myGrid

Katy Wolstencroft
University of Manchester

Background

- myGrid middleware components to support *in silico* experiments in biology

- Originally designed to support bioinformatics

  chemoinformatics

  health informatics

  medical imaging

  integrative biology

History

EPSRC funded UK eScience Program Pilot Project

MERCK

AstraZeneca

SOURCEFORGE.net

Sun microsystems

gsk GlaxoSmithKline

EMBL-EBI European Bioinformatics Institute

network inference

IBM

MANCHESTER 1824 The University of Manchester

EPISTEMICS LTD knowledge is our business

genetic change

University of Southampton

iT innovation

The Wellcome Trust

LION

THE UNIVERSITY OF NOTTINGHAM

UNIVERSITY OF NEWCASTLE UPON TYNE

myGrid

myGrid in OMII-UK

myGrid

OMII Stack

OGSA-DAI

March 2006

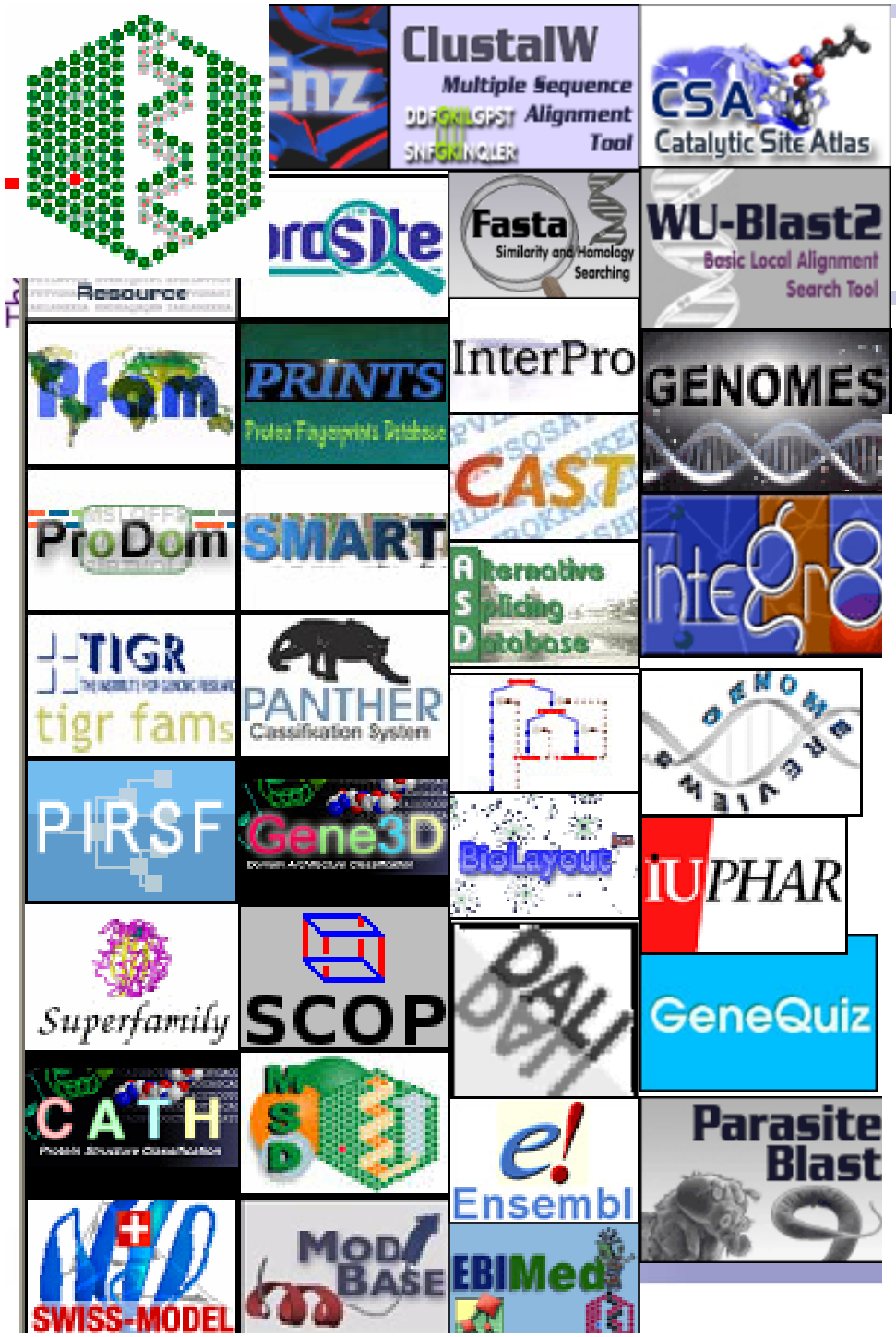
Virtual Grid of Resources

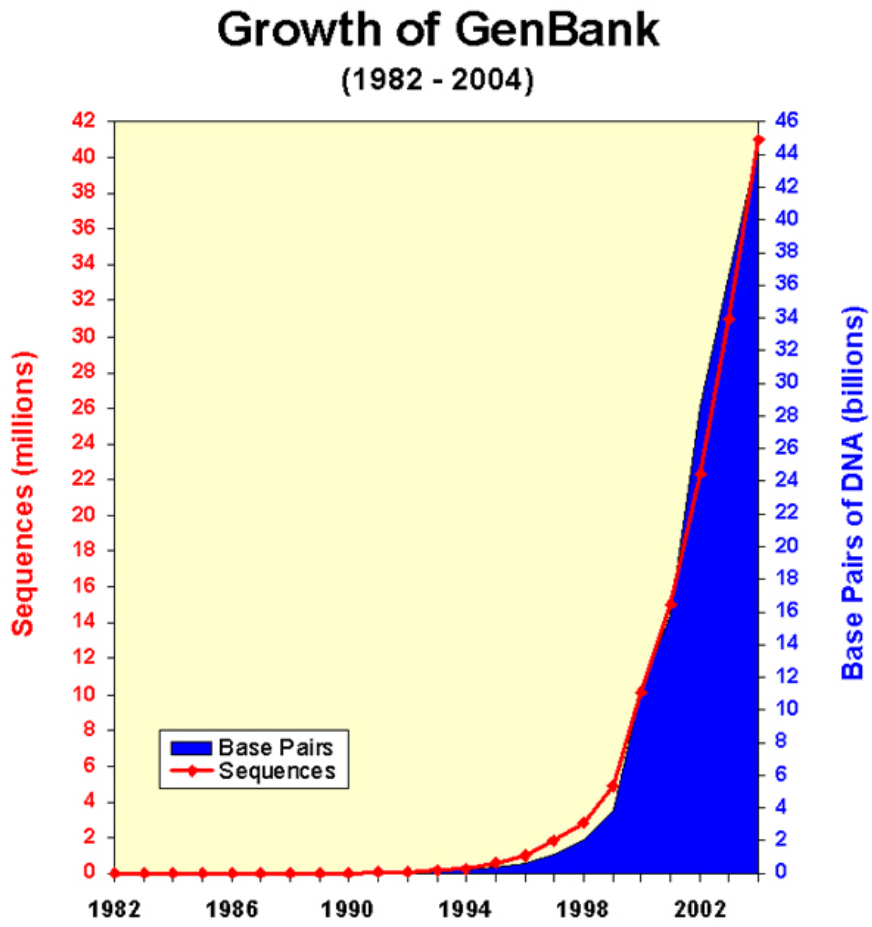
- Biology knowledge-rich
- Applying prior knowledge to new data

- myGrid middleware to enable interoperation between distributed data and resources – a grid of data – not a grid of resources

Lots of Resources

NAR 2005 – over 700 databases





The User Community

Bioinformatics is an open Community

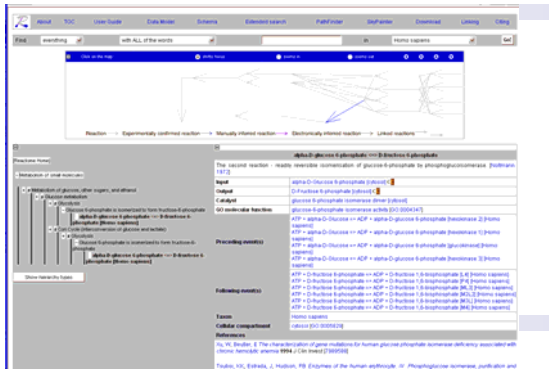
- Open access to data
- Open access to resources
- Open access to tools
- Open access to applications

