An Introduction to Running, Reusing and Sharing Workflows with Taverna

Stian Soiland-Reyes and Christian Brenninkmeijer
University of Manchester

materials by Katy Wolstencroft, Aleksandra Pawlik, Alan Williams

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
http://orcid.org/0000-0003-3156-2105

Bonn University, 2014-09-01 / 2014-09-03
http://www.taverna.org.uk/

- □ This tutorial will give you a basic introduction to reusing workflows in Taverna and my Experiment.
- □ Other tutorials will explore nested workflows, the workflow engine (iteration, looping, parallel invocation)
- □ Like in the previous tutorial workflows in this practical use small data-sets and are designed to run in a few minutes. In the real world, you would be using larger data sets and workflows would typically run for longer

Bigger Workflows: Enrichment Analysis

The previous examples were trivial, small tasks. Taverna’s real power is in iterating over large data sets

- Many experiments result in a list of genes (e.g. microarray analysis, Chip-Seq, SNP identification etc).
- In this exercise, we will use Taverna to analyse a gene set from a Chip-Seq experiment by finding and reusing existing workflows
- We will enrich our dataset by discovering:

1. Which pathways our genes are involved in
2. The functions of the genes
3. Literature evidence for the phenotype/trait of interest

Re-using workflows from myExperiment

- Go to http://www.myexperiment.org and click on ‘find workflows’
- You will see a list of the most viewed and downloaded workflow – see what the most popular workflow does by reading the description
- Change the rank to ‘Latest’ and see what has been uploaded in the last few weeks
- We will now find and download a workflow to identify the pathways each gene in our gene set is involved in

Re-using workflows from myExperiment

- □ Find the workflow called "UnigeneID to KEGG Pathways" and look at the workflow entry page (uploaded by "Aleksandra Pawlik")
- □ Download the workflow by clicking on the link: "Download Workflow" and find out what it does by reading the descriptions in myExperiment
- □ Open the workflow in Taverna by going to 'File ->Open Workflow'
- □ Run the workflow using the example values supplied (Hint: when you run the workflow you can add the example values by clicking on "Use examples")
- □ Look at the workflow output – now you will see pathway information and pathway diagrams

Combining workflows from myExperiment

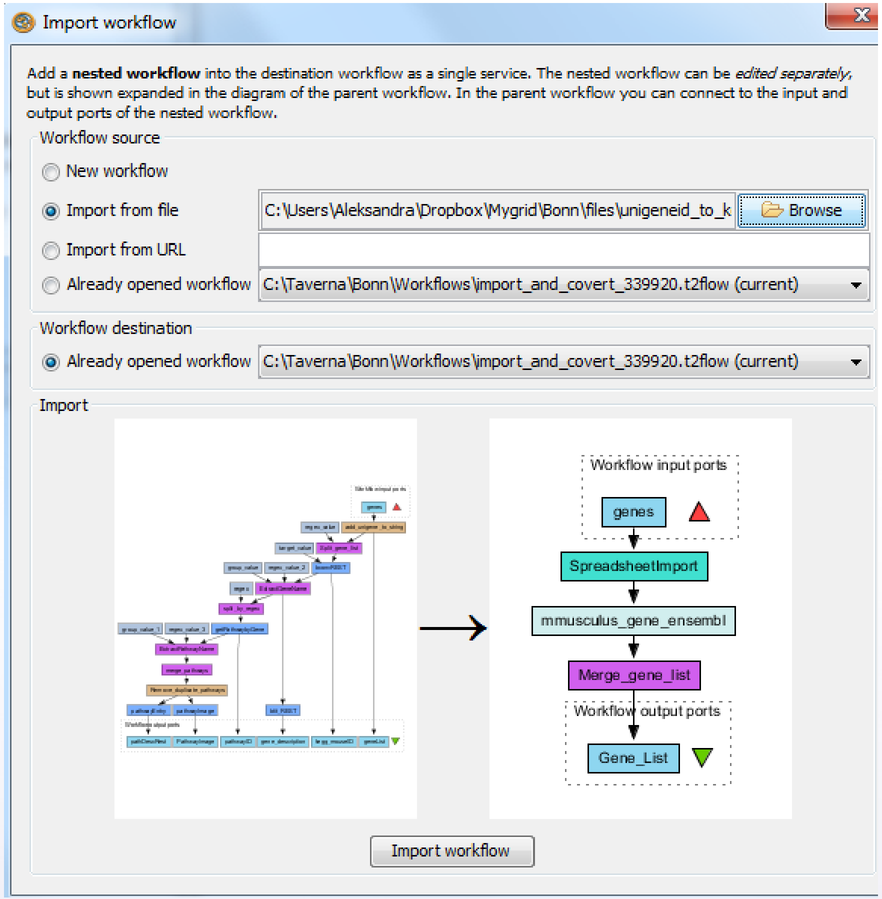
- To analyse all the genes from our ChipSeq study, we need to extract the gene list from our results file
- To make it easier to work through the example, we have provided a Chip-Seq gene list on myExperiment, you can find it under "GalaxyGeneList - short : datafile for training"
- Save this file to your local machine
- Open the file in Excel
- Save the file with a .csv extension
- As you can see, the list of genes is in column D
- Taverna can process and extract this column automatically

