MANCHESTER 1824
The University of Manchester

Introduction to Workflows with Taverna and myExperiment

Stian Soiland-Reyes and Christian Brenninkmeijer
University of Manchester

materials by Dr Katy Wolstencroft and Dr Aleksandra Pawlik

http://orcid.org/0000-0001-9842-9718
http://orcid.org/0000-0002-2937-7819
http://orcid.org/0000-0002-1279-5133
http://orcid.org/0000-0001-8418-6735

Bonn University, 2014-09-01
http://www.taverna.org.uk/

myGrid

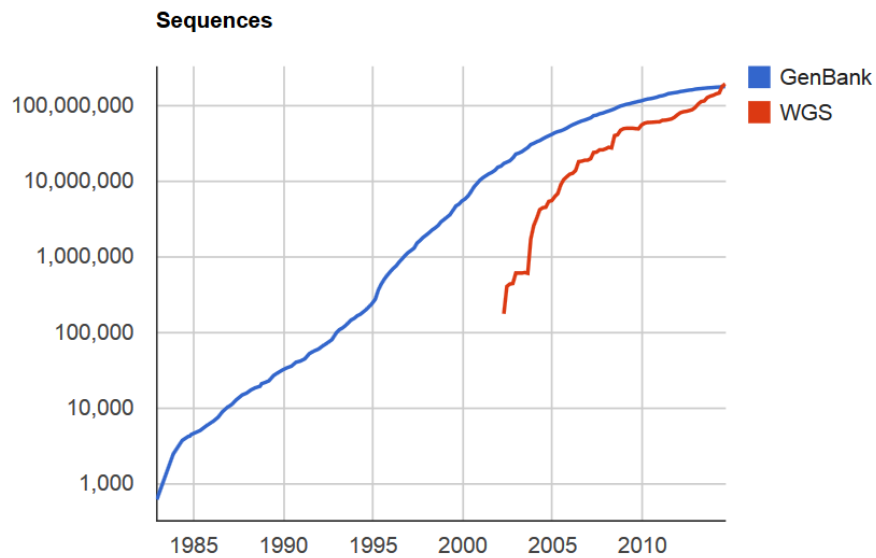
MANCHESTER 1824
The University of Manchester

Data Deluge

- ‘Omics data
- Next Gen Sequencing
- eGovernment
- World bank data
- Climate change data
- Large scale physics
  - Large Hadron collider
  - Astronomy

Lots of Resources

NAR 2014: 1552 databases
Genbank 2014-04: 172 million sequences, 162 billion basepairs
WGS 2014-04: 774 billion basepairs



myGrid

Next Generation Sequencing

- 2008-2012: 1000 Genome Project
  - A Deep Catalog of Human Genetic Variation
- 2009-: Genome 10k project
  - A genomic zoo—DNA sequences of 10,000 vertebrate species, approximately one for every vertebrate genus.
- 2012-: Human Microbiome Project
  - Characterise the microbial communities found at several different sites on the human body

Where is the data?

- Repositories run by major service providers (e.g. NCBI, EBI)
- Local project stores
- Static web pages
- Dynamic web applications
- FTP servers (!)
- Inside PDFs ☹
- Web Services ☺

NCBI
National Center for Biotechnology Information (USA)

DDBJ DNA Data Bank of Japan
Tokyo, Japan

EMBL-EBI European Bioinformatics Institute
Cambridge, UK

bio mart

SoapLab

wellcome trust sanger institute

KEGG Kyoto Encyclopedia of Genes and Genomes

MOUNT SINAI HOSPITAL

PDBj XP Protein Data Bank Japan Experimetal Version

PathPort The Pathogen Portal Web Project

emboss

moby

SeqVISTA

LION
SRS

SeqHound

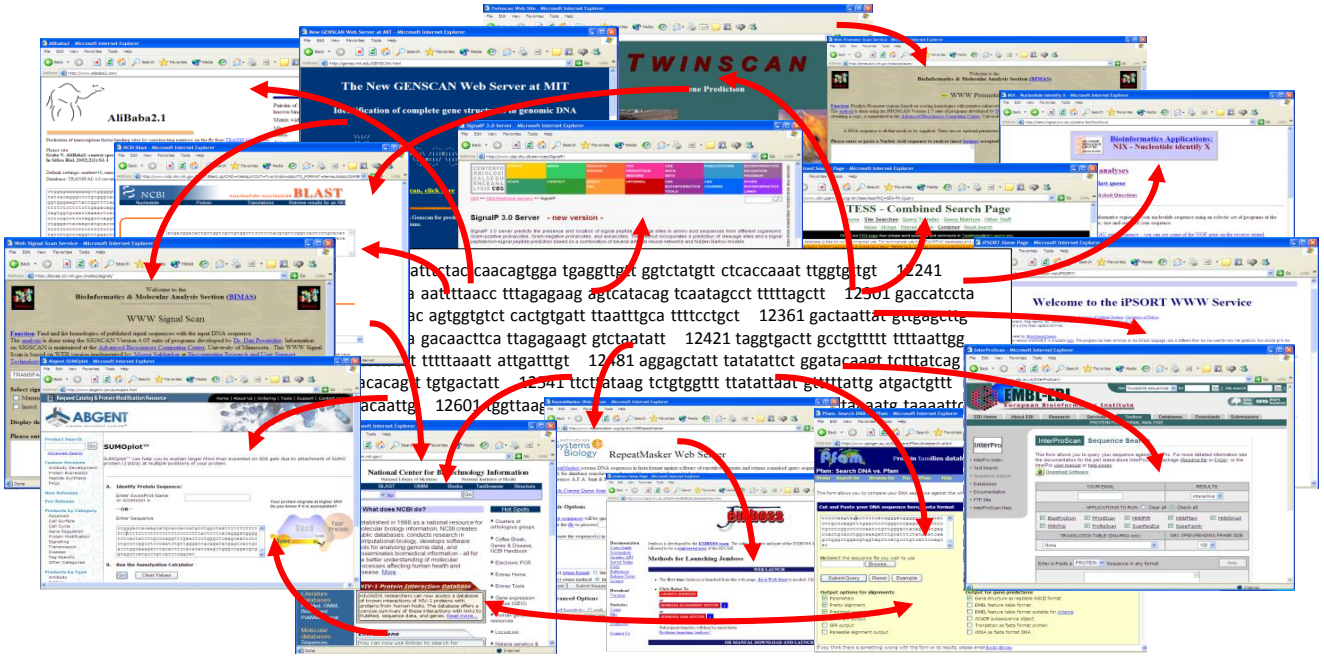
CaBIO

MANCHESTER 1824
The University of Manchester

The implicit workflow

Bioinformatics research combines:

- Data resources (public and private)
- Computational power (standard and custom)
- Researchers and collaborators



