Popular bioinformatics tools in Galaxy: III

Yanbin Yin

Spring 2013

Outline

- Hands on practice
  - EMBOSS
  - NGS data analysis: SRA read assembly

http://131.156.41.196:8080/

Hands on practice: EMBOSS

EMBOSS: European Molecular Biology Open Software Suite

NGS: QC and manipulation

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

NGS: Simulation

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Phenotype Association

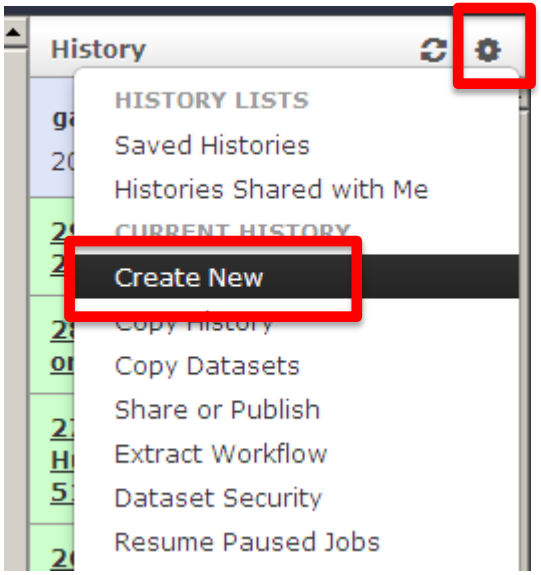
VCF Tools

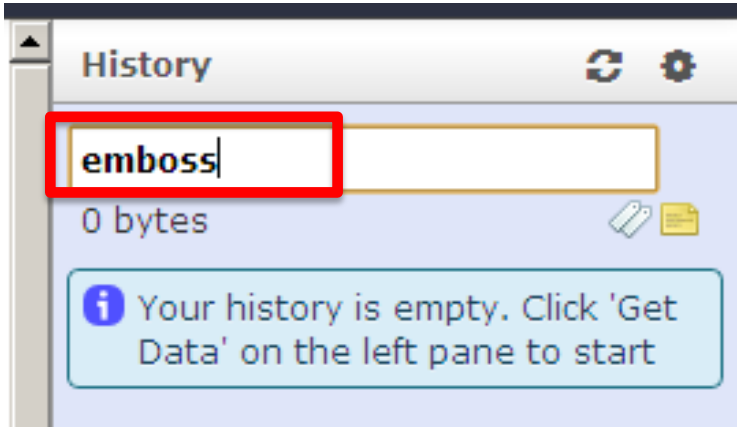
EMBOSS

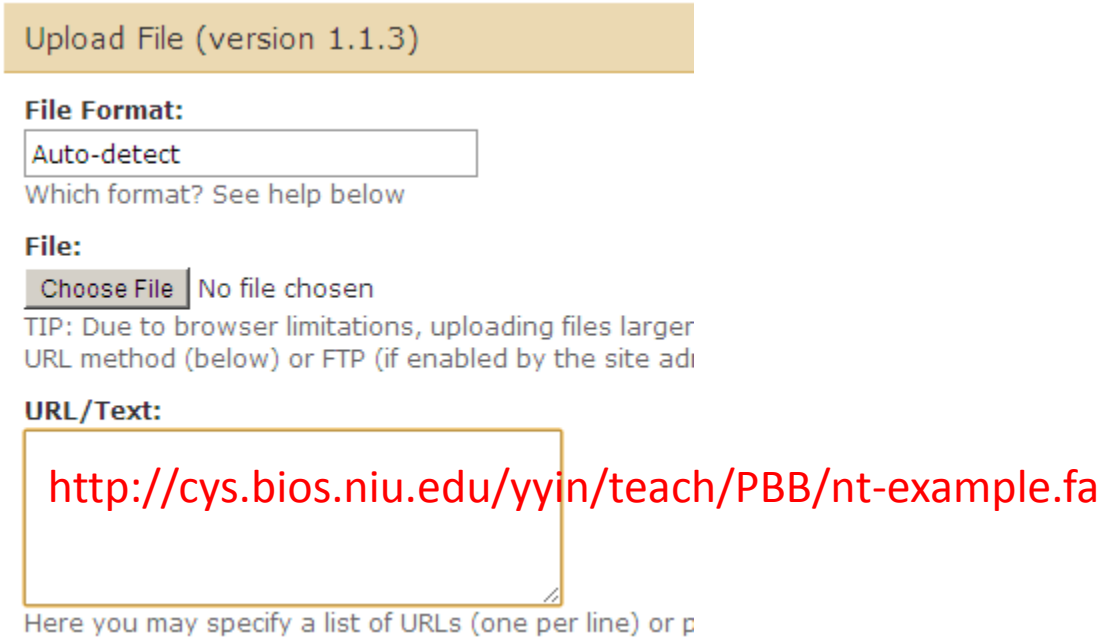
Newbler

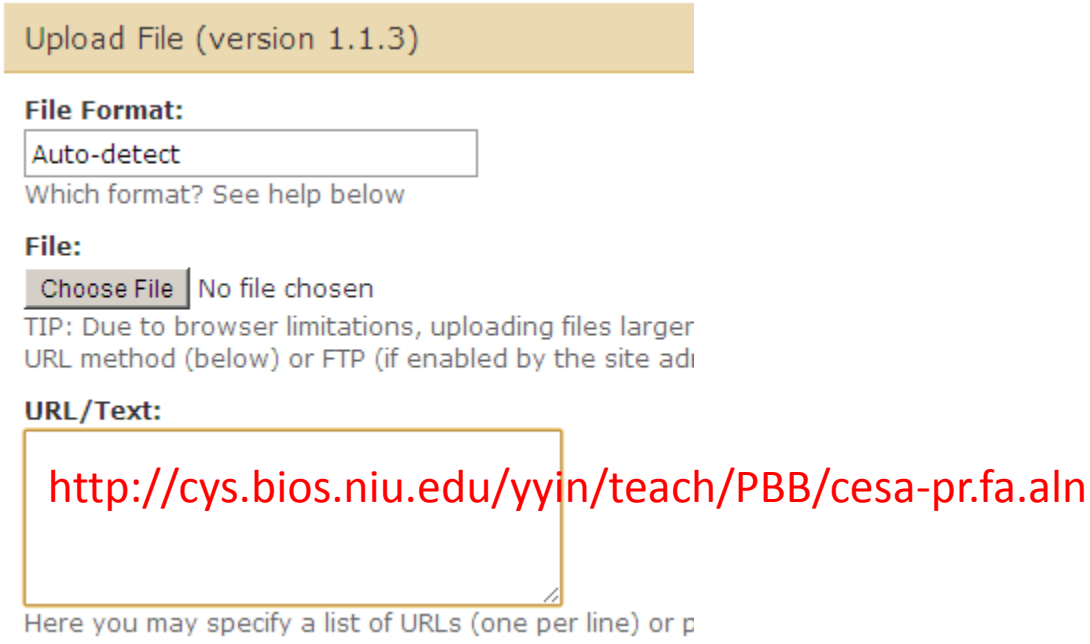
SRA tools

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

Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

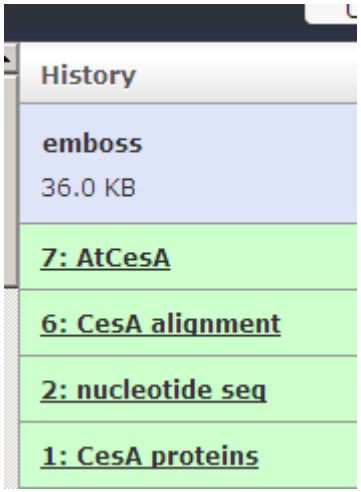
Choose File No file chosen

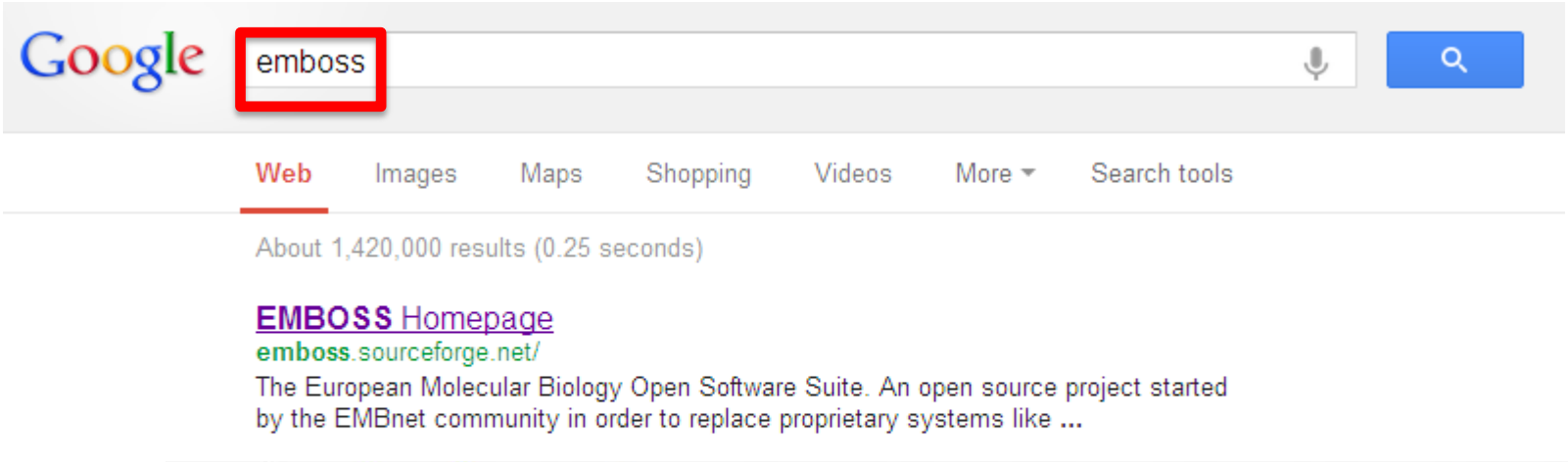
TIP: Due to browser limitations, uploading files larger than 2GI
