



*Data Capture and Information  
management platform at BeCA  
&  
CIMMYT*

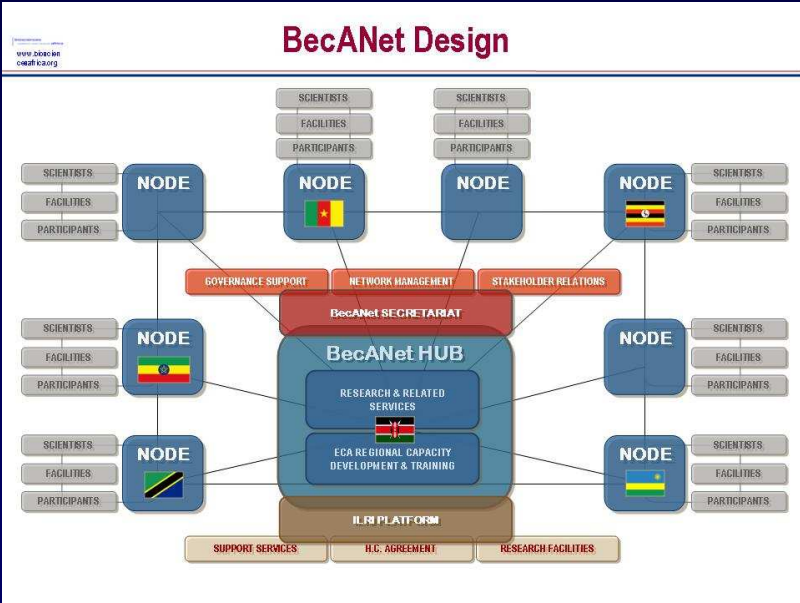
ICRISAT, Aug 2007

# Beca & Objectives

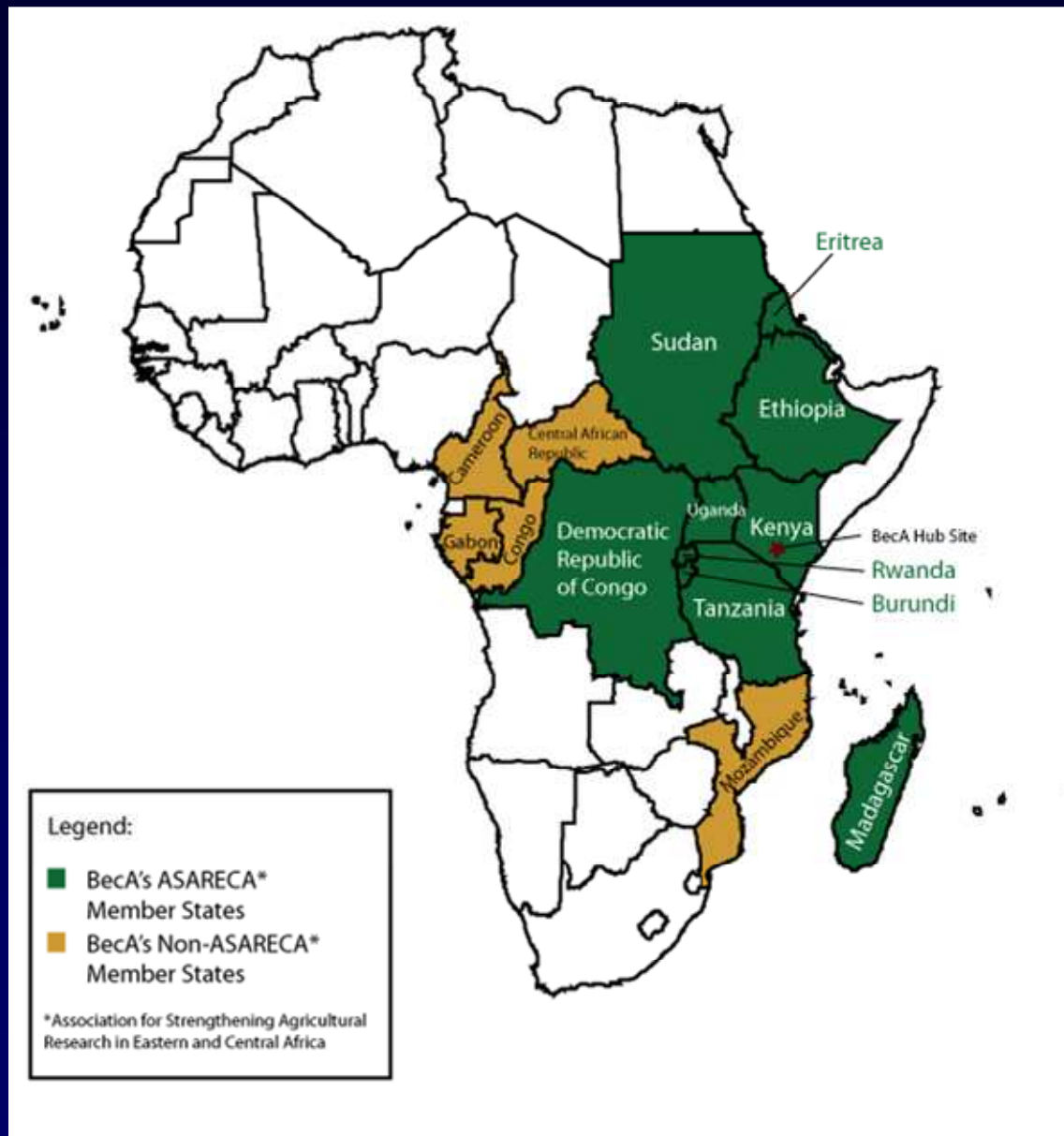
A network under the NEPAD biosciences initiative designed to catalyse and promote application of cutting edge scientific research for Africa's development

