

Introduction to UNIX

Kamil S Jaron, Marc Robinson-Rechavi

22.9. 2016

Sequencing reads of 15 genomes

\approx 1GB of plain text data / species

\approx 640938 of pages

Why UNIX?

How to check a file??

Notepad?




not well suited for big files

Office?



Use at least read-only mode

Gene name errors are widespread in the scientific literature

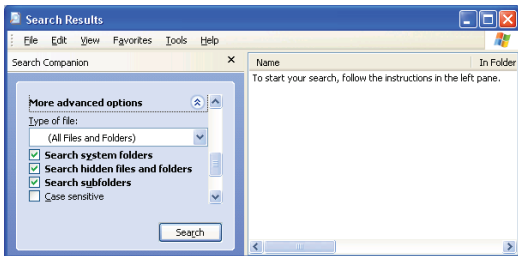
Mark Ziemann, Yotam Eren and Assam El-Osta 

Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

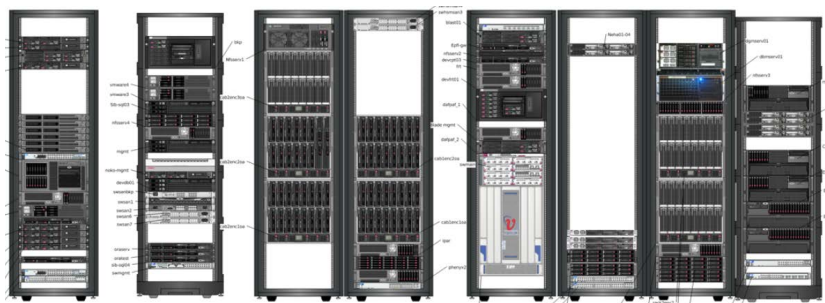
Published: 23 August 2016

Abstract

The spreadsheet software Microsoft Excel when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

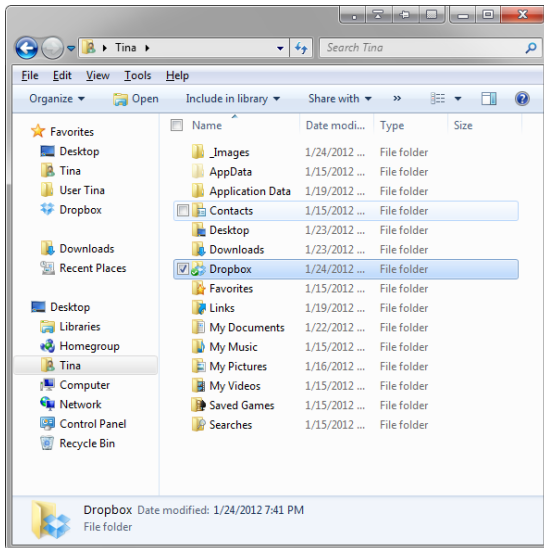


Cluster computing

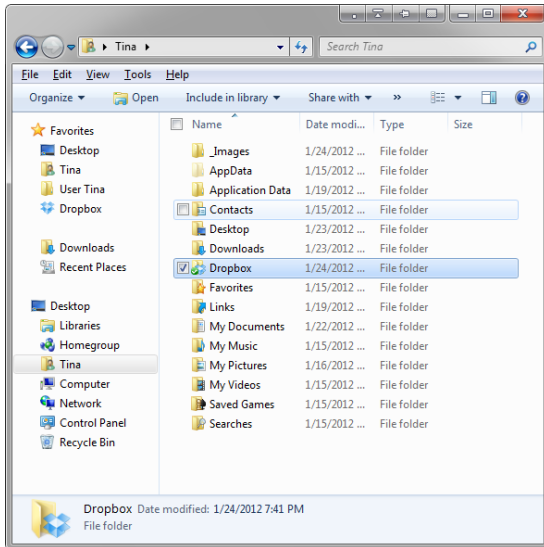


How can UNIX help us?

