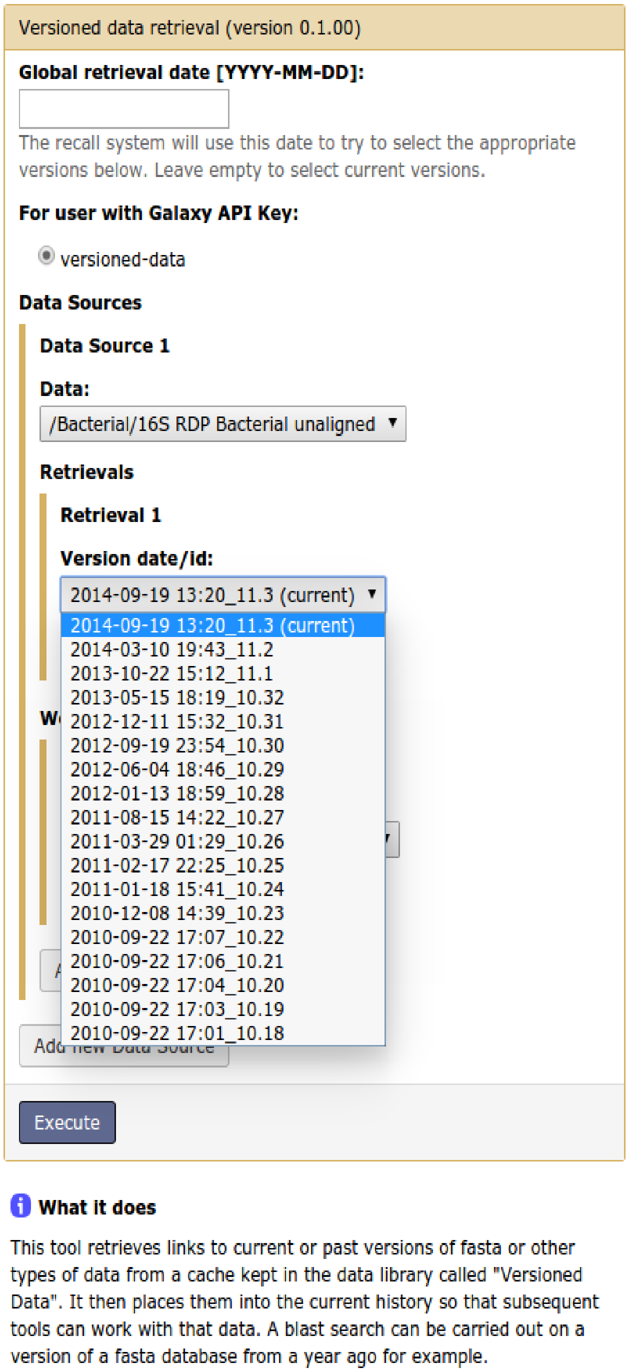
Versioned Data Galaxy and Command-line tool

Presented by https://github.com/Public-Health-Bioinformatics

March 2, 2015

This tool retrieves links to current or past versions of fasta or other types of data from a cache kept in the Galaxy data library called "Versioned Data". It then places them into the current history so that subsequent tools can work with that data. A blast search can be carried out on a version of a fasta database from a year ago for example.



Versioned data retrieval (version 0.1.00)

Global retrieval date [YYYY-MM-DD]:

The recall system will use this date to try to select the appropriate versions below. Leave empty to select current versions.

For user with Galaxy API Key:

● versioned-data

Data Sources

Data Source 1

Data: /Bacterial/16S RDP Bacterial unaligned ▼

Retrievals

Retrieval 1

Version date/id: 2014-09-19 13:20_11.3 (current) ▼

```
2014-09-19 13:20_11.3 (current)
2014-03-10 19:43_11.2
2013-10-22 15:12_11.1
2013-05-15 18:19_10.32
2012-12-11 15:32_10.31
2012-09-19 23:54_10.30
2012-06-04 18:46_10.29
2012-01-13 18:59_10.28
2011-08-15 14:22_10.27
2011-03-29 01:29_10.26
2011-02-17 22:25_10.25
2011-01-18 15:41_10.24
2010-12-08 14:39_10.23
2010-09-22 17:07_10.22
2010-09-22 17:06_10.21
2010-09-22 17:04_10.20
2010-09-22 17:03_10.19
2010-09-22 17:01_10.18
```

What it does

This tool retrieves links to current or past versions of fasta or other types of data from a cache kept in the data library called "Versioned Data". It then places them into the current history so that subsequent tools can work with that data. A blast search can be carried out on a version of a fasta database from a year ago for example.

You can select one or more files by version date or id. (This list is supplied from the Shared Data > Data Libraries > Versioned Data folder that has been set up by a Galaxy administrator).

- In the versioned data tool, user selects a data source, and then selects a version to retrieve (by date or version id).
- If a cached version of that database exists, it is linked into user's history.
- Otherwise a new version of it is created, placed in cache, and linked into history.