- Achieve excellence in science through the assembly of critical mass
- Develop African scientists' capacity to conduct, drive and fund their own research
- Increase access to state-of-the-art research facilities, reduce costs through joint activities
- Reverse brain drain, attract new investments, share knowledge

# BecA Design & Linkages



# BecA's Geographical Scope



# Core Competencies

- Bioinformatics
- Diagnostics
- Functional Genomics
- Sequencing and Genotyping
- Molecular Breeding
- Transformation
- Tissue Culture
- Vaccine Development

# BecA's Bioinformatics Platform

Provide scientist access to bioinformatics:

- applications and algorithms
- large-volume data storage and mining
- local mirror of all relevant databases
- basic training and helpdesk support
- BUILD CAPACITY

# Hardware

Paracel Cyclone is a pre-configured Linux cluster designed to optimise the performance of bioinformatics applications.

Consist of 3 components:

1. Beowulf Linux cluster - 66 Cpu / load balancing
2. Paracel Blast -150x faster than NCBI-Blast server
3. GeneMatcher2 - massively parallel 6144 CPU supercomputer designed to accelerate dynamic programming algorithms for highly sensitive and accurate genomic analysis

# Software and Applications

**DNA / Protein manipulation**

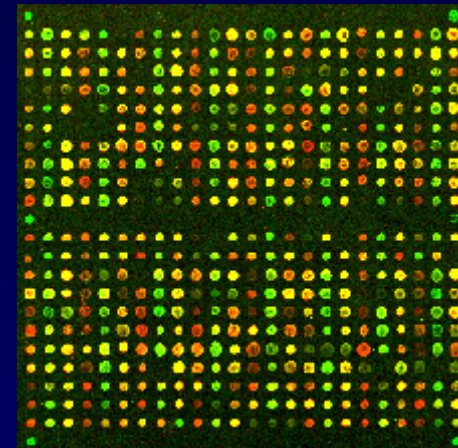
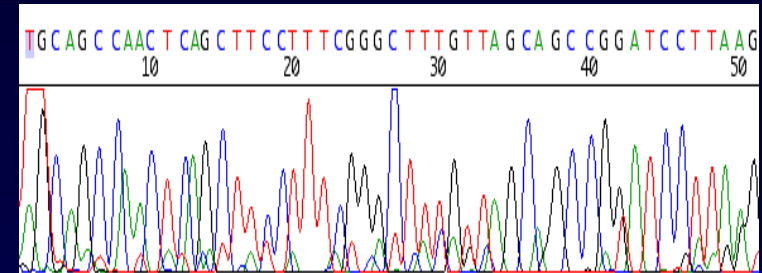
**Phylogenetic and evolutionary analysis**

