Text editors

Bioinformatics Applications (PLPTH813)

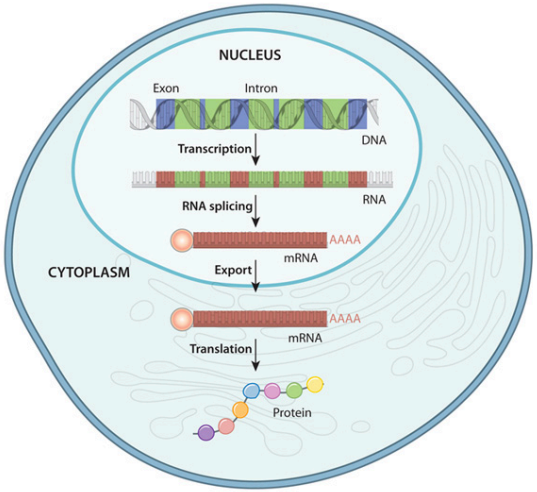
Sanzhen Liu

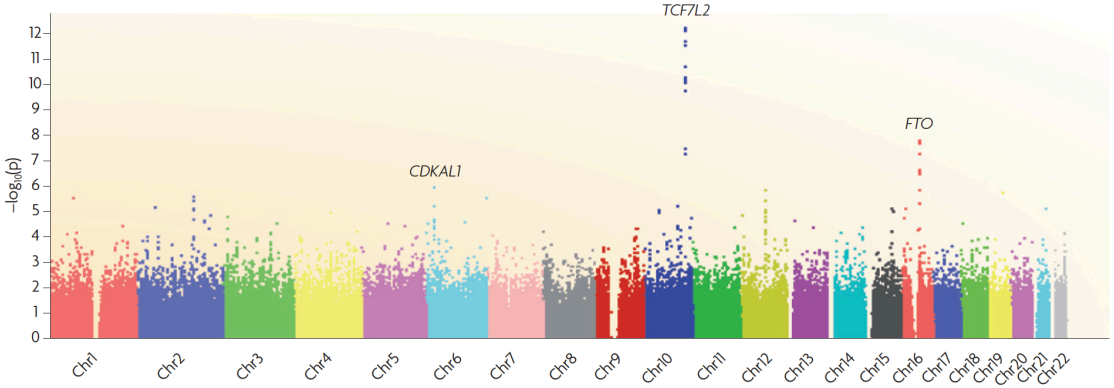
1/28/2021

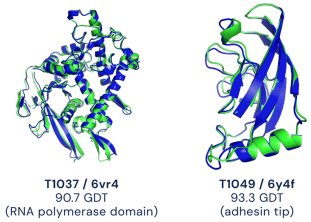
Review

```
   1 ccctgaggct tttcggagcg agctcctcaa atcgcatcca gattttcggg tccgagggaa
  61 ggaggaccct gcgaaagctg cgacgactat cttcccctgg ggccatggac tcggacgcca
 121 gcctggtgtc cagccgcccg tcgtcgccag agcccgatga ccttttttctg ccggcccgga
 181 gtaagggcag cagcggcagc gccttcactg ggggcaccgt gtcctcgtcc accccgagtg
 241 actgcccgcc ggagctgagc gccgagctgc gcggcgctat gggctctgcg ggcgcgcatc
 301 ctggggacaa gctaggaggc agtggcttca agtcgtcctc gtccagcacc tcgtcgtcta
 361 cgtcgtcggc ggctgcgtcg tccaccaaga aggacaagaa gcaaatgaca gagccggagc
 421 tgcagcagct gcgtctcaag atcaacagcc gcgagcgcaa gcgcatgcac gacctcaaca
 481 tcgccatgga tggcctccgc gaggtcatgc cgtacgcaca cggcccttcg gtgcgcaagc
 541 tttccaagat cgccacgctg ctgctggcgc gcaactacat cctcatgctc accaactcgc
 601 tggaggagat gaagcgactg gtgagcgaga tctacggggg ccaccacgct ggcttccacc
 661 cgtcggcctg cggcggcctg gcgcactccg cgcccctgcc cgccgccacc gcgcacccgg
 721 cagcagcagc gcacgccgca catcaccccg cggtgcacca ccccatcctg ccgcccgccg
 781 ccgcagcggc tgctgccgcc gctgcagccg cggctgtgtc cagcgcctct ctgcccggat
 841 ccgggctgcc gtcggtcggc tccatccgtc caccgcacgg cctactcaag tctccgtctg
 901 ctgccgcggc cgccccgctg gggggcgggg gcggcggcag tggggcgagc gggggcttcc
 961 agcactgggg cggcatgccc tgcccctgca gcatgtgcca ggtgccgccg ccgcaccacc
1021 acgtgtcggc tatgggcgcc ggcagcctgc cgcgcctcac ctccgacgcc aagtgagccg
1081 actggcgccg gcgcgttctg gcgacagggg agccaggggc cgcggggaag cgaggactgg
1141 cctgcgctgg gctcgggagc tctgtcgcga ggaggggcgc aggaccatgg actgggggtg
1201 gggcatggtg gggattccag catctgcgaa cccaagcaat gggggcgccc acagagcagt
1261 ggggagtgag gggatgttct ctccgggacc tgatcgagcg ctgtctggct ttaacctgag
1321 ctggtccagt agacatcgtt ttatgaaaag gtaccgctgt gtgcattcct cactagaact
```

