Maloof’s Rule #4 of Bioinformatics
plus
Common Problems and Solutions

Maloof Rule # 4

- Do NOT trust the computer
- Be suspicious of results that look too good (or too strange)
- Corollary: the computer is stupid

Example: Be Skeptical

- *Exercise 11b:* Use a for loop that builds on the command from 11a to count the number of unique hits in each file.

Algorithm	Time	Number of Unique Hits
megablast	0m16.054s	500
dc-megablast	0m28.169s	500
blastn	0m20.736s	500

Exercise 8: Consider the host species listed for the hits. Remembering that our samples came from a human patient, which hosts are most evolutionarily distant? Could the viruses generating these hits still have come from the patient sample

- Many people answered this with *Rhinolophus sinicus*
- What about *Streptococcus pneumoniae* and *Staphylococcus aureus ?*

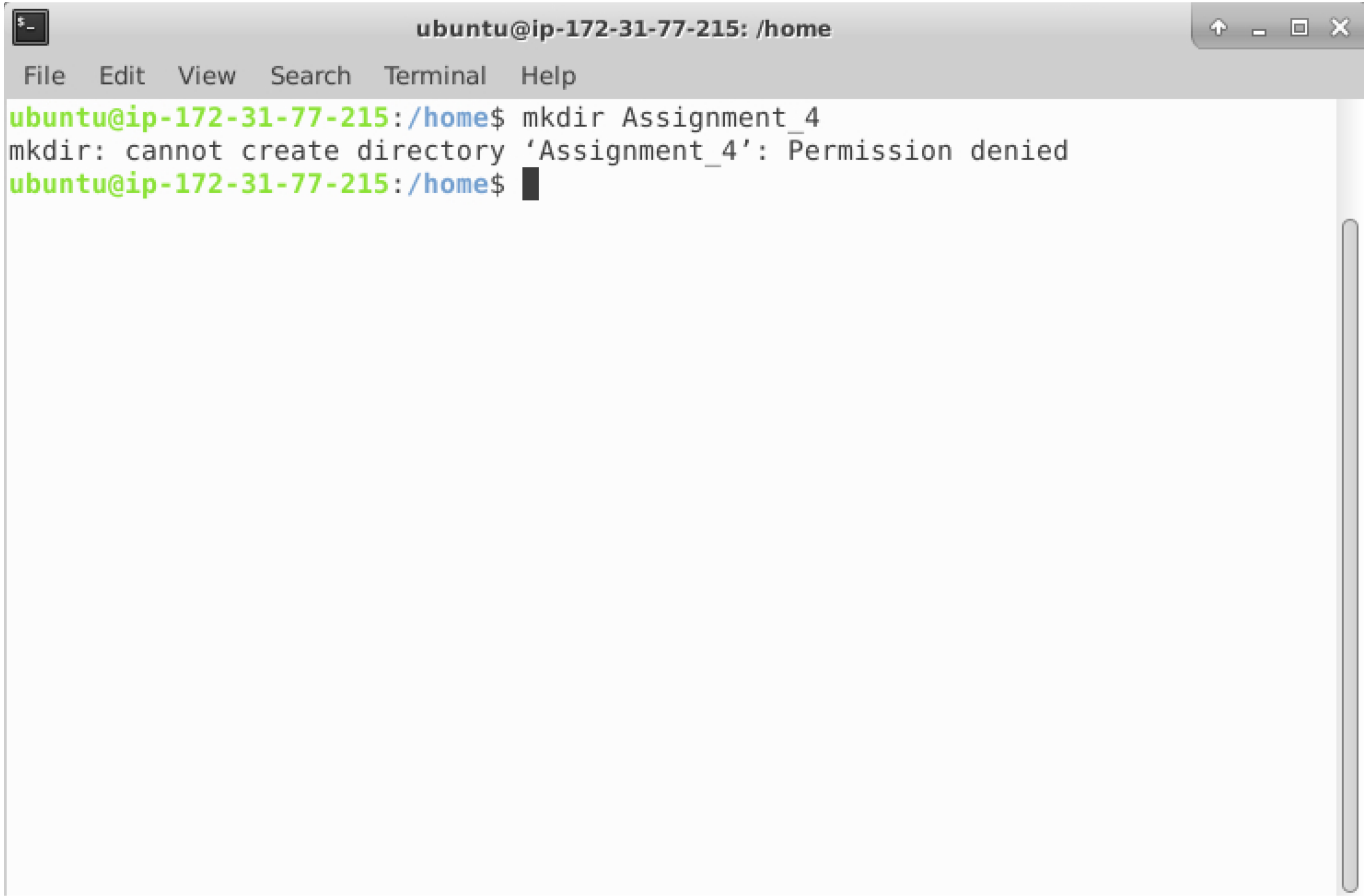
How can I be sure my html is good on Github?

- We are now requiring that you knit your .Rmd to .html
- If you click on the .html file on github you get a mess
- How to confirm your file is okay?
- On github, click on the file, and then click raw.
  - Then:
    - Copy the URL from the browser bar and paste it into https://htmlpreview.github.io/

    OR

    - Save the raw file to your computer (give it a .html extension) and open the downloaded file in your browser.

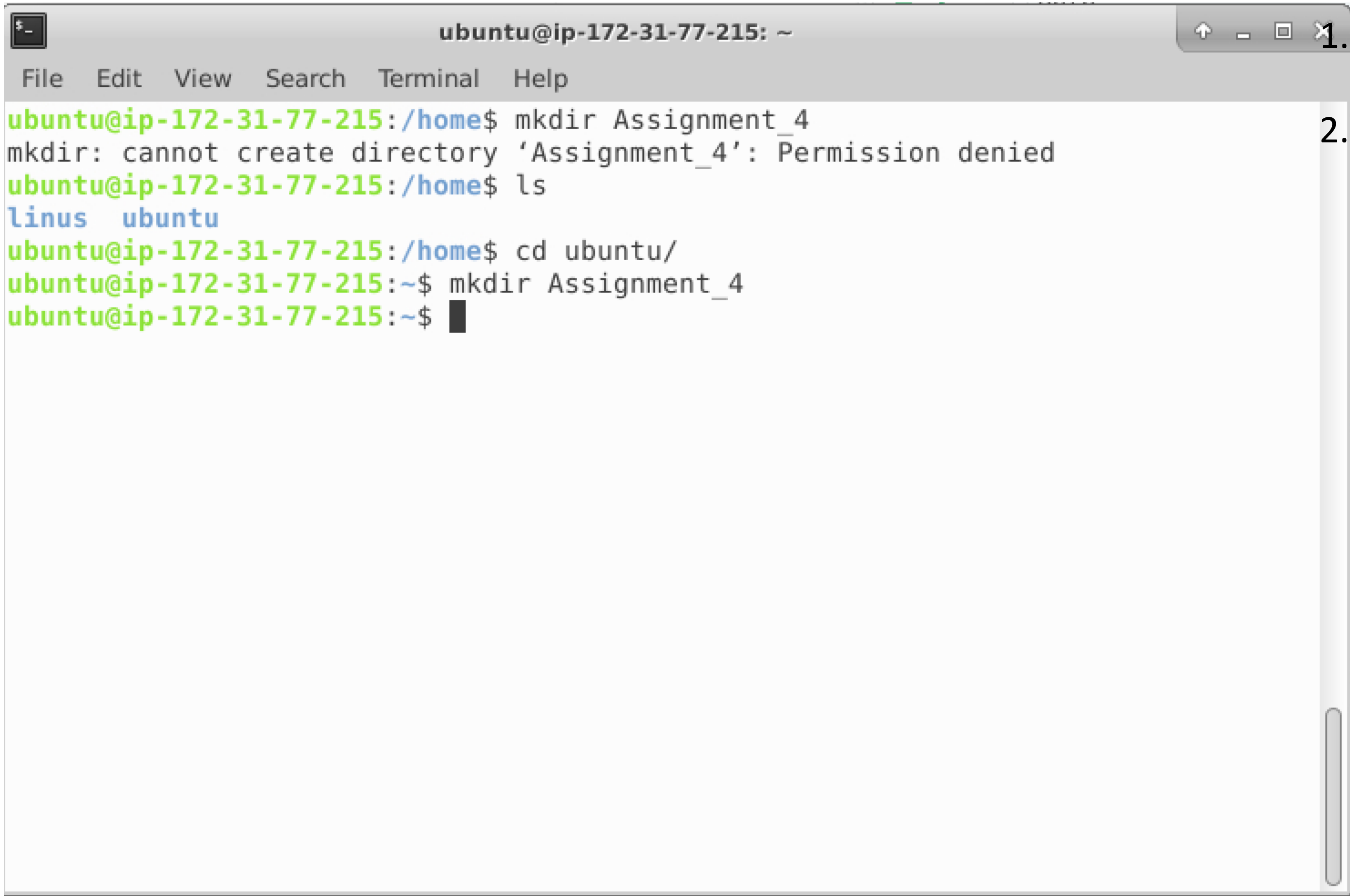
Problem



ubuntu@ip-172-31-77-215: /home
File Edit View Search Terminal Help

```
ubuntu@ip-172-31-77-215:/home$ mkdir Assignment_4
mkdir: cannot create directory ‘Assignment_4’: Permission denied
ubuntu@ip-172-31-77-215:/home$ 
```

