Bear LD Decay

ECL 243, Feb-March 2018

Notes on attempting to re-create the LD Decay Plot from Liu et al (2014).

Tassel web resources
https://bytebucket.org/tasseladmin/tassel-5-source/wiki/docs/Tassel5PipelineCLI.pdf?rev=eeb05c28894bf976c84bb110ec545647ab9a2f08

Introduction

- The point: LD associated with molecular markers uncover genes linked with certain traits
- To plot LD decay, you need a column with distance in megabase pairs (Kbp) for the data and another column with LD (r2) values
- LD decay graph interpretation from Liu et al (2014):
    - r2 value is the squared correlation coefficients between two SNPs, that gives you LD. 1=total LD, 0=no LD. In the figure, at 50kBP apart, Brown Bears have reached backround LD. Polar Bears need to go to 150kBP. So, they have higher LD, probably due to a smaller effective population size. Probaly recombination is at the same rate in both populations, so that probably means smaller Ne. **When we run Plink or Tassel, we can tell it to just look for LD 300kBP or less**

LD in Haploview

- Liu et al calculated the r2 with Haploview. According to Haploview, it "saves time by only computing pairwise LD statistics for markers within a certain distance of each other" (authors used 200, default is 500kb), we don't know how they calculated the distance between SNPs, although they did tell us they used the following parameters:
  `-maxdistance 200 -dprime -minMAF 0.01 -hwcutoff 0.001 -minGeno 0.6`
- The program will not accept the Hapmap format that we created, so we changed the data to PLINK .ped and .map files
    - now Haploview is running an error based on our "chrom" #'s because they exceed 21'–> looks like these are scaffold #'s not chromosome #s

- To do this in haploview we need to translate their scaffolds to the 21 pseudochromosomes they used.
  - scaffold: an area where a lot of contigs overlap, this can then be mapped onto the chromosome. Their data only had scaffolds, not the chromosomes
  - Pseudochromosome: created to make the population genomic and demographic analyses more conventient. Potential error: paired end reads may align to the ends of different scaffolds
- Solution: Used r to re-write the the brown bear subset data to turn the scaffolds into chromosomes
- Problem: Not enough computing power on personal computers, the program will not run.
- Next Step: Attempt to run Haploview on the cluster - contact them to have it added to Farm.

LD in Tassel

COMMAND LINE:

- To run LD with tassel in the command line, first create an alignment?
  - An alignment means that tassel can help you get your data from different taxa into one compilation that it will do work on. Unclear if this is necessary. Maybe alignment just means the dataset for each bear species.
- Use “bbsort.hmp.txt” data, which is the clean data already sorted
- command line code (that doesn’t work):

```
module load jdk
module load java
module load tassel
run_pipeline.pl -Xmx44g -h bbsort.hmp.txt -ld -ldd png -o pb_ld.png
```

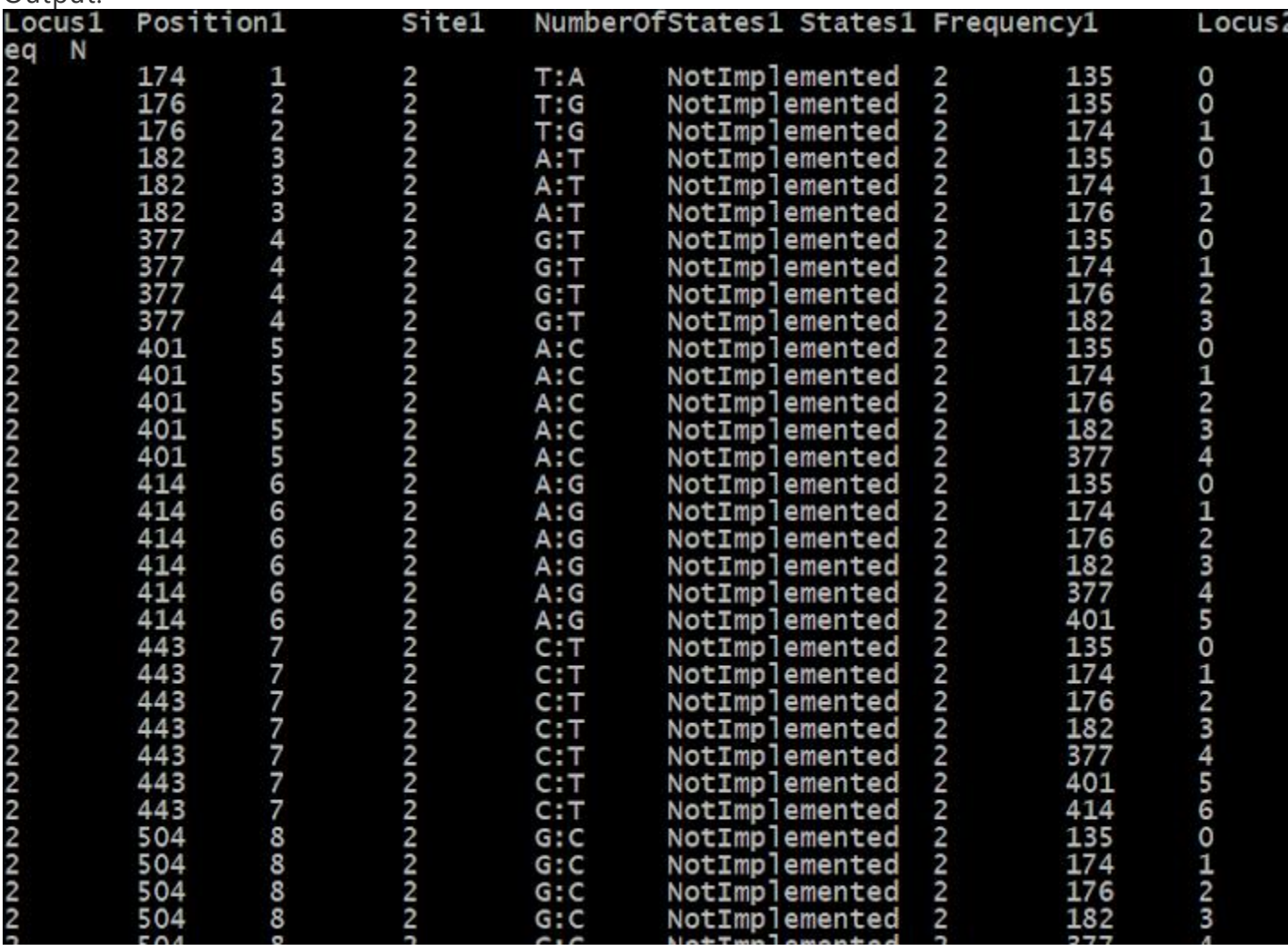
-run_pipeline.pl #executes the pipeline
-Xmx44g #how much gigs to use.
-h # means our data is in hapmap format that tassel likes
-ld #means run the ld plugin
-ldd #means output an image as a png
-o pb_ld.png #means the name of the output file is pb_ld.png

