Classical and myGrid approaches to data mining in bioinformatics

Peter Li

School of Computing Science
University of Newcastle upon Tyne

myGrid

Outline

- Real life bioinformatics use cases
  - Graves' disease
  - Williams-Beuren syndrome
- Classical approach to bioinformatics data analysis
- Bioinformatics workflows
- Using myGrid workflows for data analysis
- Issues for further work

Application scenario[1]

Graves' disease

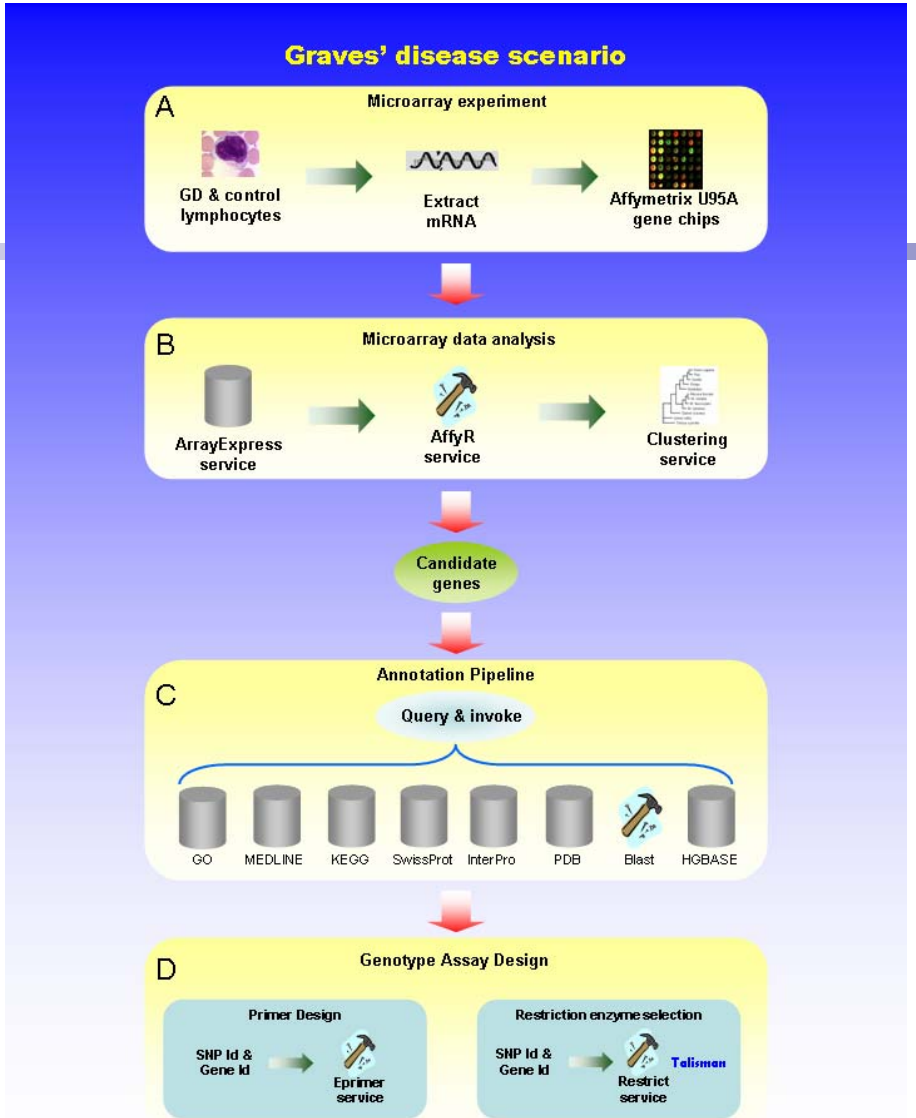
- Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle

Graves' disease

- Autoimmune thyroid disease
- Lymphocytes attack thyroid gland cells causing hyperthyroidism
- An inherited disorder
- Complex genetic basis
- Symptoms:
  - Increased pulse rate, sweating, heat intolerance
  - goitre, exophthalmos

In silico experiments in Graves' disease

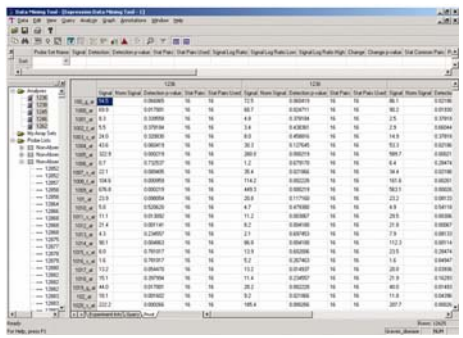
- Identification of genes:
  - Microarray data analysis
  - Gene annotation pipeline
  - Design of genotype assays for SNP variations in genes
- Distributed bioinformatics services from Japan, Hong Kong, various sites in UK
- Different data types: textual, image, gene expression, etc.



