Popular bioinformatics tools in Galaxy: II

Yanbin Yin

Spring 2013

Homework assignment 5

Watch Galaxy 101 (http://screencast.g2.bx.psu.edu/galaxy101/) and other videos I put on course page 19-Feb

Use PSU Galaxy main server or glu galaxy server:

1. Get data from UCSC: phenotype and disease group -> cosmic track -> cosmicRaw table
2. Get data: phenotype and disease -> cosmic -> cosmic (chr22)
3. Google cosmic to learn about it
4. Join the two tables according to cosmic id, design operation flow to:
   - Find what chr22 genes have the most cancer mutations (e.g. show top 10)
   - Find what organs have the most cancer mutations in chr22
5. Create workflow and apply to four other chromosomes

Write a report (in word or ppt) to include all the operations and screen shots.

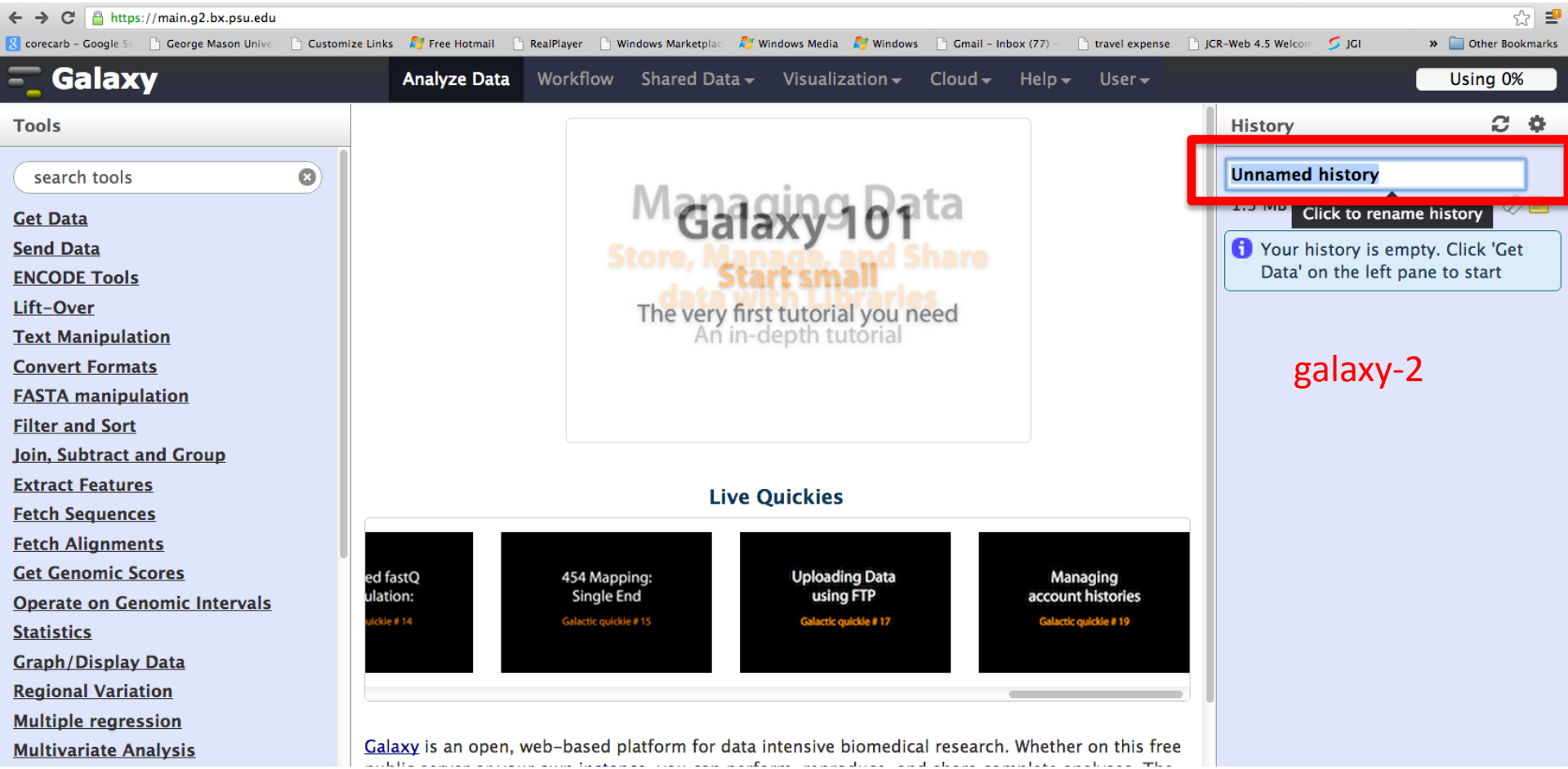
Due on March 5 (send by email)

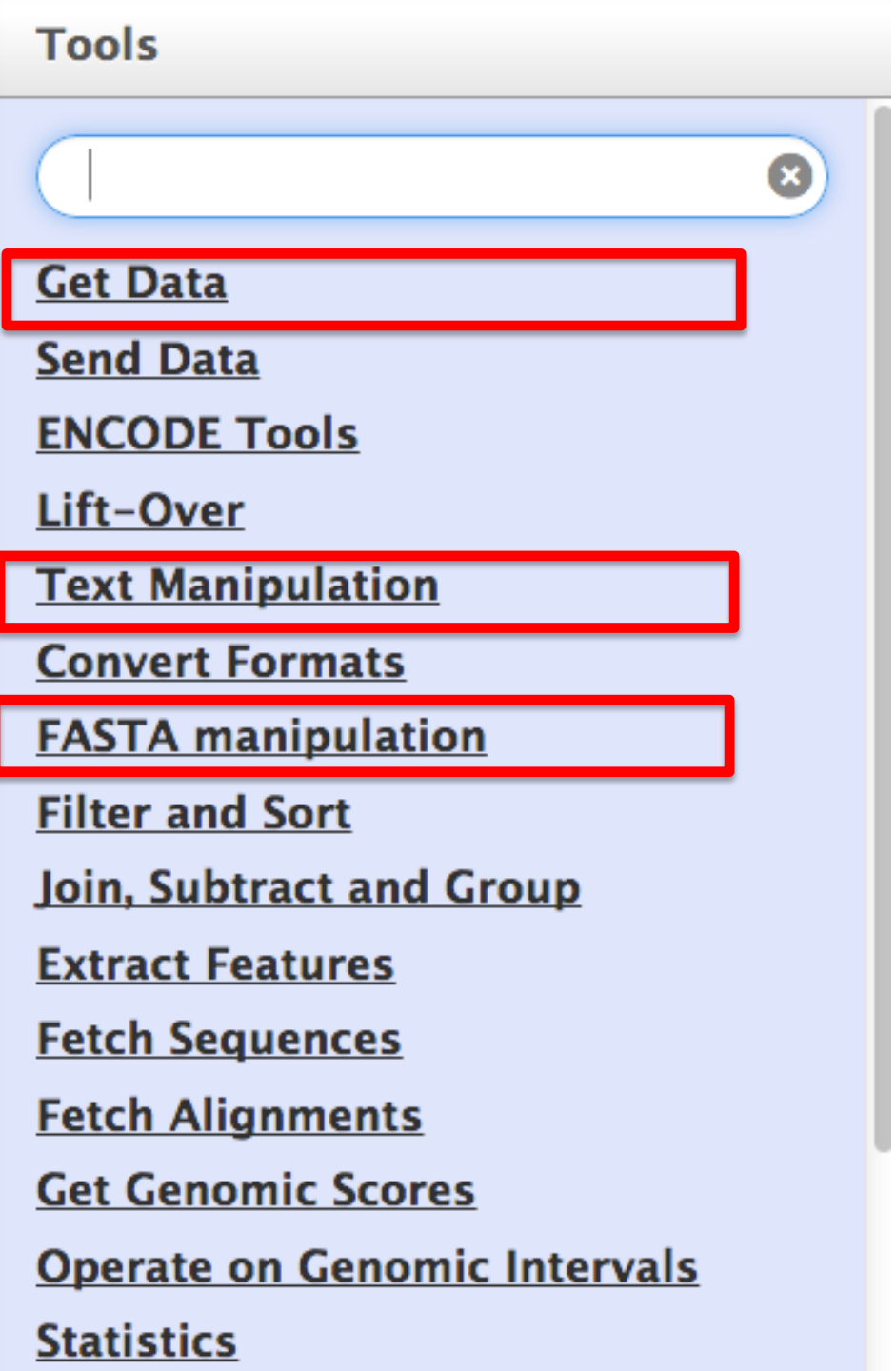
Office hour:
Tue, Thu and Fri 2-4pm, MO325A
Or email: yyin@niu.edu

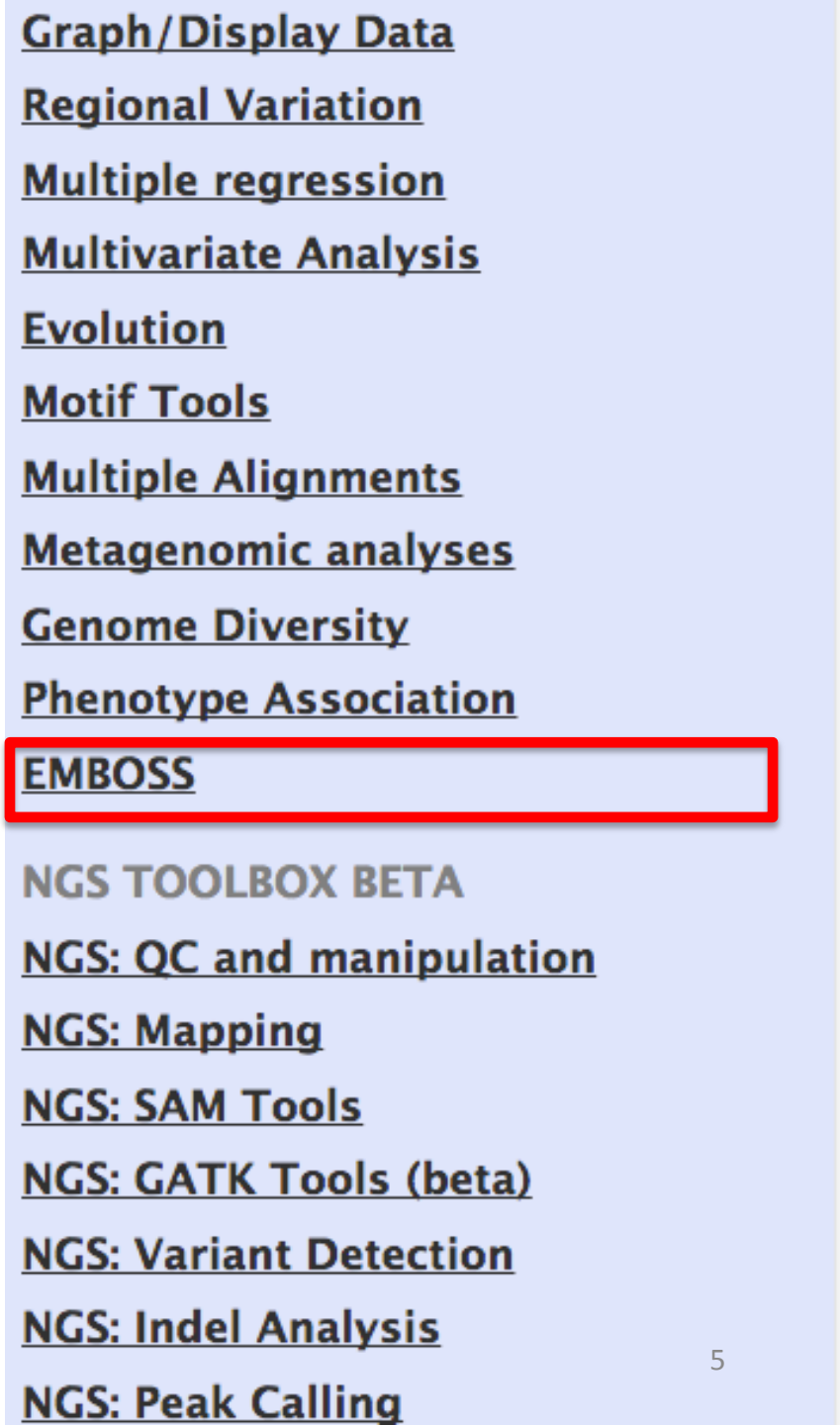
Outline

- Hands on practice
  - Introduce tools menu
  - How to upload/download
  - How to copy datasets from one history to another
  - EMBOSS

