2018 General Notebook

Author: Andrew D. Nguyen, Evolutionary Physiologist

Affiliation: University of Florida, Department of Entomology and Nematology

Contact: anbe642@gmail.com

Date started: 2018-01-01

Date end (last modified): 2018-12-31

Introduction:
Notebook for 2018 new year. It'll log the rest of my dissertation, post doc projects, meetings, papers I've read, and general project ideas.

List of projects and description

- Hsp rxn norm: Understanding how the local thermal environment shapes thermal tolerance and stress response (using Hsps as a proxy for stress) in forest ants of the genus *Aphaenogaster*. CTmax and rxn norm of Hsp expression measured across forest ants from Fl to Maine.
- Range limits: Identifying the factors/forces that set range limits in common forest ants (*Aphaenogaster picea*). Modelling + measured their cold physiology in forest ants of Maine and Vt.
- Thermal niche paper: Collaborative paper understanding how the environment shapes the ability to withstand cold and hot temperatures. In field and in a common garden, we measured upper and lower thermal limits of ants from GA-Maine (2 species).
- Stress in nature: Are ants stressed under experimental warming that projects climate change? Ants were collected from warming chambers (0-5 C increase from ambient) and we measured their stress response.
- Biological rhythms in *Rhagoletis*: Determining the relationship between behavioral rhythms in adult *Rhagoletis* and diapause exit timing + depth(eclosion and mass specific metabolic rate).
- *Rhagoletis* diapause exit: Determine the physiological parameters that lead to divergent adult emergence patterns between two host races.
- *Rhagoletis cerasi* transcriptome: Determine the adaptive shifts in the transcriptome relating to seasonal timing in low and high altitude populations.
- Proteome stability project in *Drosophila melanogaster*: Determine the physiological tactics at the molecular level that underlie differences in thermal traits and whether they've been shaped by selection at a broad scale.

Table of contents (Layout follows Page number: Date. Title of entry)

- Page 1: 2018-01-01 . Yearly goals: recap from last year and this year's
- Page 2: 2018-01-01 . Evolution meeting prep/thoughts
- Page 3: 2018-01-02 . Range limits ms edits
- Page 4: 2018-01-02 . to do list this month, 2018 january
- Page 5: 2018-01-03 . range limits ms , stats for single regressions of CCRT on Tmin
- Page 6: 2018-01-05 . UPDATE: to do list this month, 2018 january
- Page 7: 2018-01-08 . To do list
- Page 8: 2018-01-08 . Removing DS_store in github repo
- Page 9: 2018-01-09 . reading biological rhythms workshop I: Introduction by **Kuhlman et al. 2007**; Cold spring harbor
- Page 10: 2018-01-09 . Updated to do list
- Page 11: 2018-01-10 . Reading **Wadsworth et al. 2013**; Journal of Evo Biol. Developmental exit to diapause and how it relates to the evolution of insect seasonality
- Page 12: 2018-01-10 . Meeting AMellison on range limits paper
- Page 13: 2018-01-12 . **Stålhandske et al. 2014**; post winter development of a butterfly
- Page 14: 2018-01-12 . working on hsp rxn norm paper
- Page 15: 2018-01-16 . meeting with KBora on qpcr troubleshooting
- Page 16: 2018-01-16 . Transcript abundance vs. Gene expression terminology
- Page 17: 2018-01-16 . Updated to do list
- Page 18: 2018-01-17 . ECB project idea
- Page 19: 2018-01-18 . prepping meeting with Dan on 2018-01-19
- Page 20: 2018-01-18 . starting a data science course
- Page 21: 2018-01-22 . Rhago nature figures for biological rhythms project
- Page 22: 2018-01-23 . Meeting with Dan, Tom about termination ms
- Page 23: 2018-01-23 . Reading Gunter et al. 2007, The road to modularity
- Page 24: 2018-01-23 . Strategies vs tactics (Mart R. Gross 1996)
- Page 25: 2018-01-24 . revisiting thermal niche paper
- Page 26: 2018-01-25 . Phenotypic Prediction Workshop 2018
- Page 27: 2018-01-26 . Meeting with Dan
- Page 28: 2018-01-30 . Readings in Rhagoletis, pop gen, directions of gene flow
- Page 29: 2018-01-31 . Meeting with Dan , MR trajectories paper
- Page 30: 2018-02-01 . follow up analysis; partial correlation of parameters
- Page 31: 2018-02-07 . Prep meeting with Dan
- Page 32: 2018-02-12 . To do list
- Page 33: 2018-02-13 . Re-analysis of hsp rxn norm proj
- Page 34: 2018-02-14 . Update to do list
- Page 35: 2018-02-15 . Paper reading: Melo et al. Ann rev; Modularity Genes Development and Evolution
- Page 36: 2018-02-21 . Updated to do list
- Page 37: 2018-02-22 . Flow of ideas for the discussion of the hsp rxn norm ms
- Page 38: 2018-02-27 . Reading Stinchcombe et al. 2010; *Evolution*, ACROSS-ENVIRONMENT GENETIC CORRELATIONS AND THE FREQUENCY OF SELECTIVE ENVIRONMENTS SHAPE THE EVOLUTIONARY DYNAMICS OF GROWTH RATE IN IMPATIENS CAPENSIS
- Page 39: 2018-02-28 . Meeting with Tom, diapause exit ms
- Page 40: 2018-03-01 . discussion, hsp rxn norm ms
- Page 41: 2018-03-02 . dissecting out some abstracts
- Page 42: 2018-03-06 . Updates and proteostasis project development/ideas

- Page 43: 2018-03-09 & 2018-03-12 . Proj updates and proteome stability proj development
- Page 44: 2018-03-16 . Flow of ideas for evolution talk
- Page 45: 2018-03-22 . re-analysis of diapause exit in rhago
- Page 46: 2018-03-23 . Meeting with dan, plan of actions
- Page 47: 2018-04-02 . Redundancy analysis
- Page 48: 2018-04-17 . notes; rhagoletis brain transcriptome
- Page 49: 2018-04-18 . update and biological rhythms notes
- Page 50: 2018-04-24 . Reading *R. cerasi* papers
- Page 51: 2018-04-25 . Biological rhythms thoughts
- Page 52: 2018-05-01 . project update
- Page 53: 2018-05-18 . status update
- Page 54: 2018-05-18 . 2018-05-18. ECHO app, by Hannah de Los santos ; finding circadian rhythms with extended harmonic oscillators
- Page 55: 2018-05-21 . Meeting with Tom
- Page 56: 2018-05-22 .Circadian rhythm talk by Katja Lamia, scripts research
- Page 57: 2018-05-25 . Meeting with Gragland
- Page 58: 2018-05-29 . Set of tasks for HChu
- Page 59: 2018-06-04 . meeting wth Gragland, cerasi data filtering and future analyses
- Page 60: 2018-06-04 . Hchu list of projects
- Page 61: 2018-06-01 .meeting with Dan
- Page 62: 2018-06-05 . kick starting proteome stability project in Hahn lab
- Page 63: 2018-06-07 . thoughts on network analyses cerasi dataset
- Page 64: 2018-06-08 . Notes on network analyses
- Page 65: 2018-06-08 . Meeting with TPowell
- Page 66: 2018-06-12 . to do list for Hchu
- Page 67: 2018-06-25 . updated to do list
- Page 68: 2018-06-27 . Range limits writing Notes
- Page 69: 2018-06-28 . notes on messing with hipergator
- Page 70: 2018-06-28 . Meeting with Ruchir and training on ultracentrifuge
- Page 71: 2018-06-29 . Working on hipergator
- Page 72: 2018-07-02 . Writing strategies for range limits ms
- Page 73: 2018-07-03 . Different strategies for constructing weighted co-expression networks in cerasi dataset
- Page 74: 2018-07-03 . Range limits ms: re-analysis with MAT
- Page 75: 2018-07-12 . Meeting with TPowell diapause exit ms
- Page 76: 2018-07-16 . Problem with trikinetics computer: lost data
- Page 77: 2018-07-17 . Meeting with Gragland, networks
- Page 78: 2018-07-18 . SHC range edge adaptation ms
- Page 79: 2018-07-23 . Paper notes: Salachan and Sørensen 2017, JEB
- Page 80: 2018-07-25 . Proteome stability idea dump
- Page 81: 2018-07-31 . Updated to do list
- Page 82: 2018-08-08 . Circadian rhythm thoughts, data analysis
- Page 83: 2018-08-09 . Detecting rhythms in R
- Page 84: 2018-08-13 . Rhagoletis summer collections
- Page 85: 2018-08-14 . Descrepancies on rhagoletis sampling for biological rhythms
- Page 86: 2018-09-03 . Reading Murren et al. 2015; Constraints on the evolution of phenotypic plasticity: limits and costs of phenotype and plasticity
- Page 87: 2018-09-13 . Rhagoletis field collection notes 2018-08-29
- Page 88: 2018-10-26 . Thoughts on amnnat revisions
- Page 89: 2018-10-30 . more thoughts on amnat discussion tweaks
- Page 90: 2018-11-05 . running stuff on hipergator
- Page 91: 2018-11-06 . comparing qgraph with igraph (mainly speed)

- Page 92: 2018-11-09 . git version control on hipergator (computer cluster)
- Page 93: 2018-11-27 . Montanucci et al. 2011, MBE
- Page 94: 2018-11-30 . Basic wgcna code - co expression networks
- Page 95: 2018-12-06 . Meeting with Dhahn
- Page 96: 2018-12-07 . Lab meeting reading : Johnson et al. 2018

Page 1: 2017-01-01. Yearly goals: recap from last year and this year's

2017 recap of goals

1. Submit and Publish 3 manuscripts: range limits, Hsp rxn norm, and multiple stressors (also thermal niche paper).
    - Published multiple stressors ms
    - Still ned to submit and publish 3-5 other ms:
        - proteome stability proj
        - range limits
        - hsp rxn norm
        - stressed in nature (SHC is lead)
        - thermal niche paper with Lacy
2. Get a post doc. This'll probably involve learning a new study system. And also sending out tons of applications.
    - at UF! Yay!
3. Learn and build a shiny app.
    - did not do this at all
4. Learn and become more proficient in statistics (Machine learning?, Baysian, predictive modelling, mixed effects modelling, eigentensor analyses).
    - Quantitative genetics: more statistical genetics
        - just dabbled in all of this, but didn't really learn much more
        - Instead, I learned more about time series analyses, so that is something statistical
5. Form new collaborations? It'd be awesome to work with Brent Sinclair, Brent Lockwood, Joel Kingsolver, Caroline Williams, Jon Stillman, Alex Gunderson.
    - didn't form any new collabs with any of these people
    - formed one with European colleague studying temperature adaptation in an alpine ant
6. Participate in a meta-analysis? Would be cool.
    - Dan and I are coming up with a meta analysis related to biological timing or diapause
7. Learn more physiology: Q10, metabolism related topics, lipid membranes, metabolites.
    - Learned some respirometry!

2018 year goals

1. Submit and publish range limits, hsp rxn norm, and thermal niche papers.
2. Build a data science course
3. Solidify meta analysis ideas and start the project

4. Learn and become more proficient at analyzing biological rhythm data; analyze Rhagoletis biological rhythm data to a point where I have a cool story to tell.
5. Start working on european cornborers:
    - Learn rearing, diapause biology (induction, termination)
    - experiment on behavioral rhythms
    - do experiments to understand the molecular basis of behavioral rhythms, specifically focusing on period
6. Start constructing a teaching and research statement
7. Apply for and secure external funding

Page 2: 2018-01-01. Evolution meeting prep/thoughts

Presenting poster on Rhagoletis and talk on thermal adaptation in Aphaeno

thoughts on the talk - I can draw from all of my projects to construct a nice story. I want to give my talk in the evolutionary physiology symposium (s69). The flow of the story will be from field metrics of stress responses due to experimental warming and then dissecting out the possible mechanisms of those stress responses, pitched in terms of physiological strategies organisms can adopt. I want to shift away from this gene does this towards a more complete understanding of what the genes tell us about the whole organism and what they're doing!

Title:

- Physiological strategies of temperature adaptation in ants
- Responding to environmental change will require temperature adaptation involving tolerance
- Adaptive responses to environmental change will require physiological shifts in...
- Evolutionary innovations to environmental change in the common woodland ant

1. Stressed in nature - yes! Ants show a stress response to experimental warming.
    - this'll highlight the field relevance and impact !
2. Are these differences due to adaptations? If so, what kinds? And what does this response mean?
3. Ants can withstand heat stress by resistance and/or tolerance
4. Common garden experiments:
    - measuring the stress kinetics show that protection and tolerance , but not basal gxp
    - under acclimation, yes basal gxp can change
5. punchline, these types of adaptations are unique and can only be detected with a rxn norm approach, because shifts in basal are found in the majority of other studies.

Potential opening:

I'm broadly interested in how organisms operate. And we can access how organisms operate by perturbing them and observing them under different contexts. For example, one big experiment that is happening right now is climate warming. So the main question in our research group is how will organisms respond to climate warming? ...Chamber stuff...source of selection....what is the actual source of selection? does it match up? Phylo context of ctmax...mechanisms behind variation in CTmax ...phys strategies (resistance and tolerance)...

reminders from last year:

Organize evolution and biological rhtyhms meeting (tell Dan about time line)

- Evolution 2018: Montpillier France;
    - August 19-22
    - Post doc travel grant: http://evolutionmontpellier2018.org/travel-grants
        - Deadline is January 17th
- Biological rhythms meeting 2018: https://srbr.org/meetings/upcoming-meeting/registration/
    - May 12-16 2018
    - registration:
        - early(nov 15-march 1) reg non member - **$725** ; **$425 member**
        - late (march 2 to onsite) reg non member - **$825**; **$525 member**
        - early(nov 15-march 1) post doc non member - **$475** ; **$425 post doc member**
        - late (march 2 to onsite) post doc non member - **$575**; **$525 post doc member**

2017-12-22. actual meeting

Look into Cold spring harbor protocols * set up trikinetics * data analysis * data handling

Basically, look at protocols online, actually check papers that have similar data structure like ours. Look up drosophila work.

Analyze existing, publicly available datasets to see if we can recapitulate results.

How do we separate signal from noise?

Page 3: 2018-01-02. Range limits ms edits

Addressing some ms edits here:

from SHC in results section-*Determining the relationship between cold-tolerance and climate*:

> You need to test how much of the variation in Tmin is explained by the combination of Tmax and seasonality.

Not sure if this is right to do, but it seems better to just state how Tmin is correlated with Tmax and seasonality.

correlation between tmin with Tmax and sd

```
cor.test(dbio2$SD,dbio2$Tmin)

    Pearson's product-moment correlation

data:  dbio2$SD and dbio2$Tmin
t = -22.766, df = 100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9422686 -0.8773033
sample estimates:
       cor
-0.9155697
```

```
cor.test(dbio2$Tmax,dbio2$Tmin)

    Pearson's product-moment correlation

data:  dbio2$Tmax and dbio2$Tmin
t = 2.3546, df = 100, p-value = 0.0205
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.03633252 0.40558545
sample estimates:
     cor
0.229188
```

take home: Tmin has high correlation with SD and not Tmax

but i'll do the regression here to show it:

```
summary(lm(Tmin~Tmax+SD,data=dbio2))

Call:
lm(formula = Tmin ~ Tmax + SD, data = dbio2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7794 -2.3577 -0.8954  1.5918 16.5544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.3238275 11.3484853   7.783 6.98e-12 ***
Tmax         0.6692606  0.0359986  18.591  < 2e-16 ***
SD          -0.0418123  0.0008353 -50.058  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.935 on 99 degrees of freedom
Multiple R-squared:  0.964, Adjusted R-squared:  0.9633
F-statistic:  1325 on 2 and 99 DF,  p-value: < 2.2e-16
```

Dan wants G matrix--- put in supplement:

	pretreat_Temp-5	pretreat_Temp0	pretreat_Temp25	pretreat_Temp5
pretreat_Temp-5	3328.40	-2175.66	-5635.48	287.85
pretreat_Temp0	-2175.66	15856.71	-16232.99	-8876.65
pretreat_Temp25	-5635.48	-16232.99	38801.25	17335.45
pretreat_Temp5	287.85	-8876.65	17335.45	24394.16

2018-01-03 entry

```
% * <shelmscahan@gmail.com> 2017-12-26T17:27:46.052Z:
% The fact that there is a threshold does not imply that environmental change is
steep.  Even if a change is very very gradual, if an organism has a lower limit,
at some point they will hit it.  You don;t even have an indirect density
indicator of this (ie they stay really numerous right up to the boundary until
they suddenly drop out).  To say the gradient is steep would mean actually
mapping out what happens to environmental predictors across the transect from
present to absent.  You have not done this.  I have removed steep from this
paragraph (obviously if you are super opposed you could add it back over my
objection).
```

Makes no sense, this is the whole point of modelling exercise.

Page 4: 2018-01-02. to do list this month, 2018 january

1. range limits ms - send out by end of this week (jan 5)
2. hsp rxn norm ms (send out Jan 19?)
3. evolution conference submission of abstract, and also for travel grant jan 15
4. construct evolution talk, quick and dirty just to have the slides together
5. read more on time series analyses
6. biological rhythms conference - register as member and for conference just to get out of the way

---

Page 5: 2018-01-03. range limits ms , stats for single regressions of CCRT on Tmin

basal data

```
ksub<-subset(k.dat,k.dat$pretreat_Temp=="25")
ksub$coldplot<-max(ksub$treatment_recovery_s)-ksub$treatment_recovery_s
 str(ksub)
'data.frame':   21 obs. of  47 variables:
 $ Colony              : Factor w/ 21 levels "Avon","Avon_2",..: 21 5 8 4 14 13
1 2 9 3 ...
 $ Date                : Factor w/ 11 levels "","07-Jul-15",..: 3 4 4 5 6 6 7 7
8 9 ...
 $ Plot                : Factor w/ 17 levels "","19:01","19:03",..: 11 1 15 8 4
5 2 3 16 14 ...
 $ State               : Factor w/ 2 levels "Maine","Vermont": 1 1 1 1 1 1 1 1 1
1 ...
 $ County              : Factor w/ 5 levels "","Franklin ",..: 5 4 3 5 4 2 2 2 3
4 ...
 $ Town                : Factor w/ 17 levels "Avon","Bingham",..: 17 5 7 4 13 12
1 1 8 2 ...
 $ Lat                 : num  44.6 45 45.1 44.7 44.7 ...
 $ Lon                 : num  -69.3 -69.5 -68.6 -69.4 -69.8 ...
 $ Altitude_ft         : int  510 543 315 468 203 411 857 852 250 470 ...
 $ Temp                : num  20.6 22.5 27.7 23.9 23 23.6 23.9 24.6 18.8 20 ...
 $ Wind.Speed          : num  0 0 0.7 0 0 0 0 0 0 0 ...
 $ Humidity            : num  83.5 72.6 56.3 71.6 73.2 67.6 54.6 68.7 81.2 78
...
 $ Wind.Chill          : num  20.6 23.8 27.6 23.8 22.4 22.9 24 24.6 19.2 20.6
...
```

```
 $ Heat.Stress.Index   : num  20.9 22.2 31.9 25.7 22.8 22.1 24.9 28.6 18.7 20.5
...
 $ Barometric.Pressure : num  29.4 29.3 29.6 29.4 29.7 ...
 $ soil_temp           : num  14.5 15 16 15 17 15 15 14.5 15 15 ...
 $ canopy_photo        : Factor w/ 17 levels "","19:1 Avon",..: 10 16 14 7 17 4
2 3 15 13 ...
 $ Tree.Species        : Factor w/ 17 levels "","birch, conifers",..: 3 17 5 12
7 8 14 13 2 6 ...
 $ Understory          : Factor w/ 11 levels "","," new growth conifers",..: 1 5 7
2 2 1 1 6 8 1 ...
 $ Habitat.Photo.Number: Factor w/ 18 levels "","19:1 Avon",..: 10 17 14 7 18 4
2 3 15 13 ...
 $ Nest_substrate      : Factor w/ 10 levels "","Branch and leaf litter",..: 6 3
9 9 9 9 8 2 9 5 ...
 $ CTmax               : num  40.9 36.8 40.7 41.2 39.5 ...
 $ MAT                 : num  6.7 5.4 5.7 6.6 6.3 5.9 4.7 4.7 5.7 5.3 ...
 $ MDR                 : num  12.2 12.5 12.2 12.4 12.6 12.5 12.2 12.2 12 12.6
...
 $ ISO                 : num  0.29 0.29 0.29 0.3 0.3 0.3 0.29 0.29 0.29 0.29 ...
 $ SD                  : num  97.7 101.5 98.4 98.8 99.2 ...
 $ Tmax                : num  26.8 26 26.1 26.9 26.8 26.2 24.9 24.9 25.9 25.9
...
 $ Tmin                : num  -13.9 -16.3 -15.1 -14.4 -15 -15 -16.4 -16.4 -14.8
-16.4 ...
 $ TAR                 : num  40.7 42.3 41.2 41.3 41.8 41.2 41.3 41.3 40.7 42.3
...
 $ TWQ                 : num  2.1 0.6 1.2 1.9 1.6 1.3 0.1 0.1 1.3 0.7 ...
 $ TDQ                 : num  -5.1 -6.9 -6.2 -5.4 -5.8 -6 -7.4 -7.4 -6.1 -7 ...
 $ TwarmQ              : num  18.9 18 18 18.9 18.6 18.1 17 17 18 17.8 ...
 $ TminQ               : num  -6.5 -8.4 -7.6 -6.8 -7.1 -7.3 -8.7 -8.7 -7.4 -8.4
...
 $ AP                  : int  1094 1065 1046 1069 1057 1106 1102 1102 1079 1052
...
 $ PWM                 : int  116 109 113 112 108 113 113 113 118 106 ...
 $ PDM                 : int  73 67 75 69 68 73 69 69 79 64 ...
 $ PSD                 : int  12 12 12 12 11 10 12 12 12 13 ...
 $ PWQ                 : int  316 298 303 306 297 310 305 305 316 291 ...
 $ PDQ                 : int  238 217 230 226 224 244 231 231 242 211 ...
 $ PwarmQ              : int  262 280 258 265 271 279 296 296 257 286 ...
 $ PminQ               : int  254 237 257 242 237 254 243 243 270 229 ...
 $ pretreat_Temp       : int  25 25 25 25 25 25 25 25 25 25 ...
 $ pre                 : num  0 0 0 0 0 0 0 0 0 0 ...
 $ treatment_recovery_s: num  577 624 745 816 724 ...
 $ diff                : num  -577 -624 -745 -816 -724 ...
 $ coldplot            : num  541 494 373 302 394 ...
 $ hard.zero           : num  -153 -335 -386 -309 -264 ...
```

basal regression

```
summary(lm(coldplot~Tmin,data=ksub))

Call:
lm(formula = coldplot ~ Tmin, data = ksub)

Residuals:
    Min      1Q  Median      3Q     Max
-333.63  -82.44   -5.63   90.37  266.36
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -889.43     413.86  -2.149  0.04473 *
Tmin          -89.93      27.62  -3.256  0.00416 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141.1 on 19 degrees of freedom
Multiple R-squared:  0.3581,    Adjusted R-squared:  0.3244
F-statistic:  10.6 on 1 and 19 DF,  p-value: 0.004157
```

hardening data

```
zero<-subset(k.dat,k.dat$pretreat_Temp=="0")
ksub$hard.zero<-ksub$treatment_recovery_s-zero$treatment_recovery_s
nbn3<-ddply(ksub,.
(Colony,pretreat_Temp),summarize,CCRT=mean(hard.zero),Tmin=mean(Tmin))
str(nbn3)
'data.frame':   21 obs. of  4 variables:
 $ Colony       : Factor w/ 21 levels "Avon","Avon_2",..: 1 2 3 4 5 6 7 8 9 10
...
 $ pretreat_Temp: int  25 25 25 25 25 25 25 25 25 25 ...
 $ CCRT         : num  41 -72.3 -346.4 -309.2 -334.8 ...
 $ Tmin         : num  -16.4 -16.4 -16.4 -14.4 -16.3 -15.2 -13.6 -15.1 -14.8
-13.5 ...
```

hardening regression

```
summary(lm(CCRT~Tmin,data=nbn3))

Call:
lm(formula = CCRT ~ Tmin, data = nbn3)

Residuals:
    Min      1Q  Median      3Q     Max
-402.13  -96.13   26.98  137.37  236.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1346.45     516.54  -2.607   0.0173 *
Tmin           73.85      34.47  -2.142   0.0453 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 176.1 on 19 degrees of freedom
Multiple R-squared:  0.1945,    Adjusted R-squared:  0.1522
F-statistic: 4.589 on 1 and 19 DF,  p-value: 0.04534
```