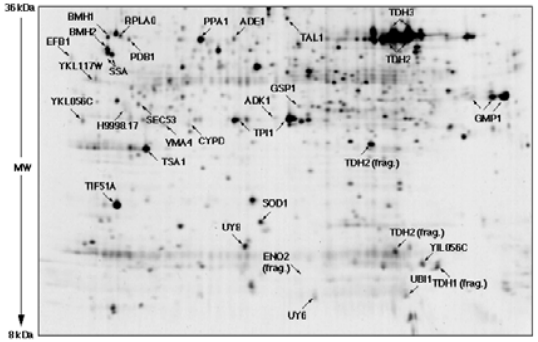
Global *in silico* biological research

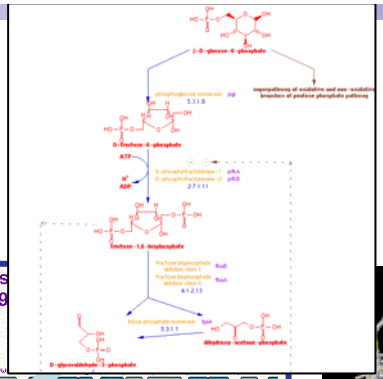
The User Community Problems

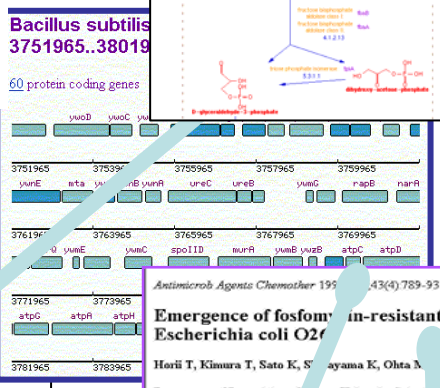
- Everything is Distributed
  - Data, Resources and Scientists
- Heterogeneous data
- Very few standards
  - I/O formats, data representation, annotation
  - Everything is a string!

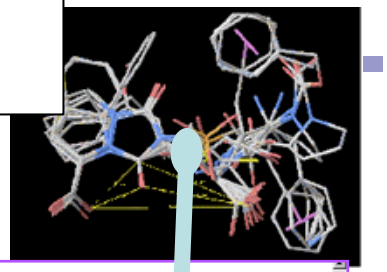
Integration of data and interoperability of resources is difficult

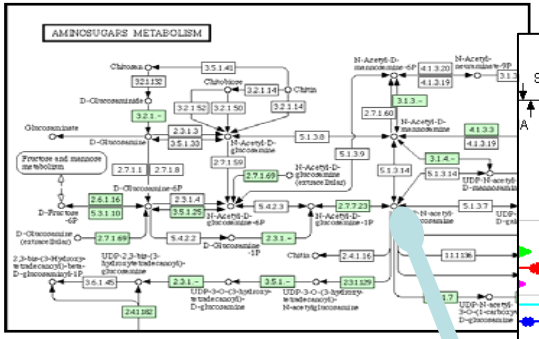


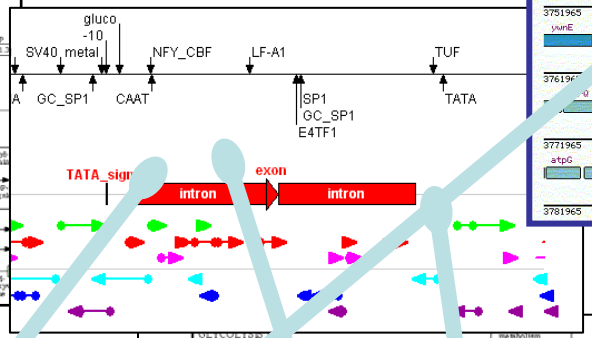

36 kDa
MW
8 kDa
BMH1
RPLA0
PPA1
ADE1
TDH3
EFB1
BMH2
TAL1
PDB1
TDH2
YKL117W
SSA
GSP1
YKL056C
ADK1
SEC53
GMP1
H9998.17
TPI1
YMA4
CYPD
TSA1
TDH2 (frag.)
TIF51A
SOD1
UYE
TDH2 (frag.)
YIL056C
ENO2 (frag.)
UBI1
TDH1 (frag.)
UYE



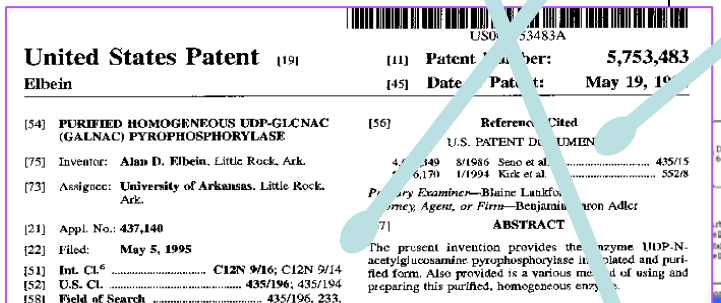

Bacillus subtilis
3751965..38019
60 protein coding genes

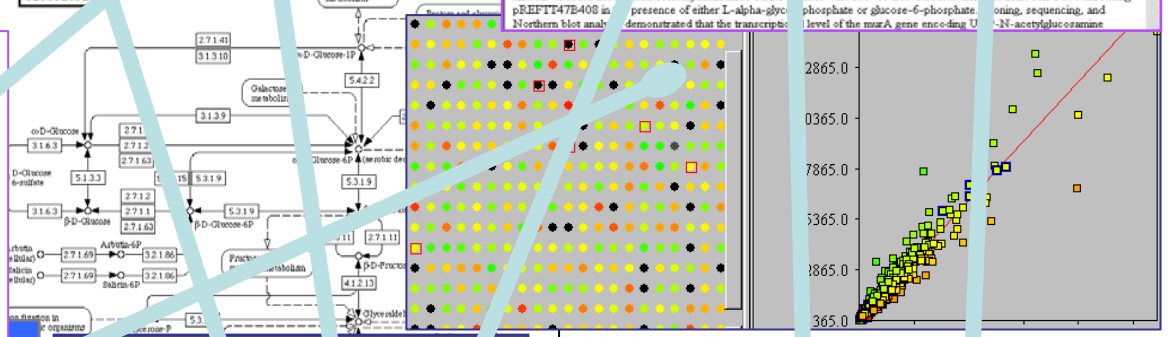


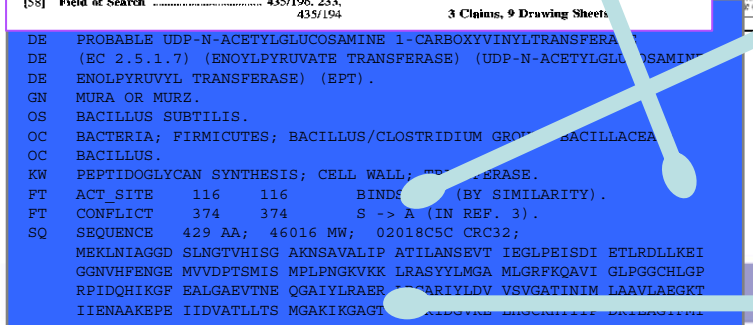

AMINOSUGARS METABOLISM

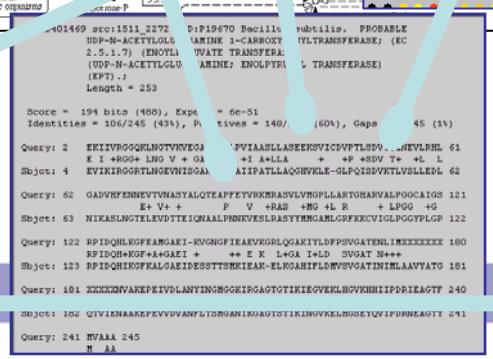

gluco
-10
SV40
metal
NFY_CBF
LF-A1
TUF
GC_SP1
CAAT
SP1
GC_SP1
E4TF1
TATA
TATA_sign
exon
intron
intron

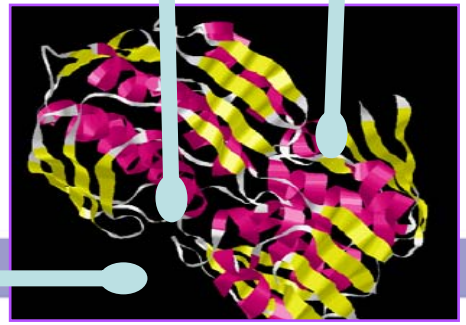

Antimicrob Agents Chemother
Emergence of fosfomycin-resistant isolates of Shiga-like toxin-producing Escherichia coli O26
Horii T, Kimura T, Sato K, Sayama K, Ohta M
Department of Bacteriology, Nagoya University School of Medicine, Japan. horii@tsuru.med.nagoya-u.ac.jp


