To Do BcAtGWAS\_RNAseq

* Col-0 bigRR
  + First run of bigRR
  + Troubleshooting done: need to rerun 500 phenos at a time
  + ~~rerun: only saved bigRR outputs from about 500 / 9k phenos~~
  + Run through z-scaling of effect estimates R script: 05\_zscale…R
* Lsmeans bigRR to T4 genome
  + Model: transcript ~ plant (fixed) + isolate
  + Wait for linux run to finish (ongoing 01/23 12pm)
    - Started 01/22 ~3pm
    - Failed 01/24 ~12pm with only ~900 phenotypes done.
  + Linux run on 500 phenos at once: done, took 16 hours (output b.csv)
  + Wait for linux batched run (500 phenos at once) to finish (started 01/25, 12pm)
    - Error: appending outputs to create huge files. Restarting 01/29.
    - Trying with writing out 1 csv per phenotype – faster?
      * And garbage collection gc()
  + Edited batched run (500 phenos at once, individual output files) is finished
    - started 01/29, 1pm
    - One pheno failed, restarted 01/30/18 1pm
    - One more pheno failed, stopped 01/30/18 7pm
    - Paused 01/31 11am to run scripts/testMethods
    - done 01/31, 9pm
  + Check lsmeans values: any large negative outliers?
    - Approach: save minimum lsmeans value per read, plot histogram of all mins
      * Mins as low as -30
    - Do same for max
      * Highs up to 12
    - Email Dan result
      * Comparing in scripts/testMethods:
        + Use default low minimums – no, why?
        + OR set all mins <-5 to equal -5 – no, weird cutoff effects
        + **OR use z-scaled lsmeans as phenotypes**
      * Analysis of results:
        + Effect size estimate: compare 3 methods – pdf plot
        + Visual comparison of manhattan plots – have saved
* Lsmeans z-scaled bigRR to B05.10 genome
  + Running bigRR
    - Started 02/02 – done 02/07
    - z-scaling lsmeans then running
    - Waiting on file compression 02/08 – 100 gb of small files.
    - 02/09 copied gzipped folder (30gb) to PC
    - should I also be running on lsmeans THEN z-scaling effect estimates?
  + Compare bigRR to T4 vs. B05.10
    - Rerunning all 12 indplants & 3 domest tomato phenotypes vs. BO5.10 on 02/08 –compare to T4 results
    - bigRR T4 effect sizes vs. B05.10 effect sizes for transcripts?
  + Test of z scaling of effect sizes
    - Use BcSolGWAS to test
    - Scatter plot z-scaled vs. original fx (3h)
    - find where threshold lands on z scale (30 min)
  + Run through z-scaling of effect estimates R script: 05\_zscale…R
    - Extract gene position information for each phenotype
      * BcinXgX <- Chr.g.position
  + How many SNPs associate with each phenotype?
    - Summary data: for each read, number of snps over: z-scaled 3, 4, 5, 6… etc.
      * Summary plots of this
        + Scatter plot: x = genome location, y = # SNPs > z thr, z = z thr color
        + Histograms: frequency of SNP counts over each z thr
        + X = z threshold, y = number of transcripts with any SNP > thr? Or y = distribution of effect sizes of top SNPs (as violin)?

Ask Dan what he meant for this plot

* + Where are the SNPs associated with phenotype?
    - Cis-effects on subset of transcripts
      * 05\_readin\_bigRRouts\_BO5.R
      * 30 chr 1 transcripts
        + Manhattan plots
        + plotted chr 1 vs. gene location: no obvious cis fx
        + send results to Dan, Monday
    - Is there a consistent effect estimate structure to all trans loci?
      * Approach: Phenotype-blind SNP locations
        + Keep only the SNP of largest effect for each phenotype
        + Lump all of these SNPs together, with their effect size info
        + Scatterplot this: x = SNP position on genome, y = effect size estimate

To get really crazy, could color points (z) by chromosome of origin for the read that gave us this SNP, and add color bars on x axis to denote chromosomes

* + - * Approach: across all bigRR outputs (transcript to B05.10), take mean at each SNP position
        + This is TOO BIG a task to open all files
        + So: 1. Loop through each SNP sequentially
        + 2. Open each file sequentially, read only the SNP of interest, save in a list
        + 3. Take mean
        + If it’s still too big, maybe can split this task per chromosome?
        + Practice this on Chromosome 1 only
    - SNP locations with transcript location
      * Chromosome 1 focus
        + X= position of top 1, 10, or 100 SNPs per Chr 1 gene
        + Y = position of gene center on Chr 1
        + Finding: very little cis effects
      * X = rank order of all transcript (phenotypes), y = rank order of TOP SNP (abs val z)
        + Don’t bother yet, based on Chr 1 won’t see synteny
        + Expect tracts of synteny for cis effects
  + Annotate SNPs? Maybe someday, but only after filtering
  + All analysis so far is on T4 SNPs, and Wei mapped transcripts to T4 genome
    - Add bigRR to B05.10?
    - email Wei whether reads are aligned to B05.10 or T4
    - ask Wei for T4 to B05.10 positional index file – convert SNP locations?
* Troubleshoot small cis effects
  + Select 1 BcBOT gene
    - Dendrogram: use all SNPs within gene ends (+- 1 kb?)
    - Are low expression isolates genetically related in this gene?
  + Crude haplotype calls per gene based on SNPs within gene
    - Problem: SNP by SNP association does not capture major cis-effects (due to minor allele/ haplotype)
    - Approach: lasso on SNP correlation matrix for locus?
    - Step one: crude haplotype assay to grab cis effects
      * fastPhase: see scripts from Michelle Stitzer
        + sensitive to a priori number of haplotypes
        + grab haplotype assignment surrounding a focal site or break into longest segments

focal site:

* + - Step two: map residuals after cis effects -- ID trans effects
* BcBOT bigRR
  + Quality check work so far
  + Check for completeness of BcBOT list previous run
    - Not sure where list came from/ how to verify
  + Annotate outputs
  + Evaluate cis/trans loci
  + Is BOT gene expression correlated with Botrytis virulence on one or more hosts?
* BcBOT Network bigRR
  + GWAS on the Botrydial network and the Arabidopsis genes that show co-expression with these transcripts
    - For At genes correlated with any Botrydial transcript – expression variation across 96 Bc isolates
      * Including only At genes with >=2 associations with the Botrydial genes?
    - For At genes correlated with only the expressed Botrydial transcripts (removed transcripts with 0 expression in XX isolates)
      * Email Wei for transcript removal details
    - Option 1: Can bigRR run each gene in these networks individually, then batch analyze downstream
    - Option 2: can bigRR run z-scaled average of all genes in a network??
  + GWAS on network z-scores of the top Bc networks on col-0 background
    - Wait for Wei to send z-scaled network variation
* General analysis
  + Data management for large files (many SNPs x many phenotypes)
    - Slow to read in/ write out: batch process split files
    - Slow to repeat calculations across all phenotypes
  + Give GEMMA another shot
    - Step one: assess progress so far and skim through old code
    - Test on BOT reads first
  + Check in with Celine about EMMAX
* Writing
  + Write methods so far
    - Modify methods description from Vivian’s paper
    - Update my analysis methods
    - Start bibliography
  + Start introduction
    - Background reading
    - Annotated bibliography / early citations
  + Start results
    - Thought map of figures and analyses that I want to include
    - Save outputs in a .ppt as I go
    - Start writing up results text
* Reading
  + PNAS Hevia 2015: Circadian clock Botrytis. In my lit folder