https://main.g2.bx.psu.edu

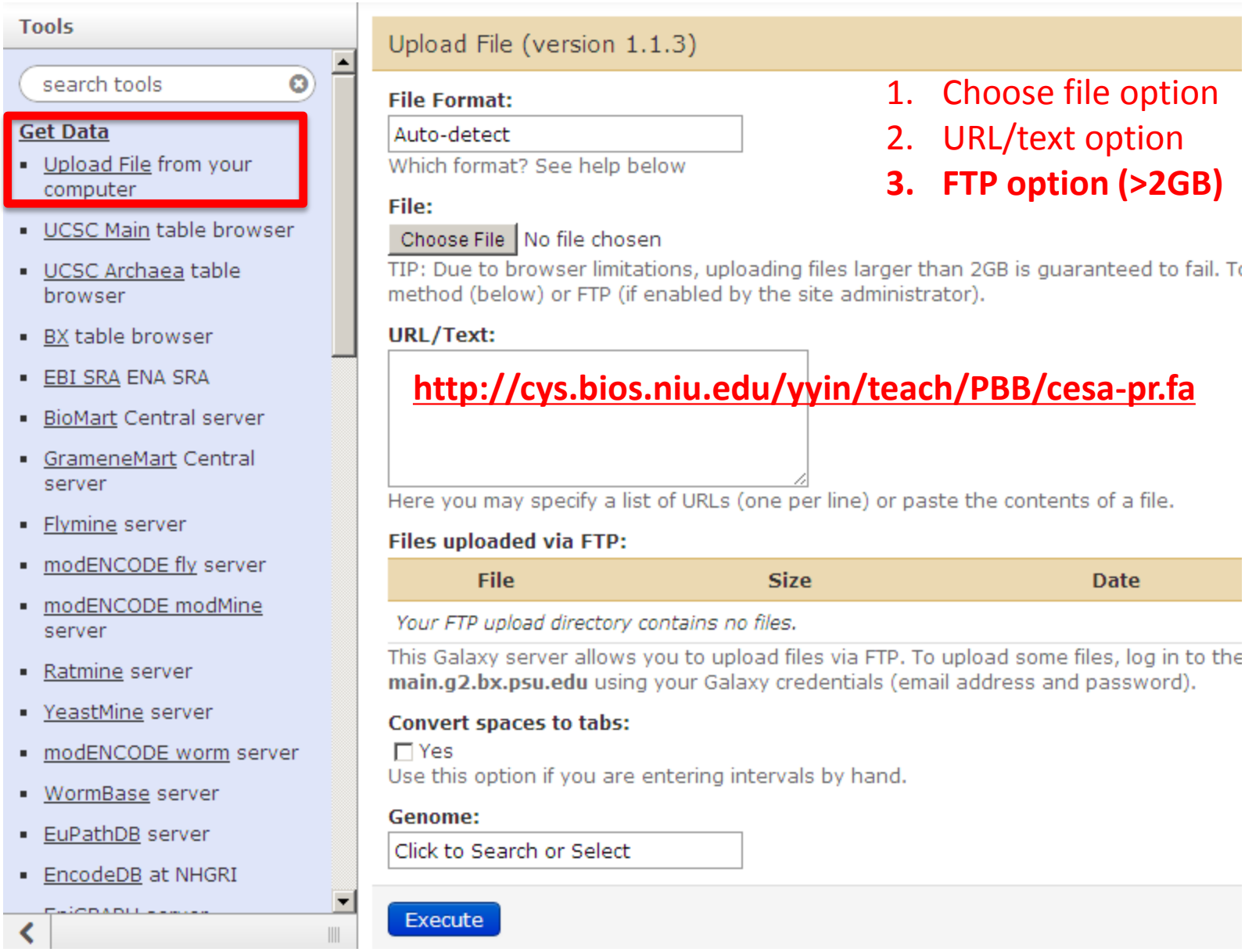



Tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics

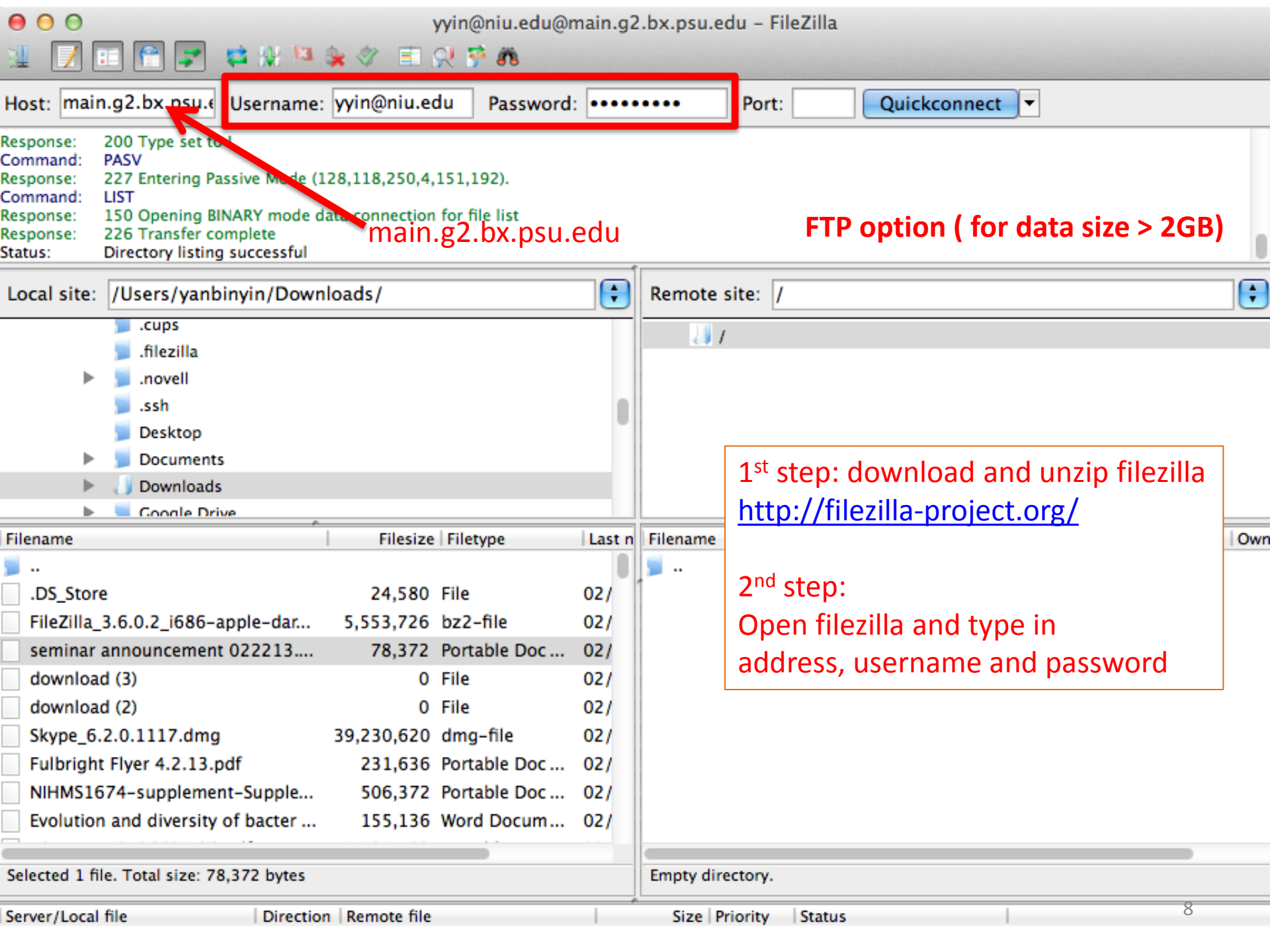

Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity
Phenotype Association
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection
NGS: Indel Analysis
NGS: Peak Calling

Hands on practice: getting data into Galaxy

https://main.g2.bx.psu.edu



http://screencast.g2.bx.psu.edu/quickie_17_ftp_upload/flow.html

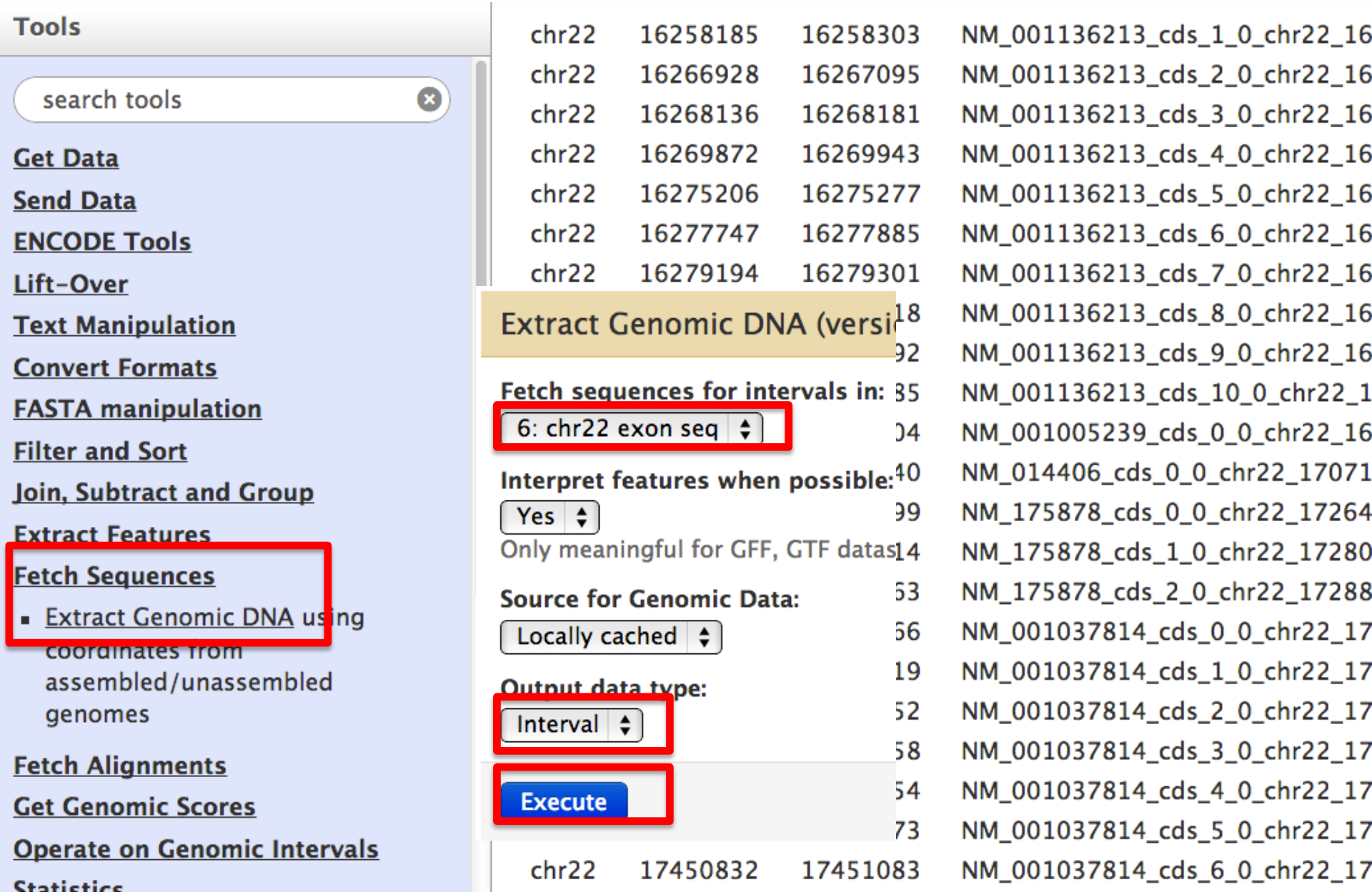

yyin@niu.edu@main.g2.bx.psu.edu - FileZilla
Host: main.g2.bx.psu.e
Username: yyin@niu.edu
Password: ••••••••
Port:
Quickconnect
Response: 200 Type set to I
Command: PASV
Response: 227 Entering Passive Mode (128,118,250,4,151,192).
Command: LIST
Response: 150 Opening BINARY mode data connection for file list
Response: 226 Transfer complete
Status: Directory listing successful
main.g2.bx.psu.edu
FTP option (for data size > 2GB)
Local site: /Users/yanbinyin/Downloads/
.cups
.filezilla
.novell
.ssh
Desktop
Documents
Downloads
Google Drive
Remote site: /
/
Filename Filesize Filetype Last n
.. .DS_Store 24,580 File 02/
FileZilla_3.6.0.2_i686-apple-dar... 5,553,726 bz2-file 02/
seminar announcement 022213.... 78,372 Portable Doc... 02/
download (3) 0 File 02/
download (2) 0 File 02/
Skype_6.2.0.1117.dmg 39,230,620 dmg-file 02/
Fulbright Flyer 4.2.13.pdf 231,636 Portable Doc... 02/
NIHMS1674-supplement-Supple... 506,372 Portable Doc... 02/
Evolution and diversity of bacter... 155,136 Word Docum... 02/
Selected 1 file. Total size: 78,372 bytes
Filename Own
..
Empty directory.
Server/Local file Direction Remote file Size Priority Status
1st step: download and unzip filezilla
http://filezilla-project.org/
2nd step:
Open filezilla and type in
address, username and password

