

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...

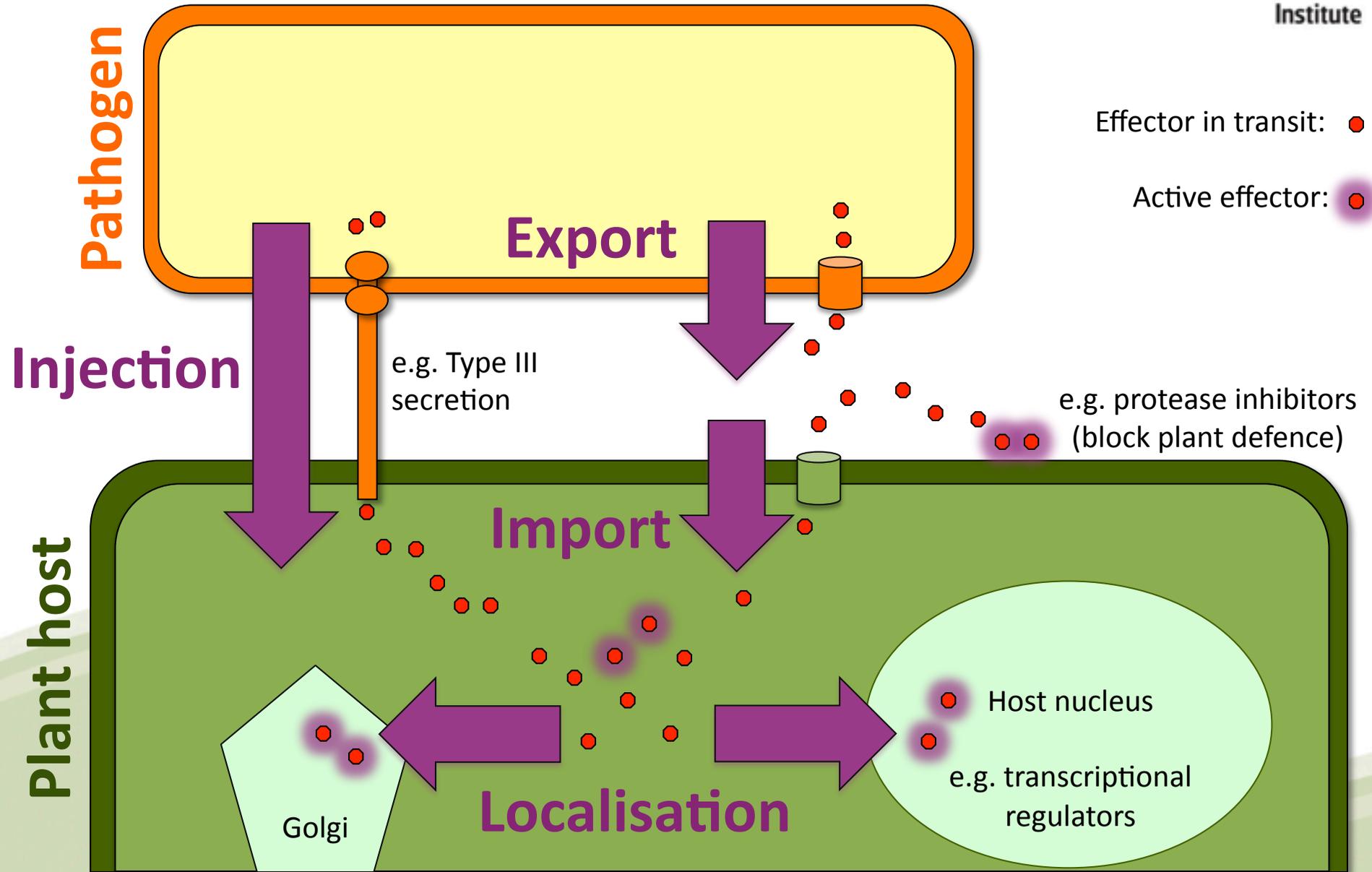


**JHI Dundee site, formerly SCRI
(Scottish Crop Research Institute)**

Plant Pathogen Genomics

- 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

- 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

- 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...

What is Galaxy?

- 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

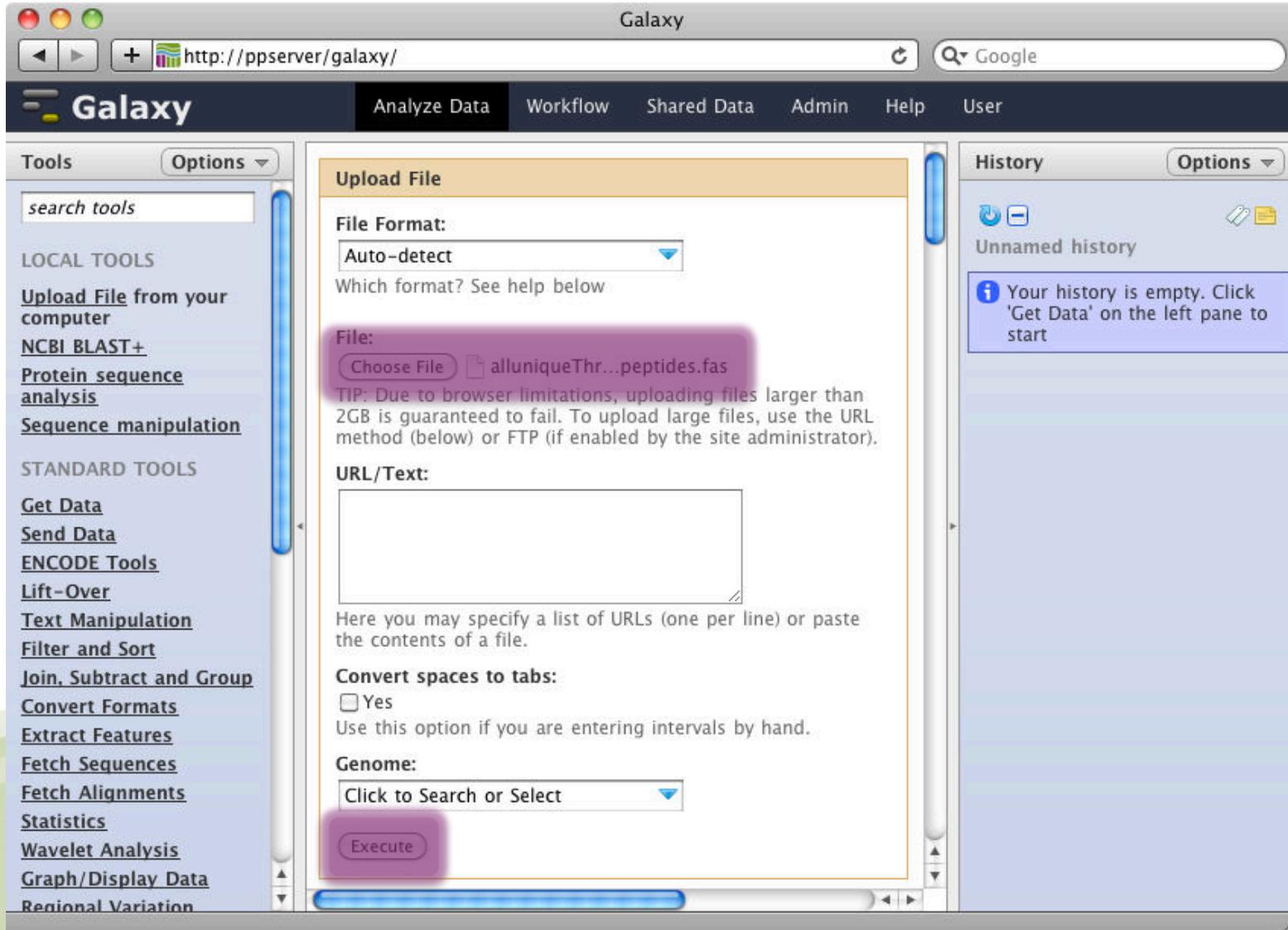
Our local Galaxy Server



The screenshot shows the Galaxy web interface in a browser window. The address bar shows `http://ppserver/galaxy/`. The main navigation bar includes "Galaxy", "Analyze Data", "Workflow", "Shared Data", "Admin", "Help", and "User". The left sidebar contains a "Tools" panel with a search box and a list of tool categories: LOCAL TOOLS, STANDARD TOOLS, and REGIONAL VARIATION. The "LOCAL TOOLS" section is highlighted, and the "Upload File from your computer" option is selected. The main content area displays a "Welcome to Galaxy" message, a "History" panel on the right, and a purple callout box with the text: "Will upload a FASTA file of predicted proteins".

Will upload a FASTA file of predicted proteins

Step 1 – Upload FASTA file



The screenshot shows the Galaxy web interface in a browser window. The address bar displays `http://ppserver/galaxy/`. The main navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', and 'User'. The left sidebar contains a 'Tools' panel with a search box and a list of tool categories: 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, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Statistics, Wavelet Analysis, Graph/Display Data, Regional Variation), and History. The central panel shows the 'Upload File' tool configuration. The 'File Format' dropdown is set to 'Auto-detect'. The 'File' section shows a 'Choose File' button and a file named 'alluniqueThr...peptides.fas'. A tip states: 'Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator)'. The 'URL/Text' section has an empty text area. The 'Convert spaces to tabs' checkbox is unchecked. The 'Genome' dropdown is set to 'Click to Search or Select'. The 'Execute' button is highlighted.

