Repeatable plant pathology bioinformatic analysis: Not everything is NGS data.

Peter Cock & Leighton Pritchard

Genomics for Non-Model Organisms

ISMB/ECCB - Workshop 6

Vienna, Austria

19 July 2011

The James Hutton Institute

JHI Plant Pathology

- We work on a range of organisms
  - Plant Viruses
  - Bacteria
  - Oomycetes
  - Fungi
  - Nematodes
  - Aphids (as virus vectors)
- Many genome sequences now available…

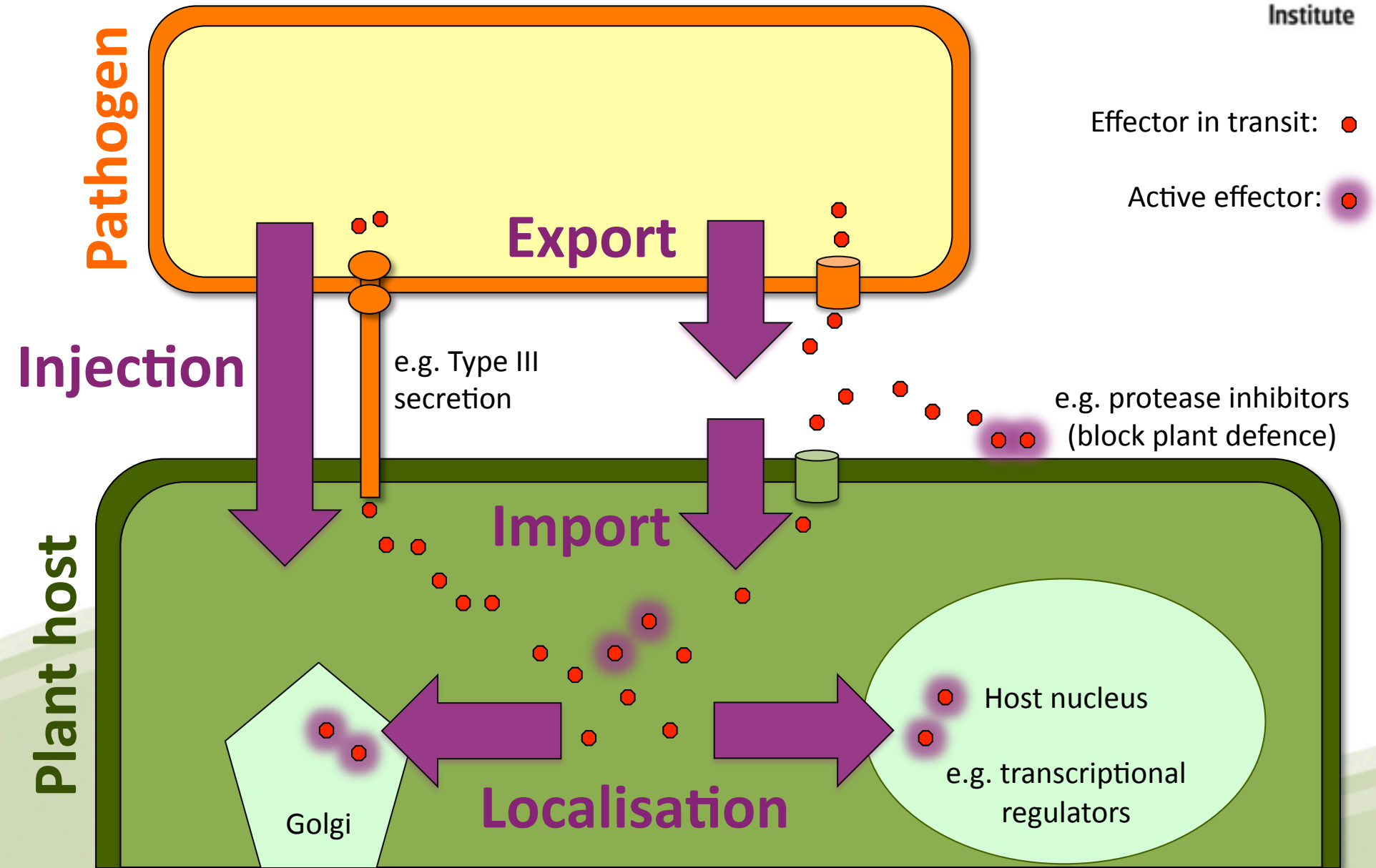
JHI Dundee site, formerly SCRI (Scottish Crop Research Institute)

Plant Pathogen Genomics

The James Hutton Institute

- Sequencing of plant pathogens becoming common
  - I will focus on the protein complement today (i.e. What happens *after* assembly of NGS data)
- Often looking at first genome for a genus
  - Automated gene annotation from other species limited
- In particular we want to find “effectors”
  - i.e. proteins used to manipulate host

Effector Protein Analysis



Example: Effector Protein Analysis

The James Hutton Institute

- Want to identify effector genes, e.g.
  - Similarity to known effectors (e.g. with BLAST)
  - Signal peptides
  - Localization signals
  - Possible horizontal gene transfer (e.g. different GC%)
- Part of larger task of automated gene annotation, e.g.
  - HMMER or RPS-BLAST domain searches
  - GO annotation with Blast2GO

Example: Nematode Effector Proteins

The James Hutton Institute

- To identify candidate effector genes:
  - Want a signal peptide (for export)
  - Don't want a transmembrane domain (not secreted)
- How? Initially used a Python script calling
  - SignalP 3.0
  - TMHMM 2.0
- Now we make this easy to use, thanks to Galaxy…

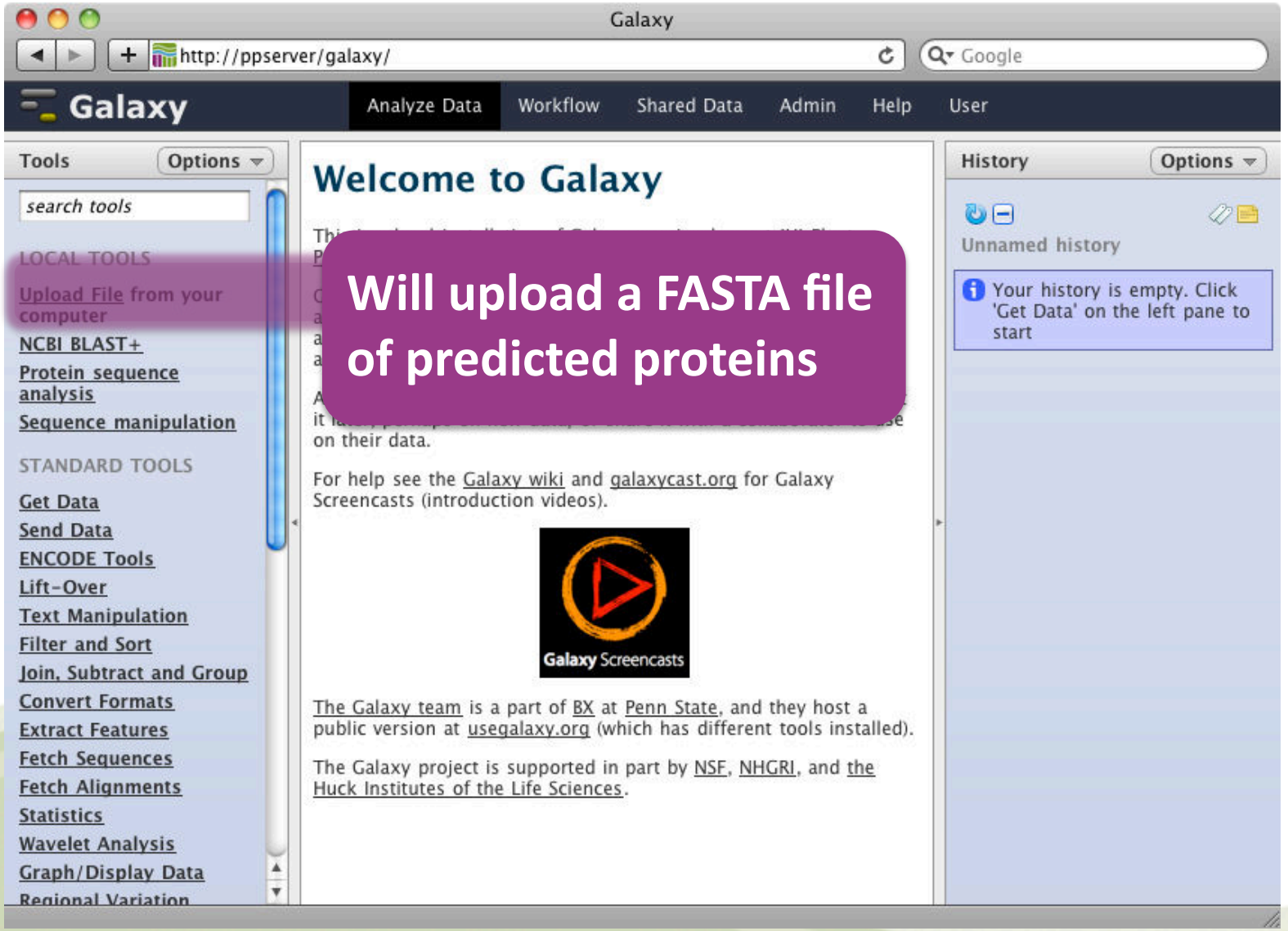
See Jones *et al.* (2009), Molecular Plant Pathology

What is Galaxy?

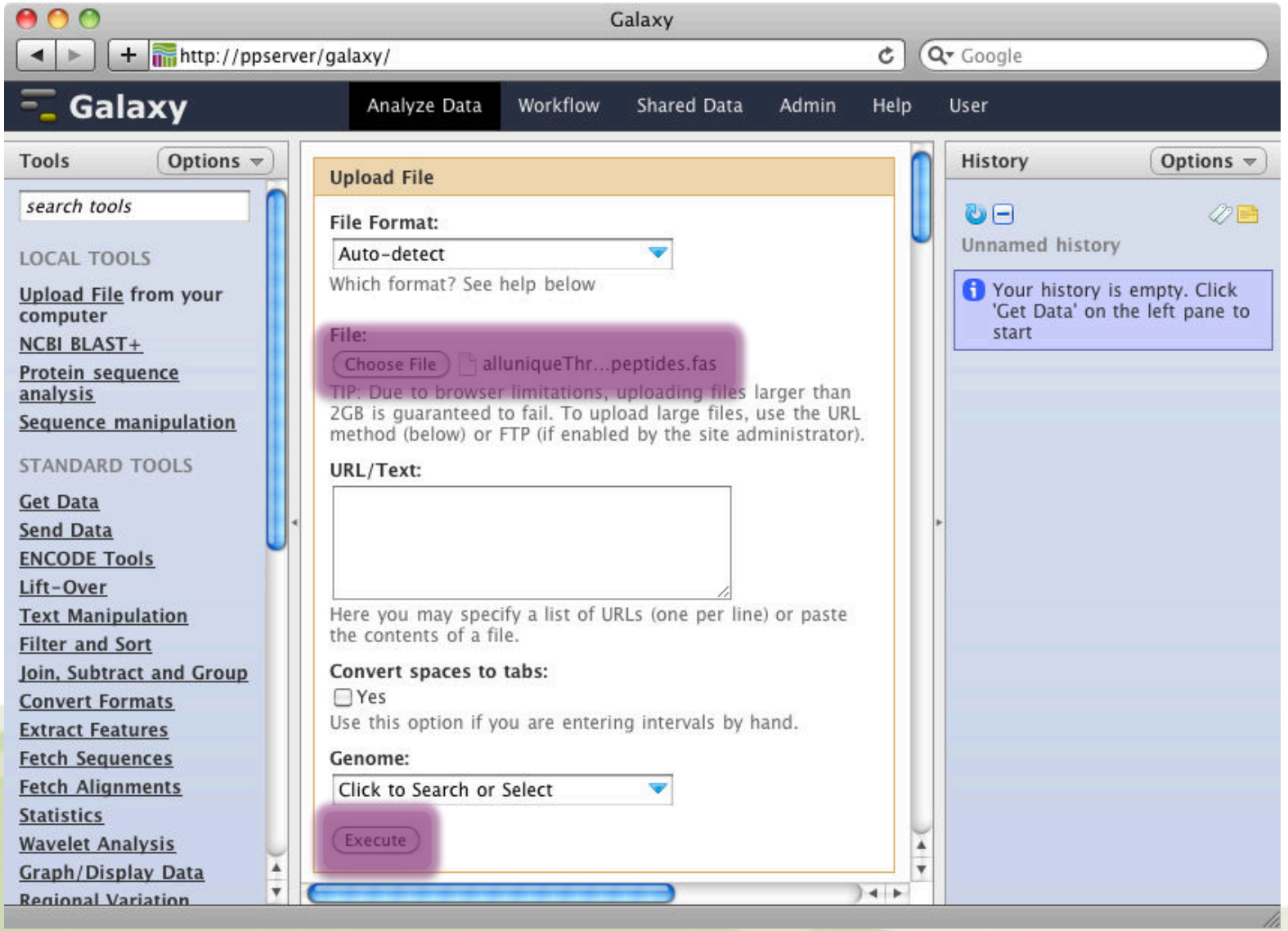
The James Hutton Institute

- Web system for running many analysis tools remotely:
  - Upload your data to Galaxy (or access shared data)
  - Run analysis on Galaxy server/cluster
  - Download results (or share them)
- Designed to link multiple analysis steps as workflows:
  - Repeatable
  - Shareable

