BcSolGWAS\_redraft\_todo.docx

1. Add GEMMA analysis
   1. Run linear model association test for DWS
      1. Manhattan plots of output
   2. Calculate relatedness matrix for DWS
   3. Run linear mixed model association test using relatedness matrix for DWS
      1. Manhattan plots of output
      2. Double check this, then add plots
      3. Use only kmat1 (same as kmat2)
      4. Double check: is this to B05.10 ???
   4. Compare bigRR T4 to GEMMA B05.10 for DWS
      1. QQ plot ?
         1. Remove excess bigRR SNPs. Pair SNPs based on location order. Plot effect size from each method
            1. This is negatively correlated, which seems weird. But it does give us a set of SNPs that are “more significant” across both methods
         2. Remove excess bigRR SNPs. Pair SNPs based on relative effect size. Plot location from each method.
            1. This is just a big mess
   5. Add plots from GEMMA for DWS
      1. Figure 7
         1. A: top 1000 significant SNPs per D, W, S - color by trait
         2. B: Venn Diagram of SNPs by trait
         3. C: Venn Diagram of genes by trait?
      2. Figure 5
   6. GEMMA for 12 phenos?
      1. Calculate relatedness matrix
      2. Run linear mixed model association test using relatedness matrix
         1. May need to rerun for LA1589? Failed using kmat1 and kmat2?
      3. Add plot: Figure 4 copy = Figure S2
         1. Script: 06\_bigRRplots\_12plants\_FigR4a.R
         2. A: Manhattan plot for LA2903
         3. B: “Manhattan” plot as # plant genotypes with p < 0.01
         4. Write into text
      4. Add plot: Figure 5 copy -- figure S3
         1. A: number of SNPs vs. plant genotypes per SNP
         2. Add gene annotation
         3. B: number of genes vs. plant genotypes per gene
      5. Double check: is this to B05.10??
2. Overlaps bigRR vs. GEMMA
   1. Script:
   2. Gene lists from bigRR
      1. Domest
      2. 12 phenos
   3. Gene lists from GEMMA
      1. Domest
      2. 12 phenos
   4. Check for gene level overlap
   5. Functional annotation of overlapping genes
      1. <https://www.broadinstitute.org/fungal-genome-initiative/botrytis-cinerea-genome-project>
      2. <https://urgi.versailles.inra.fr/Species/Botrytis/Download> gene predictions but maybe only for T4
      3. <https://urgi.versailles.inra.fr/Species/Botrytis/Sequences-Databases> idk
      4. <http://www.uniprot.org/uniprot/A6RIE1&format=html> maybe protein lookup by gene name
      5. <http://fungi.ensembl.org/index.html> idk
      6. <https://pfam.xfam.org/> pfams
      7. <https://www.ncbi.nlm.nih.gov/pubmed/28582481> try this. FUNgap
      8. <http://fungidb.org/fungidb/> idk
      9. <http://botbioger.versailles.inra.fr/botportal/index.html>
         1. Also find the annotation file for table S1 : it’s C:\Users\nesoltis\Documents\Projects\BcSolGWAS\data\GWAS\_files\05\_annotation\window2kb\AllAnnots\_byGene.xlsx
3. Revise supplementary figures
   1. For Table S1, we’ll just upload it as a .csv file rather than in the powerpoint format.
      1. annotation file for table S1 : it’s C:\Users\nesoltis\Documents\Projects\BcSolGWAS\data\GWAS\_files\05\_annotation\window2kb\AllAnnots\_byGene.xlsx
      2. now saved in resub folder
      3. to do: add bigRR/ GEMMA overlaps to S1!
   2. Same for Table S3 but it will be broken into two tables. give the B05.10 gene name as well for part A.
      1. Find S3A, S3B, copy to resub folder
      2. Add B05.10 gene names to S3A
      3. Add GEMMA results?
   3. For Figure S2, you have –log of the p-value as well as a threshold with p=0.001 but the two values don’t disagree. How was the threshold calculated? I think we need to tweak the wording.
      1. Change?
   4. Should we combine Figure S3 with the BigRR figure using a line diagram? That cuts our figures down by 1 and shows the results clearer.
      1. What does this mean?
   5. Figure S4, what is the axis on part A and do we need the thresholds if you only show significant SNPs?
      1. Change y axis: -log(p value)
4. Gene list overlap GEMMA/ bigRR
   1. Extracted gene lists with R script
   2. bigRR list shorter so: selecting 300 gene chunks
   3. plug into botportal with T4-vankan, remove hyperlinks, copy
   4. paste, remove headers, and save in new .csv
   5. reopen .csv and add into main file
   6. combine all files
      1. paste into tabs on one document
      2. remove weird characters
      3. save whole .xls file for function lookup later
      4. only keep columns:
         1. T4vanKan-BROAD
         2. B0510-BROAD gene name
         3. T4-URGI gene name
         4. B0510vanKan-BROAD gene name
         5. BcinB0510 gene name
         6. GenBank gene name
      5. filter: select down arrow for each column, check only blank
      6. select all blank rows and delete them
      7. clear filter
      8. then merge all files into one list
      9. then import to R.