Solution

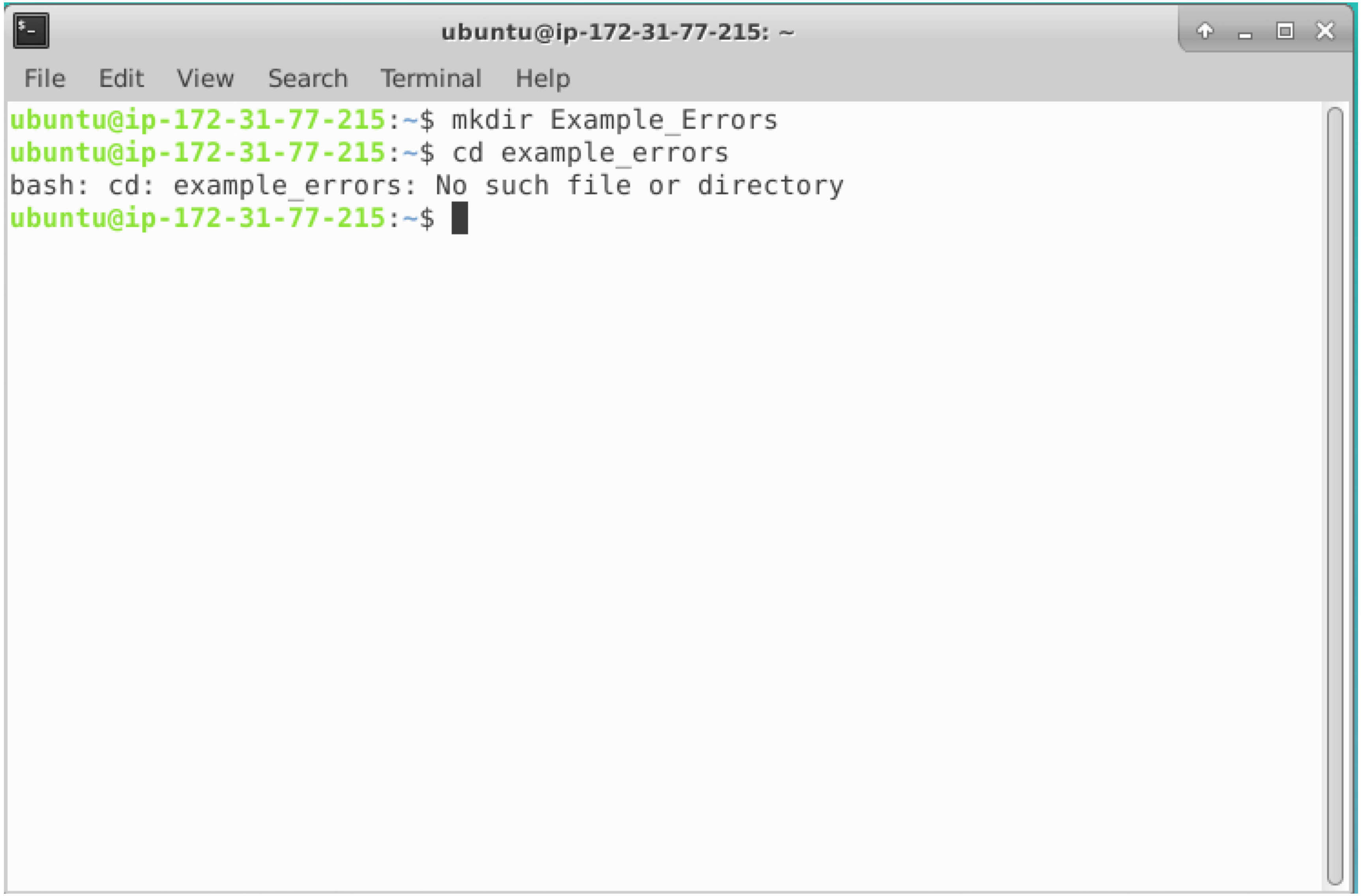


ubuntu@ip-172-31-77-215: ~
File Edit View Search Terminal Help

```
ubuntu@ip-172-31-77-215:/home$ mkdir Assignment_4
mkdir: cannot create directory ‘Assignment_4’: Permission denied
ubuntu@ip-172-31-77-215:/home$ ls
linus  ubuntu
ubuntu@ip-172-31-77-215:/home$ cd ubuntu/
ubuntu@ip-172-31-77-215:~$ mkdir Assignment_4
ubuntu@ip-172-31-77-215:~$ 
```

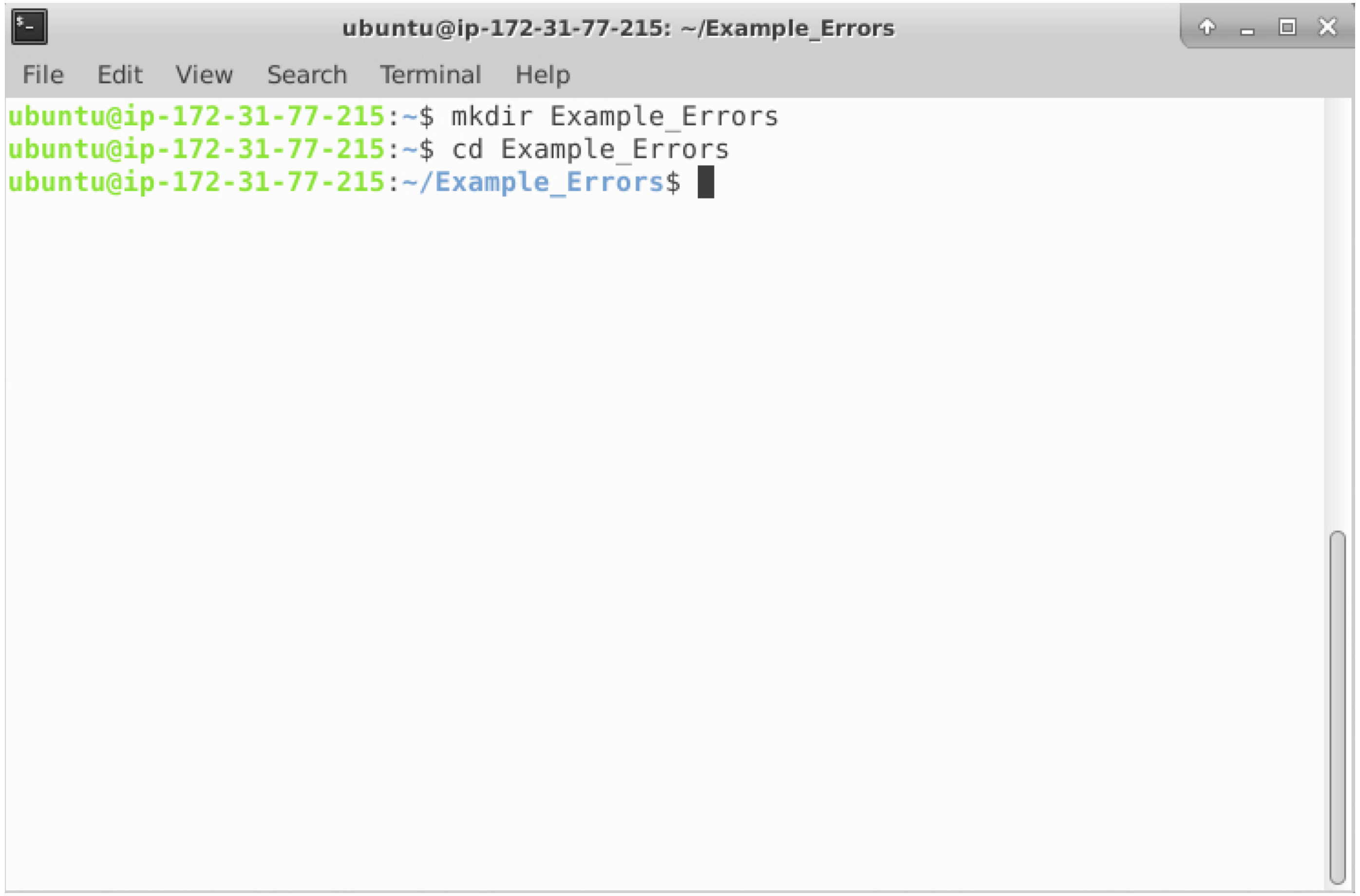
1. “/home” is not your home directory
2. “/home” is the home to everyone's home directory

Problem



```
ubuntu@ip-172-31-77-215:~$ mkdir Example_Errors
ubuntu@ip-172-31-77-215:~$ cd example_errors
bash: cd: example_errors: No such file or directory
ubuntu@ip-172-31-77-215:~$ 
```

Solution

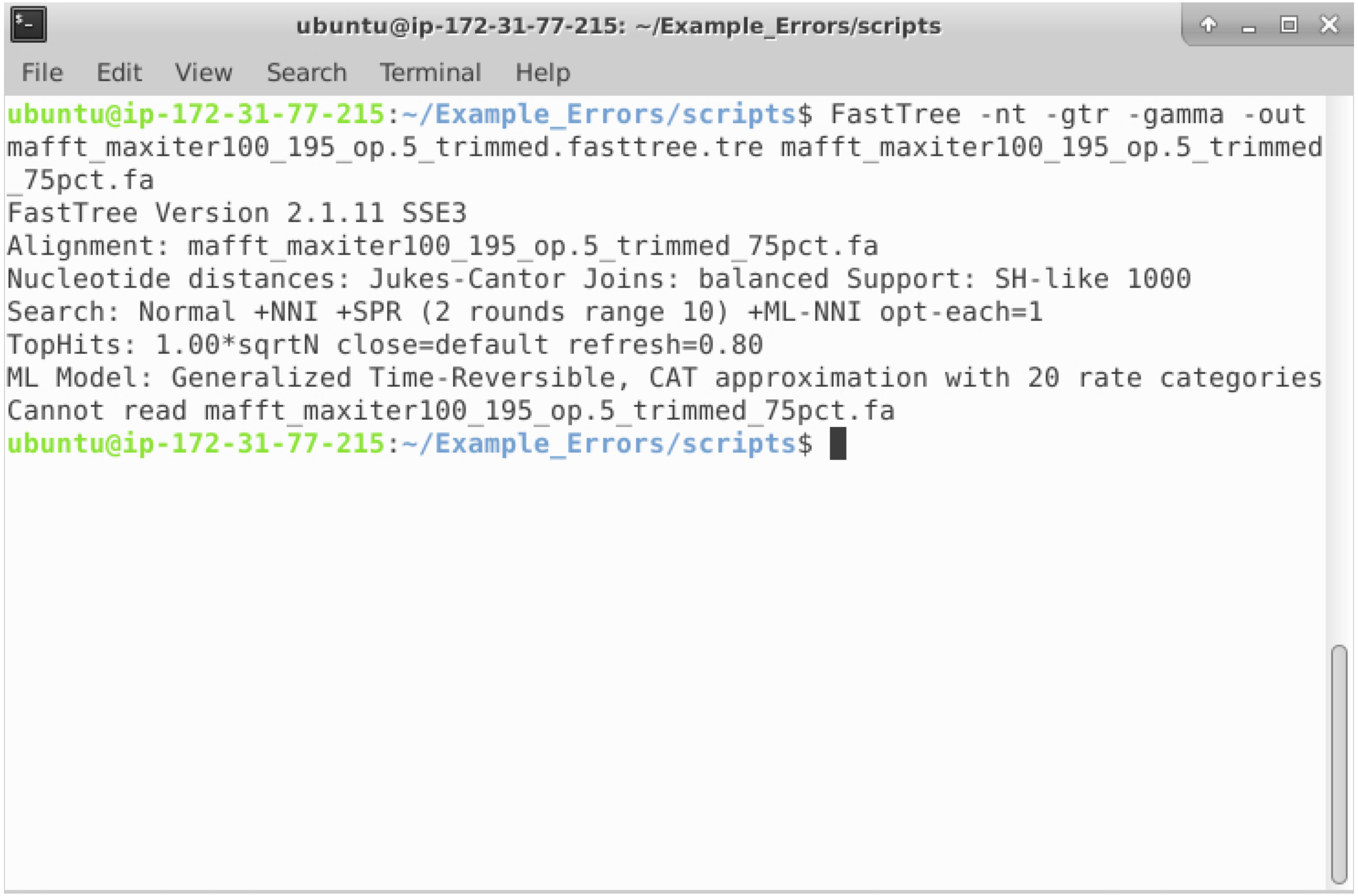


ubuntu@ip-172-31-77-215: ~/Example_Errors
File Edit View Search Terminal Help

```
ubuntu@ip-172-31-77-215:~$ mkdir Example_Errors
ubuntu@ip-172-31-77-215:~$ cd Example_Errors
ubuntu@ip-172-31-77-215:~/Example_Errors$ 
```

1. Capitalization matters when entering file names
2. When in doubt use TAB-completion, if TAB can’t find it check your path