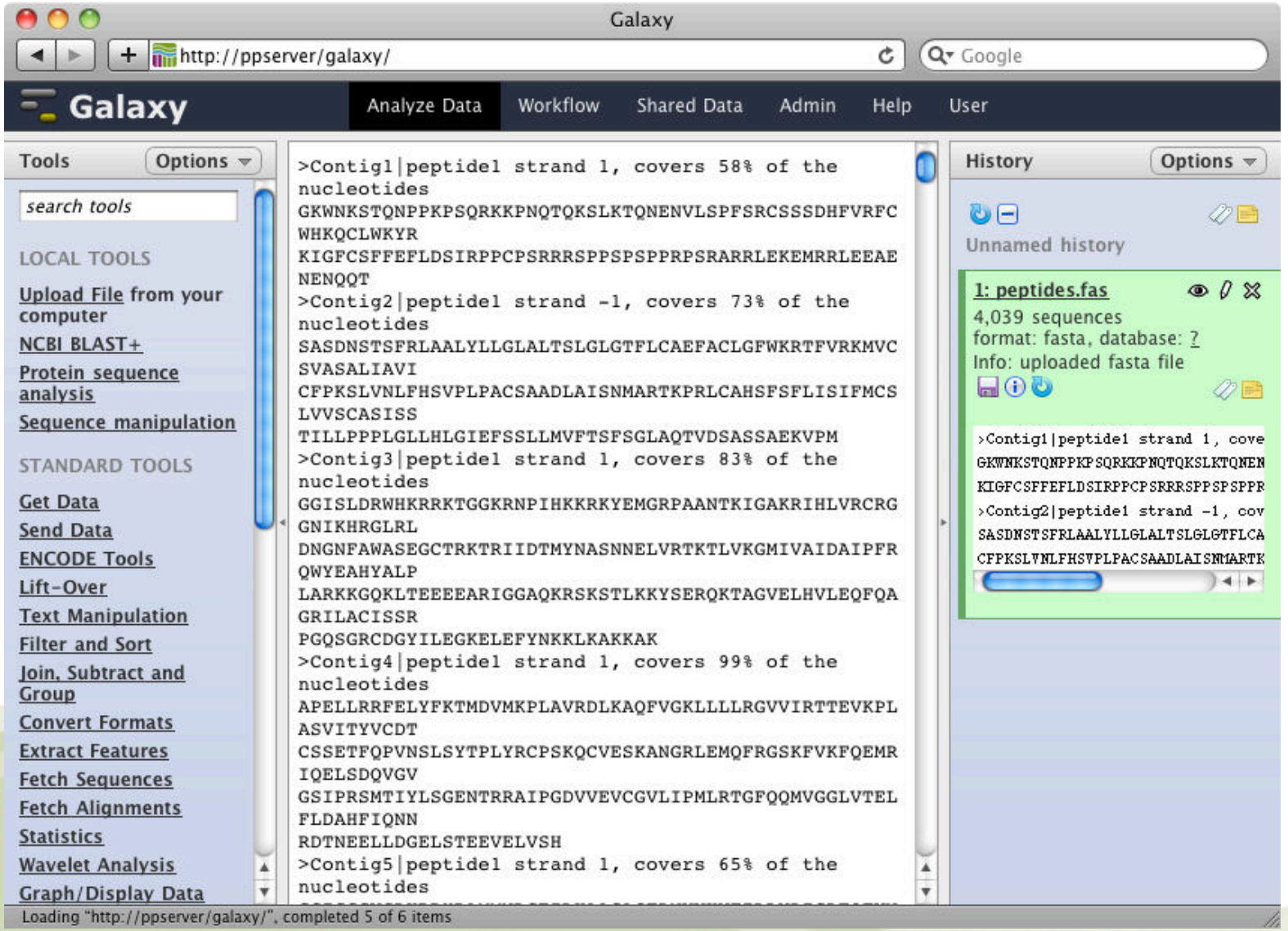
Our local Galaxy Server



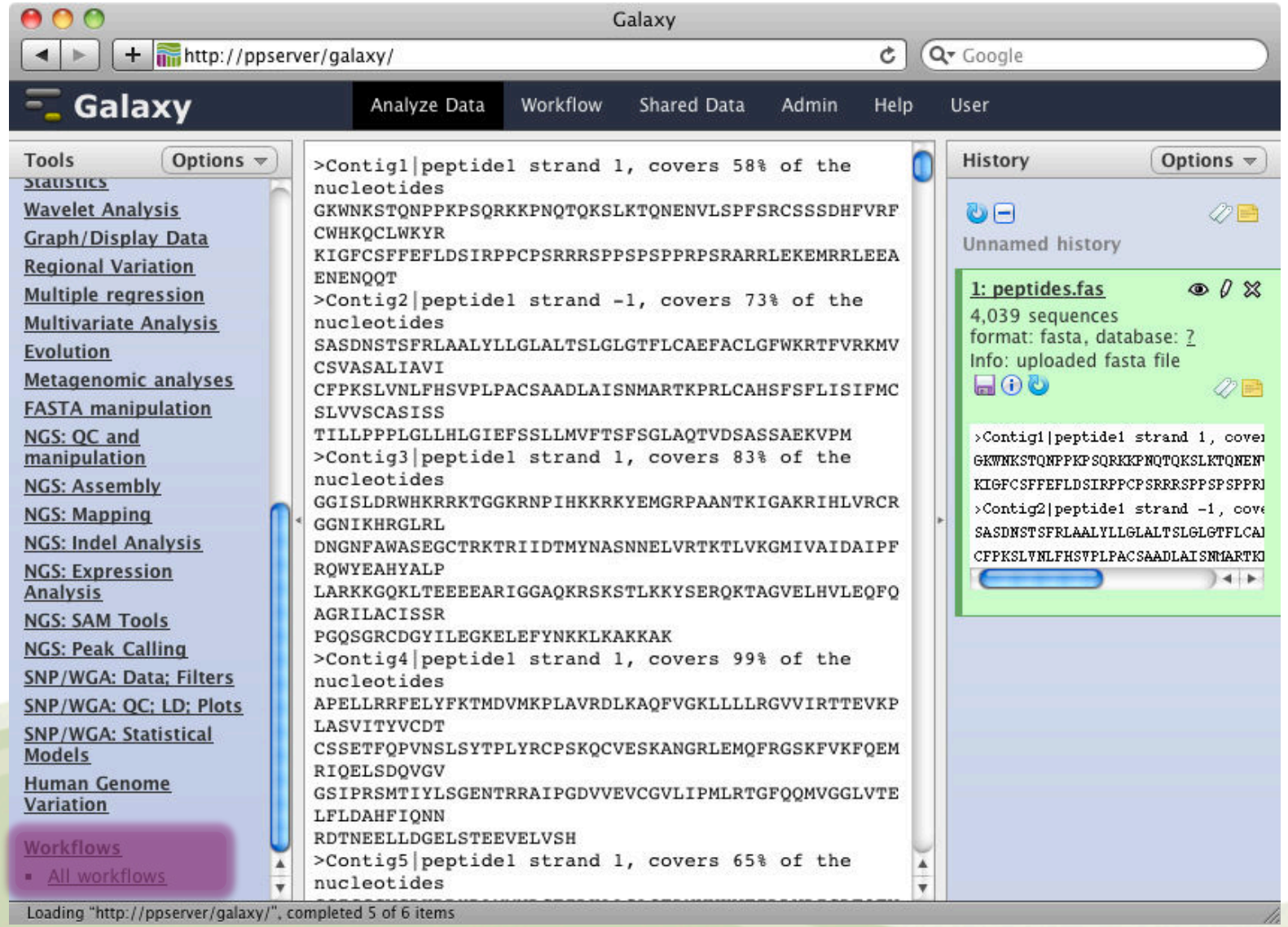
Step 1 – Upload FASTA file



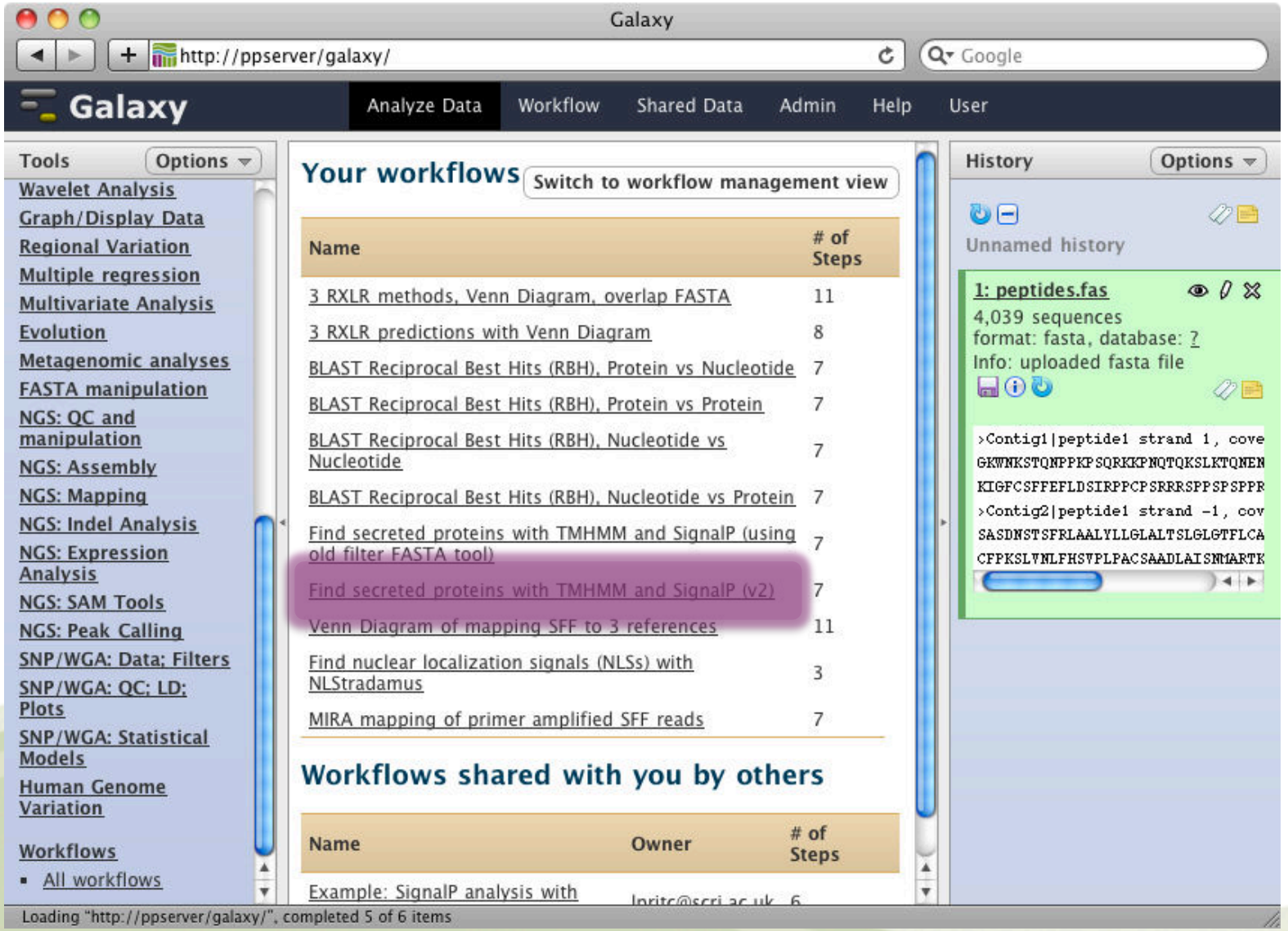
Step 1 – Upload FASTA file



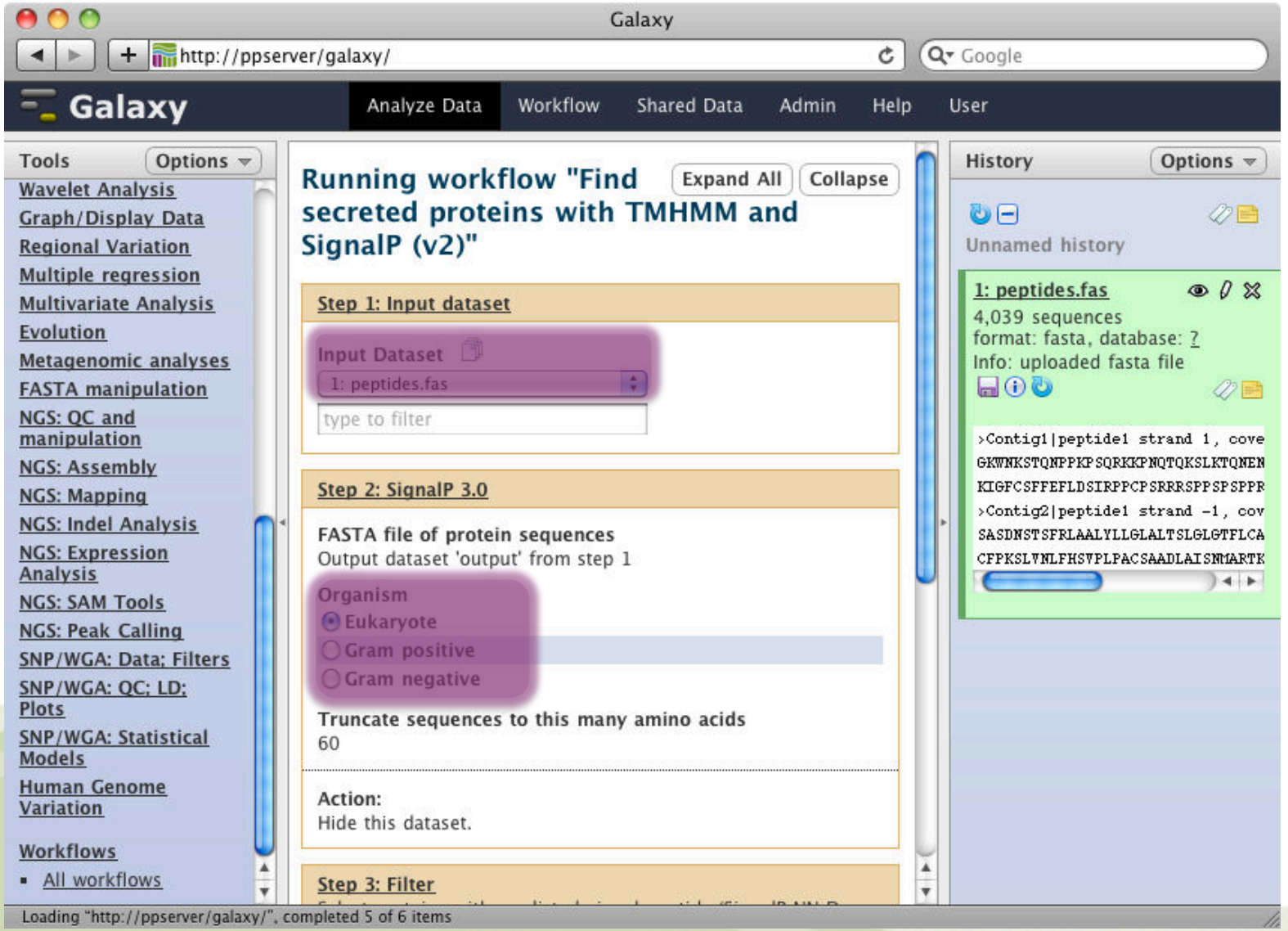
Step 2 – Run workflow



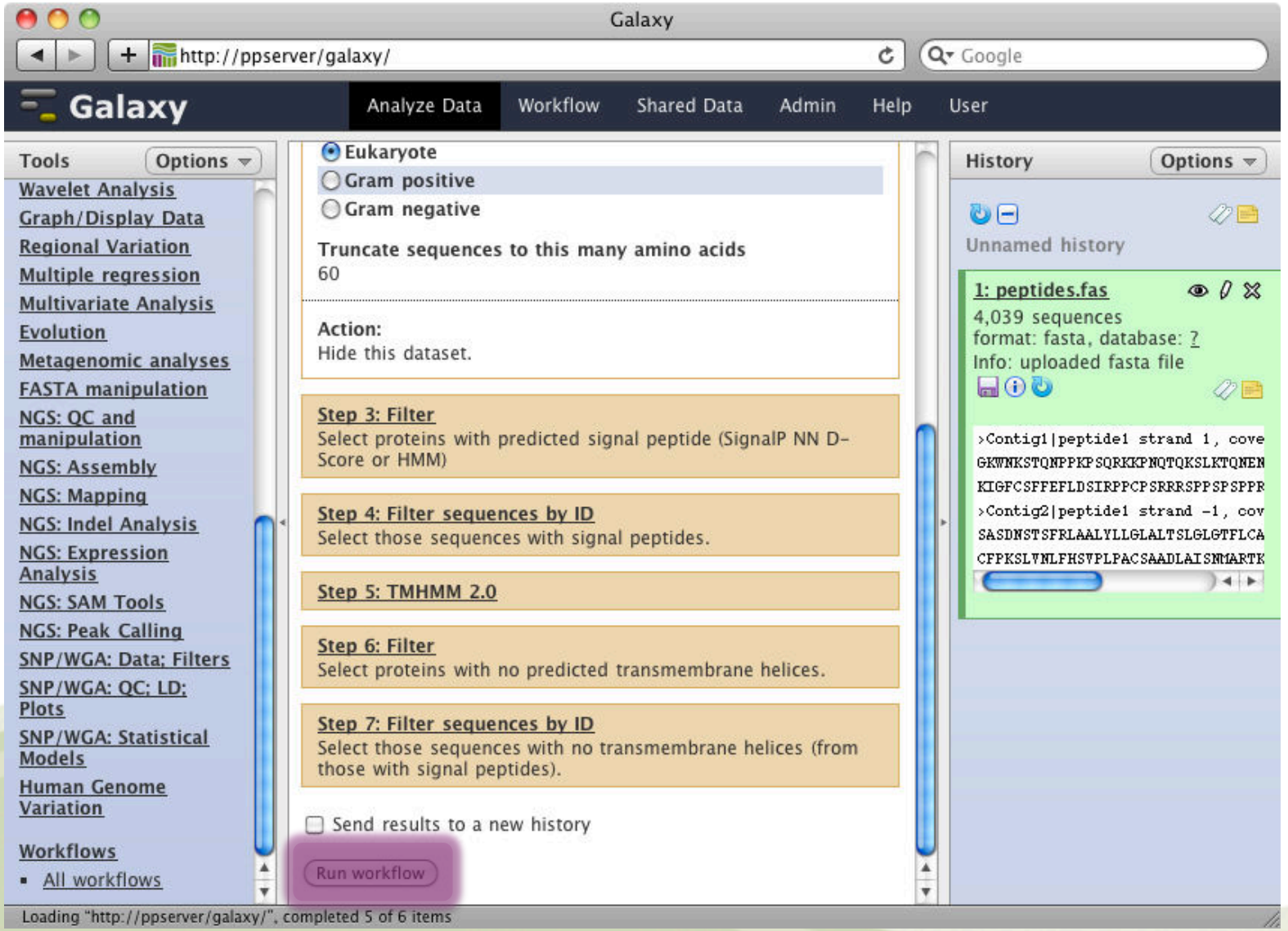
Step 2 – Run workflow



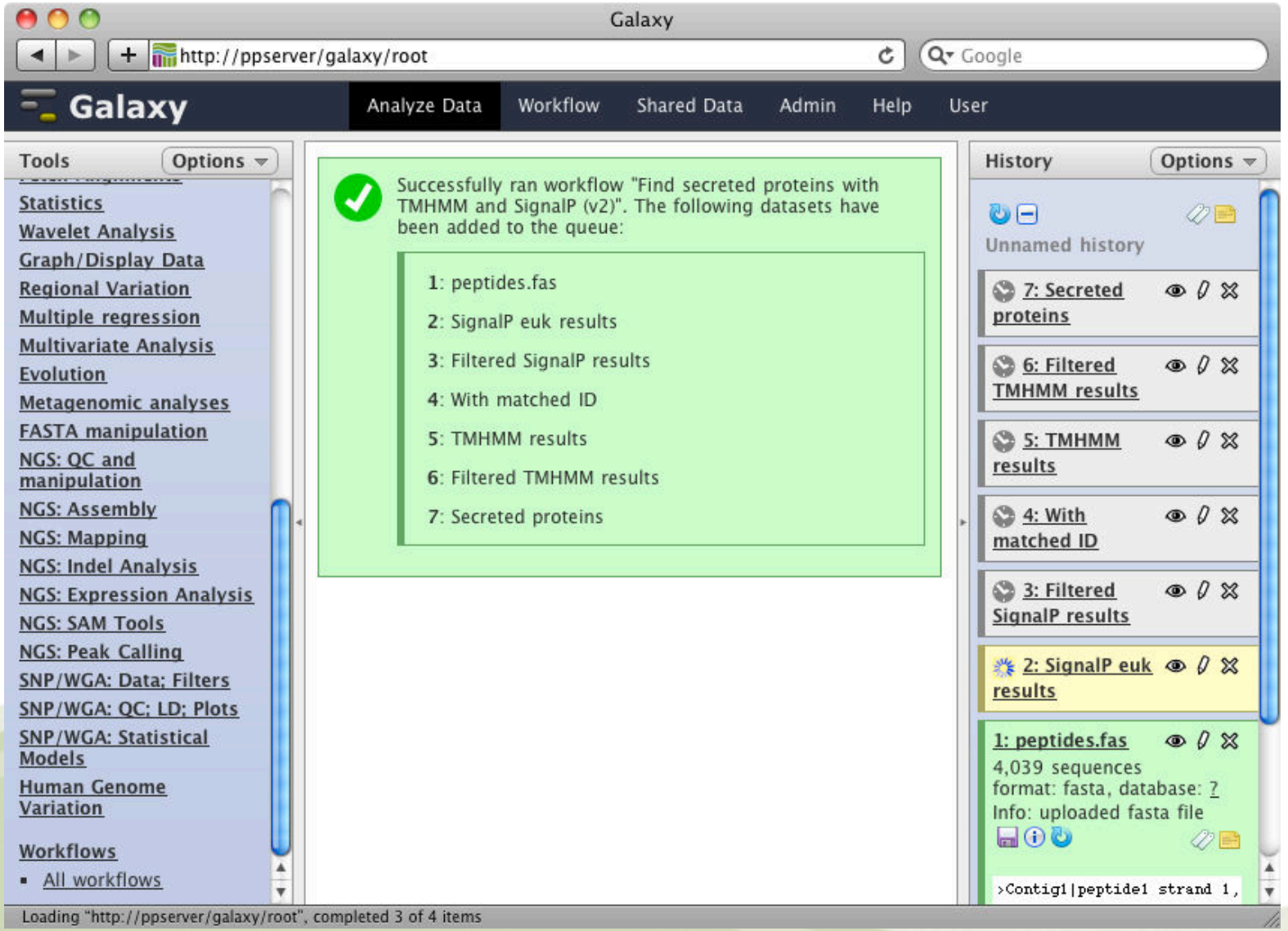
Step 2 – Run workflow



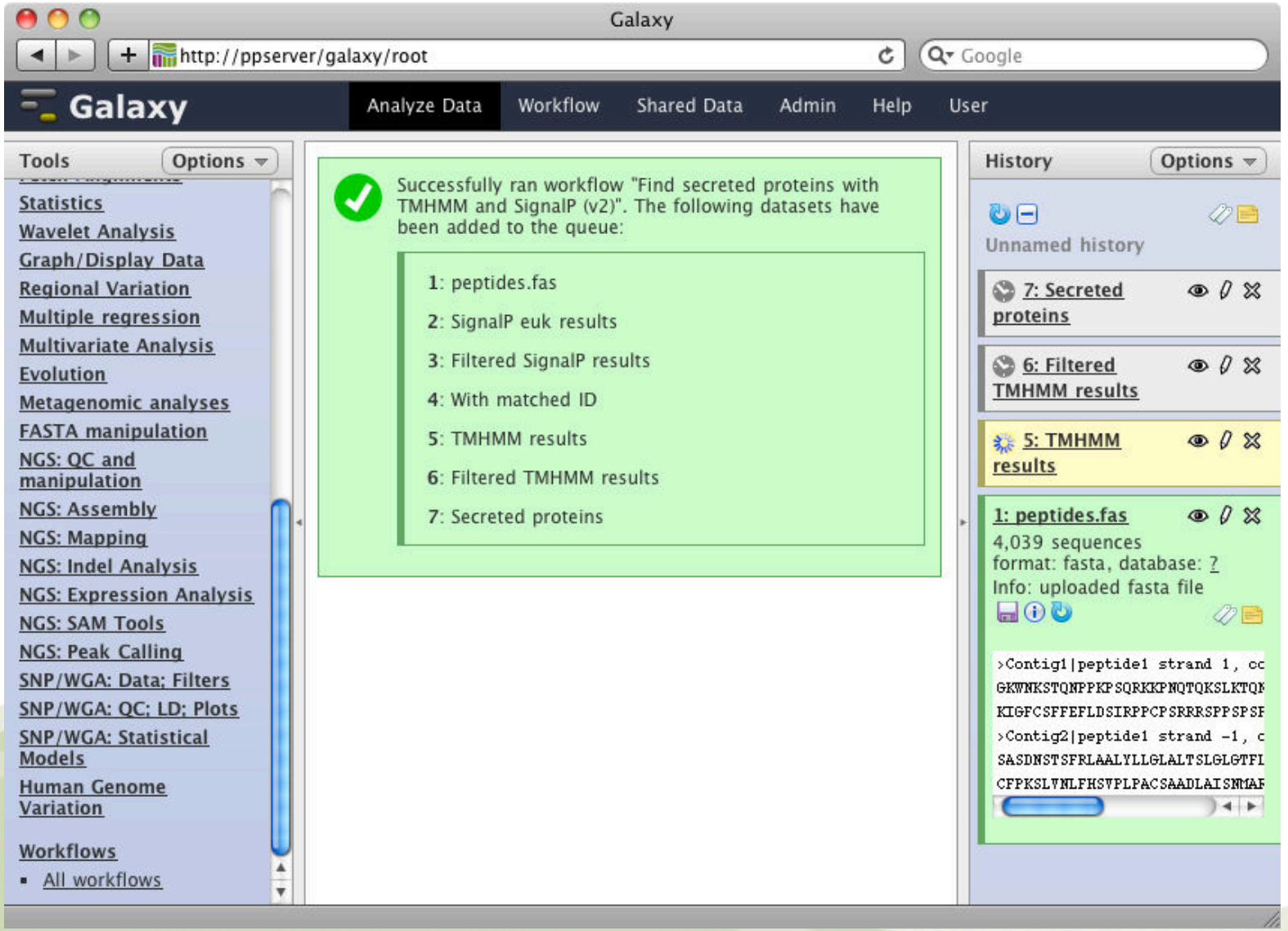
Step 2 – Run workflow



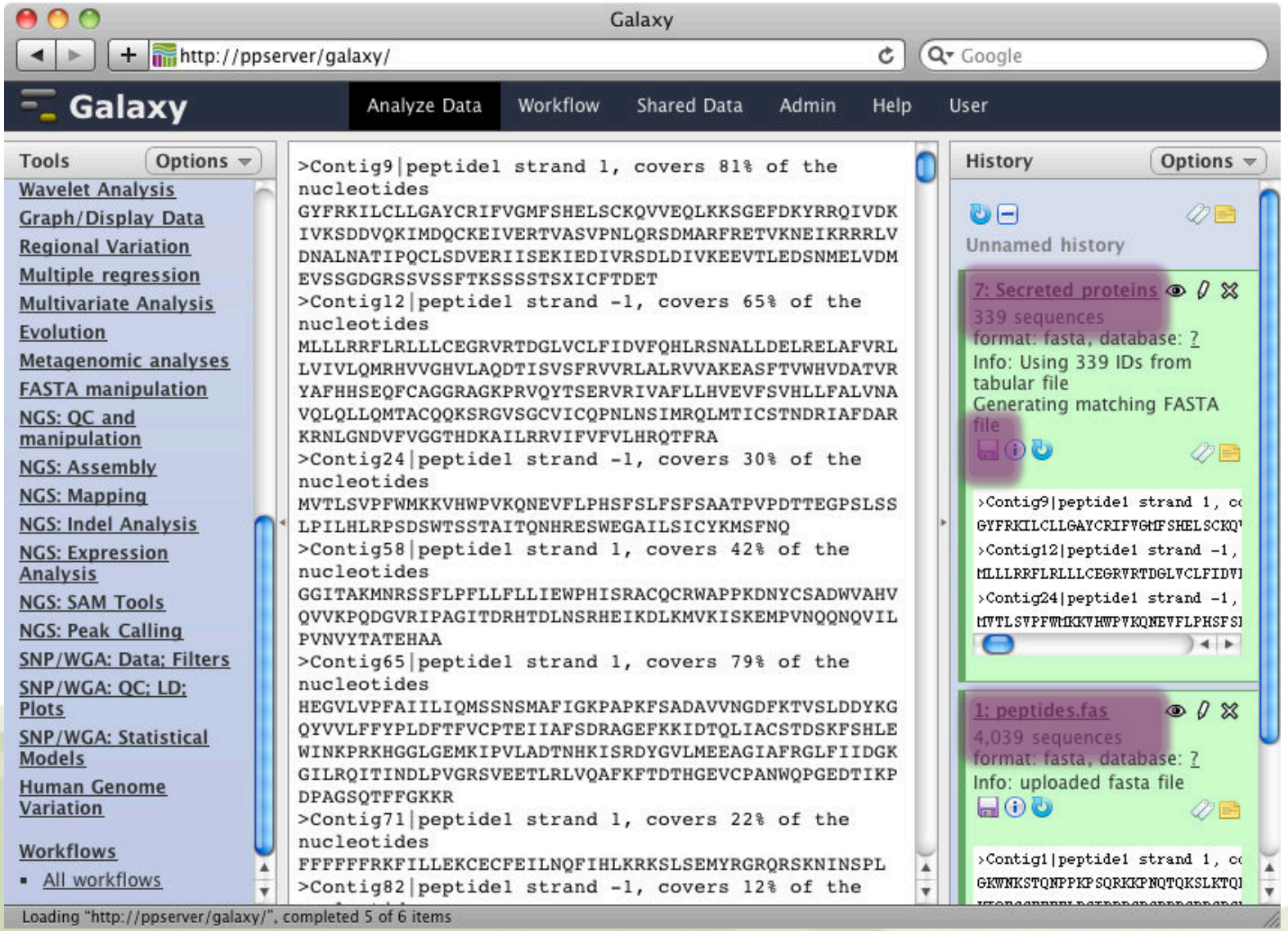
Step 2 – Run workflow