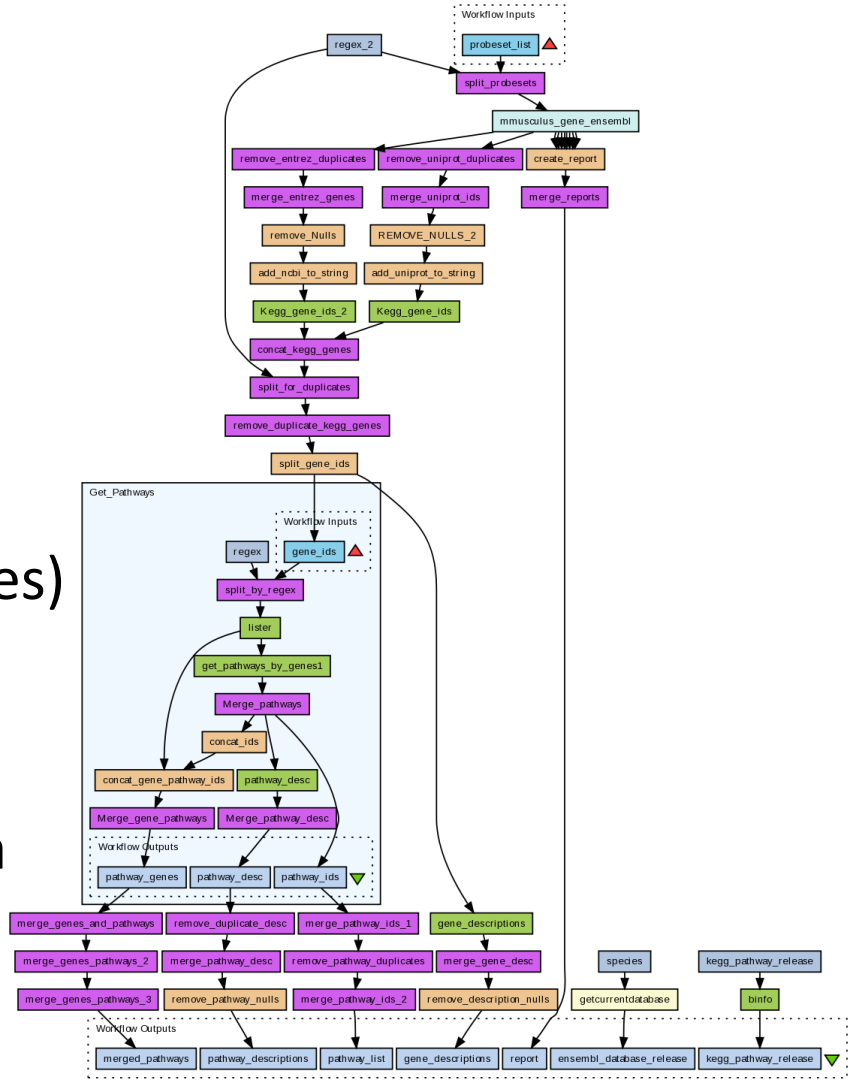
What that means for Bioinformatics

- Sequential use of distributed tools
- Incompatible input and output formats
- Challenging to record/reproduce/tweak
  - parameter selections
  - service selection
  - results of each step
- OK for one gene or one protein, but what about 10,000?
  - Analysing large data sets requires programmatic help

MANCHESTER 1824
The University of Manchester

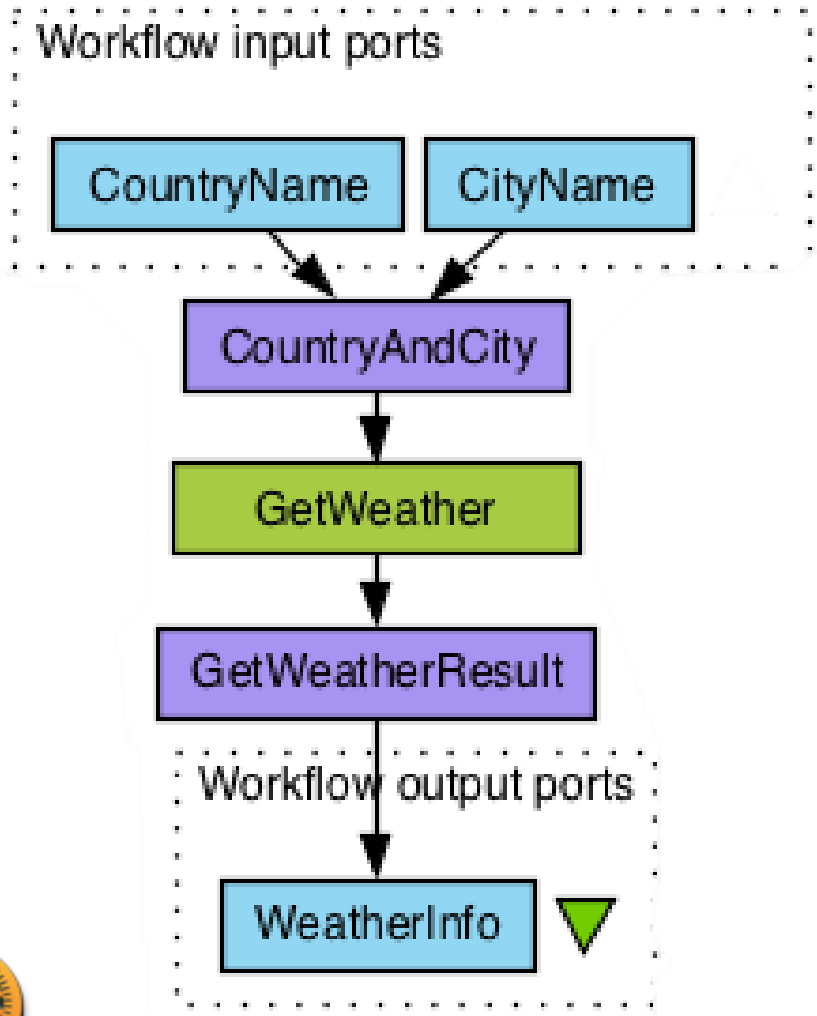
Workflow as a Solution

- Sophisticated analysis **pipeline**
- **Graphical** representation of executable analysis
- Combine a set of **services** to analyse or manage data (local or remote)
- Data **flow** from one service (boxes) to the next (connected with arrows)
- **Iteration** – process multiple data items
- Automation – **rerun** workflow



myGrid

Example Taverna Workflow



Workflow: Get the weather forecast of the day given the city and the country

Green box is a **Web Service**

Purple boxes are local **XML** services to assemble/ extract XML

Blue boxes are workflow **input** and **output** ports

Arrows define the direction of **data flow**

Workflows as a solution

- **Flow of data** from one tool to the next is automatic – just connect inputs and outputs
- Incompatibilities overcome in the workflow with helper services (*shims*)
  - Allowing new tool combinations
- Workflow engine records parameter values and algorithms – **provenance**
- Workflows can include data **integration** and **visualization**
- **Iteration** over large data sets automatic – ideal for high throughput analysis (e.g. omics)

Reproducible Research

Preventing non-reproducible research

- An array of errors

http://www.economist.com/node/21528593

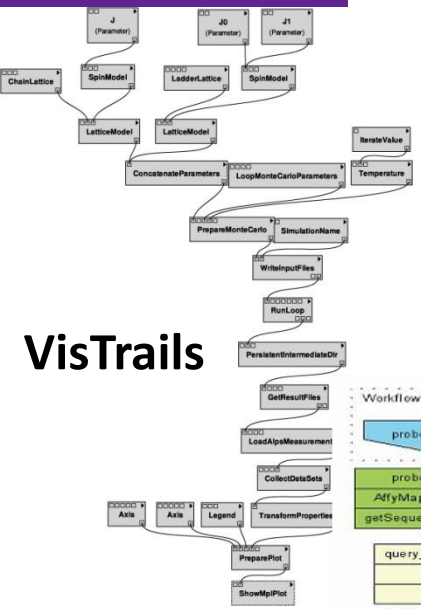
- Duke University, 2006 - Prediction of the course of a patient's lung cancer using expression arrays and recommendations on different chemotherapies from cell cultures – reported in *Nature Medicine*
- 3 different groups could not reproduce the results and uncovered mistakes in the original work

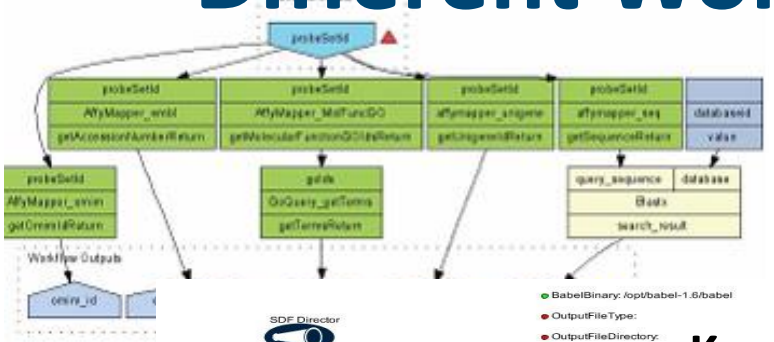
If the Analyses were done using Workflows.....