United States Patent [19]
Elbein
[11] Patent Number: 5,753,483
[45] Date Patent: May 19, 19
[54] PURIFIED HOMOGENEOUS UDP-GLCNAC (GALNAC) PYROPHOSPHORYLASE
[75] Inventor: Alan D. Elbein, Little Rock, Ark.
[73] Assignee: University of Arkansas, Little Rock, Ark.
[21] Appl. No.: 437,140
[22] Filed: May 5, 1995
[51] Int. Cl.6 C12N 9/16; C12N 9/14
[52] U.S. Cl. 435/196; 435/194
[58] Field of Search 435/196, 233, 435/194
[56] Referenced Cited
U.S. PATENT DOCUMENTS
ABSTRACT
3 Claims, 9 Drawing Sheets




DE PROBABLE UDP-N-ACETYLGLUCOSAMINE 1-CARBOXYVINYLTRANSFERASE
DE (EC 2.5.1.7) (ENOYLPYRUVATE TRANSFERASE) (UDP-N-ACETYLGLUCOSAMINE
DE ENOLPYRUVYL TRANSFERASE) (EPT).
GN MURA OR MURZ.
OS BACILLUS SUBTILIS.
OC BACTERIA; FIRMICUTES; BACILLUS/CLOSTRIDIUM GROUP; BACILLACEAE
OC BACILLUS.
KW PEPTIDOGLYCAN SYNTHESIS; CELL WALL;
FT ACT_SITE 116 116 BINDS (BY SIMILARITY).
FT CONFLICT 374 374 S -> A (IN REF. 3).
SQ SEQUENCE 429 AA; 46016 MW; 02018C5C CRC32;
MEKLNIAGGD SLNGTVHISG AKNSAVALIP ATILANSEVT IEGLPEISDI ETLRDLLKEI
GGNVHFENGE MVVDPTSMIS MPLPNGKVKK LRASYYLMGA MLGRFKQAVI GLPGGCHLGP
RPIDQHIKGF EALGAEVTNE QGAIYLRAER
IIENAAKEPE IIDVATLLTS MGAKIKGAGT





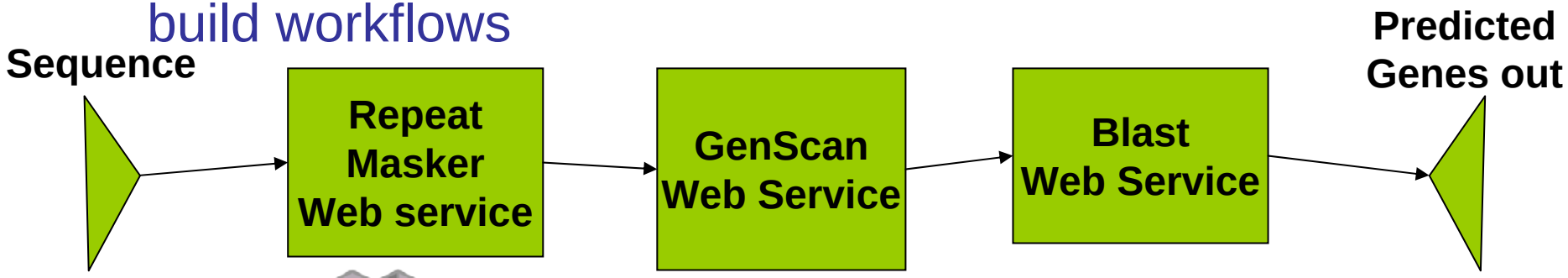
myGrid Approach - Workflows

General technique for describing and enacting a process

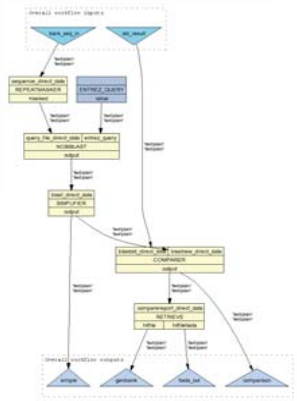
describes *what* you want to do, not *how* you want to do it

Simple language specifies how bioinformatics processes fit together – processes are web services

- High level workflow diagram separated from any lower level coding – therefore, you don't have to be a coder to build workflows



SCUFL



Taverna Workbench

Application data flow layer

Scufl graph + service introspection

Scufl + Workflow Object Model

Execution flow layer

List management; implicit iteration mechanism; MIME & semantic type decoration; fault management; service alternates

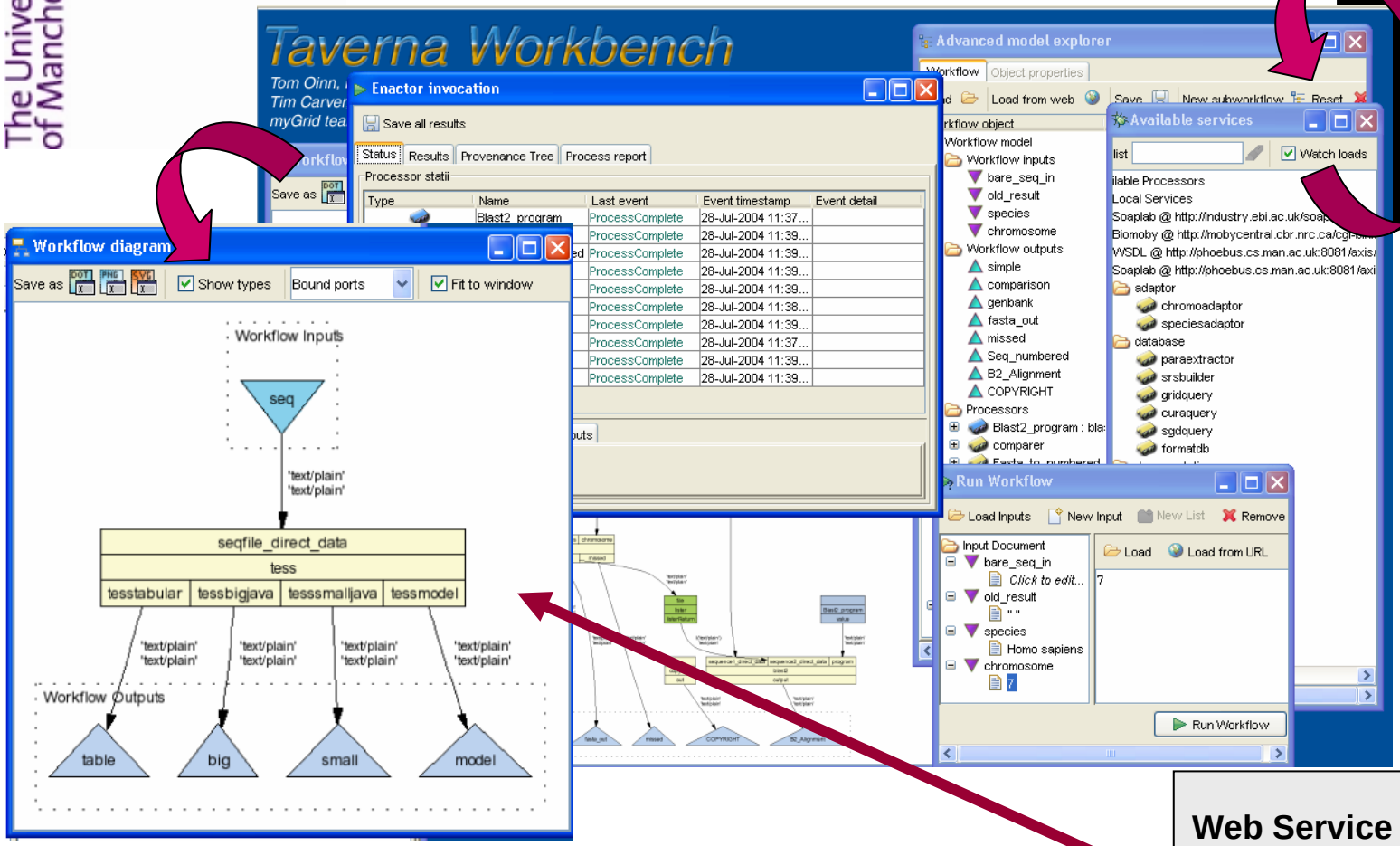
Workflow Execution

Freefluo Workflow enactor

Processor invocation layer

Processor	Processor	Processor	Processor	Processor	Processor	Processor
Bio MART	Seq Hound	Plain Web Service	Soap lab	Bio MOBY	Local App	Enactor