Step 2 – Run workflow



Step 3 – Get results



Example: Effector Protein Analysis

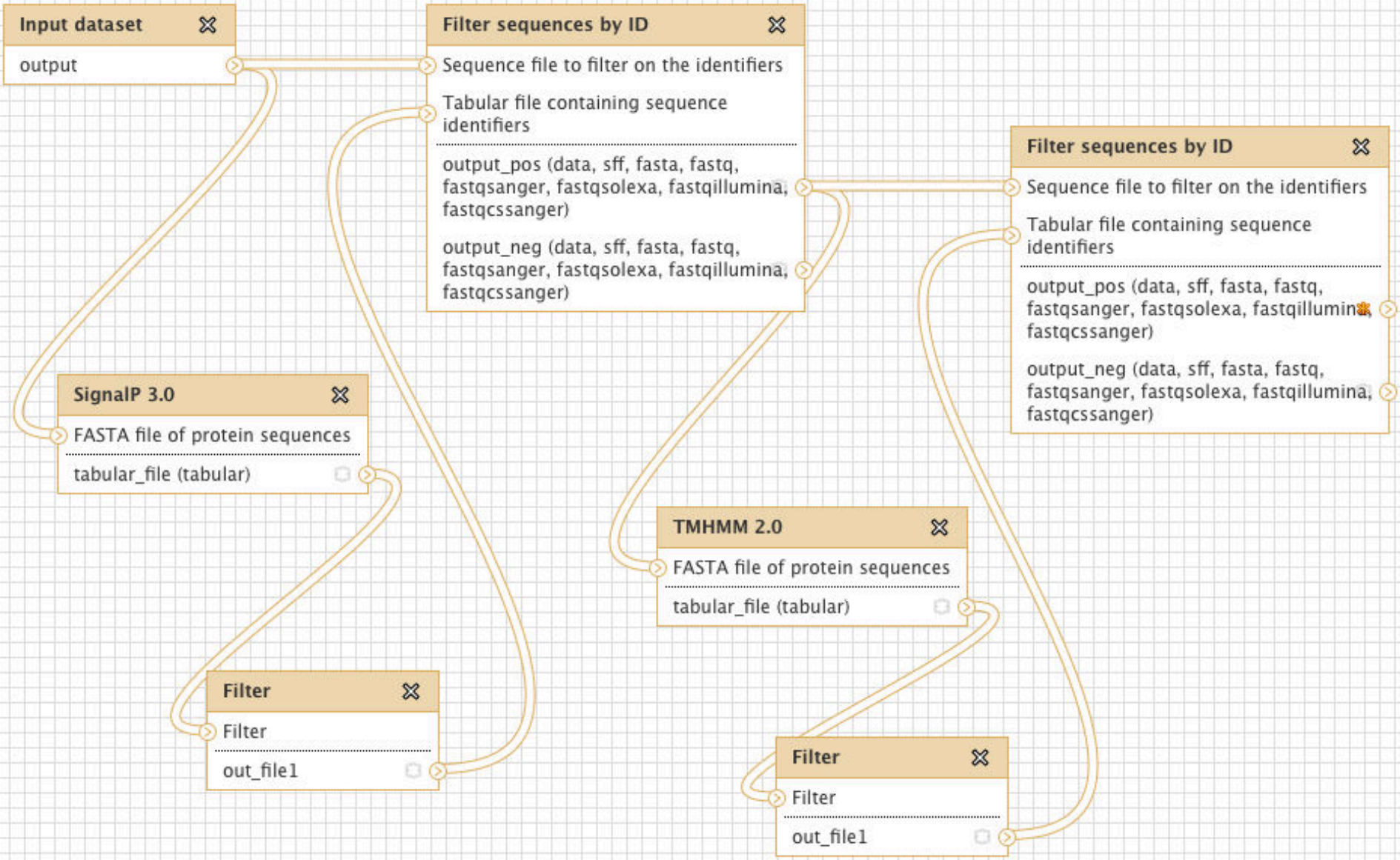
The James Hutton Institute

- To identify candidate effector genes:
  - Want a signal peptide (for export)
  - Don't want a transmembrane domain (not secreted)
- Now use Galaxy workflow calling:
  - SignalP 3.0
  - TMHMM 2.0
- I can share this workflow:
  - Easily available to anyone using our Server
  - If tools installed, could share with other Galaxy Servers
- Easy for our Biologists to apply workflow to their own data

