Sunflower Oil x Envr - r2 (shared)

A useful and not overly complex analyses to get out the door would be an analysis of the wild sunflower seed oil traits and the environmental parameters that correlate with them. A great target journal and submission type would be to the American Journal of Botany as a brief communication (research article). Those types of article are 3000 to 4000 words in length, and have no more than 4 visual items (tables or figures). This article type is concise and gives enough room to explore one or two ideas in an manuscript.

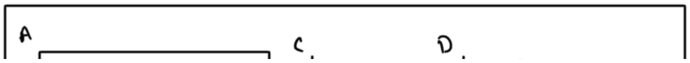
I need to get permission to use there data, find out what environmental analyses were conducted in the original dissertation, and then start building models. It's possible there are more traits in there than oil that have the large latitudinal gradient, and it may be possible to conduct a novel analyses on those data. However, just sticking to the oil and envr. traits would make for a tight brief communication.

I could get this sent out to review in two months for sure.
Authors: Karl Fetter, Max Barnhart (if he's interested), Ed McAssey, Andy Goeherty, John Burke.

Roles:
KF - fit models and write the MS
MB - collate the data and create figure 1, contextual brainstorming for introduction, edit text
AG - create shapefile of wild sunflower range.
JB - keep the lights on, edit MS

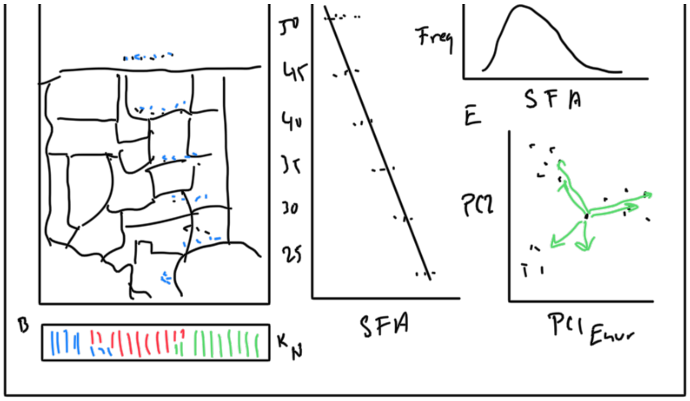
Fig. 1

Sample & Envr. Context Figure



Panels

A Sample map



B Barplot (Adapted from Pub

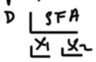
C Latitude observation

D Histogram of trait

E PCA on Envr.

Notes:

- Color points According to deme.
- Adapt structure / Admixture bar plot from Original 2016 MS (here for convienience)
- (D) could include hist of one or two other intersting envr params. eg. GSL, Temp, Precip. Params that turn out to be important.



- Include Latitude in PCA.
- Create convex hull of wild sunflower accessions from GBIF to serve as a rangemap for wild helianthus.
  - Include Range as a shapefile & provide in supplement.
  - then when people use it, they will cite the manuscript.
- Possibly something Andy Goherty can do. He can probably do a really good job quickly.

Input Data

Y = %Saturated fatty Acids

X = 19 Bioclim
Growing Season length } Abiotic -lat-longs envr.
...

G = Population Effects

Admixture Qmat., Kinship matrix?
→ From where? something difficult to get.

Grouping Effects

Deme_code
Pop_code
possibly ind_code? or family_code ← Depends on data str.

Model.

Basic Idea: $Y \sim E + G + \epsilon$ $\qquad V_p = V_G + V_E + \epsilon$

- Bayesian model
- Or, use a Random Forest (RF) model.

1) $SFA \sim Anc + E + (1 \mid Deme)$ ← grouping effect depen on how the data wer collected.

$\overset{?}{Pop}$

↓ Deme? Admixture? ...

↓

- genetic control of Qmat
- Random effect of Pop.

Depends on data format & my question. Need to think on Then may be an interesting questio

1) Basic model:

$Y \sim Anc. + E + \epsilon$

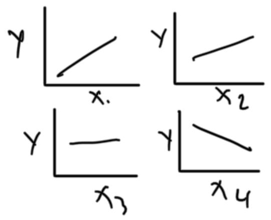
if this can be done in Bayesian or RF, do it.

↓

- How to include genetic Control w/ RF models?
- Ask Dominik

Fig 2

Plot conditional effects from model 1



} etc. · main takeaway figure.

Table 1

model output of Basic model (if using Bayesian)

[illegible]

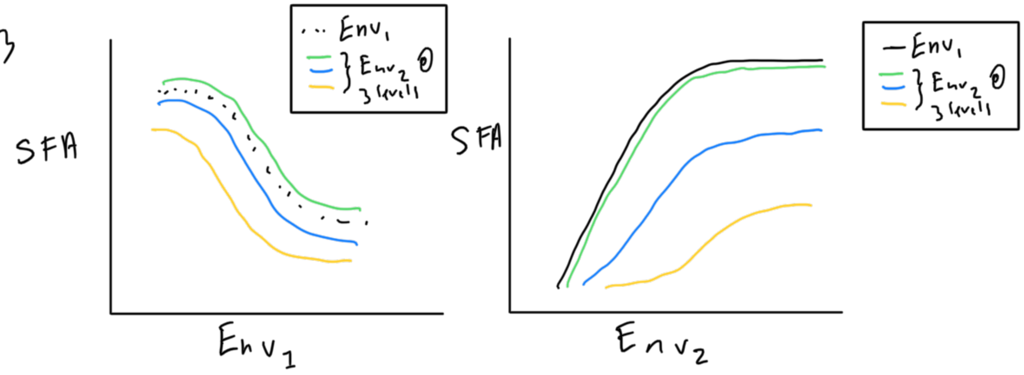
Interactions of Climate

- There are certainly interesting interactions. The idea is to explore them in some depth. Do the interactions explain more variance than the single terms?

- I don't think I can fit a full interaction model. Too many parameters. I could fit an interaction model to the subset of params that explain most variance & explore them.

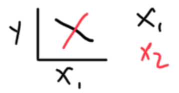
- If you take the Bayesian route plot interactions as terms or conditional effects.

Fig3



- Some visual exploration of interactions

- Would be great if some interactions were negative!



- conditional effect

Bayesian-Only Approach

- Fit All x's & genetic controls

- Take top 3 variables & fit interaction model

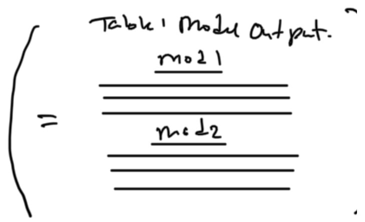
Visuals Outline

Fig 1. Sample & Exploration fig.

Fig 2. main effects plot

Fig 3. Interactions plot

Table 1. main effects table + interaction effect table



RF Approach

- Fit RF model to X's + genetic controls
- Take top 3 vars & fit Bayesian interaction model

Visuals Outline

Fig 1. Sample & Exploration fig.

Fig 2. RF output (RMSE + Cond. effs)

Fig 3. Interaction models

Table 1. Interaction model output

Pros / Cons of RF vs Bayesian

- Not sure beyond superficial ideas.

- Read &/or talk to Dominik
- Need to fit the model, first. then determine pro/cons.

Validation

Take wild seeds from populations that span the top env. gradient. If the gradient is easy to replicate in a chamber, do so. This will act as form of validation for the statistics & it connects to the mechanistic hypothesis that %SFA is under local selection for seed germination.