SHC comments to address on % hardening improvement from 25 C pre-treatment temp:

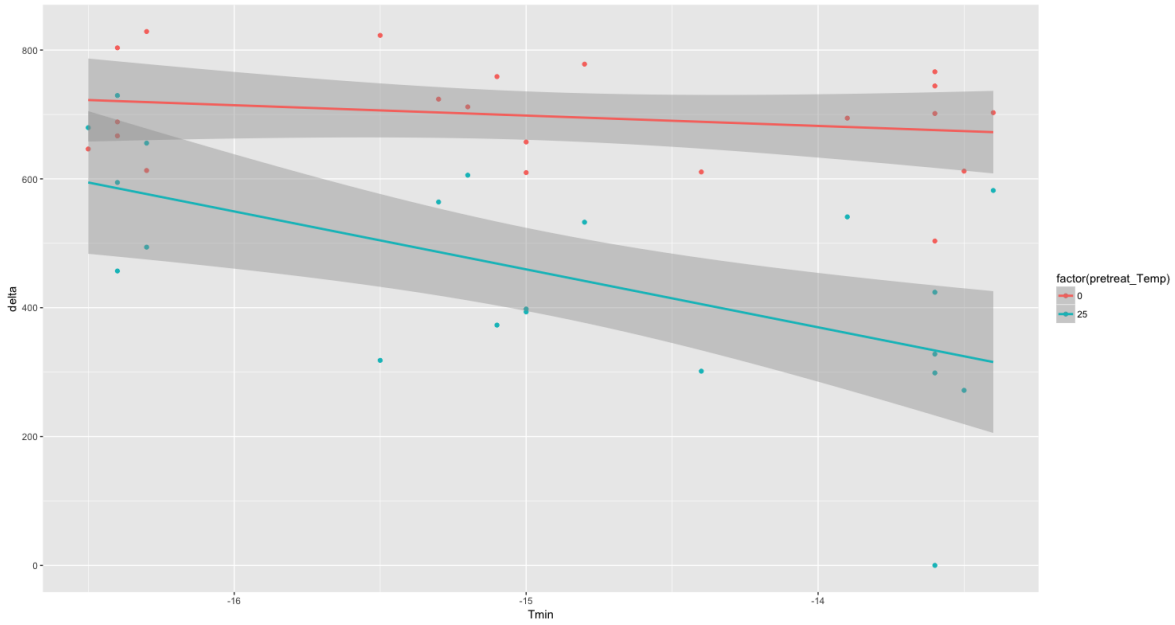
```
% What % improvement do your hardening results represent?
This would tell you whether 4.8 minutes is really too short to be able to detect
hardening given the inherent measurement error in such experiments.
If it is 10%, that would be almost a minute improvement, which seems quite
feasible.
```

how to calculate this...

```
###stitching together 0 and 25 C pre-treatment
nn<-rbind(zero,ksub)
#calculating the difference in the max CCRT (for the data which includes 0) from
all other measurements of CCRT

## this is to put everything on the same scale and have higher values as more
cold tolerant
nn$delta<-max(nn$treatment_recovery_s)-nn$treatment_recovery_s
```

lets look at the data:

```
ggplot(nn,aes(x=Tmin,y=delta,colour=factor(pretreat_Temp)))+geom_point()+stat_sm
ooth(method="lm")
```




```
##convert long to wide format to get  0 and 25 as diff columns
nn2<-melt(nn,id.vars=c("pretreat_Temp","Colony"),measure.vars=c("delta"))
wide.nn<-spread(nn2,pretreat_Temp,value)[-19,-2]


##dataset
knitr::kable(wide.nn)
```

	Colony	0	25
1	Avon	688.4167	729.4167
2	Avon_2	666.7500	594.4167
3	Bingham	803.4167	457.0000
4	Burnham	610.7500	301.5000
5	Cambridge	828.7500	494.0000
6	Canaan	711.7500	605.7500
7	cent2	503.5000	298.7500
8	Greenbush	758.7500	373.0000
9	Greenfield	778.0833	532.7500
10	Hampden	612.0000	271.7500
11	Monroe	702.7500	582.0000
12	New Portland	613.0000	655.5000
13	New Sharon	609.7500	398.0000
14	Norridgewock	657.2500	393.5000
15	Palmyra	822.7500	318.2500
16	Phillips	646.4167	679.5000
17	Saponac	723.7500	564.0833
18	ted1	766.2500	328.0000
20	ted8	701.5000	424.0000
21	Unity	694.2500	541.0000

calculating % improvement

so this is the calculation of the amount of seconds changed, scaled by the total time/value at 25 C to calculate improvement of hardening ability.

% percent changed

```
round(abs(wide.nn$`0`-wide.nn$`25`)/wide.nn$`25`,3)*100
 [1]   5.6  12.2  75.8 102.6  67.8  17.5  68.5 103.4  46.1 125.2  20.7   6.5
 53.2  67.0
[15] 158.5   4.9  28.3 133.6  65.4  28.3
```

range of % changed

```
range(round(abs(wide.nn$`0`-wide.nn$`25`)/wide.nn$`25`,3)*100)

 4.9 158.5
```

lets look at this in terms of minutes instead of seconds

```
##lets look at the amount of time that has changed itself

wide.nn$minutes_change<-abs(wide.nn$`0`-wide.nn$`25`)/60
#range of minutes
```

	Colony	0	25	minutes_change
1	Avon	688.4167	729.4167	0.6833333
2	Avon_2	666.7500	594.4167	1.2055556
3	Bingham	803.4167	457.0000	5.7736111
4	Burnham	610.7500	301.5000	5.1541667
5	Cambridge	828.7500	494.0000	5.5791667
6	Canaan	711.7500	605.7500	1.7666667

	Colony	0	25	minutes_change
7	cent2	503.5000	298.7500	3.4125000
8	Greenbush	758.7500	373.0000	6.4291667
9	Greenfield	778.0833	532.7500	4.0888889
10	Hampden	612.0000	271.7500	5.6708333
11	Monroe	702.7500	582.0000	2.0125000
12	New Portland	613.0000	655.5000	0.7083333
13	New Sharon	609.7500	398.0000	3.5291667
14	Norridgewock	657.2500	393.5000	4.3958333
15	Palmyra	822.7500	318.2500	8.4083333
16	Phillips	646.4167	679.5000	0.5513889
17	Saponac	723.7500	564.0833	2.6611111
18	ted1	766.2500	328.0000	7.3041667
20	ted8	701.5000	424.0000	4.6250000
21	Unity	694.2500	541.0000	2.5541667

ok, the range in minutes?

```
range(wide.nn$minutes_change)
[1] 0.5513889 8.4083333
```

taking this paragraph out of discussion

```
The patterns of cold hardening observed in this study suggest that \textit{A.
picea} are exposed to conditions in the field that are at or near their
physiological limits.  \textit{Aphaenogaster picea} colonies were the most cold
tolerant at 0  \degree C pre-treatment temperatures, with a substantial decline
in hardening ability when exposed to -5 \degree C, yet, soil temperatures in
Maine at ant-wintering depths typically range from 0  to -5  \degree C, with
even lower temperatures possible on a short-term basis (Schaefer et al. 2009).
Maintaining performance below 0  \degree C pre-treated temperatures may impose
significant physiological challenges due to tissue damage from freezing
selecting for both constitutive and inducible mechanisms to counter stressful
mean temperatures and pulses of more extreme conditions (Bale 2002).
```

2018-01-04: not sure if I want to expand on this idea, but leaving it here

```
negative correlation in aphaenogaster; 0 was the optimal performance for many
colonies, anything past that might be a constraint because tissues begin to
become damaged from freezing.
```

Page 6: 2018-01-05. UPDATE: to do list this month, 2018 january

1. range limits ms - send out by end of this week (jan 5)
    - done, tweaked results little, major tweaks discussion, rewrote abstract
2. hsp rxn norm ms (send out Jan 19?)
    - work on next, what needs to be done?
    - From 2018_nb_page108:
        - Rearranged figures; so results need to be rewritten; introduction needs tweaking; discussion needs to be finished
            - Figure layout:
                - Predictions figure: Fold induction vs temperature for each parameter
                - Phylogenetic analysis: ancestral trait reconstruction
                - Non-phylogenetic controlled analysis: CTmax vs environment
                - CTmax vs hsp params
                - hsp params vs habitat type
                - Methods figure: example of fitting boltzmann to hsp gxp
        - SHC has comments on results
3. evolution conference submission of abstract, and also for travel grant jan 15
    - submitted both!
4. construct evolution talk, quick and dirty just to have the slides together
5. read more on time series analyses
6. biological rhythms conference - register as member and for conference just to get out of the way

not done anything for 4-6.

Page 7: 2018-01-08. To do list

1. Set up weekly meeting with Dan
    - Fridays 3pm, starting next week
    - Dan says I need to understand the Rhagoletis system better
        - Dan will send out paper
2. Set up schedule with undergrad researchers
    - Cwu - try to set her up with Qinwen working on the ECB system ; can she get paid too ?
    - KLennon: Monday, Tuesday, Sund: 1-3pm; Thursday 10:30Am-12:30PM
    - Tat: M,W,F,Sat: 7-9PM; Thursday 9AM-11AM
    - Taariq : working on script to assemble trik data with master datasheet
3. Start working on Hsp rxn norm paper

Page 8: 2018-01-08. Removing DS_store in github repo

plaguing my existence... good solutions here

1. In terminal where project is located:

```
find . -name .DS_Store -print0 | xargs -0 git rm -f --ignore-unmatch
```

2. then add ".DS_Store" in gitignore, and save

Page 9: 2018-01-09. reading biological rhythms workshop I: Introduction by Kuhlman et al. 2007; Cold spring harbor

ref:
Kuhlman, S. J., Mackey, S. R., & Duffy, J. F. (2007). Biological Rhythms Workshop I: introduction to chronobiology. Cold Spring Harbor Symposia on Quantitative Biology, 72, 1–6. https://doi.org/10.1101/sqb.2007.72.059

This paper gives a timeline of chronobiology

a little bit of history or historical context

- An astronomer, Jean-Jacque D'Ortous deMairan (1729) made one of the first observations that organisms exhibit cyclical behavior: leaf movements in heliotrope plants continue in constant darkness. (Free run experimetn)
- Franz Halberg 1959 coined the term *circadian*, circa = about, dies= day, to refer to daily rhythms that persist under free run.
- The most well known circadian rhythms occur in single cells, whereby transcription and translation form feedback loops.
- Biological rhythms are suppose to be or presumed adaptive.
- It was Bunning 1935 who first discovered that biological rhythms are heritable and have a genetic basis; daily rhythms could be extended to seasonal rhythms.
- Ronald Konopka and Seymour Benzer 1971 mutgenized fruit flies to understand the rhythnsm of pupal eclosion and locomotor activity (have to read this paper!).These flies could be non-rhythnmic, or rhythmic over short or longer time scales. All phenotypes were found to be within the Period gene (region/locus).
- Ok, wow, we can manipulate the timing of organisms, but what are some of the functional consequences?
    - Karl von Frisch and Beling showed that bees use their clocks to visit flowers at the time of the day when the flower opens
    - Kramer showed that birds' migration patterns are rhythmic
    - DeCoursey et al. 2000 show that chipmucks modulate rhythms to avoid predation
    - Steve Reppert showed that monarch butterflies likely use biological rhythms to initiate migration and track seasonal changes

Some terminology

- "Black-box" experiments are used to probe mechanisms of clock function (Moore-Ede et al. 1982)
    - no clue what this means, wtf
- Overt outputs - observable, measurable rhythmic outputs (leaf movement)
- a rhythm is considered to be circadian if the oscillation has a 24 hour period(approx)

    - but why... the main argument is the relative role of env vs genetics on biological rhythms

Free running period

- process of organism's rhythm syncrhonizes to external cycling cue is known as entrainment
- Free run is the cyclical behavior in the absence of the cue (so darkness if light entrained)
- Diurnal orgs have shorter free run (faster clock) under high light intensity than low light intensity
    - citation?
- Free run in diurnal orgs found by Jurgen Aschoff, known as, "Aschoff's Rule" (Aschoff and Wever 1962)
- The degree of free run change is influneced by prior experience to entrainment cue (light), known as *aftereffect*. fuck these words, wtf? Keep the terms the same among disciplines --- acclimation !

Two properties of circ rhythms that are often focused in experimental studies

1. ability of rhythm to be reset, or phase-shifted, by transient exposure to time cues such as light
2. temperature-compensation

bad writing....very little clue about what they're talking about. Why would this matter? Presenting a question and hypothesis helps.

Phase shifting and phase-response curve

This is a manipulation aimed at resetting the clock such that the rhythms differ between manipulated vs unmanipulated. If you do this with an external cue (light), then it can show that the cue and changes to that cue is important for setting biological rhythms.

- If you set this manipulation at different times of the day, then it'll show that sensitivity to resetting differs at differnt parts fo the cycle (Hastings and Sweeney 1958)

Possible that the clock doesnt reset at all.

Temperature-compensation

Expectation: clock should not change despite changes in ambient temperature

What? this is garbage. One could expect this under the "hotter is better" hypothesis whereby biochemical reactions simply speed up as a function of temperature. So the clock should change by getting shorter andf aster.

uncited claim : freer un period changes little over a temp range

bad....

They need to parse out these ideas better. In the face of temperature changes, the organism can maintain their biological clock. For example, they can initiate behaviors at the same time of the day. This would be robustness of biological rhythms. But, temperature itself can act as an entrainment cue too.

Concept of Clock

Would have been nice to explain jargon before talking about it. Organizational problem here.

Chandrashekaran 1998 -

Clock - endogenous nature of rhythms (innate, rather than learned), a way of measuring time for the organism.

Bad definition. This should be described like any other trait, which has a genetic and environmental basis.

Thoughts

Aside from citations pointing to some cool studies, historical context, and defining some words (entrainment, free run), this paper doesn't have a good conceptual description of biological clocks. Why would phase shift matter, why would someone want to know that, how is it applicable to some greater concept of how organisms work? smh.

Page 10: 2018-01-09. Updated to do list

1. Read papers for discussion with Dan
    - still waiting for dan as to what they are
2. Set up schedule with undergrad researchers
    - Cwu - try to set her up with Qinwen working on the ECB system ; can she get paid too ? ---sent out coordinating email
    - KLennon: Monday, Tuesday, Sund: 1-3pm; Thursday 10:30Am-12:30PM -- got keys
    - Tat: M,W,F,Sat: 7-9PM; Thursday 9AM-11AM -- got keys
    - Taariq : working on script to assemble trik data with master datasheet (meeeting wednesday 10AM)
3. Start working on Hsp rxn norm paper
    - set up github repo, so data are accessible and can work on it from anywhere
    - still need to start....
4. thermal niche ms
    - Lchick to send me a version by end of week to edit
5. range limits ms
    - sent out,
    - submit soon? February...please!?
6. euro collab ms
    - wait for results of tests of selection, then help write results and discussion

---

Page 11: 2018-01-10. Reading Wadsworth et al. 2013; Journal of Evo Biol. Developmental exit to diapause and how it relates to the evolution of insect seasonality

ref: Wadsworth, C. B., Woods, W. A., Hahn, D. A., & Dopman, E. B. (2013). One phase of the dormancy developmental pathway is critical for the evolution of insect seasonality. Journal of Evolutionary Biology, 26(11), 2359–2368. https://doi.org/10.1111/jeb.12227

A little background: These authors are working on the european cornborer (ECB), where two strains differ in their seasonal emergence.

1. E strain = earlier emerging strain
    - this strain is bivoltine, so it has two generations per year
    - ID'd by PDD^S allele
2. Z strain = later emerging strain
    - this strain is univoltine, so 1 generation per year
    - ID'd by PDD^L allele

Main Question: What part of diapause is critical for seasonal timing? In this case, emergence.

Hypothesis: The timing of diapause termination causes the differences in emergence timing between E and Z strains.

Experimental approach:

- Induce diapause in 12:12 L:D at constant 23 C. (Glover 1992) for both strains, E and Z
- ECB diapauses as larvae; measured metabolic rates and modeled its trajectory with a function

I messed around with the function here. It is basically an exponential model with 3 parameters:

- a = presistent metabolic increase that defines the start of termination
- b = rate of metabolic increase during diapause termination
- c = initial metabolic rate during diapuse maintenance phase

Major results: diapause termination corresponds with metrics of emergence timing int he field (# breeding adults and pupae)

E strain, which has earlier emergence timing, has shorter time of diapause break (diapause termination) than Z strain. It looks like the difference amounts to 22 days (Z-E, 34-12).

Discussion points

- similar thing happens in swallowtail butterflies (*Papilio*) (Scriber & Ording 2005)
    - northern ones are univoltine with obligate pupal diapause and produce earlier summer flights
    - southern ones are bivoltine and more facultative, producing diapause and non-diapausing phenotypes

---

Page 12: 2018-01-10. Meeting AMellison on range limits paper

- In the CART model, scale the bioclim variables first.

For figure 3.

- For vertical variation, say constant for the loadings (panel b)
- panel a, additive response rather than additive variation ; because the sd doesn't change, the eigenvector remains constant
- for panel c,d ; more than just horizontal shift, both shift in level and in variance. You want to show a step function in the loadings
- simulate distributions and then compute the variance-covariance matrix and then plot the loadings
- Set up a new set of panels and set up the null expectations

2018-01-11 update: done here; look identical to my figures

For standardizing variables, seems like the consensus is no, unless you're comparing with other predictive models: 1,2

Take home:

think about how to write the patterns of perf curves and how it maps to the loadings. Look back at Kingsolver et al. 2015 and see how he writes it.

- Citizen science idea--I should write this out , later. But it deals with species distribution modeling. One problem you'll encounter is that there is error from citizens! You have to account for error of detection.
    - talk to matt fitzpatrick about proj
    - AMellison said it would be a good idea to apply for grant
- AMEllison is trying to push out genome paper with MLau

2018-01-12 update

other projects I'm on that I've forgotten about

- genome paper
- phylo paper with Bernice

Page 13: 2018-01-12. Stålhandske et al. 2014; post winter development of a butterfly

Reading to discuss with Dan.

reference:

Stålhandske, S., Gotthard, K., Posledovich, D., & Leimar, O. (2014). Variation in two phases of post-winter development of a butterfly. Journal of Evolutionary Biology, 27(12), 2644–2653. https://doi.org/10.1111/jeb.12519

Wow, awesome sentence:

> Fine- tuning of the life cycle may be particularly important to match the phenology of potential mates and resources as well as for optimizing abiotic conditions at eclosion.

Counter gradient variation is a term that has always confused me; wiki defines as:

> Countergradient variation is a type of phenotypic plasticity that occurs when the genetic components of populations cause phenotypic variation that opposes the phenotypic variation caused by an environmental gradient.

Almost not understandable and too much to unpack. It sounds more like a trait that varies along an environmental gradient that is counter to expectations. Anyway, to me, this is not a useful term at all because it muddles how we should think about the environment and how organisms can respond to it.

One prime example is growth across a lat gradient. Populations in the cooler condition (northern) that grow faster than warmer populations (southern) is considered countergradient variation because warmer temperatures are suppose to increase growth rates. But maybe, the env variation that growth rates are responding to is not solely warm temperatures.

Authors' def:

> Countergradient variation occurs when the inherent relationship between phenotype and environment given by a reaction norm is reduced or removed through genetic adaption across an environmental gra- dient (Conover & Schultz, 1995)

Totally diff definition than wikipedia. The concept is so abstract so that it is hard to interpret biologically. Angiletta's book doesn't have a good definiton.

This paper talks about other env variables as to explain the patterns described in countergradient variation....but if there are other env variables to consider other than temp, then that is cogradient variation (organisms responding with the environmental) ....see makes no sense.

> Populations that experience a shorter growing season or lower temperatures may compensate for this by a faster growth rate at any given temperature (Yamahira & Conover, 2002).

Citations for lat gradients:

> Intraspecific clines in life cycle traits often follow environmental gradients, such as temperature gradients along latitude (Bradford & Roff, 1995; Masaki, 1999; Gaston et al., 2008).

Paper is not really structured in a question-hypothesis framework. But they are interested in how the environment (latitude) shapes the variation in seasonal timing of these butterflies (*Anthocharis cardamines*).