FASTA manipulation

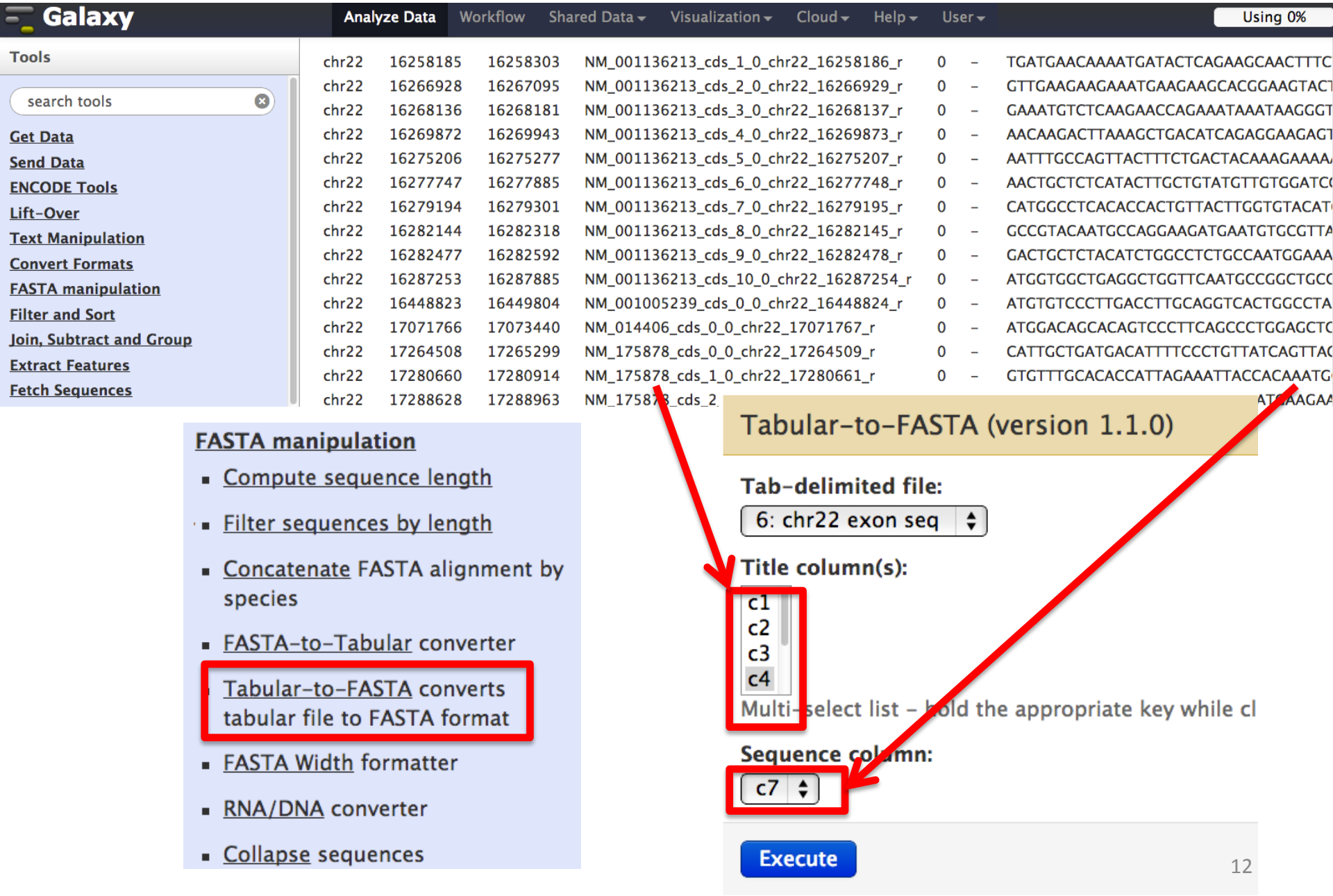
Goal: find how many exons on chr22 are shorter than 10nt

Get the chr22 exon position file

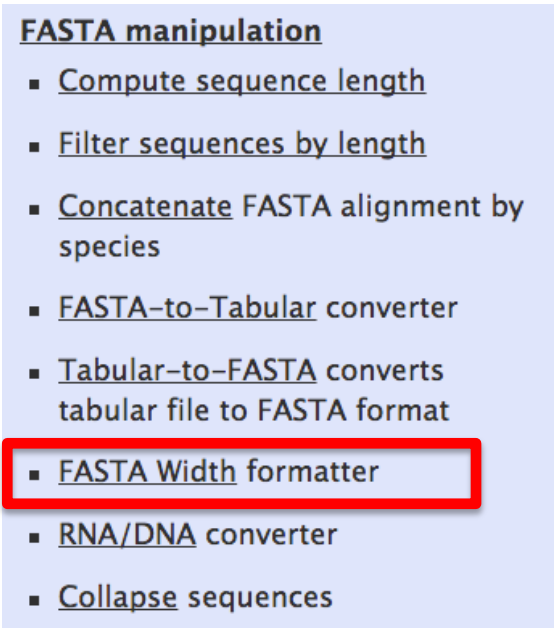
Get the exon sequences

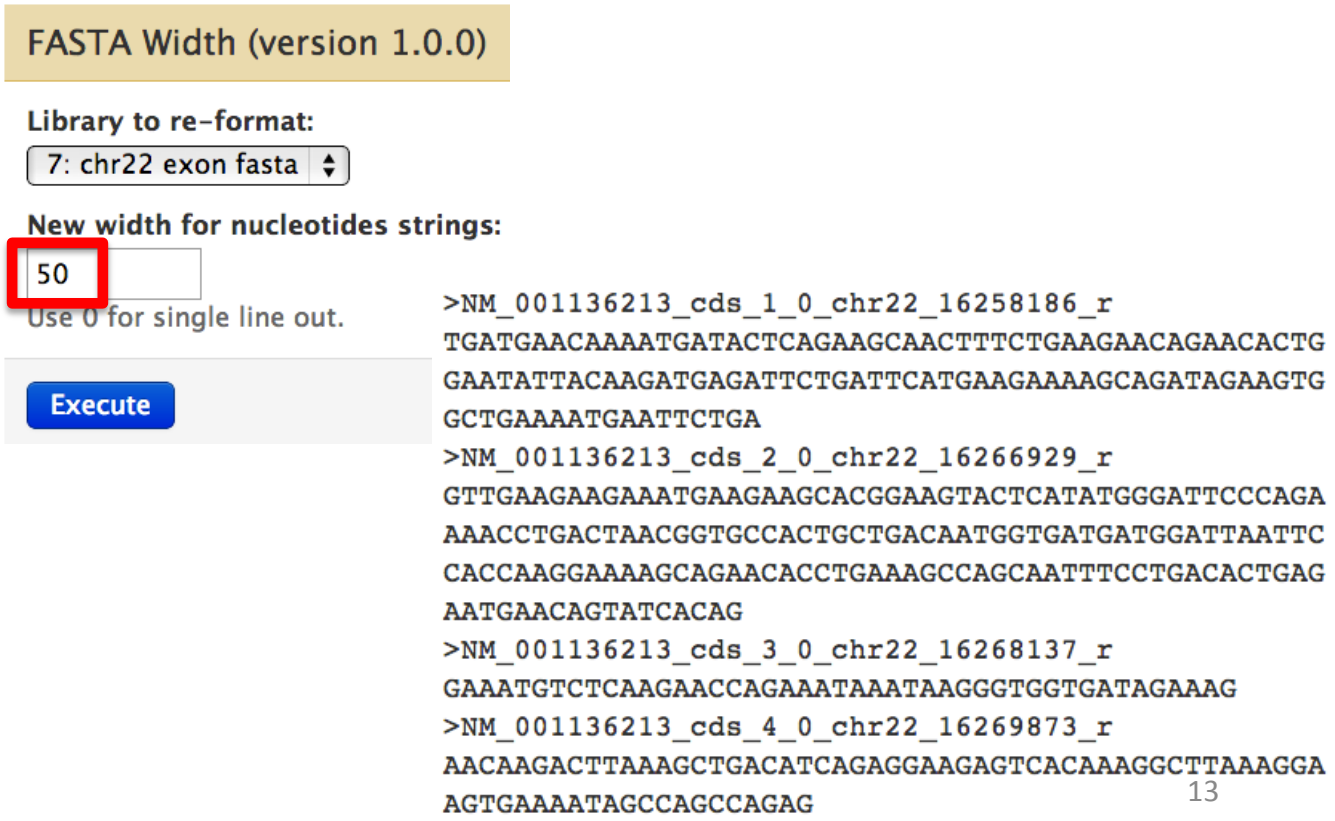


Convert to fasta sequence format



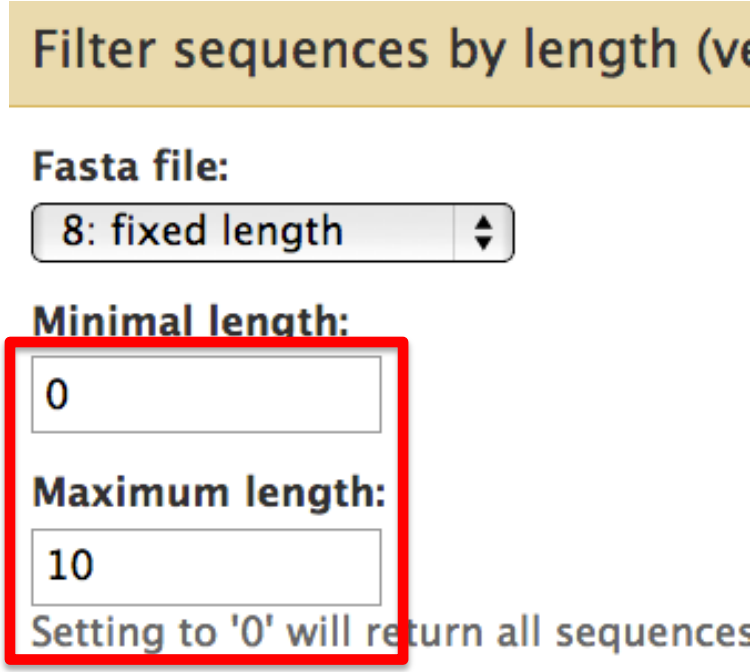
Format sequence with fixed width



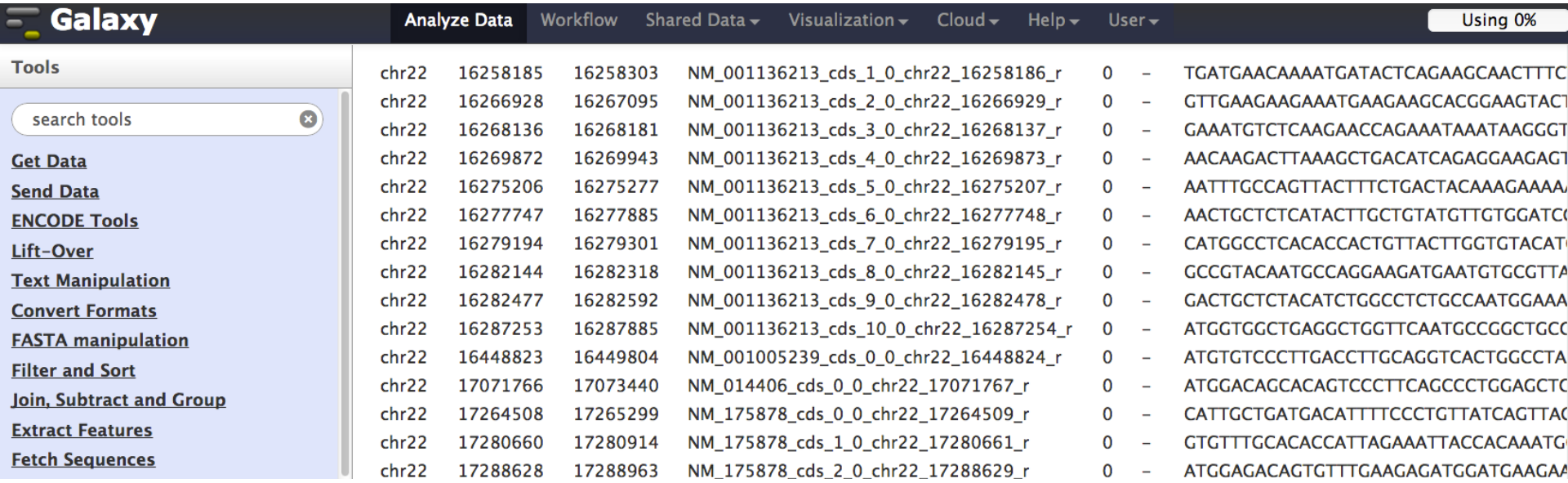


FASTA manipulation

- Compute sequence length
- Filter sequences by length
- Concatenate FASTA alignment by species
- FASTA-to-Tabular converter
- Tabular-to-FASTA converts tabular file to FASTA format
- FASTA Width formatter
- RNA/DNA converter
- Collapse sequences




```
>NM_031413_cds_17_0_chr22_18032535_f
AGCTAG
>NM_018943_cds_0_0_chr22_18593632_f
ATG
>NM_001035247_cds_0_0_chr22_19438263_r
GATGA
>NM_001035247_cds_11_0_chr22_19466606_r
ATG
>NM_005659_cds_11_0_chr22_19466606_r
ATG
>NM_000407_cds_0_0_chr22_19711093_f
ATGGGCTCCG
>NM_016449_cds_0_0_chr22_23950951_r
GGTGA
```



Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression
- Filter on ambiguities in polymorphism datasets

Filter (version 1.1.0)

Filter:

6: chr22 exon seq

Dataset missing? See TIP below.