URL method (below) or FTP (if enabled by the site administrat

URL/Text:

```
>AT2G21770.1|AT2G21770.1|cesA
MNTGGRLIAGSHNRNEFVLINADDTARIRS
AEELSGQTCKICRDEIELTDNGEPFIACNE
CAFPTCRPCYEYERREGNQACPQCGTRYK
RIKGSPRVEGDEEDDDIDDLEHEFYGMDPE
```

Here you may specify a list of URLs (one per line) or paste he




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

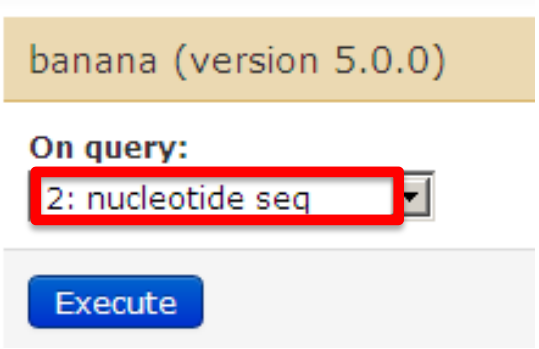
EMBOSS Applications

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

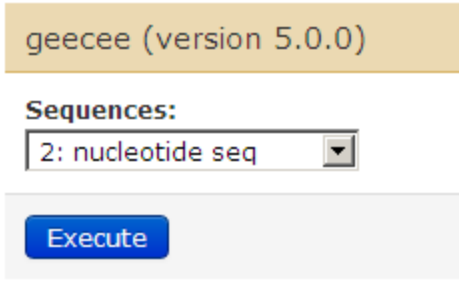


Base	Bend	Curve
t	0.0	0.0
A	11.7	0.0
A	13.9	0.0
G	14.1	0.0
A	13.3	0.0
T	10.6	0.0
A	14.9	0.0
C	17.7	0.0
C	17.7	0.0
T	22.3	0.0
C	26.9	0.0
G	18.5	0.0
A	5.0	0.0
A	1.3	0.0
A	5.9	0.0
T	9.2	0.0
A	5.9	0.0
T	1.3	0.0
T	0.0	0.0
T	3.4	0.0
T	7.8	4.2
A	5.9	5.4
T	1.3	6.8
T	5.7	7.8
T	15.3	8.6
G	19.7	9.4
C	20.7	10.3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