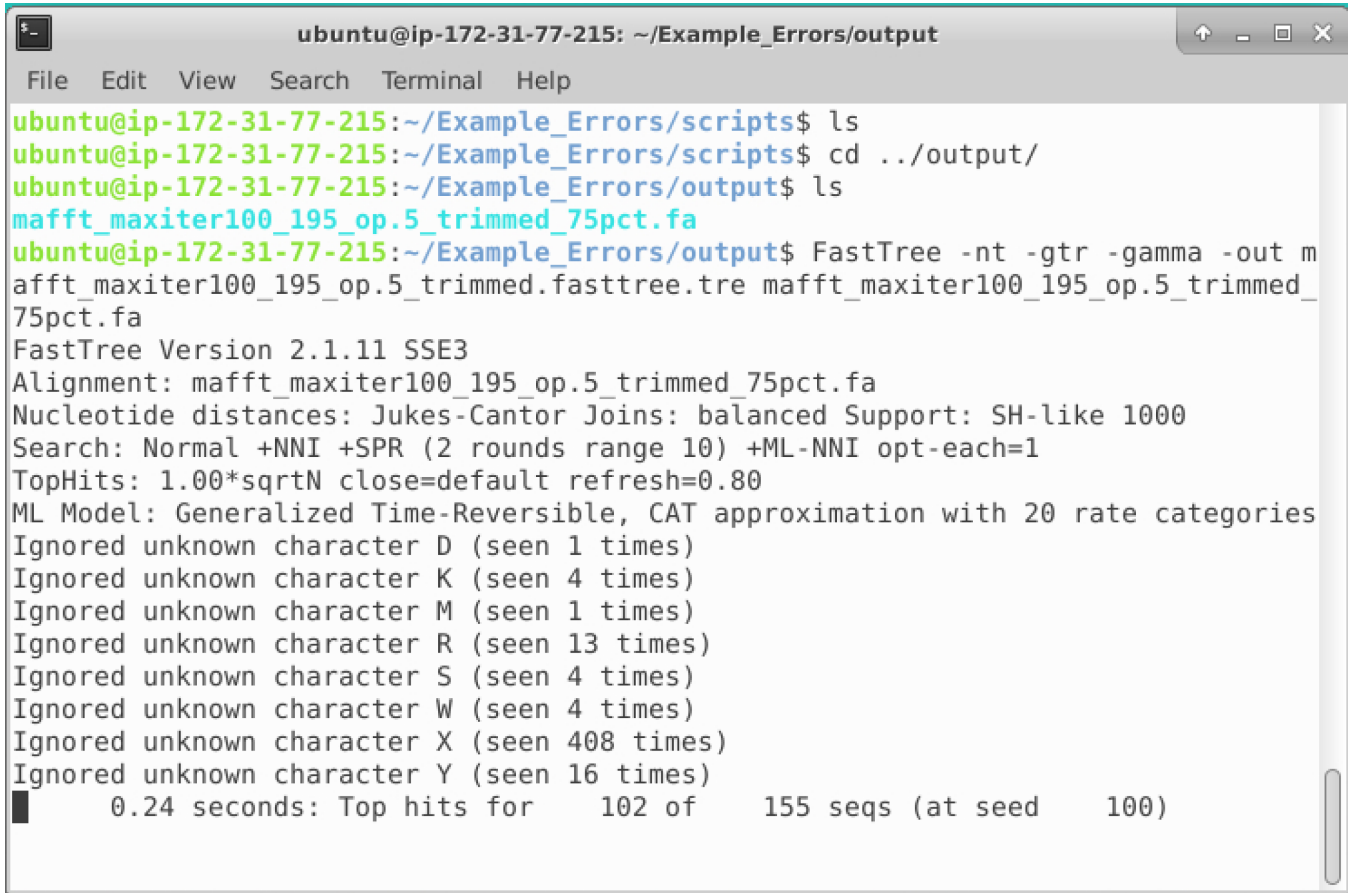
Problem



ubuntu@ip-172-31-77-215: ~/Example_Errors/scripts
File Edit View Search Terminal Help

```
ubuntu@ip-172-31-77-215:~/Example_Errors/scripts$ FastTree -nt -gtr -gamma -out
mafft_maxiter100_195_op.5_trimmed.fasttree.tre mafft_maxiter100_195_op.5_trimmed
_75pct.fa
FastTree Version 2.1.11 SSE3
Alignment: mafft_maxiter100_195_op.5_trimmed_75pct.fa
Nucleotide distances: Jukes-Cantor Joins: balanced Support: SH-like 1000
Search: Normal +NNI +SPR (2 rounds range 10) +ML-NNI opt-each=1
TopHits: 1.00*sqrtN close=default refresh=0.80
ML Model: Generalized Time-Reversible, CAT approximation with 20 rate categories
Cannot read mafft_maxiter100_195_op.5_trimmed_75pct.fa
ubuntu@ip-172-31-77-215:~/Example_Errors/scripts$ 
```

Solution



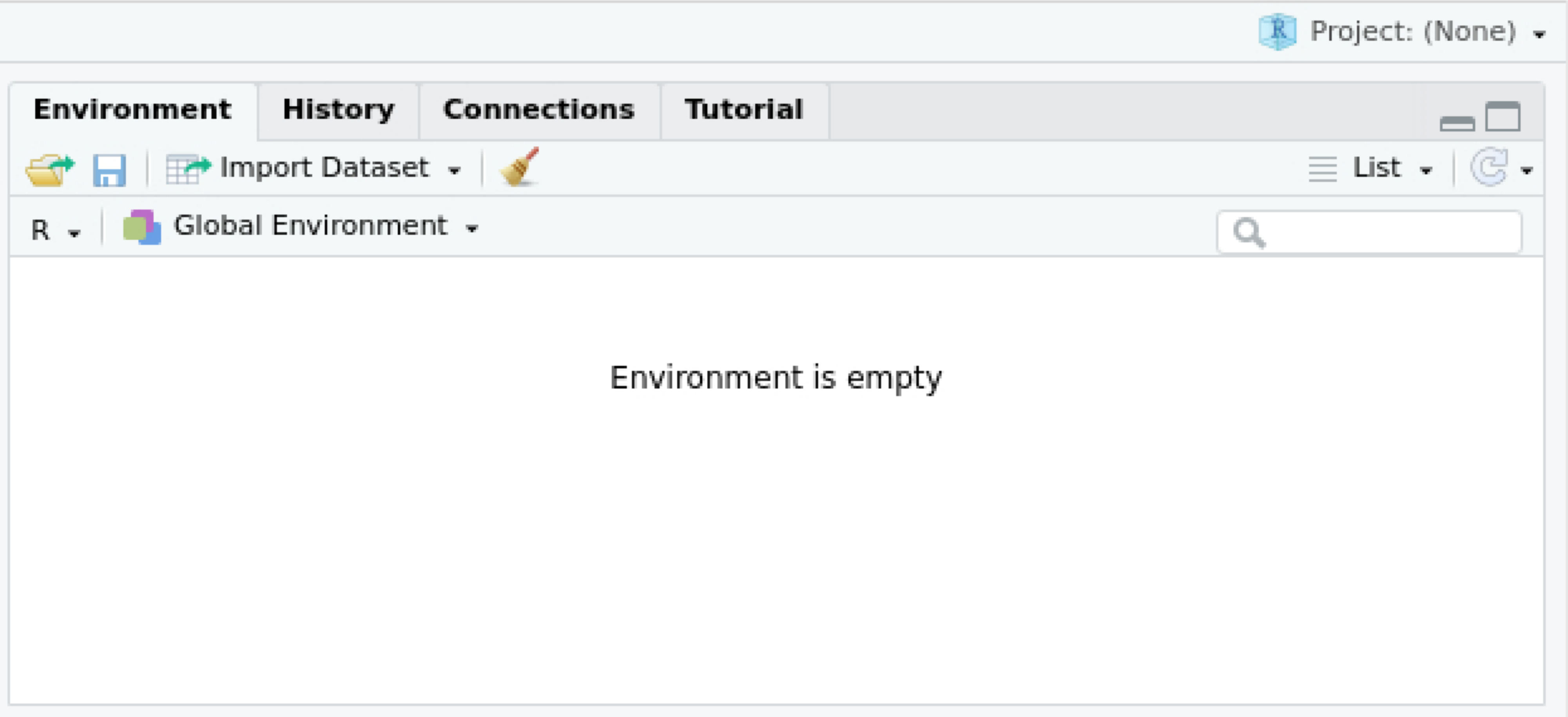
ubuntu@ip-172-31-77-215: ~/Example_Errors/output
File Edit View Search Terminal Help

```
ubuntu@ip-172-31-77-215:~/Example_Errors/scripts$ ls
ubuntu@ip-172-31-77-215:~/Example_Errors/scripts$ cd ../output/
ubuntu@ip-172-31-77-215:~/Example_Errors/output$ ls
mafft_maxiter100_195_op.5_trimmed_75pct.fa
ubuntu@ip-172-31-77-215:~/Example_Errors/output$ FastTree -nt -gtr -gamma -out m
afft_maxiter100_195_op.5_trimmed.fasttree.tre mafft_maxiter100_195_op.5_trimmed_
75pct.fa
FastTree Version 2.1.11 SSE3
Alignment: mafft_maxiter100_195_op.5_trimmed_75pct.fa
Nucleotide distances: Jukes-Cantor Joins: balanced Support: SH-like 1000
Search: Normal +NNI +SPR (2 rounds range 10) +ML-NNI opt-each=1
TopHits: 1.00*sqrtN close=default refresh=0.80
ML Model: Generalized Time-Reversible, CAT approximation with 20 rate categories
Ignored unknown character D (seen 1 times)
Ignored unknown character K (seen 4 times)
Ignored unknown character M (seen 1 times)
Ignored unknown character R (seen 13 times)
Ignored unknown character S (seen 4 times)
Ignored unknown character W (seen 4 times)
Ignored unknown character X (seen 408 times)
Ignored unknown character Y (seen 16 times)
      0.24 seconds: Top hits for    102 of    155 seqs (at seed    100)
```

1. Make sure you are in the right directory
2. Make sure the files you need have the correct path and exist
3. TAB COMPLETE

Problem

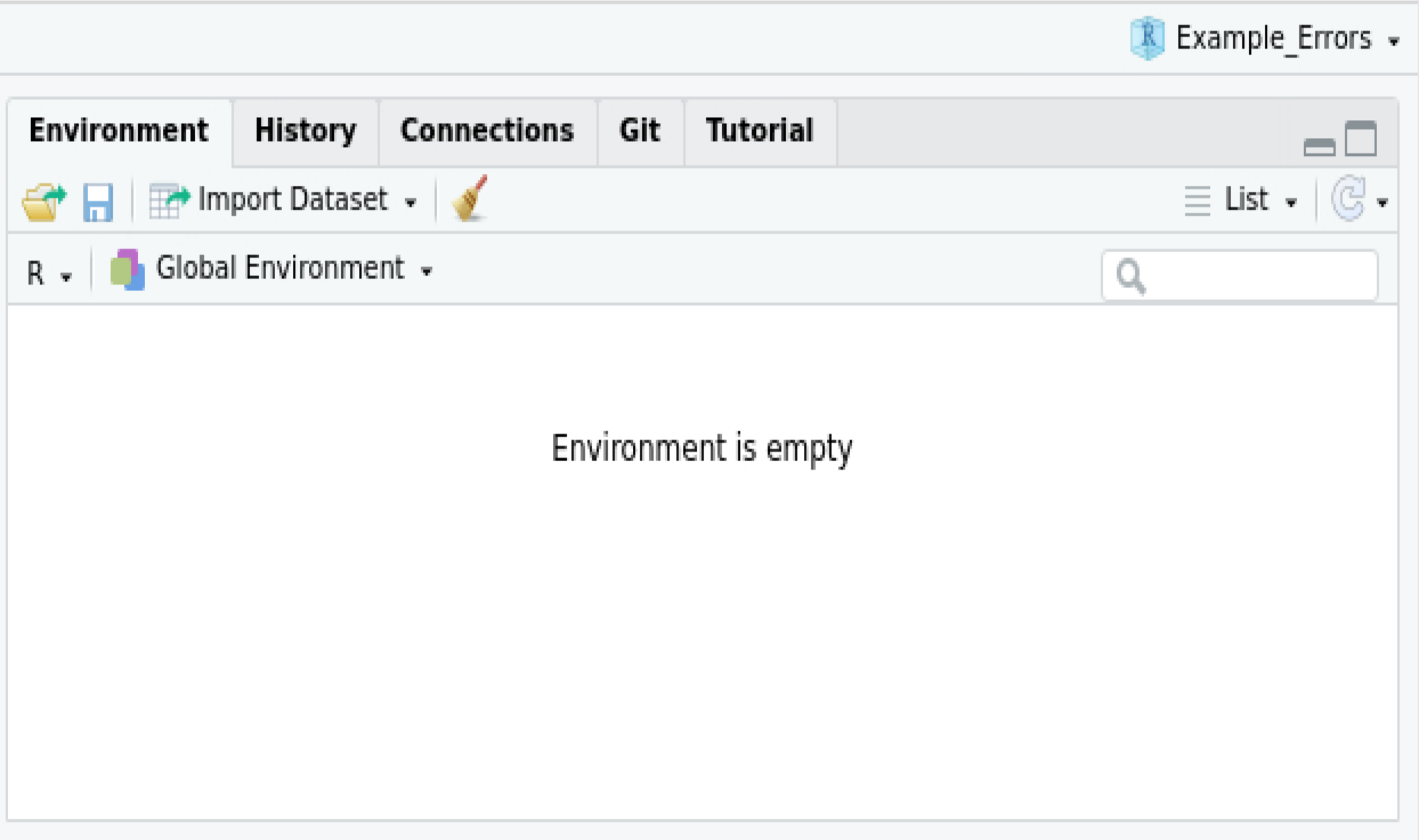
Where is the Git tab?



