Decompressing files with the tar command

Always remember to start the Biostar bioinformatics environment whenever you open a terminal window to work on the class.

```
conda activate bioinfo
```

We are going to download some bulk RNA-Seq test data and learn how to decompress it. First we will create a place to store the data.

Go to the directory you've created for working on class materials. If you haven't created a class directory yet, you can try something like this...

```
mkdir biostar_class
```

Now, go to that directory.

```
cd biostar_class
```

Create a directory for the data we are going to download.

```
mkdir RNA_Seq
```

Now, go to the RNA_Seq directory you have created.

```
cd RNA_Seq
```

Now that we're in the correct directory, we will use the "curl" command to download some bulk RNA-Seq test data.

```
curl http://genomedata.org/rnaseq-
tutorial/HBR_UHR_ERCC_ds_5pc.tar --output
HBR_UHR_ERCC_ds_5pc.tar
```

Let's take a look at this Unix command line... We know about the "curl" command. It is used to retrieve data from web sites. A similar command is "wget". Usually the Unix system will have either curl or wget installed, not both. To see which is active on your system, just type the command at the command line like this...

```
wget
```

You may see an error like this if wget is not installed.

```
-bash: wget: command not found
```

Next, try the curl command.

```
curl
```

If curl is active on the system, you may see something like this...

```
curl: try 'curl --help' or 'curl --manual' for more
information
```

We can do as the instructions say...

```
curl --help
```

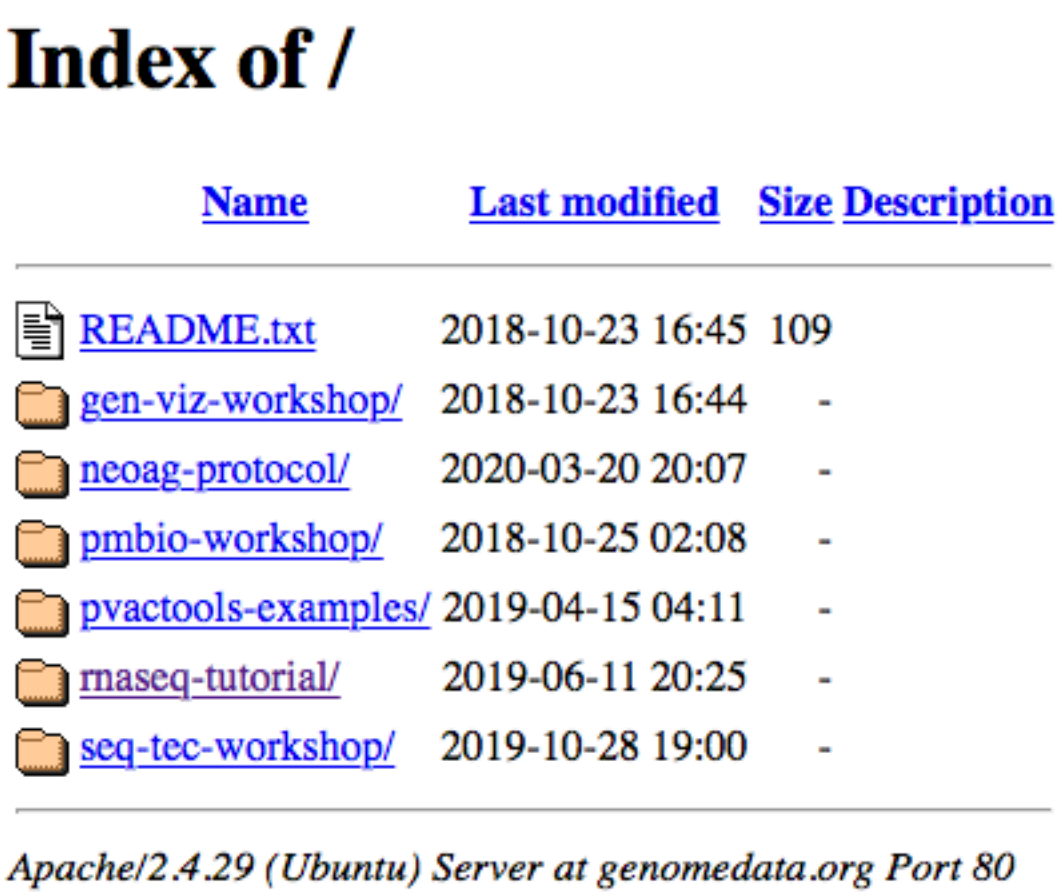
and see information on the usage of the curl command. So it looks like curl is installed on this system.

Okay, moving on. Let's take a look at this command line. We know what curl means, how about the rest of it. The URL "http://genomedata.org/rnaseq-tutorial/HBR_UHR_ERCC_ds_5pc.tar" represents the "path" to this data. Paths are a very important concept in Unix. An incorrect path can result in frustrating "file not found" errors.

```
curl http://genomedata.org/rnaseq-
tutorial/HBR_UHR_ERCC_ds_5pc.tar --output
HBR_UHR_ERCC_ds_5pc.tar
```

The path to the file "HBR_UHR_ERCC_ds_5pc.tar" is "genomedata.org/rnaseq-tutorial" which can be translated as "on the genomedata.org server, there is a directory (folder) named "rnaseq-tutorial" that contains the file HBR_UHR_ERCC_ds_5pc.tar". Notice how there are no blank spaces in the path name - Unix can not deal with spaces in file names, directories or paths.