2018







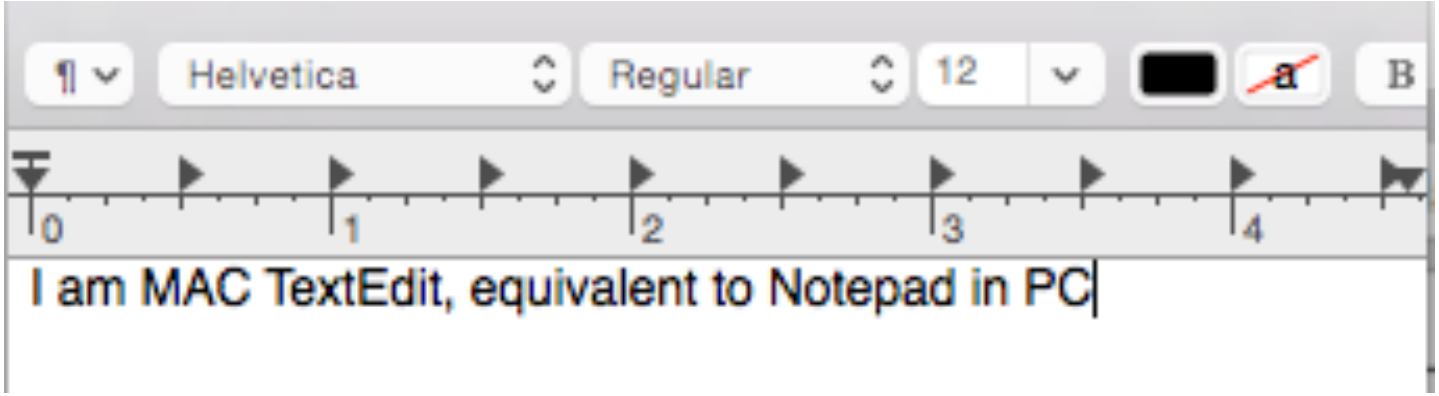
Experimental result
Computational prediction