Natural history of *Anthocharis cardamines*:

- one life cycle a year (univoltine); diapause
- lays eggs on buds and flowers of early flowering host plants (phenological specialist)
    - plants within the Brassicaceae
- Females emerge first on flowers, then males follow
- spring emergence is a function of both winter duration of diapause and the rate of post diapause development (they just want to study post witner development)
- found in europe, asia, and north africa

Experimental approach: common garden of pops collected in 6 populations spanning lat and measuring growth rates

- reared on garlic mustard
- collected eggs and placed in lab
- larvae hatched and reared to pupation in common garden
- simulated overwintering - 2 C 24 dark, 5 months

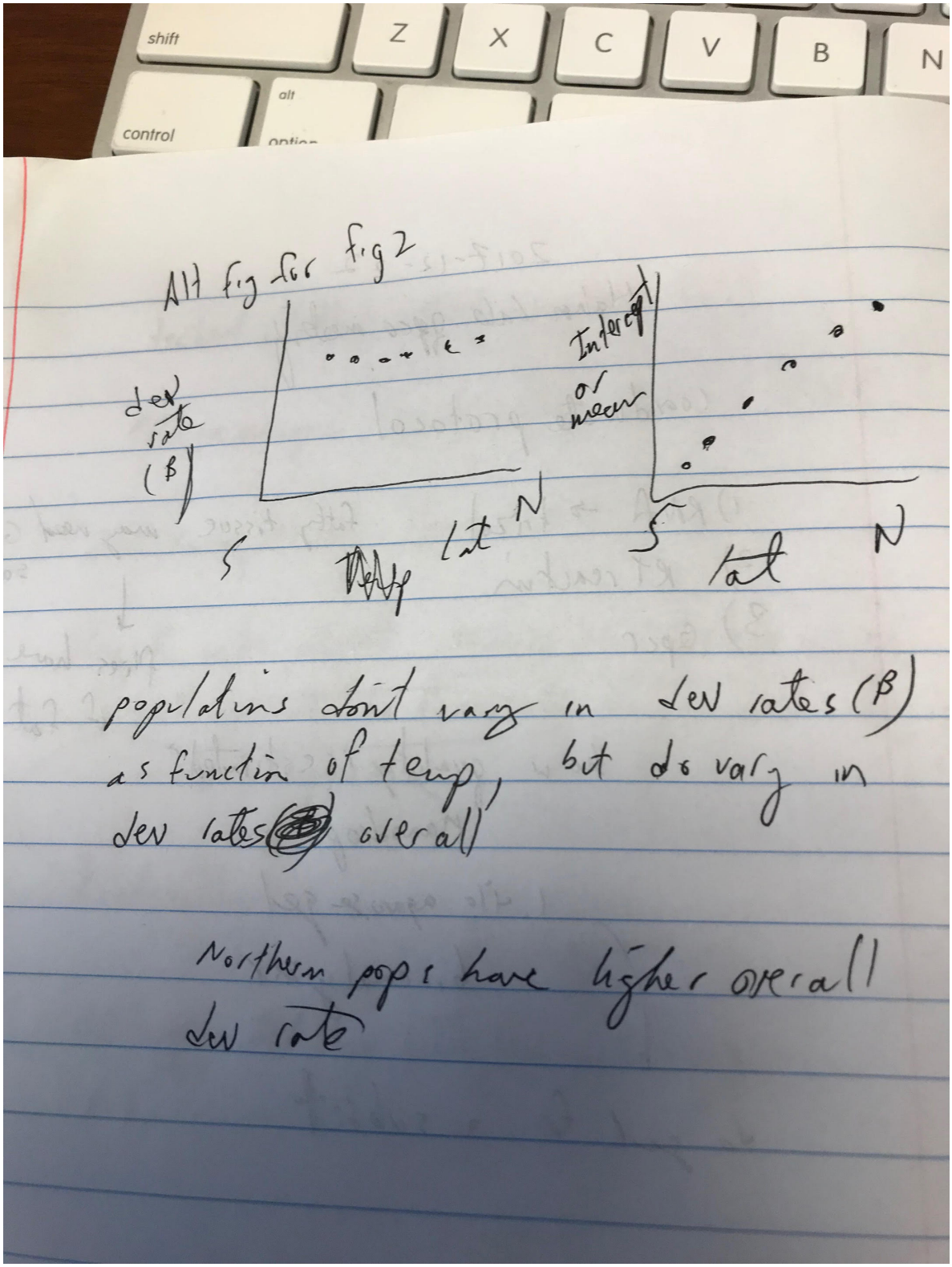
pupae then moved to 4 constant temperatures, post wintering after 5 months.

- pupae weighed every 48 hours and rate of weight loss was used as a proxy for rate of development
- stats mostly on females ; stats not explained very well.... what are they testing.

Results:

Table 1: Linear models, evaluating predictors that explain growth patterns. They only tested for local * temp interaction in the proportion variable. What about others? Ther perform model selection ? Also, what about local * initial pupal weight interaction?

How I'd show figure 2-



Northern pop have higher overall mean dev rate than southern populations. Prop in phase 1 didn't differ, it looks like. But in the model form table 1, there is a sig local X temp interaction. The y axis may be too large in units for us to see the interaction...

Table 2. local comparisons for each phase of growth; overall, most locals are diff

Fig 3. Weight loss is described as having 2 phases.

FIg 4. prop initial weight vs days for each time point. They could have displayed this better. Calculate the break point and then you'll have the time for each phase. (would be nice to have metabolic rate with these data, maybe have ot check other refs) Then test the effects of local + env(lat) on the timing of each phase.

Table 3: variance partitioning of growth for each phase, for each temperature level. First phase always explains more of the variation (> 50%!).

Thoughts/future directions

- Would have been cool if they measured life time reproductive success to determine fitness consequences of treatments.
- There needs to be a full factorial design on winter length and emergence temperatures to understand the relative importance on spring time emergence. Done is Rhagoletis for allele frequencies ; but would be nice to measure fitness
- Would have produced different plots to make it more understandable. Actually plot stuff across lat or the real env variables themselves.
- Figure 1. give this fig, you'd expect faster dev rates for sweden than uk due to the differences in the slopes of the lines. Sweden is steeper, selecting for a phenotype that chagnes in response to it.

Page 14: 2018-01-12. working on hsp rxn norm paper

- Mainly working on results and tidying up sentences.
- There are 2 abstracts, taking out the one that got voted out!

Abstract v2 Colonization into warm environments may involve mechanisms to cope with protein unfolding that is caused by high temperatures. These mechanisms include protection, tolerance, and resistance by heat shock proteins (Hsps), but the extent that contributes to physiological limits is poorly understood. We evaluated the relative importance of Hsp expression dynamics related to upper thermal limits (CTmax) of Eastern US forest ants in the genus Aphaenogaster whose geographical ranges encompass a climate gradient across two distinct forest types. Within this group, the evolutionary transition from closed canopy deciduous forests into open canopy pine forests occurred once and coincided with an increase in CTmax. Using a function-valued approach, we tested whether species divergence in CTmax between forest types corresponded with changes the Hsp expression kinetics that support the concerted roles of protection, tolerance and resistance. Protection was conferred with pre-emptive higher expression at the cooler end of the temperature gradient by slowing the overall expression rate. Greater tolerance was brought about by higher peak expression. Lastly, colonies with higher CTmax delayed their onset of expression, suggesting greater stress resistance to temperature perturbations. The upward extension of CTmax involved coordinated, adaptive, and dynamic shifts in the stress response that reflects better maintenance of protein homeostasis under temperature challenge and may have facilitated the diversification of forest ants into divergent thermal environments. These results lay out the expectation for the types of evolved mechanisms needed for surviving a warmer world in the future.

- I still need to tweak the other abstract and incorporate edits from Dan.
- Need to adjust results of topic sentences with this comment from SHC:

> It would be nice to have a statement of the objective and what you did to test it at the beginning of each results paragraph (To do X, we did Y). The methods exist, but they are at the end so your readers will not know what is going on.

- Took out this sentence in intro, leave it here:

> The protective effects of Hsps are exemplified by enhanced thermal tolerance in flies with higher experimentally increased copy number (Welte et al. 1993; Feder 1995).

Page 15: 2018-01-16. meeting with KBora on qpcr troubleshooting

KBora wants to quantify the *foraging* gene from ant brains of pogos

She has primers, but her reactions are coming up at around 30-40 CT (cycle thresholds). She wants to test out left over cDNA to see if primers work on my samples.

But lets check the melt curves: They look "good" , single peak. Primer only control wells don't have any amplification, which is good. Therefore, the good ones can be used for a standard.

Recommendations:

1. take 50 uL amouont of "good" samples (mixed)
2. pcr purify with qiagen kit
3. quantify with qubit
4. dilute, 1/100 (generally , you can be more precise), then dilute 1/10 serially (12 times, so 1 row on qpcr plate; leave last sample as water for control).

Kbora will send me melt curve pictures, amplification curves, and the datasheet. The goal is to estimate the efficiency of the primers. Once you have this value, you can use it for delta delta CT calculations of gene expression. You should also sequence the amplicon to make sure the sequence is what you think it is (map abi sequencing files to reference sequence in geneious).

Page 16: 2018-01-16. Transcript abundance vs. Gene expression terminology

Thinking about one of Dan's comments on using a different term than gene expression:

> I almost always say transcript abundance rather than expression because you don’t know whether you are seeing new transcripts being produced or just having transcripts hang around longer.

Neither term tries to distinguish between the two. Both sound right, but gene expression is used more. I think its silently known, but gene expression can be a function of new transcripts being produced and the degradation of transcripts, which ultimately influence abundance of it.

Doesn't seem like transcript abundance is a better term.

Page 17: 2018-01-16. Updated to do list

1. Read papers for discussion with Dan

refs:

- Wadsworth, C. B., Woods, W. A., Hahn, D. A., & Dopman, E. B. (2013). One phase of the dormancy developmental pathway is critical for the evolution of insect seasonality. Journal of Evolutionary Biology, 26(11), 2359–2368. https://doi.org/10.1111/jeb.12227

- Stålhandske, S., Gotthard, K., Posledovich, D., & Leimar, O. (2014). Variation in two phases of post-winter development of a butterfly. Journal of Evolutionary Biology, 27(12), 2644–2653. https://doi.org/10.1111/jeb.12519
- Ragland, G. J., Fuller, J., Feder, J. L., & Hahn, D. A. (n.d.). Biphasic metabolic rate trajectory of pupal diapause termination and post-diapause development in a tephritid fly. https://doi.org/10.1016/j.jinsphys.2008.12.013

Done, revisit before meeting to refresh. Also go over yearly goals with Dan.

2. Set up schedule with undergrad researchers
    - all done
3. Start working on Hsp rxn norm paper
    - sent out today!
4. thermal niche ms
    - Lchick to send me a version this week and make this top priority
5. range limits ms
    - sent out,
    - submit soon? February...please!?
6. euro collab ms
    - wait for results of tests of selection, then help write results and discussion

More things to do:

Once I get all of these ms out the way, I can focus more on future projects.

- Revisit data analysis in R. I still need to assemble dataset that merges master datasheet with trikinetics
    - I might need to parallel process because the data files are very large.
- Write up ideas about testing the molecular basis of eclosion shifts in ECB system so I can start working on it potentially.
- Start outlining online data science course I want to build.

Page 18: 2018-01-17. ECB project idea

Natural History

European cornborers one of the best examples of speciation. They've become reproductively isolated in time and this has produced two strains , Z and E. ECB become rep isolated by the number of generations that are fit into a year. Univoltine Z strain has a single generation in the middle of the summer, while the bivoltine E strain emerges earlier and then later in the season.

They diapause as larvae

A QTL analysis showed a suite of candidate genes related to circadian rhythms which may be responsible for the divergence of seasonal timing in ECB. In Dopman's talk at Entsoc, he mentioned that different strains mated at different times of the day. THe panel of genes incldues cry1, a few unknown contigs, and period.

In period, they inferred that one allele that interacts with tim may be responsible in the differences in post diapause development. There are two alleles:

1. G allele - associated with later post diapause development

2. T allele - associated with earlier post diapause development

(Are these alleles enriched in Z or E strains?)

The G allele codes for an **Alanine (A)**, while the T allele codes for an **Aspartic Acid (D)**. This is interesting because aspartic acid is charged and alanine is not, so there may be differences in the intra or inter molecular level interactions.

Question: How does the genetic variation in period ultimately cause divergence in post diapause development in the ECB system?

Hypothesis:

One larger hypothesis is that period sets the tempo of development in ECB; Z strains have slower tempos than E strains.

We can break this bigger hypothesis down to explanations about the biological rhythms part itself and the molecular basis of those rhythms.

1. Z strains have slower biological rhythms
    - Prediction: tau(time between peak bouts of activity that is cyclical) is shorter in Z strain than E strain.
2. The T allele (should be enriched in the Z strain) of period interacts more slowly with tim.
    - Predictions-
        - Binding affinity with tim is lower
        - It binds with more partners and has less opportunity to bind to tim
        - IT binds to the same partners, but longer, and has less opportunity to bind to tim
        - Its expressed differently--slowly (not expected since the SNP of interest is intergenic)

Experimental approach:

1. Trikinetics system- measure post diapause development in larvae and measure feeding activity in adults
2. Purify different periods and measure dissociation constants (Biacore?) with tim
    - Could use recombinant period protein with diff alleles, then do a pull down (immunoprecipitation) and measure the amount of tim (western blot)
    - Can do this in the present or absence of whole protein extracts from Z or E strains to determine the promiscuuity of period.
3. A series of immunoprecipitation experiments coupled with mass spec
    - Take 2 days worth of samples at ~2 hour intervals and immunoprecipitate period and measure what is attached to period with mass spec

Expected outcomes:

Biological rhythms

- Positive relationship between Tau (adults) and post diapause development (larvae)
- Shorter tau in Z strain than E strain
- Shorter tau in T allele than G allele in period

Tempo of molecular interactions

-

Big claim: The tempo of molecular clocks dictates broad scale physiological development that enables reproductive isolation within a season and has led to the emergence of new species.

Thoughts:

- I wonder how abundant period and tim are though; it might be difficult to isolate. Need to find antibodies
    - most antibodies are polyclonal for fruit fly period, targeting C terminus; mostly for westerns, elisas
    - here(vendor = ?, cat#=ABIN152723,) is good for IF, IHC,
    - have to align sequences to see how conserved C terminus is
        - fruit fly period via ncbi
- It would be nice to delve into the SNPs found in cry1

references:

Levy, R. C., Kozak, G. M., Wadsworth, C. B., Coates, B. S., & Dopman, E. B. (2015). Explaining the sawtooth: latitudinal periodicity in a circadian gene correlates with shifts in generation number. Journal of Evolutionary Biology, 28(1), 40–53. https://doi.org/10.1111/jeb.12562

Dopman, E. B., Robbins, P. S., & Seaman, A. (2009). COMPONENTS OF REPRODUCTIVE ISOLATION BETWEEN NORTH AMERICAN PHEROMONE STRAINS OF THE EUROPEAN CORN BORER. Evolution, 64(4), 881–902. https://doi.org/10.1111/j.1558-5646.2009.00883.x

---

Page 19: 2018-01-18. prepping meeting with Dan on 2018-01-19

1. IDP, or yearly goals
    - originally didn't include ECB system in IDP.
2. papers

refs:

- Wadsworth, C. B., Woods, W. A., Hahn, D. A., & Dopman, E. B. (2013). One phase of the dormancy developmental pathway is critical for the evolution of insect seasonality. Journal of Evolutionary Biology, 26(11), 2359–2368. https://doi.org/10.1111/jeb.12227
- Stålhandske, S., Gotthard, K., Posledovich, D., & Leimar, O. (2014). Variation in two phases of post-winter development of a butterfly. Journal of Evolutionary Biology, 27(12), 2644–2653. https://doi.org/10.1111/jeb.12519
- Ragland, G. J., Fuller, J., Feder, J. L., & Hahn, D. A. (n.d.). Biphasic metabolic rate trajectory of pupal diapause termination and post-diapause development in a tephritid fly. https://doi.org/10.1016/j.jinsphys.2008.12.013

Also, Tom's paper following up on Greg's work.

Larger/overall hypothesis to frame: Apple flies are more responsive to favorable conditions than haw flies How can they do this? Well different parts of the diapause "program" can be adaptively modified that leads to divergence in emergence timing.

Predictions:

```
1. Apples flies will terminate diapause earlier haw flies
    * and it will correlate with eclosion timing
2. Apple flies will have higher post diapause development than haw flies
3. Apple flies will have higher baseline metabolic rate
```

Tom's predictions don't have the exponential phase of thhe curve in his analyses and predictions.

Analyses:

- Paper needs to present the rational for why different phases of development were selected and how they varied between host races.
- It is also a good point to do a pca on the traits, to see how they're correlated. Then , you can make arguments about how different parts of the diapause program can potentially evolve differently.
- There needs to be a statistics section in methods so that reader can understand how the statistical approaches.

3. ECB proj idea
4. Update on Rhagoletis project
    - Monday- removing plates from fridge
5. Conferences
    - Biological rhythms conference, May 12-16 2018; Amelia Island Florida
        - registration deadline is march 1
        - Should probably write up an abstract for a talk
    - Evolution meeting, August 19-22, 2018; Montpellier, France
        - 3500 abstracts received, but they can only accept 2000

Meeting with Dan 2018-01-19

1. idp;
    - look up fellowships into the UF informatics institute/group (pam) ; UFBI
        - USDA post doctoral fellows
    - look into gordon conference or keystone meetings.

new project: brain transcripts and pool-seq data that I can explore

```
* pop bio talk scheduled feb 2, 4pm
    * check with Bob Holt to see if he can make it
    * Brett Schaeffers
    * Ben Bizer?
    * who else works on range limits?
```

2. Paper discussion
    - how do we make it more impactful? Pitch in terms developmental modules
    - look up developmental modules , Wagner Gunter
    - draft email to tpowell- ask for data

Page 20: 2018-01-18. starting a data science course

There is a really nice data science course by Christie Bahlai (who I follow on twitter). Its called reproducible quantitative methods.

It has 4 parts:

1. Data
2. Analysis
3. Communication
4. Opening your work

In my own course, I want to do something similar, but start off with the problem first. The thought experiment is this: Imagine working on a project and getting back to it in a year. Could you remember everything that you did? You'd have to log what you've done. This leads to solutions to the problem: project organization, meta data, readme file, version control, and electronic notebooks.

There is a nice paper on a modern approach to teach statistics.

paper notes

10 organization blocks for intro statistics:

1. Data tables
2. Data graphics
3. Model functions
4. Model training
5. Effect size and covariates
6. Displays of distributions
7. bootstrap replication
8. Prediction Error
9. Comparing models
10. Generalization and causality

Hmmm, too list like. The main argument in the beginning that the paper makes is that there has been a shift to learning statistics from algebra to showing stuff on the computer.

Data

go over:

1. tidy data
2. meaning of the unit of observation
3. distinction between continuous and categorical variables
4. difference between a data table and the presentation of information

Summary tables in a paper or textbook does not necessarily reflect a datatable, but it is a presentation of the info. For example, a contingency table can be broken down to individual level observations that make up the counts for each row + column.

Data graphics, model functions, and model training , effect size and covariates

use ggplot for figs and use R in general for model functions/training. Use R code to explore different aspects of the model to understand the statistics. ie look at the model output(predictions) and see if it is consistent with model parameter estimates.

Bootstrapp replication

Good way to show re-sampling and simulation in statistics.

Take homes and conclusions:

intro stats should be taught more like data science (especially because datasets are really big now), an investigative process of problem solving and decision making.

Ok, what is my layout?

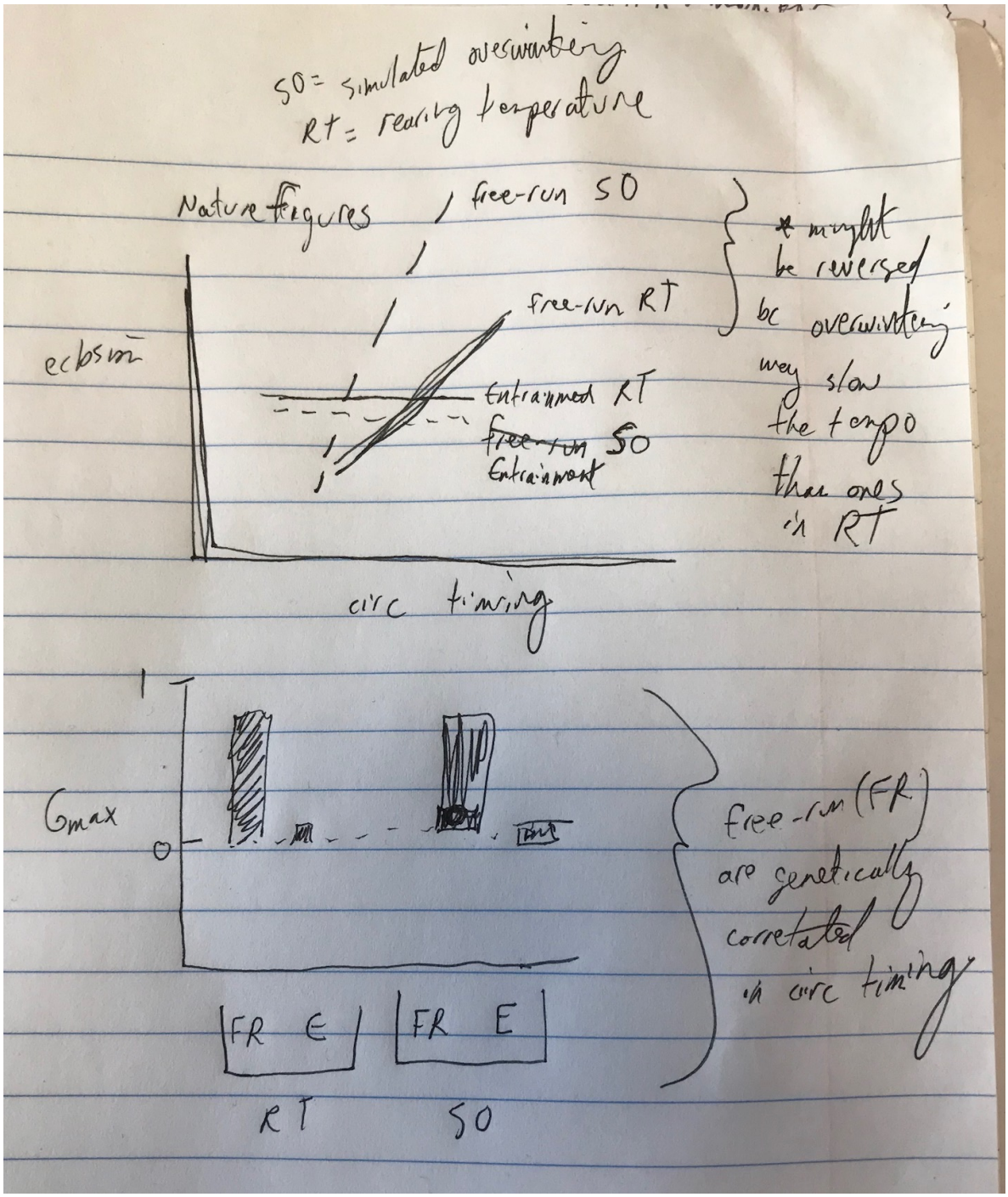
1. How do we do science? Its not really linear, but it can be highly dynamic.
2. Present the problem: it is difficult to reproduce science (for others and your future self!!!)
    - go over thought experiment
3. ..stuck..

some ideas to throw out:

1. repeat a dataset/project found github
2. have in class, student led discussions
    - using a package/function
3. have a larger data science project that will lead to a manuscript
    - leave slots open for students to work on project
    - students can even use their own dataset
4. for weekly exercises, work on a single project and explore ins and outs or work on multiple different data sets
    - data parsing exercises
    - data plotting exercises
5. have students work on each other's project to see if it is reproducible?!? cool idea
6. Encourage students to implement electronic notebook via github
    - I find that being specific about usage is a good way for students to actually adopt this
7. grading break down
    - (Bi-)weekly exercises (30%)
        - coding + answer questions
    - Electronic notebook (30%)
        - have them answer questions (replaces blog in Christie's syllabus)
    - Semester Project (30%)
    - Participation (10%)

Page 21: 2018-01-22. Rhago nature figures for biological rhythms project

Drew this up on Friday. I'm trying to think about possible outcomes of the relationship between biological timing (circadian timing) and eclosion (days) under different treatment groups. If we genotype all of the samples, we can construct a relationship matrix and feed that into a mixed effects model to quantify the genetic architecture(Gmatrix, variance -covariance matrix) for each treatment. THen we can see how correlated they are. I think I'd expect greater correlation in free-run then entrainment, showing that they intrinsically diapause and can respond to photoperiod given the opportunity.