Galaxy

http://ppserver/galaxy/ Google

Galaxy Analyze Data Workflow Shared Data Admin Help User

Tools Options

search tools

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](#)

[Join, Subtract and Group](#)

[Convert Formats](#)

[Extract Features](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Statistics](#)

[Wavelet Analysis](#)

[Graph/Display Data](#)

[Regional Variation](#)

History Options

Unnamed history

Your history is empty. Click 'Get Data' on the left pane to start

Upload File

File Format:

Auto-detect

Which format? See help below

File:

Choose File alluniqueThr...peptides.fas

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:

Yes

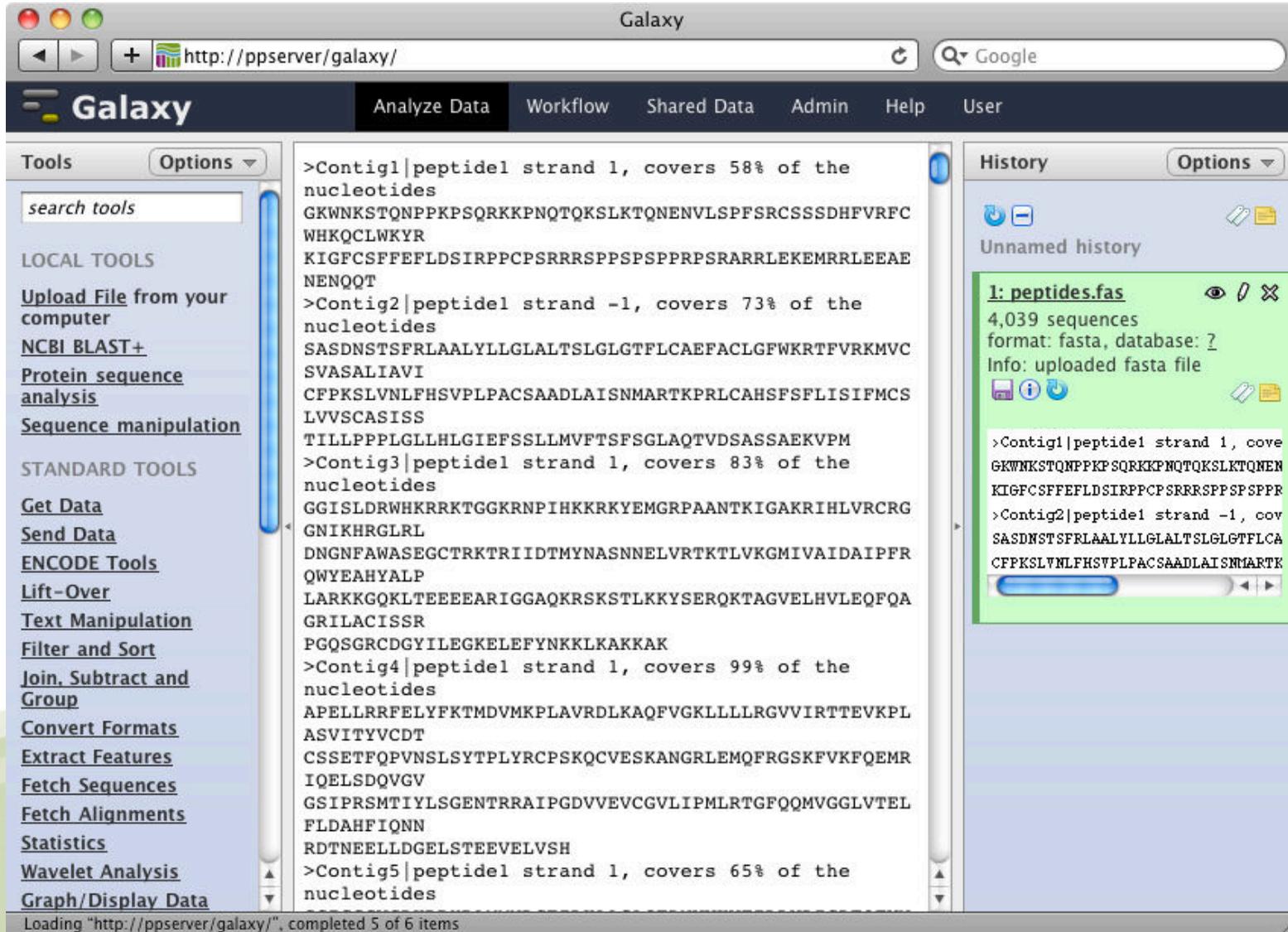
Use this option if you are entering intervals by hand.

Genome:

Click to Search or Select

Execute

Step 1 – Upload FASTA file



The screenshot shows the Galaxy web interface. The browser address bar displays `http://ppserver/galaxy/`. The main navigation bar includes **Galaxy**, **Analyze Data**, **Workflow**, **Shared Data**, **Admin**, **Help**, and **User**.

The **Tools** sidebar on the left lists various categories: LOCAL TOOLS, STANDARD TOOLS, and specific tool actions like **Upload File from your computer**, **NCBI BLAST+**, **Protein sequence analysis**, and **Sequence manipulation**.

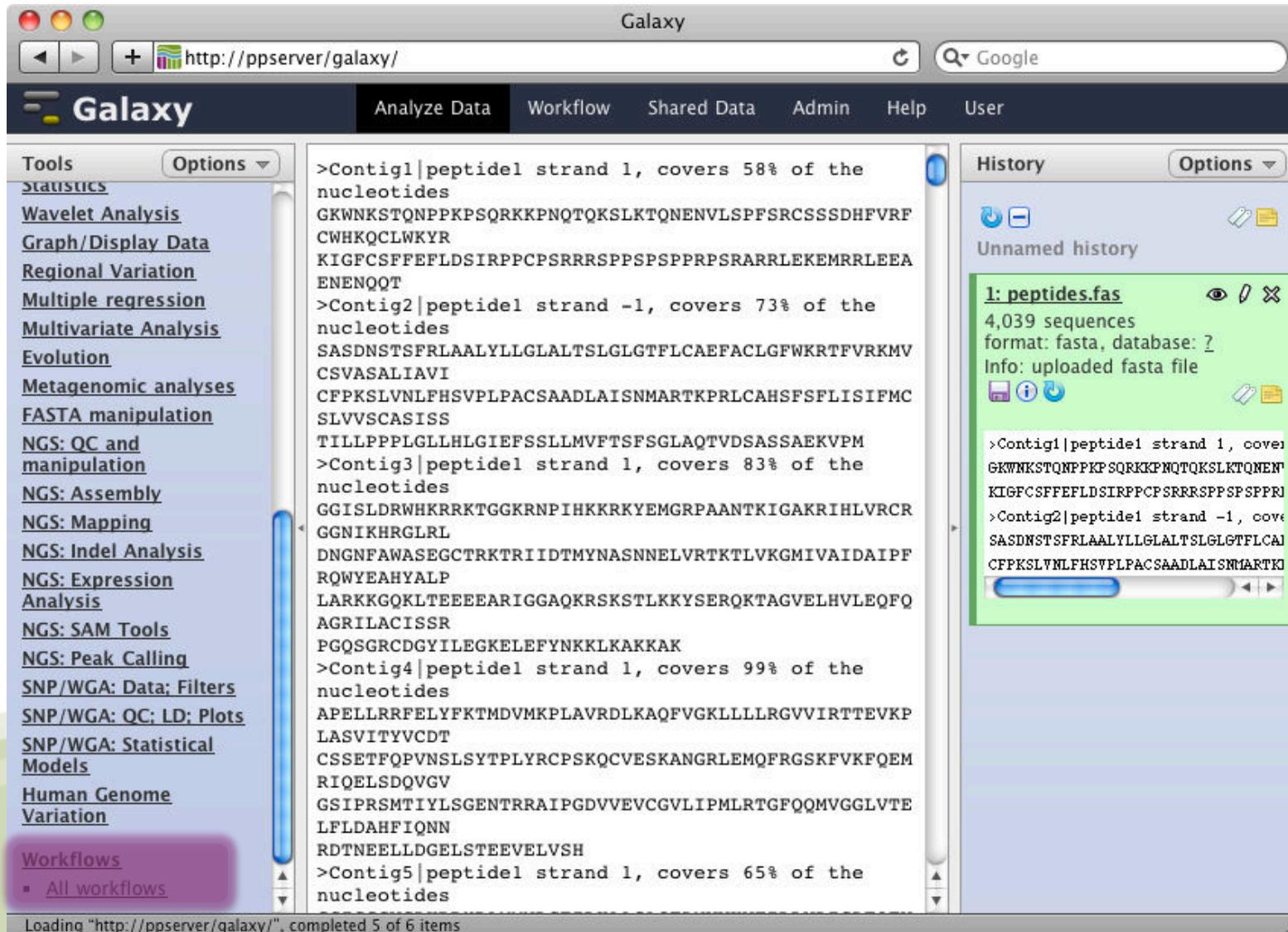
The central workspace displays the output of a FASTA file upload. It shows five contigs with their respective coverage percentages and associated nucleotide sequences:

```
>Contig1|peptidel strand 1, covers 58% of the
nucleotides
GKWNKSTQNPPKPSQRKKPNQTQKSLKTQENENLSPFSSDHFVRF
WHKQCLWKYR
KIGFCSFFEFLDSIRPPCPSRRRSPSPSPRARRLEKEMRRLEAE
NENQQT
>Contig2|peptidel strand -1, covers 73% of the
nucleotides
SASDNSTSFRLAALYLLGLALTSLGLGTFLCAEFACLGFWKRTFVRK
MVC
SVASALIAVI
CFPKSLVNLPHSVPLPACSAADLAISNMARTKPRLCAHSFSPFLIS
IFMCS
LVVSCASISS
TILLPPPLGLLHLGIEFSSLLMVFTSFSGLAQTVDSASSAEKVPM
>Contig3|peptidel strand 1, covers 83% of the
nucleotides
GGISLDRWHKRRKTGGKRNPIHKRRKYEMGRPAANTKIGAKRIHL
VRCRG
GNIKHRGLRL
DNGNFAWASEGCTRKTRI IDTMYNASNNELVRTKTLVKGMIVAID
AIPFR
QWYEAHYALP
LARKKGQKLT EEEEEARIGGAQKRKSTLKKYSERQKTAGVELHV
LEQFOA
GRILACISSR
PGQSGRCDGYILEGKELEFYNNKLLKAKKAK
>Contig4|peptidel strand 1, covers 99% of the
nucleotides
APELLRRFELYFKTMDVMKPLAVRDLKAQFVGKLLLLRGVVIRTE
VKPL
ASVITYVCDT
CSSETFQPVNSLSYTPLYRCPSKQCVESKANGRLEMQFRGSKFVK
FQEMR
IQELSDQGVG
GSIPRSMTIYLSGENTRAIPGDVVEVCGVLIPLRTGFQQMVGGL
VTEL
FLDAHFIQNN
RDTNEELLDGELSTEEVELVSH
>Contig5|peptidel strand 1, covers 65% of the
nucleotides
```

The **History** sidebar on the right shows a record for the uploaded file: **1: peptides.fasta**, containing 4,039 sequences in FASTA format. The history entry includes a preview of the first few lines of the FASTA file.