Taverna Workflow Components



Freefluo

Freefluo Workflow engine to run workflows

Web Service | **e.g. DDBJ BLAST**

SOAPLAB Web Service | **Any Application**

Scufl Simple Conceptual Unified Flow Language
Taverna Writing, running workflows & examining results
SOAPLAB Makes applications available

So many services – semantic discovery

Over 3000 services

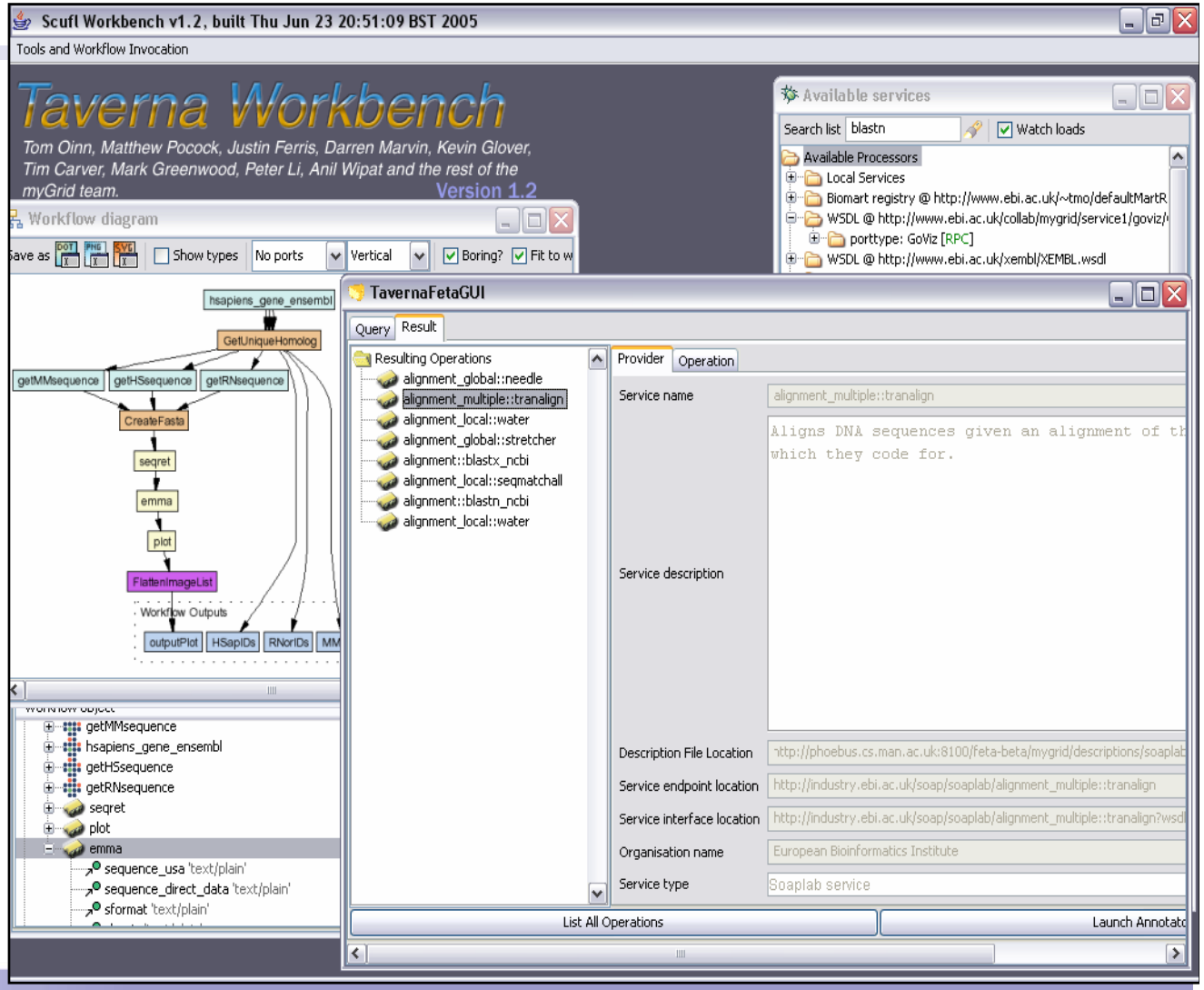
SeqHound – Database of biological sequences and tools

BioMart – Federated query system

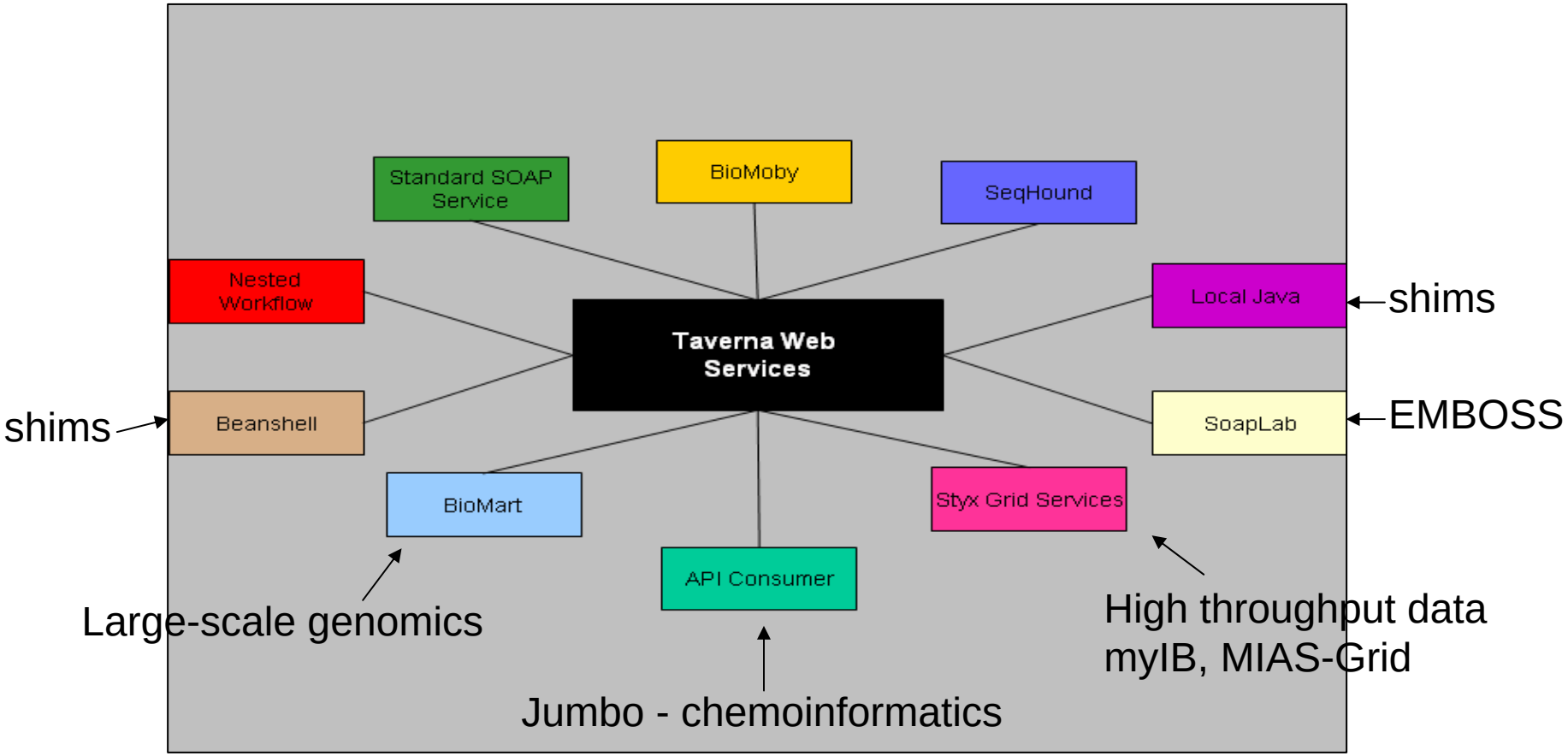
EMBOSS – Sequence analysis tools

BioMoby – Collection of web services

EBI SOAPLAB – Collection of supported services



What Services we Support



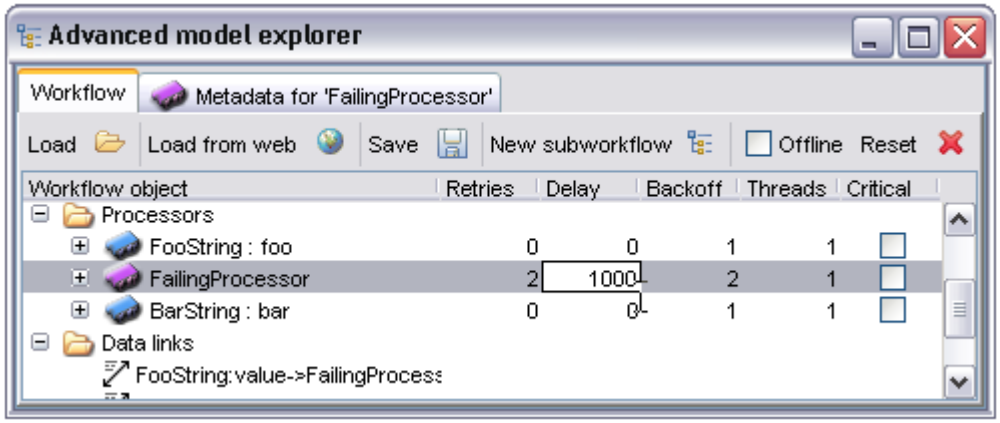
What shall I do when a service fails?