Page 22: 2018-01-23. Meeting with Dan, Tom about termination ms

Background:

Dan has invited for me to work on ms trying to identify which diapause tactics are used to shift seasonal emergence timing (eclosion).

Rhagoletis follows a few phases in which they exit diapause. IN the kostal 2006 framework, when organisms are in diapause and meet favorable conditions, they terminate diapause and then begin post-diapause development. We can determine these developmental marks using a function valued trait approach by measuring metabolic rate across days out of overwintering. The metabolic trajectory occurs in 2 or 3 phases. For the two phase model, mr intially increases and then plateaus, and this point corresponds and is interpretted as termination. Then, mr exponentially increases until eclosion, which interpretted as post-diapause development. The three phase model is the same, but there is initial drop in mr in the first few days.

Meeting:

1. Tom tried path analysis
2. PCA suggestion- good for figure display, but there can be compounding error when estimates with error are used in a follow up analysis
3. Dan wants to pitch paper in terms of developmental modules, developmental tactics/strategies (what is the difference?)
4. contingency table of diff metabolic trajectories, limitations to interpretation

Tom will send me dataset and supplementals to Dan.

Action items for me:

1. look up and into developmental modules - scholar search Gunter Wagner
2. look into the usage of tactics and strategies
3. try pca analysis when tom sends me data
4. write up flow of ideas

Page 23: 2018-01-23. Reading Gunter et al. 2007, The road to modularity

reference:

Wagner, G. P., Pavlicev, M., & Cheverud, J. M. (2007). The road to modularity. Nature Reviews Genetics, 8(12), 921–931. https://doi.org/10.1038/nrg2267

Ok, what is modularity. From Dan's discussion, the different parts of the developmental trajectories as measured by metabolic rate is a "modular" process.

First par tof abstract describes modularity as: a network of interactions that can be subdivided into autonomous, internally highly connected components...(what?)

Different Definitions:

- Variational module-a set of covarying traits that vary independently with the set of traits of interest. (So bad)
- Functional module- features that act together in performing some discrete physiological function
- Developmental module- embryo program that is quasi-autonomous ...definition is os bad.
- Quasi-autonomy - a lower than average grade of connectedness. The elements of modules are highly interaconnected but to an increased extent are unconnected with other modules - called quasi-independence

Don't see a good usage of these terms really. Traits can be correlated, just say that. But they're correlation says nothing abou ttheir interactions, which is a separate issue.

The term with best utility is a "functional module". In the ms, the work could be framed in terms where the diapause module compromises of termination and post-diapause development. Diapause timing is a functional module. Modules are often described as having features which can be or not be independent from one of another.

It can be viewed as an autonomous physiological proceess that is part of the organism.

One way to approach variational modules is to quantify the correlation matrix: a table of correlation coefficients among quantitative traits, which summarizes the correlation. Strong covarying traits are called variational modules.

Page 24: 2018-01-23. Strategies vs tactics (Mart R. Gross 1996)

ref: http://labs.eeb.utoronto.ca/gross/newGross1996.pdf

Paper on how to think about and use strategies vs tactics.

- Strategy- genetically based program (decision rule) that results in the allocation of somatic and reproductive effort of an organism (energy/development) among alternative phenotypes (tactics).
- Tactic-is a phenotype that results from a strategy. An example is the fight for accessto a mate, but alternative tactic may to sneak. The decision about which tactic is expressed is made by the strategy.

Very confusing. Strategy sounds liek a downstream afterfact of the expression of a tactic.

Maybe a strategy is an overall life history strategy (ie survive the cold) and a tactic is the mechanism by which a strategy is attained/achieved (plastic vs constitutive mechanisms)

Page 25: 2018-01-24. revisiting thermal niche paper

one of SED's comments:

> But you never really look for plasticity. Where's the model with CTmax ~ rearing T With species/population as a random effect?

I thought the model selection part would "look" for the effect of acclimation if it was there? Anyway, I'll run an anova to test for the effect of acclimation temp on CTmin/max while blocking for species.

CTmax

```
summary(aov(Ctmax~Acclimation+pc1_species,data=merg))
             Df Sum Sq Mean Sq F value   Pr(>F)
Acclimation   1   0.06    0.06   0.235    0.631
pc1_species   1  49.22   49.22 199.089 3.06e-16 ***
Residuals    36   8.90    0.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

CTmin

```
summary(aov(CTmin~Acclimation*pc1_species,data=merg))
                       Df Sum Sq Mean Sq F value   Pr(>F)
Acclimation             1   2.01    2.01   1.372    0.249
pc1_species             1 107.18  107.18  73.102 4.32e-10 ***
Acclimation:pc1_species  1   0.36    0.36   0.247    0.623
Residuals               35  51.32    1.47
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Page 26: 2018-01-25. Phenotypic Prediction workshop 2018

Schedule

- 7:30-8AM registration and breakfast
- 8:10AM opening remarks
- 8:10-9:00AM "The use of genomics in animal breding", Dr. Jack Dekkers, Iowa State University
- 9-9:50AM "Going beyond statistics - the case for biological whole genome models", Dr. Frank Technow, Dupont -Pioneer
- 9:50-10:20AM coffee break
- 10:20-11:10AM "Will Big Data Close the missing heritability gap?", Dr. Ana Vazquez, Michigan State University
- 11:10-12:00PM “High Throughput data (environmental, genomic and phenotypic covariates) for improving predictive ability under complex scenarios with applications in plant breeding.” Dr. Diego Jarquin, University of Nebraska-Lincoln

Opening remarks:

Marcio

event, 5th edition

2 main goals:

1. bring people closer together, common goal of phenotype predition
2. enable students to be exposed to these talks

questions can be posted on twiter (#UFPPW) from livestreamers

800 people registered

"The use of genomics in animal breding", Dr. Jack Dekkers, Iowa State University

Netherlands, phd in wisconsin

background is in animal breeding

PIC(pig improvement company) has been used a suite of markers starting 1991 (HAL gene). HAL gene associated with poor meat quality. Started off small number of SNPs and then more recently, many more markers were used for phenotype prediction. Genomic selection...whats that.

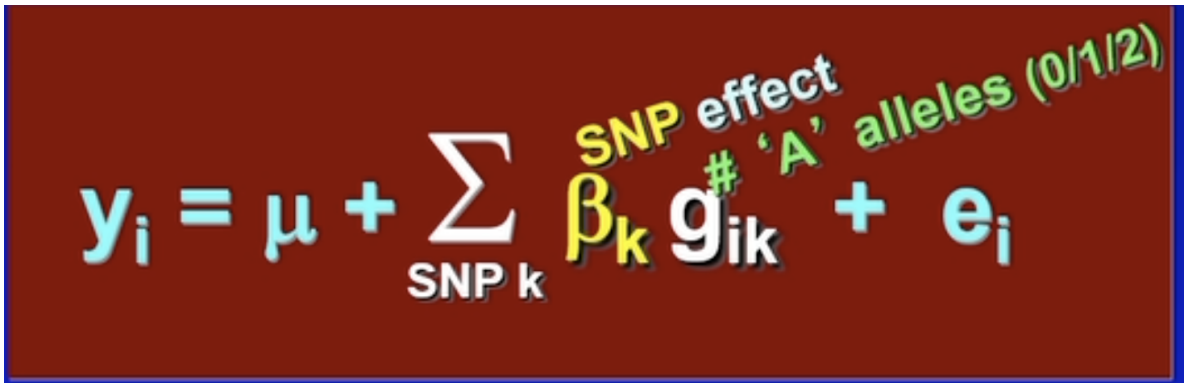
Genomic Selection

Meuwissen et al. 2001.

1. Take reference population
2. take phenotypes and genotype them (training data)
3. estimate marker effects
4. Genotype next generation and use model to predict

Genetic evaluation using high density SNPs

1. SNP effects are fitted as random vs fixed effects
    - SNPs are random effects, enabling all SNPs to be fitted simultaneously
2. General model
    - Mixed effects model- PLUB(pedigree) or GBLUP(genomic relatedness
    - use to estimate breeding values of new animals based on genotypes alone
    - Bayesian variable selection used to ID SNPs too



3. A relationship matrix including full pedigree and genomic information: Create relationship matrix from phenotype and genotype
    - H matrix

The promise of Genomic selection

- ID phenotype at earlier age
- reduce need for costly phenotyping
- Reduce generation intervals
- INcrease accuracy for "difficult" traits
    - reproduction, longevity, meat quality
    - disease resistance
    - crossbred performance in field
- Reduce rates of inbreeding/generation
    - less emphasis on family information
    - select on animal's 'own gneotypes (for markers)

Outline

1. Implementation of GS in dairy cattle
2. pig/poultry breeding programs
3. predicting the accuracy of genomic prediction (how large? structure?)
4. Genetic improvement of commercial X-bred performance

Dairy Cattle

Reliability = variance explained squared?

can genotype from embryo; selection decisions are faster(traditional way = progeny testing = 5 years vs DNA testing = 6 months)

1.5 million holstein cows genotyped

Genomic prediction = higher genetic gains (doubling)

Pig/poultry

more difficult

Typical breeding pyramid in pigs is a 3 way cross

- 3 breeding pops : 1 sire, 2 dam lines
- Genomic prediction needs to be implemented in 3 lines , which are not that large
    - each company has their own pigs to themselves
- traits are not sex limited like in cattle

Challenges:

1. multiple lines
2. limited training data per line
3. limited opportunity to reduce L
4. large # selection candidates 5 limited value of each candidate

Opportunities:

1. preselection of candidates for further testing
2. selection for female reproduction, longevity
3. feed efficiency
4. selection for carcass/meat quality
5. selection for disease resistance
6. selection for commercial crossbred performance

THe money end, What is the relative cost and benefits of genotyping vs phenotyping?

- no money, invest in phenotyping
- money, invest in genotyping

Training populations:

If heritability = 30%, you need about 5,000 individuals. Higher heritability = less individuals needed

Prediction accuracy

P<-- Genetics P <--Environment

P<-- Q(proprotion of genetic variation at the QTL that is captured by markers) P <--Environment

Accuracy = q * rq

(Dekkers 2007, JABG)

q = proportion of genetic variance at QTL captured by markers

= Markers /(markers + effective # of chromosome segments)

rq = accuracy with which markers effects are estimated

= sqrt((Nq^2h^2))/(Nq^2h^2 + Me))

Me is a key parameter if you want to get a handle of how big of apopulation you need to obtain a certain accuracy

How large is Me?

Depends on effective population size, which depends on LD (linkage equilibrium)

It also depends on L, which is the size of the chromosome

Assume chromosomes are independent, but are not, but there are covariances betewen genotype and chromosomes

by accountig for covariance, less individuals needed

The need for re-training

Train on data prior to generation 1

GBLUP is higher accuracy than PBLUP

Relatedness matters for training and test datasets

Me can be estimated from G-A (genomic and pedigree)

Variance of deviation (Var(G-A)) = average LD(r^2) over all pairs of loci = 1/Me

Me can be estimated from G-A

Relationship higher = Var(G-A) higher = Me lower = Ne smaller = increase genmic prediction accuracy

easier to predict close relatives

Conclusion

- hard to predict across breeds
- requires large datasets

Questions:

1. Why avoid inbreeding?

recessive deleterious mutants can selected out with inbreeding. so some is good. Works in corn, but in animals, its too costly.

2. Genomics can ID cause inbreeding depression.
3. What are we finding in terms of dominance? (non-additive variance)

Hard to capitalize on dominance because directed matings are limited. Breed bred lines to increase heterosis.

"Going beyond statistics - the case for biological whole genome models", Dr. Frank Technow, Dupont -Pioneer

phd from germany, bayesian methods, canada

3 sections

1. role of prediction
2. limits of predictability
3. pushing the limits

"Predictive analytics"

we can only observe phenotypes, but breeding values determine performance of next generation

and overall genetic values determine success of variety

Falconer 1989 Intr. quant genetics pg 163

Build upon predictive skills!

1. ground them in quant gen and biology
2. incoporating novel types of data and info
3. develop powerful stat methods (whole genome prediction)

Paradigm shift from QTL detection to prediction of genetic values

Whole Genome regressions

- set of genotypes and phenotypes (estimation set)
- regression of phenotype

What drives accuracy? Cooper et al. 2014, genomic prediction --see ref

Linear models:

- based on linear mixed model
- determiend by data they were fitted on
- confined to context of observed data
- extrapolation risky
- modle of biology unrealistic

Genetics is non-linear;yield is highly complex functio of :

1. traits on lower phys levels (component traits
2. Environment and management practices

This generates the fitness landscape determined by (Sewell Wright 1932)

- G x E x M (interactions with environment and management)
- physiologicla (trait) and molecular (gene level) epistasis

Why do the linear models still work?

- Constraining target populations of environment (TPE)
    - homogeneous weather and soil conditions
    - similar cropping systems
- Constraining target pop of germplasm (TGE)
    - heterotic groups
    - limiting diveristy (fixing genomic regions)

Biased sampling and making the biology easier ; takes away G x E

Hill, Goddard, and Vissher ():

- molecular pathways imply non-linear gene action
- genetic variance is mostly additive

reason for paradox: most alleles are at extreme allele frequencies (what...)

Stepping out of the comfort zone

Comfort zone- known part of bio space

- multi environment trials (MET)
- constrains make system linear
- predictable with linear models

Why step out?

- adapting to climate change and more erratic weather patterns
- increasing diversity for long-term genetic gains

MOdels should explore for adjacent possibilities

Ex: Multienvi field trials and they can test for Gx E and even G x E x M

downside of machine learning: no prior biological knowledge (a downside and upside)

Biological model (aka first principles)

1. captures univ causality of bio systems
2. transcends specific data set and provides outside context

Causal models - a priori knowledge

Thought experiemnt: what is the area of rectangle

We need data. But we need more than data. We also need prior knowledge and outside context.

Crop growth models:

Are biological models that incorporate non-linear interactions within and between all system components

- plant phys
- env and management

Models are based on concepts of

- resource capture, utliization, allocation
- results of decades of biological research
- widely used to study agronomic practices and cropping systems
- increasing interst in application to breeding and applie dgenetics

it is a set of models

Maize growth model

Environemntal inputs - temp and solar radiation

Key phys inputs - radiation use efficiency , node number

Loop through model

Sounds like we don't need any SNP information.

Merging biology wiht WGP through bayesian statistics

BioGWP - a biological whole genome prediction model

- Bayesian herarchical generalized linear model
- CGM acts as a "link function"
- genetics represented by latent physiological traits
- model for marker effects is "BayesA"
- posterioer sample with metroplis within Gibbs algorithm

Think of this as a biostat machine that can integrate diverse sources of info such as :

- harvest phenotype
- genomic data
- biological insights
- env
- management

into high resolution yield prediction

Proof of concept-

Integrating crop growth models into whole genome prediction with approximate Bayesian computation plos one

- simulated data
- demonstrate ability to work with highly non-linear systems

Which model was better?

Reduction to practice- requiremments

- need accurate crop growth model
- need env soil

ARGOS: model genotype by transgene by env interaction

- transgene affects ethylene response
- maintains silk elongation under drought
- protects yield potential
- but effects vary by genetics/phys and env background

ARGOS affects silking - growth models need to incorporate this

"Will Big Data Close the missing heritability gap?", Dr. Ana Vazquez, Michigan State University

GWAS associated with disease- genetic risk score based on GWAS significant variants

- Problem: proportion of variance explaiend by GWA-sig hits remains lower than the trait heritability ; aka missing heritability
- there could be a large number of small-effect variants not detected . starndard gwas lack power to detect
- for complex disease, these variants can explain a sizable proportion of genetic variance

Biobanks

- recruit data from thousands of patients
- pairing genomics data with electornic medical records and extensive phenotyping
- same protocol, variable def for all patienets
- all subjects are genotyped with same platform

Landscape of factors that affect prediction accuracy for complex traits

- acess effects of sample size, number of markers, LD, estimation procedure

Model: Human Height: A model complex trait

- highly heritable (.8)

- no strong QTL
- extensively studied
- strong effect of age and gender

Genotyping

- 600K SNPs *

Data from UKbiobank

- use interim release
    - 1k white people
- full release
    - 4k white people

science is so biased....

Workflow

1. pre-adjust standing height by age and gender
2. GWS in training set , do a SNP -ranking

Data partitions: Training and testing sets

- interim release
    - 102k training, 80k test
- full release
    - 400k training, rest test ; rep 5 tiems

adjusted SNP windows

3. Fit bayesian models to top - p SNPs
4. assess prediction accuracy on test set

Whole Genome regression, predictions are derived using thousands of SNPs

Challenges:

- curse of dimensionality --> bayesians, penalized to sovle
- Computation ;
- require large sample sizes

Prediction:

- ..5-.6 as functin of sample size
- prediction accuracy increasing , but slow
- highest prediction .65, r^2 0.42

More SNPs genomic heritability increases but prediction accuracy plateaus. The gap is reduc of accuracy on estimated efects.

How many markers are needed? Depends on sample size and number of SNPs

Data partition into subsets--no clue why

Prediction accuracy has opt of ~5k snps, and any more markers reduce prediction accuracy .

Increasing training size increases prediction accuracy.

“High Throughput data (environmental, genomic and phenotypic covariates) for improving predictive ability under complex scenarios with applications in plant breeding.” Dr. Diego Jarquin, University of Nebraska-Lincoln

Genomse to field idea: G x E project

genomes2fields.org

gxe is problem: use it, reduce it, ignore it

Page 27: 2018-01-26. Meeting with Dan

1. Rhago projs
    - Discuss abstract and Tom needs to send me data
        - need flow of ideas for intro/discussion
    - Dan needs to send email to Greg about Brain transcriptome
    - taking samples out of fridge
2. Details on fellowships
    - UF bioinformatics post doc.
        - app deadline is OCT 2nd for next year.
        - hs to follow the following thrust areas or themes.
        - need to work with 2 departments
3. Pop bio talk next friday feb 2
    - bob holt and brett scheffers going

Conferences:

Biological rhythms conference, May 12-16 2018; Amelia Island Florida

```
* registration deadline is march 1
* Should probably write up an abstract for a talk
```

Modules discussion:

2-3 modules, variation in termination, but post-diapause might be more environmentally driven (put discussion)

can pitch in terms of variational module

Dan, hit plant literature to check if modules are used in terms of module selection

functional modules have diff genetic variation that nat sel can act upon, but also have diff ways that can resond to environment

future readings:

Michaels 2010, Dan will send more.

Page 28: 2018-01-30. Readings in Rhagoletis, pop gen, directions of gene flow

references:

- XIE, X., MICHEL, A. P., SCHWARZ, D., RULL, J., VELEZ, S., FORBES, A. A., … FEDER, J. L. (2008). Radiation and divergence in the Rhagoletis Pomonella species complex: inferences from DNA sequence data. Journal of Evolutionary Biology, 21(3), 900–913. https://doi.org/10.1111/j.1420-9101.2008.01507.x
- Xie, X., Rull, J., Michel, A. P., Velez, S., Forbes, A. A., Lobo, N. F., … Feder, J. L. (2007). HAWTHORN-INFESTING POPULATIONS OF RHAGOLETIS POMONELLA IN MEXICO AND SPECIATION MODE PLURALITY. Evolution, 61(5), 1091–1105. https://doi.org/10.1111/j.1558-5646.2007.00091.x
- Feder, J. L., Berlocher, S. H., Roethele, J. B., Dambroski, H., Smith, J. J., Perry, W. L., … Aluja, M. (n.d.). Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis. Retrieved from http://www.pnas.org/content/pnas/100/18/10314.full.pdf

These are a series of pop gen papers aiming to understand the population dynamics of Rhagoletis pomonella. How and in which way have populations differentiated?

Feder et al. 2003 ; PNAS

One issue with the sympatric speciation story in this system is that one can't excldue allopatric speciation. Perhaps apple maggot fly alleles were initially geographically separated before sympatric incipient speciation occurred.

Well in this system, the apple flies only occur in midwest and northeastern US and Canada, where as the haw flies extend far down into Mexico. So these authors wanted to understand the direction of gene flow between populations of haws and apple. They sampled flies from a disjunct mexican pop and compared with US populations.

Their approach to the study was to compare gene trees of different populations. Specifically, 3 nuclear and 1 mitochondrial loci. They estimated divergence timing and deteremined whether topologies were likely due to lineage sorting or introgression.

Lineage sorting is the random assortment of alleles into extant taxa and this produces gene relationships that are not necessarily reflective of species relationships. Introgression can do this too, but it is a directed process where an allele migrates from one lineage to another.

Ok, so how did they distinguish lineage sorting vs introgression? They performed a likelihood ratio test, where LR = ML1/ML2. They basically evaluated maximum likelihood estimates between models to see if they're stat sig diff. Ok, how did they do this?

The null model is lineage sorting (ML1), where they let N and SN nodes have different origin times among loci . This reflects independent origins of inversions on chromosome 1-3 in the common ancestor.

The alternative model is ML2, which is the introgression model , where the derived time of origin for N and SN/M nuclear nodes are the same across loci.

Results: ((N),(SN,M)) relationships for chromosome 1-3 and (US,M) for mitoDNA.

Fig 4 makes the modelling more sense now, because the timing of how genes diverge give a good indication of incomplete lineage sorting vs gene flow. WHen alleles show shifts later in time (1.1mya), then it is more likely due to gene flow. However, it might be harder to distinguish gene flow vs ILS at 1.1 mya.

Introgression model was supported. And they argue this has something to do with diapause with no data to show for it....

Xie et al. 2007; Evolution

This study builds on previous work and looksl ike it is a snap shot or focus in on mexican and US populations of rhagos. They use 15 nuc loci and 1 mito loci.

They're pitching the story as "speciation mode plurality". wtf... I think this means that both allopatric and sympatric mechanisms contribute to speciation.

They found for some genomic regions, there is pop structure as such, ((EVTM,SMO),US) but for other genomic regions (chromosome 1-3), the US is paraphyletic as such, (US,((US,SMO),EVTM). So SMO alleles have migrated or hve been retained into US populations. The differentiation may be due to allopatry, which produced and maintained the genetic variation for host races to diversify their phenology.

They argue that speciation can involve both allopatry and sympatry...but it is possible for them to be acting in isolation. Allopatry then sympatry, not allopatry and sympatry. Their argument makes very little sense. It does change how we view how genetic diversity can be maintained and how that maintenance can facilitate divergence in future populations/lineages.

One problem- they have not established the directionality of gene flow.