With following condition:

len(c7)<= 10

Double equal signs, ==, must be used a
Select tool.

Number of header lines to skip:

0

Execute

The other way to find out: using filter option

Hands on practice: Text manipulation

Goal: find what genes on chr22 has the largest number of exons

Add one more column

chr22	16258185	16258303	NM_001136213_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	NM_001136213_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	NM_001136213_cds_3_0_chr22_16268137_r	0	-
chr22	16269872	16269943	NM_001136213_cds_4_0_chr22_16269873_r	0	-
chr22	16275206	16275277	NM_001136213_cds_5_0_chr22_16275207_r	0	-
chr22	16277747	16277885	NM_001136213_cds_6_0_chr22_16277748_r	0	-
chr22	16279194	16279301	NM_001136213_cds_7_0_chr22_16279195_r	0	-
chr22	16282144	16282318	NM_001136213_cds_8_0_chr22_16282145_r	0	-

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together

Compute (version 1.1.0)

Add expression:

c3-c2

as a new column to:

5: chr22 exon

Dataset missing? See TIP below

Round result?:

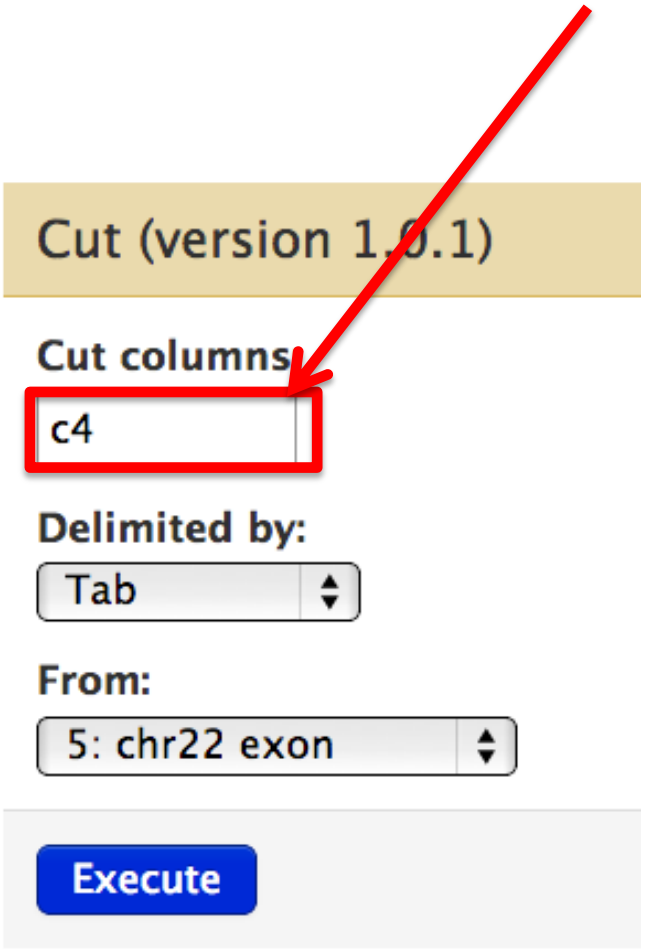
NO

Execute

chr22	16258185	16258303	NM_001136213_cds_1_0_chr22_16258186_r	0	-
chr22	16266928	16267095	NM_001136213_cds_2_0_chr22_16266929_r	0	-
chr22	16268136	16268181	NM_001136213_cds_3_0_chr22_16268137_r	0	-
chr22	16269872	16269943	NM_001136213_cds_4_0_chr22_16269873_r	0	-
chr22	16275206	16275277	NM_001136213_cds_5_0_chr22_16275207_r	0	-
		6277885	NM_001136213_cds_6_0_chr22_16277748_r	0	-
		6279301	NM_001136213_cds_7_0_chr22_16279195_r	0	-
		6282318	NM_001136213_cds_8_0_chr22_16282145_r	0	-

Text Manipulation

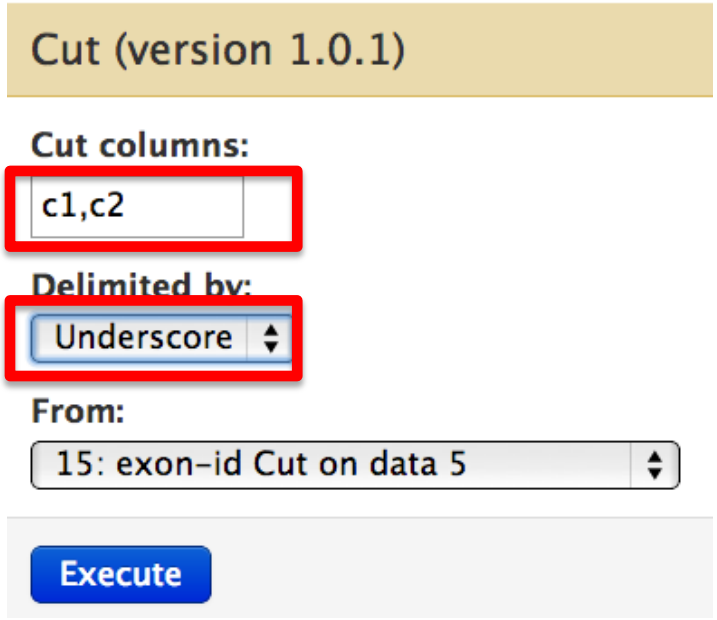
- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together
- Create single interval as a new dataset
- Cut columns from a table
- Change Case of selected columns



This column has gene ID and exon info

We can cut out this gene ID part and count how many times each ID appear

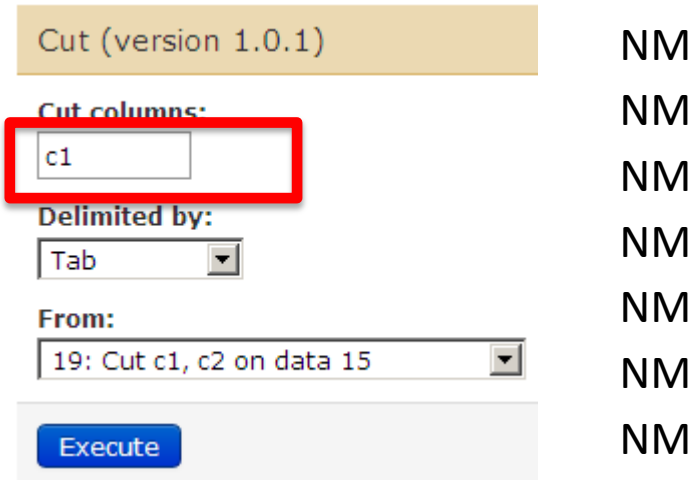
Further cut the gene ID part
The delimiter is NOT tab!!!



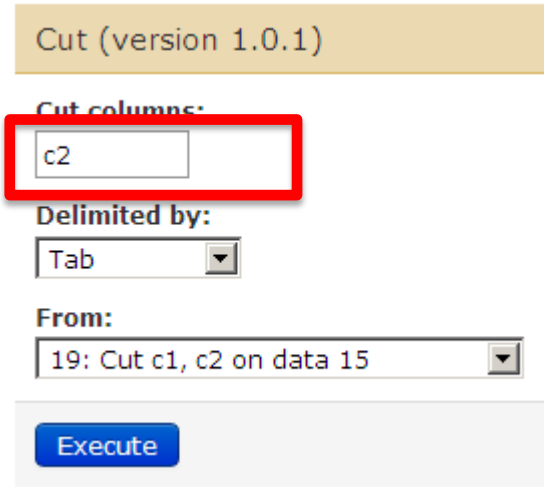

```
NM_001136213_cds_1_0_chr22_16258186_r
NM_001136213_cds_2_0_chr22_16266929_r
NM_001136213_cds_3_0_chr22_16268137_r
NM_001136213_cds_4_0_chr22_16269873_r
NM_001136213_cds_5_0_chr22_16275207_r
NM_001136213_cds_6_0_chr22_16277748_r
NM_001136213_cds_7_0_chr22_16279195_r
```

```
NM    001136213
NM    001136213
NM    001136213
NM    001136213
NM    001136213
NM    001136213
NM    001136213
```

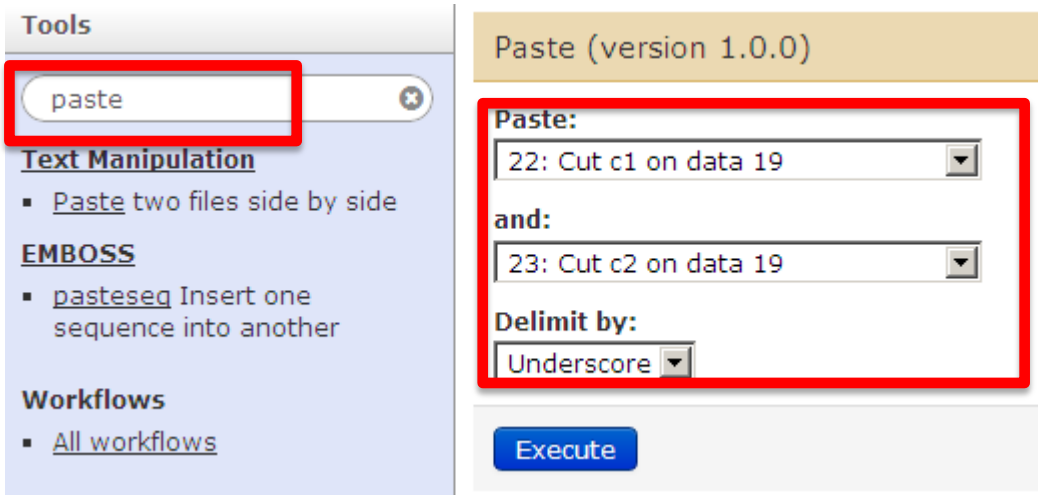
Continue to cut the two cols to save into two files and then join them using new delimiter