Environment is empty

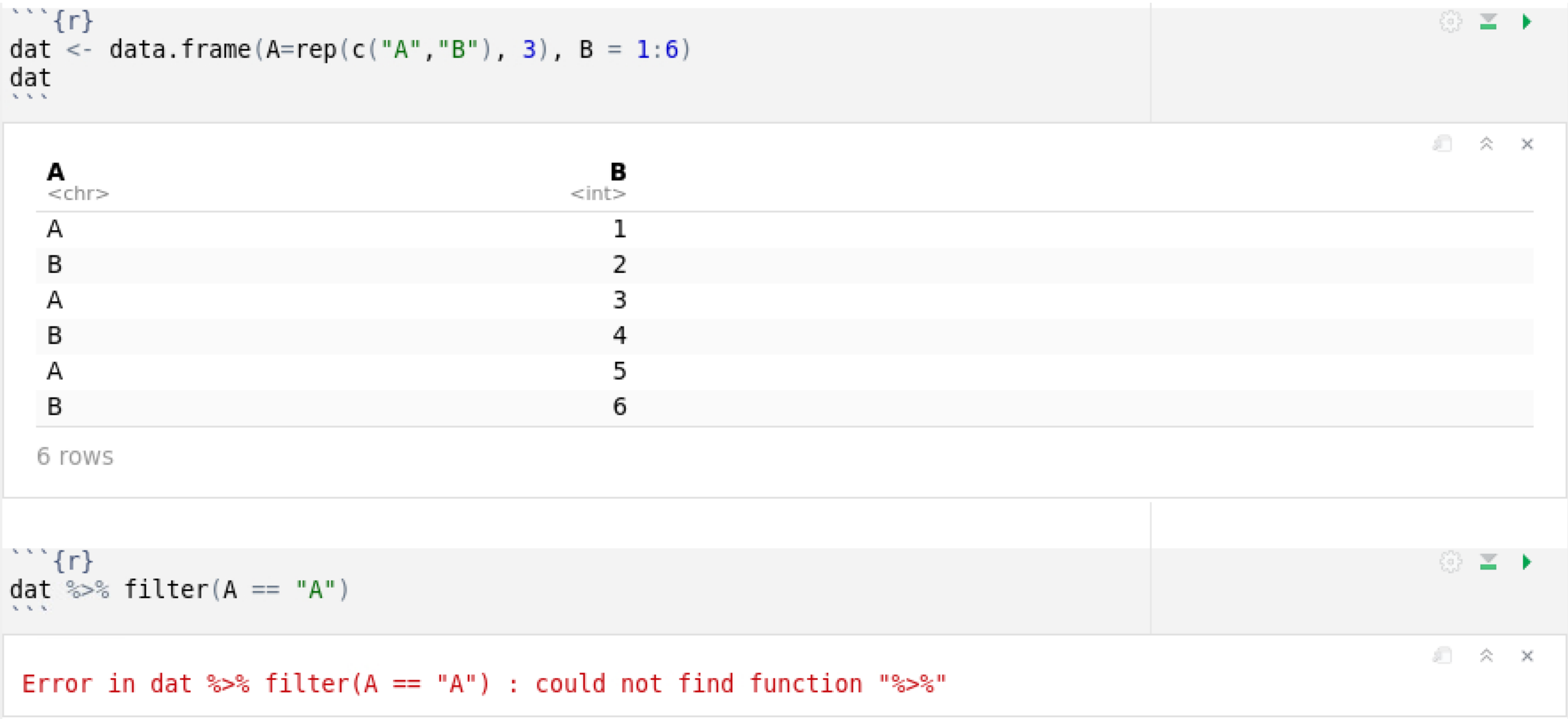
Solution

Load your project by clicking here and selecting the appropriate project



Environment is empty

Problem



~~~
```{r}
dat <- data.frame(A=rep(c("A","B"), 3), B = 1:6)
dat
```
~~~

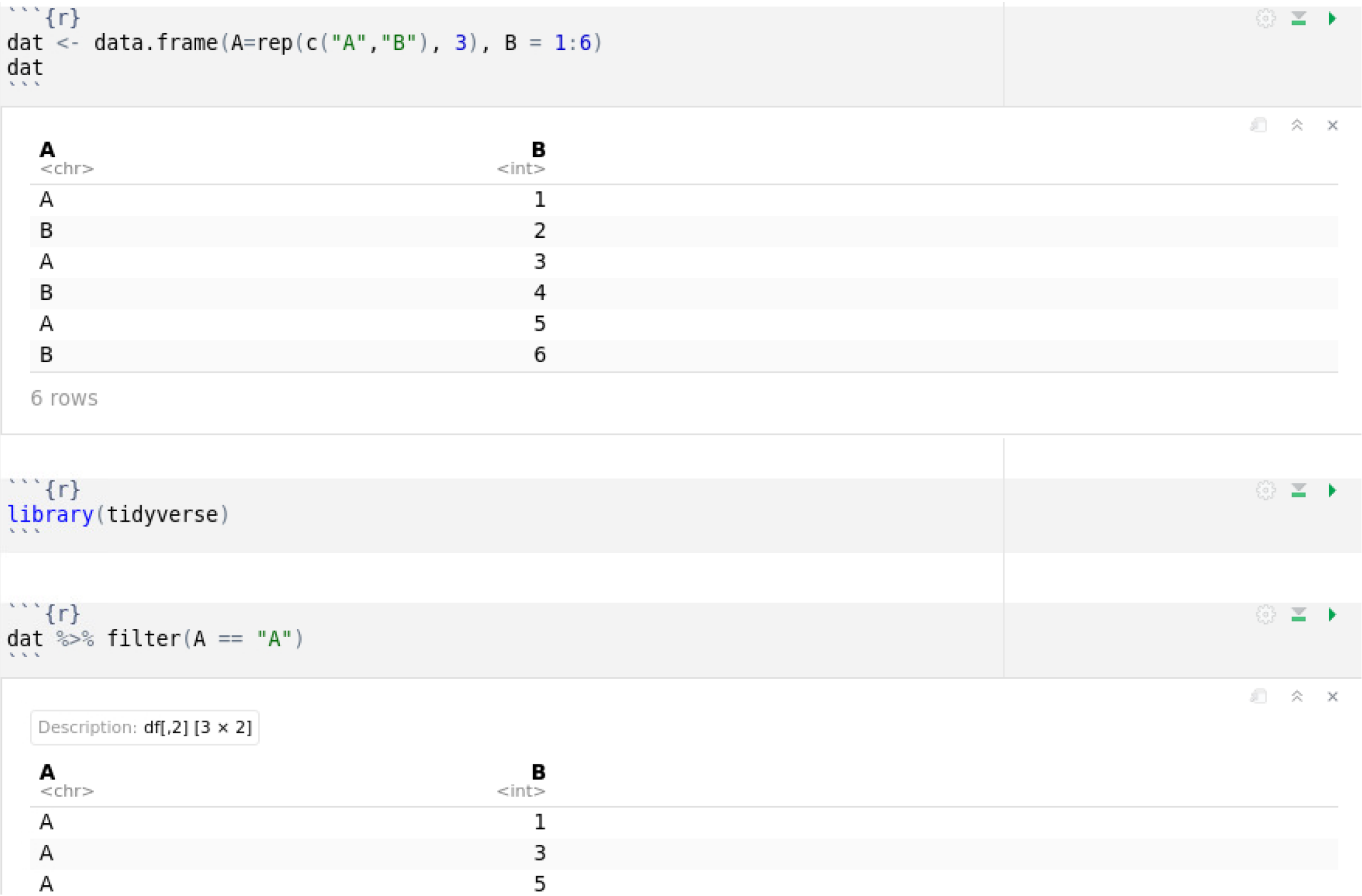
A <chr>	B <int>
A	1
B	2
A	3
B	4
A	5
B	6

6 rows

~~~
```{r}
dat %>% filter(A == "A")
```
~~~

```
Error in dat %>% filter(A == "A") : could not find function "%>%"
```

Solution



~~~
```{r}
dat <- data.frame(A=rep(c("A","B"), 3), B = 1:6)
dat
```
~~~

```
A                    B
<chr>            <int>
A                    1
B                    2
A                    3
B                    4
A                    5
B                    6
6 rows
```

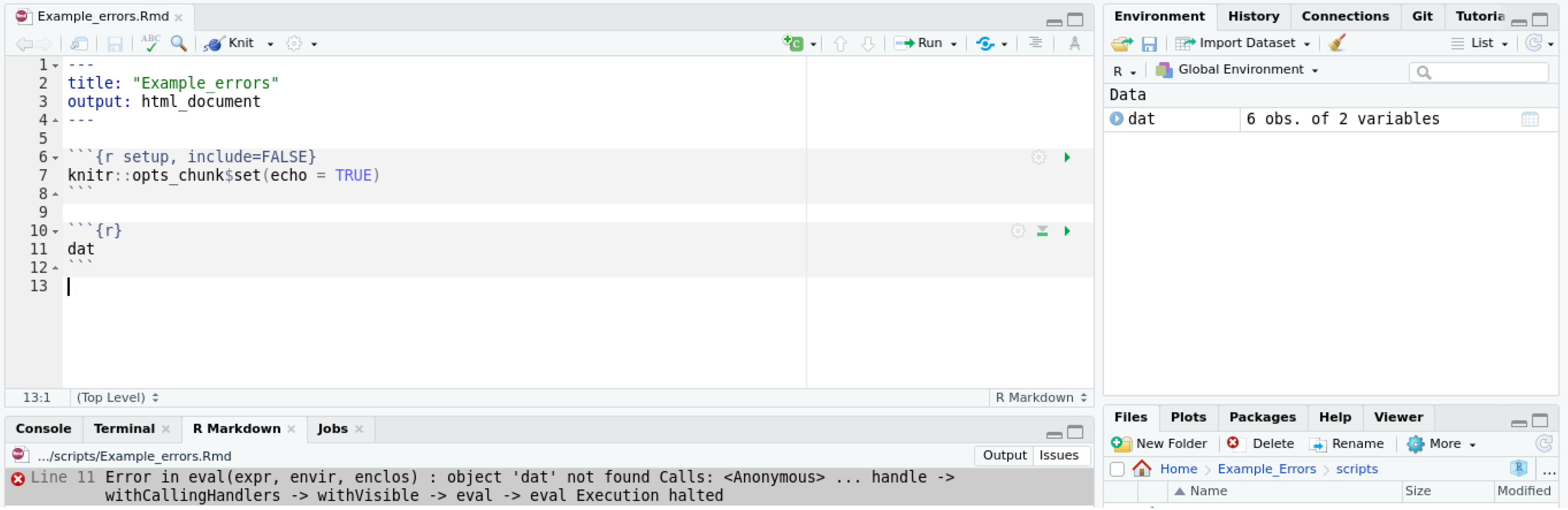
~~~
```{r}
library(tidyverse)
```
~~~

~~~
```{r}
dat %>% filter(A == "A")
```
~~~

```
Description: df[,2] [3 × 2]
A                    B
<chr>            <int>
A                    1
A                    3
A                    5
```