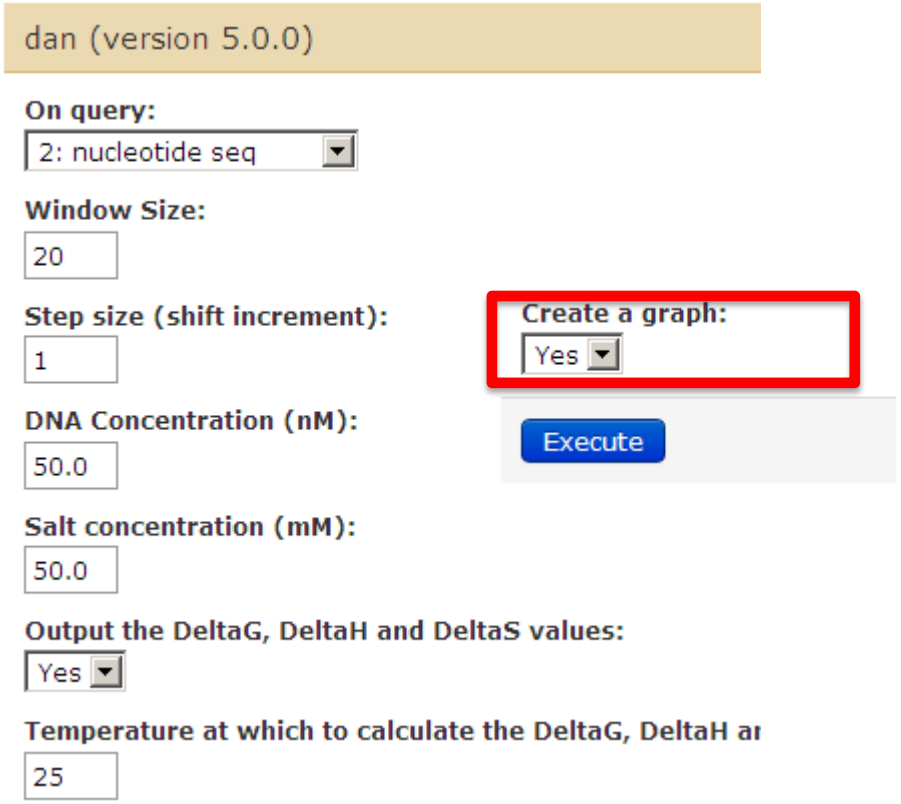
- geecee Calculates fractional GC content of nucleic acid sequences




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

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

- dan Calculates DNA RNA/DNA melting temperature

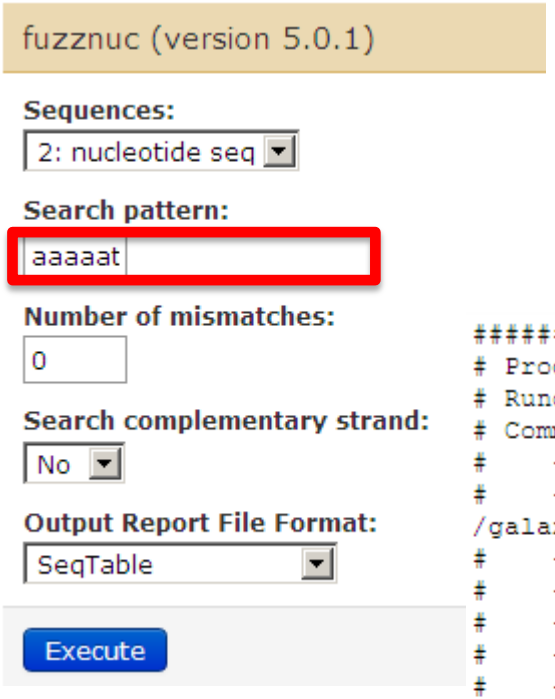


Dan calculates the melting temperature (Tm) and the percentage of G+C nucleotides for windows over a nucleic acid sequence, optionally plotting them.

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

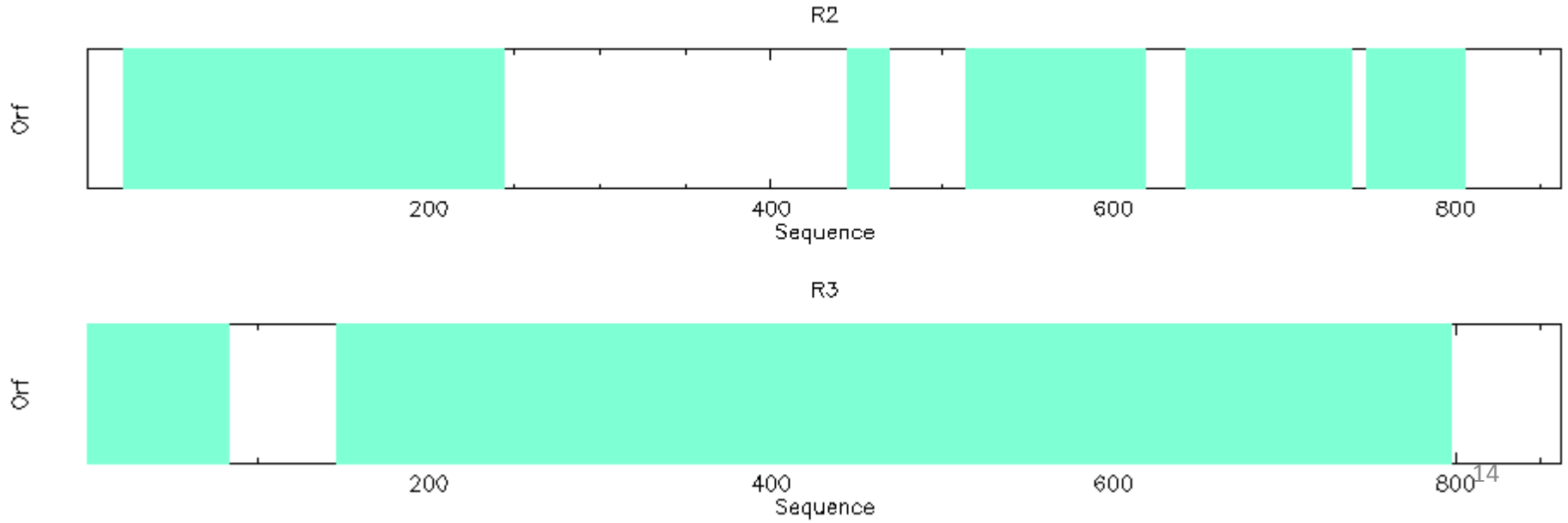
Stop codons:

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

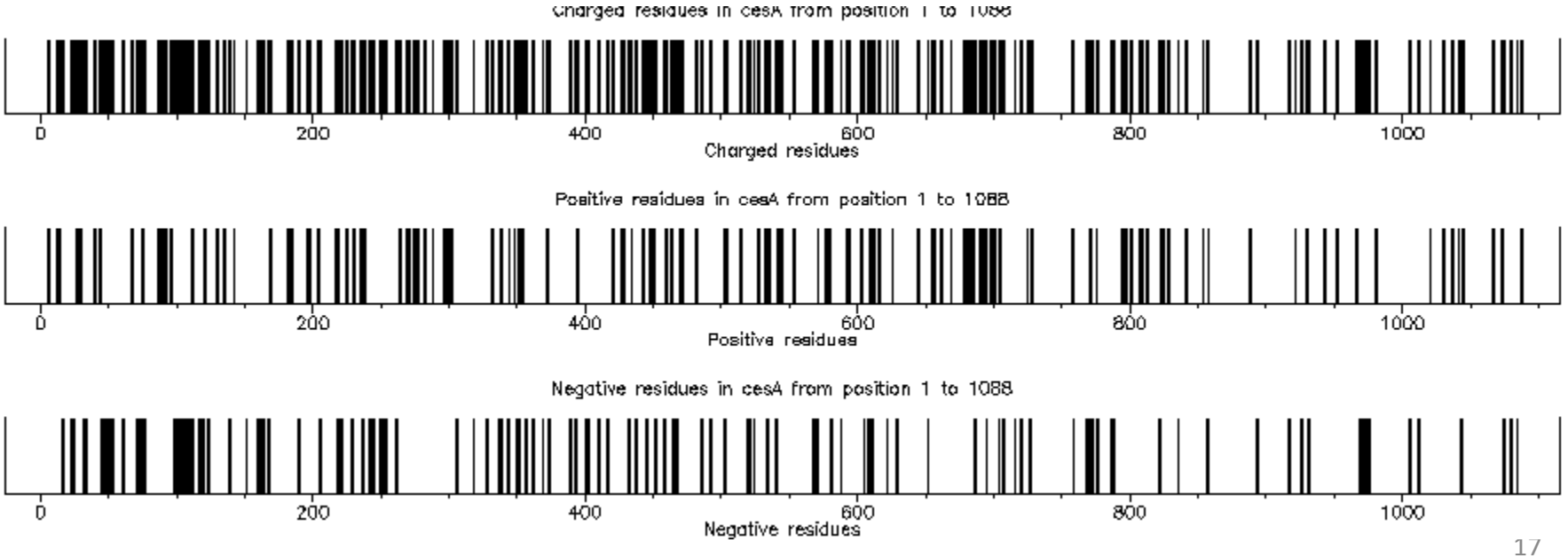
Choose a plot type:

Histogram of general properties

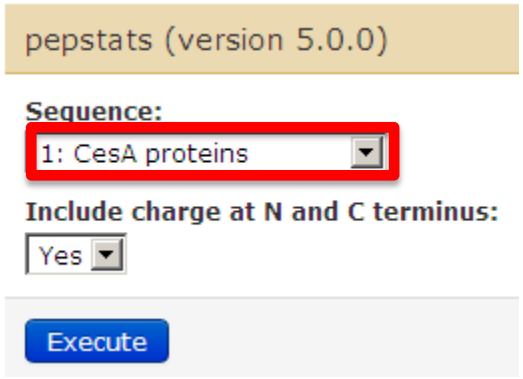
Execute

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:

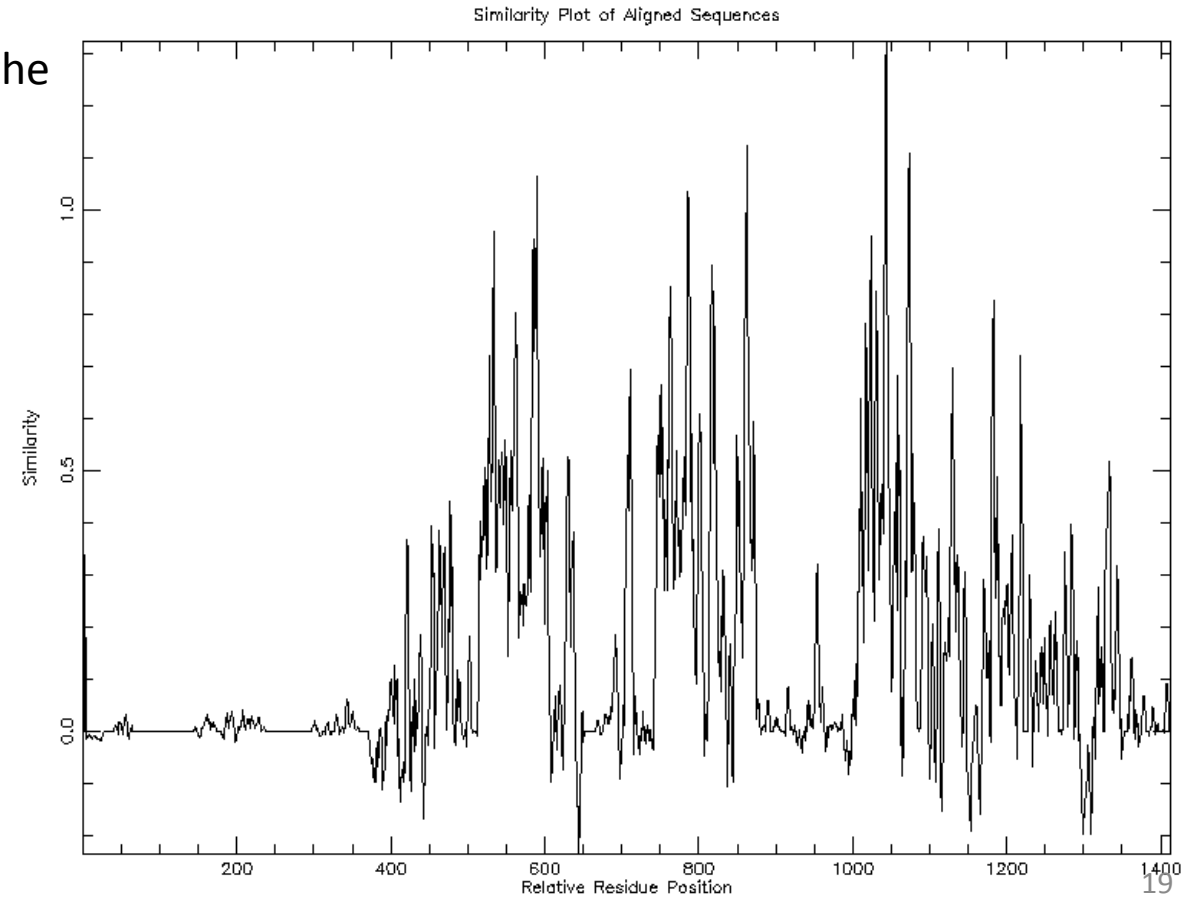
6: CesA alignment

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

Basic NGS analysis on 454 transcriptome reads

NGS: QC and manipulation

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

NGS: Simulation

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Phenotype Association

VCF Tools

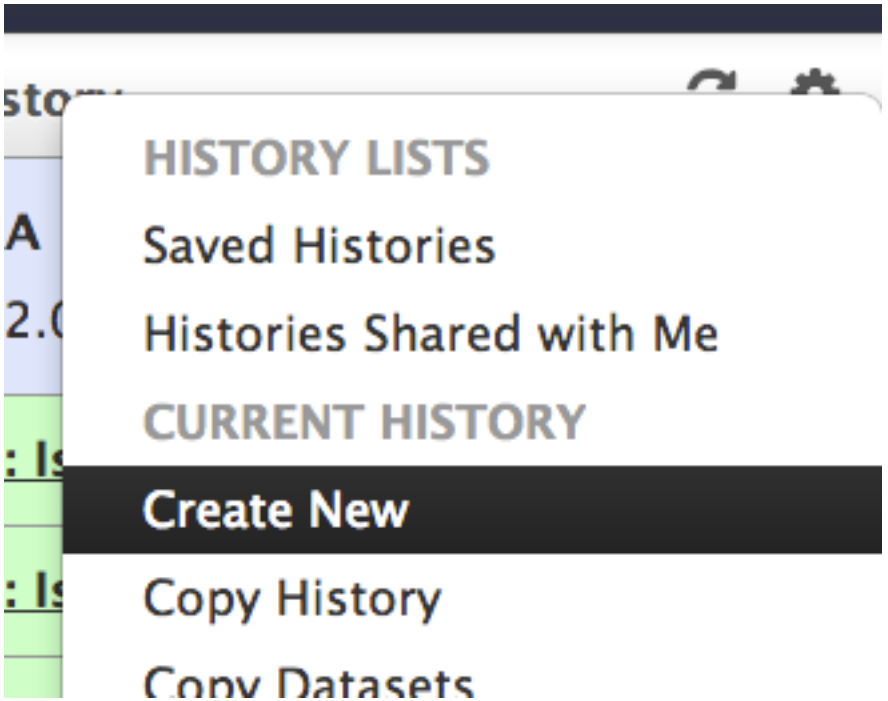
EMBOSS

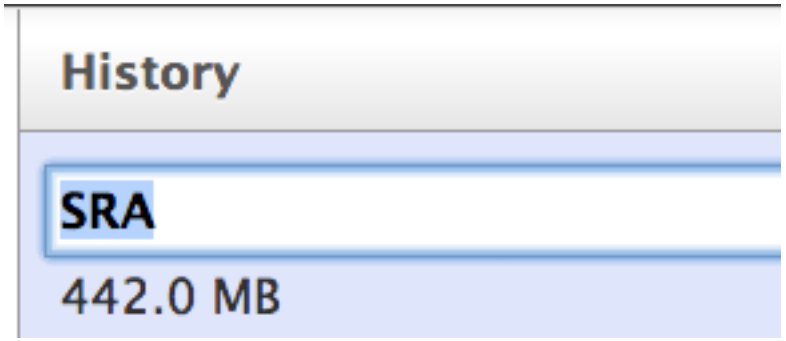
Newbler

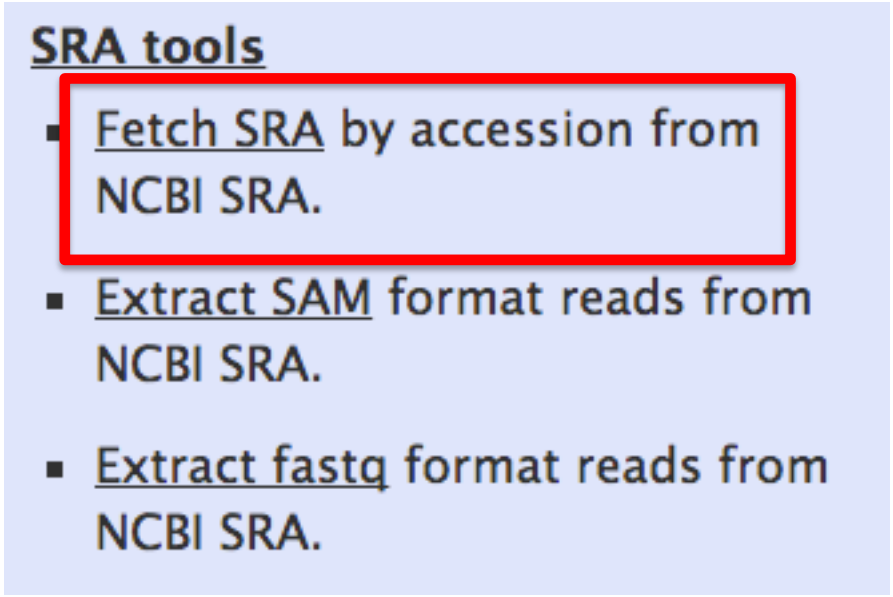
SRA tools

Not available at Galaxy main site!!!