See Jones *et al.* (2009), Molecular Plant Pathology & Kikuchi et al. (2011), PLoS Pathogens

Workflow Editor – Effector finding

The James Hutton Institute



Why Galaxy?

The James Hutton Institute

- Hi Peter, could you run a big BLAST job for me?
  - Everyone using standalone BLAST is not practical
  - Want a local BLAST web interface with multiple-query support
- Group *XXX* have just published the *YYY* genome – could you look for *ZZZ* proteins please?
  - With a suitable interface, lots of analyses are simple enough for non-bioinformaticians to run and interpret
- You remember that analysis we did last year? I want to do it again on this new genome
  - Running old scripts on new data is tedious
  - Workflows should be reproducible

Why Galaxy? Plus Points

The James Hutton Institute

- Don’t have to worry about local software installation
  - Mostly Windows here at JHI, most tools need Linux
- Uniform web based GUI for wrapped tools
  - Web interfaces are all different
  - Command line tools are scary
- Coupling tools together as sharable repeatable workflows
- Can share data files (better than email/shared drives)
- Open Source (extendable, free)
- Almost any tool can be added

Why Galaxy? Downsides

- Investment in training users
  - But interface is consistent across tools
- Bugs in Galaxy
  - Most issues arise when wrapping tools
- Missing tools
  - Have to invest time wrapping things we need

Protein Analysis Tools in our Galaxy

The James Hutton Institute

All take a FASTA protein file as input, return a tabular file.

- Sequence similarity
  - NCBI BLAST+
  - Blast2GO
- Transmembrane domains
  - TMHMM
- Signal Peptides/Motifs
  - SignalP
  - EffectiveT3
  - RXLR
- Nuclear Localisation
  - PredictNLS
  - NLStradamus
- Nucleolus Localisation
  - NoD
- Sub-cellular Localisation
  - PSORTB (*)
  - WoLF PSORT

(*) Konrad Paszkiewicz's wrapper from Galaxy Tool Shed, other wrappers written by us.

Observations from Wrapping Tools

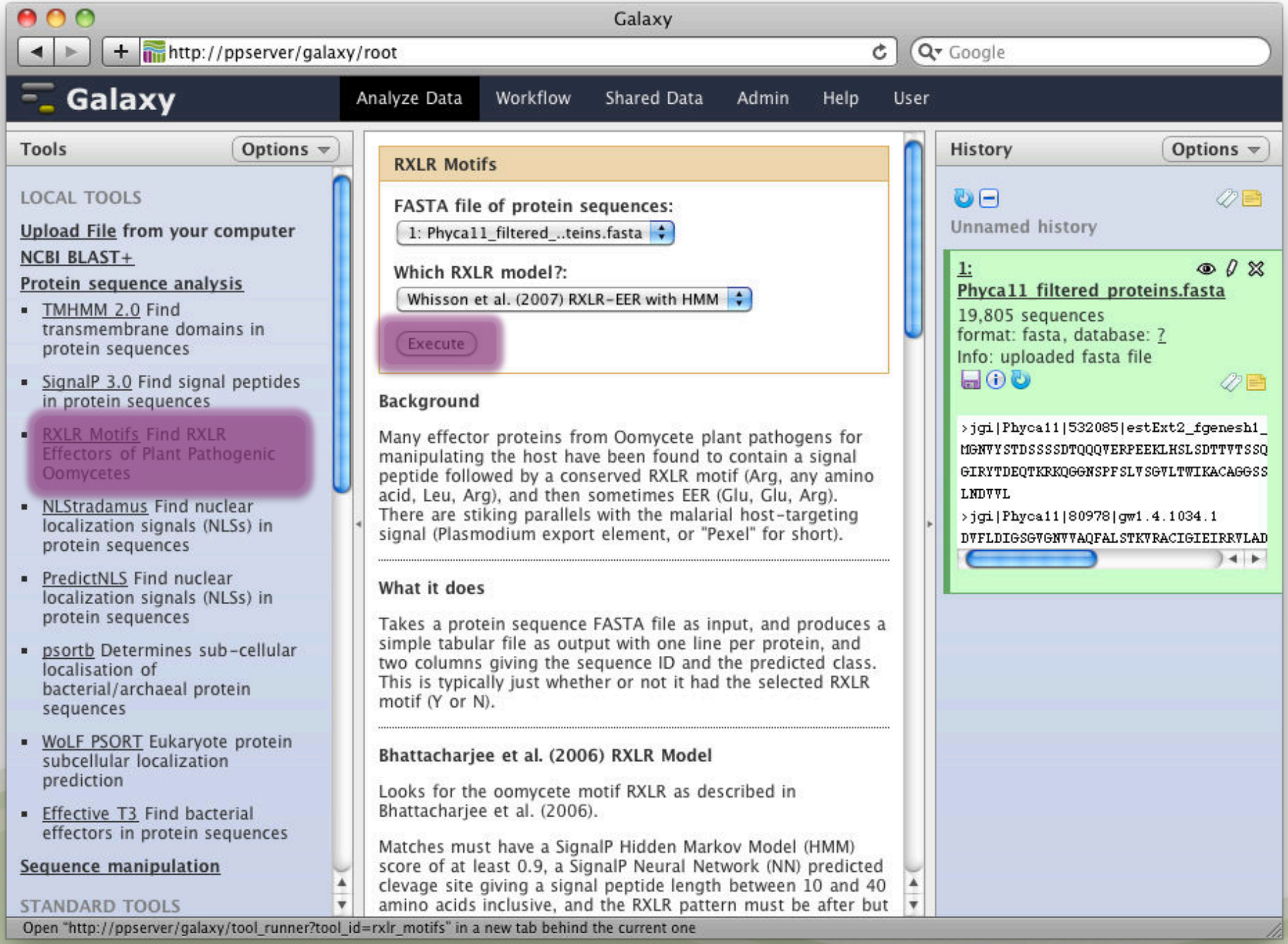
The James Hutton Institute

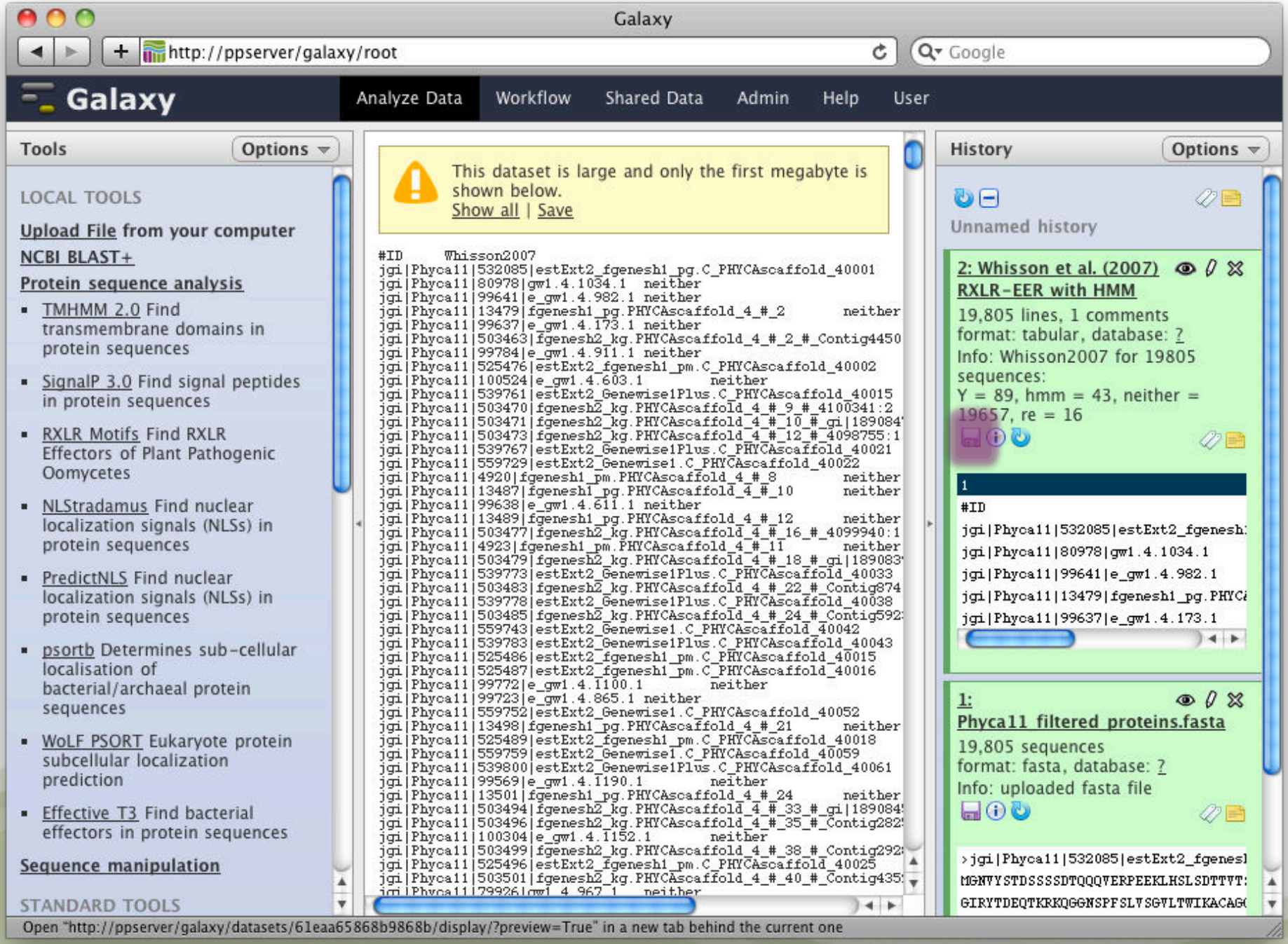
- Tabular output for Galaxy
  - Most tools' output needed reformatting
- Some tools are not threaded
  - Galaxy team working on "embarrassing parallel" case
- Interaction with tool authors can be productive and informative, and improve their tools
- To tool authors
  - Offer tabular output (if appropriate)
  - Better error handling (e.g. zero length sequences)