NM
NM
NM
NM
NM
NM
NM

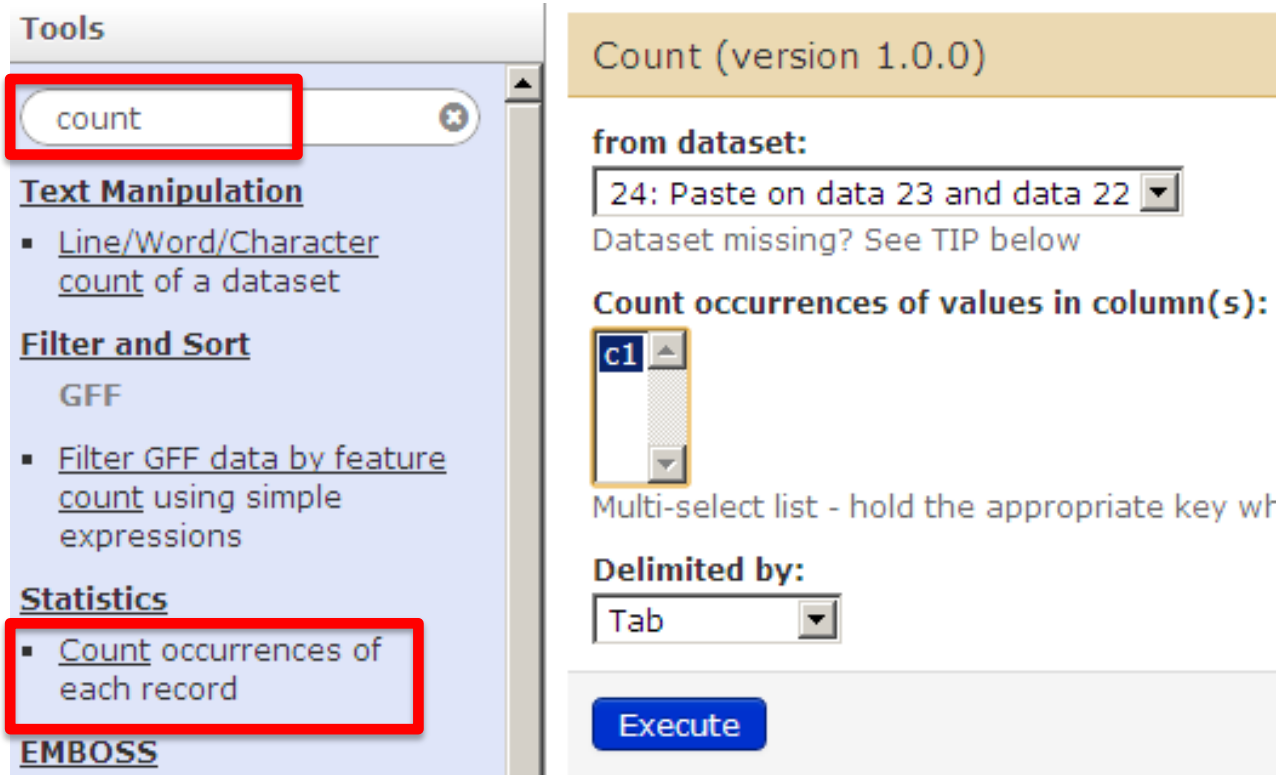


001136213
001136213
001136213
001136213
001136213
001136213
001136213
001136213

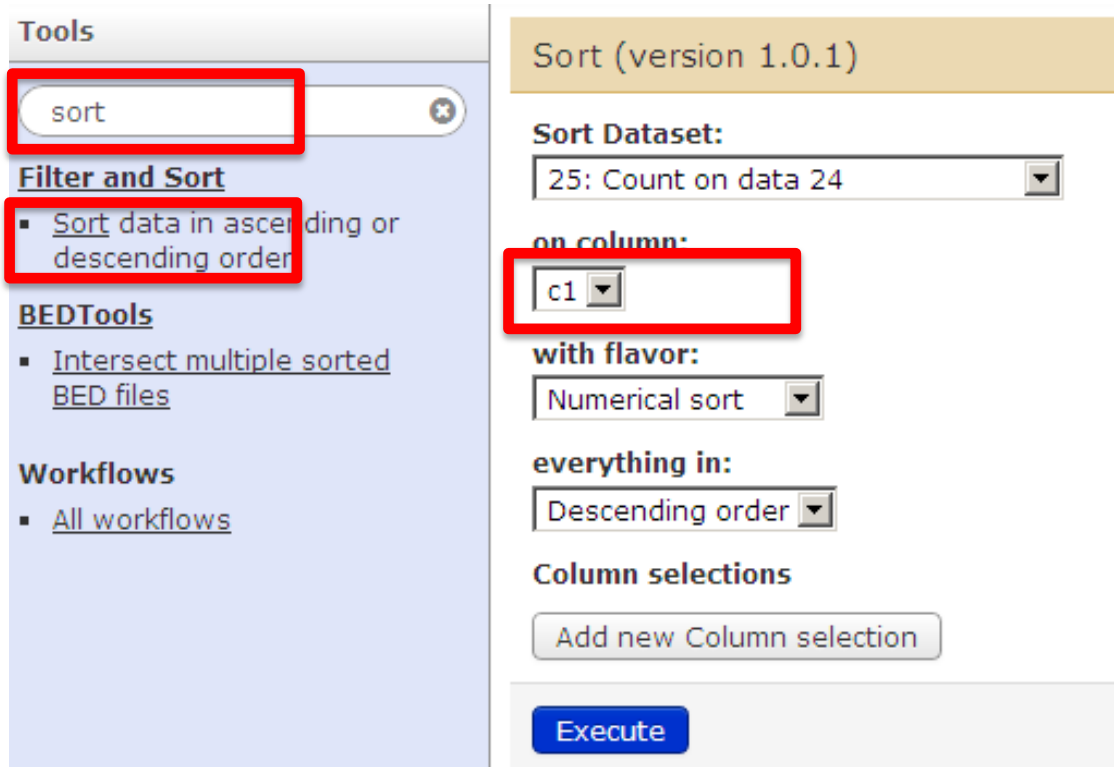


NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213
NM_001136213

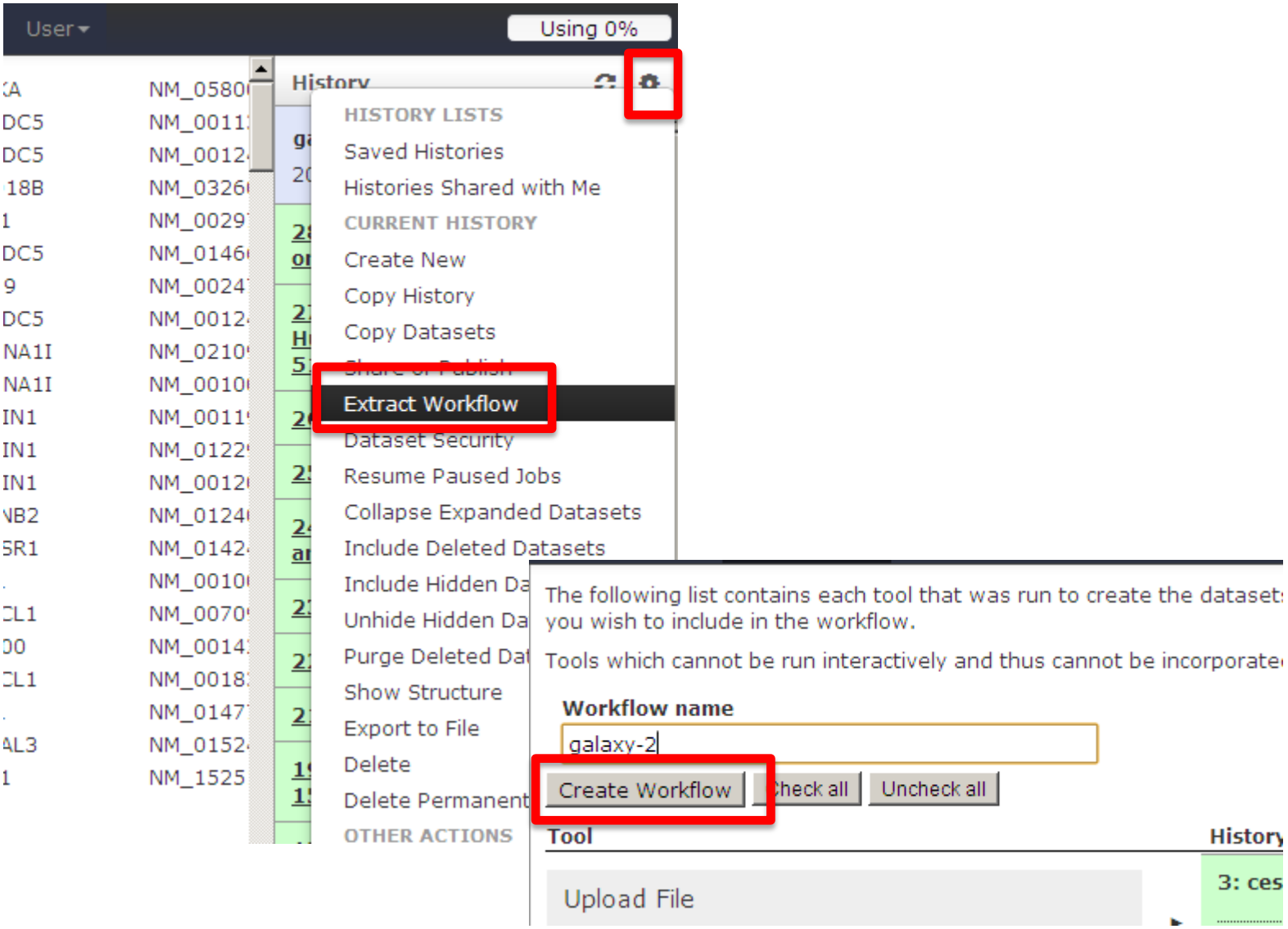
Now we can count the occurence !!!



13	NM_000026
9	NM_000106
4	NM_000185
9	NM_000262
16	NM_000268
15	NM_000343
9	NM_000355
5	NM_000362
13	NM_000395
9	NM_000398



55	NM_058004
42	NM_001136029
42	NM_001242896
42	NM_032608
41	NM_002972
41	NM_014662
40	NM_002473
39	NM_001242897
37	NM_021096

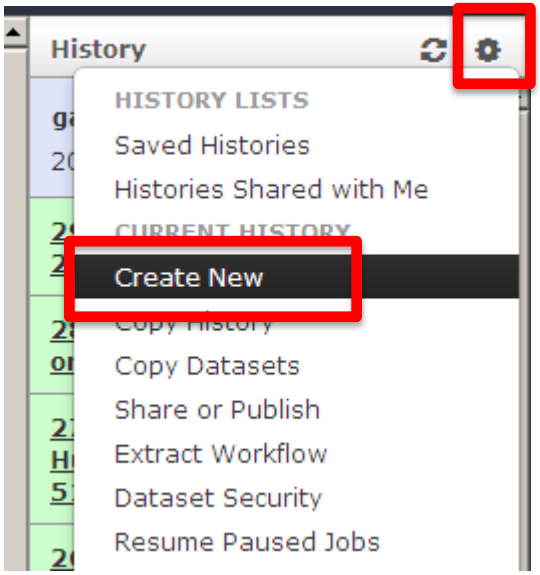

User
Using 0%
History
HISTORY LISTS
Saved Histories
Histories Shared with Me
CURRENT HISTORY
Create New
Copy History
Copy Datasets
Extract Workflow
Dataset Security
Resume Paused Jobs
Collapse Expanded Datasets
Include Deleted Datasets
Show Structure
Export to File
Delete
OTHER ACTIONS
The following list contains each tool that was run to create the dataset
you wish to include in the workflow.
Tools which cannot be run interactively and thus cannot be incorporate
Workflow name
galaxy-2
Create Workflow
Uncheck all
Tool
Upload File

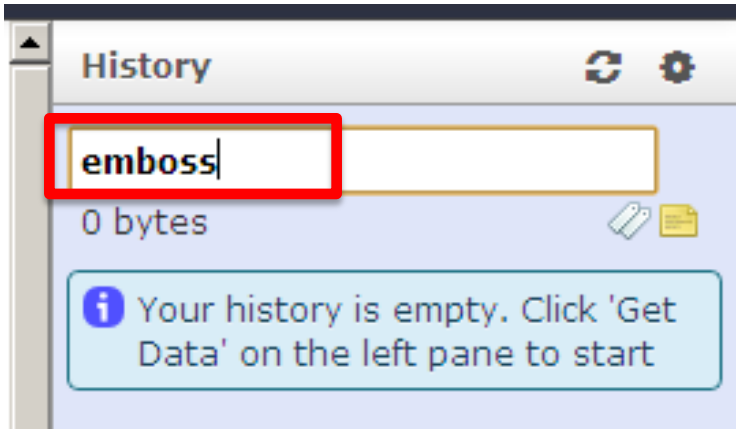
Try to run the work flow on Chr20 exons

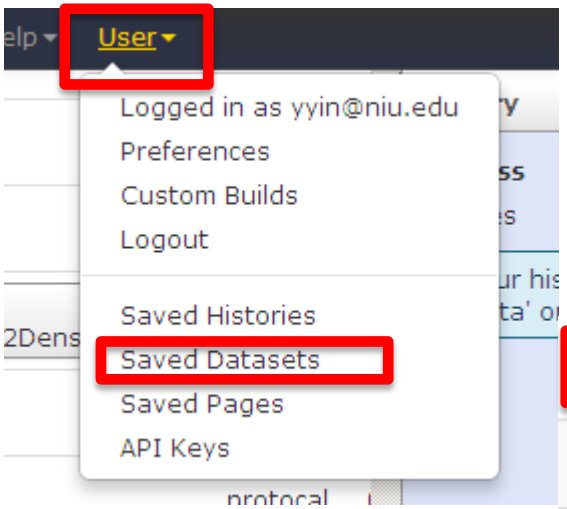
Hands on practice: EMBOSS

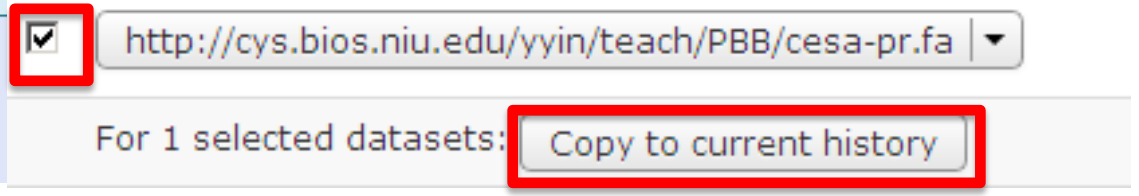
EMBOSS: European Molecular Biology Open Software Suite