- Most services are owned by other people
- No control over service failure
- Some are research level

Workflows are only as good as the services they connect!

To help - Taverna can:

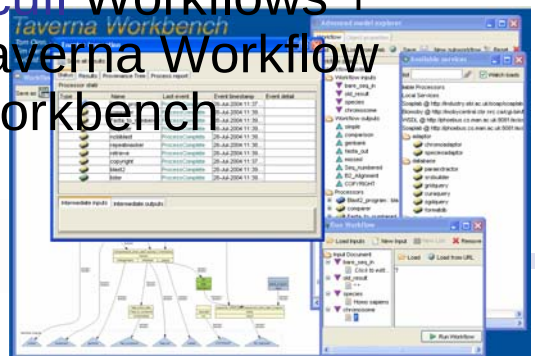
- Notify failures
- Instigate retries
- Set criticality
- Substitute services
- Instigate checkpoints for lon



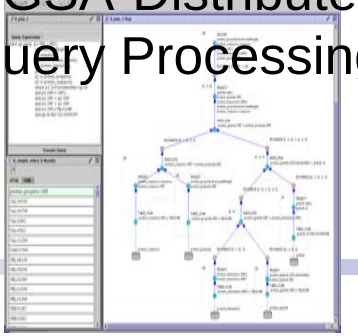
Data Management

- Workflows can generate vast amount of data - how can we manage and track it?
- Data AND metadata AND experiment provenance
- LSIDs - to identify objects
- Semantic Web technologies (RDF, Ontologies)
  - To store knowledge provenance
- Taverna workflow workbench & plugins
  - Ensure automated recording

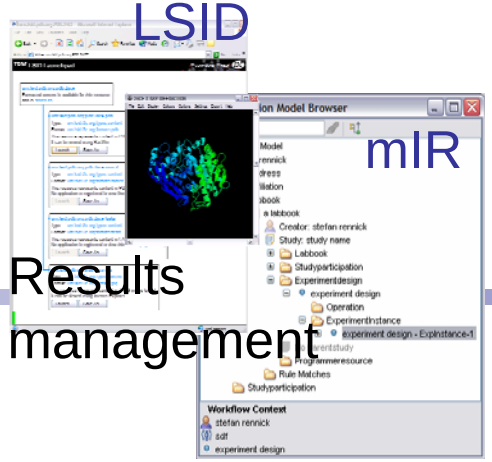
Scufl Workflows + Taverna Workflow Workbench