Combining workflows from myExperiment

- In myExperiment, find and download the workflow called “Import and convert gene list”
- This workflow will extract the list of genes in column D using Taverna’s built-in spreadsheet import tool (which can be found in the services panel, for future reference)
- The next step in the workflow converts RefSeq IDs into unigene IDs (required for the pathways workflow – converting between different types of identifiers is a common problem in bioinformatics!)
- Run the workflow. This time, in the input window, select “set file location” and set the location to the saved .csv gene list.
- Look at the workflow results

Combining workflows from myExperiment

- We will now combine the two workflows
- While you are still in the “import and convert” workflow, go to the top of the workbench and select “insert -> Nested workflow”
- In the pop-up window, select “import from file” and find the pathways workflow you downloaded earlier.
- Click on “import workflow” and the pathways workflow will appear in the main workflow diagram.



Import workflow

Add a **nested workflow** into the destination workflow as a single service. The nested workflow can be *edited separately*, but is shown expanded in the diagram of the parent workflow. In the parent workflow you can connect to the input and output ports of the nested workflow.

Workflow source

○ New workflow

● Import from file: C:\Users\Aleksandra\Dropbox\Mygrid\Bonn\files\unigeneid_to_k

○ Import from URL

○ Already opened workflow: C:\Taverna\Bonn\Workflows\import_and_covert_339920.t2flow (current)