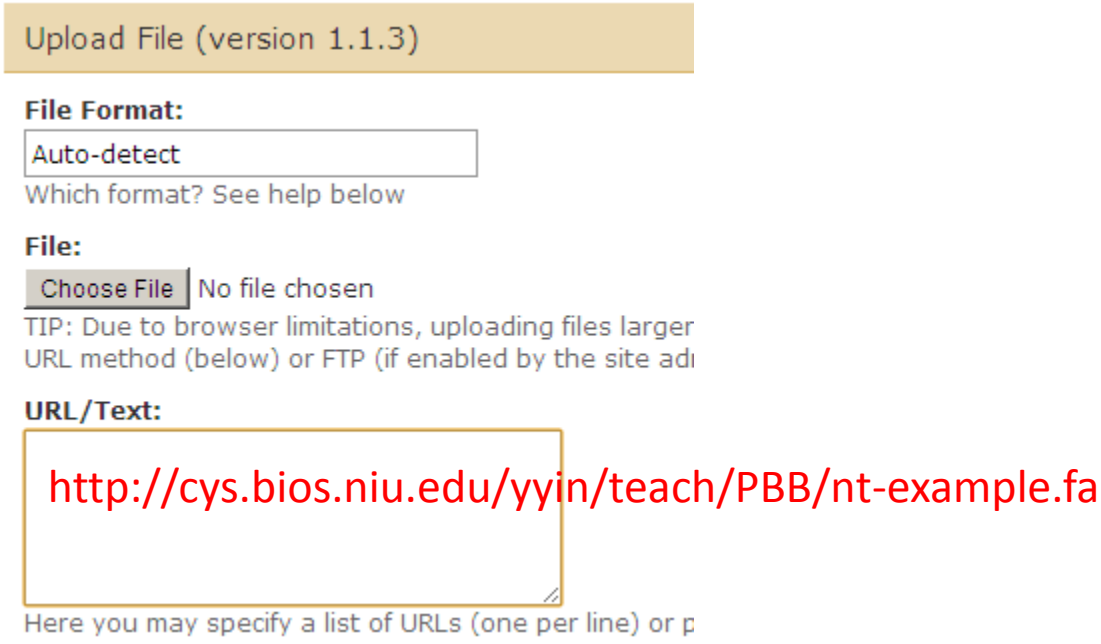
Create a new history

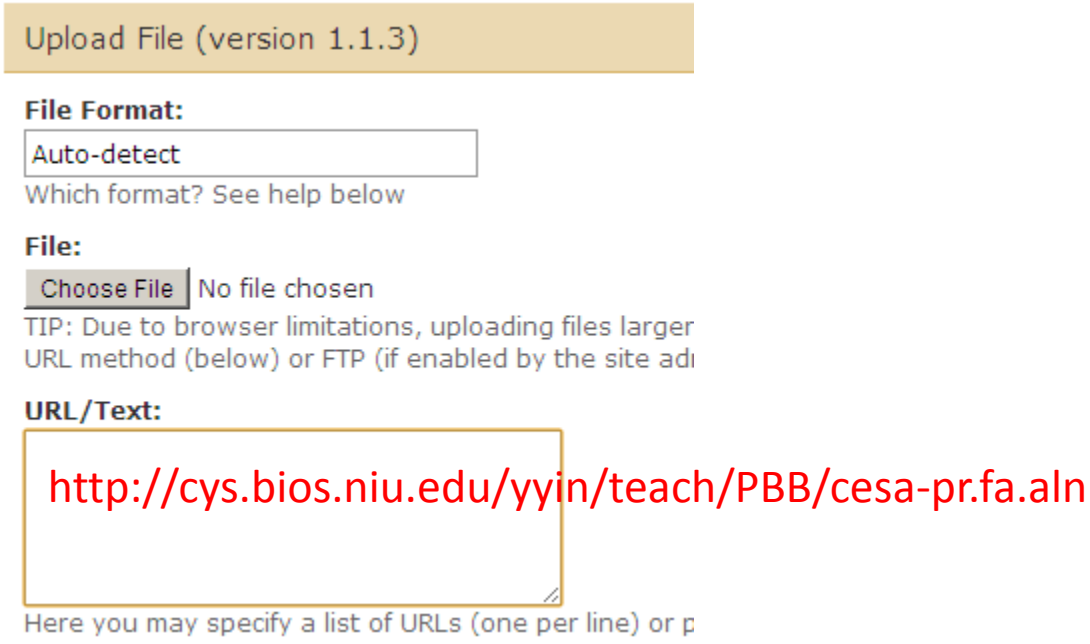




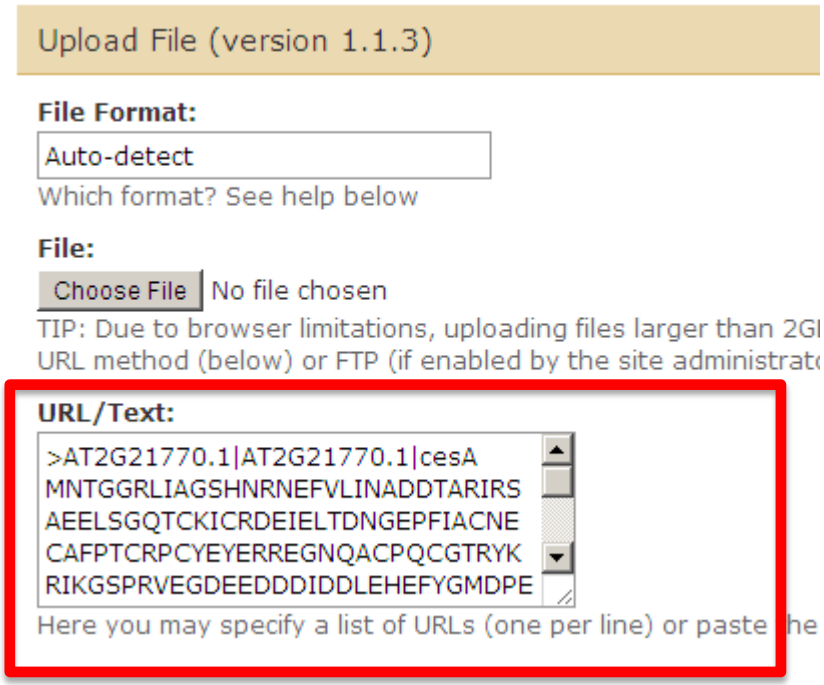


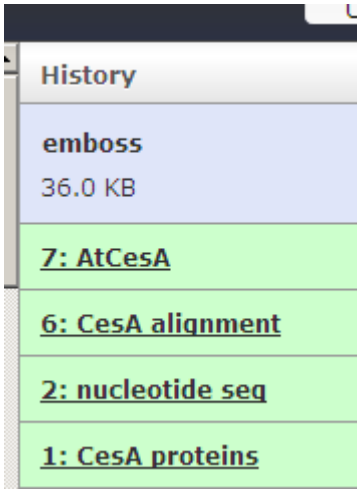


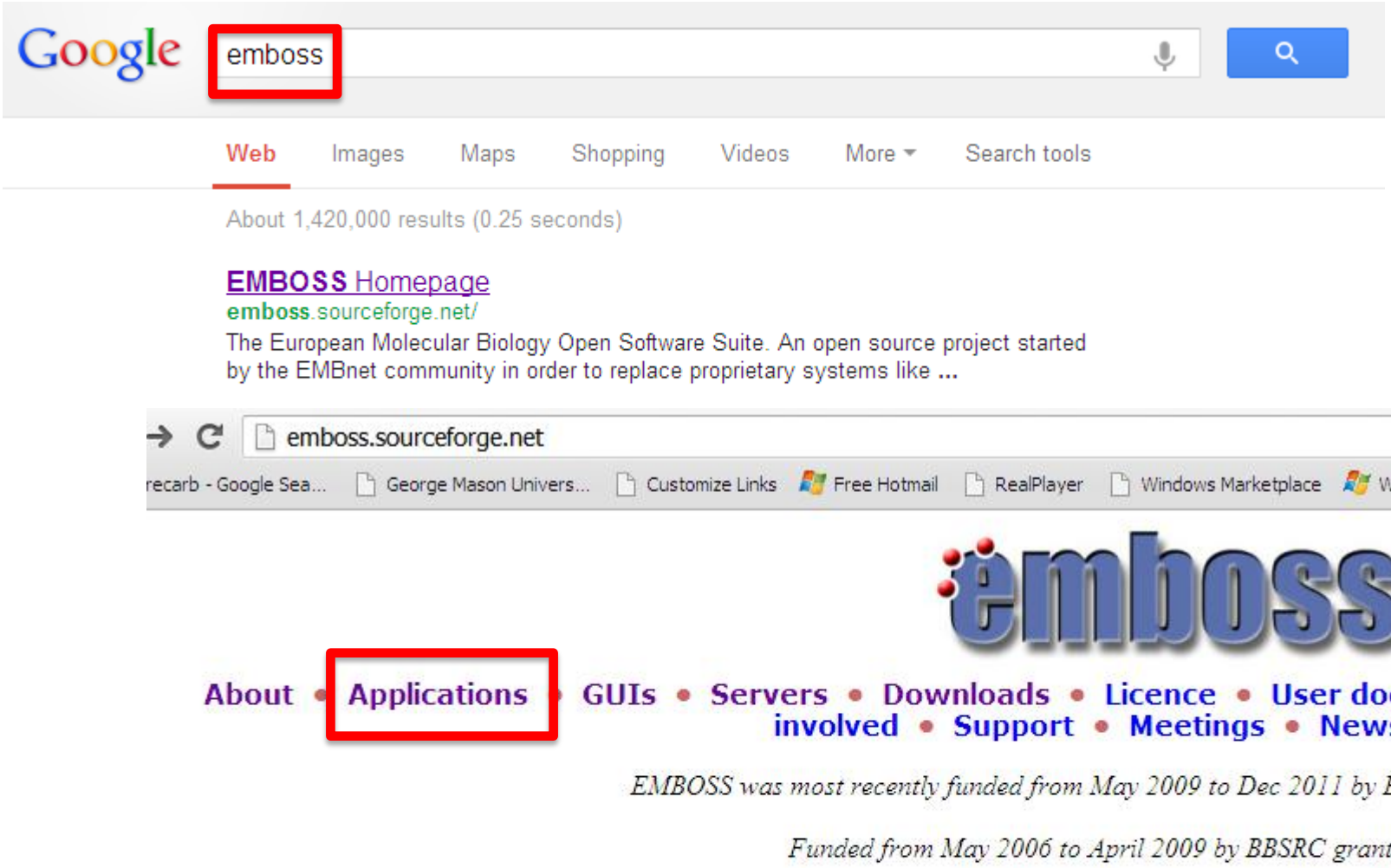

Upload File (version 1.1.3)
File Format:
Auto-detect
Which format? See help below
File:
Choose File
No file chosen
TIP: Due to browser limitations, uploading files larger
URL method (below) or FTP (if enabled by the site ad
URL/Text:
http://cys.bios.niu.edu/yyin/teach/PBB/nt-example.fa
Here you may specify a list of URLs (one per line) or p


Upload File (version 1.1.3)
File Format:
Auto-detect
Which format? See help below
File:
Choose File
No file chosen
TIP: Due to browser limitations, uploading files larger
URL method (below) or FTP (if enabled by the site ad
URL/Text:
http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa.aln
Here you may specify a list of URLs (one per line) or p