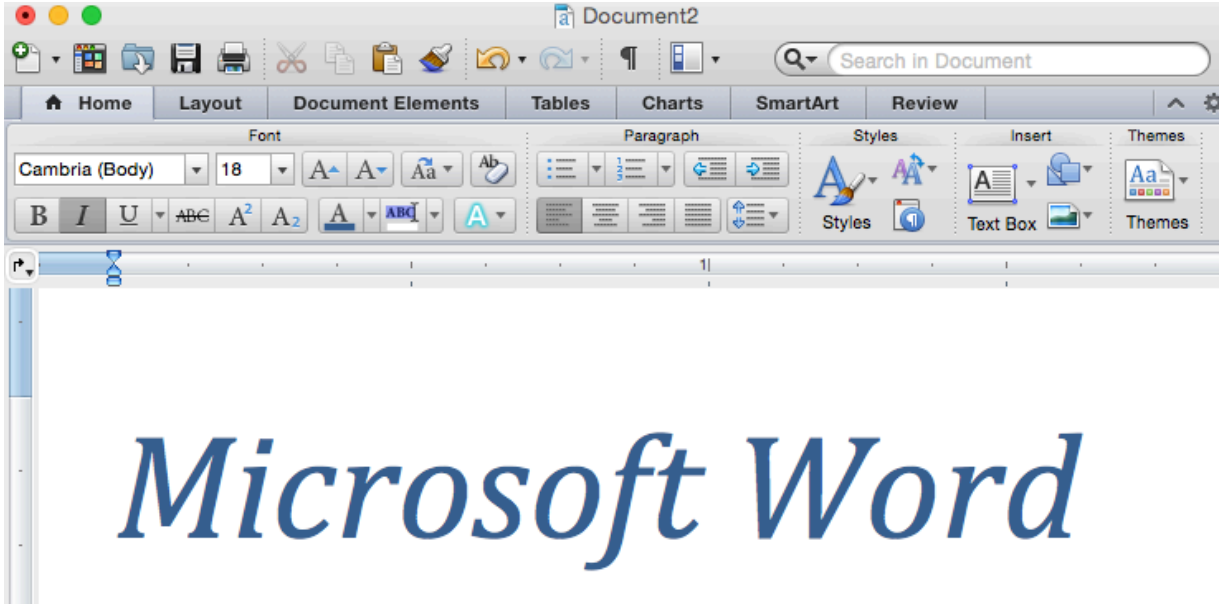
Outline

Goal: to understand how to organize data in a proper format and efficiently input and edit data.

- Formats of text data files
- Excel to generate a text file and tips in Excel
- BBEdit (Mac) Notepad++ (PC): text editor
- Regular expression
- *vi*: another text editor

Software for text editing





Text file – flat file

- **Flat file**

1. Simple format, consisting of readable characters

- ASCII (American Standard Code for Information Interchange, 128 characters)
- No rich format control (e.g. bold or Italics, etc)

2. Easy for sharing

- **The organization of data in a text file**

1. Most popular formats for tabular data: space or tab separated data file (.txt) and comma-separated values (.csv)
2. Most popular format for DNA/protein sequences: FASTA format (.fa, .fas, .fasta)

File formats

- Tab separated file (.txt)

```
name	age	>30?	gender
Josh	23	FALSE	male
Rose	35	TRUE	female
```

- Comma-separated file (.csv)

```
name,age,>30?,gender
Josh,23,FALSE,male
Rose,35,TRUE,female
```

- FASTA (.fa, .fas, .fasta)

```
>Aa1
CCATCTCATCCCTGCGTGTCTCCGACTCAG
>Aa2
CTGAGTCGGAGACACGCAGGGATGAGATGGTT
```

Text editors

- Notepad or Notepad++ (PC)
- TextEdit (Mac)
- TextWrangler (Mac)
- vi (Unix and Linux)
- Emacs
- Atom

- Word (PC and Mac): save as …
- Excel (PC and Mac): save as …
- etc

Newline – end of line (EOL)

Two types of EOL: line feed (LF) and carriage return (CR):

LF: \n
CR: \r

- LF: Unix, Linux, OS X
- CR: Mac OS up to version 9 and OS-9
- CR+LF: Microsoft Windows

http://en.wikipedia.org/wiki/Newline

Outline

- Formats of text data files
- Excel to generate a text file and tips in Excel
- BBEdit (Mac) Notepad++ (PC): text editor
- Regular expression
- *vi*: another text editor

Excel to generate a text file

name	age
Josh	23
Rose	35
Jone	18
Molly	21
Lisa	36

- copy and paste to a text editor (e.g. vi)
- save as …

Excel function - examples

Q1: =**AVERAGE**(B3:B7)

Q2: =**COUNTIF**(B3:B7, ">20")

Q3: =B3>30

Q4: search information at Table 2

1. define the Table 2: gender (control + l)

2. =**VLOOKUP**(A3, gender, 2, FALSE)