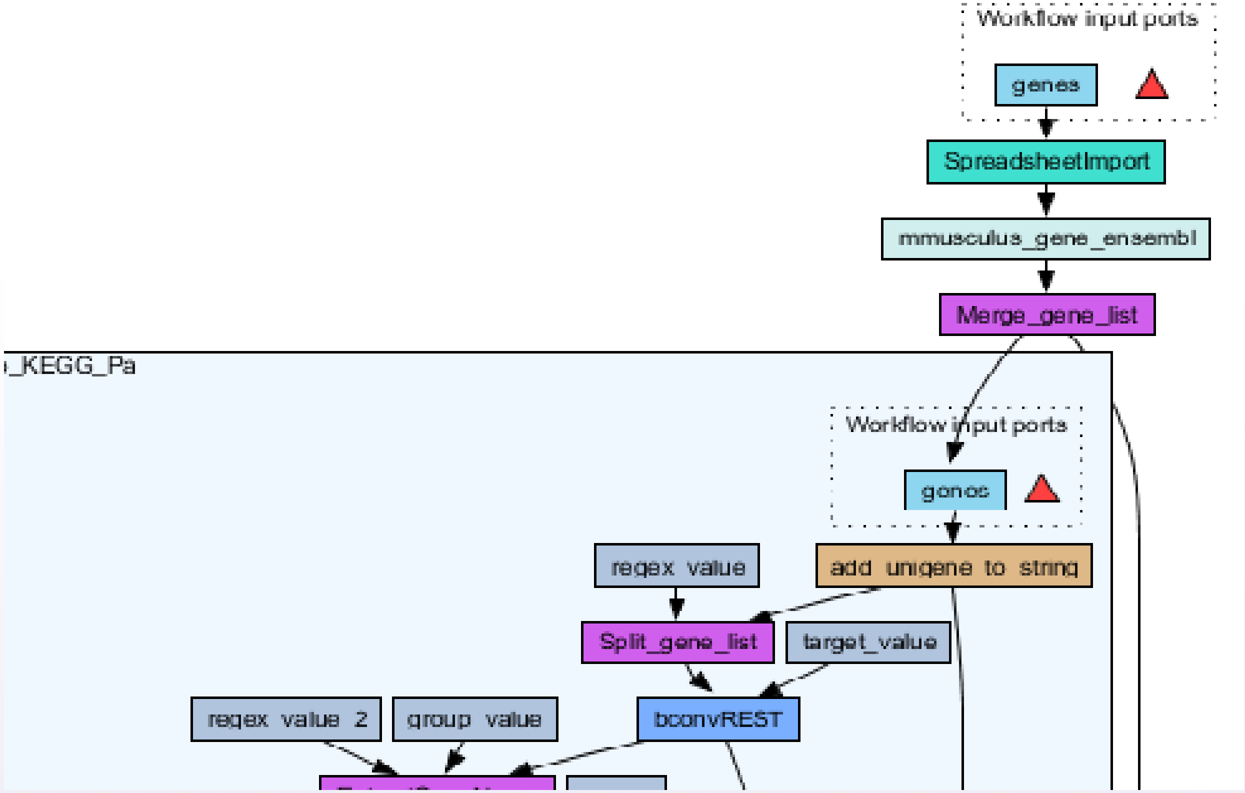
Workflow destination

● Already opened workflow: C:\Taverna\Bonn\Workflows\import_and_covert_339920.t2flow (current)

Import

Combining workflows from myExperiment

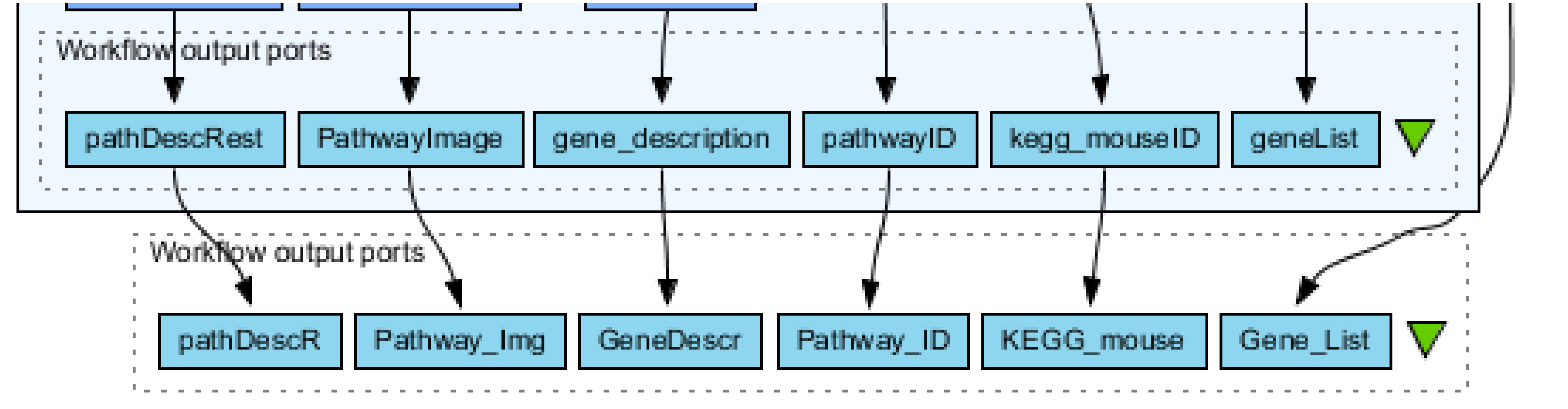
- Connect the workflows up by linking the output of the 'Merge_Gene_List' with the nested workflow input



Combining workflows from myExperiment

- Create new output ports for the Nested workflow and connect the Nested workflow outputs to the new outputs

NOTE: you don't need to connect them all, just pathway descriptions, pathway images and gene descriptions



- Save the workflow
- Run the workflow (it may take a few minutes)