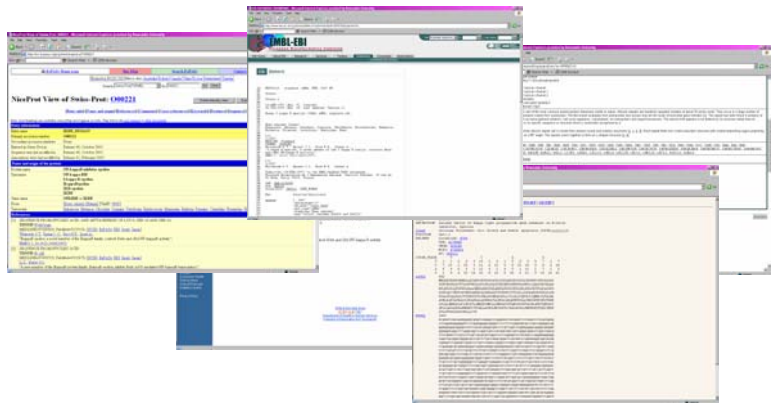
Classical approach to the bioinformatics of Graves' disease

Data Analysis - Microarray

Import microarray data to Affymetrix data mining tool, run analyses and select gene



Study annotations for many different genes Using web html based resources



Select gene and visually examine SNPS lying within gene

Experiment design to test hypotheses

Find restriction sites and design primers by eye for genotyping experiments

myGrid

Application scenario^2

Williams-Beuren Syndrome

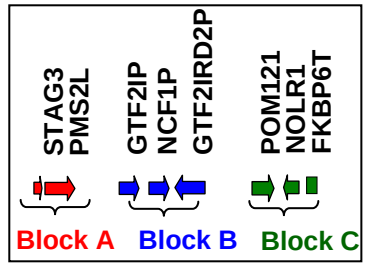
- Hannah Tipney, May Tassabehji, St Mary's Hospital, Manchester, UK
- Gene prediction; gene and protein annotation
- Services from USA, Japan, various sites in UK

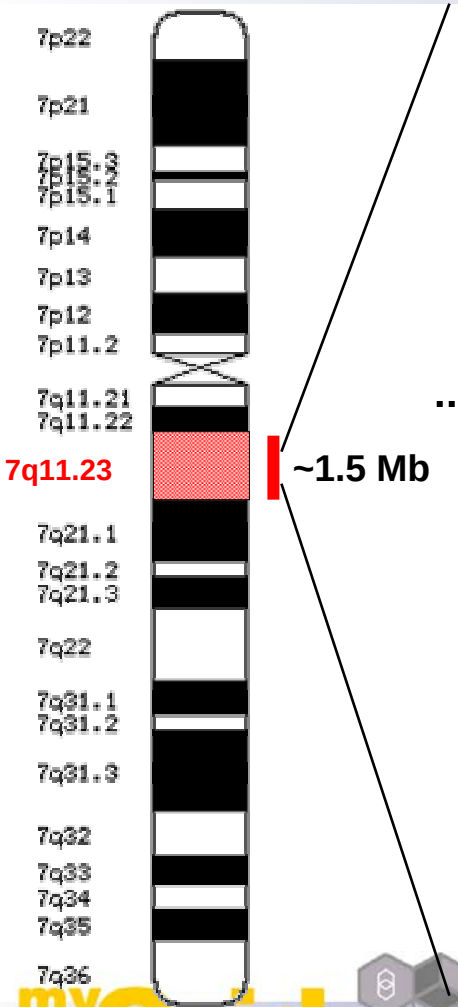
Williams-Beuren Syndrome (WBS)

- Contiguous sporadic gene deletion disorder
- 1/20,000 live births, caused by unequal crossover (homologous recombination) during meiosis
- Haploinsufficiency of the region results in the phenotype
- Multisystem phenotype – muscular, nervous, circulatory systems
- Characteristic facial features
- Unique cognitive profile
- Mental retardation (IQ 40-100, mean~60, ‘normal’ mean ~ 100)
- Outgoing personality, friendly nature, ‘charming’