1. Always make sure to load the libraries you will need
2. The simplest solution is to load them at the top of your Rmd

Problem



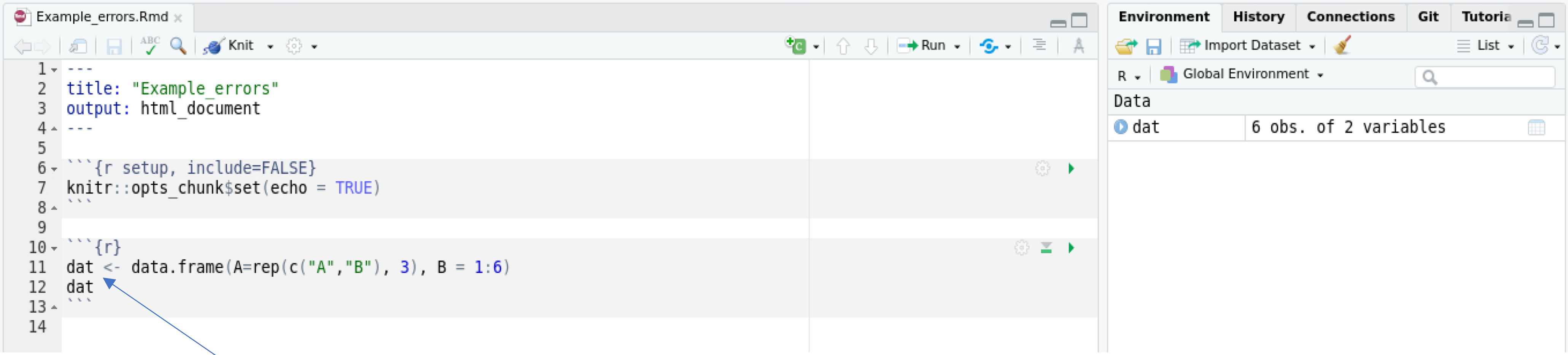
Example_errors.Rmd

~~~
---
title: "Example_errors"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

```{r}
dat
```

```

13:1 (Top Level) R Markdown

.../scripts/Example_errors.Rmd

```
Line 11 Error in eval(expr, envir, enclos) : object 'dat' not found Calls: <Anonymous> ... handle ->
        withCallingHandlers -> withVisible -> eval -> eval Execution halted
```

Data	
dat	6 obs. of 2 variables
```

# Solution

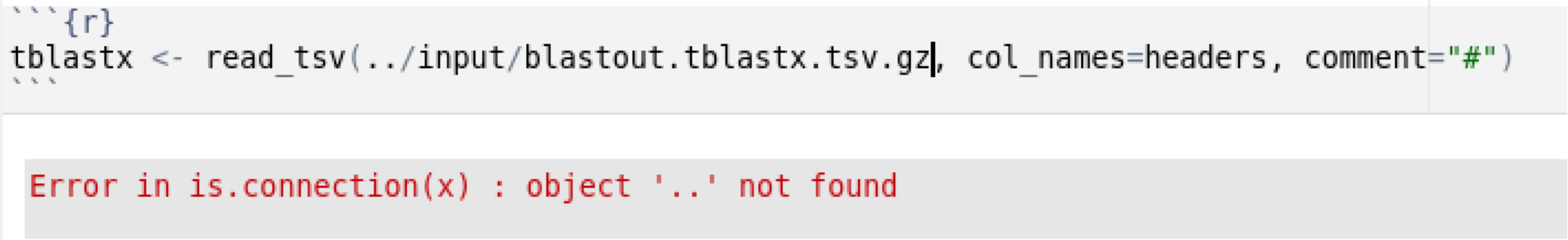


~~~
---
title: "Example_errors"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
dat <- data.frame(A=rep(c("A","B"), 3), B = 1:6)
dat
```
~~~

```
Data
dat     6 obs. of 2 variables
```

Any objects called in your Rmd file must be created in your Rmd file

Problem

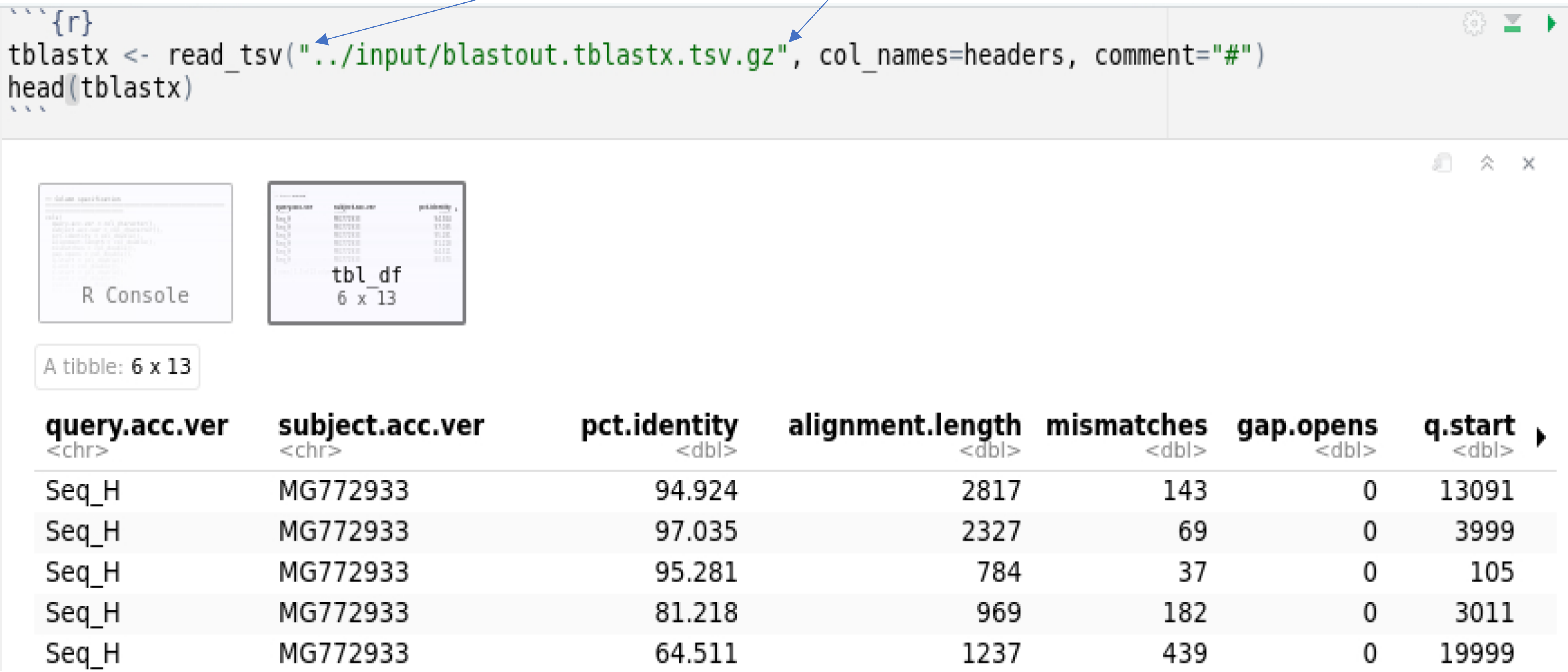


~~~
```{r}
tblastx <- read_tsv(../input/blastout.tblastx.tsv.gz, col_names=headers, comment="#")
```
~~~

```
Error in is.connection(x) : object '..' not found
```

Solution

1. File paths must be surrounded by quotation marks



~~~
```{r}
tblastx <- read_tsv("../input/blastout.tblastx.tsv.gz", col_names=headers, comment="#")
head(tblastx)
```
~~~

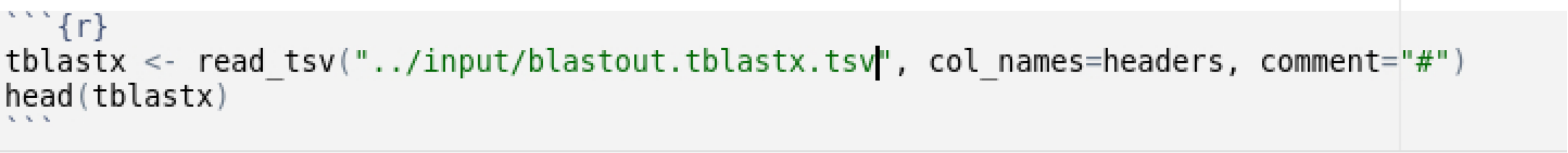
A tibble: 6 x 13

query.acc.ver <chr>	subject.acc.ver <chr>	pct.identity <dbl>	alignment.length <dbl>	mismatches <dbl>	gap.opens <dbl>	q.start <dbl>
Seq_H	MG772933	94.924	2817	143	0	13091
Seq_H	MG772933	97.035	2327	69	0	3999
Seq_H	MG772933	95.281	784	37	0	105
Seq_H	MG772933	81.218	969	182	0	3011
Seq_H	MG772933	64.511	1237	439	0	19999

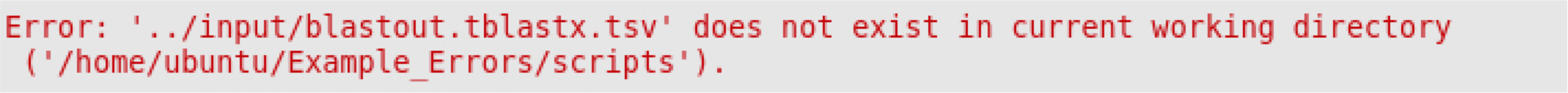
So, when do I quote in R?

- In R, generally quotes are used to enclose literal text
- If there are no quotes, R will look for an object of that name
- So, quote file paths, text for comparison or searching, etc.
- Don’t quote object names
- Don’t quote column names in the tidyverse

Problem



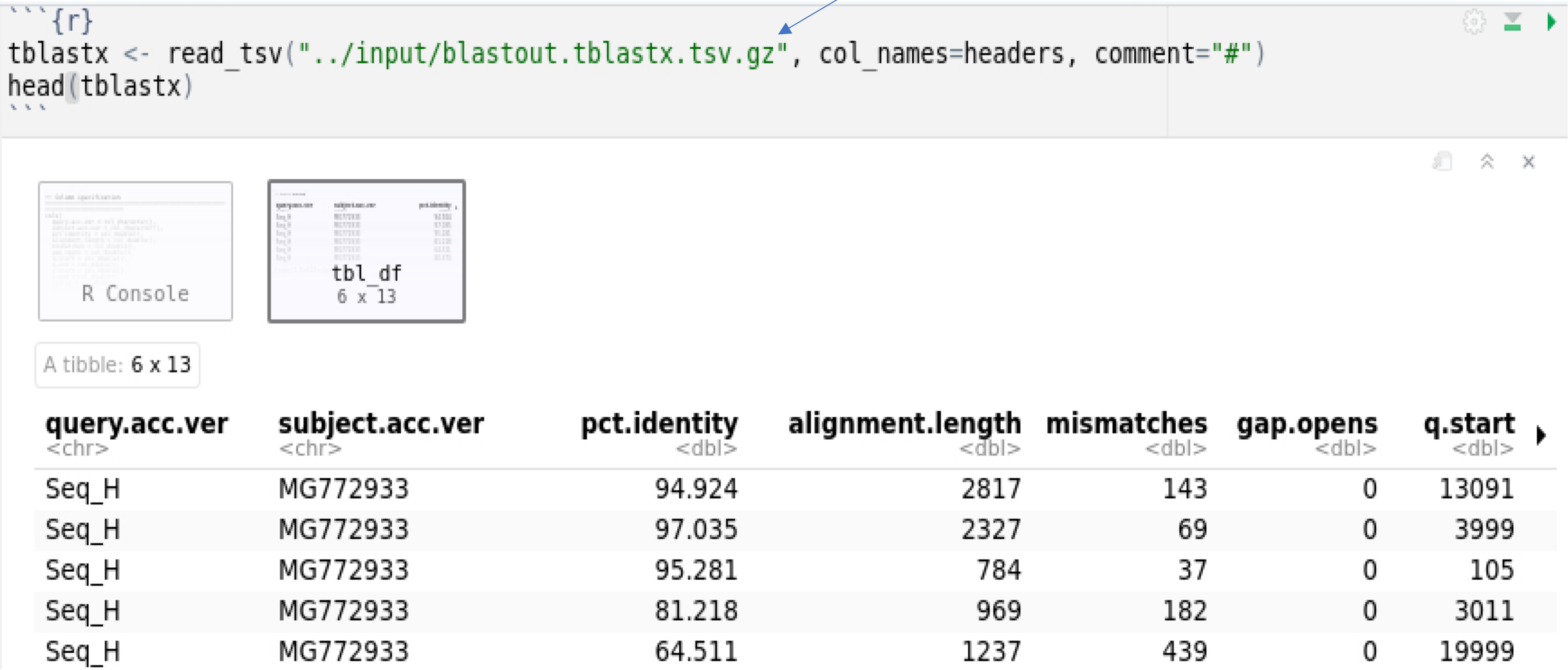
~~~
```{r}
tblastx <- read_tsv("../input/blastout.tblastx.tsv", col_names=headers, comment="#")
head(tblastx)
```
~~~



```
Error: '../input/blastout.tblastx.tsv' does not exist in current working directory
 ('/home/ubuntu/Example_Errors/scripts').
```