	A	B	C	D
1	Table 1			
2	**name**	**age**	**>30?**	**gender**
3	Josh	23		
4	Rose	35	Q3	Q4
5	Jone	18		
6	Molly	21		
7	Lisa	36		
8				
9	Table 2			
10	**name**	**gender**		
11	Josh	male		
12	Rose	female		
13	Jone	male		
14	Molly	female		
15	Lisa	female		
16				
17	Question:			
18	average age	Q1		
19	# of persons >20	Q2		

Table 1			
name	**age**	**>30?**	**gender**
Josh	23	FALSE	male
Rose	35	TRUE	female
Jone	18	FALSE	male
Molly	21	FALSE	female
Lisa	36	TRUE	female

Table 2	
name	**gender**
Josh	male
Rose	female
Jone	male
Molly	female
Lisa	female

Question:	
average age	26.6
# of persons >20	4

Useful functions in Excel

- max/min/average/sum
- len/left/right
- if/countif
- >, <, =
- & (concatenate)
- vlookup

- LEFT function Returns the leftmost characters from a text value

	A	B
1	this class is boring	=LEFT(A1, 14)&"great!"

Functions can be combined.

Problem 1

Replace the words containing “genome” with “XXX” regardless of letter case.

Genome old and new charted the emergence of agriculture. Contemporary Europeans carry DNA inherited from light-skinned, brown-eyed farmers who migrated from the Middle East beginning 7,000–8,000 years ago, in addition to more-ancient ancestry. The achievements of these early farmers — domestication of crops such as wheat and barley — are also being understood through genome sequencing.

Which software and what trick will you use?

Problem 2

Replace the words containing “genome” with “XXX” regardless of letter case (e.g., Genome = genome = genomes = Genomes).

Genomes old and new charted the emergence of agriculture. Contemporary Europeans carry DNA inherited from light-skinned, brown-eyed farmers who migrated from the Middle East beginning 7,000–8,000 years ago, in addition to more-ancient ancestry. The achievements of these early farmers — domestication of crops such as wheat and barley — are also being understood through genome sequencing. In July, a consortium reported a draft copy of the gargantuan wheat genome, which contains 124,000 genes and 17 billion nucleotides. Another group released the genomes of 3,000 rice varieties. - Science 2014