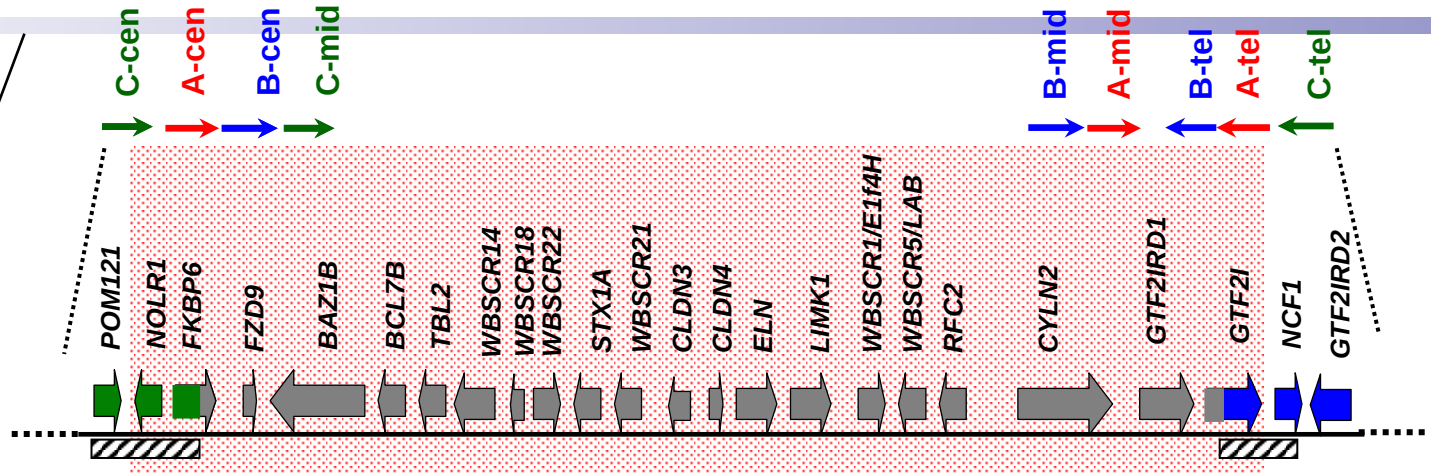
Williams-Beuren Syndrome Microdeletion

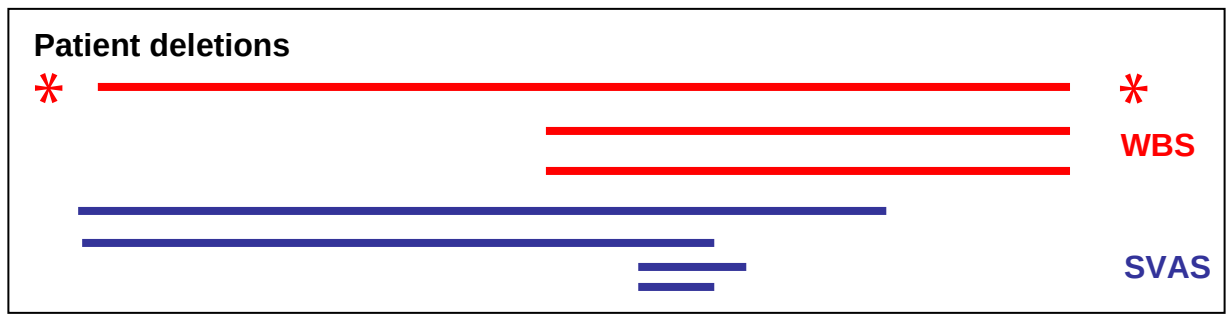
Eicher E, Clark R & She, X An Assessment of the Sequence Gaps: Unfinished Business in a Finished Human Genome. Nature Genetics Reviews (2004) 5:345-354
Hillier L et al. The DNA Sequence of Human Chromosome 7. Nature (2003) 424:157-164