Solution

1. File names must be exact
2. Use TAB-completion to be sure

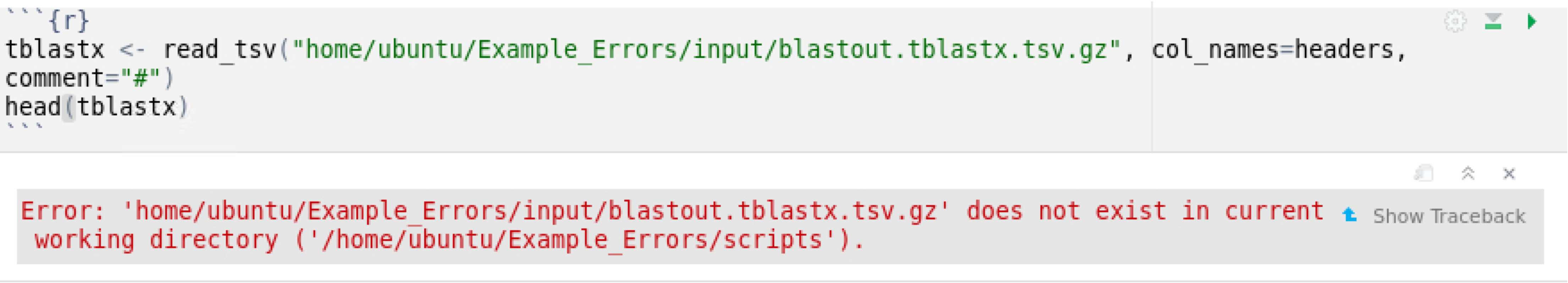


~~~
```{r}
tblastx <- read_tsv("../input/blastout.tblastx.tsv.gz", col_names=headers, comment="#")
head(tblastx)
```
~~~

A tibble: 6 x 13

query.acc.ver	subject.acc.ver	pct.identity	alignment.length	mismatches	gap.opens	q.start
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Seq_H	MG772933	94.924	2817	143	0	13091
Seq_H	MG772933	97.035	2327	69	0	3999
Seq_H	MG772933	95.281	784	37	0	105
Seq_H	MG772933	81.218	969	182	0	3011
Seq_H	MG772933	64.511	1237	439	0	19999

Problem

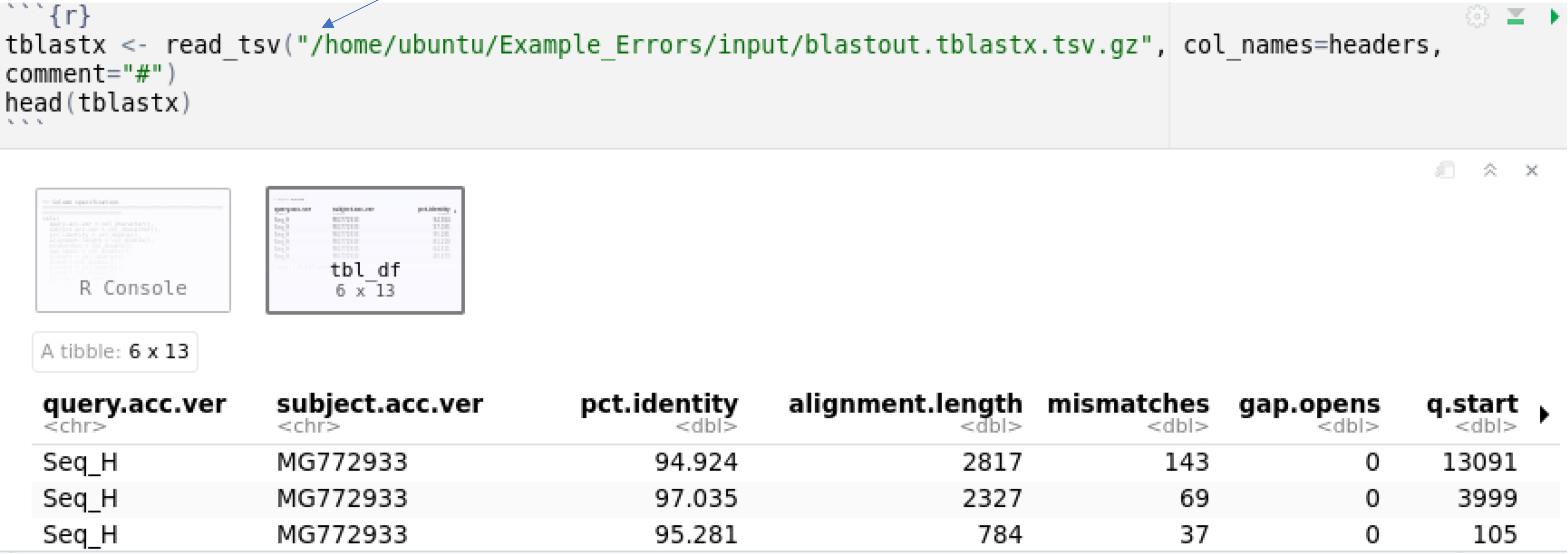


~~~
```{r}
tblastx <- read_tsv("home/ubuntu/Example_Errors/input/blastout.tblastx.tsv.gz", col_names=headers,
comment="#")
head(tblastx)
```
~~~

```
Error: 'home/ubuntu/Example_Errors/input/blastout.tblastx.tsv.gz' does not exist in current
 working directory ('/home/ubuntu/Example_Errors/scripts').
```

Solution

1. If you use the absolute path, it must start with “/”



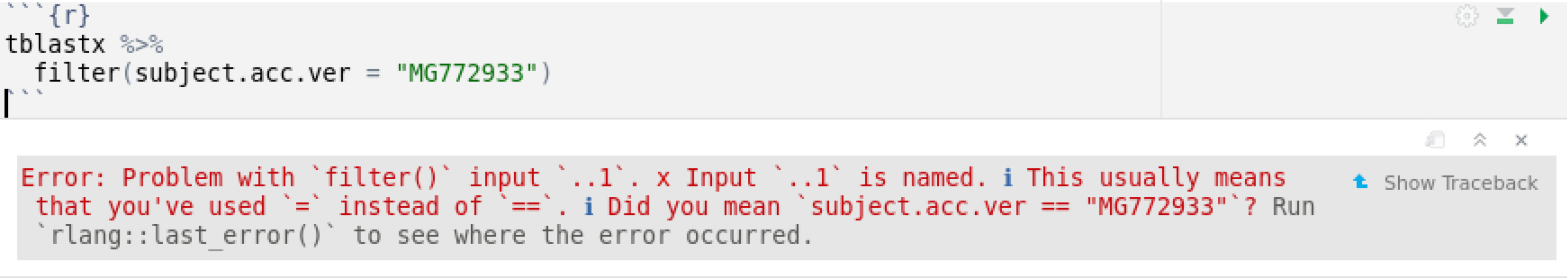
~~~
```{r}
tblastx <- read_tsv("/home/ubuntu/Example_Errors/input/blastout.tblastx.tsv.gz", col_names=headers,
comment="#")
head(tblastx)
```
~~~

A tibble: 6 x 13

query.acc.ver	subject.acc.ver	pct.identity	alignment.length	mismatches	gap.opens	q.start
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Seq_H	MG772933	94.924	2817	143	0	13091
Seq_H	MG772933	97.035	2327	69	0	3999
Seq_H	MG772933	95.281	784	37	0	105

Read the Error messages!

Problem

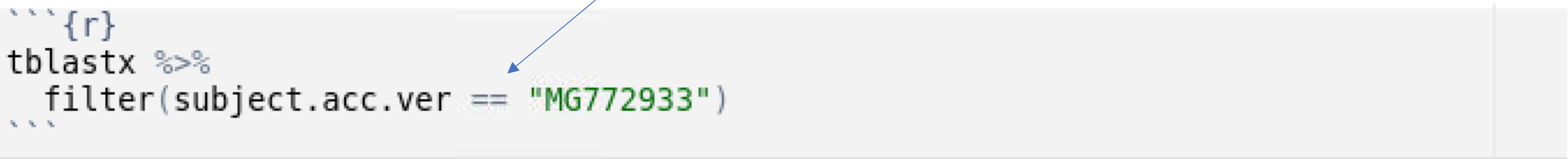


~~~
```{r}
tblastx %>%
  filter(subject.acc.ver = "MG772933")
```
~~~

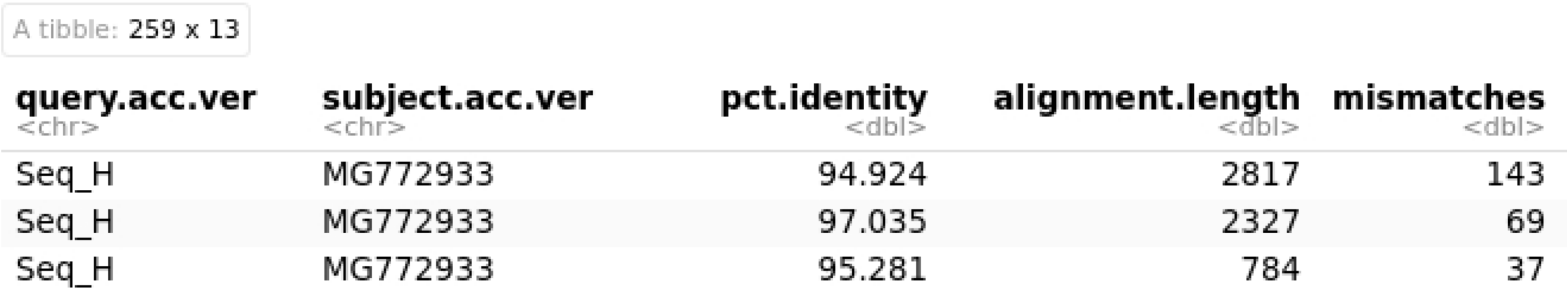
```
Error: Problem with `filter()` input `..1`. x Input `..1` is named. ℹ This usually means
 that you've used `=` instead of `==`. ℹ Did you mean `subject.acc.ver == "MG772933"`? Run
 `rlang::last_error()` to see where the error occurred.
```