Copy & past only the first protein seq






Google
emboss
Web
Images
Maps
Shopping
Videos
More
Search tools
About 1,420,000 results (0.25 seconds)
EMBOSS Homepage
emboss.sourceforge.net/
The European Molecular Biology Open Software Suite. An open source project started by the EMBnet community in order to replace proprietary systems like ...
emboss.sourceforge.net
recarb - Google Sea...
George Mason Univers...
Customize Links
Free Hotmail
RealPlayer
Windows Marketplace
Windows Media
emboss
About
Applications
GUIs
Servers
Downloads
Licence
User docs
involved
Support
Meetings
News
Credits
EMBOSS was most recently funded from May 2009 to Dec 2011 by BBSRC grant BB
Funded from May 2006 to April 2009 by BBSRC grant BB/D018358/1
About EMBOSS
Overview
Uses
FAQ
Citing EMBOSS
A high-quality package of free, Open Source software for molecular biology ... more >
Applications
EMBOSS
EMBASSY
Groups
Proposed

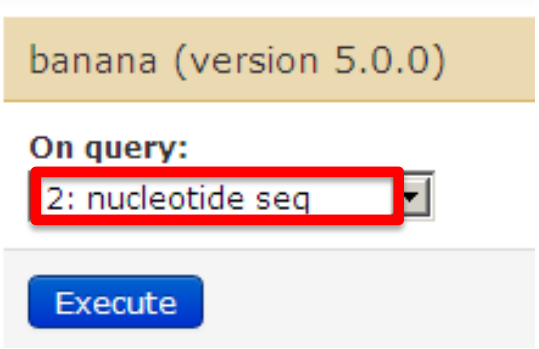
EMBOSS Applications

Contents

Hundreds of useful commands

Group	Description
Acd	Acd file utilities
Alignment	Sequence comparison and alignment
Alignment consensus	Merging sequences to make a consensus
Alignment differences	Finding differences between sequences
Alignment dot plots	Dot plot sequence comparisons
Alignment global	Global sequence alignment
Alignment local	Local sequence alignment
Alignment multiple	Multiple sequence alignment
Assembly fragment assembly	DNA sequence assembly
Data resources	Data resources
Data retrieval	Data retrieval
Data retrieval chemistry data	Chemistry data retrieval
Data retrieval feature data	Sequence feature data retrieval
Data retrieval ontology data	Ontology data retrieval
Data retrieval resource data	Resource data retrieval
Data retrieval sequence data	Sequence data retrieval

- banana Bending and curvature plot in B-DNA

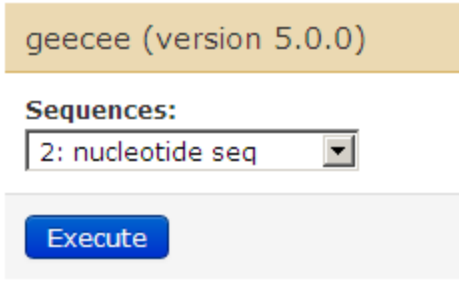


Base	Bend	Curve
t	0.0	0.0
A	11.7	0.0
A	13.9	0.0
G	14.1	0.0
A	13.3	0.0
T	10.6	0.0
A	14.9	0.0
C	17.7	0.0
C	17.7	0.0
T	22.3	0.0
C	26.9	0.0
G	18.5	0.0
A	5.0	0.0
A	1.3	0.0
A	5.9	0.0
T	9.2	0.0
A	5.9	0.0
T	1.3	0.0
T	0.0	0.0
T	3.4	0.0
T	7.8	4.2
A	5.9	5.4
T	1.3	6.8
T	5.7	7.8
T	15.3	8.6
G	19.7	9.4
C	20.7	10.3
A	12.4	11.4

banana predicts bending of a normal (B) DNA double helix, using the method of Goodsell & Dickerson, NAR 1994 11;22(24):5497-5503. The program calculates the magnitude of local bending and macroscopic curvature at each point along an arbitrary B-DNA sequence

http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/banana.html

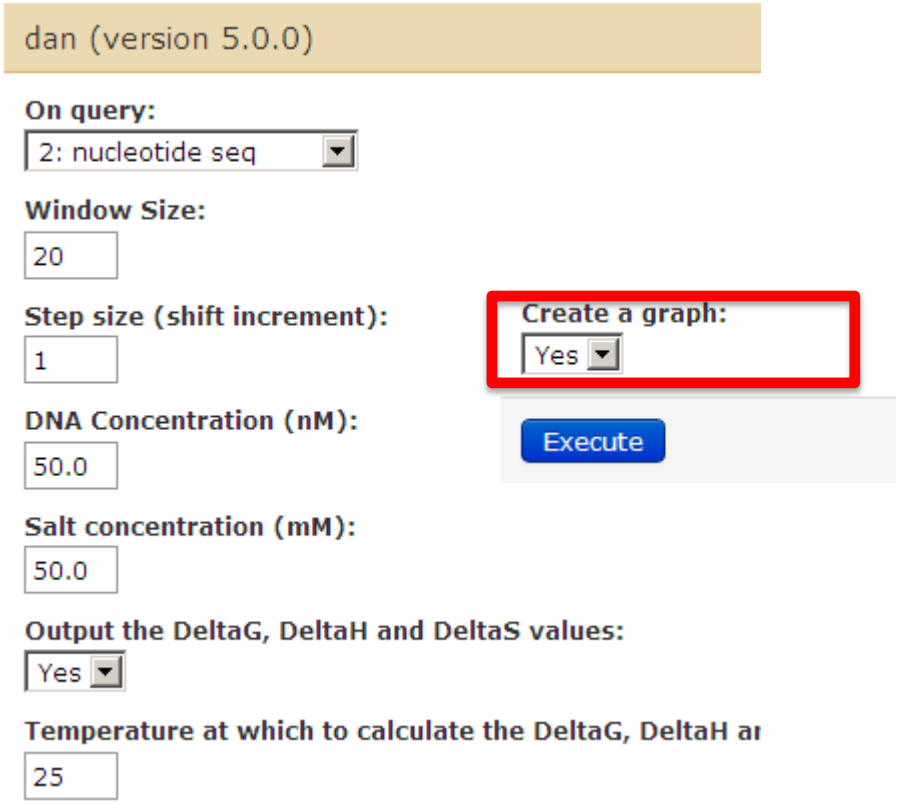
- geecee Calculates fractional GC content of nucleic acid sequences




```
#Sequence    GC content
contig00008    0.46
```

http://emboss.sourceforge.net/apps/cvs/emboss/apps/geecee.html

- dan Calculates DNA RNA/DNA melting temperature

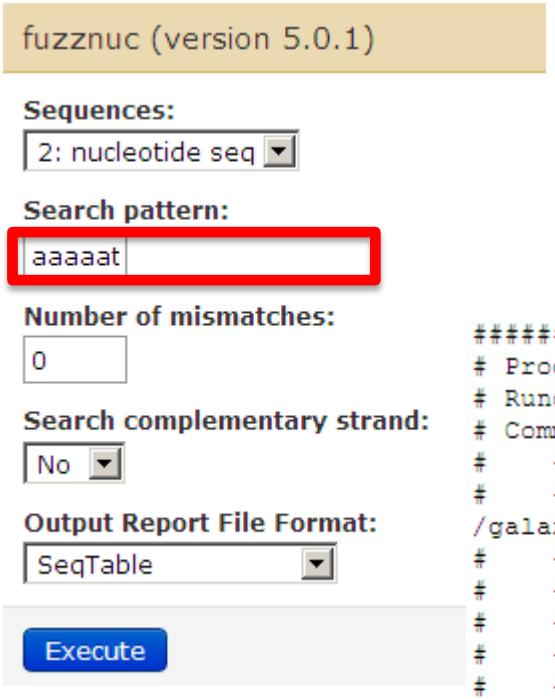


Dan calculates the melting temperature (Tm) and the percentage of G+C nucleotides for windows over a nucleic acid sequence, optionally plotting them.

http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/dan.html

- fuzznuc Nucleic acid pattern search

fuzznuc searches for a specified PROSITE-style pattern in nucleotide sequences. They can specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence. One or more nucleotide sequences are read from file. The output is a standard EMBOSS report file that includes data such as location and score of any matches




```
########################################
# Program: fuzznuc
# Rundate: Tue 19 Feb 2013 00:37:18
# Commandline: fuzznuc
#    -sequence /galaxy/main_pool/pool6/files/00
#    -outfile
/galaxy/main_pool/pool4/tmp/job_working_directo
#    -pattern aaaaat
#    -pmismatch 0
#    -complement no
#    -rformat2 seqtable
#    -auto
# Report_format: seqtable
# Report_file:
/galaxy/main_pool/pool4/tmp/job_working_directo
########################################

#=======================================
#
# Sequence: contig00008     from: 1   to: 862
# HitCount: 1
#
# Pattern_name Mismatch Pattern
# pattern1             0 aaaaat
#
# Complement: No
#
#=======================================

Start      End Pattern_name Mismatch Sequence
   95      100 pattern1             . AAAAAT
```

http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/fuzznuc.html

- plotorf Plot potential open reading frames

plotorf (version 5.0.0)

Sequence:

2: nucleotide seq

Start codons:

ATG

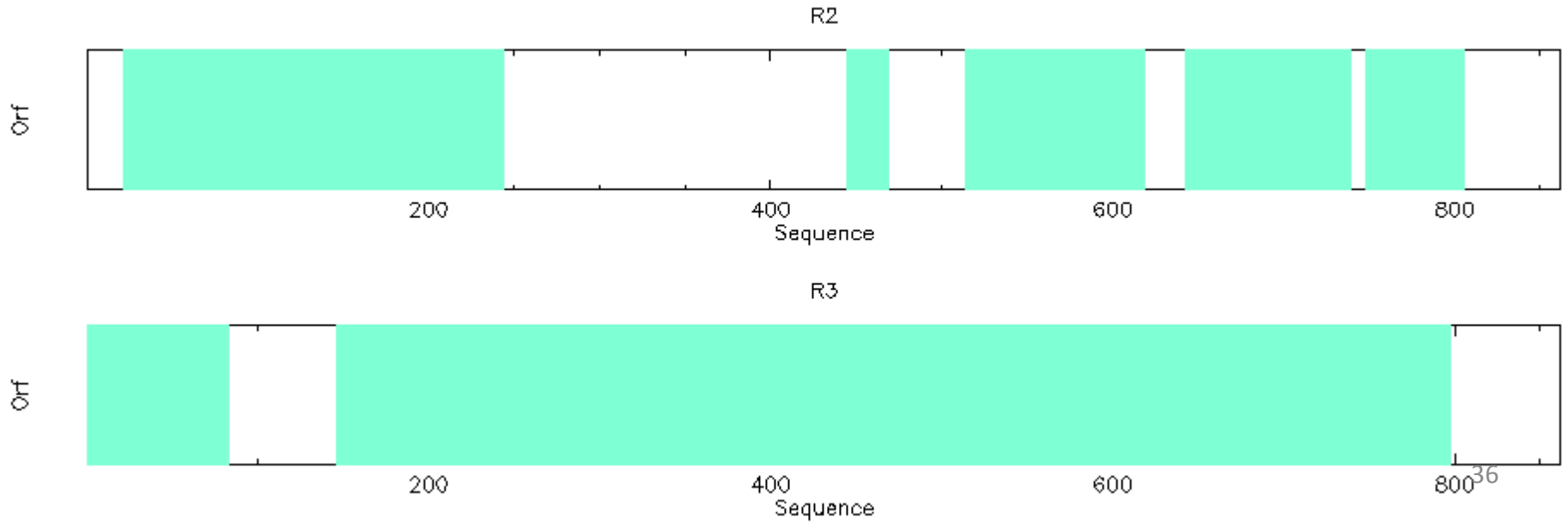
Stop codons:

TAA

Execute

plotorf plots potential open reading frames (ORFs) for an input nucleotide sequence

http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/plotorf.html



prettyseq reads a nucleotide sequence and writes an output file containing in a clean format the sequence with the translation

- prettyseq Output sequence with translated ranges

prettyseq (version 5.0.0)