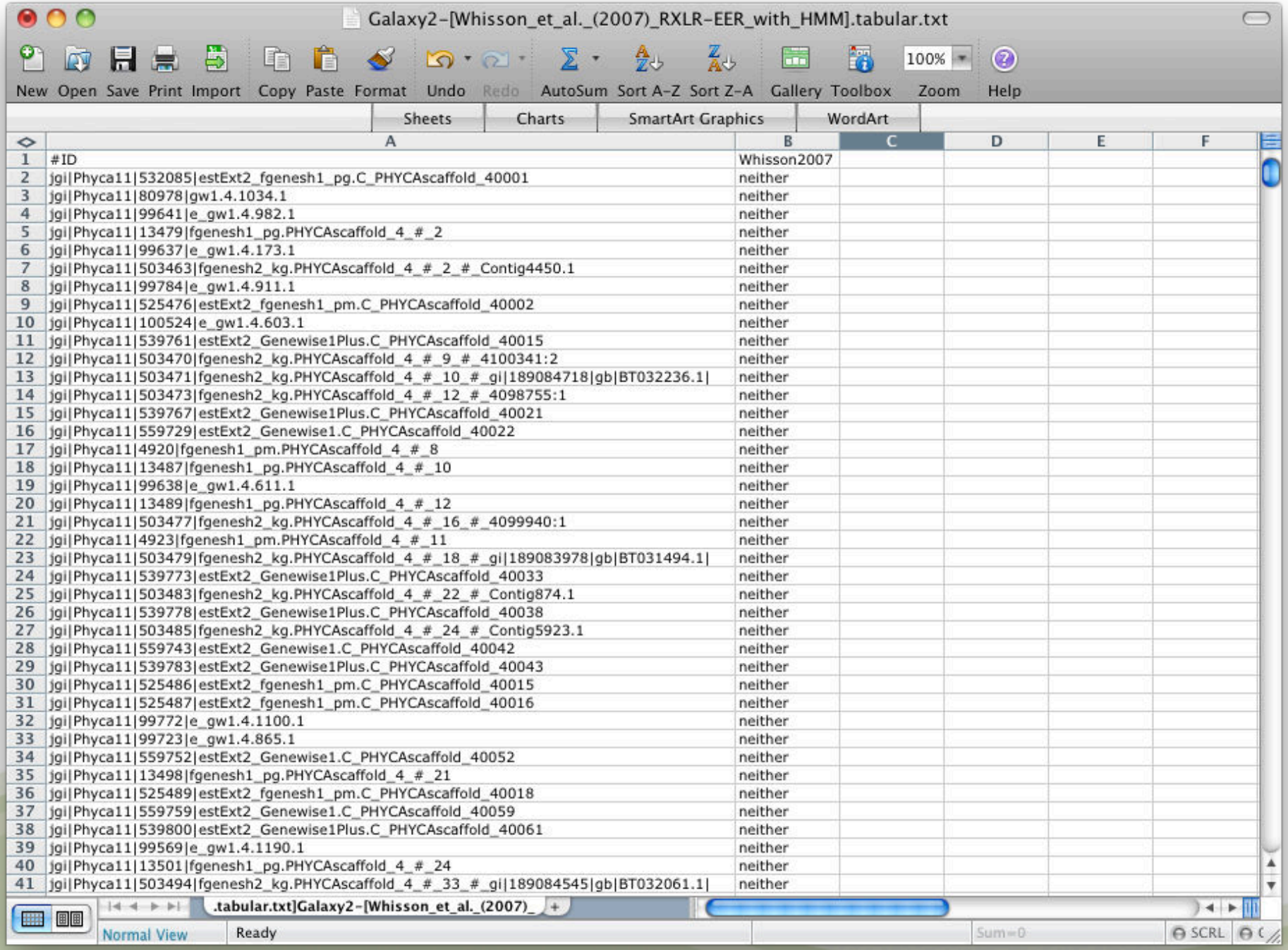
Workflow example - RXLR motifs

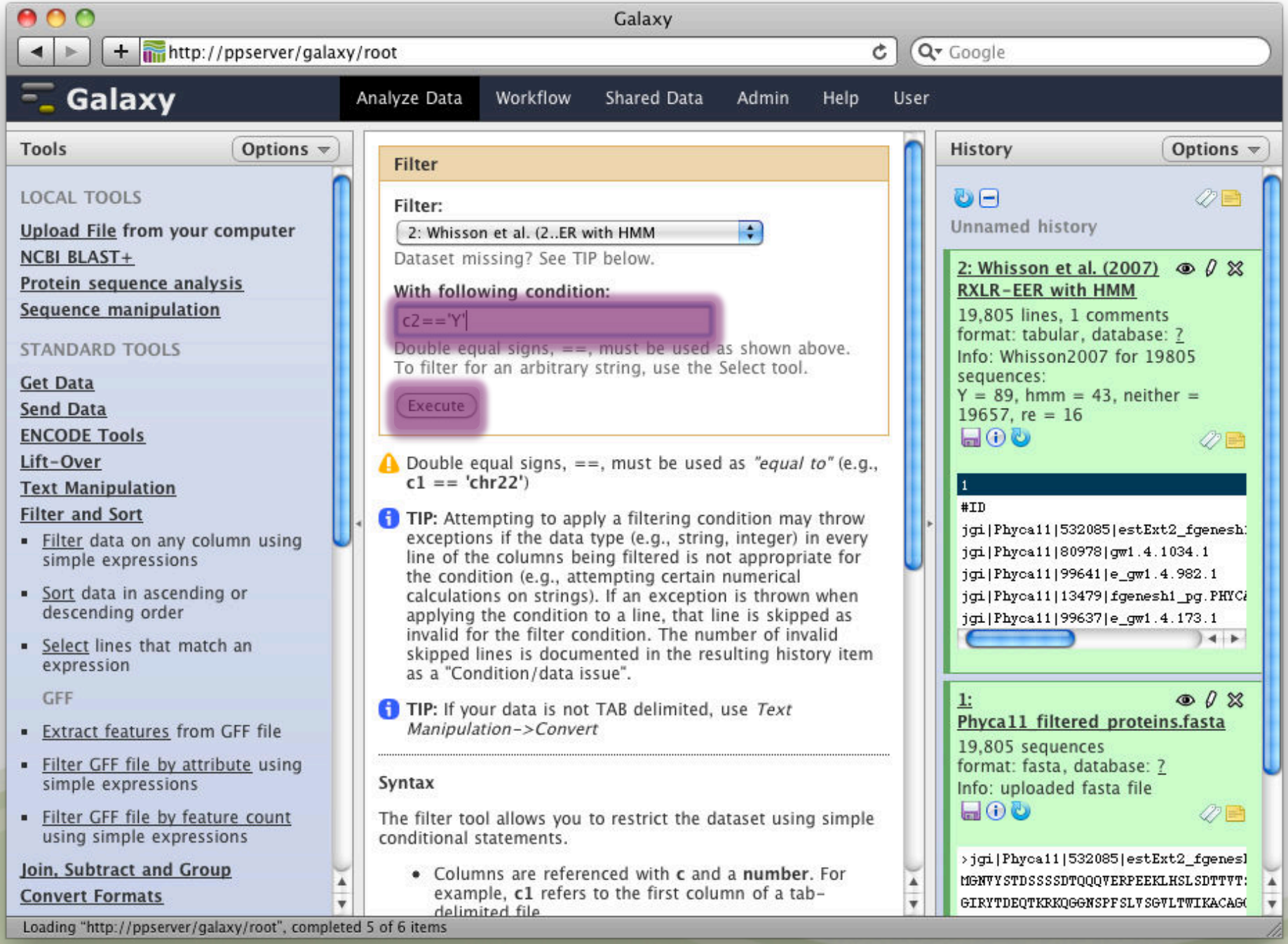
The James Hutton Institute

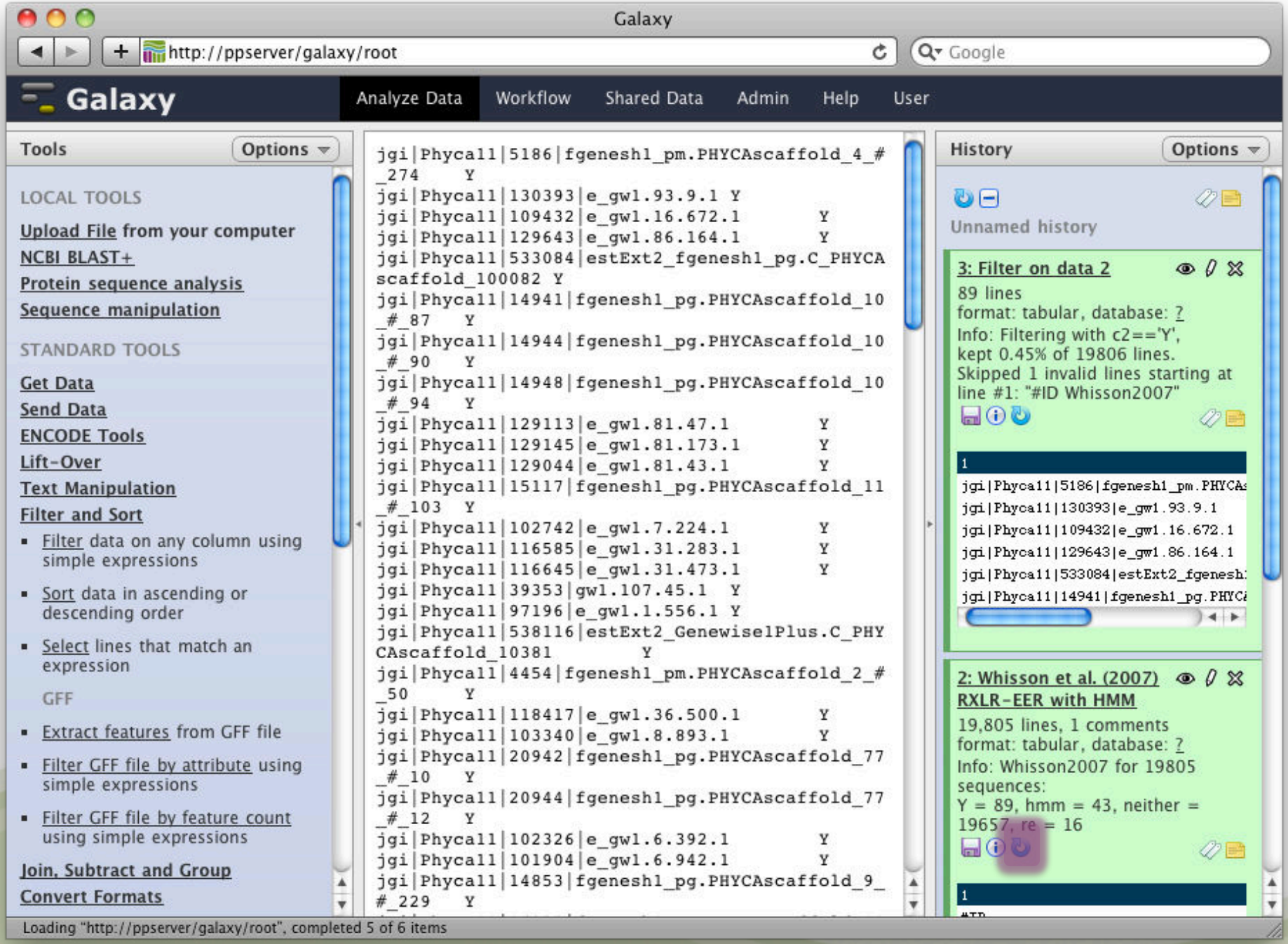
- Important translocation motif in oomycetes
- We have implemented three methods in Galaxy:
  - Bhattacharjee et al. (2006)
  - Win et al. (2007)
  - Whisson et al. (2007)
- Venn Diagram comparing the three methods

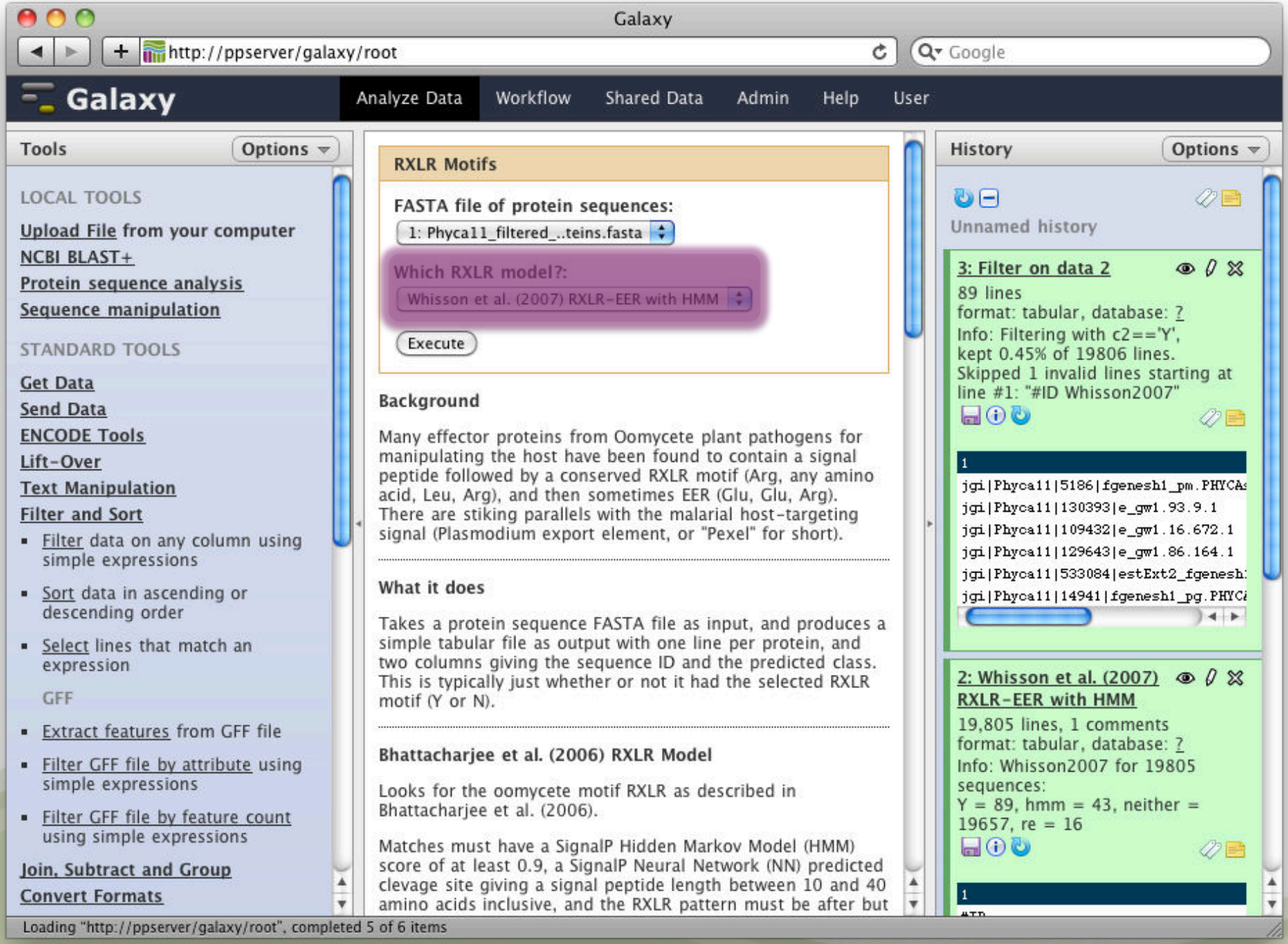

Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
LOCAL TOOLS
Upload File from your computer
NCBI BLAST+
Protein sequence analysis
TMHMM 2.0 Find transmembrane domains in protein sequences
SignalP 3.0 Find signal peptides in protein sequences
RXLR Motifs Find RXLR Effectors of Plant Pathogenic Oomycetes
NLStradamus Find nuclear localization signals (NLSs) in protein sequences
PredictNLS Find nuclear localization signals (NLSs) in protein sequences
psortb Determines sub-cellular localisation of bacterial/archaeal protein sequences
WoLF PSORT Eukaryote protein subcellular localization prediction
Effective T3 Find bacterial effectors in protein sequences
Sequence manipulation
STANDARD TOOLS
RXLR Motifs
FASTA file of protein sequences:
1: Phyca11_filtered_..teins.fasta
Which RXLR model?:
Whisson et al. (2007) RXLR-EER with HMM
Execute
Background
Many effector proteins from Oomycete plant pathogens for manipulating the host have been found to contain a signal peptide followed by a conserved RXLR motif (Arg, any amino acid, Leu, Arg), and then sometimes EER (Glu, Glu, Arg). There are stiking parallels with the malarial host-targeting signal (Plasmodium export element, or "Pexel" for short).
What it does
Takes a protein sequence FASTA file as input, and produces a simple tabular file as output with one line per protein, and two columns giving the sequence ID and the predicted class. This is typically just whether or not it had the selected RXLR motif (Y or N).
Bhattacharjee et al. (2006) RXLR Model
Looks for the oomycete motif RXLR as described in Bhattacharjee et al. (2006).
Matches must have a SignalP Hidden Markov Model (HMM) score of at least 0.9, a SignalP Neural Network (NN) predicted clevage site giving a signal peptide length between 10 and 40 amino acids inclusive, and the RXLR pattern must be after but
History
Options
Unnamed history
1:
Phyca11_filtered_proteins.fasta
19,805 sequences
format: fasta, database: ?
Info: uploaded fasta file
>jgi|Phyca11|532085|estExt2_fgenesh1_
MGNVYSTDSSSSDTQQQVERPEEKLHSLSDTTVTSSQ
GIRYTDEQTKRKQGGNSPFSLVSGVLTWIKACAGGSS
LNDVVL
>jgi|Phyca11|80978|gw1.4.1034.1
DVFLDIGSGVGNVVAQFALSTKVRACIGIEIRRVLAD
Open "http://ppserver/galaxy/tool_runner?tool_id=rxlr_motifs" in a new tab behind the current one