**Sequencing, assembly and annotation**

**Functional genomics / Microarray**

**Immunoinformatics**

**Protein structure prediction**



# Databases

**SRS server**

**Swissprot + TrEMBL**

**Other publicly available databases**

**Specialised databases**

**Storage Area Network (SAN) with storage and backup capacity of upto 3 Terrabytes**

# Capacity Building

## Training laboratory

- 18 networked computers
- MS Windows and Linux
- High speed internet connectivity

# *Theileria parva* genome annotation

- Training set for gene finders
  - highly conserved genes (132 fl + 82 partial = 266 kb)
  - cDNAs from sch-inf lymphocytes (99 fl + 65 partial = 250 kb)
- Gene finding: GlimmerM and *phat*, plus Combiner
- AnnotationStation for genes structures
- Manatee for functional assignments
  - MySQL annotation database

# *Manatee*

- Manatee is a web-based gene evaluation and genome annotation tool that can view, modify, and store annotation for prokaryotic and eukaryotic genomes.
- The Manatee interface allows biologists to quickly identify genes and make high quality functional assignments using a multitude of genome analyses tools.
- These tools consist of, but are not limited to GO classifications, BER and blast search data, paralogous families, and annotation suggestions generated from automated analysis.

# Theileria parva genome annotation

525.t01215 / 525.m06301 - Mozilla (Build ID: 2002100315)

File Edit View Go Bookmarks Tools Window Help

http://bioinfo.ilri-ken.cgiar.org/tigr-scripts/euk\_manatee/sha

Home Bookmarks Red Hat Network Support Shop Products Training

**Theileria parva** Gene Curation Page Logged into [tpa1] as etienne

Help text goes here.

**GENE CURATION INFORMATION**

**525.t01215** ( )  
 Model: 525.m06301  
 Pub Locus:  
[View BER Searches](#)  
 asmb\_id: 525

STATUS: **CURATED**

end5/end3: 151096 / 150403  
 gene length: 694  
 protein length: 206  
 mol. wt.: 23355.28  
 pl: 8.18

database: tpa1  
 feat\_name/locus:   
 New Gene

gene list pager  
 << 525.t01214 525.t01216 >>

Select Function:   
[Reload Page](#)

Gene Synonyms: None  
 Intron/Exon/UTR structure:

**GENE IDENTIFICATION**

Gene Name: orotate phosphoribosyltransferase, putative  
 Gene Symbol:   
 EC Number: 2.4.2.10

comment:  
 blastp and PFAM hits

pub\_comment:

[auto\\_comment](#)

**GENE ONTOLOGY**

None Assigned

unknown process unknown function unknown component

Add go\_id Ev\_code Reference With

525.t01215 / 525.m06301 - Mozilla (Build ID: 2002100315)

File Edit View Go Bookmarks Tools Window Help

http://bioinfo.ilri-ken.cgiar.org/tigr-scripts/euk\_manatee/sha

Home Bookmarks Red Hat Network Support Shop Products Training

**EVIDENCE PICTURE**

525.m06301  
 TIGR00336: orotate phosphoribosyltransferase  
 PF00156: Phosphoribosyl transferase domain  
 Characterized match: SP:P13298

**HMM**

**TIGR00336: orotate phosphoribosyltransferase** gene\_sym: pyrE ec#: 2.4.2.10 role\_id: 126  
 Isology: equivalog\_domain Total score: 106.0 Trusted cutoff: 100.00 Noise cutoff: 25.00 Total expect: 7.5e-28

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
<a href="#">align page</a>	19-196	1-187 / 187	106.0	7.5e-28	<input type="checkbox"/>	

[GO:0004538](#) add orotate phosphoribosyltransferase (function)  
[GO:0009220](#) add pyrimidine ribonucleotide biosynthesis (process)

**PF00156: Phosphoribosyl transferase domain** gene\_sym: none ec#: none role\_id: none  
 Isology: domain Total score: 66.2 Trusted cutoff: 2.10 Noise cutoff: 1.60 Total expect: 7e-16

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
<a href="#">align page</a>	41-194	1-174 / 174	66.2	7e-16	<input type="checkbox"/>	

No HMM-GO Suggestions To Display.