http://131.156.41.196:8080/







Read

The sequences generated by a sequencing machine from a DNA/RNA fragment.

http://www.ncbi.nlm.nih.gov/sra/SRX030762

Binary file (not text file) that are NOT human readable

SRA tools

- Fetch SRA by accession from NCBI SRA.
- Extract SAM format reads from NCBI SRA.
- Extract fastq format reads from NCBI SRA.

sion 1.0.0)

sra archive:

2: Fetch SRR072146

Split read pairs:

No

Specify alignment:

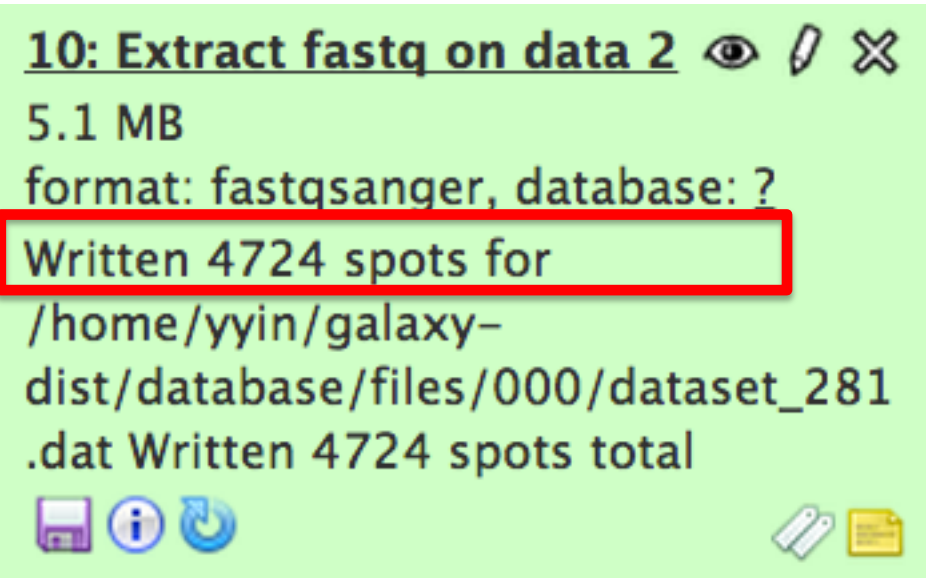
All

Execute

http://www.ncbi.nlm.nih.gov/books/NBK47528/

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

This tool extracts fastqsanger reads from SRA archives using fastq-dump. program is developed at NCBI, and is available at: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.



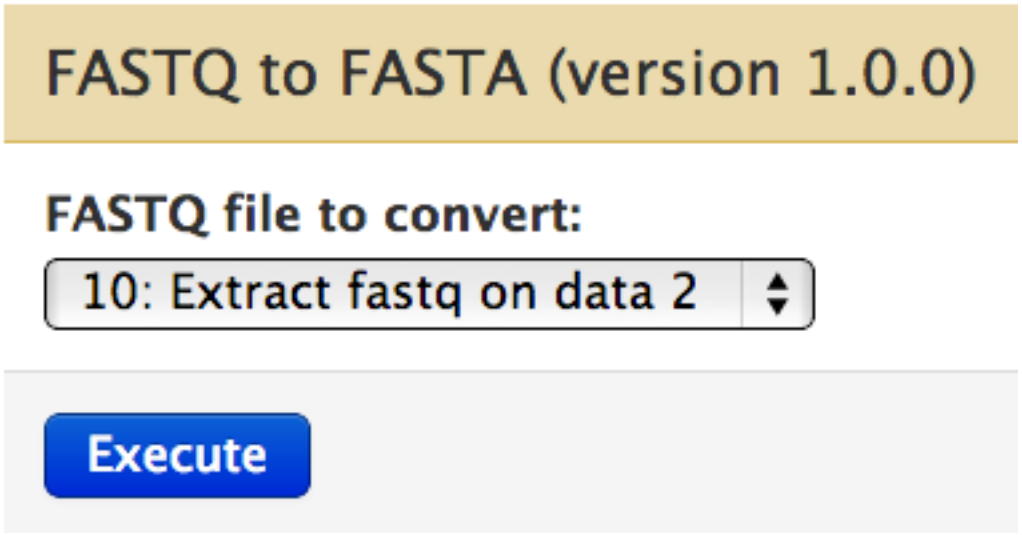
Fastq format (sequence + quality score)

```
@Fetch SRR072146.1 GJ66JU001A6PVK length=494
TCAGTTCGTCGACGCACGTCACGCGCTCNCGTCAAATGACNTCAGCAATCACTGAACNTGNAG
+Fetch SRR072146.1 GJ66JU001A6PVK length=494
447:994/,//////////////////42!///2,,,4///!/4:4/--4////7--/!///!/4
```