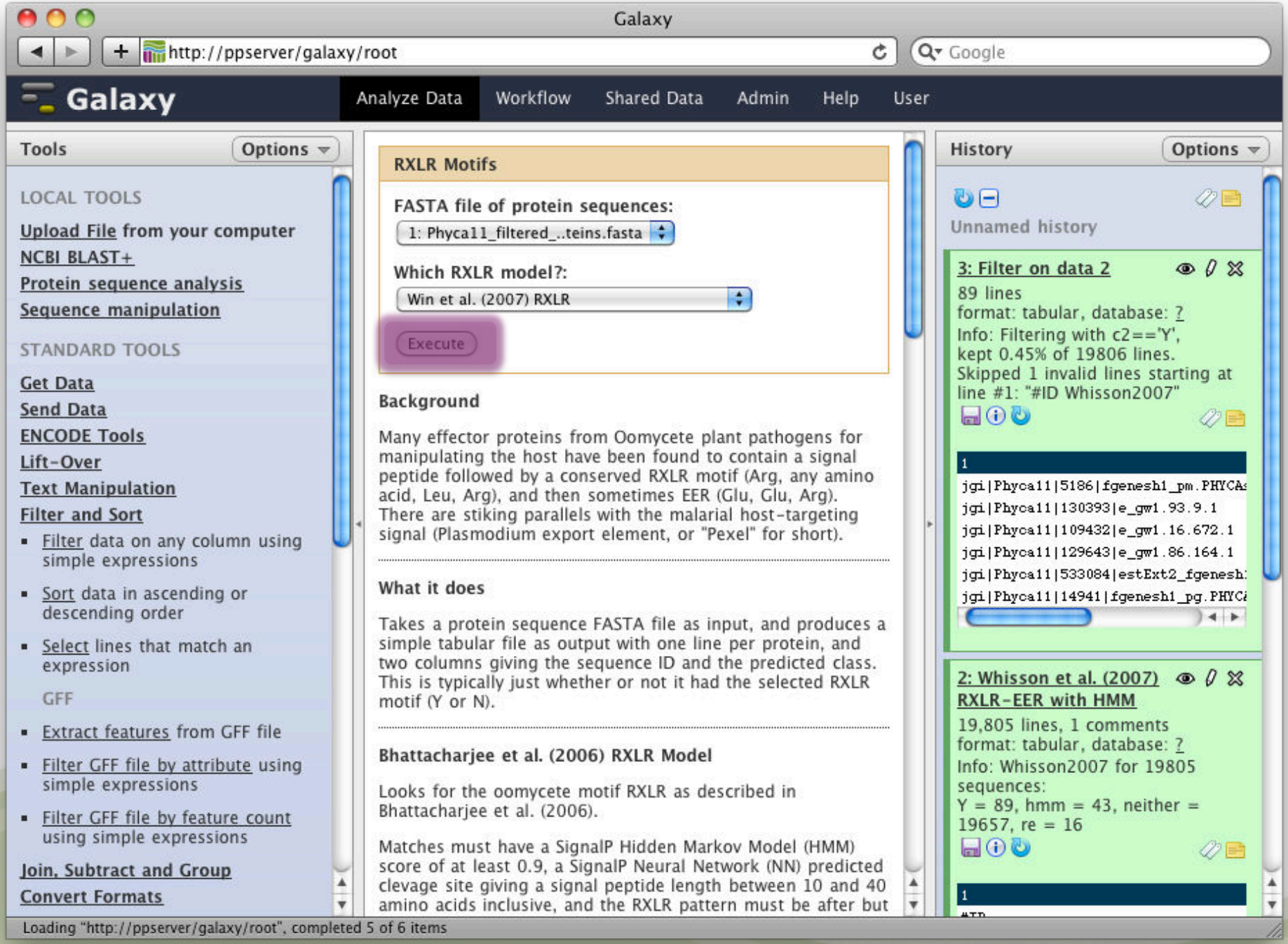

Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
LOCAL TOOLS
Upload File from your computer
NCBI BLAST+
Protein sequence analysis
TMHMM 2.0 Find transmembrane domains in protein sequences
SignalP 3.0 Find signal peptides in protein sequences
RXLR Motifs Find RXLR Effectors of Plant Pathogenic Oomycetes
NLStradamus Find nuclear localization signals (NLSs) in protein sequences
PredictNLS Find nuclear localization signals (NLSs) in protein sequences
psortb Determines sub-cellular localisation of bacterial/archaeal protein sequences
WoLF PSORT Eukaryote protein subcellular localization prediction
Effective T3 Find bacterial effectors in protein sequences
Sequence manipulation
STANDARD TOOLS
This dataset is large and only the first megabyte is shown below.
Show all | Save
#ID Whisson2007
jgi|Phyca11|532085|estExt2_fgenesh1_pg.C_PHYCAscaffold_40001
jgi|Phyca11|80978|gw1.4.1034.1 neither
jgi|Phyca11|99641|e_gw1.4.982.1 neither
jgi|Phyca11|13479|fgenesh1_pg.PHYCAscaffold_4_#_2 neither
jgi|Phyca11|99637|e_gw1.4.173.1 neither
jgi|Phyca11|503463|fgenesh2_kg.PHYCAscaffold_4_#_2_#_Contig4450
jgi|Phyca11|99784|e_gw1.4.911.1 neither
jgi|Phyca11|525476|estExt2_fgenesh1_pm.C_PHYCAscaffold_40002
jgi|Phyca11|100524|e_gw1.4.603.1 neither
jgi|Phyca11|539761|estExt2_Genewise1Plus.C_PHYCAscaffold_40015
jgi|Phyca11|503470|fgenesh2_kg.PHYCAscaffold_4_#_9_#_4100341:2
jgi|Phyca11|503471|fgenesh2_kg.PHYCAscaffold_4_#_10_#_gi|189084
jgi|Phyca11|503473|fgenesh2_kg.PHYCAscaffold_4_#_12_#_4098755:1
jgi|Phyca11|539767|estExt2_Genewise1Plus.C_PHYCAscaffold_40021
jgi|Phyca11|559729|estExt2_Genewise1.C_PHYCAscaffold_40022
jgi|Phyca11|4920|fgenesh1_pm.PHYCAscaffold_4_#_8 neither
jgi|Phyca11|13487|fgenesh1_pg.PHYCAscaffold_4_#_10 neither
jgi|Phyca11|99638|e_gw1.4.611.1 neither
jgi|Phyca11|13489|fgenesh1_pg.PHYCAscaffold_4_#_12 neither
jgi|Phyca11|503477|fgenesh2_kg.PHYCAscaffold_4_#_16_#_4099940:1
jgi|Phyca11|4923|fgenesh1_pm.PHYCAscaffold_4_#_11 neither
jgi|Phyca11|503479|fgenesh2_kg.PHYCAscaffold_4_#_18_#_gi|189083
jgi|Phyca11|539773|estExt2_Genewise1Plus.C_PHYCAscaffold_40033
jgi|Phyca11|503483|fgenesh2_kg.PHYCAscaffold_4_#_22_#_Contig874
jgi|Phyca11|539778|estExt2_Genewise1Plus.C_PHYCAscaffold_40038
jgi|Phyca11|503485|fgenesh2_kg.PHYCAscaffold_4_#_24_#_Contig592
jgi|Phyca11|559743|estExt2_Genewise1.C_PHYCAscaffold_40042
jgi|Phyca11|539783|estExt2_Genewise1Plus.C_PHYCAscaffold_40043
jgi|Phyca11|525486|estExt2_fgenesh1_pm.C_PHYCAscaffold_40015
jgi|Phyca11|525487|estExt2_fgenesh1_pm.C_PHYCAscaffold_40016
jgi|Phyca11|99772|e_gw1.4.1100.1 neither
jgi|Phyca11|99723|e_gw1.4.865.1 neither
jgi|Phyca11|559752|estExt2_Genewise1.C_PHYCAscaffold_40052
jgi|Phyca11|13498|fgenesh1_pg.PHYCAscaffold_4_#_21 neither
jgi|Phyca11|525489|estExt2_fgenesh1_pm.C_PHYCAscaffold_40018
jgi|Phyca11|559759|estExt2_Genewise1.C_PHYCAscaffold_40059
jgi|Phyca11|539800|estExt2_Genewise1Plus.C_PHYCAscaffold_40061
jgi|Phyca11|99569|e_gw1.4.1190.1 neither
jgi|Phyca11|13501|fgenesh1_pg.PHYCAscaffold_4_#_24 neither
jgi|Phyca11|503494|fgenesh2_kg.PHYCAscaffold_4_#_33_#_gi|189084
jgi|Phyca11|503496|fgenesh2_kg.PHYCAscaffold_4_#_35_#_Contig282
jgi|Phyca11|100304|e_gw1.4.1152.1 neither
jgi|Phyca11|503499|fgenesh2_kg.PHYCAscaffold_4_#_38_#_Contig292
jgi|Phyca11|525496|estExt2_fgenesh1_pm.C_PHYCAscaffold_40025
jgi|Phyca11|503501|fgenesh2_kg.PHYCAscaffold_4_#_40_#_Contig435
jgi|Phyca11|79926|gw1.4.962.1 neither
History
Options
Unnamed history
2: Whisson et al. (2007) RXLR-EER with HMM
19,805 lines, 1 comments
format: tabular, database: ?
Info: Whisson2007 for 19805 sequences:
Y = 89, hmm = 43, neither = 19657, re = 16
1
#ID
jgi|Phyca11|532085|estExt2_fgenesh
jgi|Phyca11|80978|gw1.4.1034.1
jgi|Phyca11|99641|e_gw1.4.982.1
jgi|Phyca11|13479|fgenesh1_pg.PHYC
jgi|Phyca11|99637|e_gw1.4.173.1
1:
Phyca11_filtered_proteins.fasta
19,805 sequences
format: fasta, database: ?
Info: uploaded fasta file
>jgi|Phyca11|532085|estExt2_fgenesh
MGNVYSTDSSSSDTQQQVERPEEKLHSLSDTTVT
GIRYTDEQTKRKQGGNSPFSLVSGVLTWIKACAG
Open "http://ppserver/galaxy/datasets/61eaa65868b9868b/display/?preview=True" in a new tab behind the current one


Galaxy2-[Whisson_et_al._(2007)_RXLR-EER_with_HMM].tabular.txt
New Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z Sort Z-A Gallery Toolbox 100% Zoom Help
Sheets Charts SmartArt Graphics WordArt
A B C D E F
1 #ID Whisson2007
2 jgi|Phyca11|532085|estExt2_fgenesh1_pg.C_PHYCAscaffold_40001 neither
3 jgi|Phyca11|80978|gw1.4.1034.1 neither
4 jgi|Phyca11|99641|e_gw1.4.982.1 neither
5 jgi|Phyca11|13479|fgenesh1_pg.PHYCAscaffold_4_#_2 neither
6 jgi|Phyca11|99637|e_gw1.4.173.1 neither
7 jgi|Phyca11|503463|fgenesh2_kg.PHYCAscaffold_4_#_2_#_Contig4450.1 neither
8 jgi|Phyca11|99784|e_gw1.4.911.1 neither
9 jgi|Phyca11|525476|estExt2_fgenesh1_pm.C_PHYCAscaffold_40002 neither
10 jgi|Phyca11|100524|e_gw1.4.603.1 neither
11 jgi|Phyca11|539761|estExt2_Genewise1Plus.C_PHYCAscaffold_40015 neither
12 jgi|Phyca11|503470|fgenesh2_kg.PHYCAscaffold_4_#_9_#_4100341:2 neither
13 jgi|Phyca11|503471|fgenesh2_kg.PHYCAscaffold_4_#_10_#_gi|189084718|gb|BT032236.1| neither
14 jgi|Phyca11|503473|fgenesh2_kg.PHYCAscaffold_4_#_12_#_4098755:1 neither
15 jgi|Phyca11|539767|estExt2_Genewise1Plus.C_PHYCAscaffold_40021 neither
16 jgi|Phyca11|559729|estExt2_Genewise1.C_PHYCAscaffold_40022 neither
17 jgi|Phyca11|4920|fgenesh1_pm.PHYCAscaffold_4_#_8 neither
18 jgi|Phyca11|13487|fgenesh1_pg.PHYCAscaffold_4_#_10 neither
19 jgi|Phyca11|99638|e_gw1.4.611.1 neither
20 jgi|Phyca11|13489|fgenesh1_pg.PHYCAscaffold_4_#_12 neither
21 jgi|Phyca11|503477|fgenesh2_kg.PHYCAscaffold_4_#_16_#_4099940:1 neither
22 jgi|Phyca11|4923|fgenesh1_pm.PHYCAscaffold_4_#_11 neither
23 jgi|Phyca11|503479|fgenesh2_kg.PHYCAscaffold_4_#_18_#_gi|189083978|gb|BT031494.1| neither
24 jgi|Phyca11|539773|estExt2_Genewise1Plus.C_PHYCAscaffold_40033 neither
25 jgi|Phyca11|503483|fgenesh2_kg.PHYCAscaffold_4_#_22_#_Contig874.1 neither
26 jgi|Phyca11|539778|estExt2_Genewise1Plus.C_PHYCAscaffold_40038 neither
27 jgi|Phyca11|503485|fgenesh2_kg.PHYCAscaffold_4_#_24_#_Contig5923.1 neither
28 jgi|Phyca11|559743|estExt2_Genewise1.C_PHYCAscaffold_40042 neither
29 jgi|Phyca11|539783|estExt2_Genewise1Plus.C_PHYCAscaffold_40043 neither
30 jgi|Phyca11|525486|estExt2_fgenesh1_pm.C_PHYCAscaffold_40015 neither
31 jgi|Phyca11|525487|estExt2_fgenesh1_pm.C_PHYCAscaffold_40016 neither
32 jgi|Phyca11|99772|e_gw1.4.1100.1 neither
33 jgi|Phyca11|99723|e_gw1.4.865.1 neither
34 jgi|Phyca11|559752|estExt2_Genewise1.C_PHYCAscaffold_40052 neither
35 jgi|Phyca11|13498|fgenesh1_pg.PHYCAscaffold_4_#_21 neither
36 jgi|Phyca11|525489|estExt2_fgenesh1_pm.C_PHYCAscaffold_40018 neither
37 jgi|Phyca11|559759|estExt2_Genewise1.C_PHYCAscaffold_40059 neither
38 jgi|Phyca11|539800|estExt2_Genewise1Plus.C_PHYCAscaffold_40061 neither
39 jgi|Phyca11|99569|e_gw1.4.1190.1 neither
40 jgi|Phyca11|13501|fgenesh1_pg.PHYCAscaffold_4_#_24 neither
41 jgi|Phyca11|503494|fgenesh2_kg.PHYCAscaffold_4_#_33_#_gi|189084545|gb|BT032061.1| neither
.tabular.txt]Galaxy2-[Whisson_et_al._(2007)_
Normal View
Ready
Sum=0
SCRL


Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
LOCAL TOOLS
Upload File from your computer
NCBI BLAST+
Protein sequence analysis
Sequence manipulation
STANDARD TOOLS
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF file
Filter GFF file by attribute using simple expressions
Filter GFF file by feature count using simple expressions
Join, Subtract and Group
Convert Formats
Filter
Filter:
2: Whisson et al. (2..ER with HMM
Dataset missing? See TIP below.
With following condition:
c2=='Y'
Double equal signs, ==, must be used as shown above.
To filter for an arbitrary string, use the Select tool.
Execute
Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')
TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".
TIP: If your data is not TAB delimited, use Text Manipulation->Convert
Syntax
The filter tool allows you to restrict the dataset using simple conditional statements.
Columns are referenced with c and a number. For example, c1 refers to the first column of a tab-delimited file
History
Options
Unnamed history
2: Whisson et al. (2007) RXLR-EER with HMM
19,805 lines, 1 comments
format: tabular, database: ?
Info: Whisson2007 for 19805 sequences:
Y = 89, hmm = 43, neither = 19657, re = 16
1
#ID
jgi|Phyca11|532085|estExt2_fgenesh1
jgi|Phyca11|80978|gw1.4.1034.1
jgi|Phyca11|99641|e_gw1.4.982.1
jgi|Phyca11|13479|fgenesh1_pg.PHYCA
jgi|Phyca11|99637|e_gw1.4.173.1
1: Phyca11_filtered_proteins.fasta
19,805 sequences
format: fasta, database: ?
Info: uploaded fasta file
>jgi|Phyca11|532085|estExt2_fgenesl
MGNVYSTDSSSSDTQQQVERPEEKLHSLSDTTVT
GIRYTDEQTKRKQGGNSPFSLVSGVLTWIKACAG
Loading "http://ppserver/galaxy/root", completed 5 of 6 items


Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
LOCAL TOOLS
Upload File from your computer
NCBI BLAST+
Protein sequence analysis
Sequence manipulation
STANDARD TOOLS
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF file
Filter GFF file by attribute using simple expressions
Filter GFF file by feature count using simple expressions
Join, Subtract and Group
Convert Formats
jgi|Phycal1|5186|fgeneshl_pm.PHYCAscaffold_4_#_274 Y
jgi|Phycal1|130393|e_gw1.93.9.1 Y
jgi|Phycal1|109432|e_gw1.16.672.1 Y
jgi|Phycal1|129643|e_gw1.86.164.1 Y
jgi|Phycal1|533084|estExt2_fgeneshl_pg.C_PHYCAscaffold_100082 Y
jgi|Phycal1|14941|fgeneshl_pg.PHYCAscaffold_10_#_87 Y
jgi|Phycal1|14944|fgeneshl_pg.PHYCAscaffold_10_#_90 Y
jgi|Phycal1|14948|fgeneshl_pg.PHYCAscaffold_10_#_94 Y
jgi|Phycal1|129113|e_gw1.81.47.1 Y
jgi|Phycal1|129145|e_gw1.81.173.1 Y
jgi|Phycal1|129044|e_gw1.81.43.1 Y
jgi|Phycal1|15117|fgeneshl_pg.PHYCAscaffold_11_#_103 Y
jgi|Phycal1|102742|e_gw1.7.224.1 Y
jgi|Phycal1|116585|e_gw1.31.283.1 Y
jgi|Phycal1|116645|e_gw1.31.473.1 Y
jgi|Phycal1|39353|gw1.107.45.1 Y
jgi|Phycal1|97196|e_gw1.1.556.1 Y
jgi|Phycal1|538116|estExt2_GenewiselPlus.C_PHYCAscaffold_10381 Y
jgi|Phycal1|4454|fgeneshl_pm.PHYCAscaffold_2_#_50 Y
jgi|Phycal1|118417|e_gw1.36.500.1 Y
jgi|Phycal1|103340|e_gw1.8.893.1 Y
jgi|Phycal1|20942|fgeneshl_pg.PHYCAscaffold_77_#_10 Y
jgi|Phycal1|20944|fgeneshl_pg.PHYCAscaffold_77_#_12 Y
jgi|Phycal1|102326|e_gw1.6.392.1 Y
jgi|Phycal1|101904|e_gw1.6.942.1 Y
jgi|Phycal1|14853|fgeneshl_pg.PHYCAscaffold_9_#_229 Y
History
Options
Unnamed history
3: Filter on data 2
89 lines
format: tabular, database: ?
Info: Filtering with c2=='Y', kept 0.45% of 19806 lines.
Skipped 1 invalid lines starting at line #1: "#ID Whisson2007"
1
jgi|Phyca11|5186|fgeneshl_pm.PHYCAs
jgi|Phyca11|130393|e_gw1.93.9.1
jgi|Phyca11|109432|e_gw1.16.672.1
jgi|Phyca11|129643|e_gw1.86.164.1
jgi|Phyca11|533084|estExt2_fgenesh
jgi|Phyca11|14941|fgeneshl_pg.PHYCA
2: Whisson et al. (2007) RXLR-EER with HMM
19,805 lines, 1 comments
format: tabular, database: ?
Info: Whisson2007 for 19805 sequences:
Y = 89, hmm = 43, neither = 19657, re = 16
1
Loading "http://ppserver/galaxy/root", completed 5 of 6 items

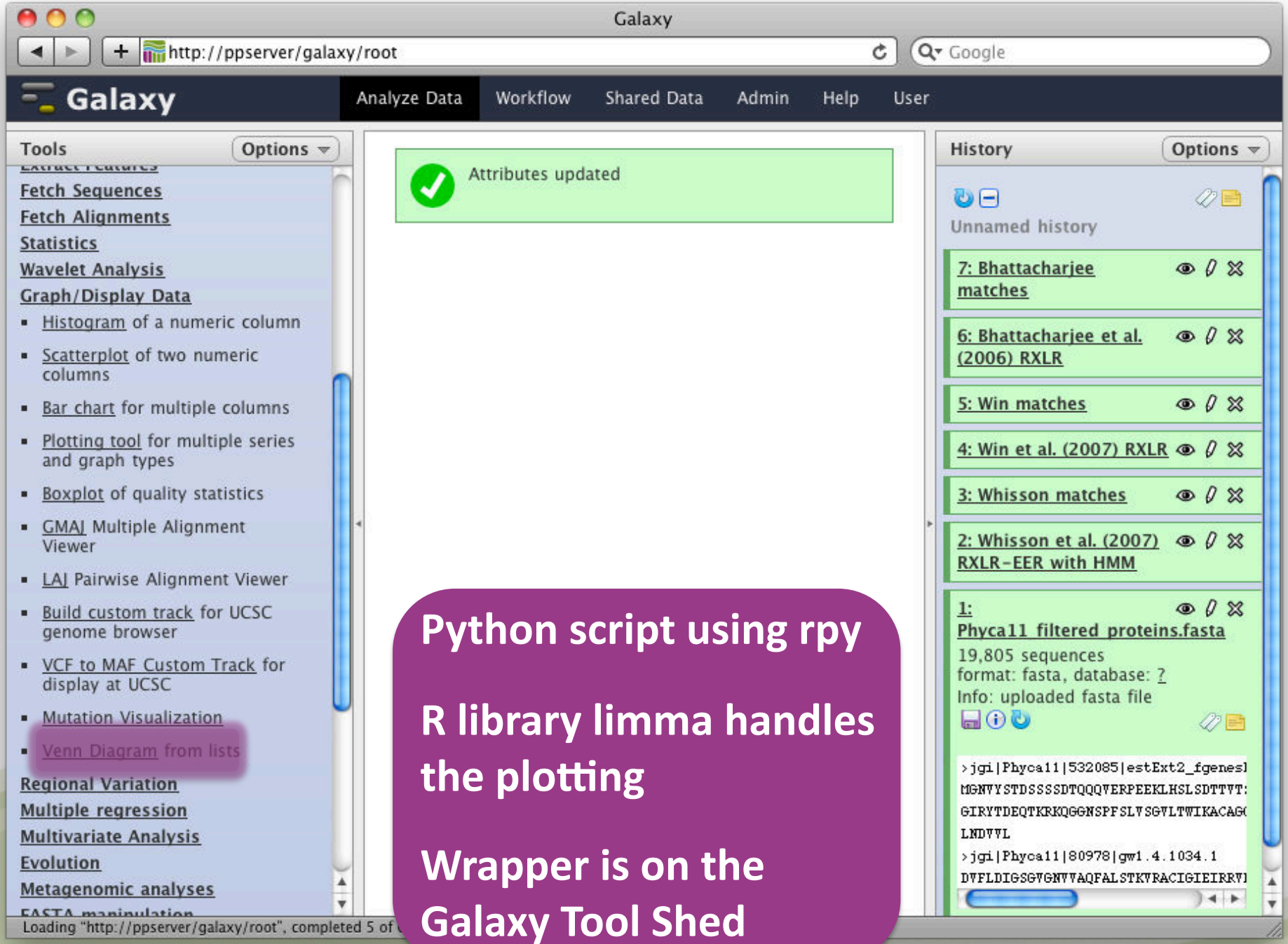

Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
LOCAL TOOLS
Upload File from your computer
NCBI BLAST+
Protein sequence analysis
Sequence manipulation
STANDARD TOOLS
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF file
Filter GFF file by attribute using simple expressions
Filter GFF file by feature count using simple expressions
Join, Subtract and Group
Convert Formats
RXLR Motifs
FASTA file of protein sequences:
1: Phyca11_filtered_..teins.fasta
Which RXLR model?:
Whisson et al. (2007) RXLR-EER with HMM
Execute
Background
Many effector proteins from Oomycete plant pathogens for manipulating the host have been found to contain a signal peptide followed by a conserved RXLR motif (Arg, any amino acid, Leu, Arg), and then sometimes EER (Glu, Glu, Arg). There are stiking parallels with the malarial host-targeting signal (Plasmodium export element, or "Pexel" for short).
What it does
Takes a protein sequence FASTA file as input, and produces a simple tabular file as output with one line per protein, and two columns giving the sequence ID and the predicted class. This is typically just whether or not it had the selected RXLR motif (Y or N).
Bhattacharjee et al. (2006) RXLR Model
Looks for the oomycete motif RXLR as described in Bhattacharjee et al. (2006).
Matches must have a SignalP Hidden Markov Model (HMM) score of at least 0.9, a SignalP Neural Network (NN) predicted clevage site giving a signal peptide length between 10 and 40 amino acids inclusive, and the RXLR pattern must be after but
History
Options
Unnamed history
3: Filter on data 2
89 lines
format: tabular, database: ?
Info: Filtering with c2=='Y', kept 0.45% of 19806 lines.
Skipped 1 invalid lines starting at line #1: "#ID Whisson2007"
1
jgi|Phyca11|5186|fgenesh1_pm.PHYCA
jgi|Phyca11|130393|e_gw1.93.9.1
jgi|Phyca11|109432|e_gw1.16.672.1
jgi|Phyca11|129643|e_gw1.86.164.1
jgi|Phyca11|533084|estExt2_fgenesh
jgi|Phyca11|14941|fgenesh1_pg.PHYC
2: Whisson et al. (2007) RXLR-EER with HMM
19,805 lines, 1 comments
format: tabular, database: ?
Info: Whisson2007 for 19805 sequences:
Y = 89, hmm = 43, neither = 19657, re = 16
1
Loading "http://ppserver/galaxy/root", completed 5 of 6 items

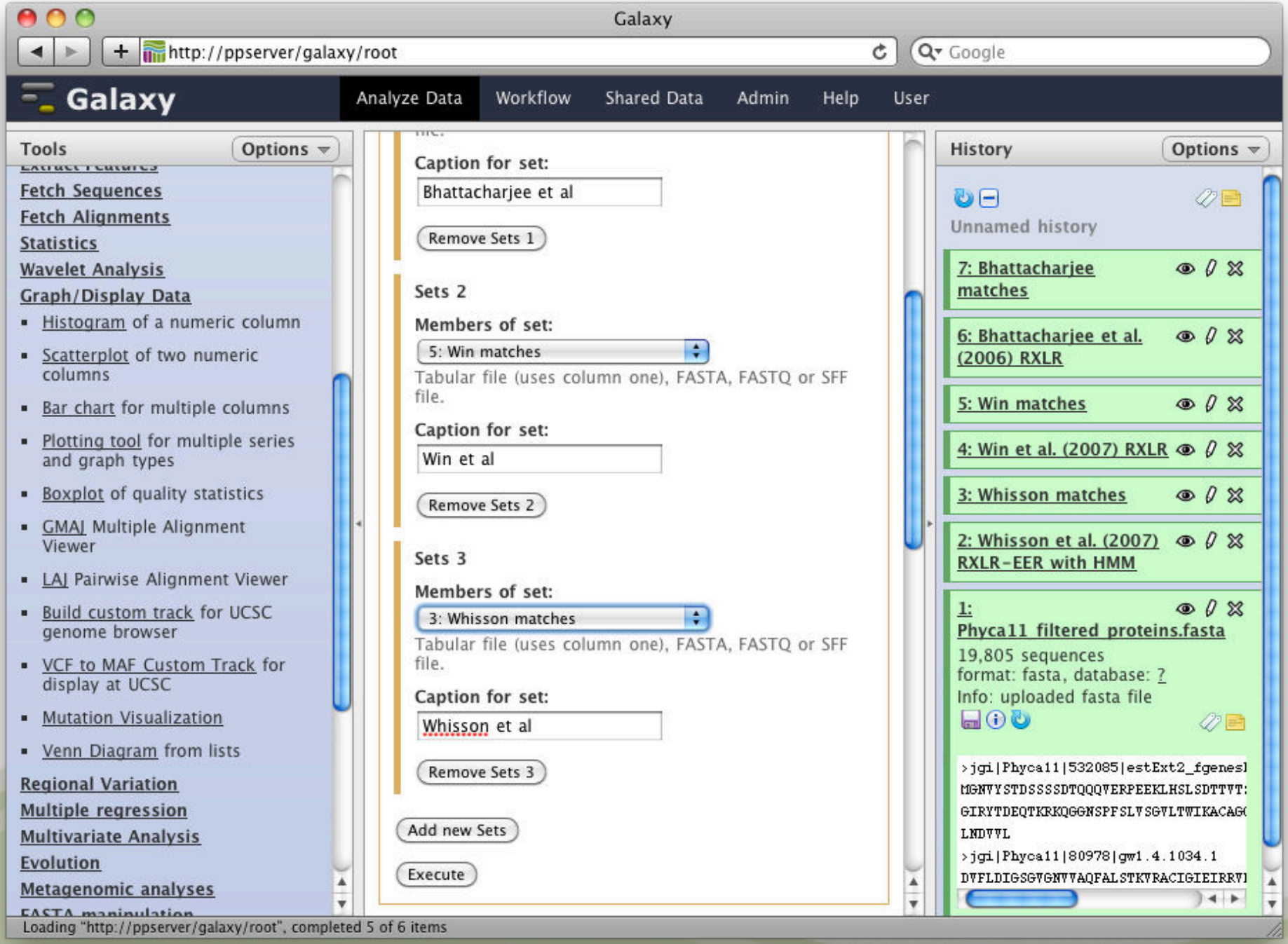

Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
LOCAL TOOLS
Upload File from your computer
NCBI BLAST+
Protein sequence analysis
Sequence manipulation
STANDARD TOOLS
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF file
Filter GFF file by attribute using simple expressions
Filter GFF file by feature count using simple expressions
Join, Subtract and Group
Convert Formats
RXLR Motifs
FASTA file of protein sequences:
1: Phyca11_filtered_..teins.fasta
Which RXLR model?:
Win et al. (2007) RXLR
Execute
Background
Many effector proteins from Oomycete plant pathogens for manipulating the host have been found to contain a signal peptide followed by a conserved RXLR motif (Arg, any amino acid, Leu, Arg), and then sometimes EER (Glu, Glu, Arg). There are stiking parallels with the malarial host-targeting signal (Plasmodium export element, or "Pexel" for short).
What it does
Takes a protein sequence FASTA file as input, and produces a simple tabular file as output with one line per protein, and two columns giving the sequence ID and the predicted class. This is typically just whether or not it had the selected RXLR motif (Y or N).
Bhattacharjee et al. (2006) RXLR Model
Looks for the oomycete motif RXLR as described in Bhattacharjee et al. (2006).
Matches must have a SignalP Hidden Markov Model (HMM) score of at least 0.9, a SignalP Neural Network (NN) predicted clevage site giving a signal peptide length between 10 and 40 amino acids inclusive, and the RXLR pattern must be after but
History
Options
Unnamed history
3: Filter on data 2
89 lines
format: tabular, database: ?
Info: Filtering with c2=='Y', kept 0.45% of 19806 lines.
Skipped 1 invalid lines starting at line #1: "#ID Whisson2007"
1
jgi|Phyca11|5186|fgenesh1_pm.PHYCA
jgi|Phyca11|130393|e_gw1.93.9.1
jgi|Phyca11|109432|e_gw1.16.672.1
jgi|Phyca11|129643|e_gw1.86.164.1
jgi|Phyca11|533084|estExt2_fgenesh
jgi|Phyca11|14941|fgenesh1_pg.PHYC
2: Whisson et al. (2007) RXLR-EER with HMM
19,805 lines, 1 comments
format: tabular, database: ?
Info: Whisson2007 for 19805 sequences:
Y = 89, hmm = 43, neither = 19657, re = 16
1
Loading "http://ppserver/galaxy/root", completed 5 of 6 items

