AtlasRDF

James Malone

November 3, 2013

1 Introduction

The AtlasRDF package exposes the data stored in the Expression Atlas RDF store at the European Bioinformatics Institute. This data concerns the reporting of genes which are differentially expressed under certain experimental and biological conditions in certain species. These conditions - often called experimental factors - include diseases such as cancer, phenotypes such as obesity, cell types such as epithelial cells, organism part such as brain and age. The data is also connected to other RDF stores such as proteins in UniProt and Pathways in Reactome, all of which this packages uses to enrich the queries available.

Here we demonstrate how to use the Atlas to extract a gene list based on a condition of interest to and to explore other data that may offer information about the condition and/or gene list using the AtlasRDF R package.

The package is loaded using

2 Extracting a gene list

First of all we will find the appropriate ontology term to query the database with. AtlasRDF uses EFO <http:/www.ebi.ac.uk/efo> to annotate data with so we will find the appropriate EFO class URI for our search term which is type II diabetes.

```
> termhits <- searchForEFOTerms("type II diabetes")
> print(termhits)
```

```
                                  efouri                      label
1 <http://www.ebi.ac.uk/efo/EFO_0001360> type II diabetes mellitus
```

If you already use another ontology, we can also find an appropriate EFO term by using the term mappings maintained by NCCBO Bioportal <http://bioportal.bioontology.org/> Here we do a simplel query of passing the URI of the class you use (in this example the URI for type II diabetes from SNOMEDCT) to find any matching terms in EFO, as mapped by BioPortal.

```
> efomappings <- getOntologyMappings("<http://purl.bioontology.org/ontology/SNOMEDCT/44054006>")
> print(efomappings)
```

```
                                  efouri
1 <http://www.ebi.ac.uk/efo/EFO_0001360>
```

We select the class URI for type II diabetes called 'type II diabetes mellitus' and extract get a gene list based on those genes differentially expressed in humans for type II diabetes. First we get the URI for the species - human in this case. Then we query for the genes for diabetes Type II.

```
> humanURI <- getTaxonURI("human")
> typeIIgenelist <- getSpeciesSpecificEnsemblGenesForExFactor("<http://www.ebi.ac.uk/efo/EFO_0001
> head(typeIIgenelist)
```

```
                                              dbXref genename       ensemblid
1 <http://identifiers.org/ensembl/ENSG00000166845> C18orf54 ENSG00000166845
2 <http://identifiers.org/ensembl/ENSG00000140521>     POLG ENSG00000140521
3 <http://identifiers.org/ensembl/ENSG00000163602>     RYBP ENSG00000163602
4 <http://identifiers.org/ensembl/ENSG00000163624>     CDS1 ENSG00000163624
5 <http://identifiers.org/ensembl/ENSG00000114315>     HES1 ENSG00000114315
6 <http://identifiers.org/ensembl/ENSG00000127526>  SLC35E1 ENSG00000127526

>
```

3 Refining the gene list

We now have a list of candidate genes of interest. We will now refine the list of approx 250 genes by finding genes which share some common signalling pathways. The following block of code may take a little while to run.

```
> #this will take a little while to run
> pathways <- getRankedPathwaysForGeneIds(typeIIgenelist[,3])
> head(pathways)

[[1]]
An object of class "pathwayresult"
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_12627.3>"

Slot "label":
[1] "Generic Transcription Pathway"

Slot "numgenes":
[1] 4

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000130684>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000167555>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000173041>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000198521>"



[[2]]
An object of class "pathwayresult"
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_118780.2>"

Slot "label":
[1] "NOTCH1 Intracellular Domain Regulates Transcription"
```

```
Slot "numgenes":
[1] 3

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000141027>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000182568>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000100603>"



[[3]]
An object of class "pathwayresult"
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_160243.1>"

Slot "label":
[1] "Constitutive Signaling by NOTCH1 PEST Domain Mutants"

Slot "numgenes":
[1] 3

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000141027>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000182568>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000100603>"



[[4]]
An object of class "pathwayresult"
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_160254.1>"

Slot "label":
[1] "Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants"

Slot "numgenes":
[1] 3

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000141027>"

$geneuri
```

```
[1] "<http://identifiers.org/ensembl/ENSG00000182568>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000100603>"



[[5]]
An object of class "pathwayresult"
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_15525.4>"

Slot "label":
[1] "Nuclear Receptor transcription pathway"

Slot "numgenes":
[1] 3

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000112033>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000186350>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000269571>"



[[6]]
An object of class "pathwayresult"
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_2155.5>"

Slot "label":
[1] "NICD traffics to nucleus"

Slot "numgenes":
[1] 2

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000182568>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000100603>"
```

We can now see a list of pathways that these genes are connected to, with the most common pathway ranked as the first element in the list.

```
> pathways[1]
```

```
[[1]]
An object of class "pathwayresult"
```

```
Slot "pathwayuri":
[1] "<http://identifiers.org/reactome/REACT_12627.3>"

Slot "label":
[1] "Generic Transcription Pathway"

Slot "numgenes":
[1] 4

Slot "genes":
$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000130684>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000167555>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000173041>"

$geneuri
[1] "<http://identifiers.org/ensembl/ENSG00000198521>"
```

Given this pathway, we can extract some further information. For instance, let us now find other experimental conditions that are connected to this pathway. First of all let's find other genes attached to this pathway.

```
> genes <- getGenesForPathwayURI(pathways[[1]]@pathwayuri)
>
```

We can ask a few other questions about these genes. For instance, what other experimental factors are these genes differentially expressed for? We can use the Atlas enrichment functions to perform this task.

