Hey Jason,

Thanks for all your work assembling this dataset.

I’ve done a few sanity checks on the dataset and have a few outlier individuals for you to look into when you do your own error checking.

**Bolded below are questions I have for you – I counted eight questions in total.** The explanation for the questions generally occurs beforehand. I’ve also made all the code and results available in a private repository on Github. If you want to see the code or additional figures, if you make a (free) Github account and tell me your account name, I can share the whole repository with you. I’ll share the main points here, but all the additional the ‘supporting information’, let’s call it, is all on Github.

-----------------------------------------------------

To decide what 2019 data to include, I went through and compared your “SRV” (survived to 2019) column to the “DEAD\_2018” column I made from the information in the file now called “GWAS\_2019\_Mortality, Weak and Not P. Virgatum Plants\_ARCHIVE.xlsx”.

I dropped anything that was called "Not virgatum" in this file, 7 individuals, 2 which “SRV” was “W”. **Do you agree that we should drop data for these 2?**

FRMI F2610 2610 J203 J203.A NA 156 NA NA W Likely not virgatum, Weak switchgrass?

FRMI F2707 2707 J421 J421.A NA 149 176 NA W Likely not virgatum, Weak switchgrass?

A few others were in disagreement on whether or not the plant died:

13 called as "W" in “SRV” but were called previously as dead (and were noted (in 'NOTES' in the ARCHIVE file) as "Dead" or "dead".) 8 have phenotype data collected in 2019.

CLMB C6920 6920 J216 A J216.A J216.A.CL

FRMI F1002 1002 J462 A J462.A J462.A.FR

FRMI F1714 1714 J458 B J458.B J458.B.FR

FRMI F2716 2716 J498 C J498.C J498.C.FR

FRMI F2724 2724 J235 A J235.A J235.A.FR

KING K1909 1909 J465 C J465.C J465.C.KG

KING K2008 2008 J597 A J597.A J597.A.KG

PKLE P5601 5601 J036 A J036.A J036.A.PK

TMPL T1303 1303 J597 A J597.A J597.A.TP

TMPL T1609 1609 J538 A J538.A J538.A.TP

TMPL T2308 2308 J504 A J504.A J504.A.TP

TMPL T2721 2721 J477 B J477.B J477.B.TP

TMPL T2803 2803 J592 C J592.C J592.C.TP

For an additional 6 individuals, SRV was called as "Y", but were noted (in 'NOTES' in the ARCHIVE file) as "Dead". All 6 have some phenotypic data collected in 2019.