**PROSITE**

No Prosite Data Available.

**ATTRIBUTES**

**SIGNAL\_P**

SignalP-2.0 Results: [Graphical Display](#) [Raw output for SP-HMM/NN](#)  
 SignalP-2.0 HMM  
 Prediction: Non-secretory protein  Curated  
 Signal peptide probability: 0.000  
 Signal anchor probability: 0.000  
 Max cleavage site probability: 0.000

# Comparative Genomics *T. annulata*

Use of Artemis and ACT developed at the Sanger Institute

Identification of unique genes in *T. parva* and *T. annulata* genomes

34 in TA and 60 in TP

Comparative analysis used to help understand mechanisms underlying transformation and cell tropism

Unequal expansion of paralogous gene families - despite high synteny and conservation

# MPSS

MPSS used to analyse the transcriptome

~1 million signatures - mapped onto the genome

Showed that majority of genes are transcriptionally active

Antisense signatures detected (~14%)

Verified by amplification and sequencing

# Other BecA projects

Develop drought stress-response enriched EST resources

Use of the Paracel Transcript Assembler

A comprehensive annotated EST collection would provide an invaluable tool facilitating

- The development of new molecular markers

- The development of oligos for microarray and real time PCR

# CIMMYT

- Storage and analysis pipeline of Microarray data
- MaxD Load and View customised to be GCP compliant and Plant MIAME compliant
  - Connection to ICIS – Germplasm and study
  - Used as a data source for other microarray analysis pipeline tools eg: TMEV

**maxdLoad2** : An extensible, MIAME-compliant database for microarray experiments

- A database schema and a software application.
- The second-generation of **maxdLoad**.
- Integrated data loading, browsing, editing and searching.
- Written in **Java™**, runs on most computers...
- Supports any **SQL92** database:  
Oracle, MySQL, Postgres, Sybase, Firebird

# Main Features

- Loading, browsing, editing and searching.
- Extensible: customisable attributes for each part of the schema.
- MIAME data capture.
- MAGE-ML data export.

# MIAME and the MGED Ontology

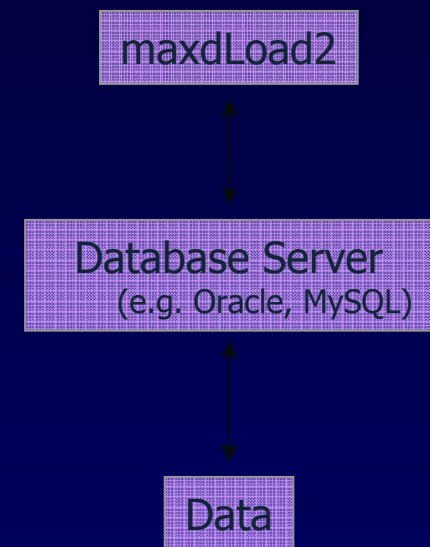
- The default configuration is designed to capture all of the meta-data required by the MIAME specification.
- Where possible, the terminology defined by the MGED Ontology is used, e.g:

```
HardwareType :  
  DNA_sequence,  
  homogenizer,  
  wash_station,  
  hybridization_chamber,  
  vortexer,  
  ...
```

- Numerical values use the units supported by the MAGE Object Model.