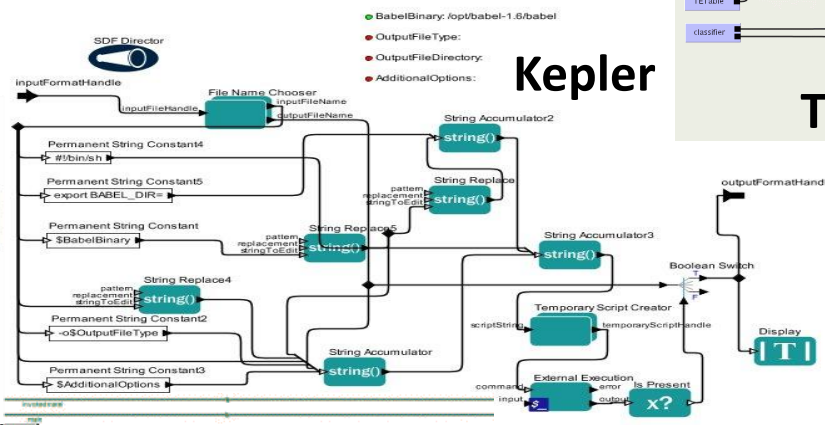
- Reviewers could re-run the *in-silico* experiments and see results for themselves
- Methods could be properly examined and criticized by inspecting the workflow
- Mistakes and opportunities could be pinpointed earlier