Xie et al. 2008; J. Evol. Biol.

goals:

1. test whether any genetic signposts of divergence exist for rhagos (what?)
2. nuclear loci are consistent with speciation mode plurality
3. whether patterns of genetic divergence vary across the genome

---

Page 29: 2018-01-31. Meeting with Dan , MR trajectories paper

- Are the MR mass corrected? check with Tom.
- Check on c and remove it to see how the loadings change or whatever
    - truncate last data point to see how c shifts
    - c might be flawed
- another analytic approach? Partial correlation and regression with eclosion

Predict based on pcs

I did stats yesterday, but I'll dump the quick and dirty findings here:

PCA table:

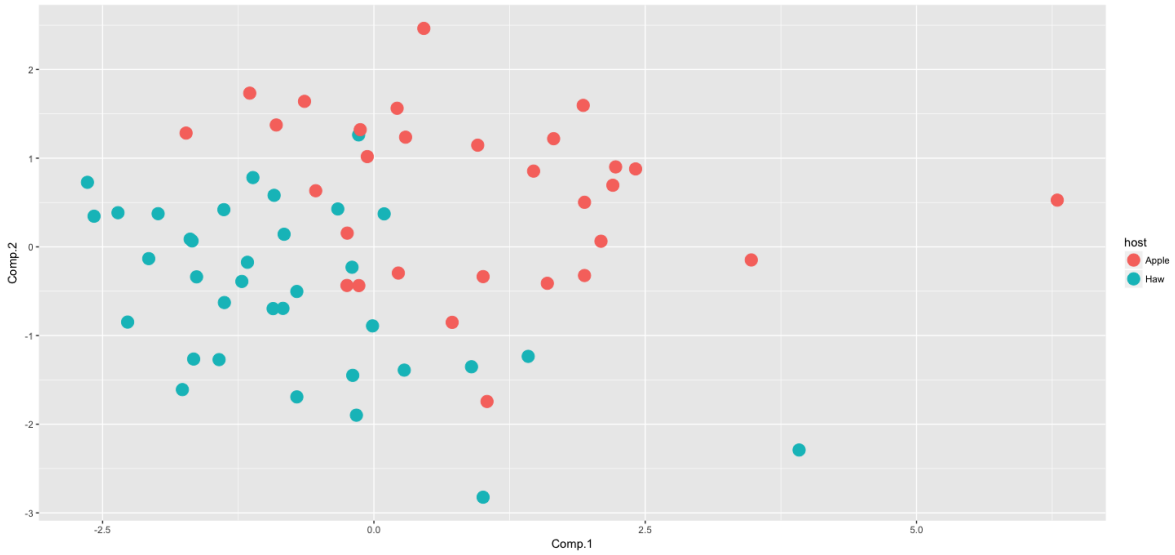
	Comp.1	Comp.2	Comp.3	Comp.4
b	0.508	0.319	0.006	0.622
term	-0.330	-0.697	-0.178	0.611
c	0.451	-0.162	-0.850	-0.185
post_dd	0.389	-0.600	0.344	-0.390
plat	0.528	-0.161	0.357	0.229

Summary: 4 pcs captured 96% of the variation in metabolic trajectories. The first pc represents the correlation between b,c,post dd, plat, that are altogether negatively correlated with termination. The second pc represents the correlation between term,c,post_dd, plat, that are altogether negatively correlated with b.

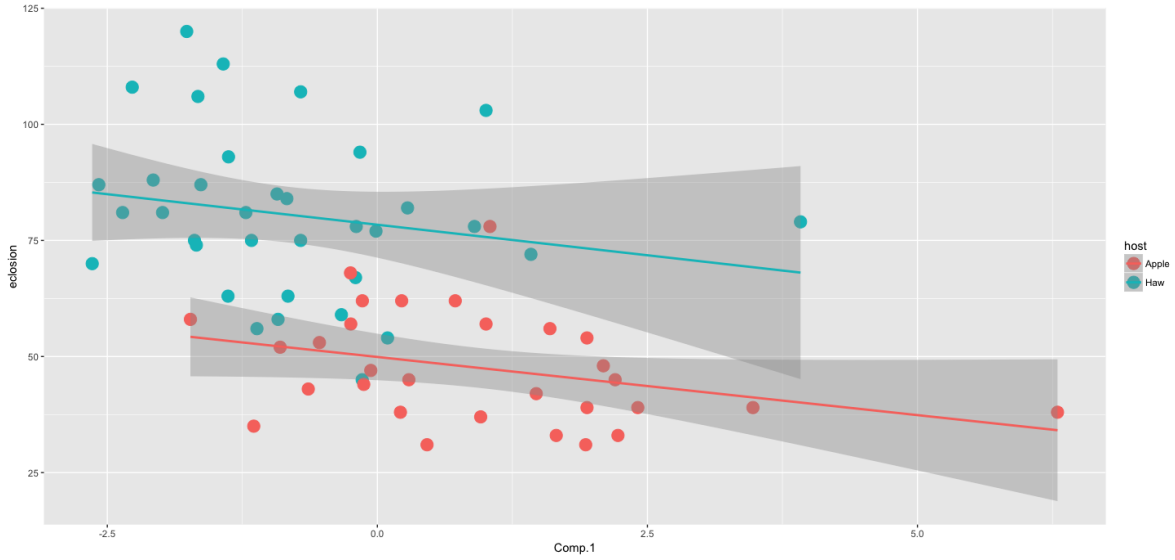
Stat model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.705	0.631	99.315	**0.000**
hostHaw	2.192	0.811	2.702	**0.009**
Comp.1	-6.294	0.299	-21.068	**0.000**
Comp.2	-14.127	0.472	-29.917	**0.000**
Comp.3	-2.100	0.388	-5.418	**0.000**
Comp.4	10.725	0.678	15.813	**0.000**
hostHaw:Comp.1	-1.203	0.537	-2.241	**0.029**
hostHaw:Comp.2	-3.213	0.715	-4.491	**0.000**

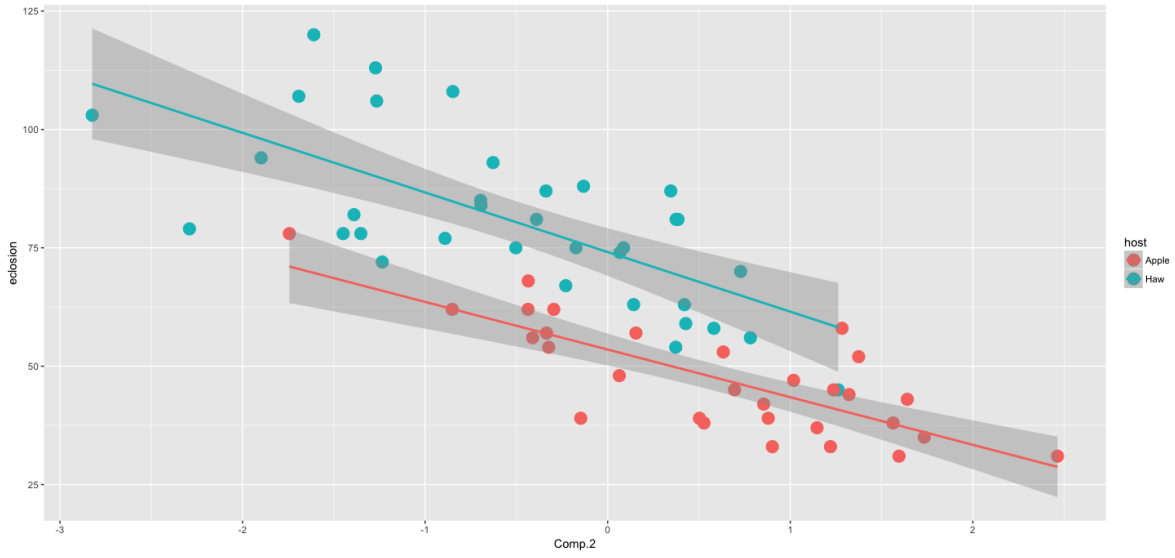
PCA biplot between host races



PC1 x host interaction



PC2 x host interaction

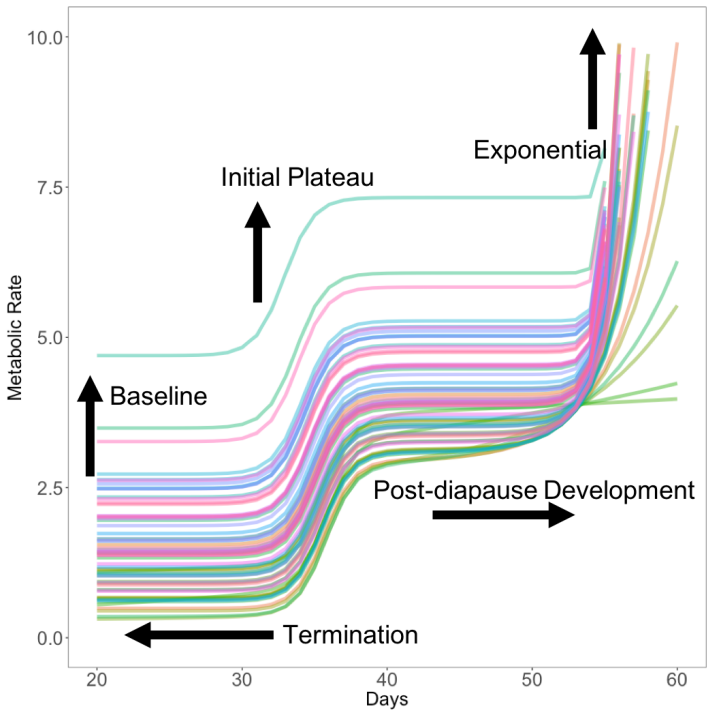


Predictions based on PCs

sample code:

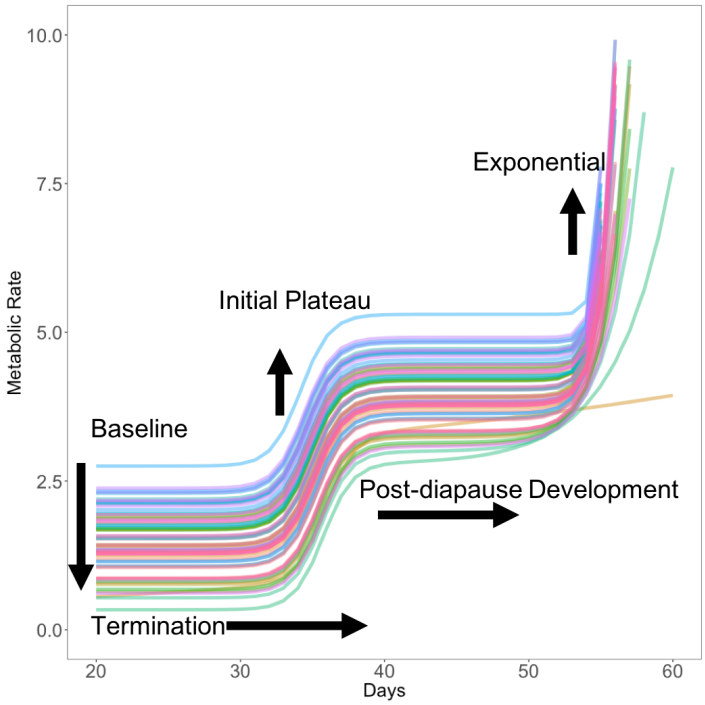
```
mean.trait<-c(1.5,35,1.3,20,4)

pc2<-data.frame(t(apply(data.frame(dat.pca$scores[,2]),1,function(x)
{x*dat.pca$loadings[,1]+mean.trait})))

pc2$X<-pc1$term+pc1$post_dd
apply(pc2,2,range)

#pred
predpc2<-data.frame(apply(pc2,1,function(x)
{Tom(b=x[1],m=x[2],p=x[5],c=x[3],X=x[6])}))

predpc2.long<-gather(predpc2,ind,mr,X1:X66)
predpc2.long$t<-rep(seq(1,60,1),65)
```

PC1:



	Comp.1	Comp.2
b	0.508	0.319
term	-0.330	-0.697
c	0.451	-0.162
post_dd	0.389	-0.600
plat	0.528	-0.161

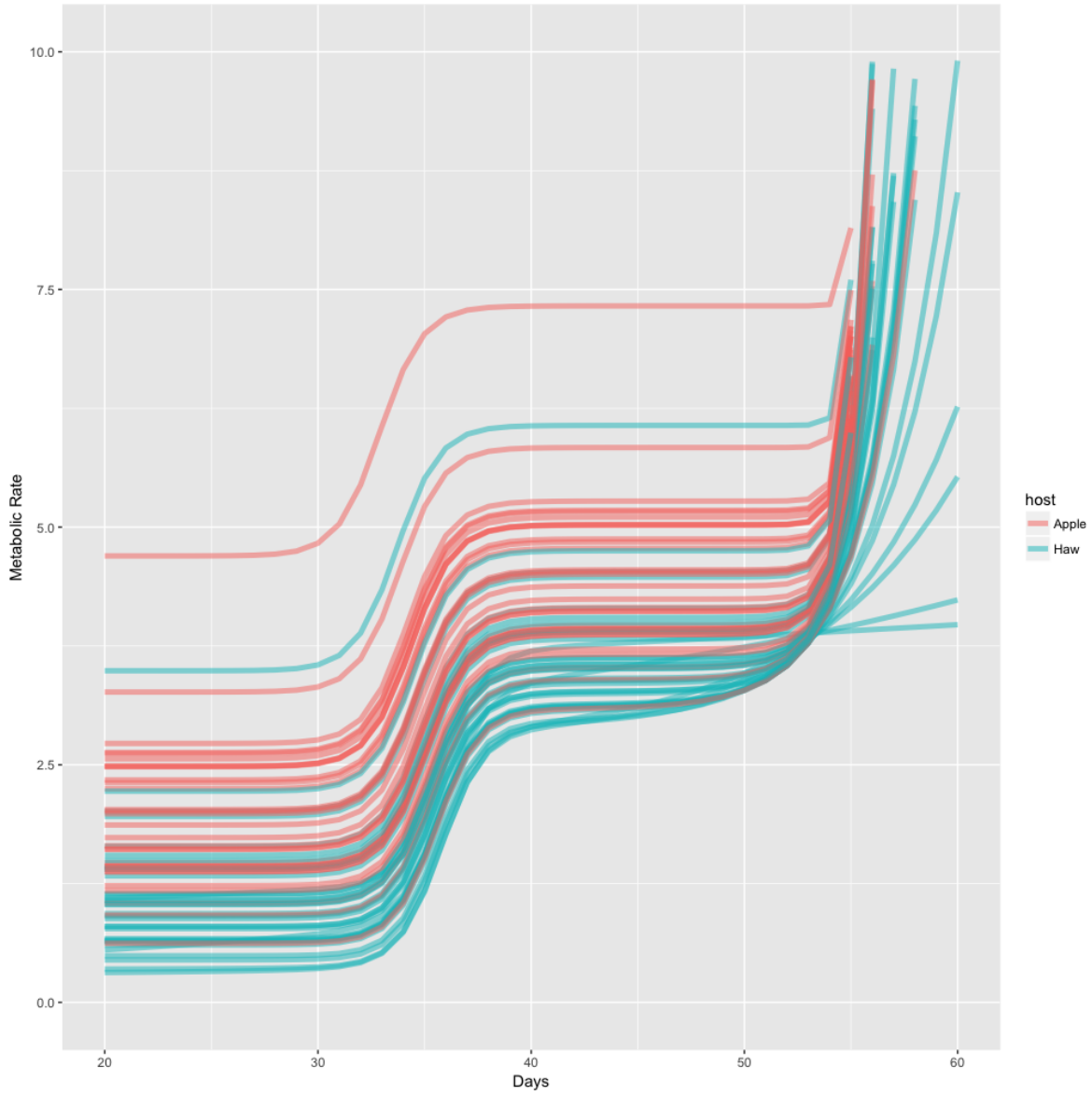
PC2:



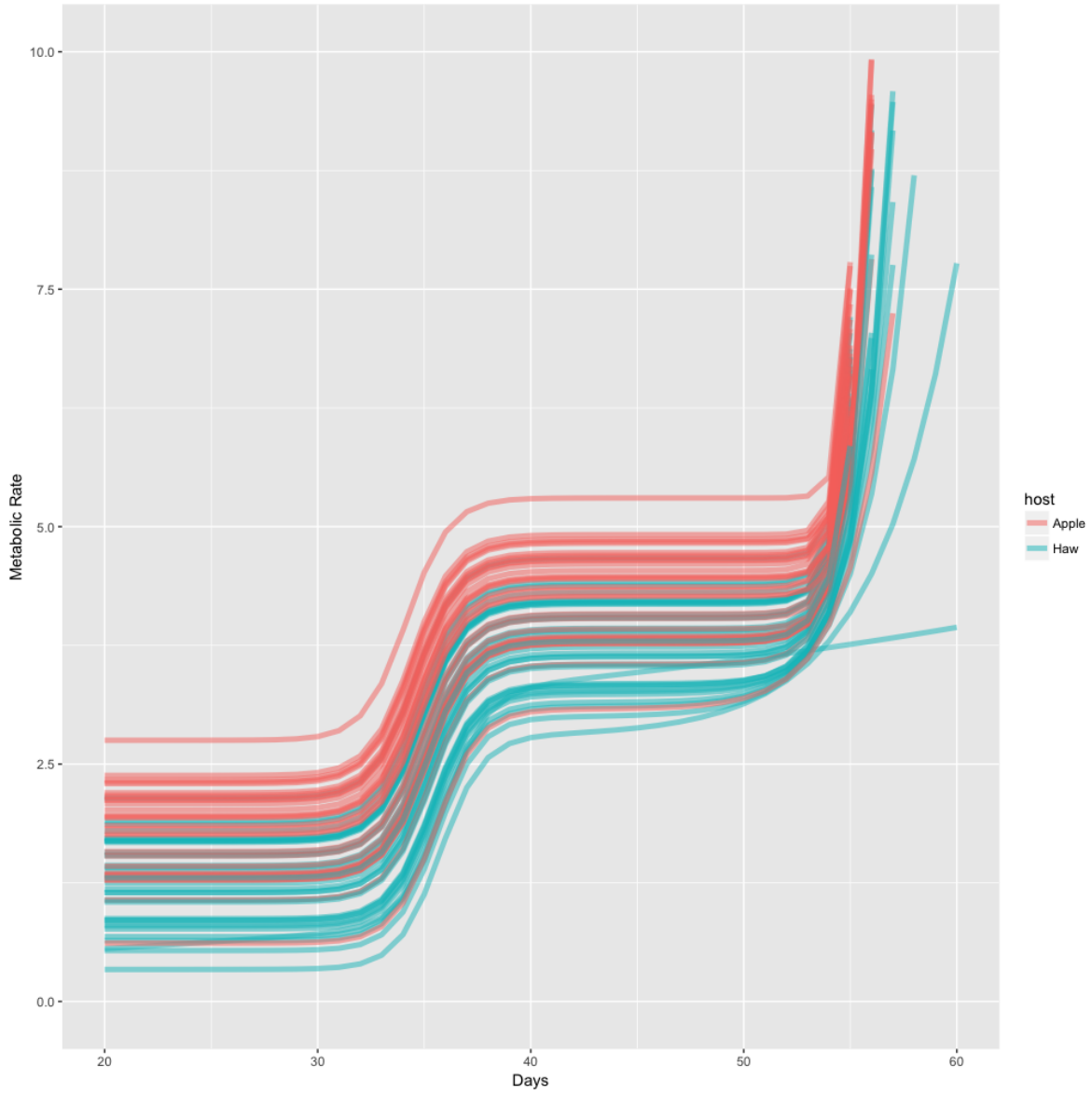
	Comp.1	Comp.2
b	0.508	0.319
term	-0.330	-0.697
c	0.451	-0.162
post_dd	0.389	-0.600
plat	0.528	-0.161

plots by host race

PC1:



PC2



Page 30: 2018-02-01. follow up analysis; partial correlation of parameters

Dan wants to know the sole effect of each parameter. I think this can be done with a multiple linear regression on the scaled parameters such that betas can be directly compared.

model construction: Testing effect of each parameter and interaction with host on eclosion

```
fullmod3<-
lm(dat$eclosion~stand.dat$term*dat$host+stand.dat$b*dat$host+stand.dat$c*dat$hos
t+stand.dat$post_dd*dat$host+stand.dat$plat*dat$host)
mp<-summary(stepAIC(fullmod3,direction="both"))
```

output:

```
knitr::kable(coef(mp))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.641181	0.6866125	92.688646	0.0000000
stand.dat$term	19.580301	0.7211073	27.153104	0.0000000
dat$hostHaw	1.540051	0.8723161	1.765474	0.0827479
stand.dat$c	-1.141651	0.3241461	-3.522026	0.0008419
stand.dat$post_dd	2.862625	0.4133899	6.924759	0.0000000
stand.dat$plat	-1.043086	0.4613302	-2.261040	0.0275236
stand.dat$term:dat$hostHaw	2.284787	0.8539093	2.675679	0.0096775

Exploring the termination * host interaction

Take residuals from a regression from previous model but without the interaction term and then plotting those residuals against eclosion

model construction

```
fullmod4<-
lm(dat$eclosion~stand.dat$term+dat$host+stand.dat$c+stand.dat$post_dd+stand.dat$plat)
leftover<-as.vector(scale(resid(fullmod4)))
```

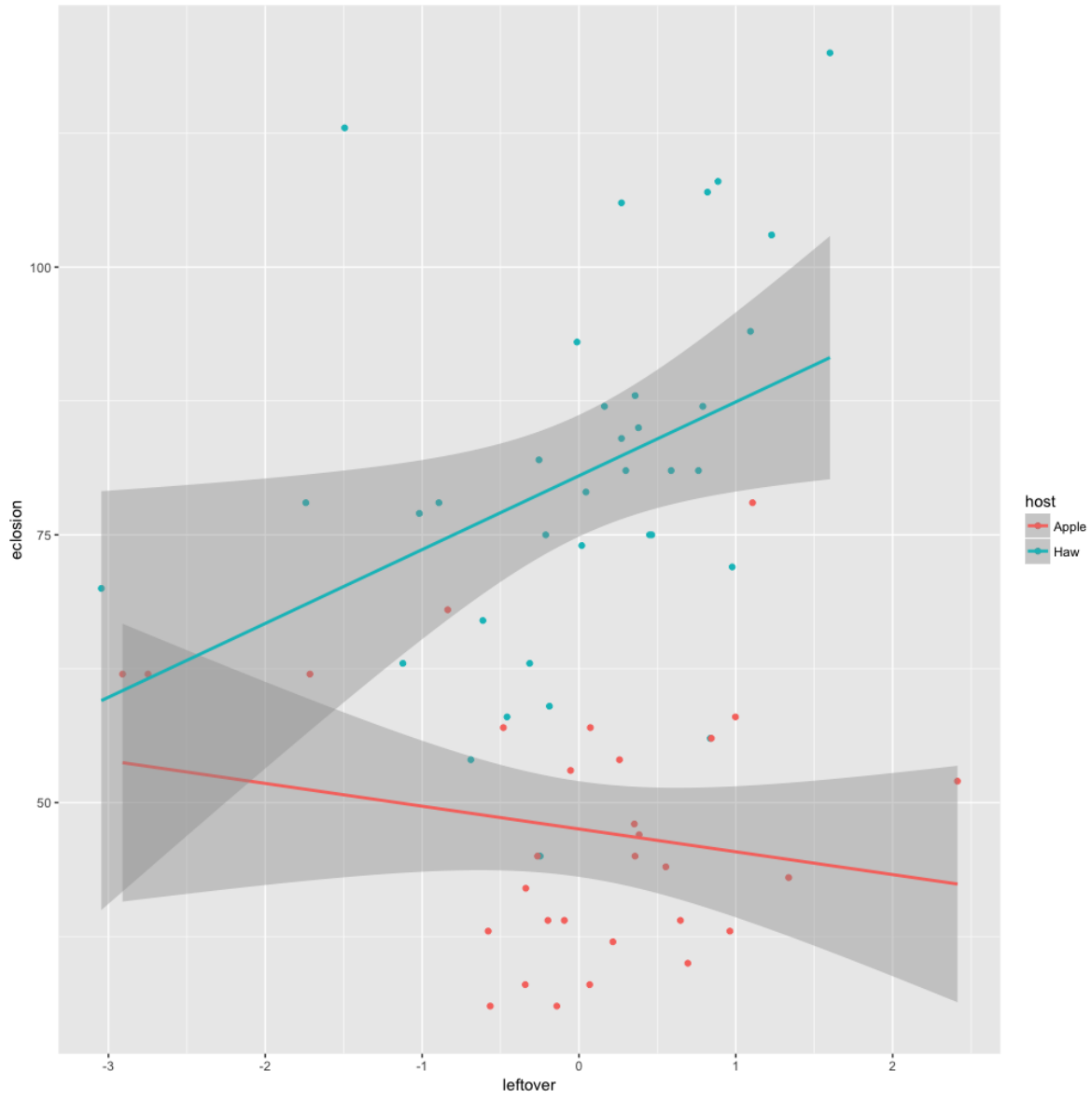
does the interaction persist?

