C_HW2: Makefile, command line exercises with yeast Version 5

Ken Youens-Clark

Abstract

Yeast is a well-characterized genome due to its small size and historical significance in genetics. The website http://yeastgenome.org/ is a dedicated resource for yeast genomics.

For this exercise, I want you to create a "Makefile" that will execute this entire pipeline if I type "make."

Citation: Ken Youens-Clark C_HW2: Makefile, command line exercises with yeast. **protocols.io** dx.doi.org/10.17504/protocols.io.ftkbnkw
Published: 13 Sep 2016

Protocol

Step 1.

Cerevisiae chromosomes

Create a directory for 'yeast':

```
$ mkdir /work/yeast
$ cd !$
```

Step 2.

Go to http://downloads.yeastgenome.org/sequence/S288C_reference/chromosomes/fasta/ and download the '.fsa' files. You can right-click on the links to copy the link location and then 'wget' the file.

Put all the '*.fsa' files into a 'fasta' directory.

Can you think of a way to easily script this, or will you just click on all 17 chromosomes?

Step 3.

Make a single whole genome file called 'fasta/genome.fa'

Step 4.

Count the chromosomes in the whole genome file. Put the number into a file called "chr-count."

HINT: Each of the original FASTA files contains a single chromosome.

Step 5.

Find size of total genome. Put the answer into a file called 'chr-size.'

HINT: Look up the command 'wc' and find out what it does. The size of the genome can be determined by counting the number of characters in the genome (not on the same line as a fasta header).

Step 6.

Download the list of cerevisiae chromosome features:
http://downloads.yeastgenome.org/curation/chromosomal_feature/SGD_features.tab

Columns:

1. Primary Standfor Gene Database ID (SGDID) (mandatory)
2. Feature type (mandatory)
3. Feature qualifier (optional)
4. Feature name (optional)
5. Standard gene name (optional)
6. Alias (optional, multiples separated by |)
7. Parent feature name (optional)
8. Secondary SGDID (optional, multiples separated by |)
9. Chromosome (optional)1
10. Start_coordinate (optional)1
11. Stop_coordinate (optional)1
12. Strand (optional)1
13. Genetic position (optional)
14. Coordinate version (optional)
15. Sequence version (optional)
16. Description (optional)

Step 7.

Count total genes ('ORF's) into a file called 'gene-count.'

Count only verified genes into a file called 'verified-genes.'

Count only uncharacterized genes into a file called 'uncharacterized-genes.'

Step 8.

Download the file 'ftp://ftp.imicrobe.us/abe487/yeast/palinsreg.txt'

1. These are detected terminator sequences in the *E. coli* genome (using the program GeSTer, if you're curious).
2. The command **grep '/G=[^]*' somefile** will find all lines that match /G=*somegenename*, where somegenename is a sequence of non-blank characters. Read the output of **man grep** and figure out how to -*o*nly print /G=*somegenename*, rather than the whole line.
3. Pipe the results of part (2) through a **cut** command to get only everything after the '='
4. Store the **unique, sorted** results of part (3) into a file named 'terminated_genes'