Biological Research-Associated Organisations identified
DNA/protein sequences submitted by Australian Research Organisations

DNA/protein seq submitted from 1000s of Organisations world-wide in EBI
EBI Collections of Australian Interest (2)

DNA/protein sequences submitted by Australian Research Organisations

- DNA/protein seq submitted from 1000s of Organisations world-wide in EBI
- 201 Australian Research Organisations
DNA/protein sequences submitted by Australian Research Organisations

DNA/protein seq submitted from 1000s of Organisations world-wide in EBI

201 Australian Research Organisations

DNA/protein seq collections from 101 Australian research Organisations
EBI Collections of Australian Interest (2)

- DNA/protein sequences submitted by Australian Research Organisations

- 101 institutions and 8 assoc higher-level (i.e. State/Federal) groupings.

- 109 collection descriptions fed to RDA

DNA/protein seq collections from 101 Australian research Organisations
Submitting to RDA descriptions of workflows performed using the NCI-SF & Bioinformatics Resource (EABR)
Submitting to RDA descriptions of workflows performed using the NCI-SF & Bioinformatics Resource (EABR)
Submitting to RDA descriptions of workflows performed using the NCI-SF & Bioinformatics Resource (EABR)

- Yabi: NCRIS 5.16 NeAT Project
- Grid enabled Internet-based environment for transparent access to HPC
- Lead by Matthew Bellgard (CCG Director and ABF Convenor)
- Project Committee: CCG, ABF, ARCS, ANDS, CSIRO, VLSCI, QFAB
Yabi Screenshot
## Yabi Workflow Manager

Omics sciences can be applied across many industries and disciplines, where scientists can search for genes that make some varieties of plants more robust. The research possible scientists must rely heavily on analysis, computation, and development of workflows. By login into Yabi, a user can access a range of tools that facilitate these processes. 

<table>
<thead>
<tr>
<th>Number</th>
<th>Name</th>
<th>Description</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Select File</td>
<td>Select a file from your workspace directory</td>
<td>Name: res Value: s1.fastq;</td>
</tr>
<tr>
<td>2</td>
<td>FastQC</td>
<td>FastQC generates a set of reports to look at the quality of a next generation dataset</td>
<td>Name: Input Value: Accepts input from tool 1 Name: --format Value: fastq Name: --threads Value: 1</td>
</tr>
<tr>
<td>3</td>
<td>fastqc_dipper</td>
<td>Removing sequencing adapters/adaptors</td>
<td>Name: -i Value: Accepts input from tool 1 Name: -o Value: dipper.fq Name: -a Value: TCGTATGCGCTTCTGCTT</td>
</tr>
<tr>
<td>4</td>
<td>FastQC</td>
<td>FastQC generates a set of reports to look at the quality of a next generation dataset</td>
<td>Name: Input Value: Accepts input from tool 3 Name: --format Value: fastq</td>
</tr>
<tr>
<td>5</td>
<td>bowtie_simple</td>
<td>Runs the bowtie aligner, with a cut-down set of configuration options, (e.g. no colour space support)</td>
<td>Name: ebiwt index Value: TAIR10 Name: output Value: aligned.sam</td>
</tr>
<tr>
<td>6</td>
<td>convert_sam_to_bam</td>
<td>Convert a sam file to the equivalent binary bam file using samtools (<a href="http://samtools.sourceforge.net">http://samtools.sourceforge.net</a>)</td>
<td>Name: Input Value: Accepts input from tool 5 Name: -o Value: aligned.bam</td>
</tr>
<tr>
<td>7</td>
<td>samtools_sort</td>
<td>Sort a bam file using samtools</td>
<td>Name: Input Value: Accepts input from tool 6 Name: Output Value: aligned.bam</td>
</tr>
<tr>
<td>8</td>
<td>Select File</td>
<td>Select a file from your workspace</td>
<td>Name: res Value: miRNA_PredFor.dff</td>
</tr>
</tbody>
</table>
Acknowledgements

- **QFAB Team**
  - Pierre Chaumeil
  - Felicity Newell
  - Troy Sadkowsky
  - Dominique Gorse
  - Jeremy Barker
  - Cas Simons
  - Nick Rhodes
  - Sarah Williams

- **CCG Group**
  - Andrew Macgregor
  - Adam Hunter
  - Matthew Bellgard

- **EMBL Australia**
  - Gavin Graham
  - Gerald Hartig
  - David Green

- **ANDS Group**
  - Jeff Christiansen
This project is supported by the Australian National Data Service (ANDS)

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative

Australian Government
Department of Innovation
Industry, Science and Research