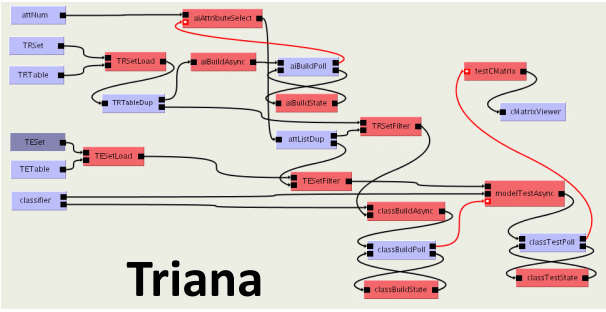
MANCHESTER 1824

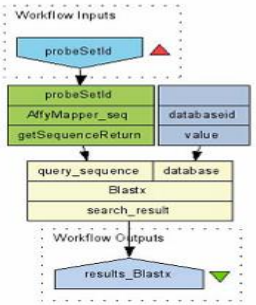
Different Workflow Systems

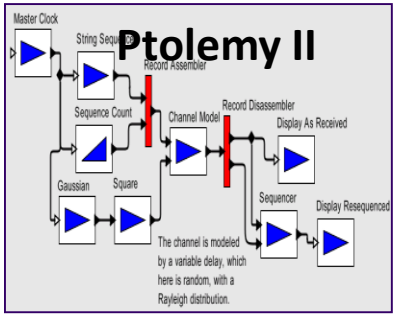


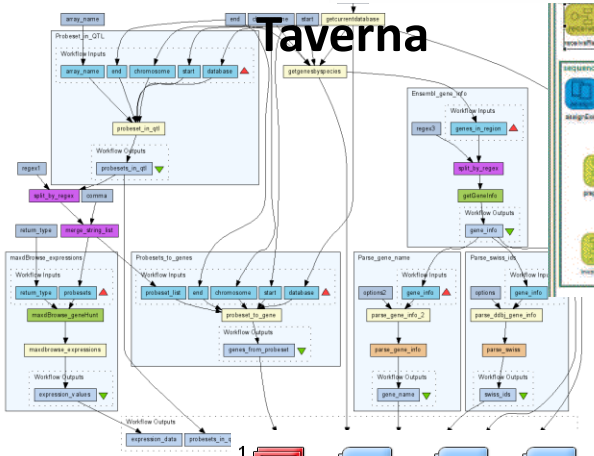


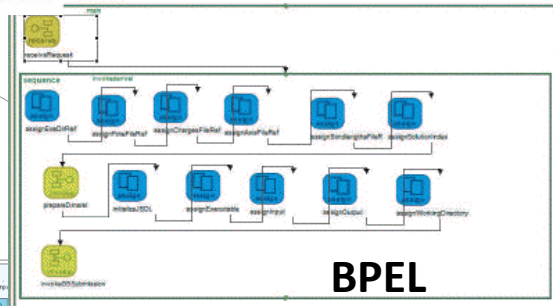


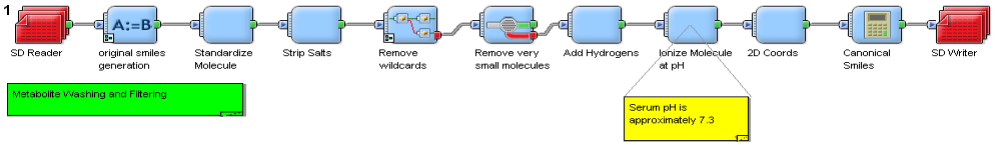


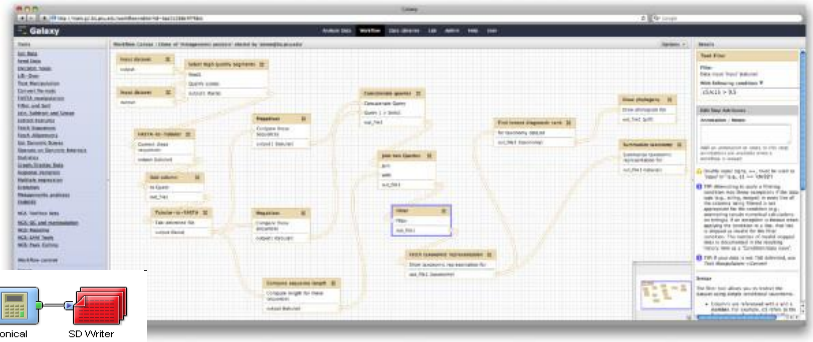












myGrid

MANCHESTER 1824
The University of Manchester

Taverna Workbench

http://www.taverna.org.uk/

Freely available, **open source**

80,000+ downloads across versions

Installers for Windows, Mac OS X, Linux

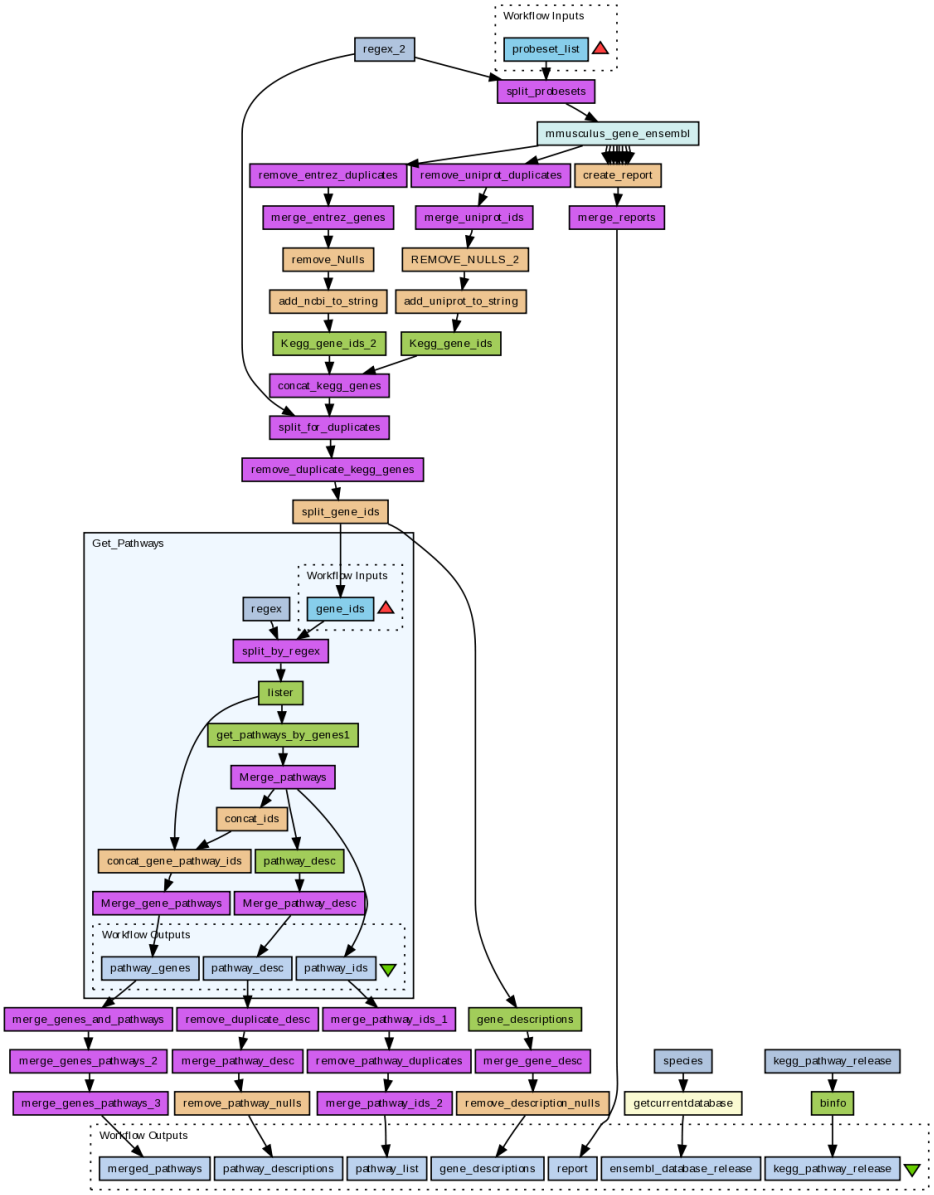
Current version: 2.5.0

Wolstencroft et al. (2013): **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud**", *Nucleic Acids Research*, **41**(W1): W557-W561. doi:10.1093/nar/gkt328

MANCHESTER 1824
The University of Manchester

Taverna Workflow System

http://www.taverna.org.uk



History:

- 2003: Taverna 0.1 (300 downloads)
- 2014: Taverna 2.5.0 (5100 downloads)

Products:

- Taverna Workbench
- Taverna Server
- Taverna Command line
- Taverna Online
- Taverna Player
- Plugins and integrations

myGrid

Taverna editions and extensibility

Taverna is a generic workflow system that can be extended by **plugins** and customized for use in different domains.

The Taverna **editions** are pre-built downloads of Taverna with plugins for the most popular domains.

- Core
- Astronomy
- **Bioinformatics**
- Biodiversity
- Digital Preservation
- Enterprise

http://www.taverna.org.uk/download/workbench/2-5/

MANCHESTER 1824
The University of Manchester

Taverna Workbench

Taverna Workbench 2.2.0

Design | Results | myExperiment

Service panel

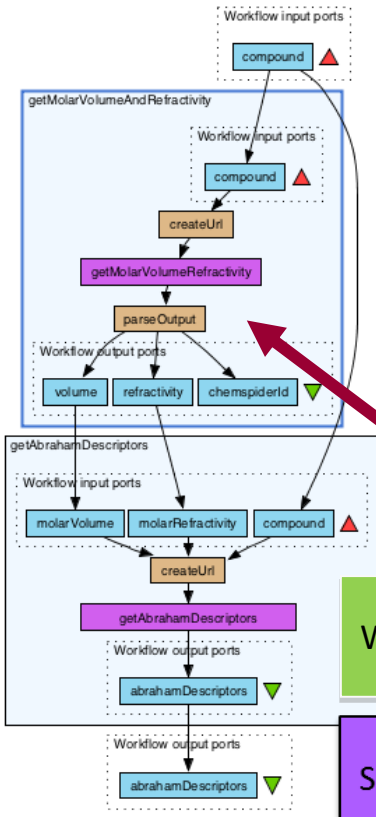
Clear

Available services
- Service templates
  - Beanshell – A service that allows Beanshell scripts, with dependenc
  - Nested workflow – A service that allows you to have one workflow
  - Rshell – A service that allows the calling of R scripts on an R server
  - SpreadsheetImport – A service that imports data from spreadshee
  - String constant – A string value that you can set

Workflow explorer | Details | Validation report

- Workflow9
  - Workflow input ports
    - compound
  - Workflow output ports
    - abrahamDescriptors
  - Services
    - getAbrahamDescriptors
      - compound
      - molarRefractivity
      - molarVolume
      - abrahamDescriptors
    - getMolarVolumeAndRefractivity
      - compound
      - chemspiderId
      - refractivity
      - volume
  - Data links
    - getMolarVolumeAndRefractivity:volume -> getAbrahamDescriptor
    - getMolarVolumeAndRefractivity:refractivity -> getAbrahamDescri

Workflow diagram



List of services

Workflow engine to run workflows

Construct and visualise workflows

Web Services — e.g. KEGG

Scripts — e.g. beanshell, R

Programming libraries — e.g. libSBML

MANCHESTER 1824
The University of Manchester

Using Tools and Services from Taverna workflows

- Web Services
  - WSDL
  - REST
- Data services
  - BioMart
- Local scripts:
  - R
  - Beanshell
  - Command line (e.g. Python, Perl)
- Other workflows
- And more..... Add your own!

myGrid

What are Web Services?

Web Services: HTTP-based programmatic access (API).

Instead of "*GET me the web page http://example.com/cat-pics*",

Web Services allow "*GET me a genome sequence http://example.com/gene/WAP_RAT*"

Connect to and use remote services from your computer in an automated way

NOT the same as services on the web (i.e. forms that shows results as a web page)

Who Provides the Services?

Open domain services and resources

- Taverna accesses thousands of services
- Third party – we don't own them – we didn't build them
- All the major providers
  - NCBI, DDBJ, EBI …
- Enforce NO common data model.

NCBI
National Center for Biotechnology Information (USA)