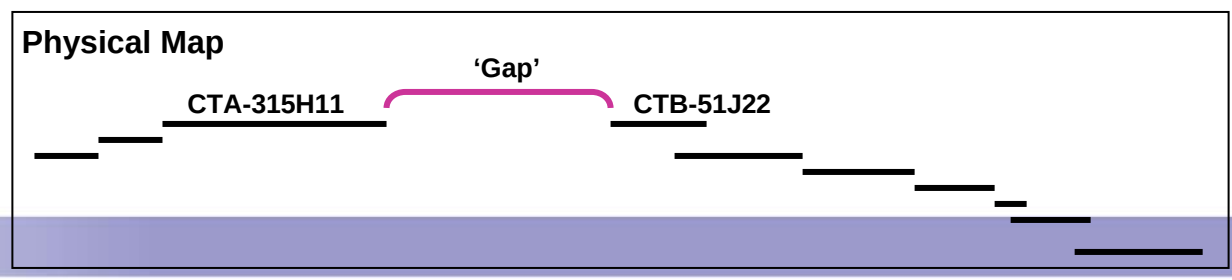




~1.5 Mb

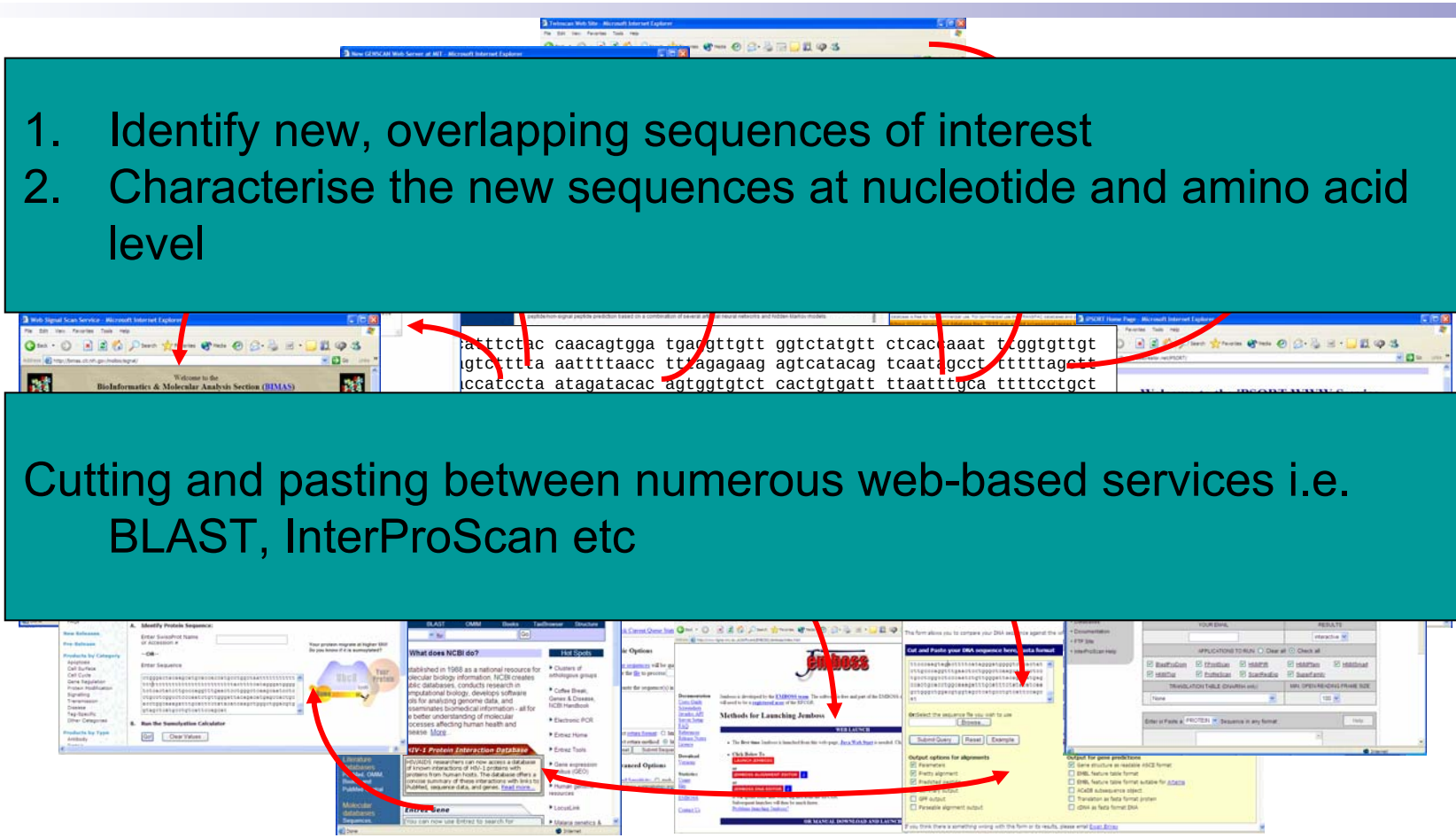






Filling a genomic gap *in Silico*

1. Identify new, overlapping sequences of interest
2. Characterise the new sequences at nucleotide and amino acid level



Cutting and pasting between numerous web-based services i.e. BLAST, InterProScan etc

myGrid

Classical approach

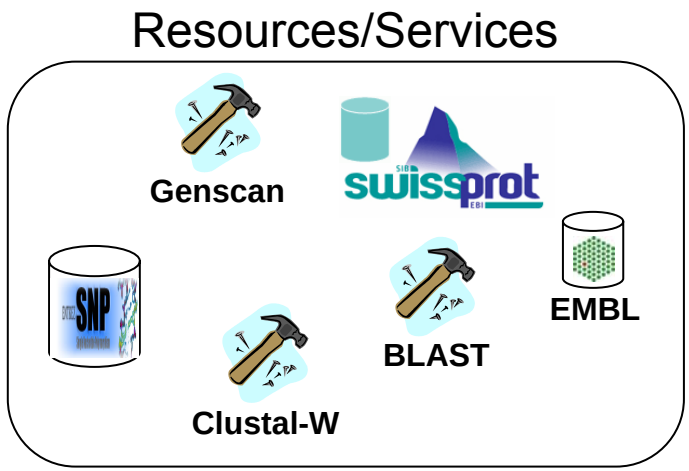
- Frequently repeated - info rapidly added to public databases
- Time consuming and mundane
- Don't always get results
- Huge amount of interrelated data is produced – handled in notebooks and files saved to local hard drive
- Much knowledge remains undocumented:
  Bioinformatician does the analysis

✓ **Advantages**:
Specialist human intervention at every step, quick and easy access to distributed services