FRMI F2519 2519 J482 A J482.A J482.A.FR

FRMI F2523 2523 J577 B J577.B J577.B.FR

FRMI F2720 2720 J465 A J465.A J465.A.FR

FRMI F3302 3302 J340 A J340.A J340.A.FR

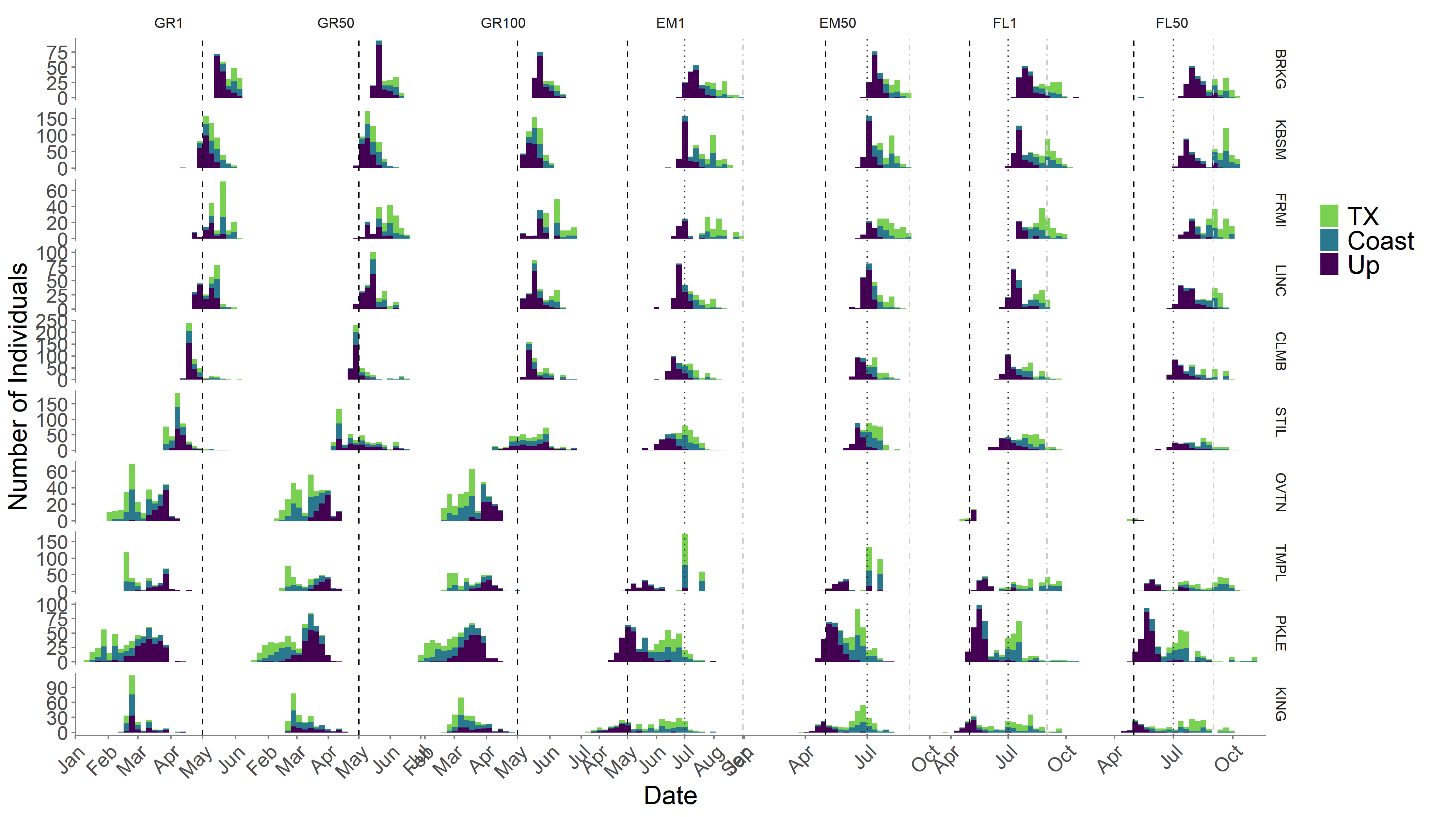
TMPL T2106 2106 J019 A J019.A J019.A.TP

OVTN V1319 1319 J580 A J580.A J580.A.VN

We’d benefit from a critical eye on the data for these individuals. **What should we do with the data for these 19 individuals?**

----------------------------------------

Dropping these individuals for now, I then looked at patterns in the measurement values by phenotype, genetic group, and site. Histograms of the phenology data by genetic group and site definitely look really interesting, and like there will be a lot of G and GxE to explore in the data:



There are dotted vertical lines on May 1st (black), July 1st (dark grey), and September 1st (light grey). The sites are ordered by latitude on the y axis, and the phenology phenotypes are ordered by timing in the season on the x axis. The colors indicate the Texas lowland ecotype (green), the coastal lowland ecotype (blue), and the upland ecotype (purple) (NB: If you divide this into genetic subpopulations, the results look very similar). I removed AP13 individuals from these histograms because they have a lot more datapoints than other individuals at each site.

A fun observation is that, for, say greenup (GR1 through GR100), you can observe GxE in the purple (upland) individuals relative to the other individuals – at northern sites, they green up first, and at southern sites they green up last.

Also, we don’t have much panicle emergence (EM) or flowering (FL) data for OVTN – I assume there’s a story there – it’d be good to know that so we can put it in the writeup in the Methods.