DDBJ DNA Data Bank of Japan
Tokyo, Japan

EMBL-EBI European Bioinformatics Institute
Cambridge, UK

biomart

SoapLab

wellcome trust sanger institute

KEGG Kyoto Encyclopedia of Genes and Genomes

MOUNT SINAI HOSPITAL

PDBjXP Protein Data Bank Japan Experimental Version

PathPort The Pathogen Portal Web Project

emboss

moby

SeqVISTA

LION

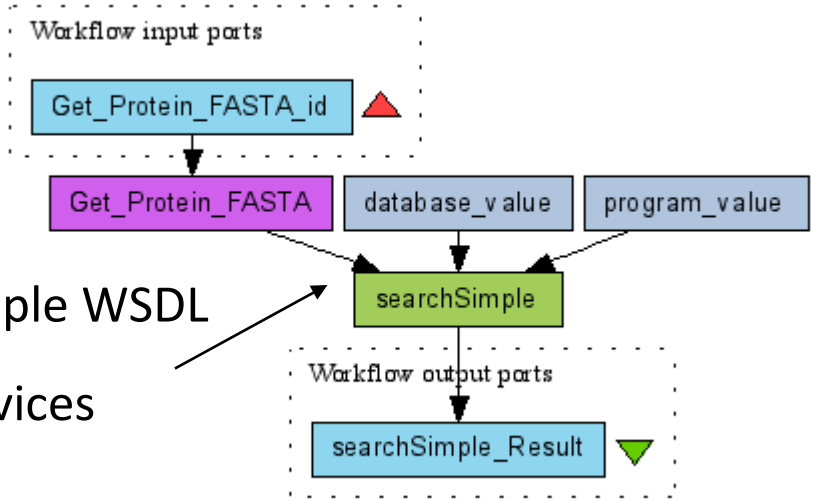
SRS

SeqHound

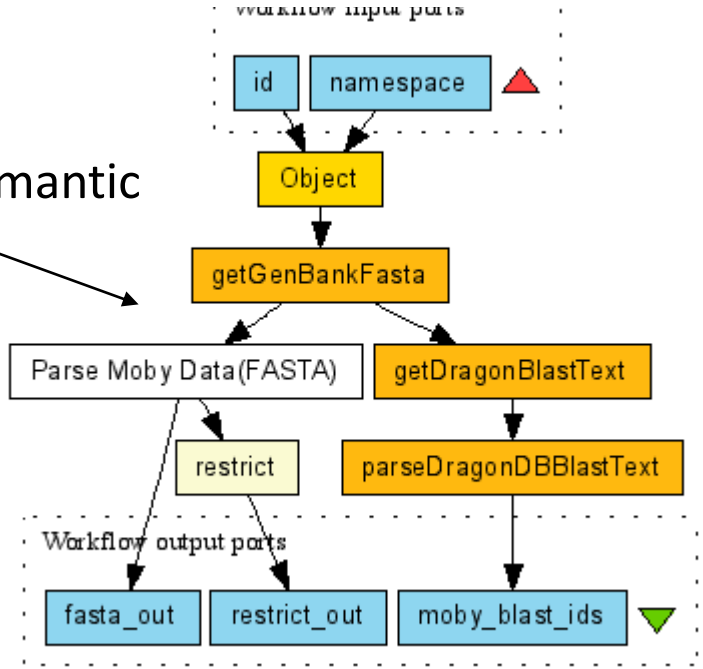
CABIO

MANCHESTER 1824
The University of Manchester

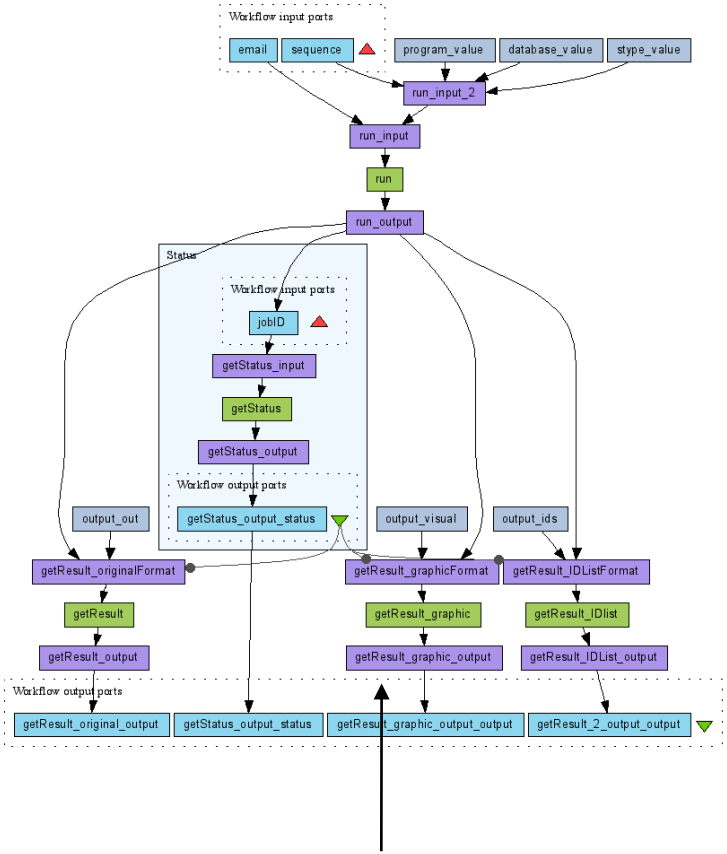
How do you use the services?



Simple WSDL services



BioMoby Semantic Services



Asynchronous services (Submit, Wait, Fetch)

myGrid

BioCatalogue beta
"The Life Science Web Service Registry"

Getting Started | About Us | Contact Us | API Docs

Sign up | Sign in

Search: Go! | Home | Services | Register a Service | Providers

Home »

SHARE

The BioCatalogue: providing a curated catalogue of Life Science Web Services

Latest Activity

Last 7 days

- vasunun **joined** the BioCatalogue
- Peter Taschner **added** a publication annotation to the Soap Service of Service: MutalyzerService
- Peter Taschner **added** a contact annotation to the Service Deployment of Service: MutalyzerService
- Peter Taschner **added** a tag annotation to Service: MutalyzerService
- Peter Taschner **added** a tag annotation to Service: MutalyzerService
- Peter Taschner **added** a tag annotation to Service: MutalyzerService
- Peter Taschner **added** a alternative name annotation to Service: MutalyzerService
- Peter Taschner **added** a documentation url annotation to the Soap Service of Service: MutalyzerService
- Peter Taschner **added** a description annotation to the Soap Service of Service:

The BioCatalogue currently has 1730 services, 130 service providers and 445 members

"Web Services are hard to find"

DISCOVER

- Find the right Web Service
- Powerful search and filtering
- Information from providers and community

More info

"My Web Services are not visible"

REGISTER

- Easily register Web Services
- Instantly available to everyone
- Providers can advertise, describe and monitor their Services

More info

"Web Services are poorly described"

ANNOTATE

- Anyone can describe and annotate
- Ongoing expert curation
- Social curation by the community

More info

"Web Services are volatile"

MONITOR

- Services change and get outdated
- BioCatalogue monitors Services
- Monitors availability and reliability

More info

Our Partners

EMBRACE Grid — Network of excellence | myexperiment | myGrid | seekda! | moby

The EMBRACE Registry and the BioCatalogue have now been merged

Site Announcements

Have your say about BioCatalogue by taking part in the BioCatalogue users's survey
By Franck Tanoh (4 days ago

BioCatalogue Maintenance - 7 December 2010 @ 9:30 am (GMT)
By Eric E. Nzuobontane (6 days ago

BioCatalogue iPhone and iPad app now available to download for free
By Franck Tanoh (2 months ago

The BioCatalogue Functional Unit paper presented at IEEE 2010 Fourth International Workshop on Scientific Workflows
By Franck Tanoh (2 months ago

The National Cancer Research Institute (NCRI) joins forces with the BioCatalogue
By Franck Tanoh (4 months ago

More

Latest Services

- MutalyzerService
- dbfetch
- graphtools
- PRANK (REST)
- EASIM (REST)

BioCatalogue.org - Services - InterProScan (REST) (REST) [Provided by: European Bioinformatics Institute (EBI)] - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.biocatalogue.org/services/2769

Most Visited | Getting Started | Latest Headlines | Post to CiteULike | Using Google Maps in ...

BioCatalogue.org - Services - InterPr...

BioCatalogue beta
"The Life Science Web Service Registry"

Getting Started | About Us | Contact Us | API Docs

Sign up | Sign in

Search: Go! | Home | Services | Register a Service | Providers

Home » Services » InterProScan (REST)

SHARE

InterProScan (REST) REST

99 0

aka JDispatcher aka InterProScan

Categories: Function Prediction

Monitoring

Annotations: 30 0 30 0

Overview | REST Endpoints (6) | Monitoring | News

Provider:
European Bioinformatics Institute (EBI)

Provider

Location:
UNITED KINGDOM

Submitter

Submitter / Source:
Hamish McWilliam (about 1 month ago)

Service Description

Base URL:
http://www.ebi.ac.uk/Tools/services/rest/iprscan

Tags

Tags (19)

bioinformatics | ebi | embl-ebi | Gene3D | HAMAP | interpro | interproscan | Panther | Pfam | PIRSF | PRINTS | ProDom | ProSite | protein domain | protein family | protein function | SMART | SUPERFAMILY | TIGRFAMs

Login to add tags

Favourited By (0)

Done

MANCHESTER 1824
The University of Manchester

What do Scientists use Taverna for?