```
pmod5<-lm(dat$eclosion~leftover*dat$host)

knitr::kable(coef(summary(pmod5)))
```

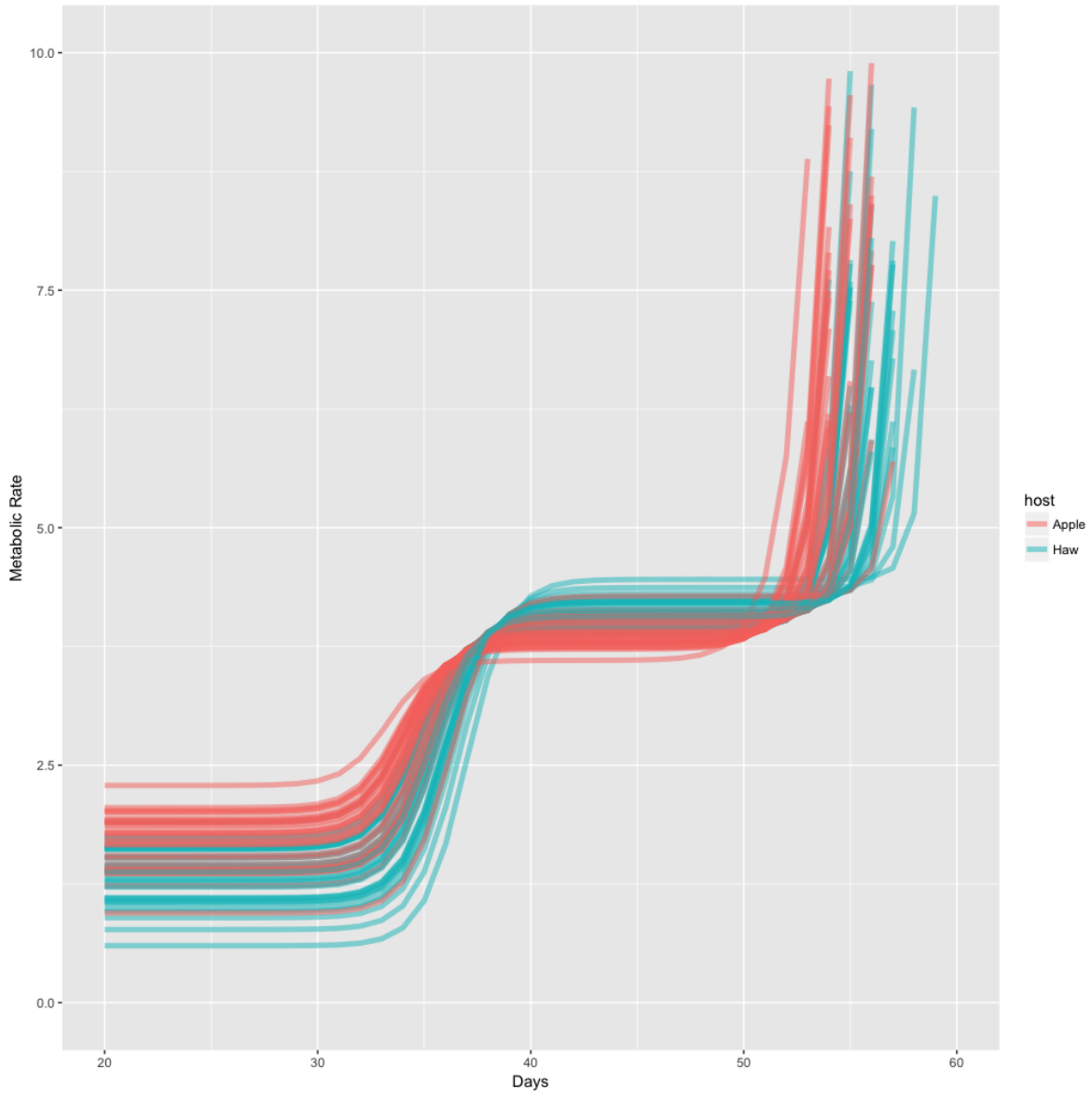
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.533333	2.663403	17.8468395	0.0000000
leftover	-2.128265	2.497493	-0.8521605	0.3974584
dat$hostHaw	32.980952	3.629607	9.0866455	0.0000000
leftover:dat$hostHaw	9.022192	3.655032	2.4684304	0.0163889

ok, make into dataframe and plot:

```
mdat<-data.frame(eclosion=dat$eclosion,leftover,host=dat$host)
ggplot(mdat,aes(x=leftover,y=eclosion,colour=host))+geom_point()+stat_smooth(method='lm')
```



fixing plotting mistake for pc2 predictions from yesterday



Page 31: 2018-02-07. Updates and Prep meeting with Dan

General Updates:

1. Gave reproducible science workshop last Friday, 2018-02-02
    - overall many students benefitted from the structure I provided
    - some students didn't even show up with laptops -.-...what
    - not enough time to go over everything
        - went over the layout of a project (publishable unit)
        - got through github and highlighted how it could be used
        - didnt get a chance to go over R, Rmarkdown, or markdown in general
        - didnt get a chance to go over metadata,
        - showed the nice features of my online notebook
2. Gave popbio talk on range limits of A. picea last Friday,2018-02-02
    - Dan suggests an additional slide describing how we need to know the ecology and the organisms themselves to ultimately understand range limits
3. Been meeting with Taariq every week to parse trikinetics data and have it line up with each unique ID (fly); script works, but is slow and may need to run the script through hipergator

Prep for meeting with Dan.

1. Circadian rhythm project: ongoing
    - taking samples out of the fridge still.
    - still a lull in eclosions for RT treated flies, is this the fraction that has a 2 year diapause?
2. Paper readings

- XIE, X., MICHEL, A. P., SCHWARZ, D., RULL, J., VELEZ, S., FORBES, A. A., … FEDER, J. L. (2008). Radiation and divergence in the Rhagoletis Pomonella species complex: inferences from DNA sequence data. Journal of Evolutionary Biology, 21(3), 900–913. https://doi.org/10.1111/j.1420-9101.2008.01507.x
- Xie, X., Rull, J., Michel, A. P., Velez, S., Forbes, A. A., Lobo, N. F., … Feder, J. L. (2007). HAWTHORN-INFESTING POPULATIONS OF RHAGOLETIS POMONELLA IN MEXICO AND SPECIATION MODE PLURALITY. Evolution, 61(5), 1091–1105. https://doi.org/10.1111/j.1558-5646.2007.00091.x
- Feder, J. L., Berlocher, S. H., Roethele, J. B., Dambroski, H., Smith, J. J., Perry, W. L., … Aluja, M. (n.d.). Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis. Retrieved from http://www.pnas.org/content/pnas/100/18/10314.full.pdf

General issues/thoughts

Phylogenetic analyses assume non-reticulate evolution(gene flow between tips) , so it is not fully appropriate to describe populations.

- Papers need a structure analysis to:
    - determine extent of admixture
    - determine population structure
- PNAS paper needed an outgroup, but subsequent papers had it
- Parsimony trees? Need ML or bayesian approach to account for secondary mutations
- Papers need to show which allele is enriched in apple (should be SN haplotypes)
- To establish stronger direction of gene flow, they should run partitioned D statstics (Eaton paper in sys biol.)
- Try coalescant models? Ones that give you timing of divergence, then compare the timing of population of alleles?

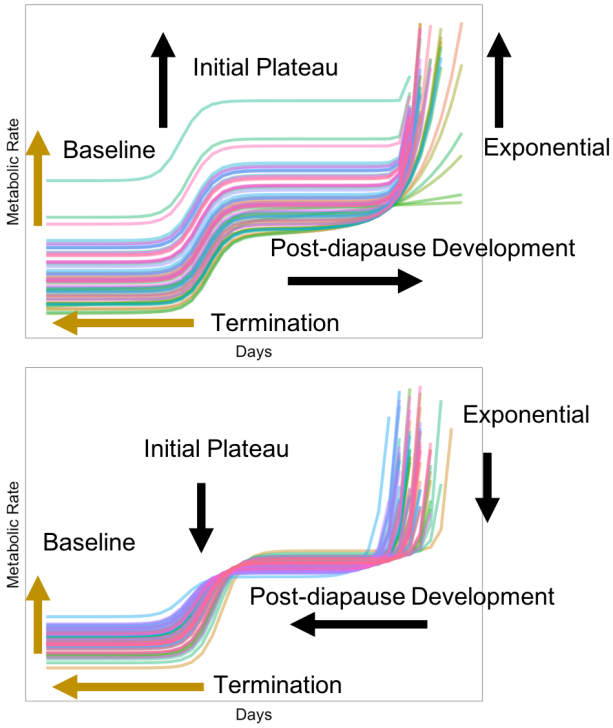
3. MR trajectory paper
    - Took a stab at intro
        - On the right track?
        - General outline:
            1. Speciation story -> divergence in emergence timing -> related to eclosion differences between host races
            2. Diapause developmental programme is highly modular with indpedent interconnected componnents that may or may not evolve independently í ragland biphasic model í related to termination, post-diapause development
            3. What drives the divergence in eclosion timing? Some say Deep diapause (Feder and Filchak), but what aspects can change?
            4. Shifts in earlier eclosion timing by apple can be achieved by becoming more responsive to favorable conditions. But how this happens within the diapause module framework remains unclear. Modules that can change :
                1. Elevate metabolic suppression during diapause
                2. Elevate metabolic rate upon meeting favorable conditions
                3. Terminate diapause earlier

4. Develop faster post-diapause, after termination
5. In this study paragraph.

```
* Data analysis- identified 2 modules with PCA
    * how is "c" influenced by resp
    * regress parameter estimates against parameter error to gauge bias
* How to proceed with Discussion?
```

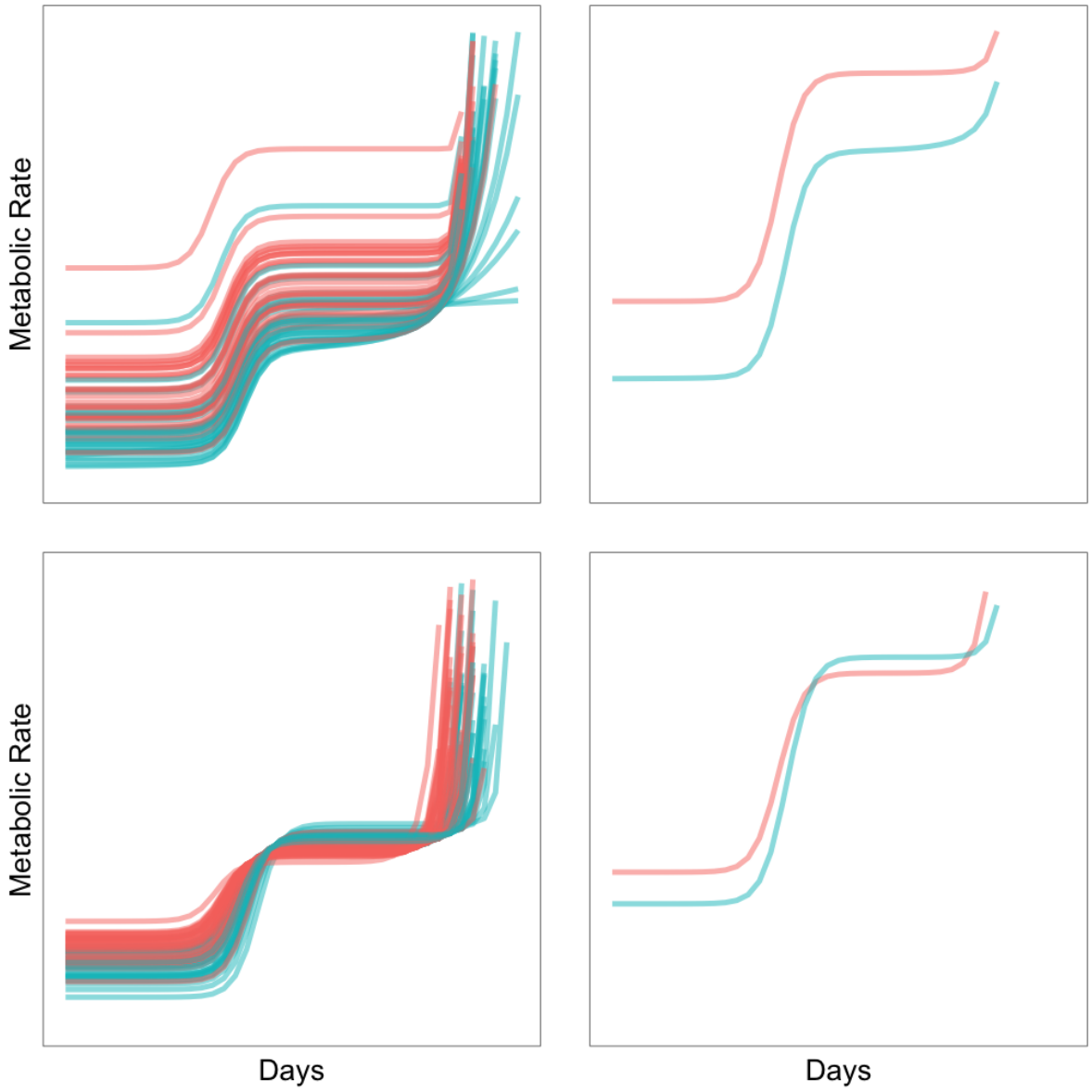
New figs:

PC1 and PC 2 comparison



	Comp.1	Comp.2
b	0.508	0.319
term	-0.330	-0.697
c	0.451	-0.162
post_dd	0.389	-0.600
plat	0.528	-0.161

Comparison among host races



4. Brain transcriptome paper
    - Greg needs to send me dataset with metadata
    - Dan needs to send me a the research plan for the project

---

Page 32: 2018-02-12. To do list

1. Biological rhythms project:
    - keep tracking eclosions and placing them into trikinetics set up.
    - Only 1 eclosion so far; eclosion needs to be calculated from days out of simulated overwintering
    - created a branch on github to preserve the project at today's time slice
2. Diapause exit proj
    - work on intro, hide detailed info on Rhago, make more broad and highlight modules in the context of diapause biology
3. Thermal niche paper:
    - read over, send to lacy
4. Hsp rxn norm paper:

- NJGotelli- wants to get rid of pca analysis and just show the correlation matrix + a multiple linear regression

5. Proteome stability project:
    - Wai sent me email asking for details- I replied
    - HPLC fractionation?
6. Biological rhythms conference
    - wrote up abstract ; give it to Dan to read.
    - Deadline is Feb 21st.
    - need to register

Page 33: 2018-02-13. Re-analysis of hsp rxn norm proj

NJG suggests that I should present a correlation matrix and use a multiple linear regression for testing the effect of hsp gxp parameters on CTmax.

One problem with multiple linear regression is that the variablies multicolinear: https://onlinecourses.science.psu.edu/stat501/node/347

Let's take a look at the dataset:

```
 str(jj2)
'data.frame':   41 obs. of  14 variables:
 $ CTmax              : num  42 41.8 40.4 41.3 40.9 ...
 $ hsc70              : num  23 22.1 22.5 22.9 23.8 ...
 $ hsp83              : num  22.4 22.1 22.3 22.4 21.6 ...
 $ hsp40              : num  24.4 24 25.1 25.2 25.4 ...
 $ colony.id2         : chr  "ALA1" "ALA4" "Avon19-1" "Bing" ...
 $ FC_hsc70_1468_max  : num  47.6 29.1 17 30.2 43.7 ...
 $ FC_hsc70_1468_slope: num  1.026 1.071 1.01 0.391 0.909 ...
 $ FC_hsc70_1468_Tm   : num  37.2 36.3 36.5 35.6 36.1 ...
 $ FC_hsp40_541_max   : num  8.5 4.62 21.96 5.79 7.06 ...
 $ FC_hsp40_541_slope : num  2.4311 1.3566 2.6054 0.8119 0.0455 ...
 $ FC_hsp40_541_Tm    : num  37.3 35.3 40.9 35.3 33.1 ...
 $ FC_Hsp83_279_max   : num  5.8 6.46 8.86 7.61 4.3 ...
 $ FC_Hsp83_279_slope : num  2.0964 1.2968 1.757 0.0497 1.0885 ...
 $ FC_Hsp83_279_Tm    : num  37 35 35.1 33.1 34.7 ...
 - attr(*, "na.action")=Class 'omit'  Named int [1:3] 12 24 43
  .. ..- attr(*, "names")= chr [1:3] "12" "24" "43"
```

Construct full linear model

```
fullmod101<-
lm(CTmax~FC_hsc70_1468_max+FC_hsc70_1468_slope+FC_hsc70_1468_Tm+FC_hsp40_541_max
+FC_hsp40_541_slope+FC_hsp40_541_Tm+FC_hsp40_541_Tm+FC_Hsp83_279_max+FC_Hsp83_27
9_slope+FC_Hsp83_279_Tm+hsc70+hsp83+hsp40,data=jj2)
knitr::kable(coef(summary(fullmod101)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.8001849	6.4812766	5.9865035	0.0000019
FC_hsc70_1468_max	0.0423297	0.0092943	4.5543913	0.0000937
FC_hsc70_1468_slope	0.1118718	0.2483733	0.4504179	0.6558775
FC_hsc70_1468_Tm	0.1079948	0.1667208	0.6477582	0.5224178
FC_hsp40_541_max	-0.0092906	0.0210301	-0.4417765	0.6620442
FC_hsp40_541_slope	-0.4047645	0.1487466	-2.7211684	0.0110578
FC_hsp40_541_Tm	0.2557764	0.0751077	3.4054617	0.0020140
FC_Hsp83_279_max	-0.0539052	0.0266171	-2.0252106	0.0524743
FC_Hsp83_279_slope	0.2774153	0.1308170	2.1206376	0.0429434
FC_Hsp83_279_Tm	-0.1136904	0.1031364	-1.1023302	0.2797033
hsc70	-0.0963486	0.1741470	-0.5532604	0.5844761
hsp83	-0.0486555	0.1294554	-0.3758475	0.7098629
hsp40	-0.1742460	0.0642824	-2.7106345	0.0113385

Checking for multicollinearity

```
vif(fullmod101)
  FC_hsc70_1468_max FC_hsc70_1468_slope     FC_hsc70_1468_Tm     FC_hsp40_541_max
 FC_hsp40_541_slope      FC_hsp40_541_Tm
           3.534310             2.668820             5.039916             3.097313
           2.447423             4.714042
   FC_Hsp83_279_max  FC_Hsp83_279_slope      FC_Hsp83_279_Tm                 hsc70
                hsp83                hsp40
           3.488256             2.150595             6.700810             1.914794
           1.807223             1.162075
```

Generally, from the online stat tutorial, variance inflation factors above 4 requires further investigation.