At the bottom of the interface, a status bar indicates: `Loading "http://ppserver/galaxy/", completed 5 of 6 items`.

Step 2 – Run workflow



Galaxy

http://ppserver/galaxy/ Google

Analyze Data Workflow Shared Data Admin Help User

Tools Options

Statistics

- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: Expression Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- Workflows
 - All workflows

```

>Contig1|peptidel strand 1, covers 58% of the
nucleotides
GKWNKSTQNPFPKPSQRKKPNQTQKSLKTQENENVLSPFSSDHFVRF
CWHKQCLWKYR
KIGFCSFFFEFLDSIRPPCPSRRRSPSPSPPPRPSRARRLEKEMRLEEA
ENENQQT
>Contig2|peptidel strand -1, covers 73% of the
nucleotides
SASDNSTSFRLAALYLLGLALTSGLGLTFLCAEFACLGFWRKTFVRKMW
CSVASALIAVI
CFPKSLVNLFHSVPLPACSAADLAISNMARTKPRCAHSFSPFLISIFMC
SLVSCASISS
TILLPPPLGLLHLGIEFSSLLMVFTSFGSLAQTVDSSASSAEKVPW
>Contig3|peptidel strand 1, covers 83% of the
nucleotides
GGISLDRWHKRRKTGGKRNPIHKKRKYEMGRPAANTKIGAKRIHLVRCR
GGNIKHRGLRL
DNGNFAWASEGCTRKTRIIDTMYNASNNELVRTKTLVKGMIVAIDAIPF
RQWYEAHYALP
LARKKGQKLTEEEEARIGGAQKRKSTLKKYSERQKTAGVELHVLEQFQ
AGRILACISSR
PGQSGRCDGYILEGKELEFYNNKLLKAKKAK
>Contig4|peptidel strand 1, covers 99% of the
nucleotides
APELLRRFELYFKTMDVMKPLAVRDLKAQFVGKLLLLRGVVIRTTEVKP
LASVITYVCDT
CSSETFPQVNSLSYTPLYRCPSKQCVESKANGRLEMQFRGSKFVKFQEM
RIQELSDQVGV
GSIPRSMTIYLSGENTRRRAIPGDVVEVCGVLIPMLRTGFPQMVGGLVTE
LFLDAHFIQNN
RDTNEELLDGELSTEEVELVSH
>Contig5|peptidel strand 1, covers 65% of the
nucleotides

```

History Options

Unnamed history

1: peptides.fas   

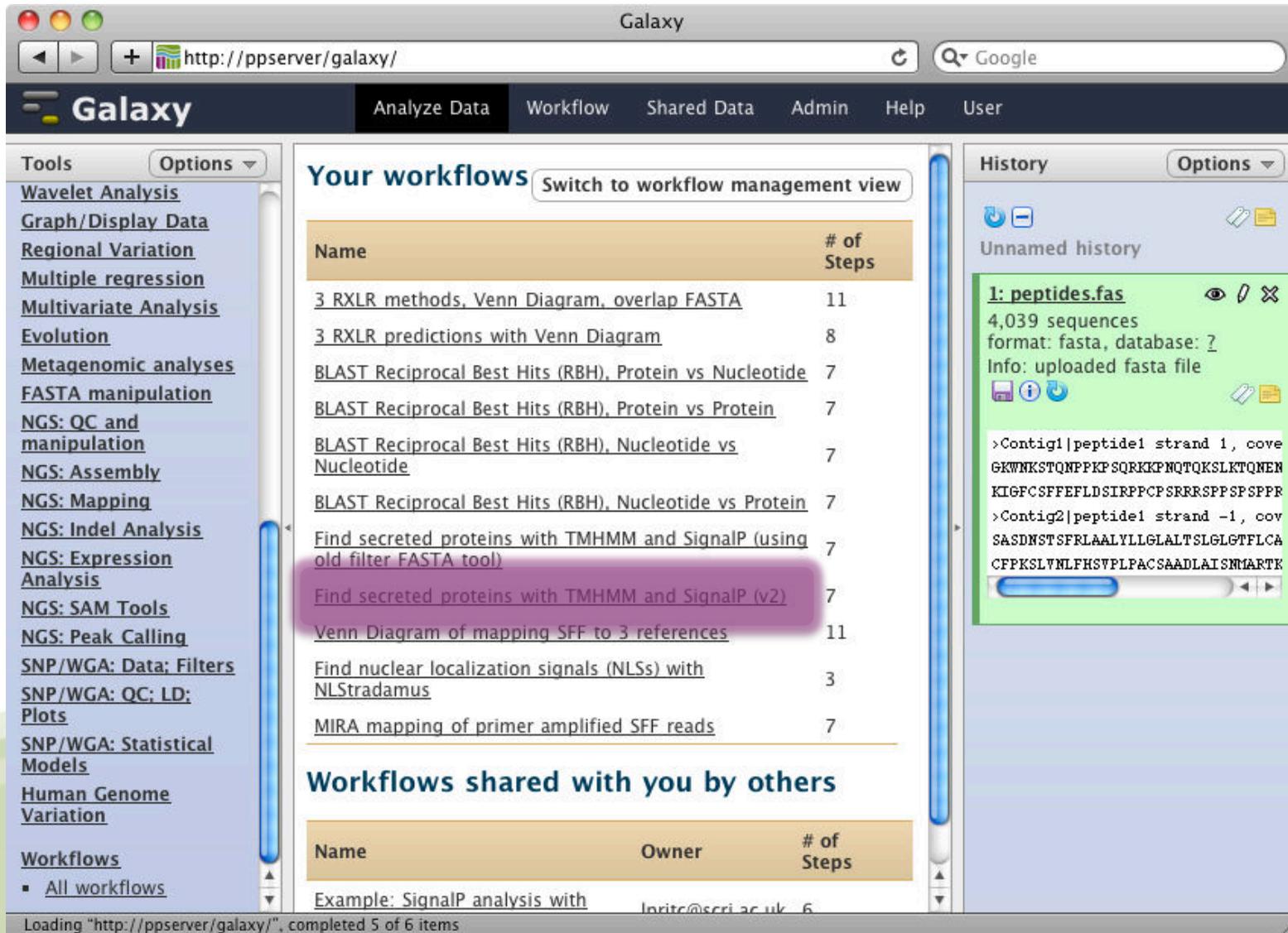
4,039 sequences
format: fasta, database: ?
Info: uploaded fasta file

>Contig1|peptidel strand 1, covers 58% of the
nucleotides
GKWNKSTQNPFPKPSQRKKPNQTQKSLKTQENENVLSPFSSDHFVRF
CWHKQCLWKYR
KIGFCSFFFEFLDSIRPPCPSRRRSPSPSPPPRPSRARRLEKEMRLEEA
ENENQQT
>Contig2|peptidel strand -1, covers 73% of the
nucleotides
SASDNSTSFRLAALYLLGLALTSGLGLTFLCAEFACLGFWRKTFVRKMW
CSVASALIAVI
CFPKSLVNLFHSVPLPACSAADLAISNMARTKPRCAHSFSPFLISIFMC
SLVSCASISS
TILLPPPLGLLHLGIEFSSLLMVFTSFGSLAQTVDSSASSAEKVPW
>Contig3|peptidel strand 1, covers 83% of the
nucleotides
GGISLDRWHKRRKTGGKRNPIHKKRKYEMGRPAANTKIGAKRIHLVRCR
GGNIKHRGLRL
DNGNFAWASEGCTRKTRIIDTMYNASNNELVRTKTLVKGMIVAIDAIPF
RQWYEAHYALP
LARKKGQKLTEEEEARIGGAQKRKSTLKKYSERQKTAGVELHVLEQFQ
AGRILACISSR
PGQSGRCDGYILEGKELEFYNNKLLKAKKAK
>Contig4|peptidel strand 1, covers 99% of the
nucleotides
APELLRRFELYFKTMDVMKPLAVRDLKAQFVGKLLLLRGVVIRTTEVKP
LASVITYVCDT
CSSETFPQVNSLSYTPLYRCPSKQCVESKANGRLEMQFRGSKFVKFQEM
RIQELSDQVGV
GSIPRSMTIYLSGENTRRRAIPGDVVEVCGVLIPMLRTGFPQMVGGLVTE
LFLDAHFIQNN
RDTNEELLDGELSTEEVELVSH
>Contig5|peptidel strand 1, covers 65% of the
nucleotides

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

Step 2 – Run workflow



The screenshot shows the Galaxy web interface. The browser address bar displays 'http://ppserver/galaxy/'. The main navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', and 'User'. On the left, a 'Tools' sidebar lists various analysis tools. The central 'Your workflows' section contains a table of workflows, with one entry highlighted in purple. The right sidebar shows the 'History' section with a preview of a workflow named '1: peptides.fas'.