Systems biology model building

Sequence analysis Protein structure prediction

Gene/protein annotation Microarray data analysis

Phylogeny Model simulations sweeps

High throughput screening Proteomics

Phenotypical studies Text mining

Public Health care epidemiology

Medical image analysis QTL studies

QSAR studies Genome Wide Association Studies

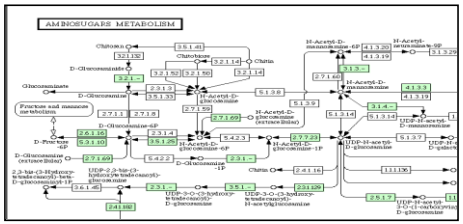
Astronomy

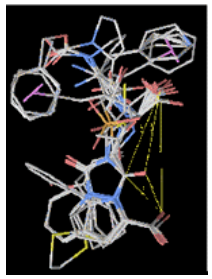
Music

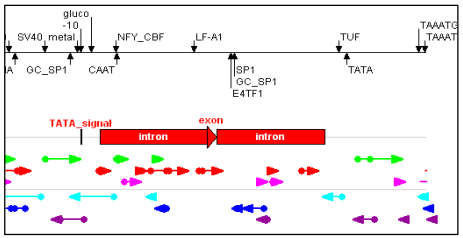
Meteorology

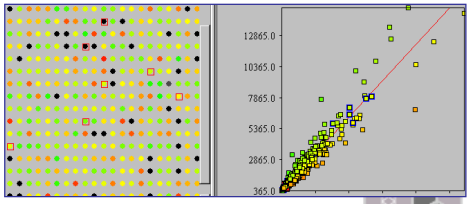
Social Science

Cheminformatics









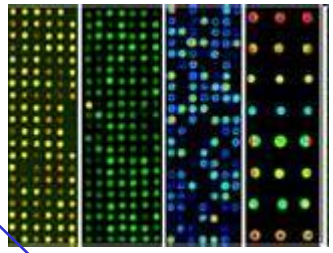
myGrid

MANCHESTER 1824
The University of Manchester

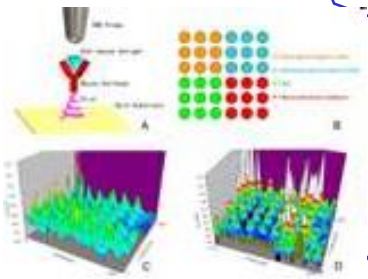
Research Example

Lymphoma Prediction Workflow

MicroArray from tumor tissue

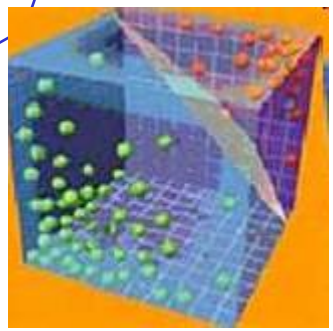


caArray



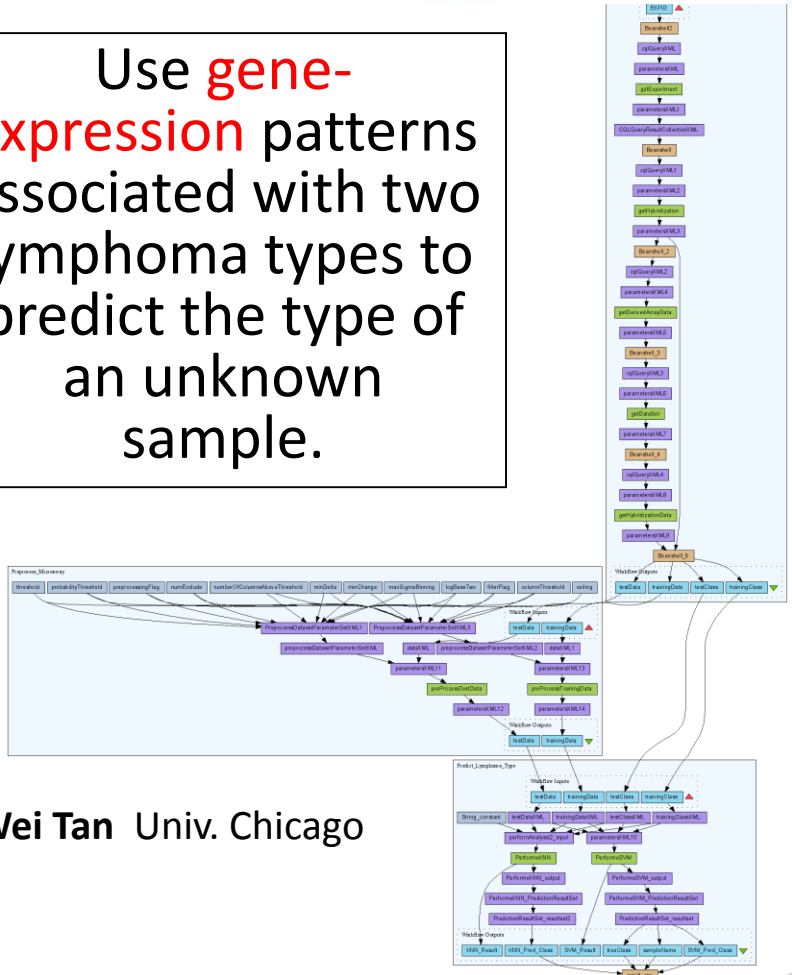
Microarray preprocessing

Lymphoma prediction



GenePattern

Use gene-expression patterns associated with two lymphoma types to predict the type of an unknown sample.