Model selection

```
fullmod102<-stepAIC(fullmod101,direction="both")
knitr::kable(coef(summary(fullmod102)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.7068907	1.9933118	18.415027	0.0000000
FC_hsc70_1468_max	0.0436305	0.0064964	6.716152	0.0000001
FC_hsp40_541_slope	-0.3756351	0.1227514	-3.060128	0.0042986
FC_hsp40_541_Tm	0.2198404	0.0466206	4.715519	0.0000400
FC_Hsp83_279_max	-0.0694954	0.0214980	-3.232640	0.0027264
FC_Hsp83_279_slope	0.2799913	0.1040755	2.690271	0.0109871
hsp40	-0.1719737	0.0585296	-2.938233	0.0058917

2018-02-14 update

NJG comments:

> I strongly suggest you drop the PCA entirely. As is always the case, it really doesn't add anything beyond the underlying correlations of the variables. I would suggest a pairwise correlation matrix (or figure from R) for all of the variables, and then perhaps a multiple regression model for each HSP with CtMax as the response variable and the four expression parameters as the predictor variables.

Yeah, lets try this again...

Cool code to make correlation matrix heat map

```
#make correlation matrix
comat<-round(cor(na.omit(jj2)),2)

##plot it with this library
library(ggcorrplot)
b<-ggcorrplot(comat, hc.order = TRUE,lab=TRUE, type = "lower",
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"))

b+theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),)
```

Stat models with hsp83

```
 summary(mod83)
```

```
Call:
lm(formula = KO_temp_worker ~ hsp83 + FC_Hsp83_279_slope + FC_Hsp83_279_Tm + 
    FC_Hsp83_279_max, data = jj1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.53807 -0.38594  0.01758  0.32602  2.01071 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        34.947040   5.117156   6.829 3.68e-08 ***
hsp83               0.008037   0.153847   0.052   0.9586    
FC_Hsp83_279_slope  0.084438   0.152880   0.552   0.5839    
FC_Hsp83_279_Tm     0.172873   0.084410   2.048   0.0473 *  
FC_Hsp83_279_max    0.019388   0.029765   0.651   0.5186    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7408 on 39 degrees of freedom
Multiple R-squared:  0.2931,    Adjusted R-squared:  0.2206 
F-statistic: 4.043 on 4 and 39 DF,  p-value: 0.007751
```

Stat models with hsp70

```
 summary(mod70)

Call:
lm(formula = KO_temp_worker ~ hsc70 + FC_hsc70_1468_max + FC_hsc70_1468_slope + 
    FC_hsc70_1468_Tm, data = jj1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.94204 -0.44307  0.07442  0.34417  1.02051 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         37.874723   5.735742   6.603 7.54e-08 ***
hsc70               -0.244822   0.142811  -1.714  0.09441 .  
FC_hsc70_1468_max    0.022367   0.008156   2.743  0.00916 ** 
FC_hsc70_1468_slope  0.322339   0.219113   1.471  0.14928    
FC_hsc70_1468_Tm     0.220416   0.120014   1.837  0.07390 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.594 on 39 degrees of freedom
Multiple R-squared:  0.5456,    Adjusted R-squared:  0.499 
F-statistic: 11.71 on 4 and 39 DF,  p-value: 2.435e-06
```

Stat models with hsp40

```
summary(mod40)

Call:
lm(formula = KO_temp_worker ~ hsp40 + FC_hsp40_541_max + FC_hsp40_541_slope + 
    FC_hsp40_541_Tm, data = jj1)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.0548 -0.3409 -0.1021  0.3422  1.8449

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        37.51526    3.15733  11.882 5.12e-14 ***
hsp40              -0.12503    0.08879  -1.408  0.16766
FC_hsp40_541_max    0.01542    0.02583   0.597  0.55435
FC_hsp40_541_slope -0.19096    0.19562  -0.976  0.33550
FC_hsp40_541_Tm     0.20127    0.07177   2.804  0.00808 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7137 on 36 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.3075,    Adjusted R-squared:  0.2305
F-statistic: 3.996 on 4 and 36 DF,  p-value: 0.008775
```

Multiple linear regression Models deviate from models considering 1 predictor at a time.

Page 34: 2018-02-14. Update to do list

1. Biological rhythms project:
    - keep tracking eclosions and placing them into trikinetics set up.
2. **Diapause exit proj**
    - work on intro, hide detailed info on Rhago, make more broad and highlight modules in the context of diapause biology
        - almost done, need to look at it one more time and then send to Dan.
3. **Thermal niche paper**:
    - **Sent to lacy**; tweaked abstract and figures
4. **Hsp rxn norm paper**:
    - NJGotelli- wants to get rid of pca analysis and just show the correlation matrix + a multiple linear regression
        - multiple linear regression diff than single regressions ; asking for NJG for response on how to proceed
        - SHC is in favor of pca bc it separates out expression kinetics from baseline gxp.
        - Fixed pca figure to make it more interprettable.
5. Proteome stability project:
    - Wai to HPLC fractionate
6. **Biological rhythms conference**
    - Abstract basically approved by Dan
    - Deadline is Feb 21st.
    - need to register
    - go over abstract 1more time and send it to Dan ---Friday
7. meeting with Brett Scheffers ; Friday Feb 16 - 2:30PM
    - agenda
        - reciprocal transplant along lat gradient in Fl ; fruit flies in madagascar

---

Page 35: 2018-02-15. Paper reading: Melo et al. Ann rev; Modularity Genes Development and Evolution

ref:

Melo, D., Porto, A., Cheverud, J. M., & Marroig, G. (2016). Modularity: Genes, Development, and Evolution. Annu. Rev. Ecol. Evol. Syst, 47, 463–86. https://doi.org/10.1146/annurev-ecolsys-121415-032409

Found new paper on modularity and trying to get a handle on what it actually means.

They describe a system being modular as:

If it can be divided into multiple sets of strongly interacting parts that are relatively autonomous with respect to each other.

For diapause, what is a module then? Is the whole diapause program 1 large module, with preparation, maintenance, and exit as distinct components. But for Rhagoletis, diapause exit can be broken down into a finer scale, such as termination and post diapause development.

Modularity can be applied to developmental processes; where modules are diff parts of embryo that interact with each other (induction and morphogenesis) or sets of interacting molecules that act independently in the patterning of multiple tissues.

The same concept can be extended to adults, where modules consist of parts that act together in the performance of some physiological function.

Variational modules are sets of traits that vary together and somewhat independently from other modules. Wait what...

The way to approach modularity - estimate its correlational structure...not anything new here

In this paper, they use genetic integration as genetic correlation. I should use the buzz word integration.

Interacting sentence here:

```
If all features of an organism are completely integrated, the parts will
be prevented from evolving independent adaptations.
```

Trait correlation has substantial consequences for their evolutionary trajectories.

This paper uses morphology a lot as an example of modularity One way

Also reading Melo & Marroig 2015; PNAS . Directional selection can drive the evolution of modularity in complex traits

ref: Melo, D., & Marroig, G. (2015). Directional selection can drive the evolution of modularity in complex traits. Proceedings of the National Academy of Sciences of the United States of America, 112(2), 470–5. https://doi.org/10.1073/pnas.1322632112

Thispaper has a good description of modularity :

Tendency for parts of many biological systems to be organized into semi-independent groups

They argue that response to selection depends on the patterns of modularity.

They approach modularity with PCA analysis of traits.

How traits are correlated is not depndent on drift and stabilizing selection. This makes sense when I think about Paccard et al. 2016(year?) where they find more traint independence at the edge of A. lyrata's range limit. They didn't investigate how selection is operating at the range limit. So, greater traint independence could be due to drift and that selection isn't operating at the range boundaries at all.

Instead, their simulations show that directional selection impacts modular structure of traits. When considering both stab and direct selection, stab is important for maintaining established trait correlation patterns.

Fig 2B. One cool interesting result if I'm understanding it right. The first pc explains consistant variation with increasing directional selction, but the second PC's variation increases with directional selection.

What they do?

They simulated 10x10 (10 traits) G matrices that are controlled by many loci in a pouplation, that are subjected mutation, recombination. They essentially did a PCA of the gmatrix to identify trait correlations reflective of 2 modules

Approach THey modelled populations under drift, stabilizing selection, and divergent natural selection.

The way they're describing module with respect to simulations is confusing. They split up the correlation within a PC(pc1 for example) such that negative or positive ones form the module.

In the methods, they describe" populations having variational modularity when the correlation between some sets of traits belonging to a module is higher than the correlation between traits of different modules. "

So they determine modularity as the AVG ratio: ratio between the within and between module average correlation.

Page 36: 2018-02-21. Updated to do list

1. Biological rhythms project:
    - Taariq sent me prelim data set, where he's associated unique ID, entrainment vs free run to the trikinetics data
        - I need to bin at different time intervals and estimate dominant periods with a spectral analysis
    - Create talk for biological rhythms conference - all set up with paper work and registration
2. **Diapause exit proj**
    - in Dan's hands
3. **Hsp rxn norm paper**:
    - Fixed pca figure to make it more interprettable.
        - redid some figures to make PCA more interpretable; fixed intro, need to move onto results
    - Make outline for discussion
4. Leading data wrangling workshop March 2 2018 for UF grad students
    - Set up rproj, simulate data
    - Manipulating data from long to wide format

- Make associated plots (Kmiller's data would be nice)

notes:

SHC wants me to present regression on deciduos forest species :

```
summary(lm(kdsub$jj.KO_temp_worker~kdsub$pchsp.scores...1.))

Call:
lm(formula = kdsub$jj.KO_temp_worker ~ kdsub$pchsp.scores...1.)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.01303 -0.29330  0.08219  0.34777  0.74836 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             41.47272    0.10029 413.525  < 2e-16 ***
kdsub$pchsp.scores...1. 0.22837    0.06394   3.571  0.00141 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4485 on 26 degrees of freedom
Multiple R-squared:  0.3291,    Adjusted R-squared:  0.3033 
F-statistic: 12.76 on 1 and 26 DF,  p-value: 0.001414
```

Page 37: 2018-02-22. Flow of ideas for the discussion of the hsp rxn norm ms

Struggling to write the discussion here, so I'm writing out my ideas to get a better sense of where I want to go.

Current structure:

1. Summary paragraph
2. Forest heterogeneity among forest types, focusign on deciduous forests
3. Flat woods thermal experiences
4. Evolutionary transition between closed to open habitat was accompanied by dynamic changes in the HSR
5. Cellular thermostat model discussion
6. Reason why hsp kinetics differ and not baseline
7. Conclusion- bringing it back to climate change

SHC comment on discussion

