MarkerPaperMetaAnalysis\_Rrr

Amanda Charbonneau

December 29, 2014

### Does days to flowering differ between native and non-native range populations of *R.r. raphanistrum*?

*R.r. raphanistrum* is native to the area around the Mediterranean, and it has been suggested that it originated on the western side where Raphanus diversity is highest. We have *R.r. raphanistrum* populations from three main geographic areas: the western side of the Mediterranean (all from France or Spain), the eastern side of the Mediterranean (all from Israel), and non-native. To test whether there is a difference in flowering time between native and non-natives, and also look for differences between native populations on either side of the Mediterranean Sea, we are modeling a subset of the data that only includes *R.r. raphanistrum* populations.

### Models for *R.r. raphanistrum*

We want to model days to flowering as a function of Geography. As our datasets were done at different times and places, this is an idealized model which attempts to account for all known sources of variation caused by differences in experimental design. To start, I'm trying five models that are all very similar to each other, and all have the same basic fixed and random effects:

#### Fixed Effects

**Geography** allows for each geographic group (east, west and non-native) to have its own intercept

**Seedstock** captures the variation in flowering time due to parental growth environment. Since most populations only have a single parental growth environment, this should only see the effect of the qtl2004 populations, which were all propagated from qtl2003 and lale2005 populations which were propagated from qtl2004.

**GrowthEnvironment** accounts for variation due to the study plant being growth in the greenhouse or field

**DOYGS** (Day of year germinated, standardized) OR **DOYPS** (Day of year planted, standardized) accounts for variation in seasons, as experiments were started at different times of year and so experianced differing growth conditions like length of day and light intensity. We don't have germination date for all plants, but know that some have a gap of weeks-months between germ and flower dates. The 'standardization' is subtracting 78 from each value, then adding 365 to values less than zero, so that all the winter days are together.

**Habitat** accounts for variation due to type of micro-habitat that each population was collected from and may be locally adapted to

**DaysVernNoNA** accounts for whether an individual was vernalized, and for how long

The interaction term **GrowthEnvironment:Geography** allows populations from each geographic region to have independent reactions to **GrowthEnvironment**

#### Random Effects

The **Geography**|**Experiment** term accounts for apparent variance among population means due to **Geography** that may be due to differences between **Experiment**s

The **Geography**:**Pop** term finds the variance among population means within each geographic region

Before running the models, here are graphs looking at the data I'm going to use to make sure everything looks like it's imported correctly and to see if any variables look bad (too many NA's, weird data points, etc.)

##### Response variable

We are modeling days to flowering as the response variable, but we have two different versions of this: **PTF** is planting to flower, which is the amount of time it took from putting the seed in soil until the subsequent plant first flowered. We also have **GTF**, which is germination to flowering. This is the amount of time from when the seedling first emerged from the soil until the plant flowered. Although these are very similar, only one, **PTF**, was recorded in all experiments, however for the experiments where both were recorded many individuals had very large lag times between planting and germination, with a range of 3 to 142 for both the dataset as a whole, and just the R.r. raphanistrum subset. (The individual that took 142 days was an MAES.)

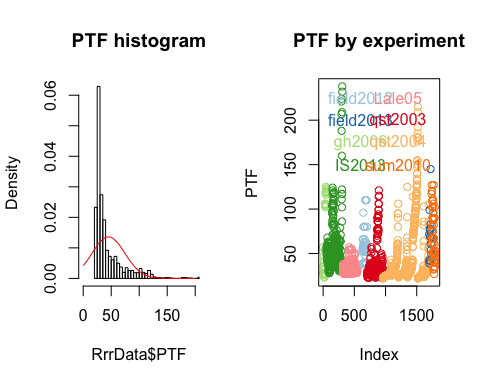
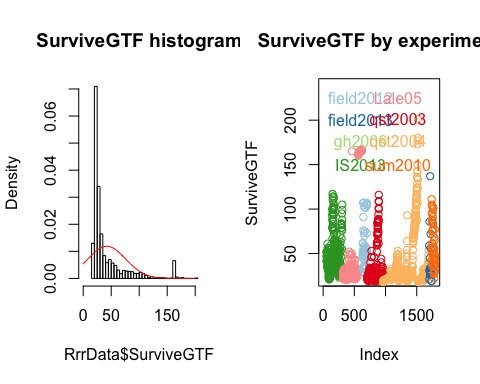
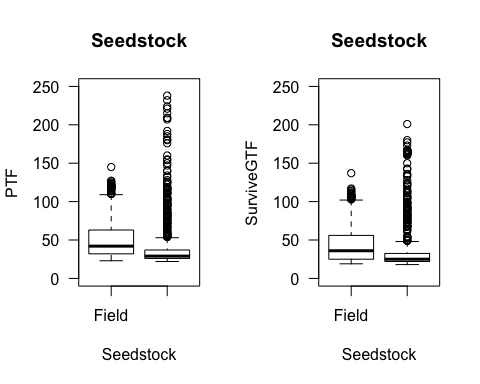
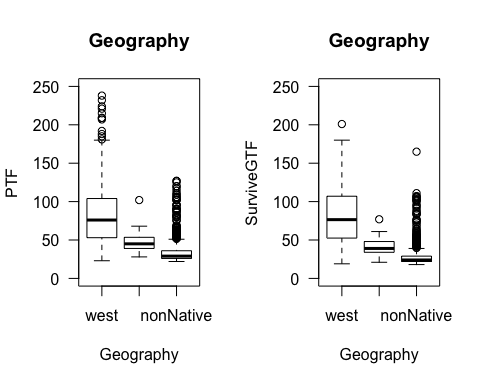
 