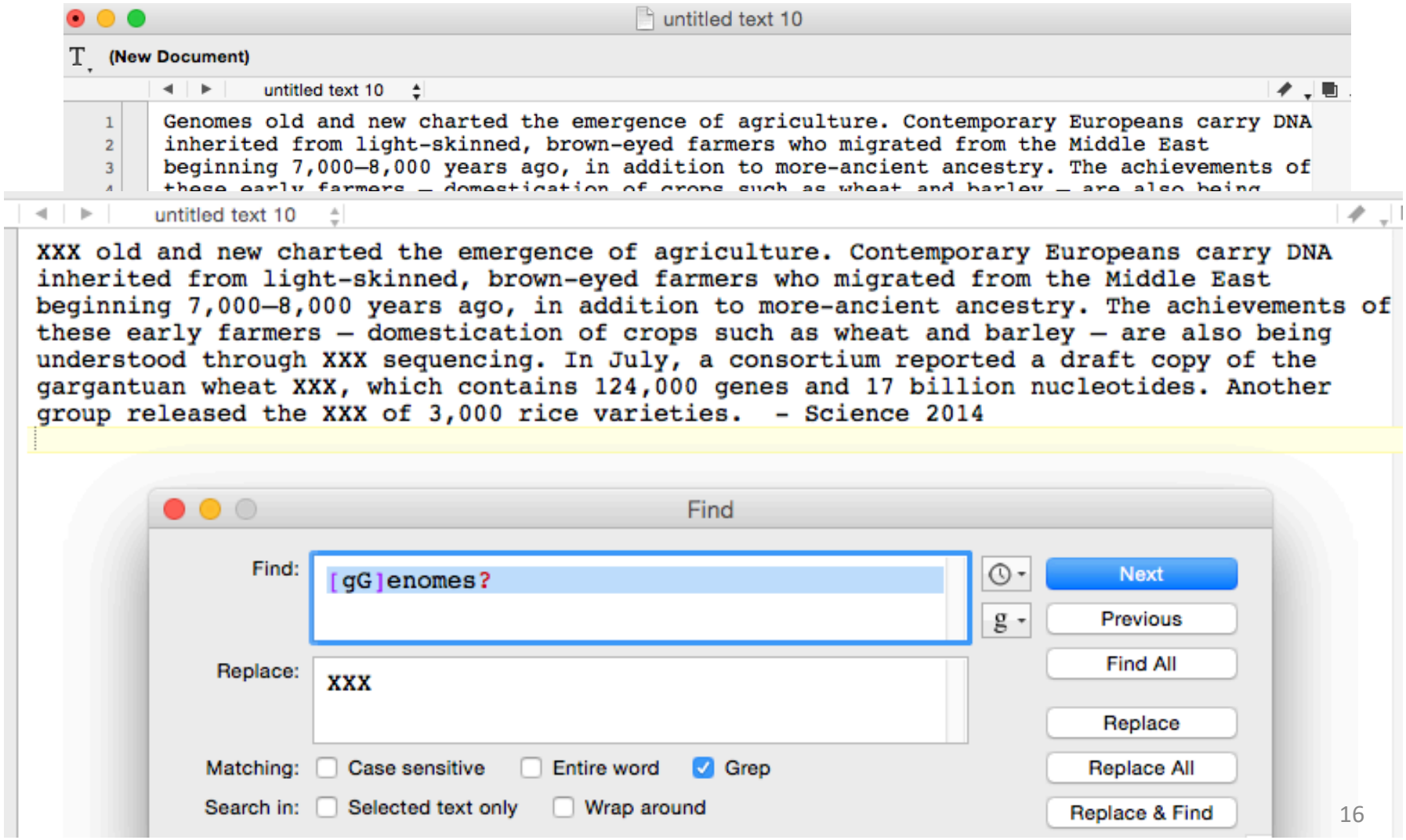
Which software and what trick will you use?

Outline

- Formats of text data files
- Excel to generate a text file and tips in Excel
- BBEdit (Mac) Notepad++ (PC): text editor
- Regular expression
- *vi*: another text editor

BBEdit

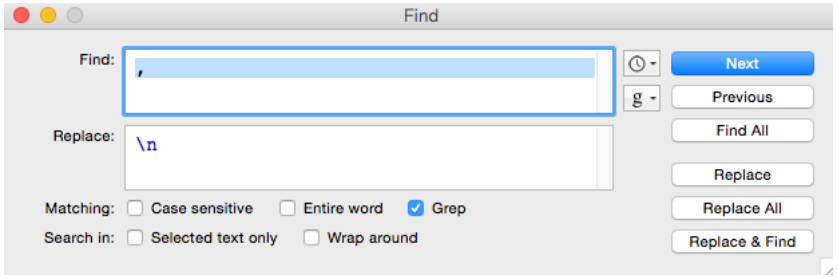
A flexible text editor with powerful functions of searching and editing.



BBEdit – more examples

Class participation 15%, Homework 15%, Midterm Exam 20%, Project 20%, Final Exam 30%

Class participation 15%
Homework 15%
Midterm Exam 20%
Project 20%
Final Exam 30%

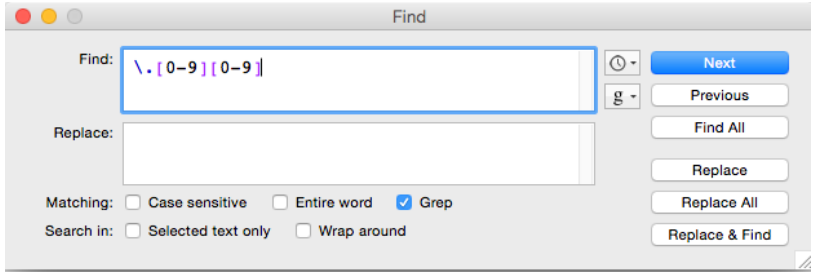


\n: end of line character (line separator)

Class participation 15.01%, Homework 15.03%, Midterm Exam 20.10%, Project 20.10%, Final Exam 30.01%

\.[0-9][0-9]

\.: the character of "."
. : any character



Regular expression

- **Regular expression** (regex or regexp) is a sequence of characters that forms a search pattern.

