HSATII Program Notes

Need to install scipy plug in for program to run –

All files and scripts available from https://github.com/iouthwaite/HSAT\_project

To find hits

1. make folder called “Genomes” – in same folder as script – can have multiple files in here as input, all should be from FASTA format saved as .txt files
2. open terminal – enter a python virtual environment if you don’t want to download programs onto your machine permanently (recommended)
3. To create a virtual environment, follow the guide: http://docs.python-guide.org/en/latest/dev/virtualenvs/
4. The last command (depends upon which directory you are in, but typically “. bin/activate”) will put you in the virtual environment you have created – this allows you to download/use programs and plug ins without affecting whole computer
5. install numpy: pip install scipy – (this is a math/stats plugin – used for beta distribution)
6. cd to folder containing Genomes, hitFinder & locusFinder in same folder, while in the virtualenv “(venv)” will be in the command line prompt.
7. can drag any files into Genomes folder – need these files in FASTA – a text file – these are individual chromosome files (ex: agm\_2)
8. Get hits: python HSAT\_hitFinder.py (can also indicate starting and ending percentages of text file – (i.e. you know pericentric regions is in middle of file) : -s 30 –e 70 (this goes after python command) – \*\* takes about an hour per chromosome file

* will get a list of hits, but this also puts this into a results file in the same folder (this will overwrite!)
* this is organized by A1, A2, B for each “hit” in chromosome file & also indicates bp location for all those hits
* 6= 6 mistakes over 24 basepairs (24-mer oligos)

IMPORTANT: A results file is automatically generated from each run; copy results file to new file (or rename) so it doesn’t overwrite the previous results!

1. python HSAT\_locusFinder.py – NOTE: need to curate “loci” by eye from text file first – copy & paste the “loci” into another text file – the goal of this is to model, or train the locusFinder to “find” HSATII loci – these loci need to be bracketed and sub-bracketed, commas and spaces in between brackets – this formatting is critical (example of this file is called HSATLoci.txt)

python HSAT\_locusFinder.py HSATLoci.txt HSATresults.txt

-c = combined results – this will allow you to combine A1, A2, B results together – does not distinguish between these

-h = minimum number of hits to define a locus (default = 10)

-u = kernel size around current hit distance for considering whether or not a group of hits is statistically significant

output of this goes to an excel file

Example flow through (for working in a project folder from the desktop)

**SETTING UP**

1. open a terminal

2. cd Desktop (if you want to do everything from a folder in your desktop)

2.5 (make a github account)

3. git pull <https://github.com/iouthwaite/HSAT_project.git>

(enter your github password etc. at the prompts)

4. Activate/make your virtualenv, you can re-activate one you have already made

learn how to here: <http://docs.python-guide.org/en/latest/dev/virtualenvs/>

5. mkdir myprojectfolder

6. Move files from the HSAT\_project folder to “myprojectfolder” by dragging them using the mac mouse

7. cd myprojectfolder (enter your new folder)

8. ls (Make sure that you have dragged HSAT\_hitFinder.py and HSAT\_lociFinder.py over, they will appear in the terminal window list if you have)

9. mkdir Genomes

10. drag all genome files (.txt FASTA format) into the Genomes folder that you want to analyze

11. your terminal line prompt should look something like this:

(venv)My-MacBook-Pro:myprojectfolder Myname$

**RUNNING HITFINDER**

12. type “python3 HSAT\_hitFinder.py”

or

“python3 HSAT\_hitFinder.py –s 20 –e 80”

if you want to look at the 20th->80th percentile if your genome files etc.

the program will print out the results in the command line AND create a results file. You should rename this file since subsequent program runs will write over it. This program can also take a while to run – about 1.5 hours per human chromosome, give or take.

13. Using the hitfinder results, make a training dataset of what you might consider to be biologically relevant loci (ex: >50 hits that are relatively close together on a genomic scale, not more then a couple kpb between hits). File should be saved as .txt

Note: MUST have brackets around groups of hits, and brackets around all the groups of hits. Only have spaces after commas. Example:

[[129368808, 129368858, 129368881, 129368908, 129368957, 129369083, 129369133, 129369231, 129369281, 129369304, 129369354, 129369377, 129369401, 129369424, 129369447, 129369550, 129369623, 129369696, 129369719, 129369842, 129369892, 129370048, 129370124, 129370229, 129370279, 129370787, 129370837, 129370910, 129370936, 129370960, 129371009],[42717672, 42717687, 42717951, 42717991, 42718022, 42718103, 42718113, 42718204, 42718253, 42718339, 42718557, 42718731, 42729342, 42729465, 42729537, 42729563, 42729610, 42729636, 42729712, 42729775, 42729826, 42729849, 42729876, 42729902, 42729978, 42730028, 42730107, 42730130, 42731614, 42731687, 42731740, 42731766, 42731790, 42731816, 42731866, 42731889, 42731939, 42731965, 42732015, 42732087, 42732186, 42732259, 42732312, 42732335, 42732362, 42732385, 42732584, 42732610, 42732822, 42732872, 42732895, 42732945, 42732968, 42733018, 42733084, 42733107, 42733157, 42733362]]

**RUNNING HSAT LOCIFINDER**

14. if “myawesomeloci.txt” is the name of the file you just made

and “results” is the name of the earlier hitFinder results file

the command is:

python3 HSAT\_lociFinder.py myawesomeloci.txt results.txt

you can add –h –u or –c afterwards

-c: combined loci from all 3 subfamilies

-h 34: minimum of 34 hits for the program to call something a loci

-u: a measure of stringency in the program; it is the +/- range of neighboring hits the program looks to when considering whether the distances between a group of hits is statistically probable. A larger number means the program may gloss over some less likely hits (larger loci), and a smaller number means the program is going to be less forgiving (smaller loci)

python3 HSAT\_locusfinder.py myawesomeloci.txt results.txt –c –h 50 –u 8

the results of this file will be written to an excel spreadsheet to make it easy to do statistical tests etc.

to make plots (one way)

PlotGenerator is a program that can help you use plotly

You will need to install plotly while inside your virtualenviroment:

pip install plotly

and register for a plotly account, and add your plotly username and password to the PlotGenerator file (near the top:

py.sign\_in('', '')

should be replaced by

py.sign\_in('myname','mypassword')

You will also need to create a file called “Genome Lengths” that contains the lengths of all the chromosomes etc. that you’re using. The names should match up to the excel file that the lociFinder generated. An example file with this name is included in the github pulldown.

Rename the excel file you want to use to make the plot:

“HSATfileforplotproduction\_c2.xls”

OR change the name in the PlotGenerator file to what you’d like

Then,

Python3 PlotGenerator

And your plot will open up in an online window