Figure 1. PTF vs. SurviveGTF.   
We can see that using SurviveGTF only slightly changes the density plot of flowering values, however it drops an entire experiment, gh2006, which appears as light green in the PTF graph. It's also obvious that many individuals within some experiments either didn't have SurviveGTF recorded, or had a very large difference in their PTF and SurviveGTF. Both IS2013 and qst2004 lose most of thier large values in the SurviveGTF graph.

##### Predictor variables

These are plots of each predictor variable against the response. For each I have plotted against **PTF** on the left and against **SurviveGTF** on the right.

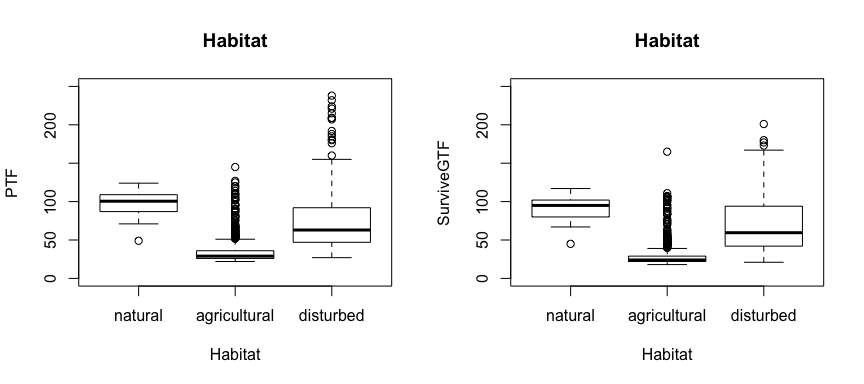
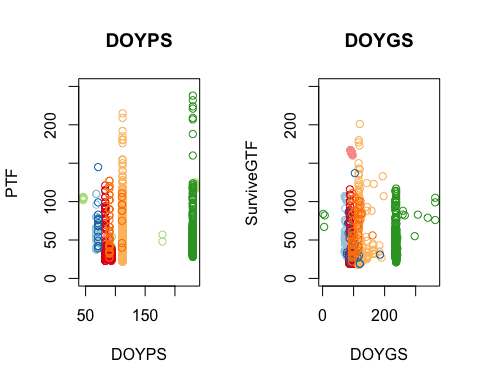


Figure 2. Variables basically unchanged by switching to SurviveGTF.   
a. Flowering time distribution of plants whose parents were field collected, versus those whose parents have reduced maternal effects/"lab domestication". b. Geographic region where the population was originally collected in the wild. c. Type of habitat where the population was originally collected in the wild. Although a few data points drop out when plants without a known germination date drop out, the distribution within each factor is basically unchanged for all of these variables.