Command line \approx explorer + toolbox of commands



bash \approx explorer



kjaron@frt:~\$



user



computer



location



\$ for user

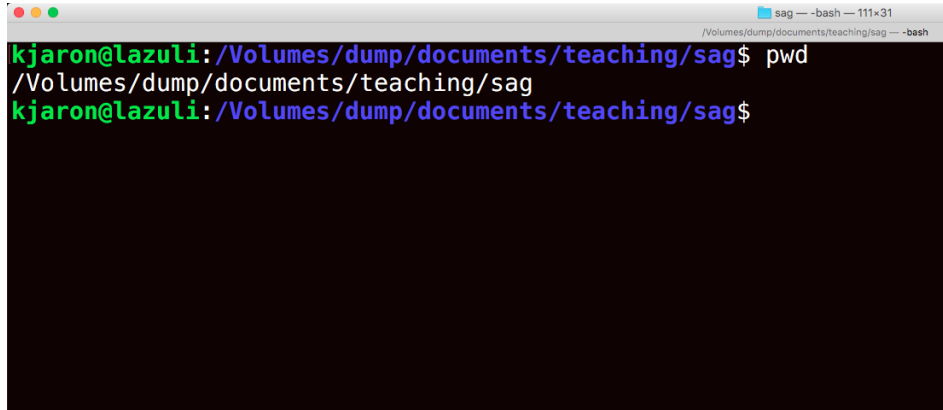
for administrator



space for a command

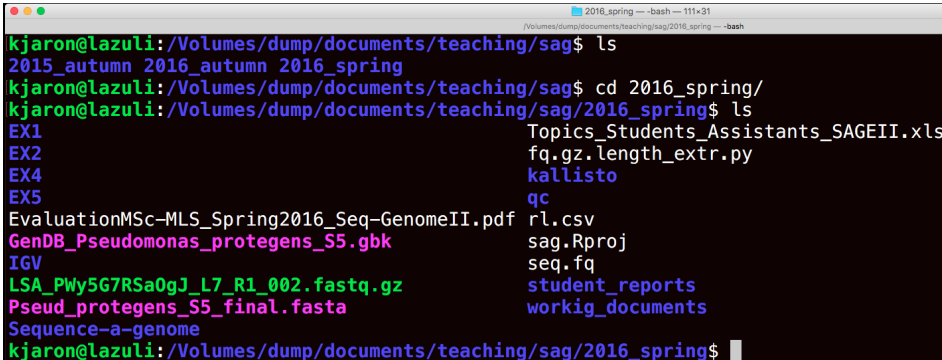


bash \approx explorer (where am I?)



```
sag -- -bash -- 111x31
/Volumes/dump/documents/teaching/sag -- -bash
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ pwd
/Volumes/dump/documents/teaching/sag
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$
```

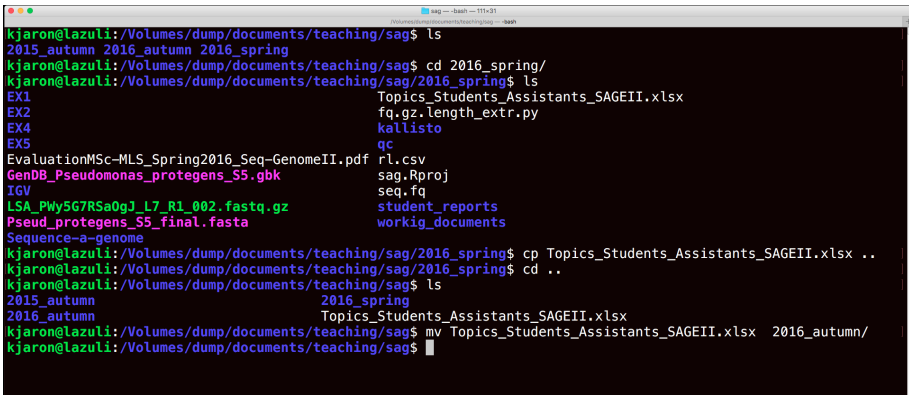
bash \approx explorer (browse directories)



A terminal window titled "2016_spring -- -bash -- 111x31" with a subtitle "/Volumes/dump/documents/teaching/sag/2016_spring -- -bash". The terminal shows a user named "kjaron@lazuli" navigating through a directory structure. The user starts in "/Volumes/dump/documents/teaching/sag" and lists files, then changes to the "2016_spring/" subdirectory and lists its contents. The files listed include directories for previous semesters, a list of topics, a genome evaluation PDF, a GenDB file, an IGV file, a fastq file, a fasta file, a sequence file, and a directory for student reports and work documents.

```
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ ls
2015_autumn 2016_autumn 2016_spring
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ cd 2016_spring/
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ ls
EX1                                Topics_Students_Assistants_SAGEII.xls
EX2                                fq.gz.length_extr.py
EX4                                kallisto
EX5                                qc
EvaluationMSc-MLS_Spring2016_Seq-GenomeII.pdf  rl.csv
GenDB_Pseudomonas_protegens_S5.gbk            sag.Rproj
IGV                                              seq.fq
LSA_PWy5G7RSa0gJ_L7_R1_002.fastq.gz          student_reports
Pseud_protegens_S5_final.fasta                workig_documents
Sequence-a-genome
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$
```

bash \approx explorer (copy and move files)



```
sag -- -bash -- 111s-01
/Volumes/dump/documents/teaching/sag -- -bash

kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ ls
2015_autumn 2016_autumn 2016_spring
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ cd 2016_spring/
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ ls
EX1                                Topics_Students_Assistants_SAGEII.xlsx
EX2                                fq.gz.length_extr.py
EX4                                kallisto
EX5                                qc
EvaluationMSc-MLS_Spring2016_Seq-GenomeII.pdf  rl.csv
GenDB_Pseudomonas_protegens_S5.gbk           sag.Rproj
IGV                                             seq.fq
LSA_PWy5G7RSa0gJ_L7_R1_002.fastq.gz          student_reports
Pseud_protegens_S5_final.fasta               workig_documents
Sequence-a-genome
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ cp Topics_Students_Assistants_SAGEII.xlsx ..
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ cd ..
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ ls
2015_autumn                2016_spring
2016_autumn                Topics_Students_Assistants_SAGEII.xlsx
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ mv Topics_Students_Assistants_SAGEII.xlsx 2016_autumn/
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$
```