Firstly, we need to download the human background data set. We require two background sets for each species we wish to perform enrichment for, since these are human genes we will download the human backgrounds. These can be found at https://github.com/jamesmalone/AtlasRDF-R/tree/master/src/backgroundsetdata in the human subfolder you will find the two background data sets required.

After downloading we need to load these files into our workspace.

```
> load("human/human_gene_list.RData") #human_genelist_bg
> load("human/human_factor_counts.RData") #human_factor_counts
```

Finally do the enrichment, passing the gene list

```
> ###do enrichment
> transcription_pathway_enrichment <- doFishersEnrichment(genes, human_genelist_bg, human_factor_
```

If you don't have access to the enrichment background data sets, you can use the example enrichment result set included. To load this data use the command:

```
> #data(transcription_pathway_enrichment)
```

You can now visualise these results in a graph.

```
> vizPvalues(transcription_pathway_enrichment, "0.00005")
```

dorsal root ganglion
4.551272e-17
differentiated
1.604334e-16
HCC70
9.597074e-15
lung adenocarcinoma
3.302170e-14
peripheral nervous system
5.097874e-14
ganglion
5.097874e-14
SW620
1.349740e-13
methotrexate
8.806241e-13
multiple sclerosis
2.129027e-12
hereditary spastic paraplegia
3.155249e-12
Down syndrome
5.526945e-12
Intestinal malformation
5.526945e-12
Eyelids malposition disorder
5.526945e-12
Secondary ectropion
5.526945e-12
Canthal anomaly
5.526945e-12
Epicanthal fold
5.526945e-12
Syndromic epicanthus
5.526945e-12
Malposition of external canthus
5.526945e-12
Syndromic keratoconus
5.526945e-12
Syndromic intestinal malformation
5.526945e-12
Keratoconus
5.526945e-12
refractory anemia
8.032289e-12
melanoma
9.421247e-12
dermatomyositis
1.071563e-11
17beta-estradiol
1.268613e-11
trigeminal ganglion
1.510773e-11
Neuro-ophthalmological disease
2.258449e-11

Genetic neuro-ophthalmological disease
2.258449e-11
chronic childhood arthritis
4.188649e-11
adrenocortical carcinoma
5.397509e-11
juvenile dermatomyositis
8.077342e-11
B cell derived cell line
1.166269e-10
glucose
1.313050e-10
glutamine
1.313050e-10
amino acid
1.313050e-10
NCI-H1975
2.018043e-10
Gorilla gorilla
3.558656e-10
"GM00719"@en
3.755884e-10
Chromosomal anomaly
3.784899e-10
Syndromic renal or urinary tract malformation
3.784899e-10
Anorectal malformation
3.784899e-10
Autosomal anomaly
3.784899e-10
Autosomal trisomy
3.784899e-10
Total autosomal trisomy
3.784899e-10
Chromosomal anomaly with cataract
3.784899e-10
Syndromic anorectal malformation
3.784899e-10
Genetic renal or urinary tract malformation
3.784899e-10
Genetic digestive tract malformation
3.784899e-10
caudate nucleus
4.069622e-10
MOLT-4
4.599782e-10
U937
4.606736e-10
Rare genetic renal disease
5.069566e-10
"metastasis to lymph node"@en
5.200810e-10
ssMCF7
5.925618e-10

Facioscapulohumeral dystrophy
6.189042e-10
occipital lobe
6.473624e-10
chronic myeloproliferative disorder
6.821629e-10
Rare strabismus and restriction syndrome
7.106703e-10
Syndrome with a symptomatic strabismus
7.106703e-10
U87
8.326930e-10
Rare genetic eye disease
1.164370e-09
SKMEL5
1.273771e-09
adrenal gland neoplasm
2.151316e-09
mixed sex population
2.237775e-09
fibroblast derived cell line
3.413414e-09
chronic myelogenous leukemia
3.471929e-09
Rare genetic skin disease
3.586947e-09
bacterial disease
4.116355e-09
HaCaT
4.296797e-09
nevus
5.515653e-09
benign neoplasm
5.776003e-09
medulla oblongata
6.819602e-09
interleukin-4 (Homo sapiens)
7.267976e-09
interleukin (Homo sapiens)
7.267976e-09
interleukin
7.267976e-09
Muscular dystrophy
7.935404e-09
Genetic neuromuscular disease
7.935404e-09
Genetic skeletal muscle disease
7.935404e-09
Progressive muscular dystrophy
7.935404e-09
bladder carcinoma
1.039201e-08
hepatocellular carcinoma
1.056329e-08

