A Systematic Strategy for Large-Scale Analysis of Genotype-Phenotype Correlations

Paul Fisher[1§], Cornelia Hedeler[1], Katherine Wolstencroft[1], Helen Hulme[1], Harry Noyes[2], Stephen Kemp[2], Robert Stevens[1] and Andrew Brass[1, 3]

[1] School of Computer Science, Kilburn Building, University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[2] School of Biological Sciences, Biosciences Building, University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK
[3] Faculty of Life Science, Michael Smith Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

[§]Corresponding author: pfisher@cs.manchester.ac.uk

The process of linking genotype and phenotype plays a crucial role in understanding the biological processes which contribute to overall cellular, tissue and organism responses, particularly when under a disease state. Researchers have discovered single-gene lesions for a large number of simple Mendelian traits; however, it has proved much more difficult to discover genes underlying genetically complex traits.

The use of Quantitative Trait Loci data is increasingly being used to aid in the discovery of candidate genes involved in phenotypic variation. Tens to hundreds of genes, however, may lie within even well defined QTL; it is therefore vital that the identification, prioritisation and functional testing of candidate Quantitative Trait genes (QTg) is carried out systematically without bias introduced from prior assumptions about candidate genes or the biological processes that may be influencing the observed phenotype. With the advent of microarrays, researchers are able to directly examine the expression of all genes under a QTL.

The scale of data being generated by such high-throughput experiments has led some investigators to follow a hypothesis-driven approach, where the triage and selection of candidate genes is based on some prior knowledge or assumption. Although these techniques for candidate gene identification are valid, they run the risk of overlooking genes that have less obvious associations with the phenotype. The complexity of multigenic traits can also lead to problems when attempting to identify the varied processes involved in the phenotype. By making selections based on prior assumptions of what processes may be involved, the pathways, and therefore genes, that may actually be involved in the phenotype can be overlooked or missed entirely.

A further complication is that the use of *ad hoc* methods for candidate gene identification are inherently difficult to replicate and are compounded by poor documentation of the methods used to generate and capture the data from such investigations in published literature. The use of 'link-integration' through any number of data resources further exacerbates the problem of capturing the methods used for obtaining *in silico* results since it is often difficult to identify the essential data in the chain of hyperlinked resources.

With an ever increasing number of institutes offering programmatic access to their resources in the form of web services, however, experiments previously conducted manually can now be replaced by automated experiments, capable of processing a far greater volume of data. By reconstructing the original investigation methods in the form of workflows, we are now able to pass data directly from one service to the next without the need for any interaction from researchers. This enables us to process the data in a much more systematic, un-biased, and explicit manner.

We propose a methodology that revises the known pathways that intersect a QTL and those derived from a set of differentially expressed genes. This methodology is implemented systematically through the use of web services and workflows. For the purpose of implementing this systematic pathway driven approach, we have chosen a service based infrastructure, ^{my}Grid, coupled with workflow technology provided by the Taverna workbench.