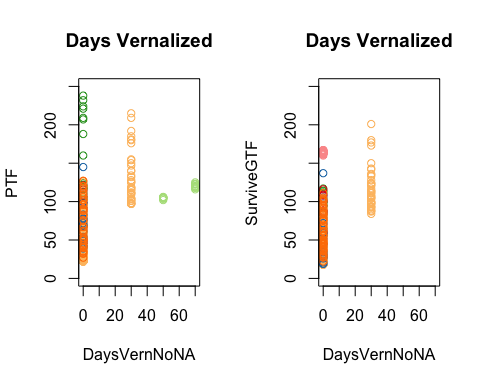


Figure 3.   
Variables signifigantly altered by using SurviveGTF.  
a. Day of year individual was planted/germinated. Since in most experiments, all individuals were planted at once, DOYP is effectivly a factor that is mostly colinear with experiment. In fact, the on expeiment with varied planting dates, is the gh2006, which has no germination data. DOYG has a much wider spread of dates for each population, so it will behave more like a continuous variable as intended and probably have a better estimate. b. Days Vernalized. This variable was already somewhat suspect because although we are treating it as continuous, in practise it is more like a factor. It happens that the plants that experianced 50 and 70 days of vernalization were also ones that didn't have germination recorded, so in the SurviveGTF model, this is effectively now vernalized: yes or no.

#### Modeling

##### Model 1:

###### Response variable: Days from planting to flowering

Because we know that **PTF** incorporates a lot of germination latency, this model will likely over-esimate time to flowering considerably. However, we don't have germination dates for one experiment, and I don't know how having a large number of NA's in the response would effect the model. So, this uses planting to flowering as the response, and DOYPS (day of year planted, standardized), which were recorded in all experiments.

Rrr\_lmer\_1 <- lmer( PTF ~ Geography + Seedstock + GrowthEnvironment + DOYPS + Habitat + DaysVernNoNA + GrowthEnvironment:Geography + (1 + Geography|Experiment) + (1|Geography:Pop), data=RrrDataGTF )

##### Model 8:

###### Model 2, but without Habitat and the **Geography**|**Experiment** random effect changed to just 1|**Experiment**

Rrr\_lmer\_8 <- lmer( GTF ~ Geography + Seedstock + GrowthEnvironment + DOYGS + DaysVernNoNA + GrowthEnvironment:Geography + (1|Experiment) + (1|Geography:Pop), data=RrrDataGTF )

##### Model 13:

###### Response variable: Days from germination to flowering

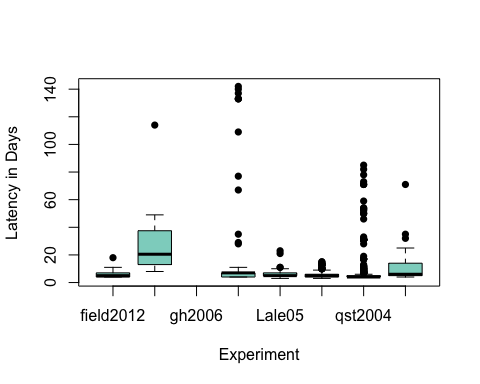
Here, I'm using germination to flowering (or last alive date) as the response, which will be more accurate than planting to flowering, but have many more NA's.

Rrr\_lmer\_13 <- lmer( SurviveGTF ~ Geography + Seedstock + GrowthEnvironment + DOYGS + DaysVernNoNA + GrowthEnvironment:Geography + (1|Experiment) + (1|Geography:Pop), data=RrrDataSurvive )

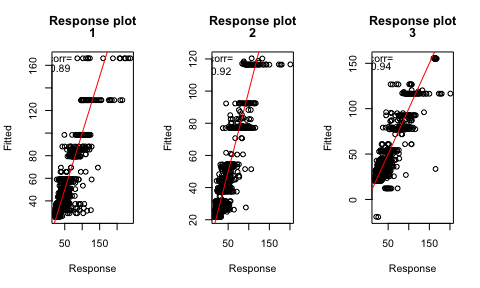
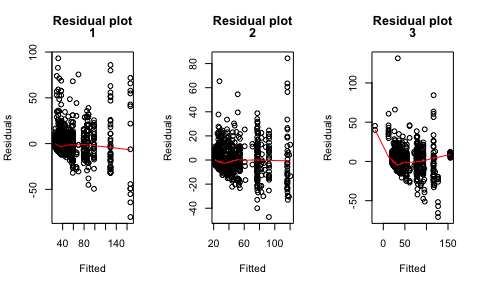
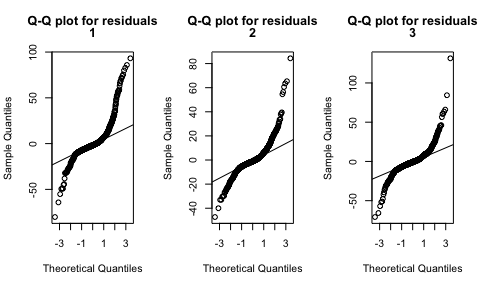
#### Model Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Fixed Effects** (std.err) | M1: Response PTF | M2: Response GTF | M3: Response SurviveGTF |
| GeographyWest | 83.9 (40.48) | 49.9 (11.47) | 88.6 (21.26) |
| GeographyEast | 15.8 (41.1) | -27.5 (16.65) | -43.7 (31.48) |
| GeographyNonNative | -16.4 (8.78) | -23.8 (8.21) | -61.5 (14.98) |
| SeedstockGreenhouse | 106.3 (6.3) | 13.7 (4.45) | -18.6 (5.17) |
| GroEnvirGreenhouse | 23 (51.49) | 8.2 (12.66) | -53.4 (23.22) |
| DOYGS or DOYPS | -0.1 (0.07) | 0 (0.02) | 0 (0.02) |
| DaysVernNoNA | 1.7 (0.09) | 1.3 (0.06) | 1.3 (0.07) |
| GeoEast:GroEnvirGreenhouse | -15 (53.78) | 4.2 (17.25) | 34.4 (32.58) |
| GeoNonNative:GroEnvirGreenhouse | -24.7 (6.46) | -15 (3.91) | 70 (2.15) |