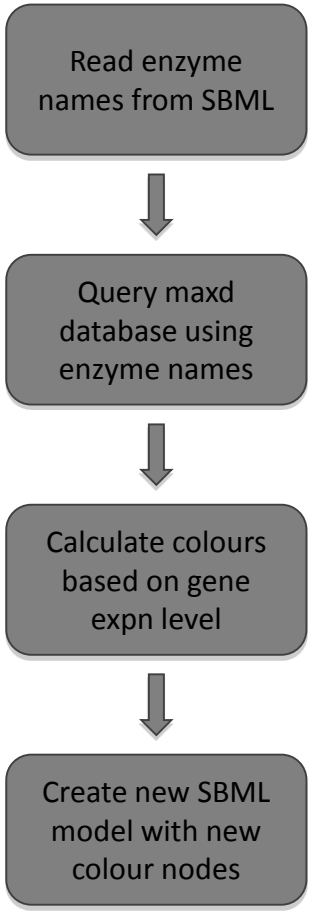


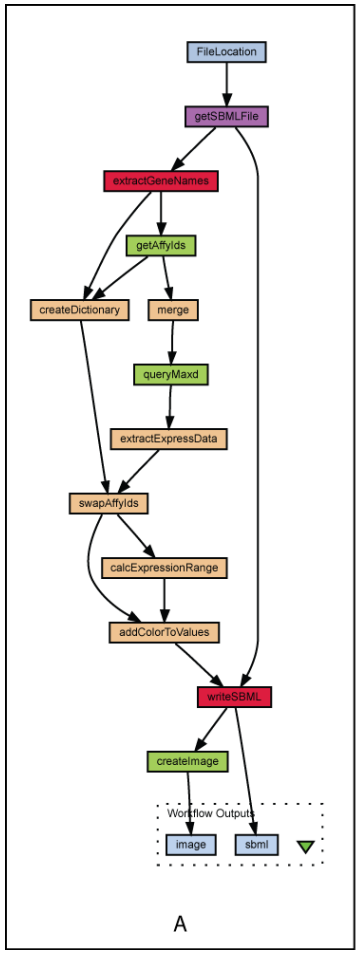
Wei Tan Univ. Chicago

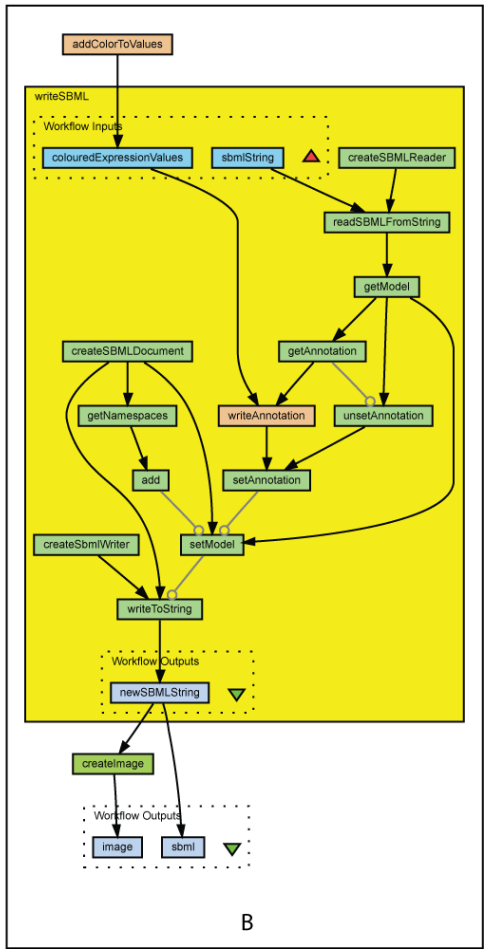
Ack. Juli Klemm, Xiaopeng Bian, Rashmi Srinivasa (**NCI**)
Jared Nedzel (**MIT**)

MANCHESTER 1824
The University of Manchester

Systems Biology Data Integration







Mapping transcriptomics data onto SBML models

Peter Li, Doug Kell, U Manchester

myGrid

MANCHESTER 1824
The University of Manchester

Workflows are …

… records and protocols (i.e. your *in silico* experimental method)

… know-how and intellectual property

… hard work to develop and get right

…..re-usable methods (i.e. you can build on the work of others)

So why not share and re-use them

my experiment

myGrid

MANCHESTER 1824
The University of Manchester

Workflow Repository

my experiment

About | Mailing List | Publications

Logout | Give us Feedback | Invite

Home | Users | Groups | Workflows | Files | Packs | Services | Topics

Workflows | Search

Home » Workflows

BOOKMARK

Workflows

Search filter terms

« previous | 1 | 2 | 3 | ... | 191 | next »

Sort by: Rank

Showing 1905 results. Use the filters on the left and the search box below to refine the results.

Search

Filter by type

Taverna 2	863
Taverna 1	645
RapidMiner	171
Kepler	43
Bioclipse Scri...	34
GWorkflowDL	24
LONI Pipeline	22
BioExtract Server	16
Trident (Packa...	10
LabTrove Tem...	9

Filter by tag

Taverna 2

Pathways and Gene annotations for QTL region (v7)

View

Download (v7)

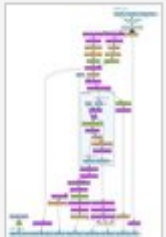
Original Uploader

Paul Fisher

Created: 19/11/09 @ 18:18:52 | **Last updated:** 02/09/11 @ 11:44:57

Credits: Paul Fisher

License: Creative Commons Attribution-Share Alike 3.0 Unported License



This workflow searches for genes which reside in a QTL (Quantitative Trait Loci) region in the mouse, Mus musculus. The workflow requires an input of: a chromosome name or number; a QTL start base pair position; QTL end base pair position. Data is then extracted from BioMart to annotate each of the genes found in this region. The Entrez and UniProt

New/Upload

Workflow | GO

Katy Wolstencroft

My Profile [edit]
My Messages
My Memberships (5)
My History
My News

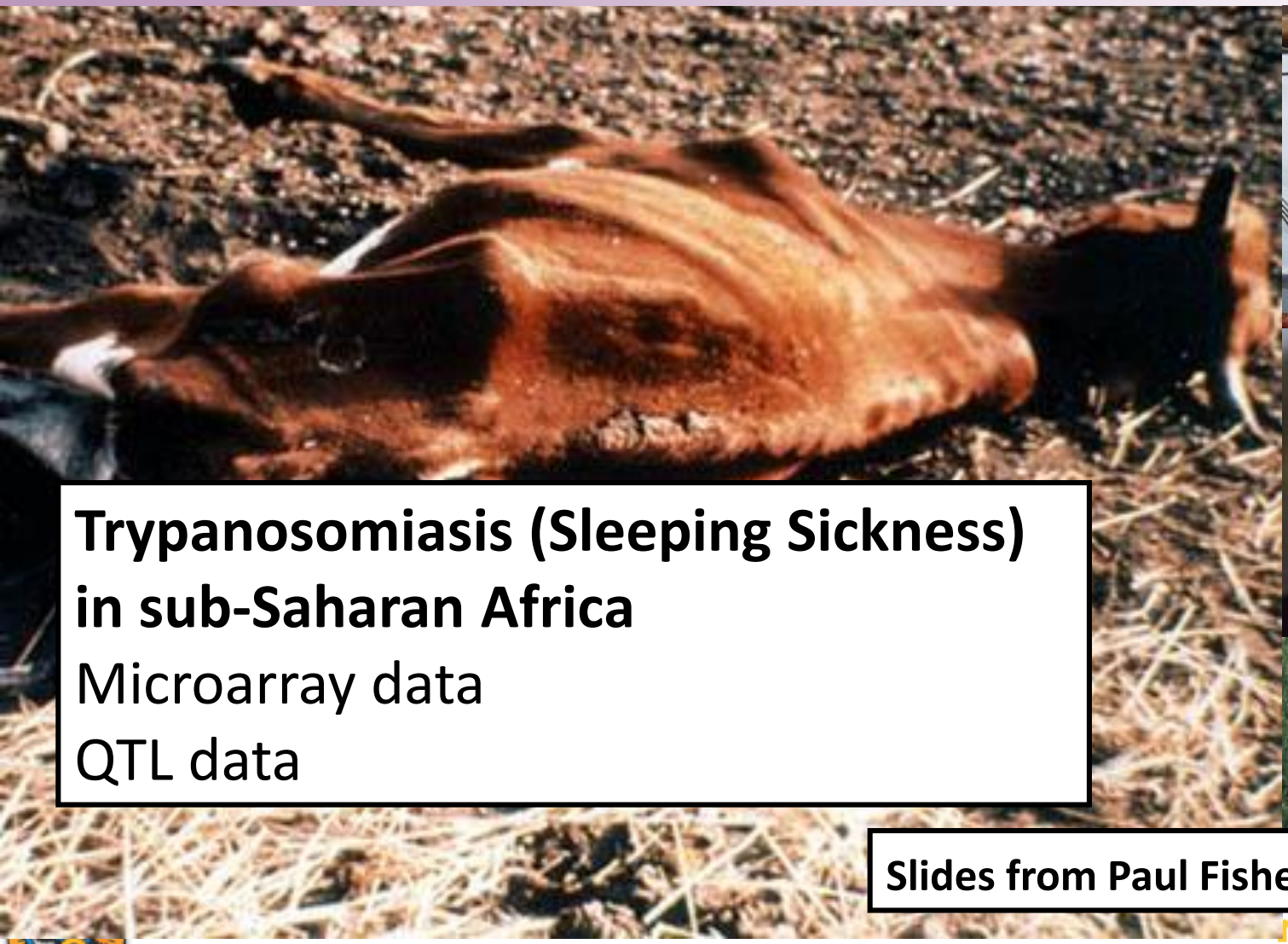
4 new friendship requests

Onlyhakanboz
mihaiionita_me
Pankaj chauhan
Hanny

Just Enough Sharing….