- this all ran, **but I don't know where the output went!** Maybe I messed something up and need to tell it to turn off the plug-in? Also, it gives back an error loading the bb data set bc one of the chromosomes has an illegal value

- Alternate code with different data set:

```
module load jdk
module load java
module load tassel
run_pipeline.pl -Xmx24g -h sortex.hmp.txt -ld -export pb_sub_LD
```

Output:



Locus1	Position1	Site1	NumberOfStates1	States1	Frequency1	Locus2
eq N						
2	174 1	2	T:A	NotImplemented	2 135	0
2	176 2	2	T:G	NotImplemented	2 135	0
2	176 2	2	T:G	NotImplemented	2 174	1
2	182 3	2	A:T	NotImplemented	2 135	0
2	182 3	2	A:T	NotImplemented	2 174	1
2	182 3	2	A:T	NotImplemented	2 176	2
2	377 4	2	G:T	NotImplemented	2 135	0
2	377 4	2	G:T	NotImplemented	2 174	1
2	377 4	2	G:T	NotImplemented	2 176	2
2	377 4	2	G:T	NotImplemented	2 182	3
2	401 5	2	A:C	NotImplemented	2 135	0
2	401 5	2	A:C	NotImplemented	2 174	1
2	401 5	2	A:C	NotImplemented	2 176	2
2	401 5	2	A:C	NotImplemented	2 182	3
2	401 5	2	A:C	NotImplemented	2 377	4
2	414 6	2	A:G	NotImplemented	2 135	0
2	414 6	2	A:G	NotImplemented	2 174	1
2	414 6	2	A:G	NotImplemented	2 176	2
2	414 6	2	A:G	NotImplemented	2 182	3
2	414 6	2	A:G	NotImplemented	2 377	4
2	414 6	2	A:G	NotImplemented	2 401	5
2	443 7	2	C:T	NotImplemented	2 135	0
2	443 7	2	C:T	NotImplemented	2 174	1
2	443 7	2	C:T	NotImplemented	2 176	2
2	443 7	2	C:T	NotImplemented	2 182	3
2	443 7	2	C:T	NotImplemented	2 377	4
2	443 7	2	C:T	NotImplemented	2 401	5
2	443 7	2	C:T	NotImplemented	2 414	6
2	504 8	2	G:C	NotImplemented	2 135	0
2	504 8	2	G:C	NotImplemented	2 174	1
2	504 8	2	G:C	NotImplemented	2 176	2
2	504 8	2	G:C	NotImplemented	2 182	3

- But what does it mean???

GUI:

- Input data as Hapmap format and select "Sort positions"

- ran LD with a subset of the polar bear and brown bear data (sortex)
    - Output:

R2BinMin	R2BinMax	Count
0	0.01	0
0.01	0.02	0
0.02	0.03	0
0.03	0.04	0
0.04	0.05	0
0.05	0.06	0
0.06	0.07	0
0.07	0.08	0
0.08	0.09	0
0.09	0.1	0
0.1	0.11	0
0.11	0.12	0
0.12	0.13	0
0.13	0.14	0
0.14	0.15	0
0.15	0.16	0
0.16	0.17	0
0.17	0.18	0
0.18	0.19	0
0.19	0.2	0
0.2	0.21	0
0.21	0.22	0
0.22	0.23	0
0.23	0.24	0

Problem

- Not sure how to interpret this output and it looks like the analysis did not run correctly
- In order to get any real data from this we need to have genetic distance between SNPs