Solution

“=“ is used to assign a value to an object
“==“ is used to compare two objects



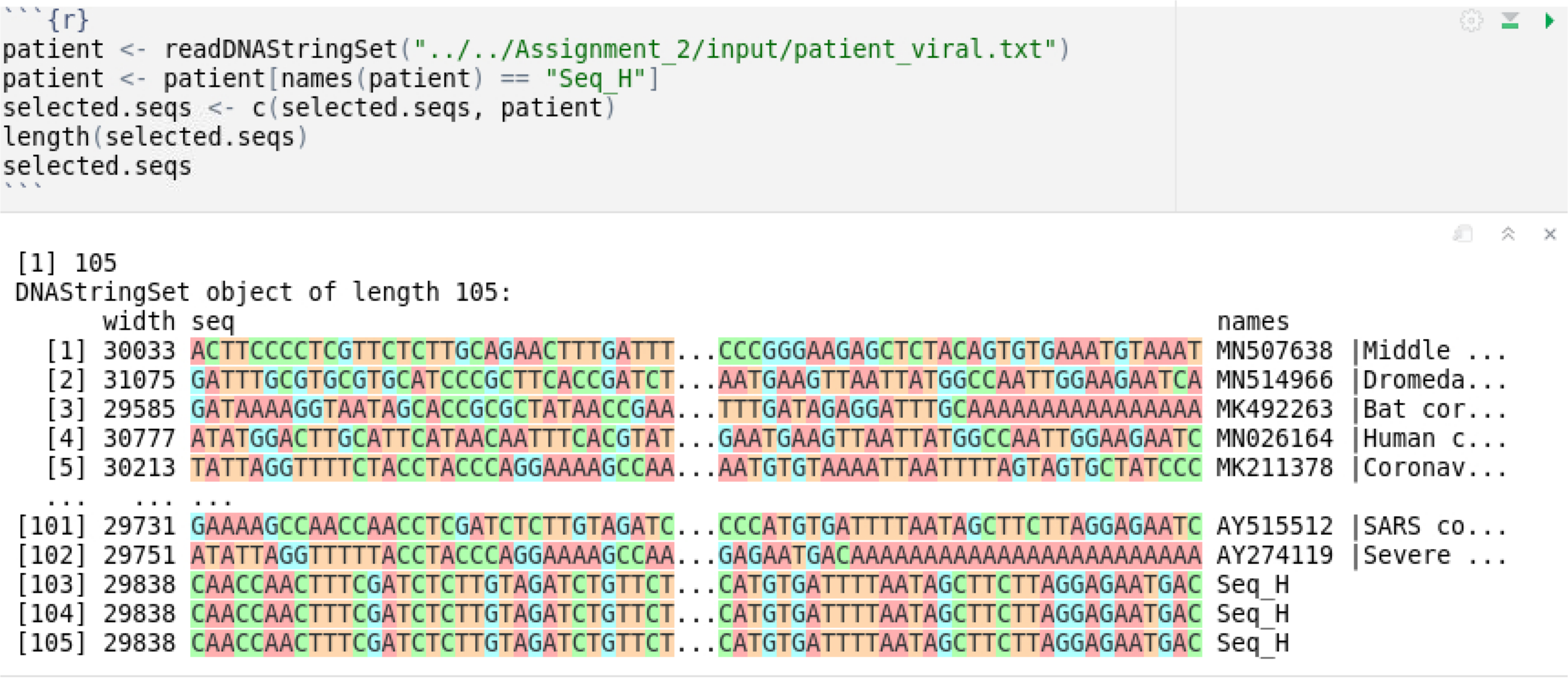
~~~
```{r}
tblastx %>%
  filter(subject.acc.ver == "MG772933")
```
~~~



A tibble: 259 x 13

query.acc.ver	subject.acc.ver	pct.identity	alignment.length	mismatches
<chr>	<chr>	<dbl>	<dbl>	<dbl>
Seq_H	MG772933	94.924	2817	143
Seq_H	MG772933	97.035	2327	69
Seq_H	MG772933	95.281	784	37

Problem

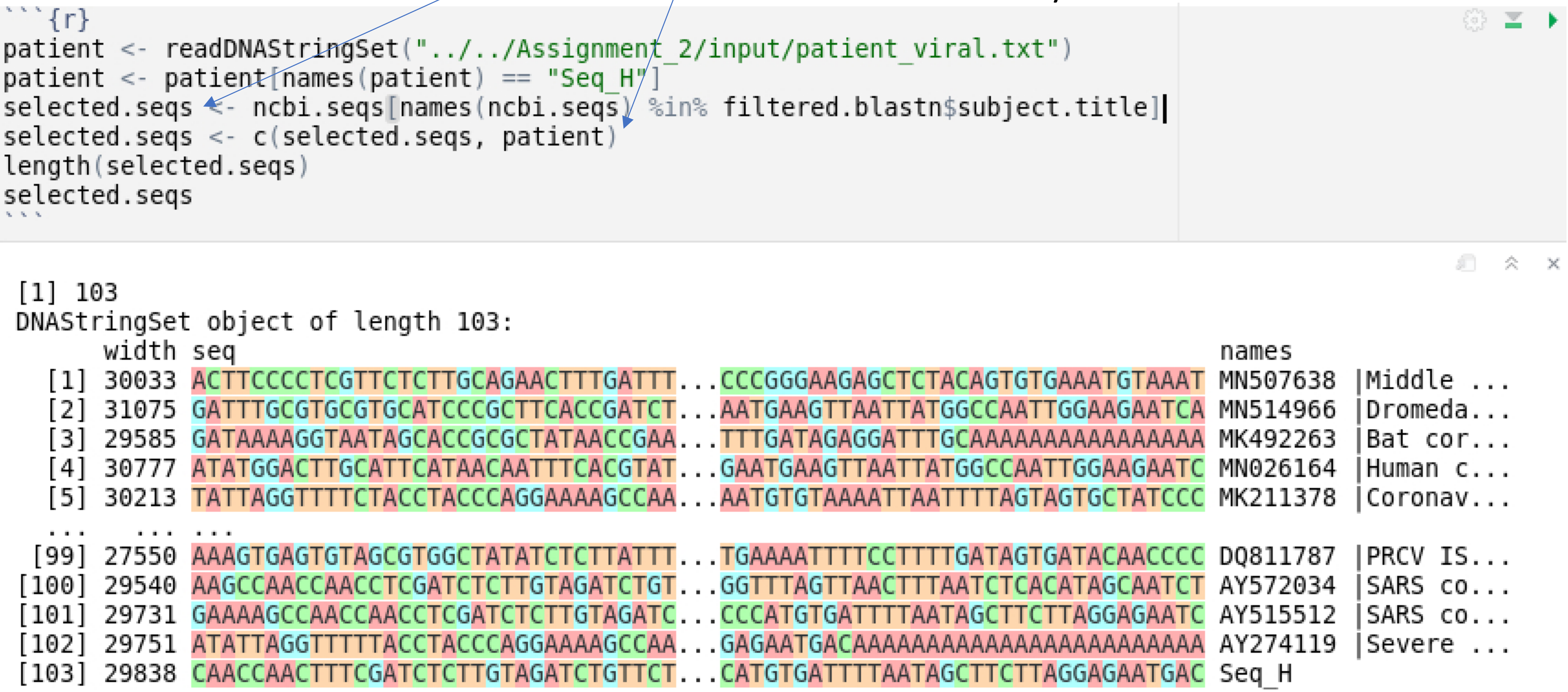


~~~
```{r}
patient <- readDNAStringSet("../../Assignment_2/input/patient_viral.txt")
patient <- patient[names(patient) == "Seq_H"]
selected.seqs <- c(selected.seqs, patient)
length(selected.seqs)
selected.seqs
```
~~~

```
[1] 105
DNAStringSet object of length 105:
      width seq                                                   names               
  [1] 30033 ACTTCCCCTCGTTCTCTTGCAGAACTTTGATTT...CCCGGGAAGAGCTCTACAGTGTGAAATGTAAAT MN507638 |Middle ...
  [2] 31075 GATTTGCGTGCGTGCATCCCGCTTCACCGATCT...AATGAAGTTAATTATGGCCAATTGGAAGAATCA MN514966 |Dromeda...
  [3] 29585 GATAAAAGGTAATAGCACCGCGCTATAACCGAA...TTTGATAGAGGATTTGCAAAAAAAAAAAAAAAA MK492263 |Bat cor...
  [4] 30777 ATATGGACTTGCATTCATAACAATTTCACGTAT...GAATGAAGTTAATTATGGCCAATTGGAAGAATC MN026164 |Human c...
  [5] 30213 TATTAGGTTTTCTACCTACCCAGGAAAAGCCAA...AATGTGTAAAATTAATTTTAGTAGTGCTATCCC MK211378 |Coronav...
  ...   ... ...
[101] 29731 GAAAAGCCAACCAACCTCGATCTCTTGTAGATC...CCCATGTGATTTTAATAGCTTCTTAGGAGAATC AY515512 |SARS co...
[102] 29751 ATATTAGGTTTTTACCTACCCAGGAAAAGCCAA...GAGAATGACAAAAAAAAAAAAAAAAAAAAAAAA AY274119 |Severe ...
[103] 29838 CAACCAACTTTCGATCTCTTGTAGATCTGTTCT...CATGTGATTTTAATAGCTTCTTAGGAGAATGAC Seq_H
[104] 29838 CAACCAACTTTCGATCTCTTGTAGATCTGTTCT...CATGTGATTTTAATAGCTTCTTAGGAGAATGAC Seq_H
[105] 29838 CAACCAACTTTCGATCTCTTGTAGATCTGTTCT...CATGTGATTTTAATAGCTTCTTAGGAGAATGAC Seq_H
```

Solution

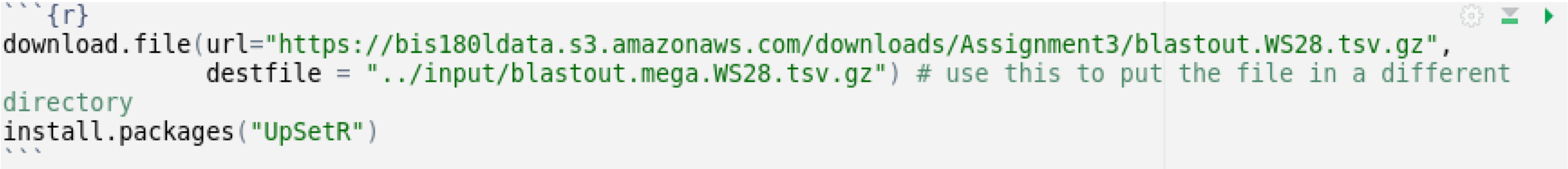
1. If you run this line multiple times it will add Seq_H to selected.seqs each time
2. Recreate selected.seqs if you want to remove the additional Seq_H entries
3. When knitting, this will not be an issue as each line is only ran once



~~~
```{r}
patient <- readDNAStringSet("../../Assignment_2/input/patient_viral.txt")
patient <- patient[names(patient) == "Seq_H"]
selected.seqs <- ncbi.seqs[names(ncbi.seqs) %in% filtered.blastn$subject.title]
selected.seqs <- c(selected.seqs, patient)
length(selected.seqs)
selected.seqs
```
~~~

```
[1] 103
DNAStringSet object of length 103:
      width seq                                                names
  [1] 30033 ACTTCCCCTCGTTCTCTTGCAGAACTTTGATTT...CCCGGGAAGAGCTCTACAGTGTGAAATGTAAAT MN507638 |Middle ...
  [2] 31075 GATTTGCGTGCGTGCATCCCGCTTCACCGATCT...AATGAAGTTAATTATGGCCAATTGGAAGAATCA MN514966 |Dromeda...
  [3] 29585 GATAAAAGGTAATAGCACCGCGCTATAACCGAA...TTTGATAGAGGATTTGCAAAAAAAAAAAAAAAA MK492263 |Bat cor...
  [4] 30777 ATATGGACTTGCATTCATAACAATTTCACGTAT...GAATGAAGTTAATTATGGCCAATTGGAAGAATC MN026164 |Human c...
  [5] 30213 TATTAGGTTTTCTACCTACCCAGGAAAAGCCAA...AATGTGTAAAATTAATTTTAGTAGTGCTATCCC MK211378 |Coronav...
  ...   ... ...
 [99] 27550 AAAGTGAGTGTAGCGTGGCTATATCTCTTATTT...TGAAAATTTTCCTTTTGATAGTGATACAACCCC DQ811787 |PRCV IS...
[100] 29540 AAGCCAACCAACCTCGATCTCTTGTAGATCTGT...GGTTTAGTTAACTTTAATCTCACATAGCAATCT AY572034 |SARS co...
[101] 29731 GAAAAGCCAACCAACCTCGATCTCTTGTAGATC...CCCATGTGATTTTAATAGCTTCTTAGGAGAATC AY515512 |SARS co...
[102] 29751 ATATTAGGTTTTTACCTACCCAGGAAAAGCCAA...GAGAATGACAAAAAAAAAAAAAAAAAAAAAAAA AY274119 |Severe ...
[103] 29838 CAACCAACTTTCGATCTCTTGTAGATCTGTTCT...CATGTGATTTTAATAGCTTCTTAGGAGAATGAC Seq_H
```