The Versioned Data form starts with an optional top-level "Global retrieval date" which is applied to all selected databases. This can be overridden by a retrieval date or version that you supply for a particular database.

Finally, if you just select a data source to retrieve, but no global retrieval date or particular versions, the most recent version of the selected data source will be retrieved.

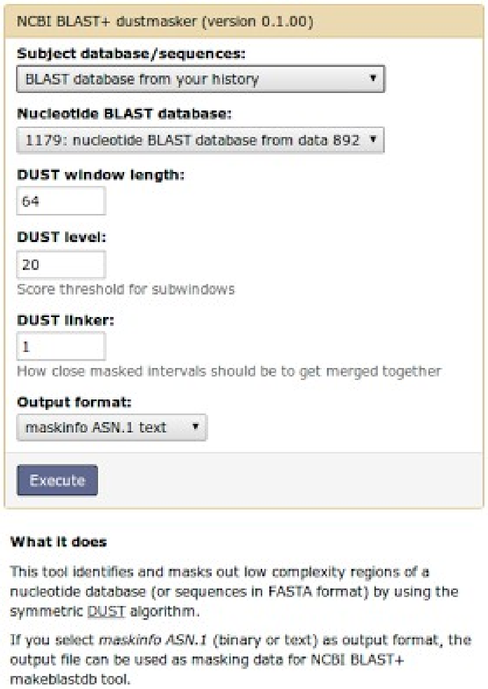
The caching system caches both the versioned data and workflow data that the tool generates. If you request versioned data or derivative data that isn't cached, then(depending on the size of the archive) it may take time to regenerate.

Generation of workflow data

The Workflows section allows you to select one or more pre-defined workflows to execute on the versioned data. Currently this includes any workflow that begins with the phrase "Versioned: ". The results are placed in your history for use by other tools or workflows.

Currently workflow parameters must be entirely specified ("canned"), when the workflow is created/updated, rather than being specified at runtime. This means that a separate workflow with fixed settings must be predefined for each desired retrieval process (e.g.a blastdb with regions of low complexity filtered out, which requires a few steps to execute -dustmasker + makeblastdb etc).

Any user that needs more specific parameters for a reference database creation can just invoke the tools/steps after using the Versioned Data tool to retrieve the raw fasta data. The only drawback in this case is that the derivative data can't be cached - it has to be redone each time the tool is run.



NCBI BLAST+ dustmasker (version 0.1.00)

Subject database/sequences:
BLAST database from your history

Nucleotide BLAST database:
1179: nucleotide BLAST database from data 892

DUST window length:
64

DUST level:
20
Score threshold for subwindows

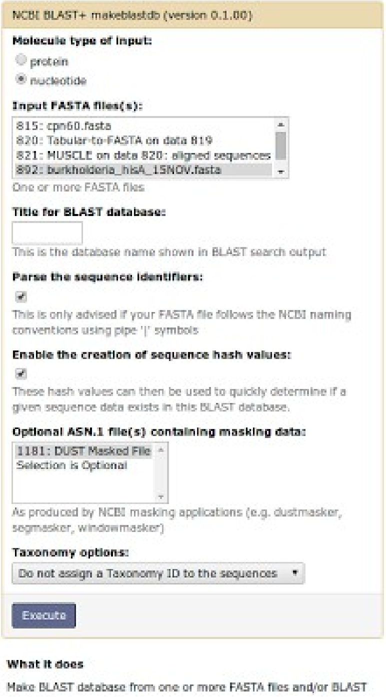
DUST linker:
1
How close masked intervals should be to get merged together

Output format:
maskinfo ASN.1 text

What it does

This tool identifies and masks out low complexity regions of a nucleotide database (or sequences in FASTA format) by using the symmetric DUST algorithm.

If you select *maskinfo ASN.1* (binary or text) as output format, the output file can be used as masking data for NCBI BLAST+ makeblastdb tool.



NCBI BLAST+ makeblastdb (version 0.1.00)

Molecule type of input:
○ protein
● nucleotide

Input FASTA files(s):
815: cpn60.fasta
820: Tabular-to-FASTA on data 819
821: MUSCLE on data 820: aligned sequences
892: burkholderia_hisA_15NOV.fasta
One or more FASTA files

Title for BLAST database:
This is the database name shown in BLAST search output

Parse the sequence identifiers:
☑
This is only advised if your FASTA file follows the NCBI naming conventions using pipe '|' symbols

Enable the creation of sequence hash values:
☑
These hash values can then be used to quickly determine if a given sequence data exists in this BLAST database.

Optional ASN.1 file(s) containing masking data:
1181: DUST Masked File
Selection is Optional
As produced by NCBI masking applications (e.g. dustmasker, segmasker, windowmasker)

Taxonomy options:
Do not assign a Taxonomy ID to the sequences

What it does

Make BLAST database from one or more FASTA files and/or BLAST