Your workflows Switch to workflow management view

Name	# of Steps
3 RXLR methods, Venn Diagram, overlap FASTA	11
3 RXLR predictions with Venn Diagram	8
BLAST Reciprocal Best Hits (RBH), Protein vs Nucleotide	7
BLAST Reciprocal Best Hits (RBH), Protein vs Protein	7
BLAST Reciprocal Best Hits (RBH), Nucleotide vs Nucleotide	7
BLAST Reciprocal Best Hits (RBH), Nucleotide vs Protein	7
Find secreted proteins with TMHMM and SignalP (using old filter FASTA tool)	7
Find secreted proteins with TMHMM and SignalP (v2)	7
Venn Diagram of mapping SFF to 3 references	11
Find nuclear localization signals (NLSs) with NLStradamus	3
MIRA mapping of primer amplified SFF reads	7

Workflows shared with you by others

Name	Owner	# of Steps
Example: SignalP analysis with	lnrite@scri.ac.uk	6

History Options

Unnamed history

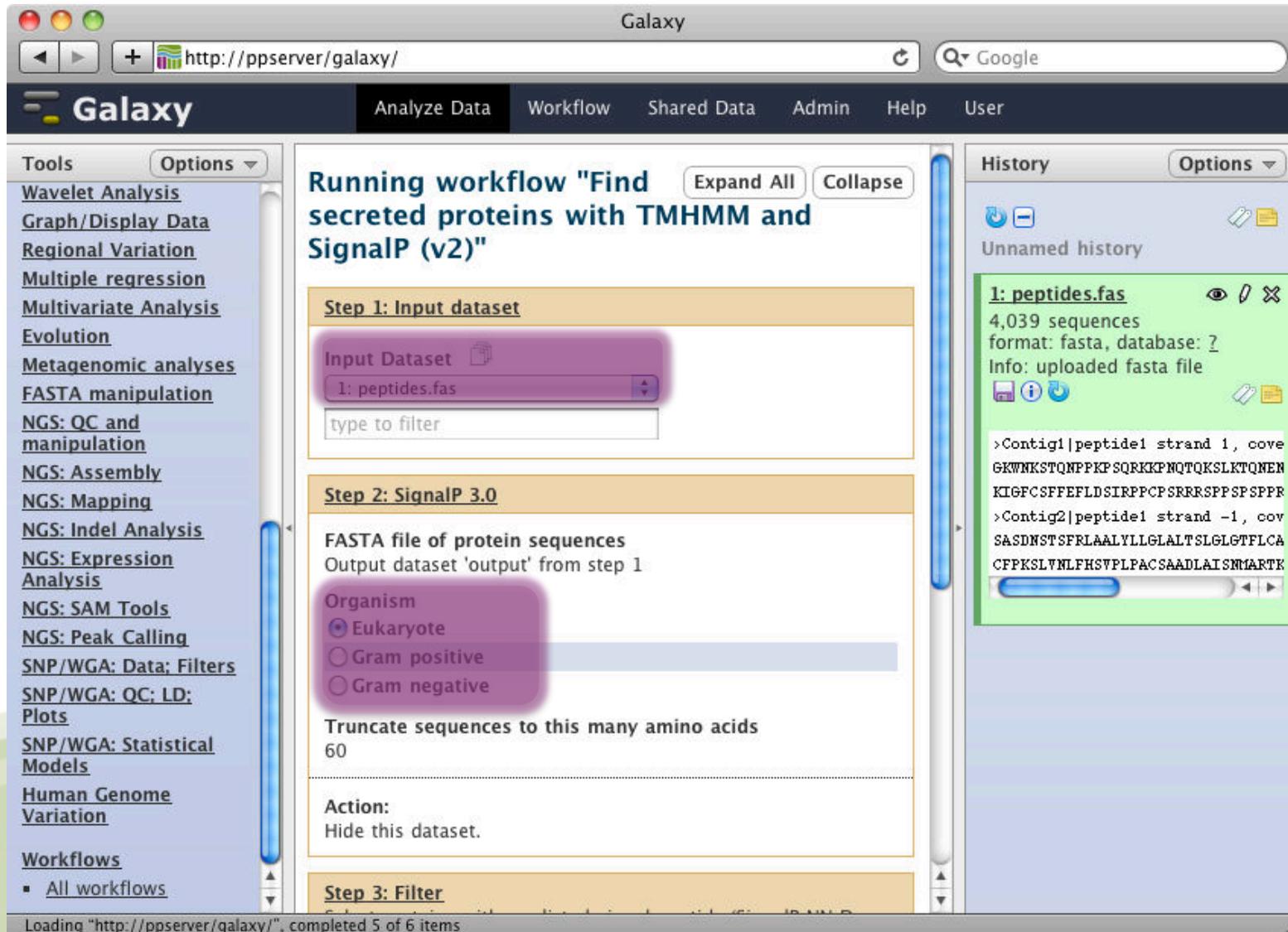
1: peptides.fas 👁 ✂ ✕

4,039 sequences
format: fasta, database: ?
Info: uploaded fasta file

```
>Contig1|peptide1 strand 1, cov
GKWNKSTQMPKPSQRKPNQTKSLKQEN
KIGFCSPFEFLDSIRPPCSRSSPPSPSPR
>Contig2|peptide1 strand -1, cov
SASDNSTSFRLAALYLLGLALTSGLGTFLLCA
CFPKSLVNLFSVPLPACSAADLAISMARK
```

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

Step 2 – Run workflow



Galaxy

http://ppserver/galaxy/ Google

Galaxy Analyze Data Workflow Shared Data Admin Help User

Tools Options

- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: Expression Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- SNP/WGA: Data: Filters
- SNP/WGA: QC: LD: Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- Workflows
 - All workflows

Running workflow "Find secreted proteins with TMHMM and SignalP (v2)"

Expand All Collapse

Step 1: Input dataset

Input Dataset

1: peptides.fas

type to filter

Step 2: SignalP 3.0

FASTA file of protein sequences

Output dataset 'output' from step 1

Organism

- Eukaryote
- Gram positive
- Gram negative

Truncate sequences to this many amino acids

60

Action:

Hide this dataset.

Step 3: Filter

History Options

Unnamed history

1: peptides.fas

4,039 sequences

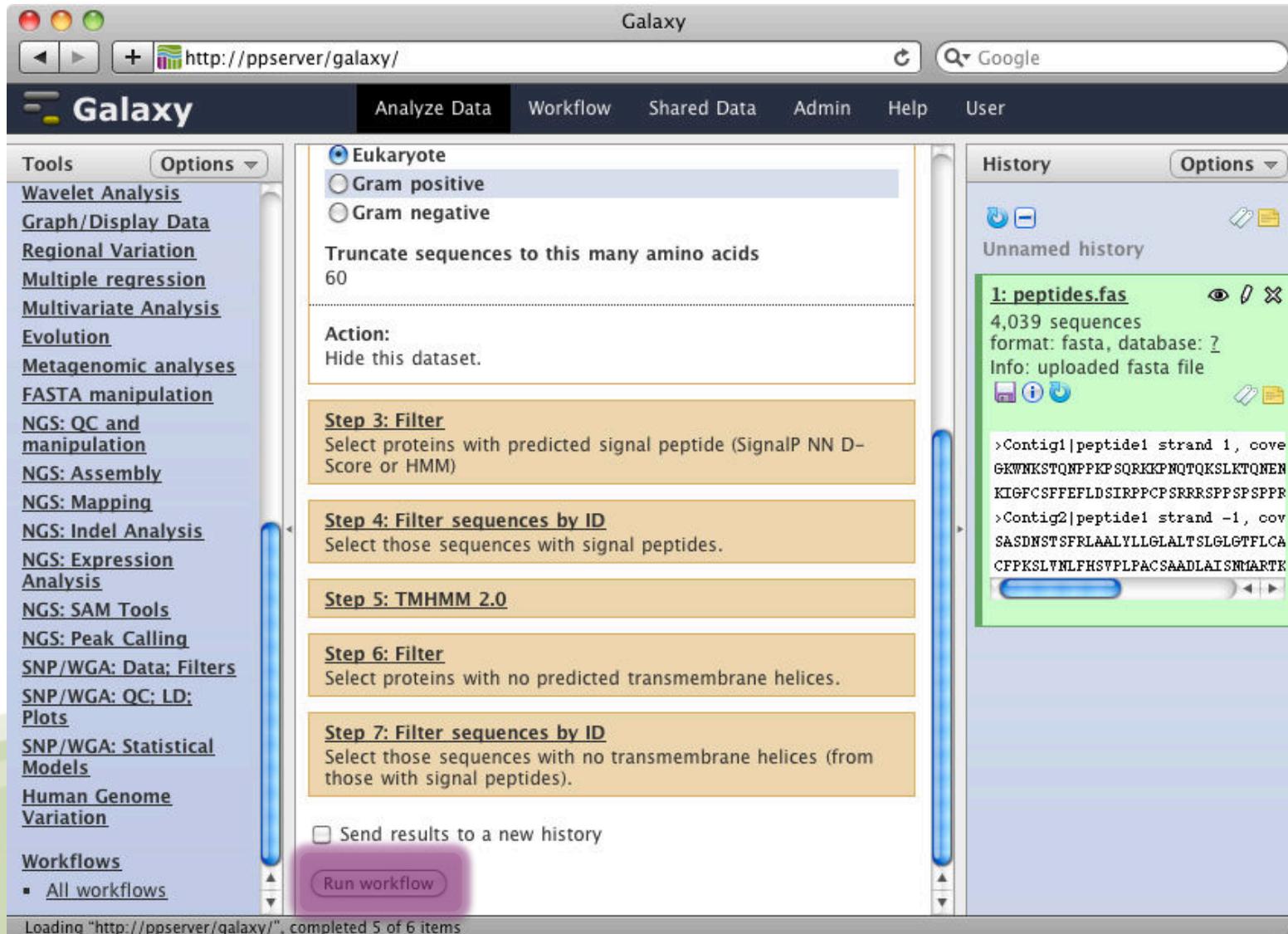
format: fasta, database: ?

Info: uploaded fasta file

```
>Contig1|peptide1 strand 1, cove
GKWNKSTQMPKPKSQRKPNQTKSLKQEN
KIGFCSFFEFLDSIRPPCSRSSPPSPSPR
>Contig2|peptide1 strand -1, cov
SASDNSTSFRLAALYLLGLALTSIGLGTFLCA
CFPKSLVNLFHSVPLPACSAADLAISMARK
```

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

Step 2 – Run workflow



The screenshot shows the Galaxy web interface with a workflow configuration. The browser address bar shows `http://ppserver/galaxy/`. The navigation menu includes **Analyze Data**, **Workflow**, **Shared Data**, **Admin**, **Help**, and **User**.