OGSA-Distributed Query Processing



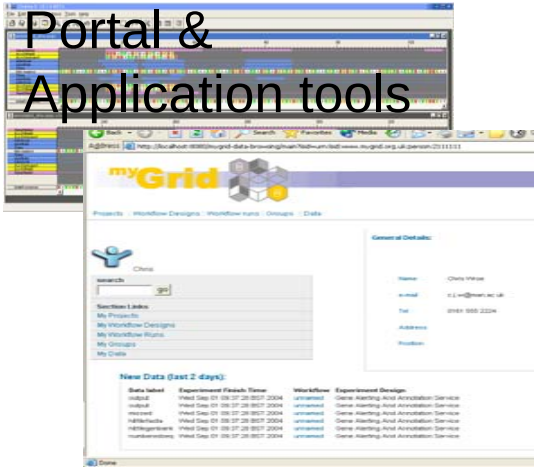
LSID

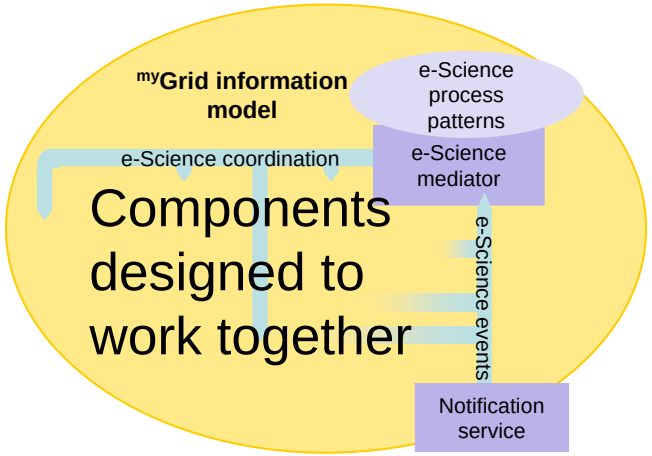


mIR

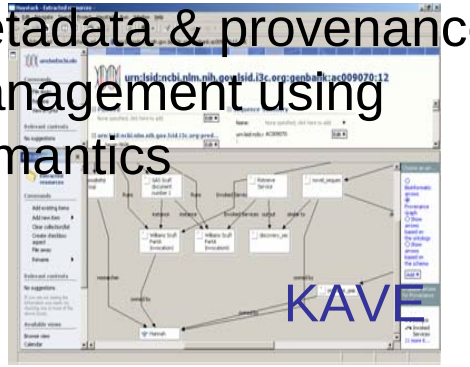
Results management

Portal & Application tools



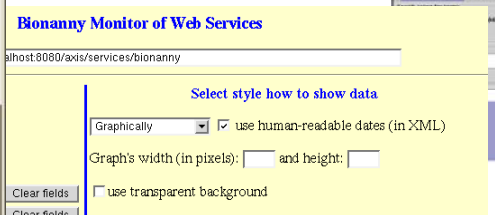


Metadata & provenance management using semantics



KAVE

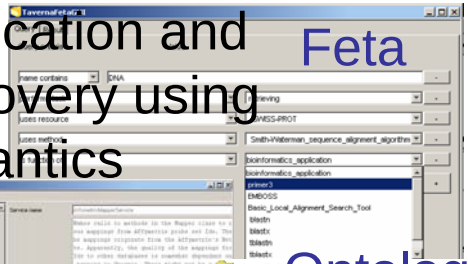
Service management



Text Mining Services

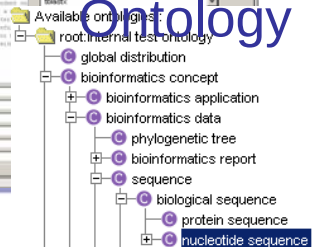
Termino

Publication and Discovery using semantics



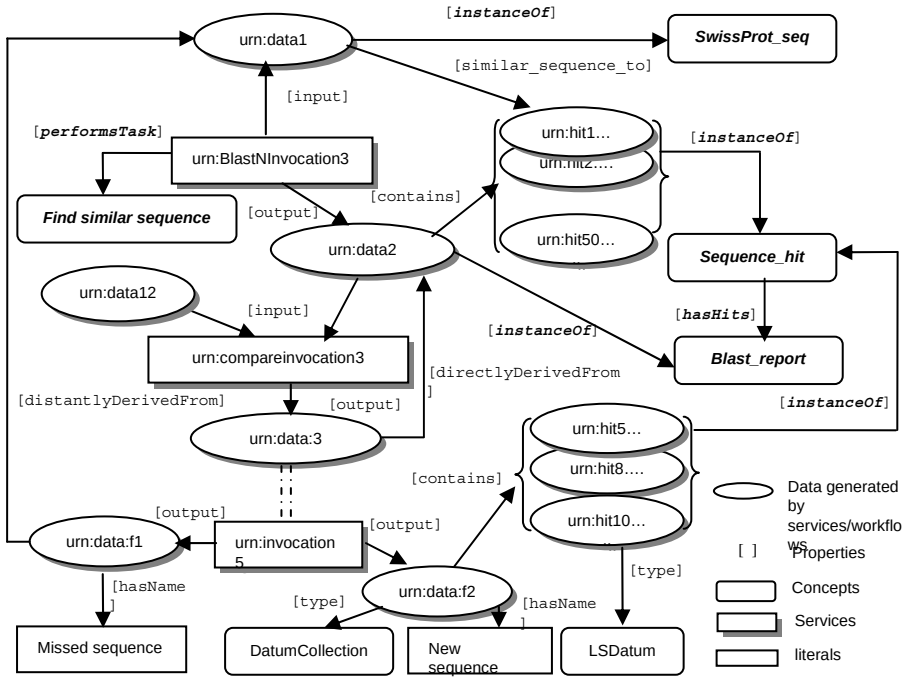
Feta

Pedro



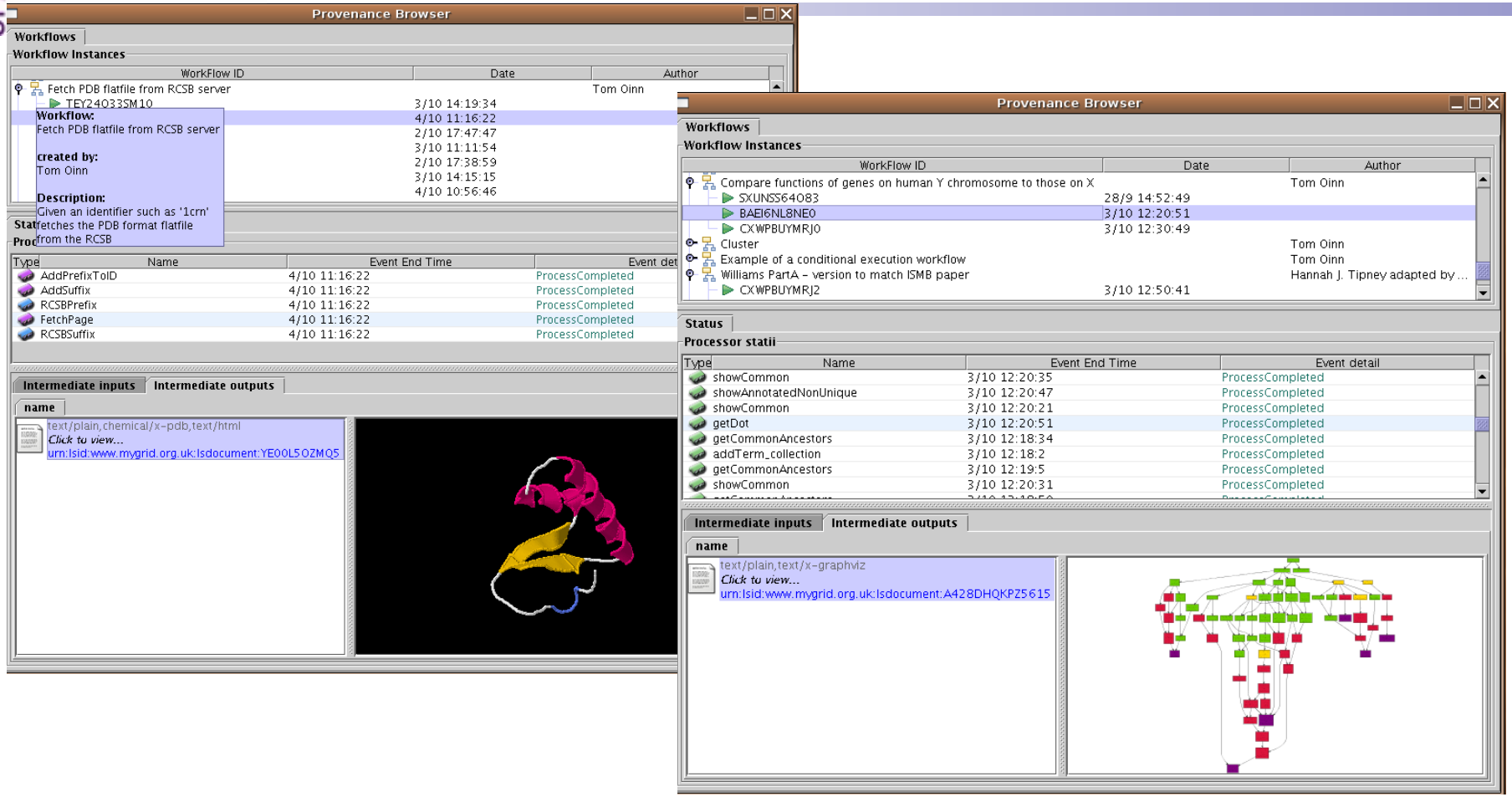
Ontology

KAVE Data and metadata management

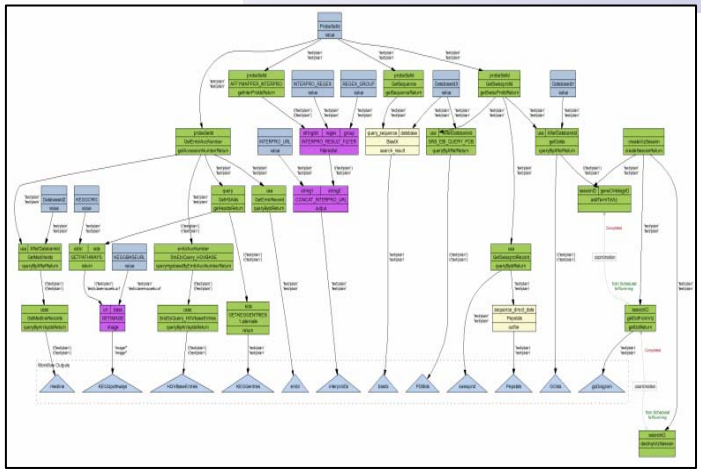


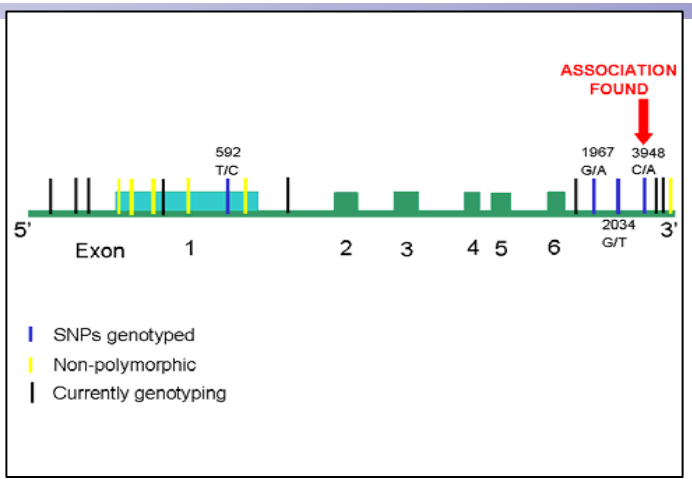
- Life Science Identifiers
- Information Model
- File management
- Support for custom database building
- Provenance metadata capture using RDF
- SRB integration
- OGSA-DAI integration