- myExperiment can provide a central location for workflows from one community/group
- myExperiment allows you to say
  - Who can look at your workflow
  - Who can download your workflow
  - Who can modify your workflow
  - Who can run your workflow
- Ownership and attribution

The Wellcome Trust Funded Host-Pathogen Project



Steve Kemp

Andy Brass

Paul Fisher

Trypanosomiasis (Sleeping Sickness) in sub-Saharan Africa

Microarray data

QTL data

Slides from Paul Fisher

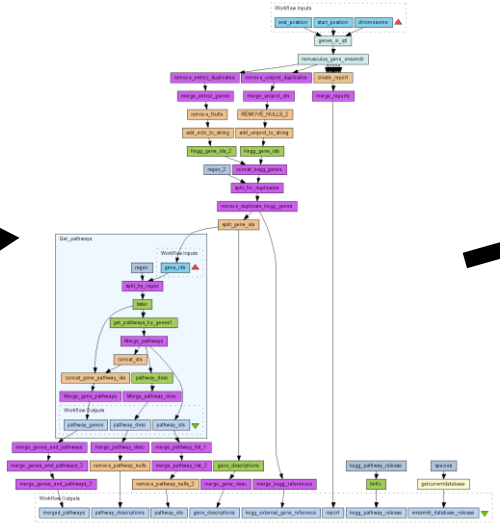
http://www.genomics.liv.ac.uk/tryps/

MANCHESTER 1824
The University of Manchester

Reuse, Recycle, Repurpose Workflows

Dr Paul Fisher

Identify *QTg and pathways* implicated in resistance to **Trypanosomiasis** in **cattle**



Dr Jo Pennock

Identify the QTg and pathways of **colitis** and **helminth** infections in the **mouse** model

PubMed ID: 20687192

doi:10.1186/1471-2164-14-127

Another Host, Another Parasite...but the SAME Method

- Mouse whipworm infection - parasite model of the human parasite - *Trichuris trichuria*

Understanding Phenotype

- Comparing resistant vs susceptible strains – Microarrays

Understanding Genotype

- Mapping quantitative traits – Classical genetics QTL

Joanne Pennock, Richard Grencis
University of Manchester

Workflow Results

- Identified the biological pathways involved in sex dependence in the mouse model, previously believed to be involved in the ability of mice to expel the parasite.
- Manual experimentation: Two year study of candidate genes, processes unidentified
- Workflow experimentation: Two weeks study – identified candidate genes

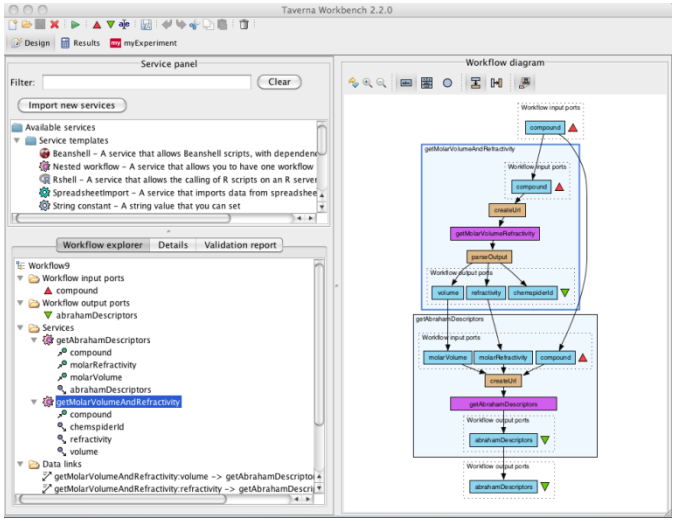
Joanne Pennock, Richard Grencis
University of Manchester

Workflow Success

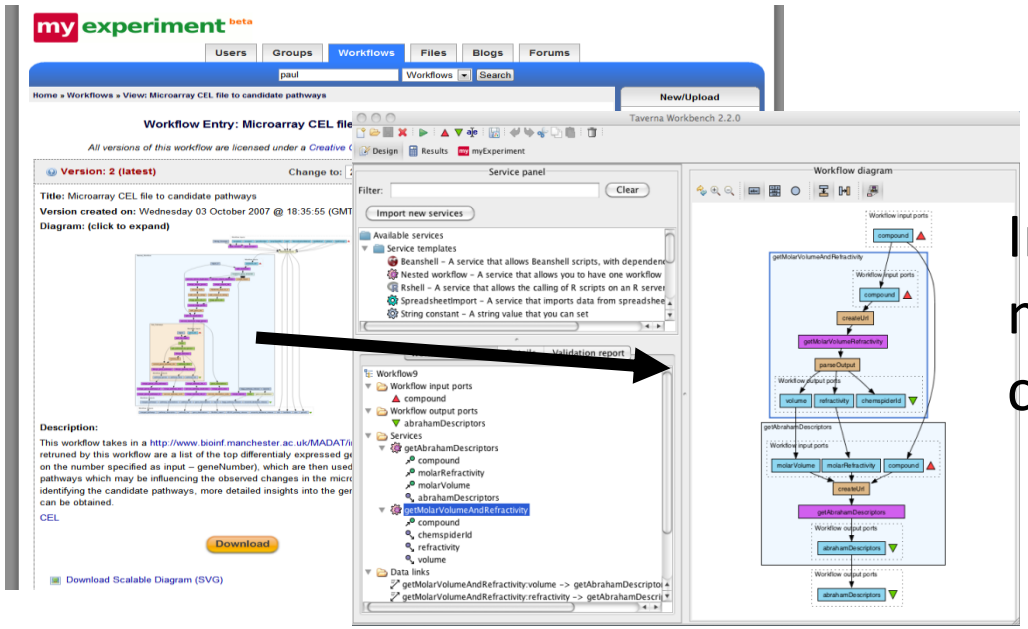
- Workflow analysed each piece of data *systematically*
  - Eliminated user bias and premature filtering of datasets
- The size of the QTL and amount of the microarray data made a manual approach impractical
- Workflows capture exactly where data came from and how it was analysed
- Workflow output produced a manageable amount of data for the biologists to interpret and verify
  - "make sense of this data" -> "does this make sense?"

MANCHESTER 1824
The University of Manchester

Spectrum of Users



Advanced users design and build workflows (informaticians)



Intermediate users reuse and modify existing workflows or components

Others “replay” workflows through web page

myGrid

MANCHESTER 1824
The University of Manchester

A Collection of Tools

Workflow Repository

myexperiment

Service Catalogue

BioCatalogue
"The Life Science Web Service Registry"

Activity and Service Plug-in Manager

SoapLab
KnowARC
biomart
moby
BeanShell
R

Workflow GUI Workbench and 3rd party plug-ins

Taverna

Provenance Store

W3C PROV

Workflow Server

Secure Service Access, and Programming APIs

Client User Interfaces

SEEK

Web Portals

E-Laboratories

Programming and APIs

Bioclipse
OSGi Alliance

myGrid

Summary – Workflow Advantages

- Informatics often relies on **data integration** and large-scale **data analysis**
- Workflows are a mechanism for **linking** together resources and analyses
- Promote **reproducible** research
- Find and use successful analysis methods *developed by others* with myExperiment

More Information

- Taverna
  - http://www.taverna.org.uk
- myExperiment
  - http://www.myexperiment.org
- BioCatalogue
  - http://www.biocatalogue.org

Tutorials

- Using Taverna to design and build workflows
- Reusing workflows from myExperiment
- Finding and using different services:

  REST, Xpath, Beanshell, R, …
- Exploring the workflow engine: iteration, looping, retries, parallel invocation
- Web: Taverna Online, Taverna Player
- Interactions
- Components