GO Associations

- There are many different tools we could use to find Gene Ontology associations for your gene list
- For example, we could simply modify the BioMart/Ensembl service in the ‘Import and convert gene list’ workflow we have already used
- Reload the ‘Import and Convert gene list’ workflow
- Right-click on the ‘mmusculus_gene_ensembl’ service and select ‘Copy’
- Paste an extra copy of this service into the same workflow diagram

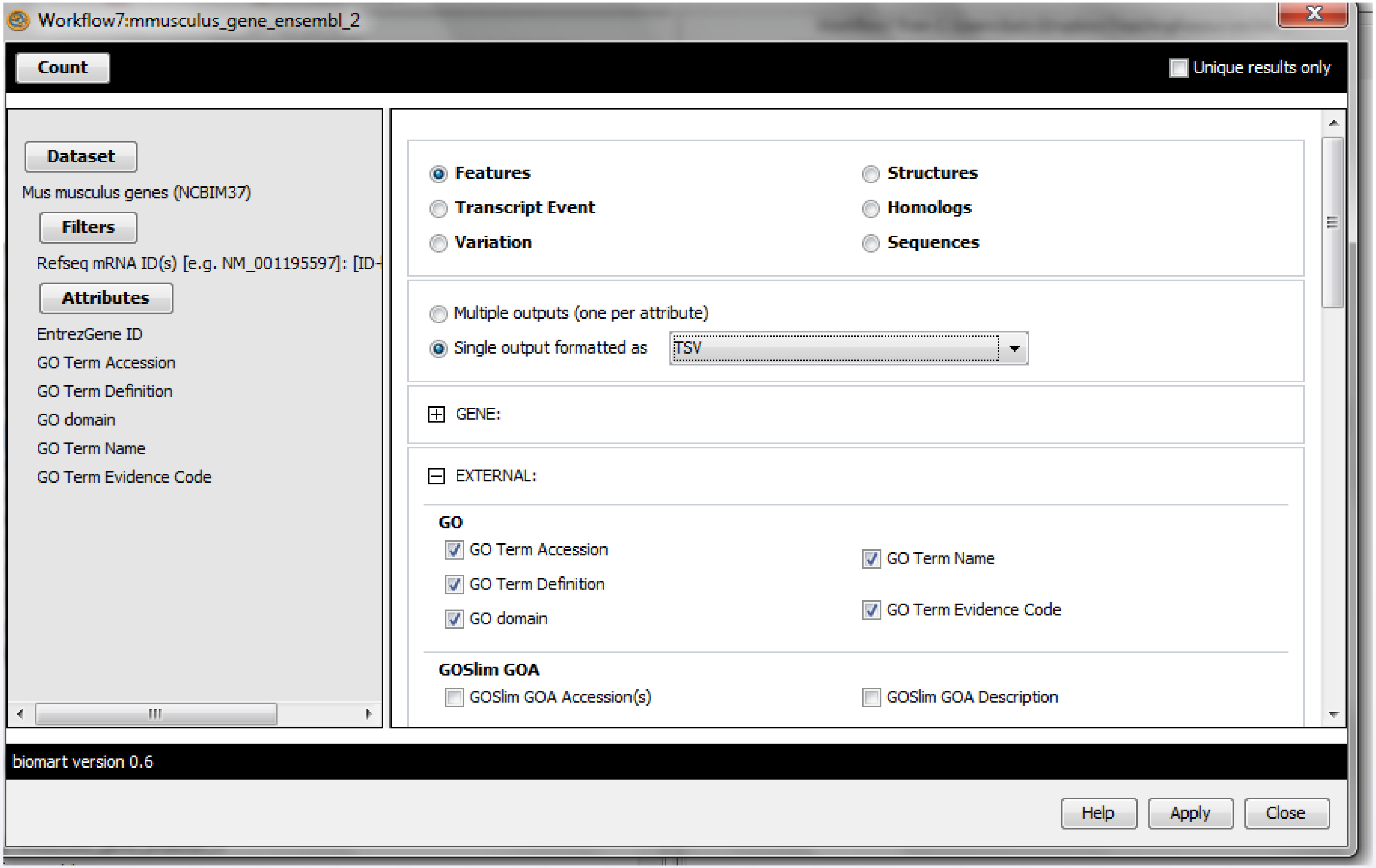
GO Associations

- This is a BioMart service. It allows you to retrieve omics data from ENSEMBL and other genomics resources. If you are familiar with BioMart, you will see the interface in Taverna is very similar to the web interface
- We will modify the BioMart query to find all GO associations for each gene associated with a Chip-Seq peak
- Right-click on the new copy of the service and select ‘Configure BioMart Query’