5. Address reviewer comments
   1. Broad comments: reviewer 1
      1. Phenology of domesticated vs. wild tomato
         1. Phenology is a mechanism of disease resistance -- incorporated into design (rather than flaw in experiment)
         2. should check how we wrote about making sure the plants were mid-way between flowering and germination to minimize phenological differences and see if we need to strengthen the comment about how the leaflet and leaf position did not matter in the model. We should also write in the phenology is an ecological form of resistance known as evasion and that we consider this as valid to study as classical resistance mechanisms.
      2. Selection of leaflets
         1. 2nd true leaf and younger
      3. Inoculation droplet position
         1. Quantify lesion area AND lesion shape. 2 separate measures, interested in both (incl. how much does the lesion track the vasculature). Simply focusing on lesion area here.
      4. Inoculation solution
         1. Spore concentration
         2. Spore germination -- timing
      5. Replication
         1. THREE replicates at TWO timepoints per interaction
         2. How to justify this as sufficient?
         3. Test removing 1 replicate -- change conclusions?
      6. GWA mapping on gapless B05.10 genome
         1. bigRR?
         2. Done in GEMMA
      7. Annotation windows
         1. Smaller windows
         2. Test for assigning >1 gene to SNP?
   2. Broad comments: reviewer 2
      1. Only include GENETIC factors in % variance explained by model (omit experiment, block)
      2. Kinship across tomato lines - sampling bias?
         1. Maloof paper
         2. cite the Maloof/Zamir paper to say that these tomato genotypes are well distributed in the distribution from wild to domestication and do not cluster meaning they are as randomized a sample as possible.
      3. Figure 3 redundant. No phenotypic bottleneck. Less discussion of variance in lesion size
      4. Figure 1c-h an interaction of 2 genomes? Random non-genetic variation?
      5. Interaction effect of 2 genomes: use random effect
      6. Table 2. Wilcoxon with FDR
         1. Include table 2
         2. Clarify FDR correction of p-value
         3. Is population structure of Botrytis considered?
         4. Independence of individuals
         5. Null distribution of Wilcoxon statistic?
      7. Account for population structure
         1. Standard mixed-model GWAS?
      8. Figure 5a
         1. Statistical analysis: overlap > expectation?
      9. Domestication GWAS
         1. Account for population structure
         2. Why genetic effect domest =/= wild
      10. Remove table 1 redundancy (SS vs. F-value)
      11. Report p = XX NOT p < XX.
      12. Figure 2: too small a wild vs. domesticated effect
   3. Reviewer 3 comments
      1. Phylogenetic distance in host?
      2. Line 283: “we identified a significant increase in the resistance of wild tomato in comparison to domesticated tomato”
         1. Reword: observe expected decreased resistance in domesticated tomato --- not direct change/ causative
      3. Pathogen Specialization to Host Variation section
         1. Retitle. There is specialization to genotypes within the host variation (host x pathogen genotypic interactions)
      4. Include caveats in focusing on 2/97 isolates with significant domestication effect
      5. Figure 5
         1. Better explain in figure legend + text
         2. Explain inset graphs
         3. Line 406: how to interpret inset graphs for “levels of overlap exceed the expected overlap due to random chance”
      6. Ancestry of domesticated genotypes
         1. What is known about origins? Multiple domestication events?
   4. Additional edits from Celine
      1. Redraw black/ grey figures as color figures (esp. GWAS)
      2. Figure 1
         1. put into the text that we inoculated close to the central vasculature and measured both lesion size and lesion shape. While both traits showed genetic variation due to the plant and the pathogen that we are going to focus on lesion size in this manuscript. This approach tells the reader that both are informative but we don’t have time to talk about both.
         2. We should however comment tighter in the figure legend of Figure 1 about the statistical test used to make these calls.
      3. Cartoon to explain experimental design
      4. Figure 3 highlight different patterns with colors -- according to statistical threshold
         1. Figure 3 could be colored using a rank plot like Vivian did where she colored the isolates based on their lesion in Wild and kept that color across to domestication. It helps visualize the interaction a bit better.
   5. Add in GEMMA methods
      1. “The GWA was performed using the options to create a centered relatedness matrix (−gk 2) and perform all three possible frequentist tests: Wald, likelihood ratio and score (−fa 4). The relatedness matrix was incorporated to control for family structure among the discovery cohort and was constructed using a linkage-disequilibrium (LD)-pruned set of markers from the imputed genomewide SNP data (100 SNP windows, sliding by 25 SNPs along the genome, pruned at r2 > 0.2; PLINK command –indep-pairwise 100 25 0.2) [45]. SNPs were pruned prior to GWA using the default GEMMA parameters of MAF < 1 % and missingness < 95 %. The genomewide significance cutoff using an adjusted Bonferroni correction based on the effective number of independent tests in our data was p <1.86 × 10−6 [58]. P-values between 1.68 × 10−4 and 1.68 × 10−6 were considered to indicate moderate association. ”
   6. Cover letter