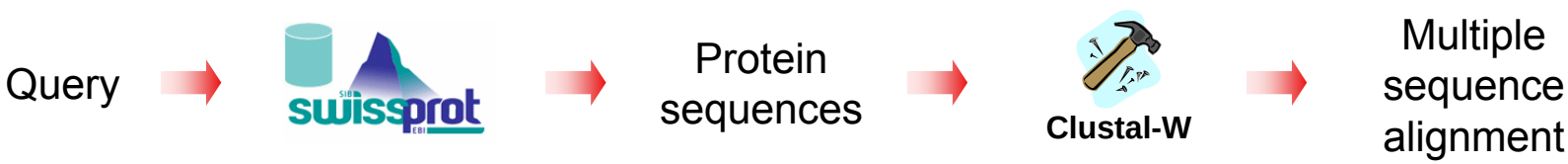
× **Disadvantages**:
Labour intensive, time consuming, highly repetitive and error prone process, tacit procedure so difficult to share both protocol and results

In silico experiments in bioinformatics

Bioinformatics analyses - *in silico* experiments - workflows



Example workflow: Investigate the evolutionary relationships between proteins



myGrid

Why workflows and services?

Workflow = general technique for describing and enacting a process
Workflow = describes *what* you want to do, not *how* you want to do it
Web Service = *how* you want to do it
Web Service = automated programmatic internet access to applications

- Automation
  - Capturing processes in an explicit manner
  - Tedium! Computers don't get bored/distracted/hungry/impatient!
  - Saves repeated time and effort
- Modification, maintenance, substitution and personalisation
- Easy to share, explain, relocate, reuse and build
- Available to wider audience: don't need to be a coder, just need to know how to do Bioinformatics
- Releases Scientists/Bioinformaticians to do other work
- Record
  - Provenance: what the data is like, where it came from, its quality
  - Management of data (LSID - Life Science IDentifiers)

myGrid

- EPSRC e-Science pilot research project
- Manchester, Newcastle, Sheffield, Southampton, Nottingham, EBI and industrial partners.
- 'Targeted to develop open source software to support personalised *in silico* experiments in biology on a Grid.'

Which means enabling scientists to….

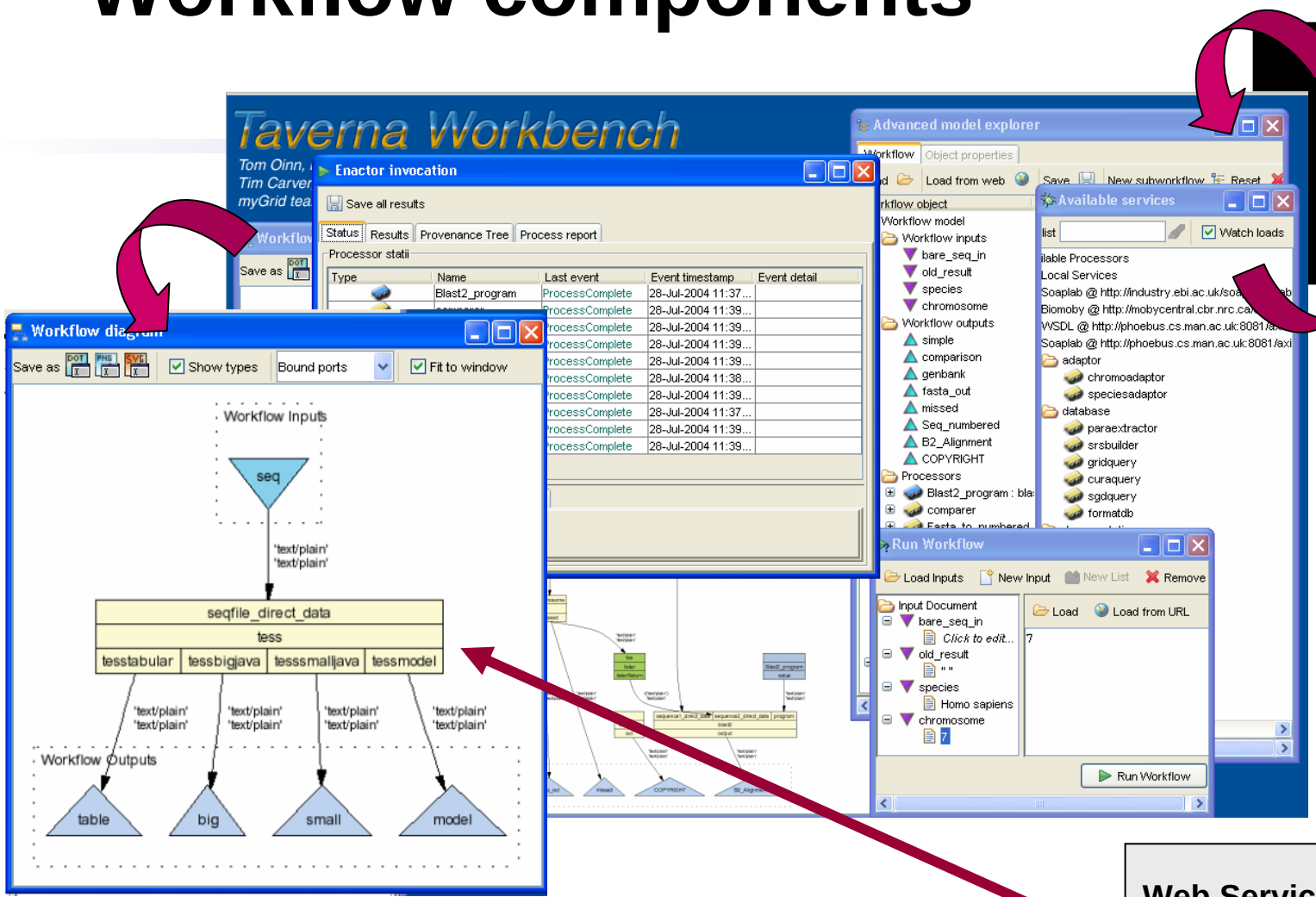
Distributed computing – machines, tools, databanks, people