Search Genome or genomes:

[gG]enomes?

[] : a single character of a range indicated in the square brackets
?: no matches or just one match

More regex characters

Wildcard

\w : letters, numbers, and _
. : any character except \n \r
? : no matches or just one match
+ : one or more matches
* : any character
\d : numerical digits
\t : Tab
\r : return; also used as the generic end-of-line in BBEdit
\n : line-feed character; also used as the generic end-of-line in Notepad++
\s : space, tab, or end of line
[A-Z]: a single character of the ranges indicated in square brackets
[^A-Z]: a single character including all characters not in the brackets.
Note that this will include \n unless otherwise specified.

^ : match the start of the line, i.e., the position before the first character
$: match the last position before the end-of-line character

Regular expression (I)

\t : a tab character

\r (or \n): end-of-line

Potato,apple,orange

Regexp	Replace
,	\t

Potato apple orange

Regexp	Replace
\t	\n

Potato

apple

orange

Regular expression (II)

- **^** beginnings
- **$** endings

Potato

apple

orange

Regexp	Replace
^	-

-Potato

-apple

-orange

Potato

apple

orange

Regexp	Replace
$	s

Potatos

apples

oranges

Regular expression (III)

- **\w** a ***w***ord character, including letters, numbers and underscore
- **\d** : numerical ***d***igits

I have 5 apples.

Regexp	Replace
^\w	We

We have 5 apples.

I have 5 apples.

Regexp	Replace
\d	a lot of

I have a lot of apples.

Regular expression (IV)

+ : 1 or more previous regular expression

? : 0 or 1 previous regular expression

. : any character except \n \r

potato,apple,orange

Regexp	Replace
p+	-

-otato,a-le,orange

potato,apple,orange

Regexp	Replace
p?	-

--o-t-a-t-o-,-a---l-e-,-o-r-a-n-g-e

potato,apple,orange

Regexp	Replace
p.	-

-tato,a-le,orange

Regular expression (V)

[A-Z] : any single letter

NspI

[AG]CATG[CT]

select 2012, 2013, 2014 201[2-4]

{} : specify a range of numbers to repeat the match of the immediately preceding character.

Poly A (12 A in a row) A{12}

Poly A (10-12 A in a row) A{10,12}

Poly A (>=10 A in a row) A{10,}

Regular expression (VI)

|

|: or

hello|hi Match either hello or hi in the text

Polling Questions

K-[Ss]tate|KSU

^[AGCT]+

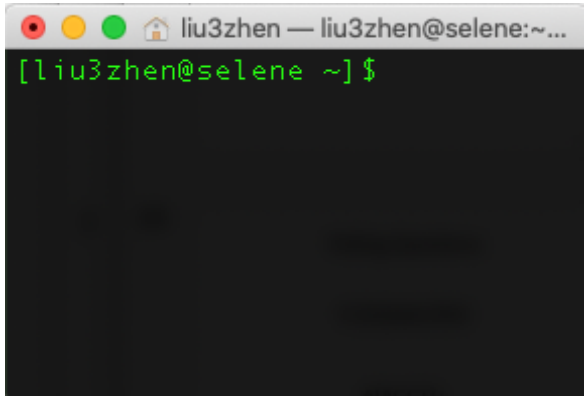
Regular expression

- Regular expression is for pattern searches
- It is commonly employed in programming languages
- The rules vary depending on the specific implementation (or programming languages or versions) in use.

Does Google provide search with regular expressions?