Provenance Browsing in Taverna



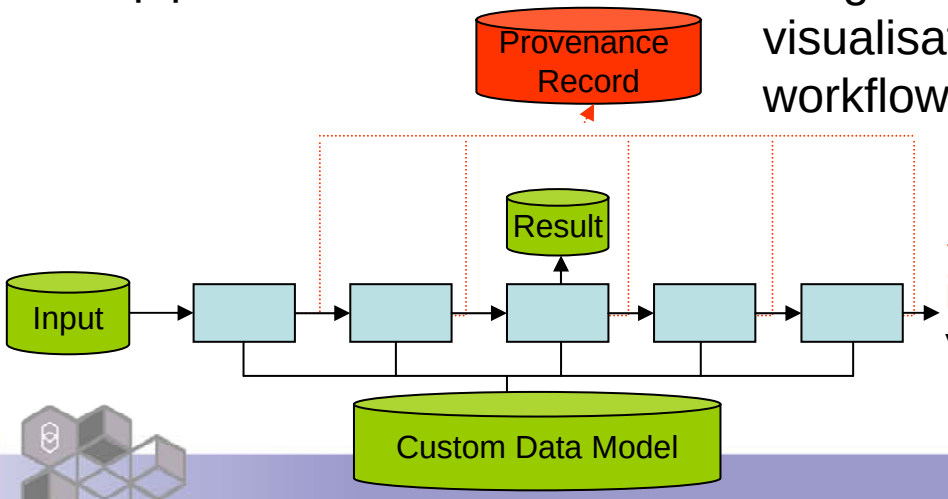
Results Integration





Gene annotation pipeline workflow

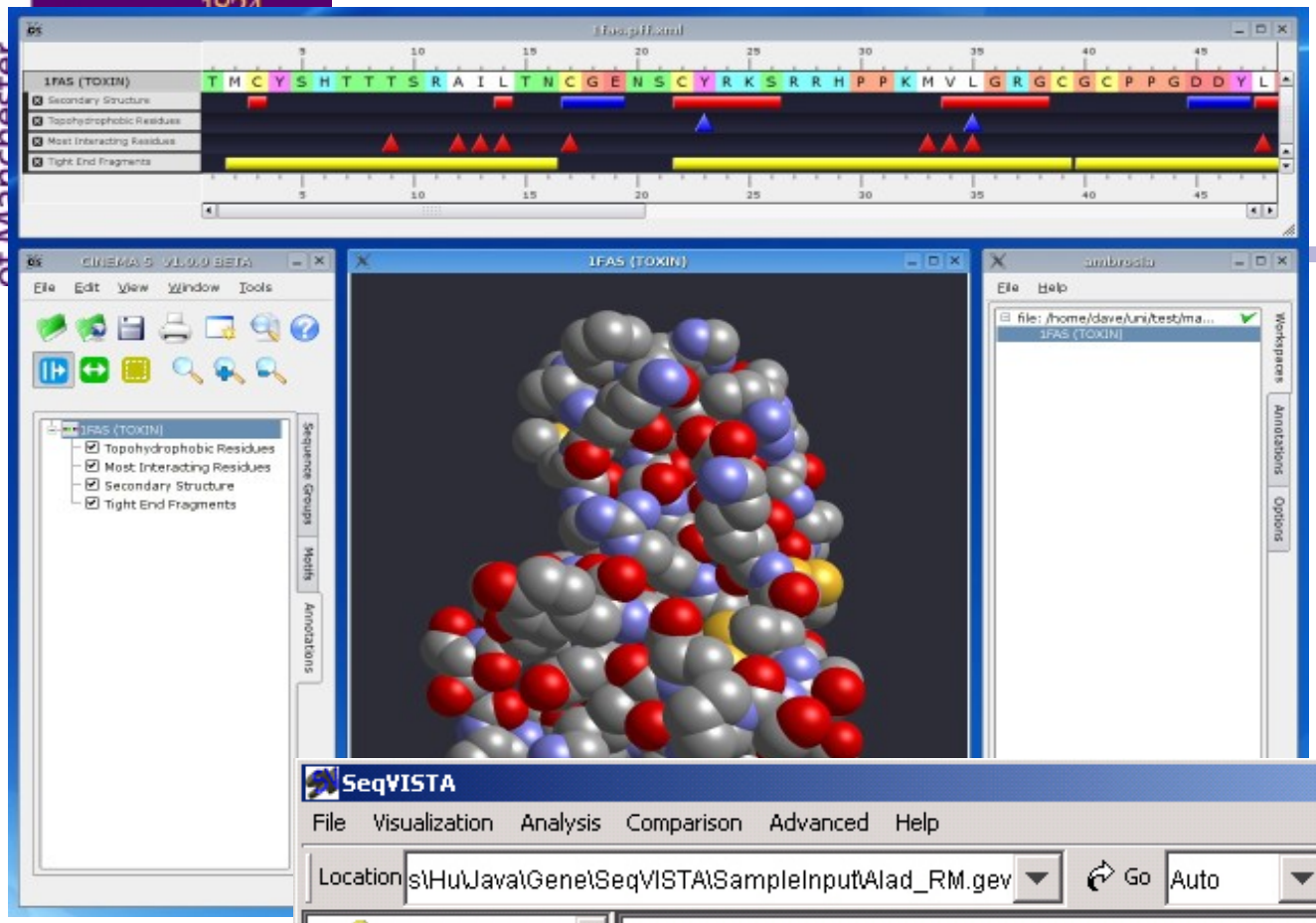
Integration and visualisation of GD annotation workflow results