relative paths

. # this directory

.. # parent directory

~ # my home directory

absolute path

/ # root directory

bash \approx explorer (remove files)

```
2016_autumn --- bash --- 111x31
/Volumes/dump/documents/teaching/sag 2016_autumn --- bash
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ ls
2015_autumn 2016_autumn 2016_spring
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ cd 2016_spring/
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ ls
EX1                                Topics_Students_Assistants_SAGEII.xlsx
EX2                                fq.gz.length_extr.py
EX4                                kallisto
EX5                                qc
EvaluationMSc-MLS_Spring2016_Seq-GenomeII.pdf  rl.csv
GenDB_Pseudomonas_protegens_S5.gbk           sag.Rproj
IGV                                             seq.fq
LSA_Phy5G7RSa0gJ_L7_R1_002.fastq.gz          student_reports
Pseud_protegens_S5_final.fasta                workig_documents
Sequence-a-genome
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ cp Topics_Students_Assistants_SAGEII.xlsx ..
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_spring$ cd ..
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ ls
2015_autumn      2016_spring
2016_autumn      Topics_Students_Assistants_SAGEII.xlsx
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ mv Topics_Students_Assistants_SAGEII.xlsx 2016_autumn/
kjaron@lazuli:/Volumes/dump/documents/teaching/sag$ cd 2016_autumn/
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_autumn$ rm Topics_Students_Assistants_SAGEII.xlsx
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_autumn$ ls
1_unix_local.Rmd      1_unix_local.pdf      3_qc_and_trimming.Rmd  unix_presentation.Rmd
1_unix_local.html     2_unix_cluster.Rmd    3_qc_presentation.Rmd  unix_presentation.html
kjaron@lazuli:/Volumes/dump/documents/teaching/sag/2016_autumn$
```


command -<parameters> <arguments>

Examples:

ls -lah #list long, all, human readable

ls -la .. #list in parent directory

cp -r <what_dir> <where> # recursive

rm -rf <what_dir> # -||- , force

careful with this one...

OK, try it!

OK, try it!

```
# Auto Completion by <tab>  
cd /<tab><tab>    # lists all in root  
cd ~/k<tab><tab>  # lists all in home
```

MyUnix → Sequence a Genome II → 1_unix_local.pdf

```
# Auto Completion by <tab>
cd /<tab><tab>      # lists all in root
cd ~/k<tab><tab>    # lists all in home
# Command history
<arrow_up>         # last excuted command
<Ctrl+R>           # full-text search
```

What about that toolbox?

How to check a file??

Notepad?



not well suited for big files

Office?



Use at least read-only mode

Look at a plain-text file

command <text_file>

head

tail

cat # catenate

less # > more; text reader

wc # word count

tr # transform

grep # global regular expression print

IO streams - a way how to build pipes!

```
command <text_file>  
      == cat <text_file> | command
```

head

tail

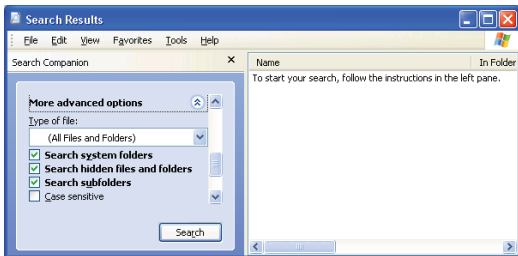
cat # catenate

less # > more

wc # word count

tr # transform

grep # global regular expression print



IO streams!

```
ls | wc -w
```

Streams in the service of genomics!

```
grep ">" seq.fasta | wc -l
```

input file

```
grep -v ">" seq.fasta | \  
tr -d '\n' | wc -l
```

```
>seq1
```

```
CGATCGTCGTAGCTACGAT
```

```
>seq2
```

```
ACCGATCAAACCGTCGTAA
```

```
grep -v ">"
```

```
grep -v ">" seq.fasta | \  
tr -d "\n" | wc -l
```

```
CGATCGTCGTAGCTACGAT  
ACCGATCAAACCGTCGTAA
```

```
tr -d "\n"
```

```
grep -v ">" seq.fasta | \  
tr -d "\n" | wc -l
```

CGATCGTCGTAGCTACGATACCGATCAAACCGTCGTAA

```
wc -l
```

```
grep -v ">" seq.fasta | \  
tr -d "\n" | wc -l
```

Stream redirection overview

```
grep ">" seq.fasta          # print
grep ">" seq.fasta | wc      # pipe
grep ">" seq.fasta > file    # write
grep ">" seq.fasta >> file   # append
```


Your turn

command <text_file>

head

tail

cat # catenate

less # > more; text reader

wc # word count

tr # transform

grep # global regular expression print