Another way to get to this data file is via the WWW. Open a browser window and enter "http://genomedata.org". You will see an index page listing all the directories on this server. It should look something like this.



Index of /

Name	Last modified	Size	Description
README.txt	2018-10-23 16:45	109	
gen-viz-workshop/	2018-10-23 16:44	-	
neoag-protocol/	2020-03-20 20:07	-	
pmbio-workshop/	2018-10-25 02:08	-	
pvactools-examples/	2019-04-15 04:11	-	
rnaseq-tutorial/	2019-06-11 20:25	-	
seq-tec-workshop/	2019-10-28 19:00	-	

Apache/2.4.29 (Ubuntu) Server at genomedata.org Port 80

Find the "rnaseq-tutorial" folder and click on it. Now you will see something like this.

Index of /rnaseq-tutorial

Name	Last modified	Size	Description
Parent Directory		-	
ERCC_Controls_Analysis.txt	2018-10-23 16:49	4.0K	
GRC-human-build38_human_90_38-rna_seq_annotation.tgz	2018-10-23 16:48	5.8G	
GRC-human-build38_human_90_38-rna_seq_annotation.tgz.md5	2018-10-23 16:49	87	
HBR_UHR_ERCC_ds_5pc.tar	2018-10-23 16:49	111M	
Integrative_Assignment_RNA.tar.gz	2019-02-27 20:23	12G	
RSeQC.zip	2018-10-23 16:49	422M	
annotations/	2018-10-23 16:49	-	
data.tar.gz	2018-10-23 16:49	58M	
fasta/	2018-10-23 16:48	-	
illumina_multiplex.fa	2018-10-23 16:49	161	
multiqc_report.html	2019-06-11 20:25	1.1M	
practical.tar	2018-10-23 16:48	347M	
testdata/	2019-03-15 18:44	-	
trinity_trinotate_tutorial.Toronto2017.tar.gz	2018-10-23 16:49	2.5G	

Apache/2.4.29 (Ubuntu) Server at genomedata.org Port 80

If you look closely, you will find a file named "HBR_UHR_ERCC_ds_5pc.tar". What happens if you click on this link? Does it download? Can you open a tar file in the Mac environment? How about on PC? How would you do it?

Okay, let's take a look at the file name "HBR_UHR_ERCC_ds_5pc.tar". What does the ".tar" extension mean? tar refers to "tape archive", and is the most commonly used Unix method to compress files. How do we deal with tar files? On a Unix system, we can decompress .tar files using the tar command, like this.

```
tar xvf filename.tar
```

So for our file, the command would be

```
tar xvf HBR_UHR_ERCC_ds_5pc.tar
```

What does "xvf" mean? If we check the "man" page for tar, we could find out...

```
man tar
```

"x" means - extract to disk from the tar (tape archive),

"v" means - produce verbose output. When using this flag tar will list each file name as it is read from the tar (tape archive).

"f" (file) means read the tar (tape archive) from or to the specified file.

You will sometimes see the command used this way with a "-" in front of the flags.

```
tar -xvf filename.tar
```

OR

```
tar xvf filename.tar
```

Both of these notations produce the same results.

There are also lots more flags that can be used - see the man page.

What happens when you run the tar command?

```
tar xvf HBR_UHR_ERCC_ds_5pc.tar
```

You should see each of the files listed as the tar is decompressed. There should be a total of 12 files in this tar. Note that each file now has the extension .fastq.gz. What does this tell you about the files? They are fastq formatted files, and they are "zipped", which is another form of compression. Instead of "unzipping" all these files with the "gunzip" command, we can peek inside them with the "zcat" command. On Mac systems, you may need to use "gzcat" instead.

```
gzcat UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-
chr22.read1.fastq.gz | head -n 8
```

In this case, we are "piping" - with the pipe symbol "|", the results of the "zcat" command into the "head" command and selecting to see the top 8 lines of the file (-n 8).

The results should show the top 8 lines of the ".fastq.gz" file, which consists of two fastq files (remember each fastq file has 4 lines). Something like this...

```
@HWI-ST718_146963544:8:1212:5958:93757/2
TTATGGGATTCGATCAACAGAGAGTAACAGAGTATTATTATGTTATTTTATTCTGTGTGTATTT

+
CCCFFFFFHHHHHJJJJJJJJJJJHIJJJJHICFGIIJJJIIIIJJJJJJHHJJJIJIJIJJJJ

@HWI-ST718_146963544:7:2308:7250:88065/2
CTAGCATTCACATGCATGTTGCTACAGTACAATTGATTCATTAATTAACTTTAGCCAATTACTT

+
```

```
@@@FDFFFHGHHGIFIIJGGGIJJGHJHGIJJGIEIJJIIIJEHGIGIJ>FHIJIGHIJIJJJJ
```

(don't worry if your data is not exactly the same as the example)

Keep in mind, there are several Unix commands that can be used to look at the contents of files, each has it's own flags/options and is used slightly differently. For example:

```
less
more
cat
head
tail
```

To see how each of them works, you can look at the man pages.

```
man less
man more
man cat
man head
man tail
```

Now, if you DID want to do the final decompression on these files, you would use the "gunzip" command.

```
gunzip UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-
chr22.read1.fastq.gz
```

Keep in mind however, that many of the downstream data analysis steps can be done on ".gz" compressed files, so no need to remove this final compression as it will just take up lots of space in your directories.

See here for more information on the bulk RNA-Seq test data set.

+ Add a custom footer

+ Add a custom sidebar

Clone this wiki locally

https://github.com/AmyStonelake/BTEP.wiki.git