Tools (left sidebar):

- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: Expression Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- Workflows
 - All workflows

Workflow Configuration (center):

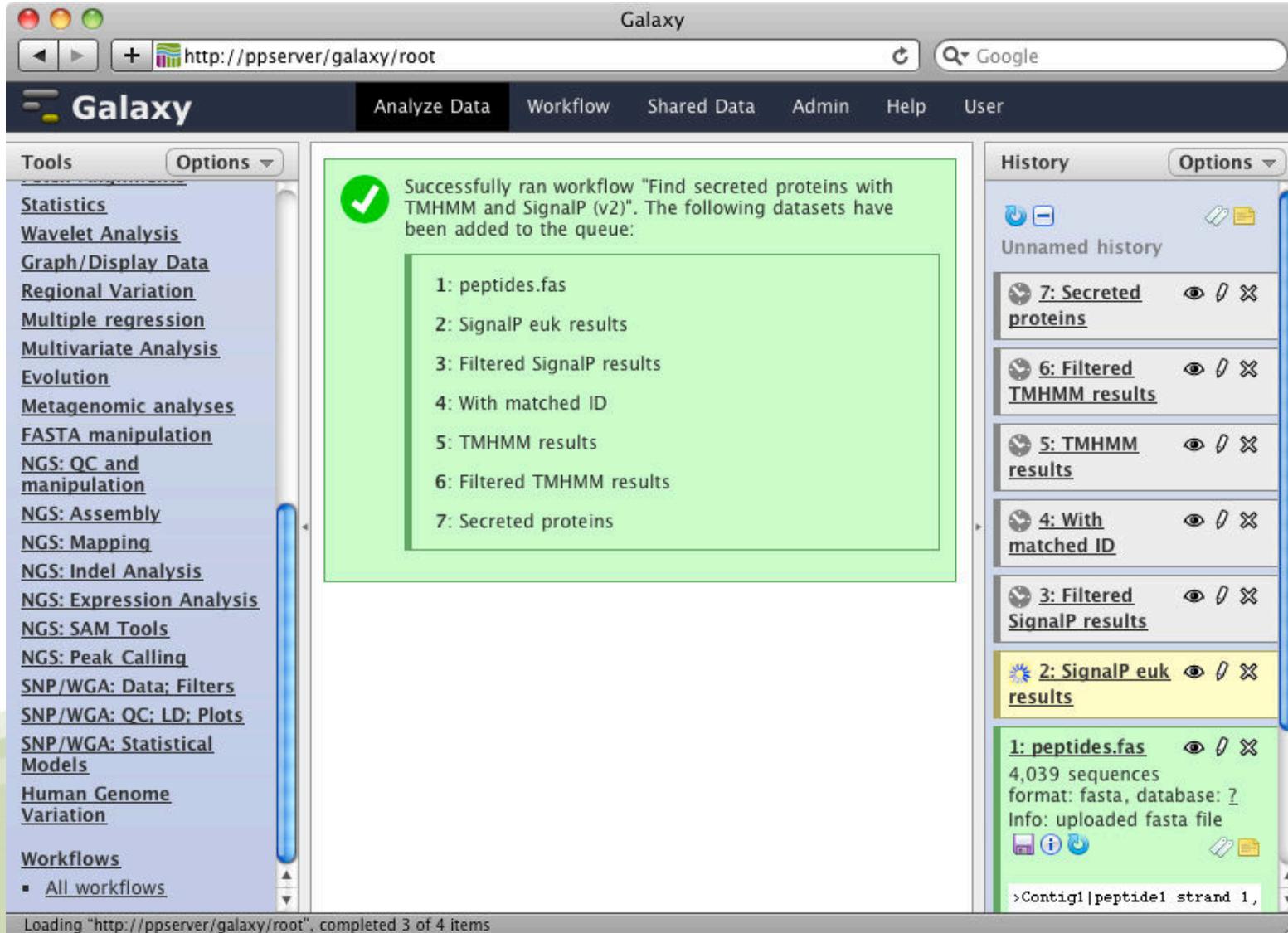
- Eukaryote
- Gram positive
- Gram negative
- Truncate sequences to this many amino acids: 60
- Action: Hide this dataset.
- Step 3: Filter**
Select proteins with predicted signal peptide (SignalP NN D-Score or HMM)
- Step 4: Filter sequences by ID**
Select those sequences with signal peptides.
- Step 5: TMHMM 2.0**
- Step 6: Filter**
Select proteins with no predicted transmembrane helices.
- Step 7: Filter sequences by ID**
Select those sequences with no transmembrane helices (from those with signal peptides).
- Send results to a new history
- Run workflow** (button)

History (right sidebar):

- Unnamed history
- 1: peptides.fas** (4,039 sequences, format: fasta, database: ?)
- Info: uploaded fasta file
- Contig1|peptide1 strand 1, cov
 GKWNKSTQMPKPSQRKPNQTKSLKQEN
 KIGFCFFFEFLDSIRPPCSRSSPPSPSPR
- Contig2|peptide1 strand -1, cov
 SASDNSTSFRLAALYLLGLALTSGLGTFLLCA
 CFPKSLVNLFSVPLPACSAADLAISMARK

At the bottom, a status bar indicates: Loading "http://ppserver/galaxy/", completed 5 of 6 items

Step 2 – Run workflow



The screenshot shows the Galaxy web interface. At the top, the browser address bar displays "http://ppserver/galaxy/root". The main navigation bar includes "Galaxy" and menu items: "Analyze Data", "Workflow", "Shared Data", "Admin", "Help", and "User".

On the left, a "Tools" sidebar lists various analysis categories such as "Statistics", "Wavelet Analysis", "Graph/Display Data", "Regional Variation", "Multiple regression", "Multivariate Analysis", "Evolution", "Metagenomic analyses", "FASTA manipulation", "NGS: QC and manipulation", "NGS: Assembly", "NGS: Mapping", "NGS: Indel Analysis", "NGS: Expression Analysis", "NGS: SAM Tools", "NGS: Peak Calling", "SNP/WGA: Data; Filters", "SNP/WGA: QC; LD; Plots", "SNP/WGA: Statistical Models", "Human Genome Variation", and "Workflows".

The central workspace displays a green notification box with a checkmark icon, stating: "Successfully ran workflow 'Find secreted proteins with TMHMM and SignalP (v2)'. The following datasets have been added to the queue:"

- 1: peptides.fas
- 2: SignalP euk results
- 3: Filtered SignalP results
- 4: With matched ID
- 5: TMHMM results
- 6: Filtered TMHMM results
- 7: Secreted proteins

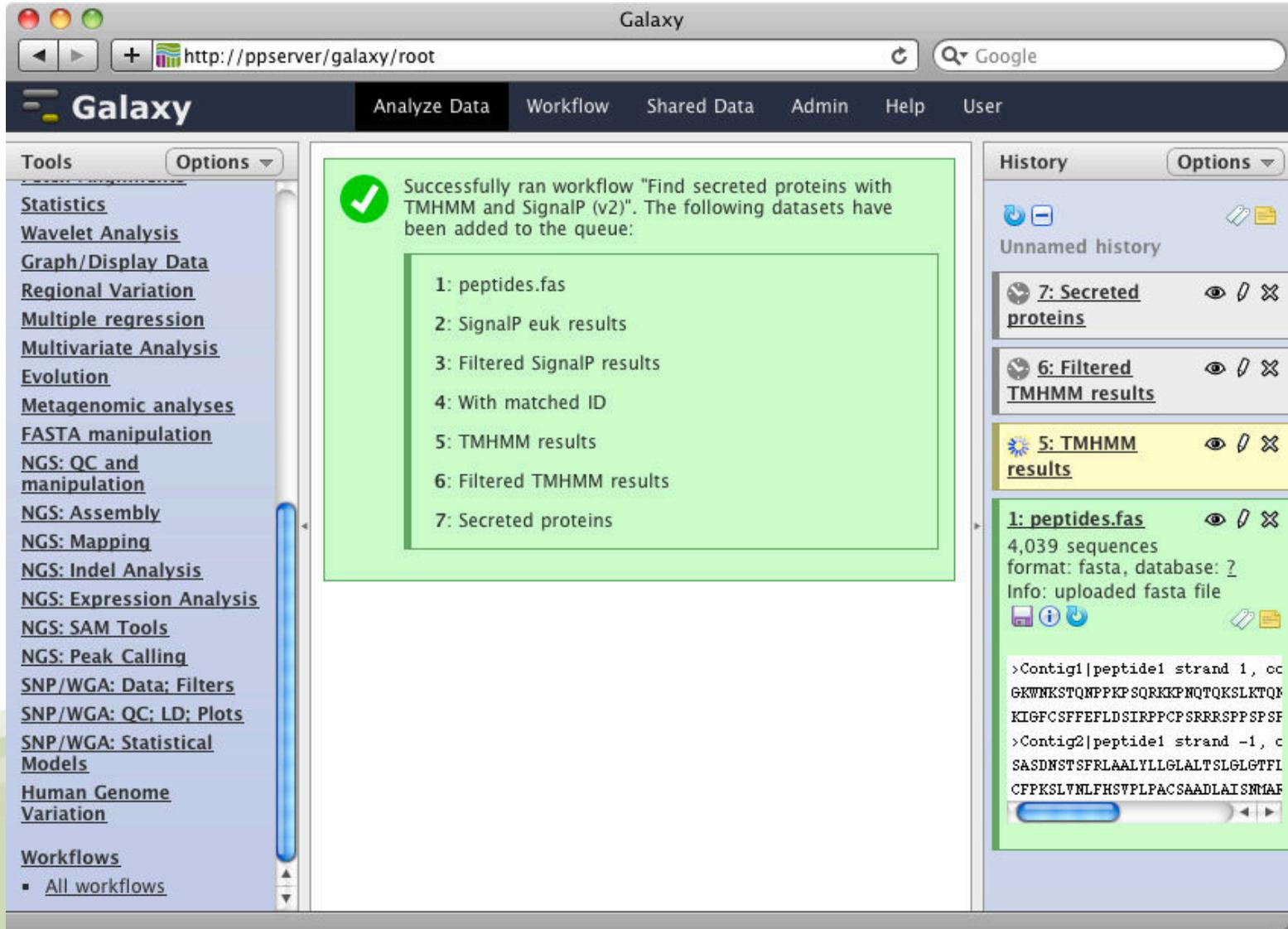
On the right, the "History" panel shows a list of datasets in reverse chronological order:

- 7: Secreted proteins
- 6: Filtered TMHMM results
- 5: TMHMM results
- 4: With matched ID
- 3: Filtered SignalP results
- 2: SignalP euk results
- 1: peptides.fas

The details for the first dataset, "1: peptides.fas", are expanded, showing: "4,039 sequences", "format: fasta, database: ?", "Info: uploaded fasta file", and a preview of the first line: ">Contig1|peptide1 strand 1,".

