Tips and Tricks – December 8th, 2022

Top 7 Tips for Bioinformatics Success

ICG | INSTITUTE FOR COMPARATIVE GENOMICS

Dalhousie University is located in Mi’kma’ki, the ancestral and unceded territory of the Mi’kmaq. We are all Treaty people.

Most of bioinformatics involves simple text files

-input/output

-separated by tabs, commas, semicolons, spaces

Most bioinformatics programs are poorly described and often die on the vine

-created by grad students and postdocs

-funding life cycle

Most bioinformatics problems can be solved multiple ways

-need to string together multiple programs

Every bioinformatics problem is unique/the same

-variations on a theme

We deal with big datasets

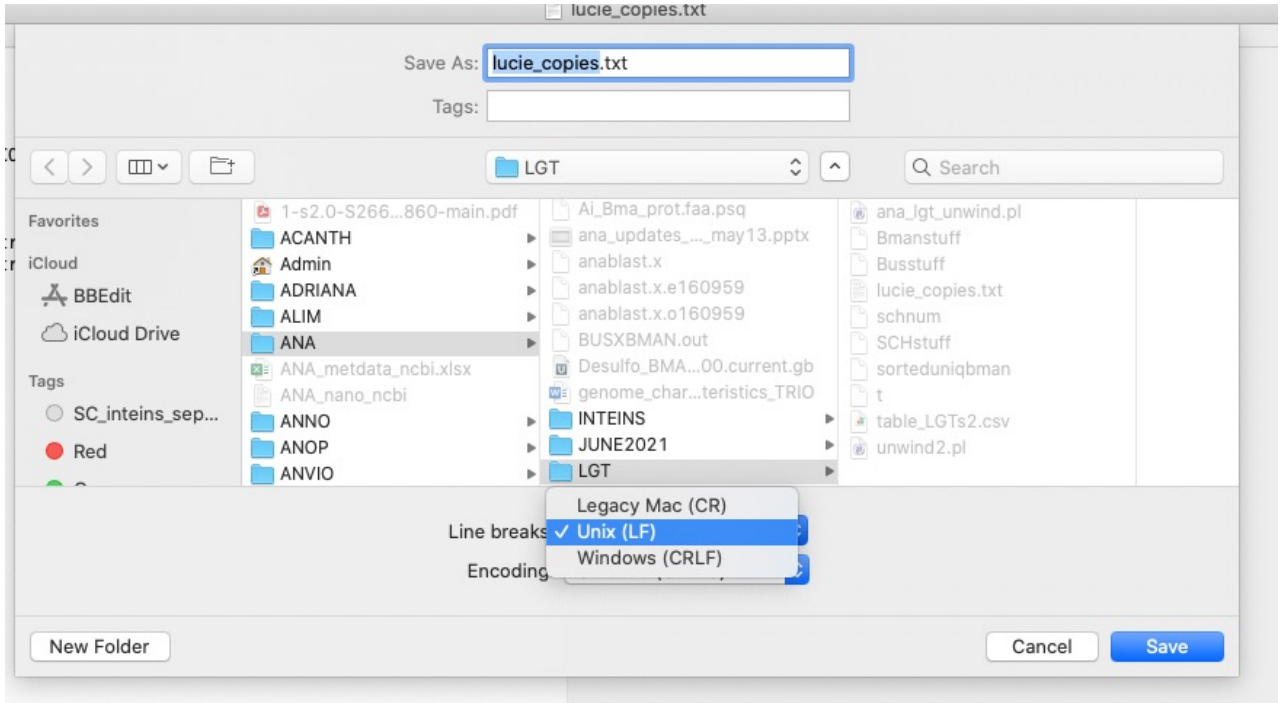
1. Get a proper simple text app/program

Most personal computers have programs for looking at text – Word, texteditor

- -these programs introduce hidden characters
  - -writing scripts
  - -opening output and modifying it

- -these hidden characters make bioinformatics program do crazy things
  - -won't work
  - -truncate output

- -some options
  - -BBEdit (be careful what format you save as)
  - -pico (UNIX/linux)
  - -nano (UNIX/linux)
  - -don't bother with Vi