Smarter workflow design incorporating visualisation

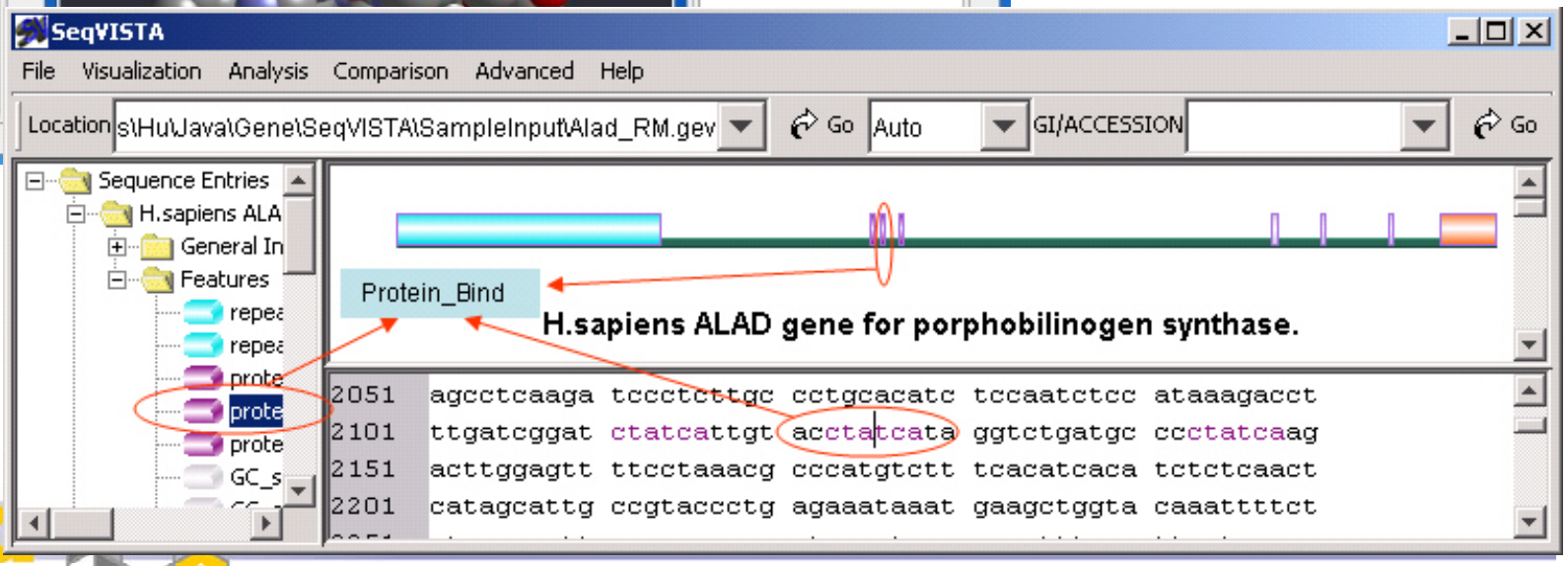
VBI collaboration

Visualisation



Utopia

SeqVista



Applications

Resistance to trypanosomiasis in cattle in Kenya

Andy Brass, Paul Fisher – University of Manchester

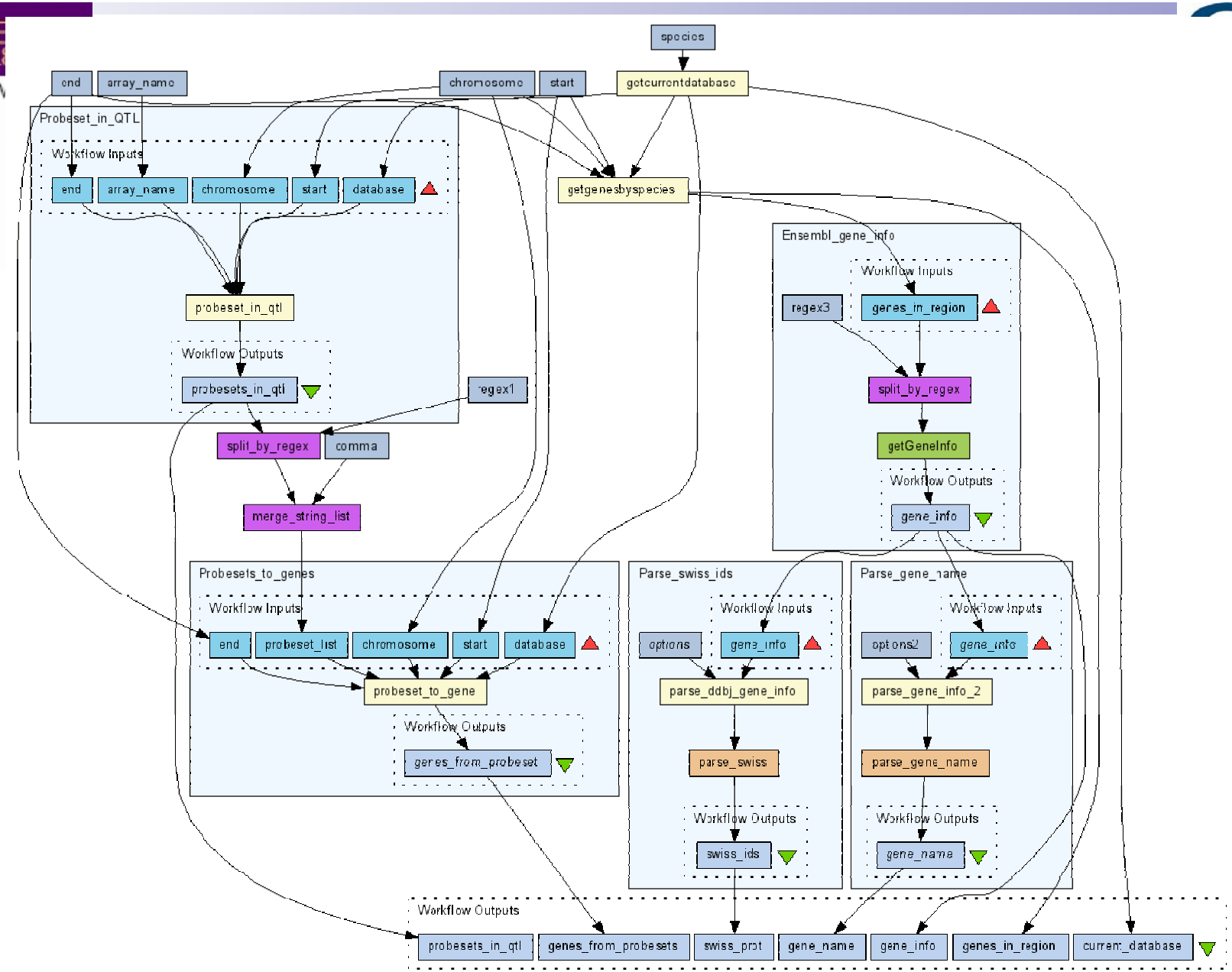
Microarray

QTL

SNPs

Metabolic pathway analysis

Need to access microarray data, genomic sequence information, pathway databases AND integrate the results


species
end
array_name
chromosome
start
getcurrentdatabase
Probeset_in_QTL
Workflow Inputs
end
array_name
chromosome
start
database
probeset_in_qtl
Workflow Outputs
probesets_in_qtl
getgenesbyspecies
Ensembl_gene_info
Workflow Inputs
regex3
genes_in_region
split_by_regex
getGeneInfo
Workflow Outputs
gene_info
regex1
split_by_regex
comma
merge_string_list
Probesets_to_genes
Workflow Inputs
end
probeset_list
chromosome
start
database
probeset_to_gene
Workflow Outputs
genes_from_probeset
Parse_swiss_ids
Workflow Inputs
options
gene_info
parse_ddbj_gene_info
parse_swiss
Workflow Outputs
swiss_ids
Parse_gene_name
Workflow Inputs
options2
gene_info
parse_gene_info_2
parse_gene_name
Workflow Outputs
gene_name
Workflow Outputs
probesets_in_qtl
genes_from_probesets
swiss_prot
gene_name
gene_info
genes_in_region
current_database

myGrid Alliance: Applications

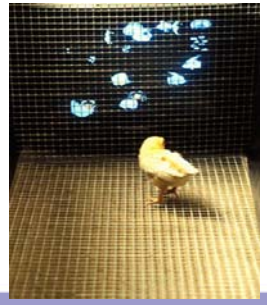
Large user community – over 15600 downloads

PsyGrid

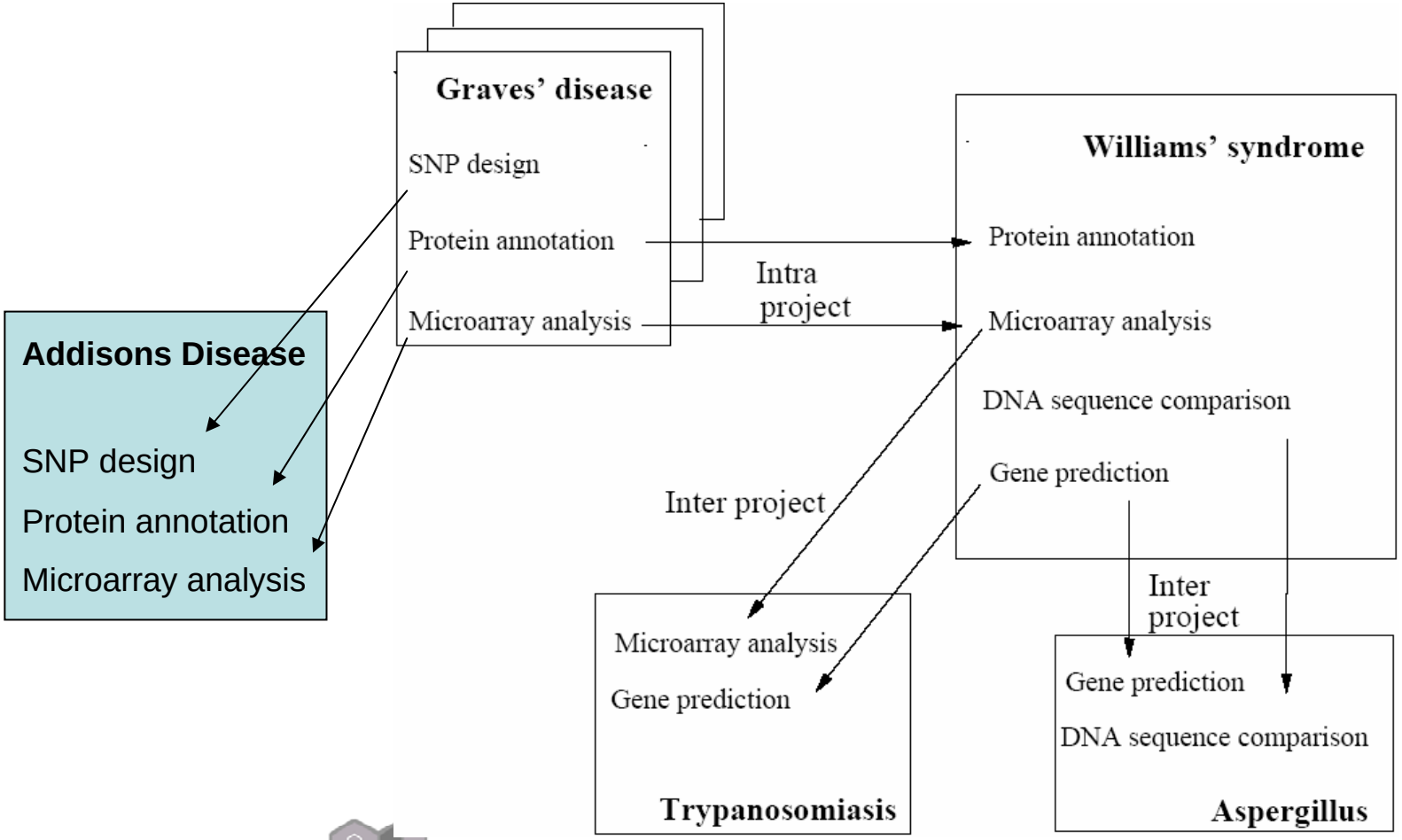
Small molecules,
Murray-Rust, Cambridge

Mias-Grid

Chicken genome
Roslin Institute



Workflow Reuse



Taverna is now OMII-UK

- Taverna 1.3.1 production Sept 2006
  - Packaging, Installation, Deployment, Maintenance, Testing
  - GridSAM, GRIMOIRES, BioMOBY integration
  - Semantic content for registry
  - Smoothed integration of discovery and metadata management
  - Security AA for KAVE data and metadata management
- Taverna 2.0 Spring 2007
  - Redevelopment of the plug in and enactor framework, improved iteration events, data management
- Close collaboration with pioneers
- Incremental rollouts to early adopters

Taverna in OMII-UK

- Development of Taverna 2.0
  - reworking of the processor model to include duel execution semantics incorporating data and control flow
  - enhanced support for long-running workflows
  - large scale data transfer
  - improved provenance collection with nested workflows and complex iterations
  - fully distributed workflow enactment and authoring

Acknowledgements

- Carole Goble and the myGrid team
- OMII-UK
- All of our users