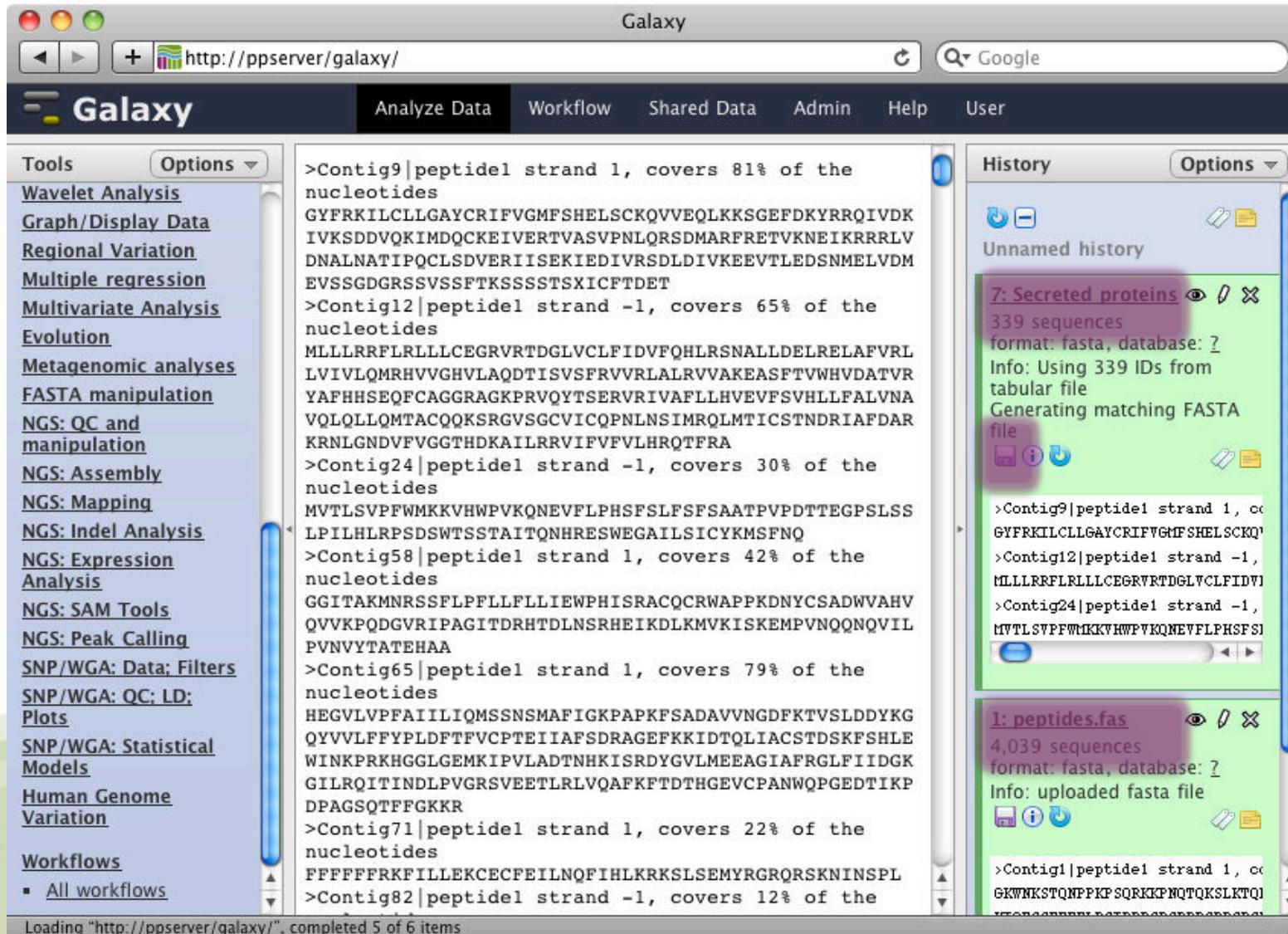
At the bottom of the interface, a status bar indicates: "Loading 'http://ppserver/galaxy/root', completed 3 of 4 items".

Step 2 – Run workflow



The screenshot shows the Galaxy web interface. At the top, there's a navigation bar with 'Galaxy' and menu items: 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', and 'User'. Below this is a 'Tools' sidebar on the left with various analysis categories like 'Statistics', 'Wavelet Analysis', 'Graph/Display Data', etc. The main content area displays a green success message: 'Successfully ran workflow "Find secreted proteins with TMHMM and SignalP (v2)". The following datasets have been added to the queue:'. Below the message is a numbered list of datasets: 1: peptides.fas, 2: SignalP euk results, 3: Filtered SignalP results, 4: With matched ID, 5: TMHMM results, 6: Filtered TMHMM results, and 7: Secreted proteins. On the right, the 'History' panel shows a list of datasets: '7: Secreted proteins', '6: Filtered TMHMM results', '5: TMHMM results' (highlighted in yellow), and '1: peptides.fas'. The '1: peptides.fas' dataset is expanded to show details: '4,039 sequences', 'format: fasta, database: ?', 'Info: uploaded fasta file', and a preview of the FASTA sequence content.

Step 3 – Get results



The screenshot shows the Galaxy web interface. The browser address bar displays `http://ppserver/galaxy/`. The main navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Admin', 'Help', and 'User'. On the left, a 'Tools' sidebar lists various analysis options such as 'Wavelet Analysis', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Metagenomic analyses', 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Assembly', 'NGS: Mapping', 'NGS: Indel Analysis', 'NGS: Expression Analysis', 'NGS: SAM Tools', 'NGS: Peak Calling', 'SNP/WGA: Data; Filters', 'SNP/WGA: QC; LD; Plots', 'SNP/WGA: Statistical Models', 'Human Genome Variation', and 'Workflows'. The central workspace displays the results of a peptide analysis, showing coverage percentages and nucleotide sequences for several contigs. The right-hand 'History' panel shows a list of recent jobs, including '7: Secreted proteins' (339 sequences) and '1: peptides.fas' (4,039 sequences).

Tools Options ▾

- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Indel Analysis
- NGS: Expression Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models
- Human Genome Variation
- Workflows
 - All workflows

History Options ▾

Unnamed history

7: Secreted proteins   
 339 sequences
 format: fasta, database: ?
 Info: Using 339 IDs from tabular file
 Generating matching FASTA file

1: peptides.fas   
 4,039 sequences
 format: fasta, database: ?
 Info: uploaded fasta file

Galaxy Analyze Data Workflow Shared Data Admin Help User

Galaxy

http://ppserver/galaxy/ Google

>Contig9|peptidel strand 1, covers 81% of the nucleotides
 GYFRKILCLLGAYCRIFVGMFSHELSCQVVEQLKKSGEFDKYRRQIVDK
 IVKSDDVQKIMDQCKEIVERTVASVPLNQRSDMARFRETVKNEIKRRRLV
 DNALNATIPQCLSDVERI ISEKIEDIVRSDLDIVKEEVTLEDSNMELVDM
 EVSSGDGRSSVSSFTKSSSSTSXCFTDET

>Contig12|peptidel strand -1, covers 65% of the nucleotides
 MLLLRFLRLLLCEGRVRTDGLVCLFIDVQHLRSNALLDELRELAFVRL
 LVIVLQMRHVVGHVLAQDTISVSFRVRLALRVVAKEASFTVWHVDATVR
 YAFHHSEQFCAGGRAGKPRVQYTSERVRIVAFLLHVEVFSVHLLFALVNA
 VQLQLQMTACQKSRGVSVCVICPNLNSIMRQLMTICSTNDRIAFDAR
 KRNLGNDVFPVGGTHDKAILRRVIFVFLHRQTFRA

>Contig24|peptidel strand -1, covers 30% of the nucleotides
 MVTLSVPFWMKKVHWPVKQNEVFLPHSFLSFSASAATPVPDTEGPSLSS
 LPILHLRPSDSWTSSTAITQNHRESWEGAILSICYKMSFNQ

>Contig58|peptidel strand 1, covers 42% of the nucleotides
 GGITAKMNRSSFLPFLFLLEIWP HISRACQCRWAPPKDNYSADWVAHV
 QVVKPQDGVRI PAGITDRHTDLNSRHEIKDLKMKV KISKEMPVNQQNVIL
 PVNVYTATEHAA

>Contig65|peptidel strand 1, covers 79% of the nucleotides
 HEGVLVPPFAI ILIQMSSNSMAFIGKPAKPSADAVVNGDFKTVSLDDYKG
 QYVVLFFYPLDFTFVCPT EIIAFSDRAGEFKKIDTQLIACSTDSKFSHLE
 WINKPRKHGGLGEMKIPVLADTNHKISR DYGLMEEAGIAFRGLF IIDGK
 GILRQITINDLPVGRSVEETLRLVQAFKFTDTHGEVCPANWQP GEDTIK P
 DPAGSQTFPGKKR

>Contig71|peptidel strand 1, covers 22% of the nucleotides
 FFFFFFFRKFILLEKCECFEILNQFIHLKRKSLSEMYRGRQRSKNINSPL