Provenance and data management

Workflow enactment and notification

A virtual lab 'workbench', a toolkit which serves life science communities.

Workflow components



Freefluo
Workflow engine to run workflows

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

Web Service	**e.g. DDBJ BLAST**
SOAPLAB Web Service	**Any Application**

The workflow experience

Have workflows delivered on their promise? **YES!**

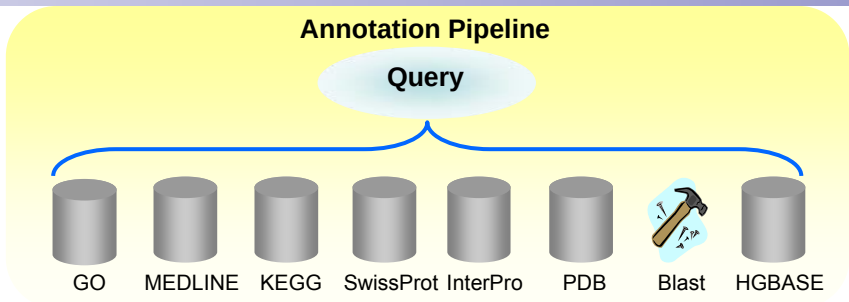
- Correct and biologically meaningful results
- Automation
  - Saved time, increased productivity
  - But you still require humans!
- Sharing
  - Other people have used and want to develop the workflows
- Change of work practises
  - *Post hoc* analysis. Don't analyse data piece by piece receive all data all at once
  - Data stored and collected in a more standardised manner
  - Results management

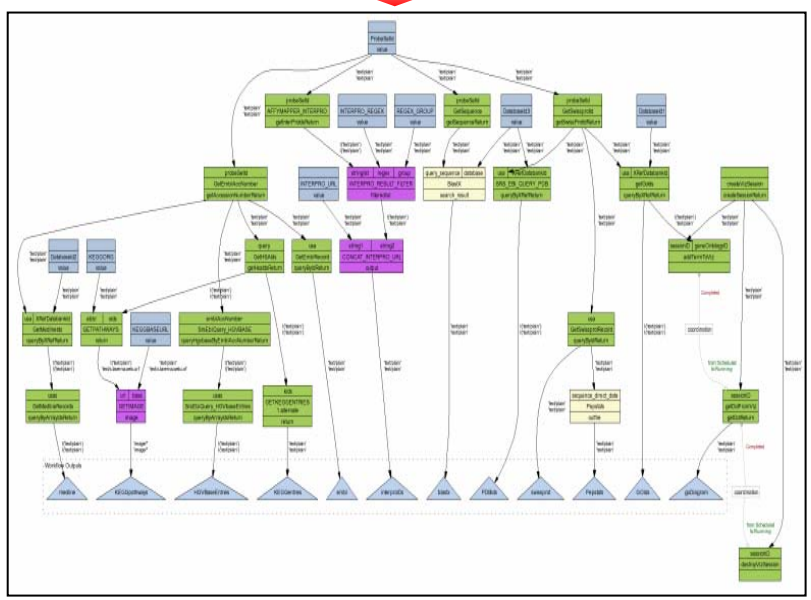
The workflow experience

- Activation Energy versus Reusability trade-off
  - Lack of ‘available’ services, levels of redundancy can be limited
  - But once available can be reused for the greater good of the community
- Instability of external bioinformatics web services
  - Research level
  - Reliant on other peoples servers
  - Taverna can retry or substitute before graceful failure
- Need Shim services in workflows

Modelling *in silico* experiments as workflows requires Shims

- Unrecorded ‘steps’ which aren’t realised until attempting to build something
- Enable services to fit together
- Semantic, syntactic and format typing of data in workflow
- Data has to be filtered, transformed, parsed for consumption by services





myGrid

Shims

‘I want to identify new sequences which overlap with my query sequence and determine if they are useful’