liver disease
1.056329e-08
Rare genetic developmental defect during embryogenesis
1.136407e-08
saline
1.284213e-08
frontal lobe
1.359377e-08
spinal cord
1.746833e-08
Rare palpebral, lacrimal system and conjunctival diseases
2.269362e-08
Rare palpebral disease
2.269362e-08
Rare genetic palpebral, lacrimal system and conjunctival disease
2.269362e-08
A549
2.606638e-08
lung cancer cell line
2.934397e-08
MDAMB468
3.231395e-08
thalamus
3.265892e-08
inflammatory bowel disease
3.946729e-08
Rare genetic neurological disease
4.123526e-08
Rare otorhinolaryngological malformation
4.405727e-08
Syndrome or malformation associated with head and neck malformations
4.405727e-08
Genetic head and neck malformation
4.405727e-08
clear cell sarcoma of the kidney
4.513018e-08
small intestine
4.971445e-08
Huntington disease
5.400857e-08
Oculomotor apraxia or related oculomotor disease
5.400857e-08
Lens and zonula anomaly
8.242084e-08
Rare cataract
8.242084e-08
Syndromic cataract
8.242084e-08
Genetic lens and zonula anomaly
8.242084e-08
germ cell tumor
9.707078e-08
testicular seminoma
9.707078e-08

parietal lobe
1.045082e-07
adrenal gland
1.074964e-07
adrenal cortex
1.074964e-07
heart component
1.137984e-07
Defect in innate immunity
1.186904e-07
Primary immunodeficiency
1.186904e-07
Chronic granulomatous disease
1.186904e-07
Genetic immune deficiency with skin involvement
1.186904e-07
Functional neutrophil defect
1.186904e-07
Rare genetic immune disease
1.186904e-07
Rare dystonia
1.272960e-07
Heredodegenerative disease with dystonia as a major feature
1.272960e-07
Genetic dementia
1.272960e-07
Genetic neurodegenerative disease
1.272960e-07
Rare genetic movement disorder
1.272960e-07
Genetic neurodegenerative disease with dementia
1.272960e-07
week
1.367558e-07
cardiovascular system
1.669147e-07
hippocampus
1.756621e-07
amygdala
2.096179e-07
trichostatin A
2.133855e-07
BJAB
2.189406e-07
estradiol
2.670330e-07
hypothalamus
3.082404e-07
Burkitts lymphoma
3.356678e-07
temporal lobe
3.398621e-07
sevoflurane
3.454125e-07

adenoma
3.639875e-07
muscular system
4.564782e-07
muscle
4.564782e-07
myocardium
4.569026e-07
atrial myocardium
4.569026e-07
gastric fundus
4.862617e-07
cardiac ventricle
5.012813e-07
caecum
5.014048e-07
salivary gland
5.246266e-07
A498
6.047078e-07
basal-like carcinoma
6.225189e-07
propofol
6.396943e-07
Caki2
6.445676e-07
BC-1
6.992697e-07
lymph node
7.284753e-07
ITM-E6E7
9.577759e-07
spleen
1.018491e-06
Myopathy with eye involvement
1.325031e-06
Myasthenic syndrome with eye involvement
1.325031e-06
Duchenne muscular dystrophy
1.325031e-06
genetic disorder
1.387374e-06
ulcerative colitis
1.651803e-06
duodenum
2.356772e-06
Hodgkins lymphoma
2.596379e-06
6-propyl-2-thiouracil
3.092639e-06
Rare genetic cardiac disease
3.333536e-06
Familial dilated cardiomyopathy
3.333536e-06

```
                                  Neuromuscular disease with dilated cardiomyopathy
                                                                        3.333536e-06
                                                vulvar intraepithelial neoplasia
                                                                        3.648883e-06
```

There are a lot of resuts here, even with a low p-value threshold, so let's filter some types of factors of interest. Let's look at just disease factors these genes are enriched for. The ontology class for disease is efo:EFO_0000408 so let's include only subtypes of this class, i.e. diseases

We can find the class in efo by using the search function since we will not usually know what the identifier of an EFO class is when we perform this filtering.

```
> filteredgenes <- includeOnlySubclasses("efo:EFO_0000408", transcription_pathway_enrichment)
```

Though we still have quite a lot of results, even from this filter, so let's take the top 20 most enriched experimental factors.

```
> sortedset <- orderEnrichmentResults(filteredgenes)
> vizgraph <- vizPvalues(sortedset[1:20], 0.000001)
```

This points us to other factors that may be of interest to this pathway.