NGS: QC and manipulation

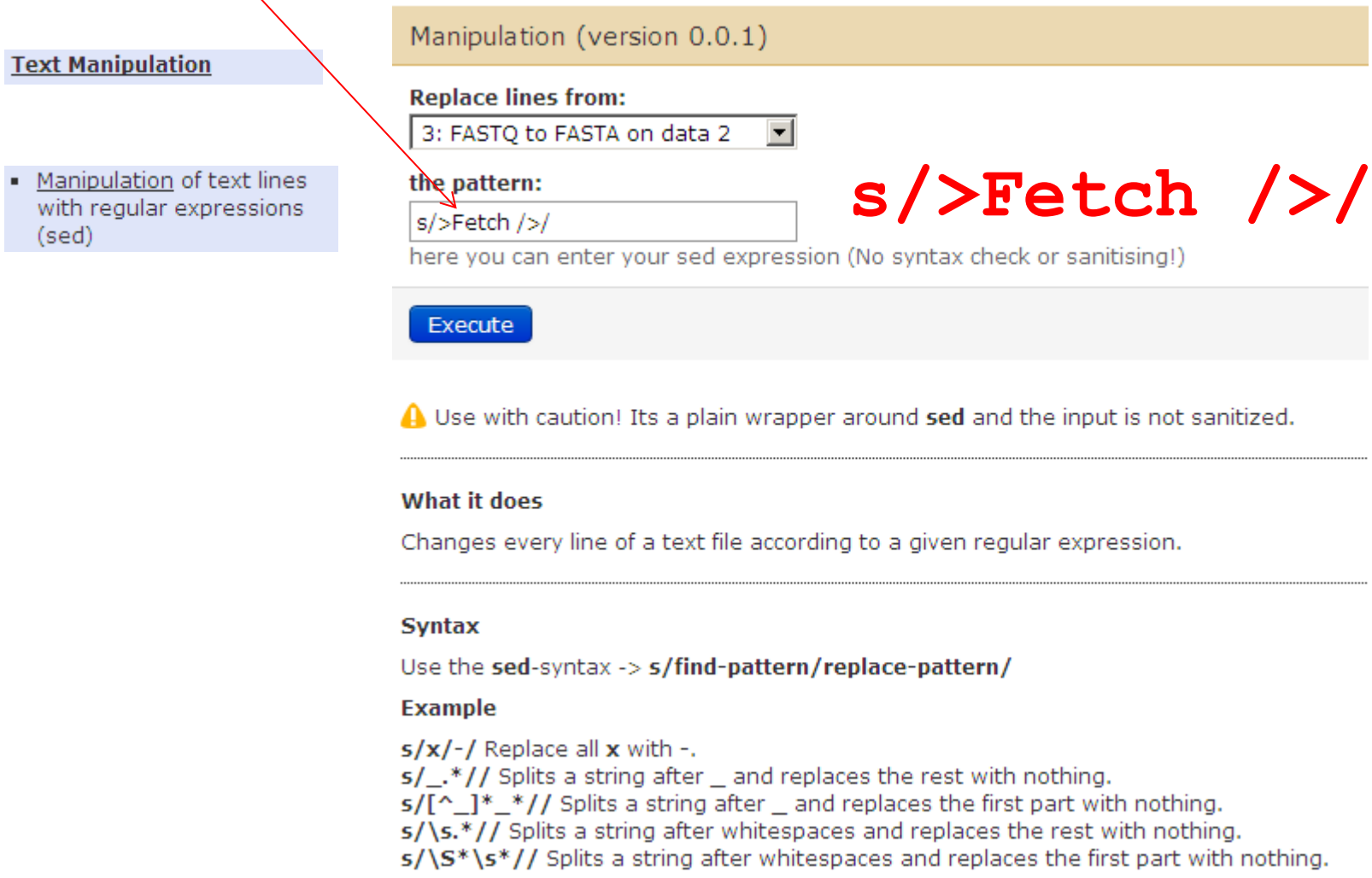
GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- FASTQ interlacer on paired end reads
- FASTQ de-interlacer on paired end reads
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter




```
>Fetch SRR072146.1 GJ66JU001A6PVK length=49
TCAGTTCGTCGACGCACGTCACGCGCTCNCGTCAAATGACNTCAGCAATCACTGAACNTGNAGCTGATTGAAT
>Fetch SRR072146.2 GJ66JU001DA9IB length=440
TCAGTTCAAGATACCTGCTGTGACNGCNTNAGTNTCACNACCGANCAATCAAACGCAAGATTGAAAAGTNTAA
>Fetch SRR072146.3 GJ66JU001DS8LY length=529
TCAGTCCNCCGANCCTGNGNTGCTGNTGGCTTGNGTCGNGCCGCTGCGNCTATCGATGAGNTCCGGNTCGGGA
>Fetch SRR072146.4 GJ66JU001BBLUI length=504
TCAGGTACGTACGATACTGAGTNGTACAGCAAGCGCGCAACATAGTATCTGGGAGTATGGAAGCATGGACATC
>Fetch SRR072146.5 GJ66JU001BDTUJ length=522
TCAGTTTCCGACCATCGAGTGCGCCTTCGGCGATGCGCCGCTAGCGCGGAAGGGAACCTTCATAACGATCAGG
```

>Fetch SRR072146.1 GJ66JU001A6PVK length=494



Newbler

- runMapping Map Roche/454 reads to a reference using Newbler
- Sff File Select reads to include o exclude from one or more input Sff files
- runAssembly De novo assembly of Roche/454 reads using Newbler
- runMapping cDNA Map Roche/454 reads to a reference using Newbler
- runAssembly cDNA De novo assembly of Roche/454 cDNA reads using Newbler
- Sff to Fastq Converter Convert SFF to Fastq

runAssembly cDNA (version 1.0.0)

Newbler version:

(•) default

Unpaired Reads Sff Files

Add new Unpaired Reads Sff Files

Unpaired Reads Fasta Files

Add new Unpaired Reads Fasta Files

Paired Reads Sff Files

Add new Pai

Paired Reads

Add new Pai

[-paired_read

(•) no

( ) [-paired_re

Unpaired Reads Fasta Files

Unpaired Reads Fasta Files 1

SE Fasta file:

4: Manipulation on data 3

Remove Unpaired Reads Fasta Files 1

[-pair] Output pairwise overlaps:

(•) no

( ) [-pair] yes

[-it] Specify the maximum number of isotigs in an isog

100

The following job has been successfully added to the queue:

4: runAssembly cDNA on data 3

5: Read Status

6: Alignment Info

7: All Contigs (Fasta)

8: All Contigs (Qual454)

9: Contig Graph

10: Trim Status

11: Isotigs (Fasta)

12: Isotigs (Qual454)

13: Isotigs (Agp)

14: Isotig Layout

You can check the statu… **History** pane. When th… if completed successfull…

[-ml] Minimum overlap length - The minimu… alignment step. The value can either be a n… length. In the case of a percentage, simply … value. Allowed values: 1 or greater:

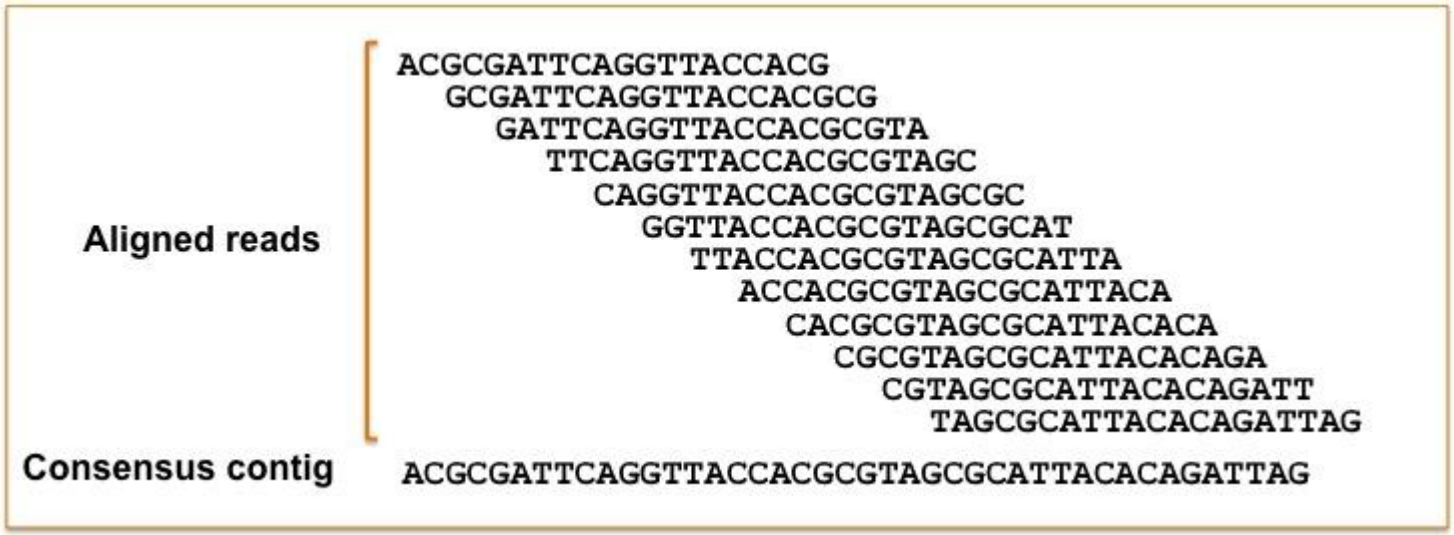
40

[-mi] Minimum overlap identity - The percer… alignment step. Allowed values: 0 or greate…

90

<u>Overlap:</u>

The relationship between two reads, the ends of which have highly similar sequences. The minimum length allowed for the corresponding sequence is an important parameter in assembly.




```
/****************************************************************************
**
**       454 Life Sciences Corporation
**          Newbler Metrics Results
**
**       Date of Assembly: 2013/02/20 23:45:43
**       Project Directory: /home/yyin/galaxy-
dist/database/files/000/dataset_402_files
**       Software Release: 2.7 (20120228_1408)
**
****************************************************************************/

/*
**  Input information.
*/

runData
{
        file
        {
                path = "/home/yyin/galaxy-
dist/database/files/000/dataset_401.dat";

                numberOfReads = 4724, 2567;
                numberOfBases = 2455416, 1350324;
        }

}

/*
**  Operation metrics.
*/

runMetrics
```

Contigs = exons
Isotigs = transcripts
Isogroups = genes

http://contig.wordpress.com/category/newbler-output/
http://contig.wordpress.com/2010/09/21/running-newbler-de-novo-transcriptome-assembly-ii-the-output-files/

These Ns are unknown bases that were not read out by the sequencing machines.

They have the worst quality and we want to get ride of them

emboss NUCLEIC (

Predictions of genes and other genomic features

Program name	Description
checktrans	Reports STOP codons and ORF statistics of a protein
getorf	Finds and extracts open reading frames (ORFs)
marscan	Finds matrix/scaffold recognition (MRS) signatures in DNA sequences
plotorf	Plot potential open reading frames in a nucleotide sequence
showorf	Display a nucleotide sequence and translation in pretty format
sixpack	Display a DNA sequence with 6-frame translation and ORFs
syco	Draw synonymous codon usage statistic plot for a nucleotide sequence
tcode	Identify protein-coding regions using Fickett TESTCODE statistic
wobble	Plot third base position variability in a nucleotide sequence

Technology	Read length (bp)	Error rate	Native paired-end read support	Refs
ABI/Solid	75	Low (~2%)	Yes	93
Illumina/Solexa	100–150	Low (<2%)	Yes	94
IonTorrent	~200	Medium (~4%)*	No	94
Roche/454	400–600	Medium (~4%)*	No	94
Sanger	Up to ~2,000 bp	Low (~2%)	Yes	
Pacific Biosciences	Up to ~15,000‡	High (~18%)	Yes (in strobe read mode)	39

*454 and Ion Torrent technologies are prone to errors in homopolymer regions, which are segments of the genome in which the same nucleotide is repeated multiple times[94]. ‡Pacific Biosciences instruments produce reads with an exponential distribution of read lengths, only a few of which reach the multi-kb range[10,11].

NATURE REVIEWS | **GENETICS**
VOLUME 14 | MARCH 2013 | **157**

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400 ~ 900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

http://www.hindawi.com/journals/bmri/2012/251364/tab1/

Next class: Bioinformatics softwares run on Windows