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Effects** std.dev | M1: Response PTF | M2: Response GTF | M3: Response SurviveGTF |
| Geography:Pop | 9.6 | 14.94 | 28.79 |
| Geographywest by Experiment | 66.97 | 16.09 | 30.27 |
| Geographyeast by Experiment | 68.04 | NA | NA |
| GeographynonNative by Experiment | 4.13 | NA | NA |

As you can see from the tables above, model one gives a much higher point estimate for base days to flowering, as well as for the **Geography** and **GrowthEnvironment** variables. This is not surprising since the plants that had long latency between planting and germination tended to be from a only a few populations and experiments:  


Models 2, 3, 4 and 5 all have very similar results to each other. Models 2 and 3, are, in fact, identical. This is not terribly surprising since gh2006 doesn't have **SurviveGTF** data. Model 2 used **SurviveGTF** as the response, and model 3 excluded the gh2006 dataset, so I expected them to correlate with each other. Similarly, I expected Model 5 to have analogous estimates. The estimates for the effect of east are small across all models, so removing the ability to assign effect to east shouldn't have much effect. However, I didn't expect any of them to be so similar to Model 4, which excluded all of the vernalized plants. Although the point estimates differ slightly, between 2/3, 4, and 5, all the differences are less than a day, which is biologically insignifigant, and below our detection threshold since we checked plants daily at most. So for inference, models 2, 3, 4 and 5 are basically indistinguishable from one another. You can also see that they all explain the data equally well:

However, if we look at the variance inflation factor, that is, how much each predictor is contributing to the overall variance, both the **PTF** model and models 2-4 with **PTF** have a lot of variance inflation at the predictors using **Geography**. (All of them are exactly the same, this is the one for **SurviveGTF**)

However, changing from the three leveled **Geography** to the two leveled **Native** decreases the VIF by quite a bit:

## GVIF Df GVIF^(1/(2\*Df))  
## Geography 154.588707 2 3.526099  
## Seedstock 1.480117 1 1.216601  
## GrowthEnvironment 13.447571 1 3.667093  
## DOYGS 3.122067 1 1.766937  
## Habitat 5.569463 2 1.536220  
## DaysVernNoNA 1.443283 1 1.201367  
## Geography:GrowthEnvironment 262.010493 2 4.023275

## GVIF Df GVIF^(1/(2\*Df))  
## Geography 154.593676 2 3.526127  
## Seedstock 1.479664 1 1.216414  
## GrowthEnvironment 13.407736 1 3.661657  
## DOYGS 3.121269 1 1.766711  
## Habitat 5.568258 2 1.536137  
## DaysVernNoNA 1.443262 1 1.201358  
## Geography:GrowthEnvironment 261.787241 2 4.022417

## GVIF Df GVIF^(1/(2\*Df))  
## Geography 84.669841 2 3.033417  
## Seedstock 1.397951 1 1.182350  
## GrowthEnvironment 6.833598 1 2.614115  
## DOYGS 3.182598 1 1.783984  
## Habitat 5.639713 2 1.541041  
## DaysVernNoNA 1.433113 1 1.197127  
## Geography:GrowthEnvironment 121.051688 2 3.316979