GO Associations

- The inputs (or filters) already accept RefSeq Ids from our input file, but we need to modify the outputs (or attributes)
- Select ‘Attributes’ and expand the …‘External’… section.
- Select ‘Go Term Accession’, ‘GO Term Definition’ and ‘Go Domain’
- Unselect ‘UniGeneID’ and select ‘RefSeq mRNA’…
- (See screenshot on the next slide for an example)

GO Associations



□ Unique results only

Dataset: Mus musculus genes (NCBIM37)

Filters: Refseq mRNA ID(s) [e.g. NM_001195597]: [ID-

Attributes:
EntrezGene ID
GO Term Accession
GO Term Definition
GO domain
GO Term Name
GO Term Evidence Code

● Features ○ Structures
○ Transcript Event ○ Homologs
○ Variation ○ Sequences

○ Multiple outputs (one per attribute)
● Single output formatted as TSV

⊞ GENE:

⊟ EXTERNAL:

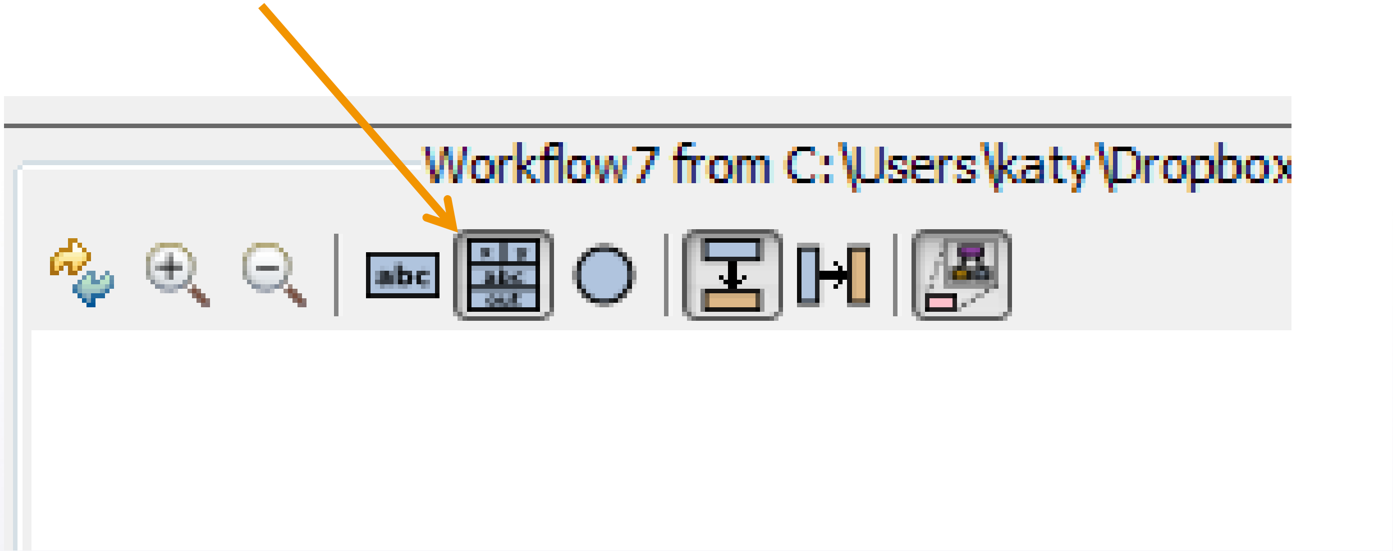
GO
☑ GO Term Accession ☑ GO Term Name
☑ GO Term Definition ☑ GO Term Evidence Code
☑ GO domain