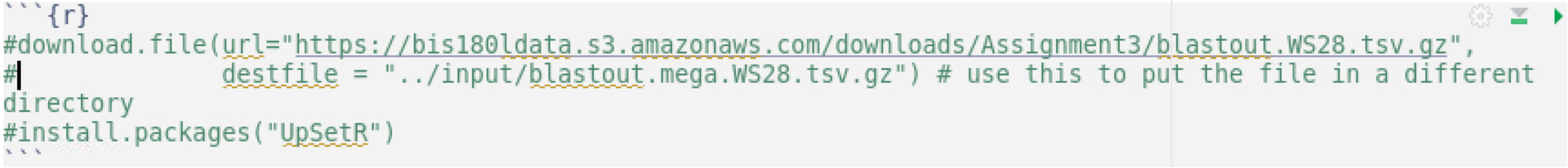
Problem



~~~
```{r}
download.file(url="https://bis180ldata.s3.amazonaws.com/downloads/Assignment3/blastout.WS28.tsv.gz",
               destfile = "../input/blastout.mega.WS28.tsv.gz") # use this to put the file in a different
directory
install.packages("UpSetR")
```
~~~

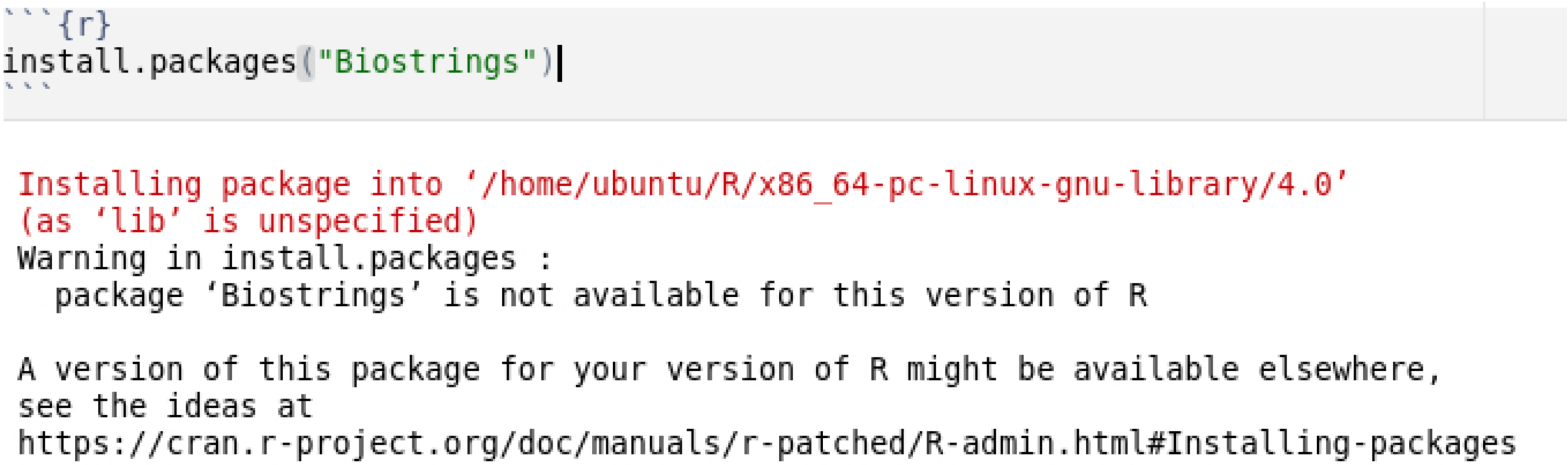
Solution

Once you install a package or download a file, there’s no need to reinstall or download each time you run/knit your code. Comment the commands out to save time



~~~
```{r}
#download.file(url="https://bis180ldata.s3.amazonaws.com/downloads/Assignment3/blastout.WS28.tsv.gz",
#              destfile = "../input/blastout.mega.WS28.tsv.gz") # use this to put the file in a different
directory
#install.packages("UpSetR")
```
~~~

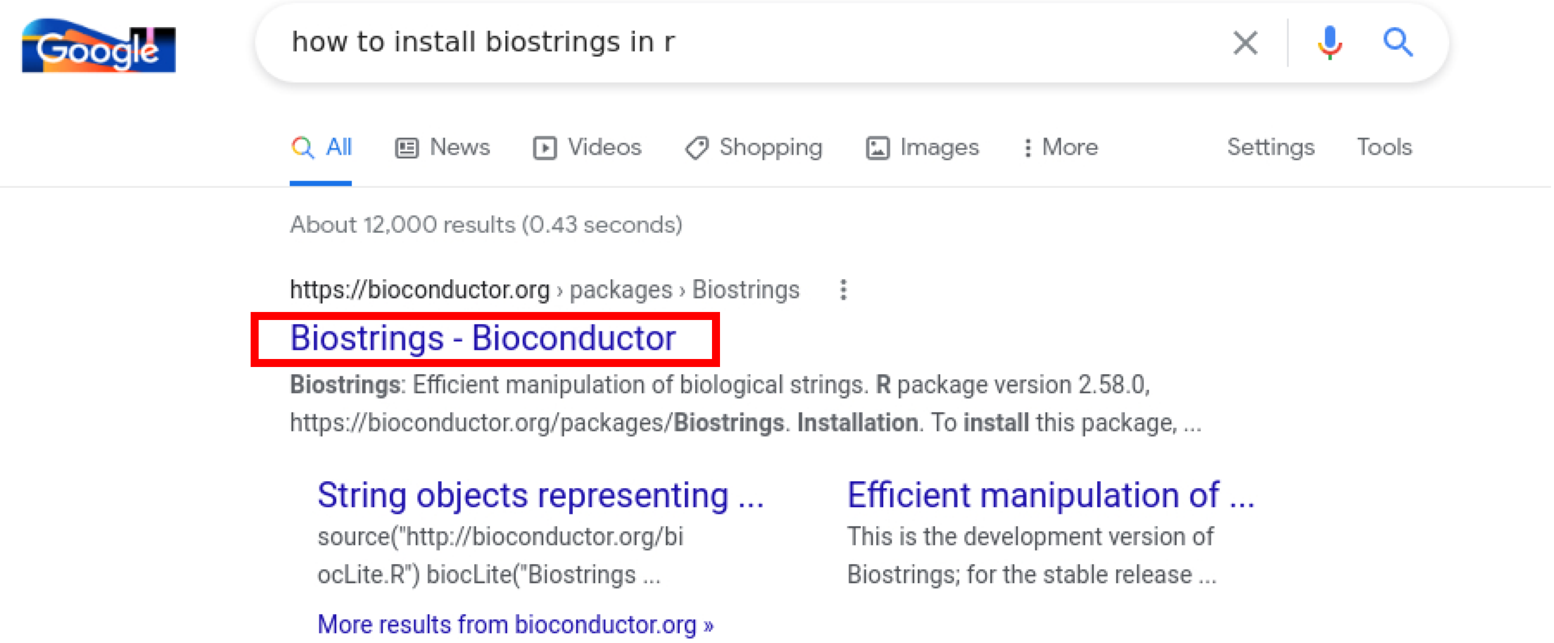
Problem



~~~
```{r}
install.packages("Biostrings")
```
~~~

```
Installing package into ‘/home/ubuntu/R/x86_64-pc-linux-gnu-library/4.0’
(as ‘lib’ is unspecified)
Warning in install.packages :
  package ‘Biostrings’ is not available for this version of R

A version of this package for your version of R might be available elsewhere,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

Solution



how to install biostrings in r

About 12,000 results (0.43 seconds)

https://bioconductor.org › packages › Biostrings

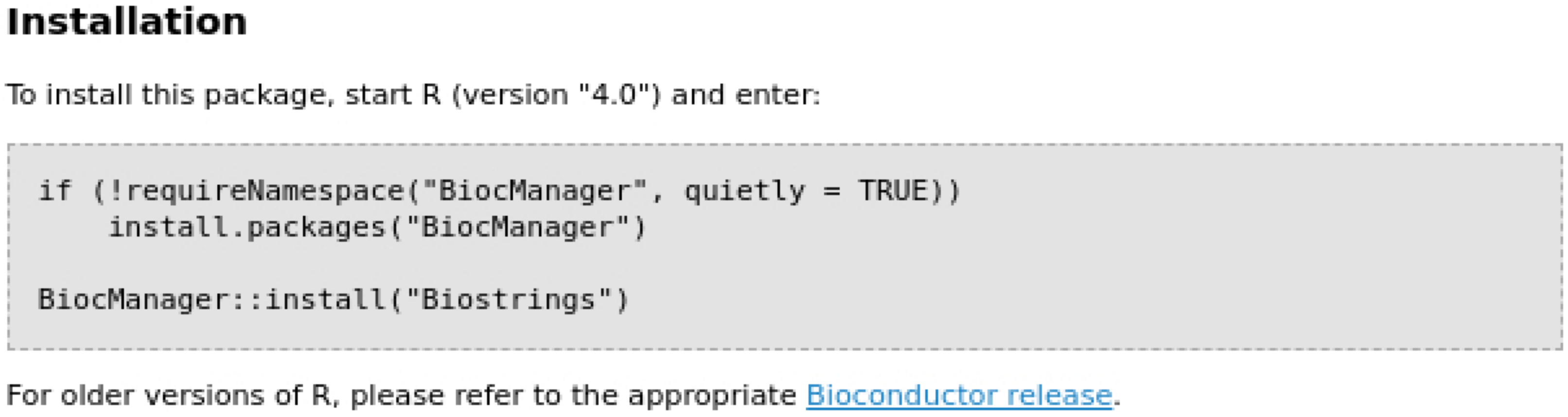
Biostrings - Bioconductor

Biostrings: Efficient manipulation of biological strings. **R** package version 2.58.0, https://bioconductor.org/packages/**Biostrings**. **Installation**. To **install** this package, ...

String objects representing ...
source("http://bioconductor.org/biocLite.R") biocLite("Biostrings ...

Efficient manipulation of ...
This is the development version of Biostrings; for the stable release ...

More results from bioconductor.org »



Installation

To install this package, start R (version "4.0") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("Biostrings")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Not all R libraries are stored in CRAN

Some require additional steps to install

Overall Takeaways

- Be skeptical
- Know your working directory
- Check file paths
- Use tab complete
- Inspect your objects
- Read error messages