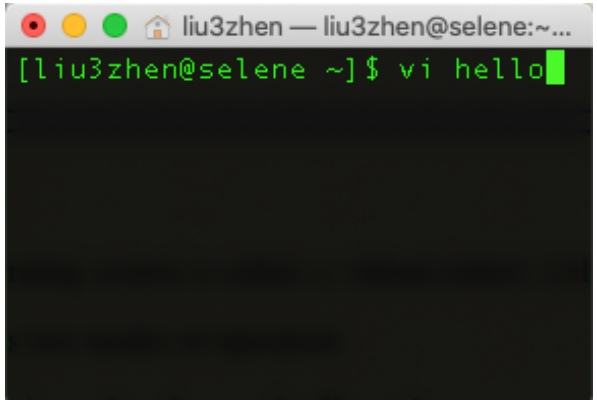
"genome * sequencing"

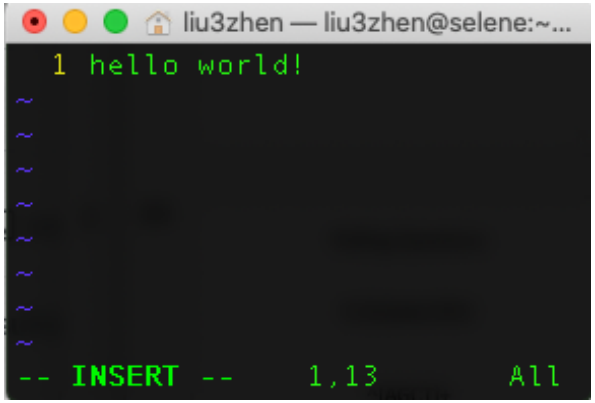
vi is a text editor created for the Unix operating system.
- fast and powerful



In a Unix/Linux system, any “words” typed are commands

What can we do if we need to type data or codes?





vi is a command to execute a program

vi

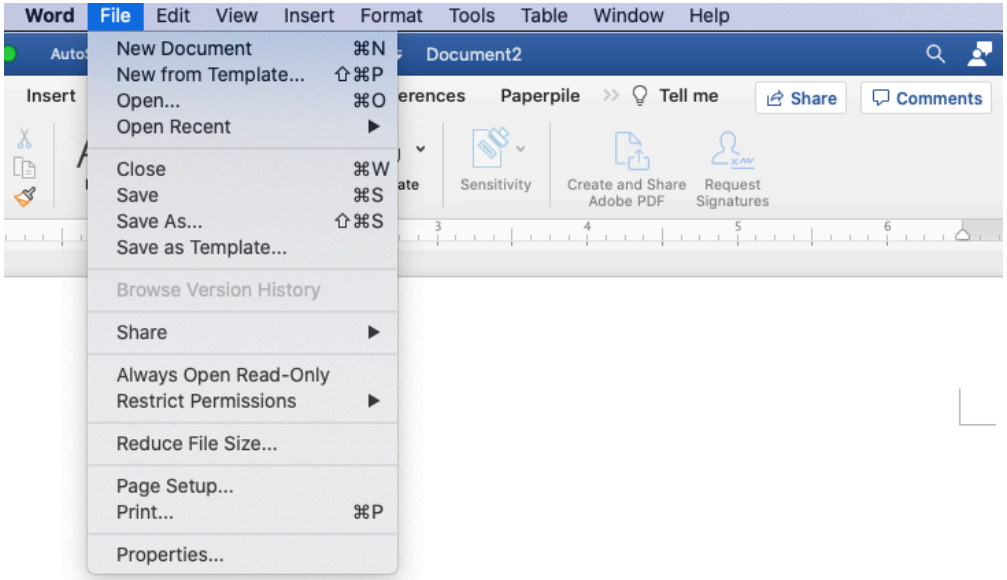
- *vi* has two modes:

  1. insert mode (edit as other text editors)

  2. command mode (commands that control the edit session).

  switch modes by using “i” and “ESC” key

Your keyboard controls “everything”.



Actions in command mode

Search: to search content using “/”