Sequence i.e. last known 3000bp

Mask

BLAST

Simplify and Compare

Retrieve

Sequence database entry
Fasta format sequence
Genbank format sequence

Identify new sequences and determine their degree of identity

Lister

Old BLAST result

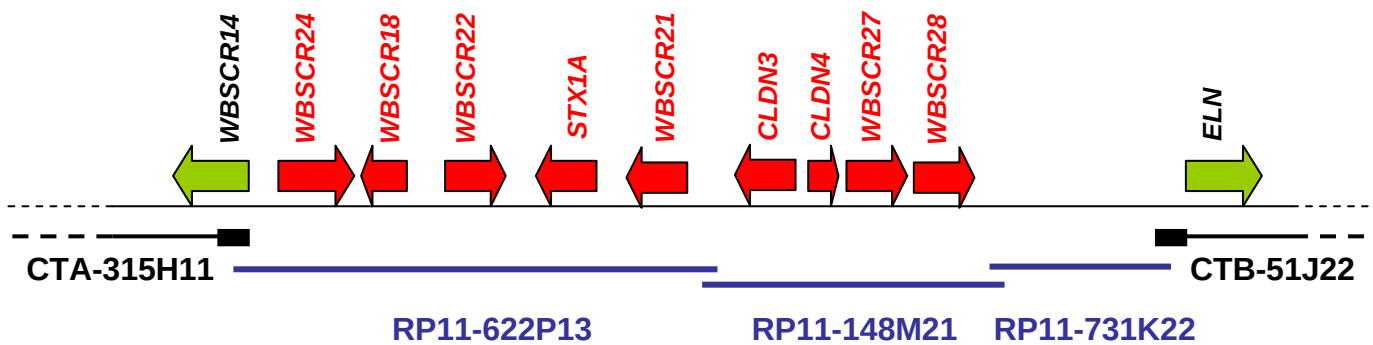
BLAST2

Alignment of full query sequence V full ‘new’ sequence

myGrid

Biological results from WB syndrome

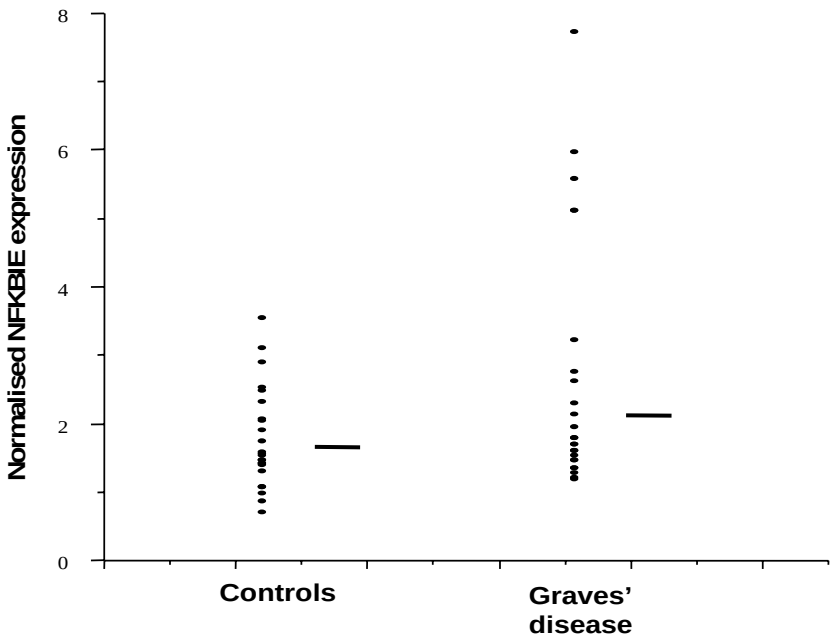
Four workflow cycles totalling ~ 10 hours
The gap was correctly closed and all known features identified



314,004bp extension

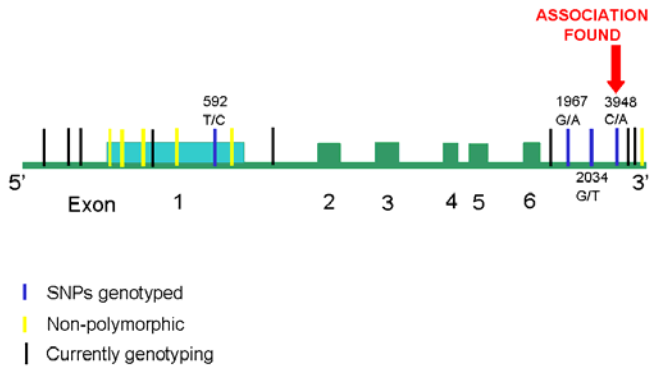
All nine known genes identified
(40/45 exons identified)