GOSlim GOA
□ GOSlim GOA Accession(s) □ GOSlim GOA Description

biomart version 0.6

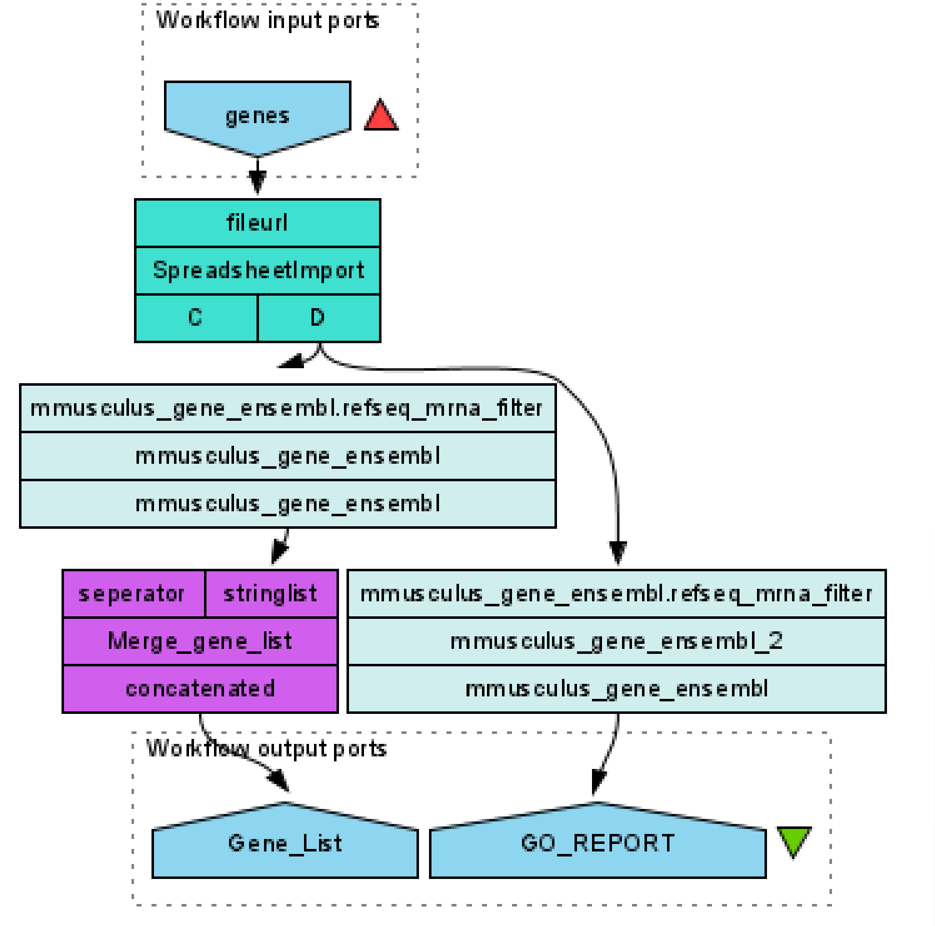
GO Associations

- Click ‘apply’ to save your changes, and ‘close’, to go back to Taverna
- At the top of the workflow diagram, change the workflow view to show all ports by clicking on the table icon



GO Associations

- Connect your new service tc the workflow by linking the ‘D’ output port of the spreadsheet service to the input of your new service
- Make the new output ports and connect them as shown to your new service



GO Associations

- Save the workflow by going to ‘File -> Save Workflow’
- Run the workflow
  - Set file location to the GalaxyGeneList file saved earlier
- (Download and) view the GO report
  - In “GO_REPORT” tab select “Value 1” and press “Save Value”
  - Save as a .txt or as a .tsf (tab separated) file

Simple Text Mining

- So far we have looked at enriching the genomic information, but we could also use workflows for running data analyses (e.g. aligning mouse genes with human homologues) or performing literature searches
- Think about the ways you could extend this analysis with literature searches (e.g. Correlations between pathways, genes, GO terms, phenotypes etc)
- Search myExperiment for workflows involving text mining, using the search terms "text mining" and "Pubmed"

Text Mining

- Find and open the workflow “Phenotype to pubmed”
- One of the services is no longer available in the nested workflow (the faded-out service). Taverna checks the availability of each service when you load the workflow and when you run it
- In this case, the workflow will still run without the final nested workflow (clean text)
- Delete the ‘clean text’ nested workflow (by selecting it and right-clicking), and reconnect the workflow output
- Run the workflow with the search term ‘erythropoiesis’ (or a phenotype term to describe the disease you are studying)