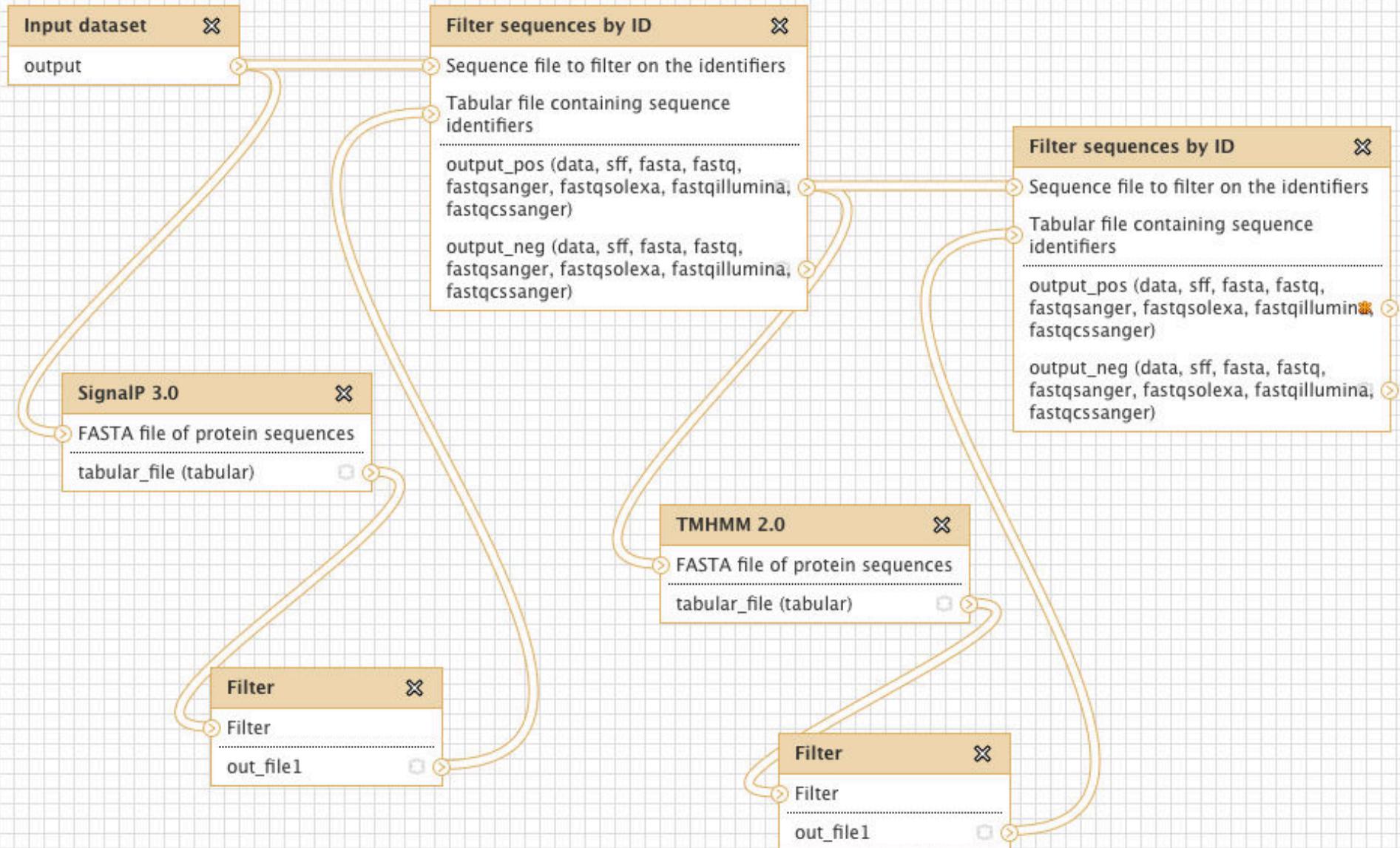
>Contig82|peptidel strand -1, covers 12% of the

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

Example: Effector Protein Analysis

- 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

Workflow Editor – Effector finding



Why Galaxy?

- 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

- 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

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

Observations from Wrapping Tools

- 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

- 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

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

Open "http://ppserver/galaxy/tool_runner?tool_id=rxlr_motifs" in a new tab behind the current one

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 striking 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 cleavage 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_fgenes1_
MGNVYSTDSSSSDTQQQVERPEEKLSLSDTTVTSSQ
GIRYTDQTKRKQGGNSPFLVSGVLTWIKACAGGSS
LNDVVL
> jgi|Phyca11|80978|gw1.4.1034.1
DVFLDIGSGVGNVVAQFALSTKVRACIGIEIRRVLAD
```

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_fgenes1_pg.C.PHYCAscaffold_40001
jgi|Phyca11|80978|gw1.4.1034.1 neither
jgi|Phyca11|99641|e_gw1.4.982.1 neither
jgi|Phyca11|13479|fgenes1_pg.PHYCAscaffold_4_#_2 neither
jgi|Phyca11|99637|e_gw1.4.173.1 neither
jgi|Phyca11|503463|fgenes2_kg.PHYCAscaffold_4_#_2_#_Contig4450
jgi|Phyca11|99784|e_gw1.4.911.1 neither
jgi|Phyca11|525476|estExt2_fgenes1_pm.C.PHYCAscaffold_40002
jgi|Phyca11|100524|e_gw1.4.603.1 neither
jgi|Phyca11|539761|estExt2_Genewise1Plus.C.PHYCAscaffold_40015
jgi|Phyca11|503470|fgenes2_kg.PHYCAscaffold_4_#_9_#_4100341:2
jgi|Phyca11|503471|fgenes2_kg.PHYCAscaffold_4_#_10_#_gi|189084
jgi|Phyca11|503473|fgenes2_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|fgenes1_pm.PHYCAscaffold_4_#_8 neither
jgi|Phyca11|13487|fgenes1_pg.PHYCAscaffold_4_#_10 neither
jgi|Phyca11|99638|e_gw1.4.611.1 neither
jgi|Phyca11|13489|fgenes1_pg.PHYCAscaffold_4_#_12 neither
jgi|Phyca11|503477|fgenes2_kg.PHYCAscaffold_4_#_16_#_4099940:1
jgi|Phyca11|4923|fgenes1_pm.PHYCAscaffold_4_#_11 neither
jgi|Phyca11|503479|fgenes2_kg.PHYCAscaffold_4_#_18_#_gi|189083
jgi|Phyca11|539773|estExt2_Genewise1Plus.C.PHYCAscaffold_40033
jgi|Phyca11|503483|fgenes2_kg.PHYCAscaffold_4_#_22_#_Contig874
jgi|Phyca11|539778|estExt2_Genewise1Plus.C.PHYCAscaffold_40038
jgi|Phyca11|503485|fgenes2_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_fgenes1_pm.C.PHYCAscaffold_40015
jgi|Phyca11|525487|estExt2_fgenes1_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|fgenes1_pg.PHYCAscaffold_4_#_21 neither
jgi|Phyca11|525489|estExt2_fgenes1_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|fgenes1_pg.PHYCAscaffold_4_#_24 neither
jgi|Phyca11|503494|fgenes2_kg.PHYCAscaffold_4_#_33_#_gi|189084
jgi|Phyca11|503496|fgenes2_kg.PHYCAscaffold_4_#_35_#_Contig282
jgi|Phyca11|100304|e_gw1.4.1152.1 neither
jgi|Phyca11|503499|fgenes2_kg.PHYCAscaffold_4_#_38_#_Contig292
jgi|Phyca11|525496|estExt2_fgenes1_pm.C.PHYCAscaffold_40025
jgi|Phyca11|503501|fgenes2_kg.PHYCAscaffold_4_#_40_#_Contig435
jgi|Phyca11|79926|gw1.4.967.1 neither
```

History

Options ▾

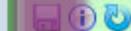


Unnamed history

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_fgenes1_pg.C.PHYCAscaffold_40001
jgi|Phyca11|80978|gw1.4.1034.1
jgi|Phyca11|99641|e_gw1.4.982.1
jgi|Phyca11|13479|fgenes1_pg.PHYCAscaffold_4_#_2
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_fgenes1_pg.C.PHYCAscaffold_40001
MGNVYSTDSSSSDTPQQQVERPEEKLHSLSDTTVT:
GIRYTDQETKRRKQGGNSPFLVSGVLTWIKACAGC
```

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 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				

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')

ℹ **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".

ℹ **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

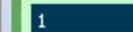
Options ▾



Unnamed history

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

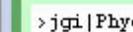
19,805 lines, 1 comments
format: tabular, database: ?
Info: Whisson2007 for 19805 sequences:
Y = 89, hmm = 43, neither = 19657, re = 16



1
#ID
jgi|Phyca11|532085|estExt2_fgenes|
jgi|Phyca11|80978|gw1.4.1034.1
jgi|Phyca11|99641|e_gw1.4.982.1
jgi|Phyca11|13479|fgenes1_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_fgenes|
MGNVYSTDSSSSDTQQQVERPEEKLHSLSDTTVT:
GIRYTDQETKRKQGGNSPFLVSGVLTWIKACAGC
```

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|Phyca11|5186|fgenesh1_pm.PHYCAscaffold_4_#
_274 Y
jgi|Phyca11|130393|e_gwl.93.9.1 Y
jgi|Phyca11|109432|e_gwl.16.672.1 Y
jgi|Phyca11|129643|e_gwl.86.164.1 Y
jgi|Phyca11|533084|estExt2_fgenesh1_pg.C_PHYCA
scaffold_100082 Y
jgi|Phyca11|14941|fgenesh1_pg.PHYCAscaffold_10
_#_87 Y
jgi|Phyca11|14944|fgenesh1_pg.PHYCAscaffold_10
_#_90 Y
jgi|Phyca11|14948|fgenesh1_pg.PHYCAscaffold_10
_#_94 Y
jgi|Phyca11|129113|e_gwl.81.47.1 Y
jgi|Phyca11|129145|e_gwl.81.173.1 Y
jgi|Phyca11|129044|e_gwl.81.43.1 Y
jgi|Phyca11|15117|fgenesh1_pg.PHYCAscaffold_11
_#_103 Y
jgi|Phyca11|102742|e_gwl.7.224.1 Y
jgi|Phyca11|116585|e_gwl.31.283.1 Y
jgi|Phyca11|116645|e_gwl.31.473.1 Y
jgi|Phyca11|39353|gwl.107.45.1 Y
jgi|Phyca11|97196|e_gwl.1.556.1 Y
jgi|Phyca11|538116|estExt2_GenewiselPlus.C_PHY
CAscaffold_10381 Y
jgi|Phyca11|4454|fgenesh1_pm.PHYCAscaffold_2_#
_50 Y
jgi|Phyca11|118417|e_gwl.36.500.1 Y
jgi|Phyca11|103340|e_gwl.8.893.1 Y
jgi|Phyca11|20942|fgenesh1_pg.PHYCAscaffold_77
_#_10 Y
jgi|Phyca11|20944|fgenesh1_pg.PHYCAscaffold_77
_#_12 Y
jgi|Phyca11|102326|e_gwl.6.392.1 Y
jgi|Phyca11|101904|e_gwl.6.942.1 Y
jgi|Phyca11|14853|fgenesh1_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|fgenesh1_pm.PHYCAs
jgi|Phyca11|130393|e_gwl.93.9.1
jgi|Phyca11|109432|e_gwl.16.672.1
jgi|Phyca11|129643|e_gwl.86.164.1
jgi|Phyca11|533084|estExt2_fgenesh1
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

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 striking 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 cleavage 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|fgenesht1_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_fgenesht
jgi|Phyca11|14941|fgenesht1_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
***
```

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 striking 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 cleavage 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|fgenesht1_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_fgenesht
jgi|Phyca11|14941|fgenesht1_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
***
```

Next few steps omitted...