lucie_copies.txt
Save As: lucie_copies.txt
Tags:
LGT
Search
Favorites
iCloud
BBEdit
iCloud Drive
Tags
SC_inteins_sep...
Red
1-s2.0-S266...860-main.pdf
ACANTH
Admin
ADRIANA
ALIM
ANA
ANA_metdata_ncbi.xlsx
ANA_nano_ncbi
ANNO
ANOP
ANVIO
Ai_Bma_prot.faa.psq
ana_updates_..._may13.pptx
anablast.x
anablast.x.e160959
anablast.x.o160959
BUSXBMAN.out
Desulfo_BMA...00.current.gb
genome_char...teristics_TRIO
INTEINS
JUNE2021
LGT
ana_lgt_unwind.pl
Bmanstuff
Busstuff
lucie_copies.txt
schnum
SCHstuff
sorteduniqbman
t
table_LGTs2.csv
unwind2.pl
Line breaks
Encoding
Legacy Mac (CR)
Unix (LF)
Windows (CRLF)
New Folder
Cancel
Save

```
BMAN


14562.t1 2.6e-50 14563.t1
FAD-DEPENDENT OXIDOREDUCTASE DOMAIN-CONTAINING PROTEIN 1        FAD-DEPENDENT OXIDOREDUCTASE DOMAIN-CONTAINING PROTEIN 1
KS


g6567.t1 keep separate, different intron arrangement
g6568.t1 keep separate, different intron arrangement
g6569.t1 not related


g7818.t1 KS

g7819.t1 KS

g7820.t1 KS
g7821.t1 KS

g13853.t1 KS

g13854.t1 KS


g13882.t1 KS

g13883.t1 KS



g10752.t1 KS

g10753.t1 KS

g10754.t1 KS

g10755.t1 KS

|



Busselton

g24225.t1
g24227.t1

g25577.t1

g4830.t1
g4831.t1


g6069.t1

g6070.t1
g6071.t1


g8076.t1
g8077.t1
g8078.t1
```

```
/home/curtisba@perun4> more lucie_copies_example.txt
g8078.t11 KS relatedte, different intron arrangementIN 1                FAD-DEPENDENT OXIDOREDUCTASE DOMAIN-CONTAINING PROTEIN 1
/home/curtisba@perun4>
```

2. Learn a programming language

-almost all input/output is simple text format

-need to manipulate/parse the text

-extract what you need from the large volume

-change format/content for additional programs

-PERL

-Python

-R

-?

3. Keep a personal repository of scripts for simple things you do frequently

-folder on your local computer or perun

-copy, change and paste

-or find one that has something similar

-github

-ICG has one

https://github.com/Dalhousie-ICG/icg-shared-scripts

GitHub - Dalhousie-ICG/icg-shared-scripts: General repository for members of ICG to share, store and find useful scripts

-perun wiki

https://perun.biochem.dal.ca/user-wiki/doku.php?id=start

novigit update README

- perun_scripts
- Extract_contigs_left2remains.py
- LICENSE
- NCBI_genome_download.py
- NCBI_genome_download_using_cen...
- README.md
- Remove_short_contigs_fasta_files_i...
- TreeFINISHER_ete3_v1-1.py
- aa_recoding.pl
- add_intergenic_space_features.py
- add_intron_features.py
- add_orfs_to_intergenic_regions.py
- alignment_pruner.pl
- calcCARSC.py
- calcNARSC.py
- colorFastq.pl
- compare_assemblies.py
- concatenateRenameAlignment.pl
- count_tripartitions.py
- fastaNamesSizes.pl
- fix_genes_with_false_introns.py
- getCodingDensity.pl
- getIntergenicSpace.pl
- getRecords.pl
- mafft_and_trimal.sh
- parseSplitcounts.pl
- parse_tripartition_counts.py

- run_blastn_vs_nt.sh
- run_braker2.sh
- run_busco3.sh
- run_busco5_augustus.sh
- run_busco5_metaeuk.sh
- run_bwa_mem.sh
- run_bwa_mem2.sh
- run_bwa_meme.sh
- run_evm.sh
- run_guppy_gpu.sh
- run_hisat2_dna.sh
- run_hisat2_rnaseq.sh
- run_medaka.sh
- run_ngmlr.sh
- run_pasa.sh
- run_pilon.sh
- run_ploidyNGS.sh
- run_repeatmasker.sh
- run_repeatmasker_tmpdir.sh
- run_trinity_genome_guided.sh

4. Think beyond perun

-other resources

-Digital Research Alliance of Canada (Compute Canada)

-Your own computer (MACs)

-anaconda

-Online resources

-Galaxy

-local linux box

-google (you are not alone/unique)

5. Don’t settle for default

-explore the options

-each genome is unique and presents unique problems

-most of the defaults based on model systems

-read the manual !!!!!!!

-github

-on the command line -h -help --h –help

blast

-task (blastn, blastn-short, dc-megablast (discontinuous), megablast)

-dust

-num_alignments, num_descriptions

-perc_identity

6. Explore the world of emboss

-a large suite of simple but useful programs for manipulating sequence data

- -infoseq
  - -get basic information (length, gc%
- -extractseq
  - -extract sequences regions
- -seqretsplit
  - -split multifasta files
- -seqret
  - -change formats
    - -over 50 input/output formats
    - -not just sequence (phylogenetics)
- -transeq
  - -translate nucleotide sequences into protein sequences
- -sizeseq
  - -order multifasta files based on length of individual fastas

7. **Simple but powerful unix commands**

-many have great options

ls -lt

-long form, most recent

grep --color

grep –n (line number of appearance)

wc -l (word count, just lines)

sort (lots of options -r,-n,-k)

uniq –c (count number of occurrences collapsed into uniq)

uniq –d (display duplicated occurrences)

cut –f1 (cut simple text by column)

find (search function in directories)

-string them together using |

-grep blah file |cut –f1,2 |sort |uniq –c |sort -n

Questions?

A recording of this session will be available from
https://perun.biochem.dal.ca/downloads/icgvideo/TAT/

We are looking for new presenters in 2023

-don't have to be an expert

-new software

-new procedure

-contact Wanda Danilchuk

ICG INSTITUTE FOR COMPARATIVE GENOMICS