Sequence:
2: nucleotide seq

Add a ruler:
Yes

Number translations:
Yes

Number DNA sequence:
Yes

Width of screen:
60

Execute

```
PRETTYSEQ of contig00008 from 1 to 862

         ---------|---------|---------|---------|---------|---------|
       1 taagatacctcgaaatattttatttgcattttattttgcattgagaatggtaactacatc 60
       1 *  D  T  S  K  Y  F  I  C  I  L  F  C  I  E  N  G  N  Y  I   19

         ---------|---------|---------|---------|---------|---------|
      61 acgctggcttccactgagcacattctcagggatgaaaaatggaccgaattccagcctctt 120
      20 T  L  A  S  T  E  H  I  L  R  D  E  K  W  T  E  F  Q  P  L   39

         ---------|---------|---------|---------|---------|---------|
     121 ttacaaaatgccagccccaaaatttagcaagaggaccagttgtcaggattcaaggccacc 180
      40 L  Q  N  A  S  P  K  I  *  Q  E  D  Q  L  S  G  F  K  A  T   11

         ---------|---------|---------|---------|---------|---------|
     181 tacacagaagctatgtacttgataaggtccaccacacggttgctgtagccccactcgtta 240
      12 Y  T  E  A  M  Y  L  I  R  S  T  T  R  L  L  *  P  H  S  L   4

         ---------|---------|---------|---------|---------|---------|
     241 tcataccacgcgacaaacttcatgaatttttcgtttaacgcaatgccggcctttgcgtcg 300
       5 S  Y  H  A  T  N  F  M  N  F  S  F  N  A  M  P  A  F  A  S   24

         ---------|---------|---------|---------|---------|---------|
     301 aagatactcgacctgctatcaccaatcaggtctgaagagacgaggtcctcttctgtgtat 360
      25 K  I  L  D  L  L  S  P  I  R  S  E  E  T  R  S  S  S  V  Y   44

         ---------|---------|---------|---------|---------|---------|
     361 ccaagaatacctttcagtgctccctctgattctttctttataacagattttacttcctcg 420
      45 P  R  I  P  F  S  A  P  S  D  S  F  F  I  T  D  F  T  S  S   64

         ---------|---------|---------|---------|---------|---------|
     421 taagaagttgctttctcgagccttacagtcagatcgacaactgatacatcagcagtaggc 480
```

http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/prettyseq.html

garnier is an implementation of the original Garnier Osguthorpe Robson algorithm (GOR I) for predicting protein secondary structure

- garnier Predicts protein secondary structure

garnier (version 5.0.0)

Sequences:

7: AtCesA

In their paper, GOR mention that if yc you are analyzing, you can do better i which provide 'decision constants' (d sheet (extend) terms. So, idc=0 says various combinations of dch,dcs offse

idc 0

Output Report File Format:

TagSeq

Execute

```
      .   10    .   20    .   30    .   40    .   50
     MNTGGRLIAGSHNRNEFVLINADDTARIRSAEELSGQTCKICRDEIELTD
helix                H    HHHHHHHHHHHHH           HHH
sheet E    EEEE       EEEE                    EEEEE
turns   T           T                  TTTTTT         TT
 coil  C CC    CCCCC                                  C
      .   60    .   70    .   80    .   90    .  100
     NGEPFIACNECAFPTCRPCYEYERREGNQACPQCGTRYKRIKGSPRVEGD
helix      HH                                        H
sheet                     EE        E     EEE
turns T  TT  TTTTTTTTT TTT  TTTTTTTT TTTTT   TTT
 coil  CC             C                         CCCCCCC
      .  110    .  120    .  130    .  140    .  150
     EEDDDIDDLEHEFYGMDPEHVTEAALYYMRLNTGRGTDEVSHLYSASPGS
helix HHHHHHHHHHHHH    HHHHHHHHHH
sheet                             EEEE       EEEEE
turns             TT                T           TT   T
 coil               CC               CCCCCC       CCC C
      .  160    .  170    .  180    .  190    .  200
     EVPLLTYCDEDSDMYSDRHALIVPPSTGLGNRVHHVPFTDSFASIHTRPM
helix          H   HHHHH                       HH
sheet EEEEEEE           EEEEEE       EEEEEE
turns        TT TTT           TT           TTT    T   TT
 coil                          CCCCC          C  C CCC
      .  210    .  220    .  230    .  240    .  250
     VPQKDLTVYGYGSVAWKDRMEVWKKQQIEKLQVVKNERVNDGDGDGFIVD
helix               HHHHHHHHHHHHHHHHHHH H
sheet    EEEEEE    E                   E EEE       EEEEE
turns   T      TT                           TTTTT
 coil CC         CC                              CC
      .  260    .  270    .  280    .  290    .  300
     ELDDPGLPMMDEGRQPLSRKLPIRSSRINPYRMLIFCRLAILGLFFHYRI
```

http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/garnier.html

- pepinfo Plots simple amino acid properties in parallel

pepinfo (version 5.0.0)

Sequence:

7: AtCesA

Window size for hydropathy averaging:

9

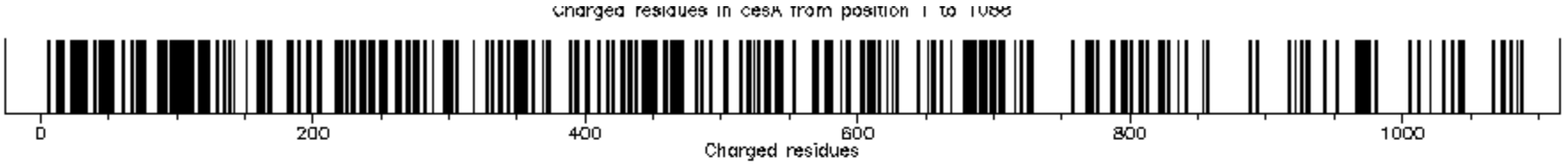
Choose a plot type:

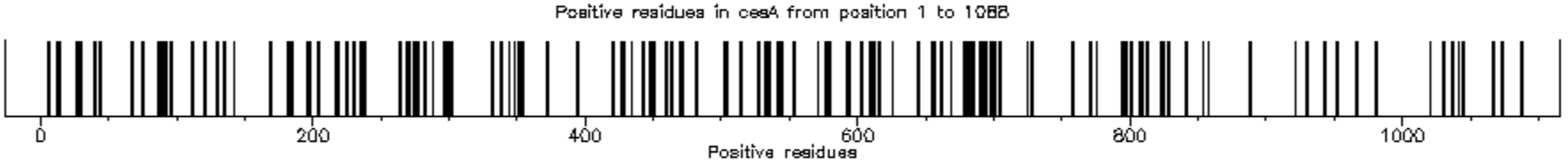
Histogram of general properties

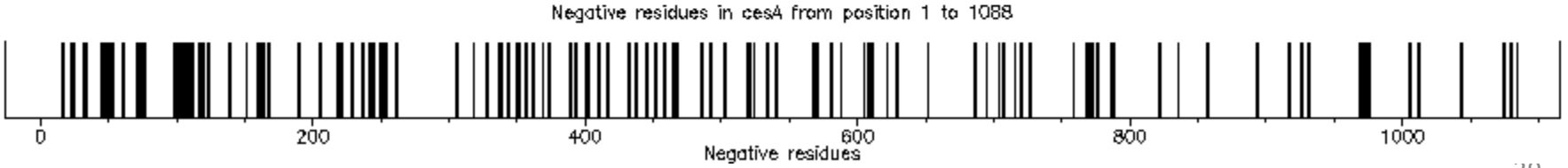
Execute

pepinfo plots various amino acid properties in parallel for an input protein sequence

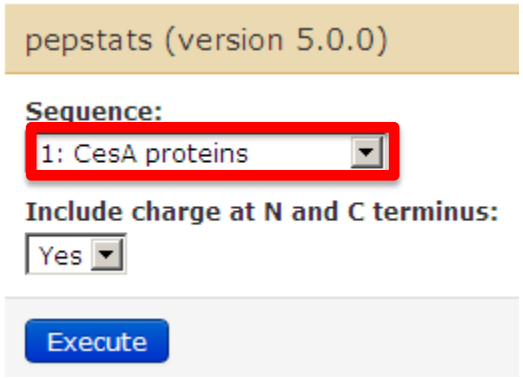
http://emboss.sourceforge.net/apps/release/6.4/emboss/apps/pepinfo.html







- pepstats Protein statistics




```
PEPSTATS of cesA from 1 to 108

Molecular weight = 123446.86             Residues = 1088
Average Residue Weight  = 113.462        Charge   = 5.5
Isoelectric Point = 6.3610
A280 Molar Extinction Coefficient  = 211800
A280 Extinction Coefficient 1mg/ml = 1.72
Improbability of expression in inclusion bodies = 0.695
```

```
PEPSTATS of cslA from 1 to 534

Molecular weight = 61558.14              Residues = 534
Average Residue Weight  = 115.277        Charge   = 20.0
Isoelectric Point = 9.4005
A280 Molar Extinction Coefficient  = 109670
A280 Extinction Coefficient 1mg/ml = 1.78
Improbability of expression in inclusion bodies = 0.790
```

- plotcon Plot quality of conservation of a sequence alignment

plotcon (version 5.0.0)

Sequence:

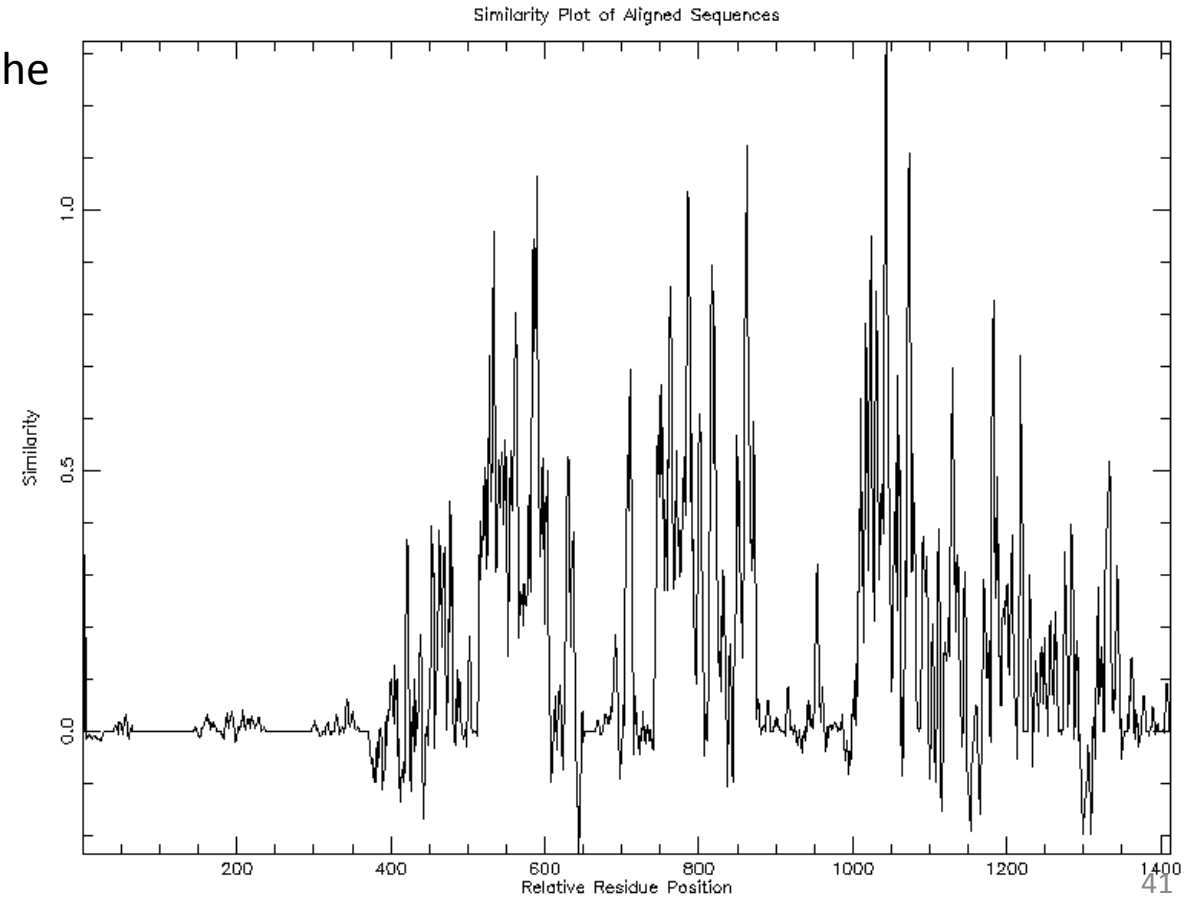
6: CesA alignment

Number of columns to average alignment quality over:

4

Execute

plotcon reads a sequence alignment and draws a plot of the sequence conservation within windows over the alignment



http://emboss.sourceforge.net/apps/release/6.1/emboss/apps/plotcon.html

Next class: Galaxy III - basic NGS analysis