Next few steps omitted…

- Repeated RXLR search & filter using other two models
- Labelled some history entries

The James Hutton Institute


Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
Fetch Sequences
Fetch Alignments
Statistics
Wavelet Analysis
Graph/Display Data
Histogram of a numeric column
Scatterplot of two numeric columns
Bar chart for multiple columns
Plotting tool for multiple series and graph types
Boxplot of quality statistics
GMAJ Multiple Alignment Viewer
LAJ Pairwise Alignment Viewer
Build custom track for UCSC genome browser
VCF to MAF Custom Track for display at UCSC
Mutation Visualization
Venn Diagram from lists
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Attributes updated
History
Options
Unnamed history
7: Bhattacharjee matches
6: Bhattacharjee et al. (2006) RXLR
5: Win matches
4: Win et al. (2007) RXLR
3: Whisson matches
2: Whisson et al. (2007) RXLR-EER with HMM
1: Phyca11_filtered_proteins.fasta
19,805 sequences
format: fasta, database: ?
Info: uploaded fasta file
>jgi|Phyca11|532085|estExt2_fgenesl
MGNVYSTDSSSSDTQQQVERPEEKLHSLSDTTVT:
GIRYTDEQTKRKQGGNSPFSLVSGVLTWIKACAG
LNDVVL
>jgi|Phyca11|80978|gw1.4.1034.1
DVFLDIGSGVGNVVAQFALSTKVRACIGIEIRRV
Python script using rpy
R library limma handles the plotting
Wrapper is on the Galaxy Tool Shed
Loading "http://ppserver/galaxy/root", completed 5 of


Galaxy
http://ppserver/galaxy/root
Google
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Tools
Options
Fetch Sequences
Fetch Alignments
Statistics
Wavelet Analysis
Graph/Display Data
Histogram of a numeric column
Scatterplot of two numeric columns
Bar chart for multiple columns
Plotting tool for multiple series and graph types
Boxplot of quality statistics
GMAJ Multiple Alignment Viewer
LAJ Pairwise Alignment Viewer
Build custom track for UCSC genome browser
VCF to MAF Custom Track for display at UCSC
Mutation Visualization
Venn Diagram from lists
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Caption for set:
Bhattacharjee et al
Remove Sets 1
Sets 2
Members of set:
5: Win matches
Tabular file (uses column one), FASTA, FASTQ or SFF file.
Caption for set:
Win et al
Remove Sets 2
Sets 3
Members of set:
3: Whisson matches
Tabular file (uses column one), FASTA, FASTQ or SFF file.
Caption for set:
Whisson et al
Remove Sets 3
Add new Sets
Execute
History
Options
Unnamed history
7: Bhattacharjee matches
6: Bhattacharjee et al. (2006) RXLR
5: Win matches
4: Win et al. (2007) RXLR
3: Whisson matches
2: Whisson et al. (2007) RXLR-EER with HMM
1: Phyca11_filtered_proteins.fasta
19,805 sequences
format: fasta, database: ?
Info: uploaded fasta file
>jgi|Phyca11|532085|estExt2_fgenesl
MGNVYSTDSSSSDTQQQVERPEEKLHSLSDTTVT:
GIRYTDEQTKRKQGGNSPFSLVSGVLTWIKACAG
LNDVVL
>jgi|Phyca11|80978|gw1.4.1034.1
DVFLDIGSGVGNVVAQFALSTKVRACIGIEIRRV
Loading "http://ppserver/galaxy/root", completed 5 of 6 items

Galaxy

http://ppserver/galaxy/root

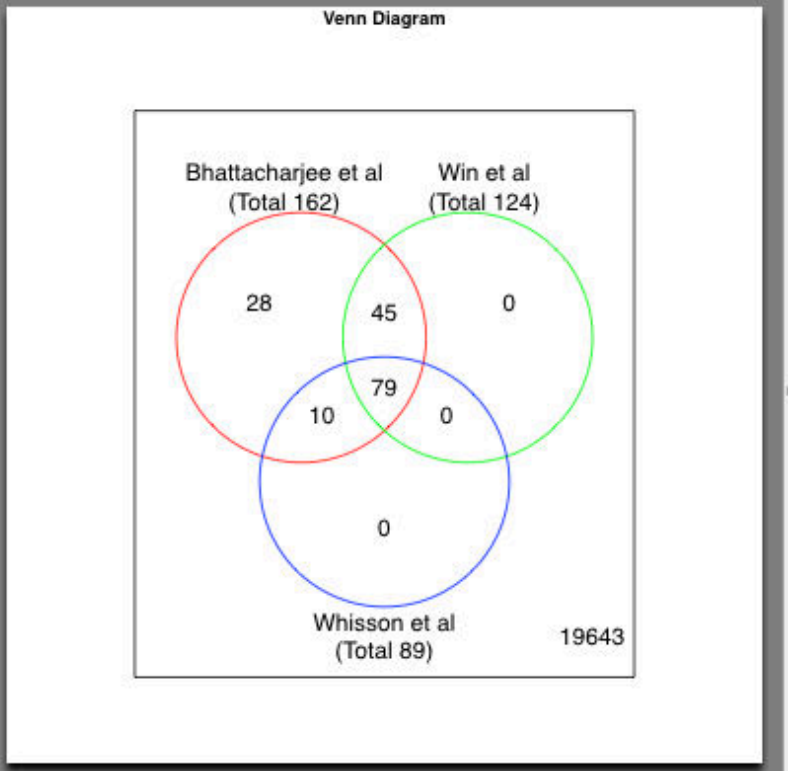
Google

Galaxy | Analyze Data | Workflow | Shared Data

Save this analysis as a workflow

Tools | Options

- Fetch Sequences
- Fetch Alignments
- Statistics
- Wavelet Analysis
- Graph/Display Data
  - Histogram of a numeric column
  - Scatterplot of two numeric columns
  - Bar chart for multiple columns
  - Plotting tool for multiple series and graph types
  - Boxplot of quality statistics
  - GMAJ Multiple Alignment Viewer
  - LAJ Pairwise Alignment Viewer
  - Build custom track for UCSC genome browser
  - VCF to MAF Custom Track for display at UCSC
  - Mutation Visualization
  - Venn Diagram from lists
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation



Options

8: Venn Diagram on data 7, data 1, and others
19.5 Kb
format: pdf, database: ?
Info: Doing 3-way Venn Diagram
Total of 19805 IDs
162 in Bhattacharjee et al
124 in Win et al
89 in Whisson et al
Image in pdf format

7: Bhattacharjee matches

6: Bhattacharjee et al. (2006) RXLR

5: Win matches

4: Win et al. (2007) RXLR

3: Whisson matches

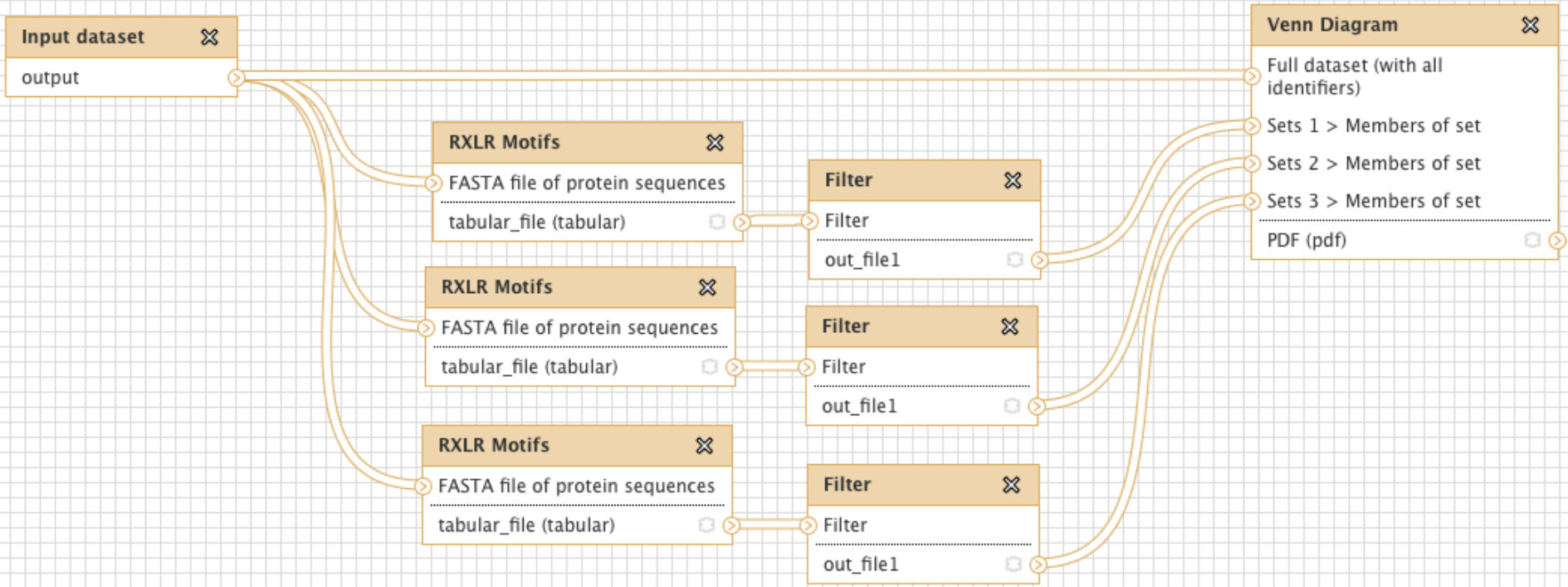
2: Whisson et al. (2007) RXLR-EER with HMM

1:

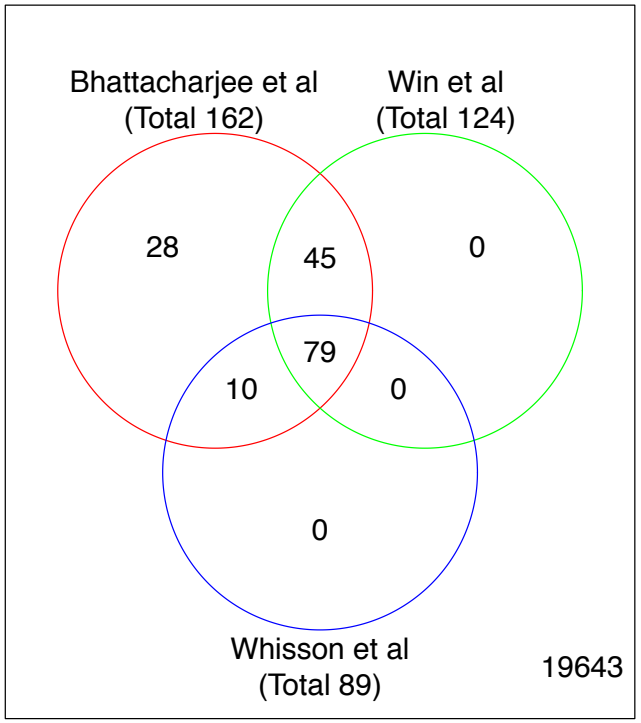
Loading "http://ppserver/galaxy/root", completed 5 of 6 items

Workflow editor view

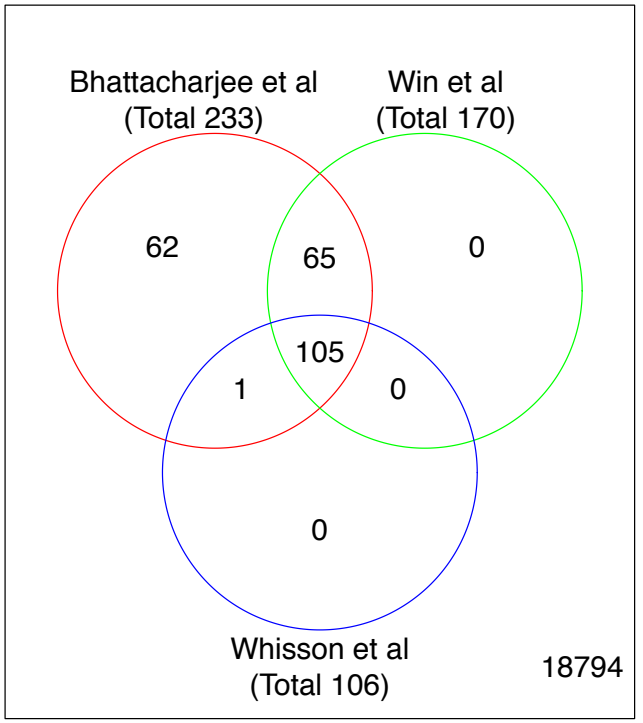
The James Hutton Institute



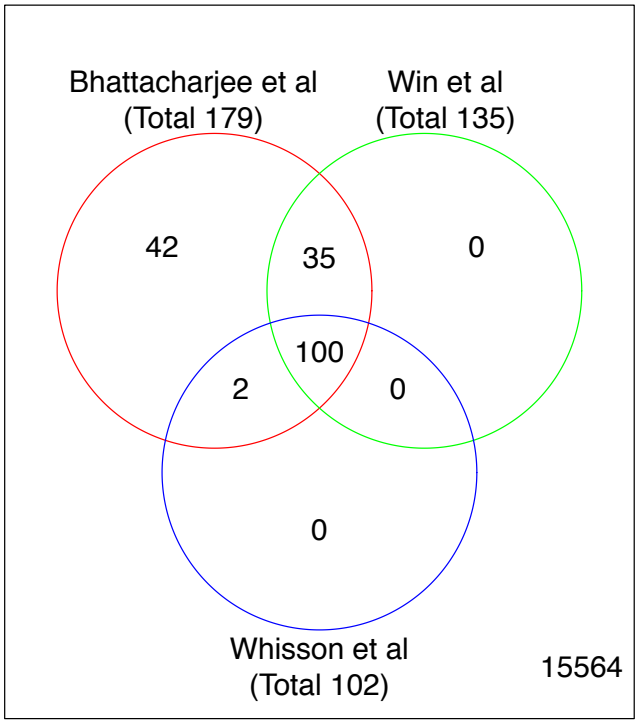
RXLRs in *Phytophthora* draft genomes



P. capsici
(19,805 proteins)



P. sojae
(19,027 proteins)



P. ramorum
(15,743 proteins)

Acknowledgements

- Helpful tool authors:
  - Alex Nguyen (NLStradamus)
  - Laszlo Kajan (PredictNLS)
  - Peter Troshin, Michelle Scott (NoD)
- JHI Testers:
  - John Jones, Remco Stam, Julietta Jupe
- The Galaxy Developers & mailing list community

The James Hutton Institute

Conclusions

The James Hutton Institute

- We like Galaxy
  - Consistent web interface for combining many tools
  - Open source and extensible
  - We can add tools of local interest
- Our Galaxy tool wrappers are on the Galaxy Tool Shed
  - http://usegalaxy.org/community
- Questions?