GD results: Differential expression and variations of the I kappa B-epsilon gene



Mean NFKBIE expression levels -

- Controls: 1.60 +/- 0.11 (SEM)
- GD: 2.22 +/- 0.20 (SEM)
- P=0.0047 (T-test)



3' UTR SNP – 3948 C/A

	Controls n=922	GD n=796
C allele	724 (78.5%)	575 (72.2%)
A allele	198 (21.5%)	221 (27.8%)

- MnI restriction site
- $\chi^2 = 9.1$, $p = 0.0025$, Odds Ratio = 1.4

myGrid

Conclusions

- It works – a new tool has been developed which is being utilised by biologists
- More regularly undertaken, less mundane, less error prone
- More systematic collection and analysis of results
- Increased productivity
- Services: only as good as the individual services, lots of them, we don't own them, many are unique and at a single site, research level software, reliant on other peoples services
- Activation energy

Issues and future directions[1]

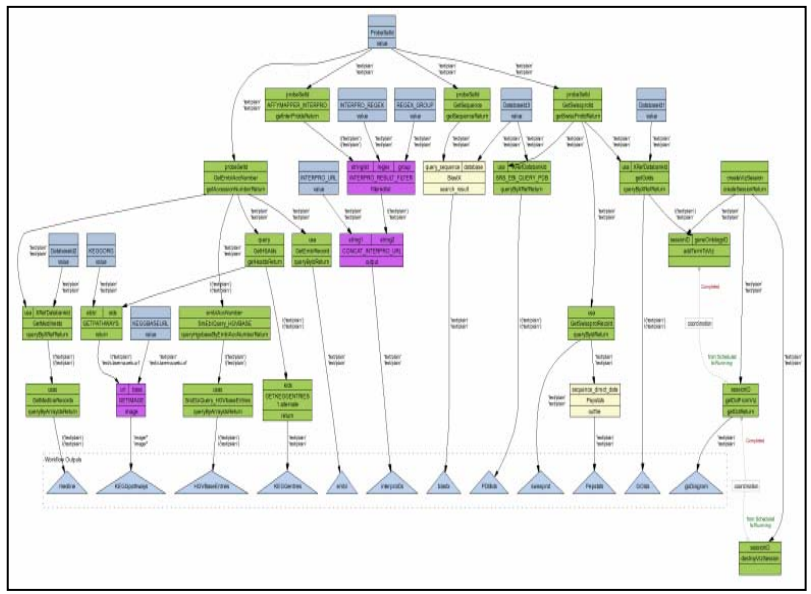
- Transfer of large data sets between services (microarray data)
  - Passing data by value breaks Web services
  - Streaming (Inferno)
  - Pass by reference and use third party data transfer (GridFTP, LSID)

Issues and future directions[2]

- Data visualisation
  - How to visualise results mined from data using workflows?

Workflow results

- Large amounts of information (or datatypes)
- Results are implicitly linked within itself
- Results are implicitly linked outside of itself
- Genomic sequence is central co-ordinating point, but there are a number of different co-ordinate systems
- Need holistic view



myGrid

What's the problem?

- No domain model in myGrid
- We need a model for visualisation
- But domain models are hard
- It's not clear that the domain model should be in the middleware

What have we done!?

- Bioinformatics PM (pre myGrid)
- One big distributed data heterogeneity and integration problem

What have we done!?

- Bioinformatics PM (post myGrid)
- One big data heterogeneity and integration problem

Initial Solutions

- Take the data
- Use something (Perl script or an MSc student) to map the data into a (partial) data model
- Visualise results which are linked via HTML pages

A second solution

- Start to build visualisation information into the workflow, using beanshell scripts.
- http://www.mrl.nott.ac.uk/~sre/workflowblatest
- But what if we change the workflow?

Summary

- Domain models are hard
- Workflows can obfuscate the model
- Visualisation requires one
- We can build some knowledge of a domain model into the workflow
- Is there a better way?

Acknowledgements

Core

Matthew Addis, Nedim Alpdemir, Neil Davis, Alvaro Fernandes, Justin Ferris, Robert Gaizaukaus, Kevin Glover, Carole Goble, Chris Greenhalgh, Mark Greenwood, Yikun Guo, Ananth Krishna, Peter Li, Phillip Lord, Darren Marvin, Simon Miles, Luc Moreau, Arijit Mukherjee, ***Tom Oinn***, Juri Papay, Savas Parastatidis, Norman Paton, Terry Payne, Matthew Pocock Milena Radenkovic, Stefan Rennick-Egglestone, Peter Rice, Martin Senger, Nick Sharman, Robert Stevens, Victor Tan, Anil Wipat, Paul Watson and Chris Wroe.

Users

Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle, UK

Hannah Tipney*, *May Tassabehji, Andy Brass, St Mary's Hospital, Manchester, UK

Postgraduates

Martin Szomszor, Duncan Hull, Jun Zhao, Pinar Alper, John Dickman, Keith Flanagan, Antoon Goderis, Tracy Craddock, Alastair Hampshire

myGrid