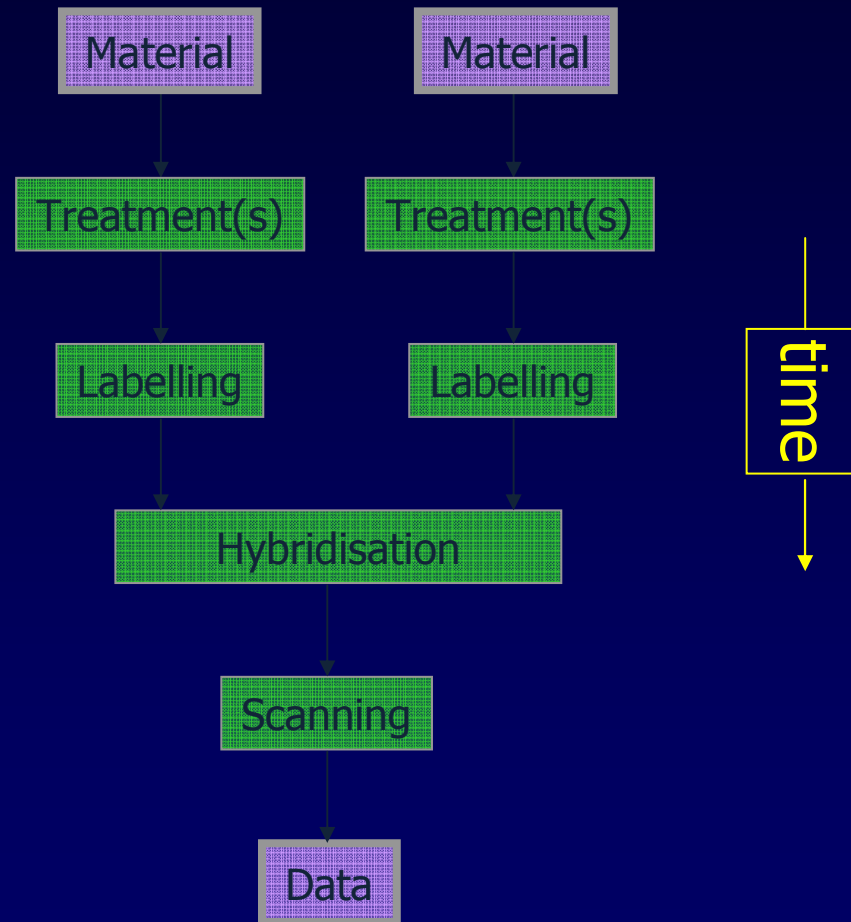
# System Architecture

- **maxdLoad2** is NOT accessed via a web-browser
- It is a stand-alone application, written in Java (this makes it very portable).
- **maxdLoad2** and the database server can run on the same machine, no network connection or web server is needed.
- However, **maxdLoad2** and the database server can be on separate machines connected via a network.



# Microarray Experiment Workflow

- A typical microarray experiment is a sequence of steps starting with one or more 'BioMaterials' and ending up with a big pile of numbers.
- These steps can be thought of as transformations:  
material A + treatment = material B  
and combinations:  
image + scanning = data
- Each of the steps needs to be recorded in the database.
- Many of the steps will be standardised, for example, the protocol used for labelling. They will only have to be defined once.



# What is the database made of?

- **BioMaterials**  
Sources, Samples, TreatedSamples, Extracts, LabelledExtracts
- **Protocols**  
SamplingProtocol, ScanningProtocol, LabellingProtocol, etc.
- **Arrays**  
ArrayTypes, Features, Reporters & Genes
- **Hybridisations**  
Experiments, Measurements, Images & Hybridisations

# The User Interface



# The User Interface

- These buttons control which mode the software is in (create, browse, find, edit or load)
- These buttons are used to open the form used to input or explore the data for each of the database components
- The arrows show how the components are interconnected
- These buttons access the other main features: import, export, options and the built-in help system.



# The User Interface

- Clicking on one of the boxes opens a form in which the full details of an instance can be viewed (and edited)

The screenshot displays the 'BROWSE LABELLEDEXTRACTS' application window. On the left, a list of 344 items is shown, with 'Study51-i2:post/1/eBL1' selected. The main area shows the details for this selected item:

- Name:** Study51-i2:post/1/eBL1
- Extract:** Study51-i2:post/1/eB
- LabellingProtocol:** Labelling Protocol - L1
- Attributes:**
  - External Identifier:** A section with the instruction 'If this item exists in some external database, this attribute can be used to refer to it' and a 'Name' input field.
  - Application of the Protocol:** A section with the instruction 'How was the selected LabellingProtocol used to create this LabelledExtract, when was it done and who did it?'. It includes:
    - Type Of Action:** labeling
    - Performed By:** Fred
    - Performed On:** Day 19, Month 6, Year 2001
  - Comments:** A large text area for notes.
  - Description:** A large text area for a detailed description.

At the bottom of the window, there are buttons for 'Edit', 'Delete', 'Export', 'Help', and 'Close'. A filter bar at the bottom left shows '344 items, no filter used' and a 'Select all' button.

**Thank you for your attention!!**