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

The screenshot shows the Galaxy web interface. At the top, the browser address bar displays "http://ppserver/galaxy/root". The main navigation bar includes "Galaxy" and "Analyze Data", "Workflow", "Shared Data", "Admin", "Help", and "User". On the left, a "Tools" sidebar lists various analysis tools, with "Venn Diagram from lists" highlighted. The central workspace displays a green notification box with a checkmark and the text "Attributes updated". On the right, a "History" panel shows a list of jobs, including "Phyca11 filtered proteins.fasta" and "Whisson et al. (2007) RXLR-EER with HMM". A purple callout box is overlaid on the bottom center of the interface.

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
- LAI 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

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

Attributes updated

History Options

Unnamed history

- 7: Bhattacherjee matches
- 6: Bhattacherjee 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_fgenes|  
MGNVYSTDSSTQQQVERPEEKLSLSDTTVT:  
GIRYTDEQTKRKQGGNSPFLVSGVLTWIKACAG  
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

Tools

Options ▾

Fetch SequencesFetch AlignmentsStatisticsWavelet AnalysisGraph/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
- LAI 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 VariationMultiple regressionMultivariate AnalysisEvolutionMetagenomic analysesFASTA manipulation

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_fgenes|
MGNVYSTDSSTQQQVERPEEKLSLSDTTVT:
GIRYTDQTKRKQGGNSPFLVSGVLTWIKACAG
LNDVVL
```

```
> jgi|Phyca11|80978|gw1.4.1034.1
DVFLDIGSGVGNVVAQFALSTKVRACIGIEIRRV
```

Save this analysis
as a workflow

Tools

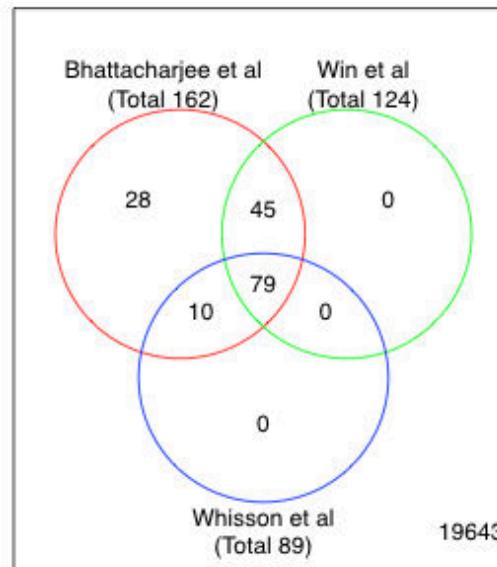
Options ▾

EXTRACT FEATURES[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](#)

Venn Diagram



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](#)

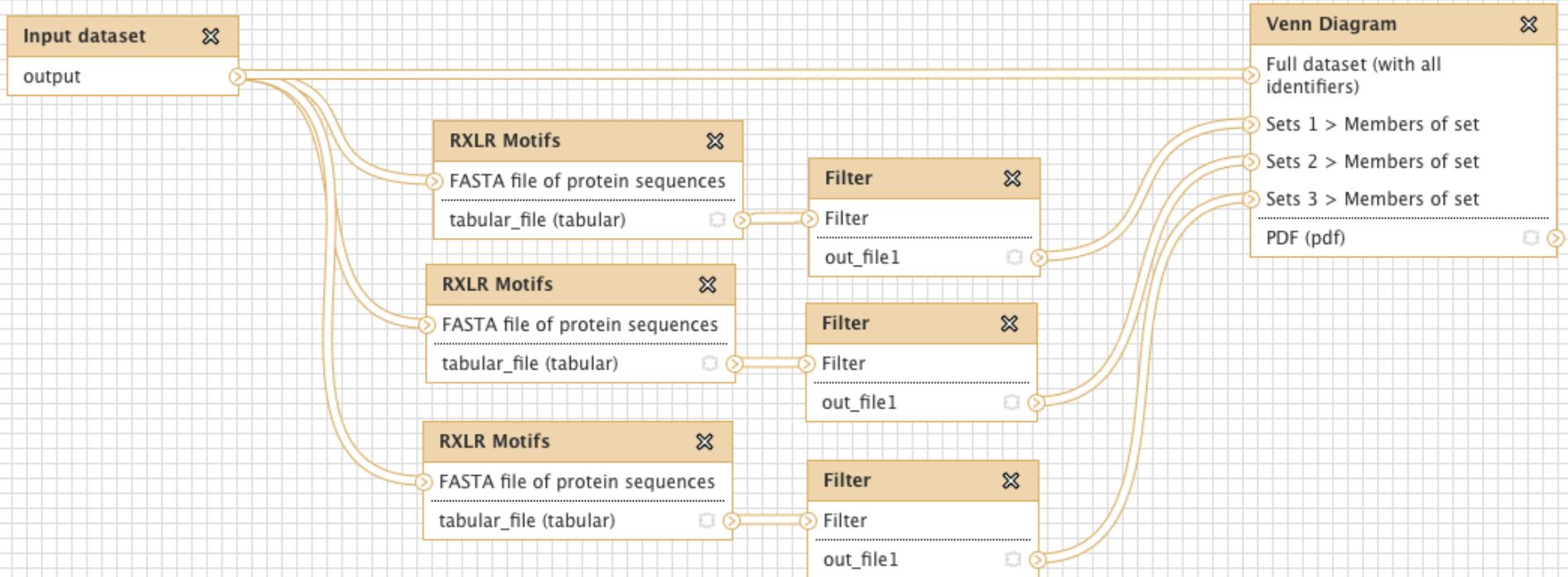
4: [Win et al. \(2007\) RXLR](#)

3: [Whisson matches](#)

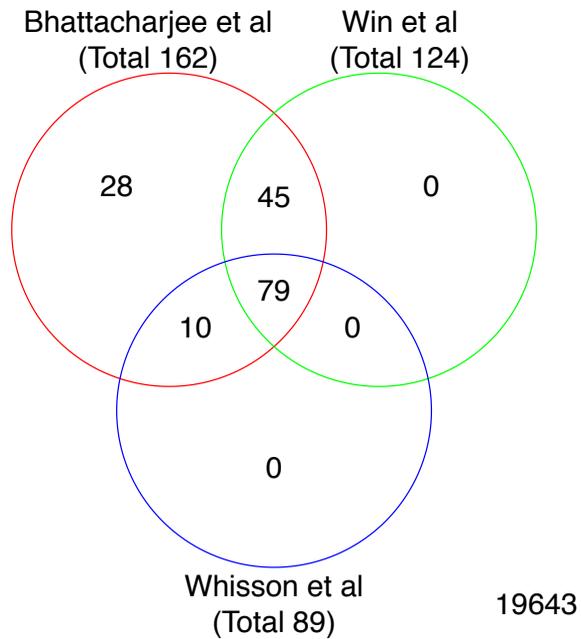
2: [Whisson et al. \(2007\) RXLR-EER with HMM](#)

1:

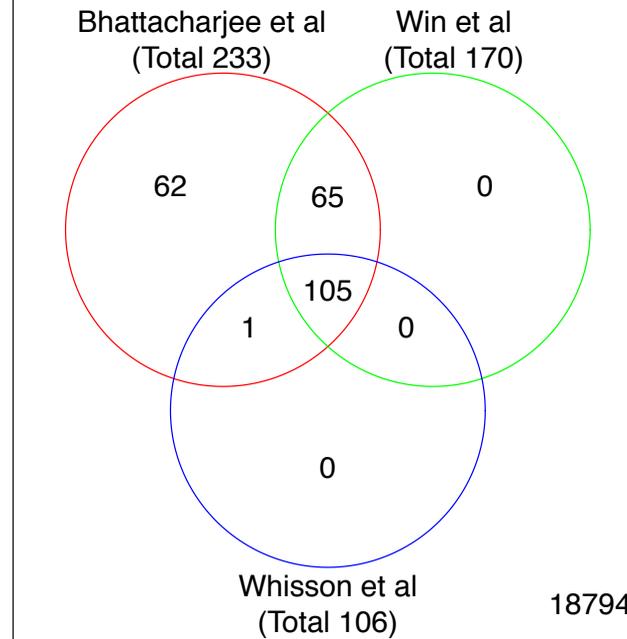
Workflow editor view



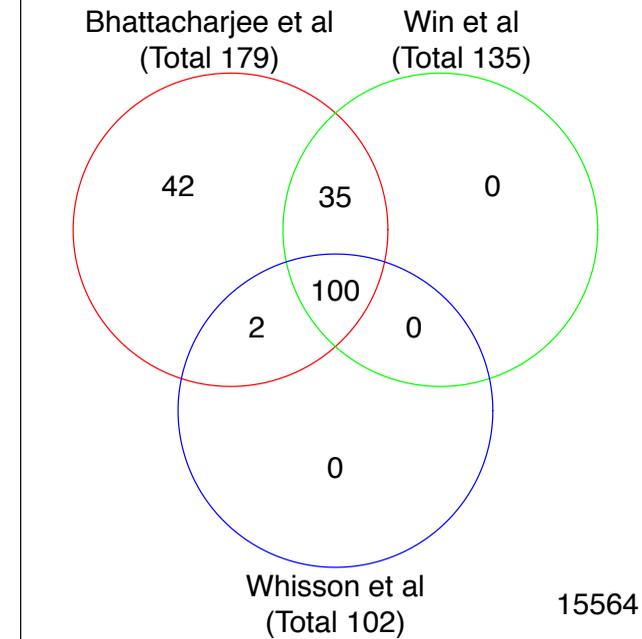
RXLRs in *Phytophthora* draft genomes



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

Conclusions

■ 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?