- /<text or regular expression>

Delete contents for example by lines

Copy and **paste**

Command list

ZZ	Exit, saving changes	t<x>	Up to <x> forward
Q	Enter ex mode	T<x>	Back up to <x>
<ESC>	End of insert	<x>|	Go to column <x>
:<cmd>	Execute ex command	w,W	Forward one word
:!<cmd>	Shell command	b,B	Back one word
^g	Show filename/size	e,E	End of word
^f	Forward one screen	^h	Erase last character
^b	Back one screen	^w	Erase last word
^d	Forward half screen	^?	Interrupt
^u	Backward half screen	~	Toggle character case
<x>G	Go to line <x>	a	Append after
/<x>	Search forward for <x>	i,I	Insert before
?<x>	Search backward for <x>	A	Append at end of line
n	Repeat last search	o	Open line below
N	Reverse last search	O	Open line above
]]	Next section/function	r	Replace character
[[	Previous section/function	R	Replace characters
%	Find matching () { or }	d	Delete
^l	Redraw screen	dd	Delete line
^r	Refresh screen	c	Change
z<CR>	Current line at top	y	Yank lines to buffer
z-	Current line at bottom	C	Change rest of line
^e	Scroll down one line	D	Delete rest of line
^y	Scroll up one line	s	Substitute character
``	Previous context	S	Substitute lines
H	Home window line	J	Join lines
L	Last window line	x	Delete after
M	Middle window line	X	Delete before
+	Next line	Y	Yank current line
hjkl	Cursor movement:	p	Put back lines
	left/down/up/right	P	Put before
0	Beginning of line	<<	Shift line left
$	End of line	>>	Shift line right
f<x>	Find <x> forward	u	Undo last change
F<x>	Find <x> backward	U	Restore current line

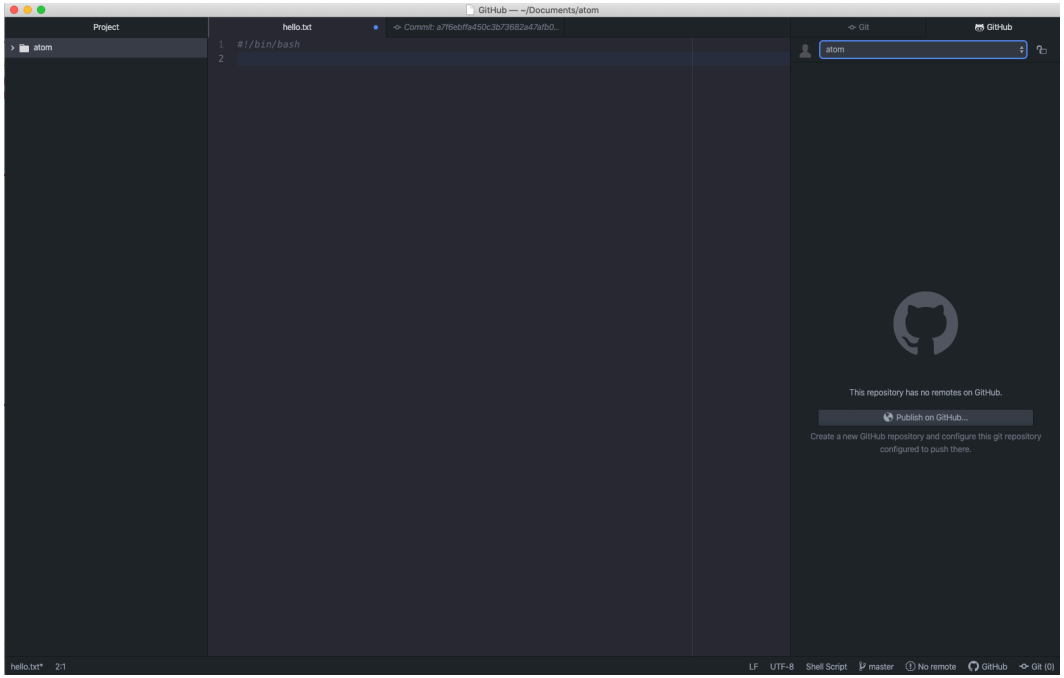
ex command

q	Quit
q!	Quit, discard changes
r <f>	Read in file <f>
sh	Invoke shell
vi	Vi mode
wq	Write and quit
w <f>	Write file <f>
w! <f>	Overwrite file <f>

https://kb.iu.edu/d/afdc

Atom (atom.io)

1. A modern desktop text editor
2. Version control through git and Github
3. …



Goal of today's lab

- Familiar to Excel functions
- Try *vi* at Beocat
- Practice using regular expression in BBEdit

for PC, download the software "putty" and "notepad++"
32 or 64 bits

for mac, download "BBEdit"