**Any story for the missing data for EM and FL in OVTN?**

-----------------------------------

Next, to look for outliers, I mostly plotted phenotypes against one another – say, EM1 vs EM50 – with the expectation that EM50 values should be equal or larger than EM1 values. Or, in general, that:

GR1 < GR50 < GR100 < EM1 < EM50 < FL1 < FL50

A close up of a map

Description automatically generated So, in the plots below, any points that are below the black diagonal line are potential errors.A close up of a map

Description automatically generated

We don’t see any points below the line for Greenup (nor for comparisons between GR1, GR50, and GR100).

A close up of a map

Description automatically generated

A drawing of a face

Description automatically generatedThere are a few points below the line for panicle emergence. If you break down this plot by site, you see that all of these points are in TMPL. In general, TMPL has some issues with EM1 and EM50 data. Both were only collected two-three times at later timepoints. We’ll have to think about how this might affect the analyses **– any story with EM at TMPL?**. Also, KING has an interesting delay in EM50 relative to EM1.

A close up of a map

Description automatically generatedFlowering also has a few points below the line – these are split between BRKG, TMPL, and OVTN. There’s also a spread of points above the line, particulalry for PKLE and KING.

A close up of a logo

Description automatically generated

Here are the individuals where EM50 is smaller than EM1:

SITE PLOT\_GL PLOT\_LC ACC PLANT\_ID EM1 EM50 EMdiff

STIL S1110 1110 J587 J587.A 186 185 -1

STIL S2907 2907 J073 J073.A 190 184 -6

STIL S3603 3603 J210 J210.A 188 186 -2

TMPL T1417 1417 J321 J321.A 182 180 -2

TMPL T2101 2101 J586 J586.B 199 182 -17

TMPL T2103 2103 J008 J008.A 182 126 -56

TMPL T2111 2111 J326 J326.B 199 19 -180

TMPL T2203 2203 J614 J614.B 142 141 -1

TMPL T2411 2411 J270 J270.A 199 99 -100

**Are any of these EM datapoints errors?**

Here are the individuals where FL50 is smaller than FL1:

| **SITE**  <fctr> | **PLOT\_GL**  <chr> | **PLOT\_LC**  <dbl> | **ACC**  <chr> | **PLANT\_ID**  <chr> | **FL1**  <dbl> | **FL50**  <dbl> | **FLdiff**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- |
| BRKG | B1406 | 1406 | J635 | J635.B | 289 | 193 | -96 |
| BRKG | B1409 | 1409 | J482 | J482.A | 231 | 133 | -98 |
| TMPL | T1409 | 1409 | J427 | J427.B | 154 | 56 | -98 |

**Are any of these FL datapoints errors?**

**Do you want any of the FL or EM datapoints from KING or PKLE that are well above the 1:1 line?**

A picture containing text

Description automatically generatedThere are also interesting patterns when you compare greenup, emergence, and flowering times by plant and by site. Particularly when you color the measurements by ecotype, you can see that there are big differences in how the phenology of the three ecotypes is responding at different latitudes.

A close up of text on a white background

Description automatically generated

A close up of text on a white background

Description automatically generated

So for the GxE study, here’s what I think are the most interesting quantities:

Greenup (just as Julian days, or possibly as a time to greenup after the site has a certain number of growing degree days)

Time to bolting/Vegetative growing time (EM50 – GR50, or possibly EM1 – GR1)

Panicle emergence date (also could tie this back to temperature by making it a function of growing degree days)

Flowering date (also want to tie this back to site weather somehow as above)

Time to flowering (FL50 – GR50, or possibly FL1 – GR1).

Tiller count EOS 2019 (I’ve done tiller counting now and it’s very imprecise; I’m wondering about putting this on a natural log scale for GWAS, so it’d be more like doubling the tiller number)

Tiller addition rate (Tillers added per vegetative growing day or total growing day (TC\_EOS / (FL50 – GR50), or something else).)

Panicle height EOS 2019

**Any comments on these, or additional suggestions (especially informed by how these plants actually grow in the field)?**