Grace Campidilli

6/8/2021

Summary of files

1. Filter and gather data
   1. Genotype data
      1. Rscript: FINAL\_bonnmu\_data\_filtering.R
      2. Input:
         1. BonnMu.fromMaizeGDBbrowser.gff3
      3. Output:
         1. cleaned\_bonnMu\_rm\_double.RData
   2. Phenotype data
      1. Rscript: none
      2. Input:
         1. <https://images.maizegdb.org/BonnMuImages/>
      3. Output
         1. Bonnmu\_phenotypes.csv (originally on google sheets)
2. Identify which genes are most highly correlated with traits by performing GLMs in rTASSEL and extracting p-values for each gene/trait regression
   1. Prepare genotype data
      1. Rscript: FINAL\_rtassel\_prep.R
      2. Input:
         1. cleaned\_bonnMu.RData
      3. Output:
         1. hapmaps/bonnmu\_hapmap\_FREQ\_’x’PLUS.hmp.txt

Note: ‘x’ represents the the minimum number of different F2 families that must have an insertion in the same gene

for that gene to be included in downstream analyses;

Ran scripts for x = 3, 4, and 5

* 1. Prepare phenotype data
     1. Rscript: FINAL\_rtassel\_prep.R
     2. Input:
        1. bonnmu\_phenotypes.csv
     3. Output:
        1. logit\_phenoDF.RData

Note: FINAL PHENO DATA FILE

* + - 1. logit\_phenoDF.txt
      2. phenoDF\_updated.RData (not logit transformed)
  1. Perform GLM in rTASSEL
     1. Rscript: none

Note: Procedure for using rTASSEL software for analysis is commented at the end of FINAL\_rtassel\_prep.R

* + 1. Input:
       1. hapmaps/bonnmu\_hapmap\_FREQ\_’x’PLUS.hmp.txt
       2. logit\_phenoDF.txt
    2. Output:
       1. glm\_stats/glm\_stats\_FREQ\_’x’PLUS.csv

1. Use B73 reference genome to further contextualize Mutator insertion data in cleaned\_bonnMu
   1. Assign a genomic category tag to each insertion in cleaned\_bonnMu to identify where insertions are inserting relative to known regions of a gene: promoter, 5’UTR, cds, intron, 3’UTR, intergenic.
      1. Rscript: FINAL\_insertion\_categorization.R
      2. Input:
         1. cleaned\_bonnMu\_rm\_double.RData
         2. Zm-B73-REFERENCE-GRAMENE-4.0\_Zm00001d.2.gff3.gz
      3. Output:
         1. bp\_insertion\_summary.RData
         2. cbm\_cat\_updated.RData

Note: FINAL GENO DATA FILE

“categorized bonnMu = cbm”

* 1. Figure to visualize the frequency that insertions are inserting into genomic regions
     1. Rscript: fig\_for\_prop.R
     2. Input:
        1. pheno\_DF\_updated.RData
        2. bp\_insertion\_summary.RData
     3. Output:
        1. insertion\_df.csv : summary table about genomic categories and amount of insertions in each category
        2. freq\_figure.png

1. Perform 2 multiple linear regressions - chlorosis and germination traits are the response.

Model: chlor\_pheno ~ copy\_num + utr5 + cds + intron + utr3 + promoter + upstream\_of\_promoter + intergenic

* 1. Chlorosis and germination regressions
     1. Rscript: FINAL\_mlm\_for\_traits.R
     2. Input:
        1. glm\_stats/glm\_stats\_FREQ\_’x’PLUS.csv
        2. cleaned\_bonnMu\_rm\_double.RData
        3. cbm\_cat\_updated.RData
     3. Output:
        1. lm\_outputs folder

Ex. chlor\_lm\_sumary\_FREQ\_5PLUS\_CUTOFF\_5PER.txt