```
Is this really the key result?  Or is it that you have demonstrated a
surprisingly large role of Hsp expression evolution in this process that is new
and exciting?  I am not claiming I know what that key fnding is, but this spot
is the most important location in your entire paper.  Think very carefully about
what you want to put here.
```

Ok, lets try to restructure it:

1. Summary paragraph

    - Colonizing thermally stressful habitats requires ability to behav and phys adjust to stress
    - Maybe contrast between divergent habitats where it is expected for baseline expression to increase, and non necessarily kinetics. I also should think about why other groups may not find strong differences in kinetics-it could be due to unsampled variation or missing variation that can be best captured with a function valued trait approach
    - Summary of results - CT max and habits
    - Maybe say something about how this study is good because it decomposes the variation at different levels of biological organization
    - may need to mention the life history of ants: they usually stay in place and colony persistance willd epend on their ability to take the heat. Although, their offspring can migrate northward. Oh, maybe discuss how thermally adapted alleles can move northward.
2. How thermal ecology shapes thermal traits, the direction we expect them to evolve and what types of mechanisms we'd expect to underlie their evolution. For example, how do organisms extend CTmax?
    - How clades evolve CTmax? Constrained? Largest signal was from habitats, not latitude.
    - hit on the fact that ants are long lived and display low plasticity for upper thermal limits (which is true for most insects), so they have to rely on evolution .
    - compare and contrast temperate regions, with tropical regions. We'd expect open habitats to resemble mechanisms in tropical regions perhaps because they experience chronic temperature stress
        - think about how to integrate sub saharan african ants and their thermal phys
3. Transition into flat woods, accompanied by changes in kinetics, not baseline. Counter to what we expect.
    - End on explained variation
    - touch on the remaining variation that could be explained
4. Discuss in the context of the thermostat model
5. Different strategies to cope with heat stress may have facilitated niche expansion in terrestrial ectotherms.
    - compare baseline expression used in sub saharan african ants, vs temperate forest ants
    - variability selects for plasticity
    - might be aproblem for cliamte change, if temperatures are increasing and becoming more stable.
    - Ants colonies don't move very far within their lifetime, so they have to take the heat where they are

---

Page 38: 2018-02-27. Reading Stinchcombe et al. 2010; *Evolution*, ACROSS-ENVIRONMENT GENETIC CORRELATIONS AND THE FREQUENCY OF SELECTIVE ENVIRONMENTS SHAPE THE EVOLUTIONARY DYNAMICS OF GROWTH RATE IN IMPATIENS CAPENSIS

ref: Stinchcombe, J. R., Izem, R., Heschel, M. S., McGoey, B. V., & Schmitt, J. (2010). ACROSS-ENVIRONMENT GENETIC CORRELATIONS AND THE FREQUENCY OF SELECTIVE ENVIRONMENTS SHAPE THE EVOLUTIONARY DYNAMICS OF GROWTH RATE IN IMPATIENS CAPENSIS. Evolution, no-no. https://doi.org/10.1111/j.1558-5646.2010.01060.x

Cool paper by John Stinchcombe. Take home: the expression of the genetic architecture can change depending on the environment. In this case, they manipulated plants grown in shade or sun; at different densities.

table 4 is interesting and something I've been thinking about. If you plot each PC loading vs density , color coded by shade and sun. You can basically see the shift in the genetic architecture.

Page 39: 2018-02-28. Meeting with Tom, diapause exit ms

Meeting about data analysis. My quick and dirty approach found that the error of the estimate correlated with the estimates themselves. How can we control/account for this?

1. Regress estimate on error of the estimates. lm(Estimate ~ Error)
2. Deming regression, where you can do single regressions and include measurement error.

Action items:

1. Tom to send us dataset with updated parameters and error
2. I will re-analyze and explore whether the error matters
    - If it does, try approaches outlined above
3. Tom will read the intro and provide feedback
4. Set the intro, methods, and results. Send to Greg for feedback.

Page 40: 2018-03-01. discussion, hsp rxn norm ms

Tweaking outline;

1. Summary paragraph
    - Colonizing thermally stressful habitats requires ability to behav and phys adjust to stress
    - Maybe contrast between divergent habitats where it is expected for baseline expression to increase, and non necessarily kinetics. I also should think about why other groups may not find strong differences in kinetics-it could be due to unsampled variation or missing variation that can be best captured with a function valued trait approach
    - Summary of results - CT max and habits
    - Maybe say something about how this study is good because it decomposes the variation at different levels of biological organization
    - may need to mention the life history of ants: they usually stay in place and colony persistance willd epend on their ability to take the heat. Although, their offspring can migrate northward. Oh, maybe discuss how thermally adapted alleles can move northward.

2. How thermal ecology shapes thermal traits, the direction we expect them to evolve and what types of mechanisms we'd expect to underlie their evolution. For example, how do organisms extend CTmax?
    - How clades evolve CTmax? Constrained? Largest signal was from habitats, not latitude.
    - hit on the fact that ants are long lived and display low plasticity for upper thermal limits (which is true for most insects), so they have to rely on evolution .
    - compare and contrast temperate regions, with tropical regions. We'd expect open habitats to resemble mechanisms in tropical regions perhaps because they experience chronic temperature stress
        - think about how to integrate sub saharan african ants and their thermal phys
3. Transition into flat woods, accompanied by changes in kinetics, not baseline. Counter to what we expect.
    - highlight function valued trait approach and how it can effectively detect subtle differences in rxn norm
    - End on explained variation
    - touch on the remaining variation that could be explained
4. **This should lead to other mechanisms -- focus on proteome stability and mention cool follow up approaches**
    - cite the 2 papers that have proteome stability data out there
    - protein homeostasis is also a function of protein degradation, and hsp chaperoning activity
5. Different strategies to cope with heat stress may have facilitated niche expansion in terrestrial ectotherms.
    - compare baseline expression used in sub saharan african ants, vs temperate forest ants
    - variability selects for plasticity
    - might be aproblem for cliamte change, if temperatures are increasing and becoming more stable.
    - Ants colonies don't move very far within their lifetime, so they have to take the heat where they are

Ok, what are my topic and ending sentences for each paragraph ? s

1. Colonization of thermally stressful environments depends on the ability to behaviorally and physiologically cope with temperature stress
    - Physiological adaptation played an important role in facilitating niche expansion in aphaeno
    - transition from closed to open forests was accompanied by increase of upper thermal limits
    - We evaluated the ability of these ants to resist and tolerate thermal perturbations and found shifts relating to both using a function valued trait approach and not in the types of mechanisms expected.
    - We found dynamic ,coordinated shfits in the Hsp expression parameters, not baseline expression , which also related to species divergence between habitats, and also when removing open canopy forested colonies; using a function valued trait approach
    - Aphaenos likely use a common mechanism to cope with temperature stress in temperate forest ecosystems, opposite what is found in extremely warm

habitats(cataglyphis).

2. Temperate Forest environments generate considerable heterogeneity in thermal microclimates that influences the thermal experiences of ectotherms (Scheffers et al. 2014).
    - our largest signal in shifts in ctmax were due to habitat, and less so latitude
        - solar radiation is htting forest dwelling insects more in open than closed habitat
        - closed canopy forest species may benefit from bogert effect and not open canopy forest
    - Expectations for the types of mechanisms to evolve; basal vs induced
3. Transition into open canopy forests and overall extension of CTmax was accompanied by kinetics(induced responses) rather than baseline, counter to what we expect
    - By decomposing gene expression into functional parameters, we had high power to uncover variation in the expression pattersn fo hsps
    - this could be why, some studies fail to see divergence in induction rather than baseline
    - Altogether kinetics explained less than half of the variation, suggesting other mechanisms are important, such as proteome , metabolome, and cell membrane dynamics
4. Protein homeostasis good candidate to understand a more full picture of temperature adaptation (is it the source for resistance and tolerance? resistance and tolerance depends on homeostasis?)
    - depends on the functional stability of the proteome
        - degradation
        - translation
        - chaperoning
    - follow up approaches: multiplex proteomics to find melting temperatuers
    - How do all of these mechanisms interact?
5. Different strategies to cope with heat stress may have facilitated niche expansion in terrestrial ectotherms
    - tempearate systems should evolve induction, troopical systems should evolve baseline
    - it is costly to constantly upregulate
    - variability selects for plasticity and
    - Ants don't move, so they must rely on current mechanisms
    - this may be a problem for future climate change when we expect more chronicly higher temperatures

ok, now to fill in the whole discussion...

Page 41: 2018-03-02. Dissecting out abstracts Amnat

ref:

Angert, A. L., Bayly, M., Sheth, S. N., & Paul, J. R. (2017). Testing Range-Limit Hypotheses Using Range-Wide Habitat Suitability and Occupancy for the Scarlet Monkeyflower (Erythranthe cardinalis). The American Naturalist, E000–E000. https://doi.org/10.1086/695984

Determining the causes of geographic range limits is a fundamental problem in ecology, evolution, and conservation biology. Range limits arise because of fitness and dispersal limitation, which yield contrasting predictions about habitat suitability and occupancy of suitable habitat across geographic ranges. If a range edge is limited primarily by fitness, occupancy of

suitable habitat should be high, habitat suitability should decline toward the edge, and no suitable habitat should exist beyond it. In contrast, a range edge limited primarily by dispersal should have unoccupied but suitable habitat at and beyond the edge. We built ecological niche models relating occurrence records for the scarlet monkeyflower (Erythranthe cardinalis) to climatic variables and applied these models to independent data from systematic, range-wide surveys of presence and absence to estimate the availability and occupancy of climatically suitable habitat. We found that fitness limitation predominated over dispersal limitation, but dispersal limitation also played a role at the poleward edge. These results are consistent with the hypothesis that dispersal limitation is more important along shallow environmental gradients and also suggest that synergy between dispersal and fitness limitation can contribute to colonization failure. The framework used here is validated by independent data and could be readily applied to inferring causes of range limits in many other species.

THe first two sentences describe theory and background info.

Determining the causes of geographic range limits is a fundamental problem in ecology, evolution, and conservation biology. Range limits arise because of fitness and dispersal limitation, which yield contrasting predictions about habitat suitability and occupancy of suitable habitat across geographic ranges.

Expectations based on theory and background info:

If a range edge is limited primarily by fitness, occupancy of suitable habitat should be high, habitat suitability should decline toward the edge, and no suitable habitat should exist beyond it. In contrast, a range edge limited primarily by dispersal should have unoccupied but suitable habitat at and beyond the edge.

So they've used the first 4 sentences to set up the background and expectations. Next, they get into their analytical approach and end with the justification.

We built ecological niche models relating occurrence records for the scarlet monkeyflower (Erythranthe cardinalis) to climatic variables and applied these models to independent data from systematic, range-wide surveys of presence and absence to estimate the availability and occupancy of climatically suitable habitat.

They state the result.

We found that fitness limitation predominated over dispersal limitation, but dispersal limitation also played a role at the poleward edge.

They put their results in a greater context and broaden to how these methods can be applied to other systems.

These results are consistent with the hypothesis that dispersal limitation is more important along shallow environmental gradients and also suggest that synergy between dispersal and fitness limitation can contribute to colonization failure. The framework used here is validated by independent data and could be readily applied to inferring causes of range limits in many other species.

---

ref: Burgess, S. C., Snyder, R. E., & Rountree, B. (2017). Collective Dispersal Leads to Variance in Fitness and Maintains Offspring Size Variation within Marine Populations. The American Naturalist, 000–000. https://doi.org/10.1086/695879

Variance in fitness is well known to influence the outcome of evolution but is rarely considered in the theory of marine reproductive strategies. In coastal environments, turbulent mesoscale eddies can collect larvae into packets, resulting in collective dispersal. Larvae in packets return to the coast or are lost offshore in groups, producing variance in fitness. Using a Markov process to calculate fixation probabilities for competing phenotypes, we examine the evolution of offspring size and spawning duration in species with benthic adults and pelagic offspring. The offspring size that provides mothers with the highest mean fitness also generates the greatest variance in fitness, but pairwise invasion plots show that bet-hedging strategies are not evolutionarily stable; maximizing expected fitness correctly predicts the unique evolutionarily stable strategy. Nonetheless, fixation can take a long time. We find that selection to increase spawning duration as a risk avoidance strategy to reduce the negative impacts of stochastic recruitment success can allow multiple offspring sizes to coexist in a population for extended periods. This has two important consequences for offspring size: (1) coexistence occurs over a broader range of sizes and is longer when spawning duration is longer because longer spawning durations reduce variation in fitness and increase the time to fixation, and (2) longer spawning durations can compensate for having a nonoptimal size and even allow less optimal sizes to reach fixation. Collective dispersal and longer spawning durations could effectively maintain offspring size variation even in the absence of good and bad years or locations. Empirical comparisons of offspring size would therefore not always reflect environment-specific differences in the optimal size.

This abstract starts directly off with the problem.

Variance in fitness is well known to influence the outcome of evolution but is rarely considered in the theory of marine reproductive strategies.

Then goes over background info on larvae and how there potentially variance in fitness.

In coastal environments, turbulent mesoscale eddies can collect larvae into packets, resulting in collective dispersal. Larvae in packets return to the coast or are lost offshore in groups, producing variance in fitness.

Next, the methods or approach sentence where they're modelling competing phenotypes related to fitness

Using a Markov process to calculate fixation probabilities for competing phenotypes, we examine the evolution of offspring size and spawning duration in species with benthic adults and pelagic offspring.

Results sentences-

The offspring size that provides mothers with the highest mean fitness also generates the greatest variance in fitness, but pairwise invasion plots show that bet-hedging strategies are not evolutionarily stable; maximizing expected fitness correctly predicts the unique evolutionarily stable strategy. Nonetheless, fixation can take a long time. We find that selection to increase spawning duration as a risk avoidance strategy to reduce the negative impacts of stochastic recruitment success can allow multiple offspring sizes to coexist in a population for extended periods.

What the results mean-

This has two important consequences for offspring size: (1) coexistence occurs over a broader range of sizes and is longer when spawning duration is longer because longer spawning durations reduce variation in fitness and increase the time to fixation, and (2) longer spawning durations can compensate for having a nonoptimal size and even allow less optimal sizes to reach fixation.

Broadening out the larger context-

Collective dispersal and longer spawning durations could effectively maintain offspring size variation even in the absence of good and bad years or locations. Empirical comparisons of offspring size would therefore not always reflect environment-specific differences in the optimal size.

---

ref:
Burgess, S. C., Snyder, R. E., & Rountree, B. (2017). Collective Dispersal Leads to Variance in Fitness and Maintains Offspring Size Variation within Marine Populations. The American Naturalist, 000–000. https://doi.org/10.1086/695879

Climate change is expected to favor smaller-bodied organisms through effects of temperature on physiological performance and food-web interactions, so much so that smaller body size has been touted as a universal response to global warming alongside range shifts and changing phenology. However, climate change involves more than warming. It is multivariate, and the interplay between climate variables may result in less straightforward predictions. We present a model that considers the simultaneous effect of multiple variables (temperature, CO2, and moisture) on herbivore body sizes within a tritrophic food web comprised of vegetation, herbivores, and a shared predator. The model accounts for climate effects on animal behavior, plant and animal metabolism, and plant quality to explore emergent effects on herbivore body size. Our analysis reveals that some common multivariate climate change scenarios may favor larger-bodied herbivores, challenging previous findings of shifts toward small-bodied herbivores in the face of rising temperatures.

Opening with background info, to set up a contrast in the next sentence.

Climate change is expected to favor smaller-bodied organisms through effects of temperature on physiological performance and food-web interactions, so much so that smaller body size has been touted as a universal response to global warming alongside range shifts and changing phenology.

Presents the problem (that climate change will impose selection from many different sources) and contrasts the previous sentence.

However, climate change involves more than warming. It is multivariate, and the interplay between climate variables may result in less straightforward predictions.

Approach and methods sentence

We present a model that considers the simultaneous effect of multiple variables (temperature, CO2, and moisture) on herbivore body sizes within a tritrophic food web comprised of vegetation, herbivores, and a shared predator. The model accounts for climate effects on animal behavior, plant and animal metabolism, and plant quality to explore emergent effects on herbivore body size.

Ends with a results sentence and no real conclusion. Maybe the result is the conclusion. But we have no clue under what variables can lead to larger bodied herbivores to do well...

Our analysis reveals that some common multivariate climate change scenarios may favor larger-bodied herbivores, challenging previous findings of shifts toward small-bodied herbivores in the face of rising temperatures.

Thoughts

Abstracts for Amnat present background that lead to a gap in knowledge or problem or expectation. They usually have 1 methods/approach sentence and variable results/conclusion sections depending on what they want to highlight.

Page 42: 2018-03-06. Updates and proteostasis project development/ideas

Met with Dan yesterday to talk about project ideas for undergrad in the summer. The project focuses on dissecting out the evolutionary tactics organisms use to cope with heat stress at the molecular level.

Action items:

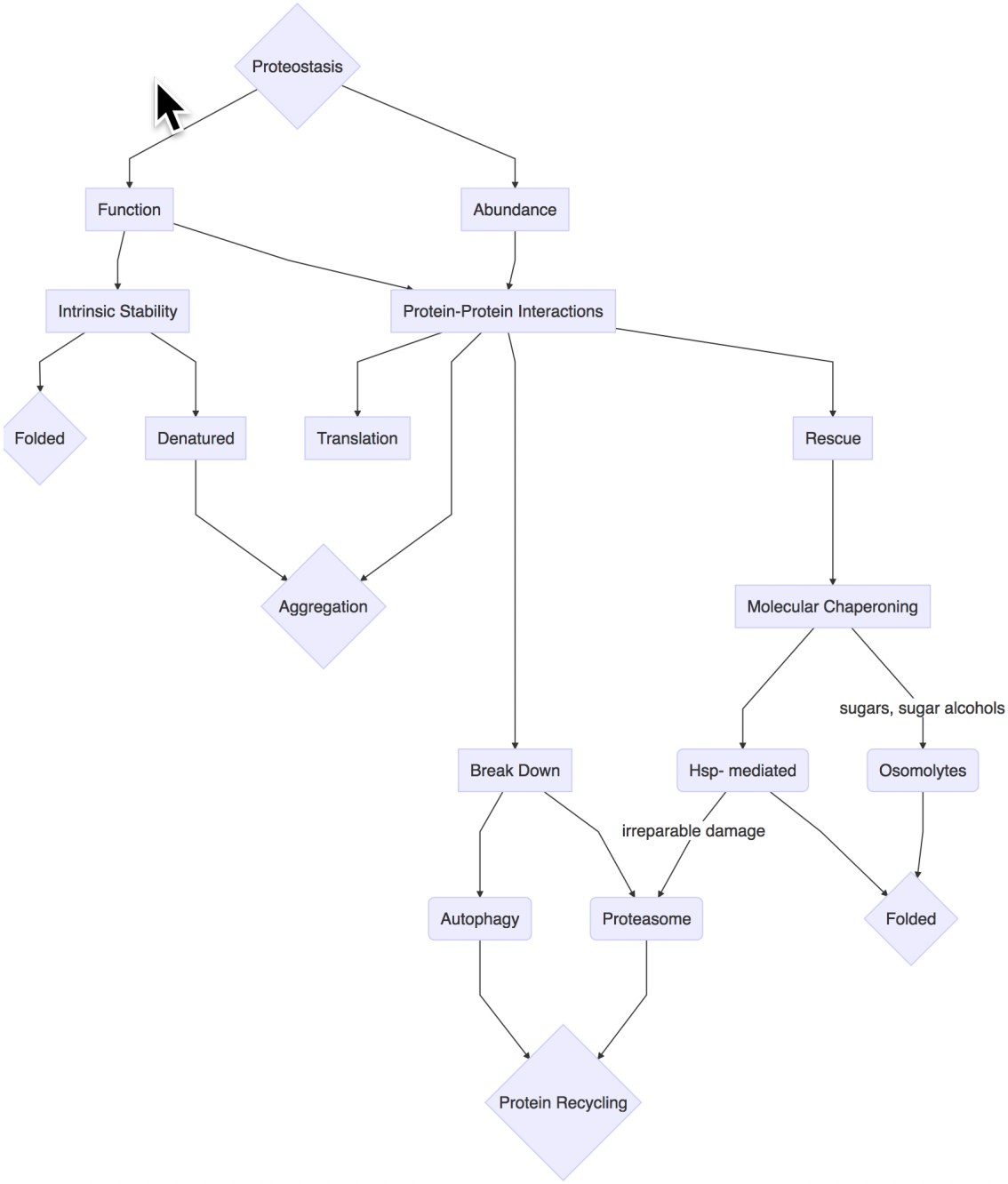
1. Send Dan two papers on proteome stability

refs:

- Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P. J., von Mering, C., … Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science (New York, N.Y.), 355(6327), eaai7825. https://doi.org/10.1126/science.aai7825
- Mikhail Savitski, P. M., M Reinhard, F. B., Franken, H., Werner, T., Fälth Savitski, M., Eberhard, D., … Drewes, G. (n.d.). Tracking cancer drugs in living cells by thermal profiling of the proteome. https://doi.org/10.1126/science.1255784

2. Scour the literature to see what is known, what has been done, to figure out how we can contribute to the field. I'm making a google sheets to organize the literature, here.
    - Dan wants to do something with prediction to try to pitch the idea for NIH funding.
    - What has been done in the DGRP?

Flow diagram of the determinants of proteostasis (protein homeostasis)



Types of questions we can ask

1. What are the relative contributions of different molecular tactics underlying proteostasis under temperature stress?
2. What evolutionary tactics at the molecular level enable adaptive changes in thermal tolerance?

Hypotheses

- Organisms rely on more stable proteome (resistance) than induced responses (tolerance) to cope with heat stress.

Page 43: 2018-03-09 & 2018-03-12. Proj updates and proteome stability proj development

1. Hsp rxn norm ms-Waiting on abstract, intro, results edits from SHC and NJG

2. THermal niche ms- Lchick to send out
3. Range limits ms- Waiting on discussion edits from SHC and NJG
4. Biological rhythms - still tracking eclosions from simulated overwintering
5. Rhagoletis Brain transcriptome proj- Dan to send me background info on the proj
6. Proteome stability proj- write up a research plan

WRiting up research plan.

Background:

The way organisms cope with temperature stress is multi-faceted and dynamical. Temperature stress can vary over the short and long term over an organism's lifespan. Ultimately, this stress causes macromolecular damage that disrupts biological function and can lead to negative fitness consequences.

Therefore, organisms should respond to temperature stress by diminishing the encounter of macromolecular damage (resistant) and/or controlling macromolecular damage (tolerance). One of the most important macromolecules are proteins, which provide structural and enzymatic functions inside the organism and the degree to which resistance and tolerance mechanisms are important depend on the magnitude and duration of stress. Short term heat bouts selects for resistance tactics such as evolving greater protein stability, while long term heat stress selects for tolerance by changing protein expression, or degrading and rescuing protein function. How organisms utilize a set a of tactics at the molecular level to face short term or long term stress is poorly understand.

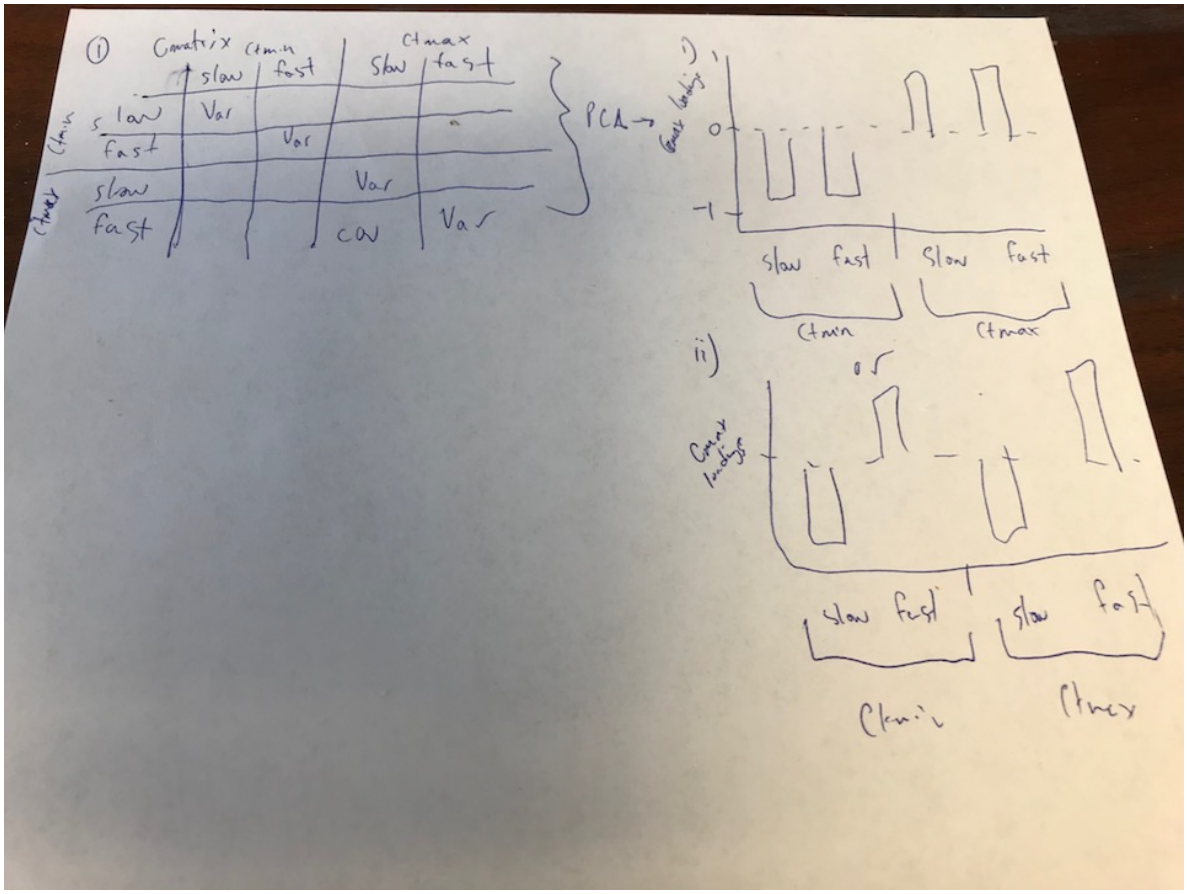
However, quantitative genetic analyses of upper thermal limits (in fruit flies) suggests independent and unique genetic components between dealing with hot temperatures over short and longer time scales.

Study System: Fruit flies

- DGRP http://dgrp2.gnets.ncsu.edu/
- DSPR http://wfitch.bio.uci.edu/~dspr/

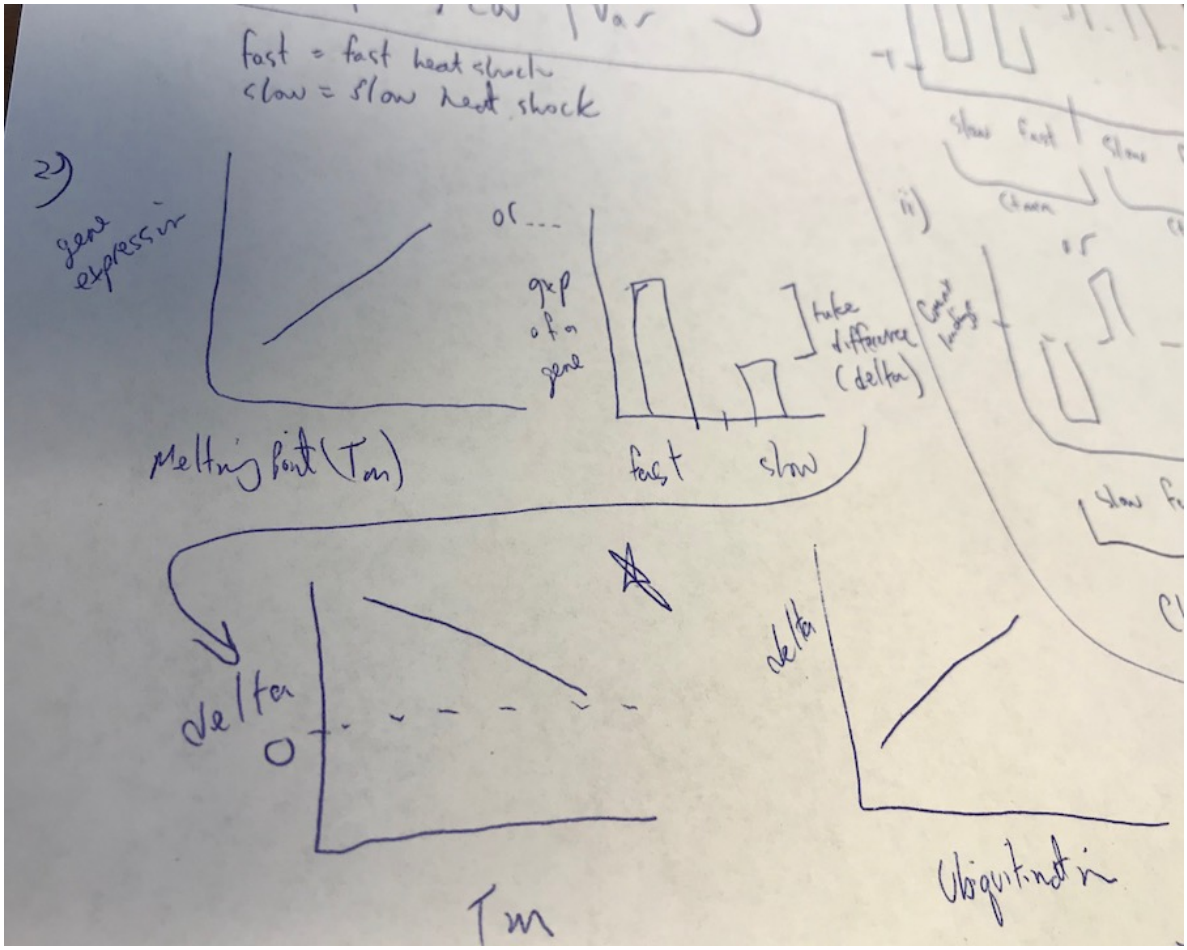
Main Questions:

1. What is the arrangement of the genetic architecture of thermal performance?
    - **Hypothesis:** Thermal limits under short term and long term stress share common and unique genetic variation.
        - **Approach:** Measure Ctmax and CTmin under fast ramp, slow ramp, hardening ability and estimate the genetic architecture (variance-covariance G matrix)
        - **Expected outcomes:** PCA decomposition of the Gmatrix will show different trait correlations.
            1. All of the CTmax related traits could be negative correlated with CTmin along the first PC.
            2. On the other hand, all of the treatment types (slow , fast, hardening) could share the same genetic arthicture.



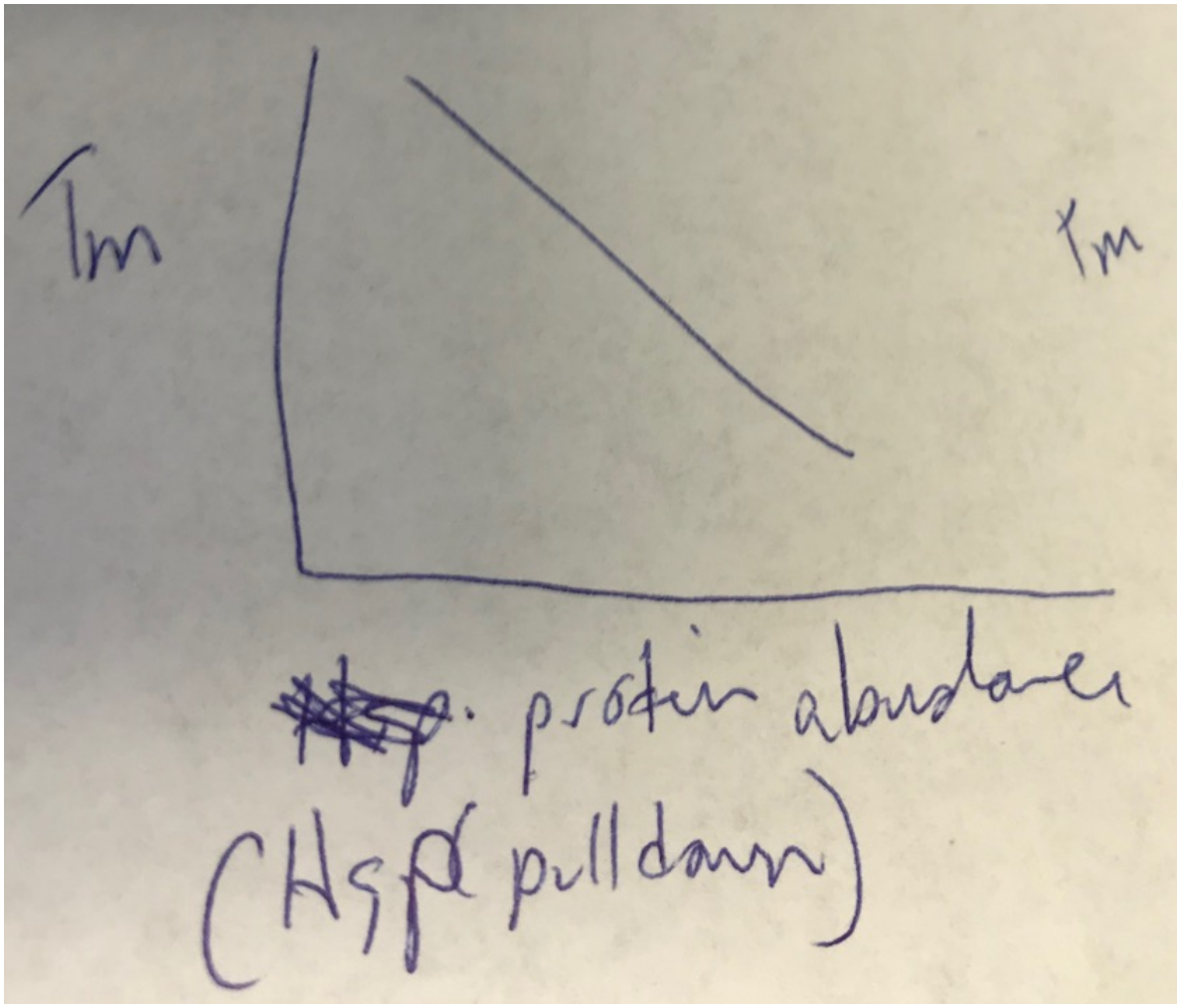
Ultimate inference-- There is common and divergent genetic basis of thermal traits This leads to the next question...

2. What are the relative contributions of molecular level tactics that organisms can use to respond to temperature stress?
    - **Hypothesis:** Thermally labile proteins are either not expressed(or expression becomes altered) and become broken down rapidly under an acclimation response.
        - **Approach:** Measure the thermal stability of using multiplex proteomics under short and long term heat stress and determine the shift in relative abundance of proteins. The mass spec will give the melting curves of every protein that can be detected.
        - **Expected outcomes:**
            - proteins with lower Tm (melting point) have lower expression in long term vs short term responses.

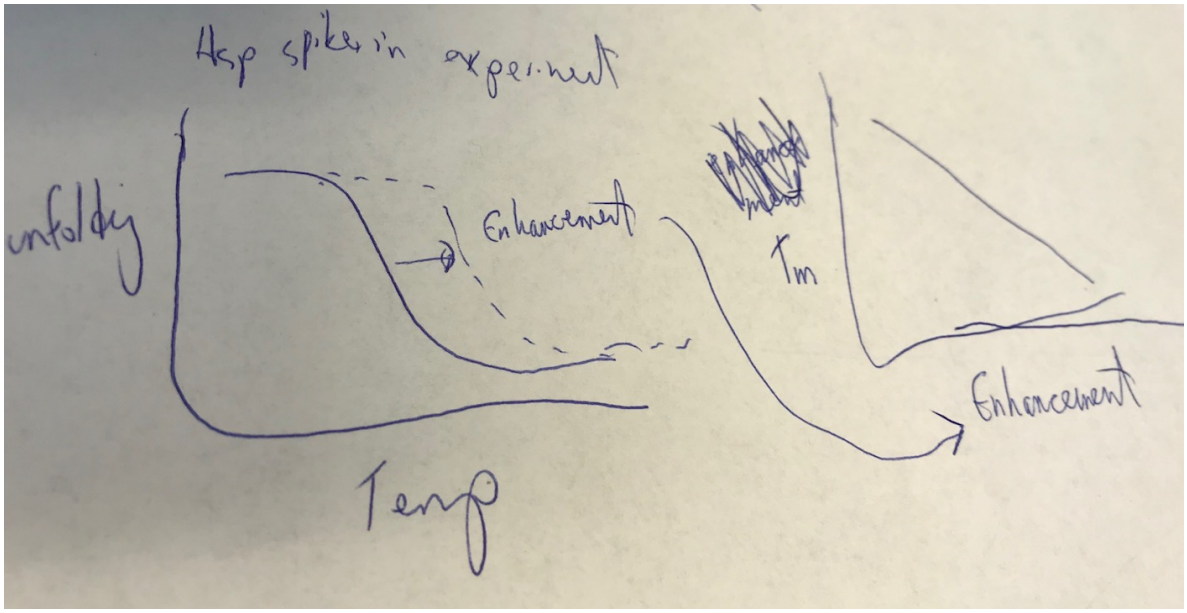


- **Hypothesis:** More thermally labile proteins require molecular chaperoning (hsp mediated or osmolyte mediated)
    - **Approach**- Metabolomics to measure the whole suite of metabolics between short and long term temperature stress. Coupled with this experiment would be a pull down of Hsps90 and 70 to determine the relative effects of Hsp mediated vs osmolyte mediated chaperoning.
    - **Expected outcomes:** Either osmolytes or hsp mediated chaperoning plays a larger role (for which proteins? Is it protein specific?).

What to expect for Hsp pull down experiment:



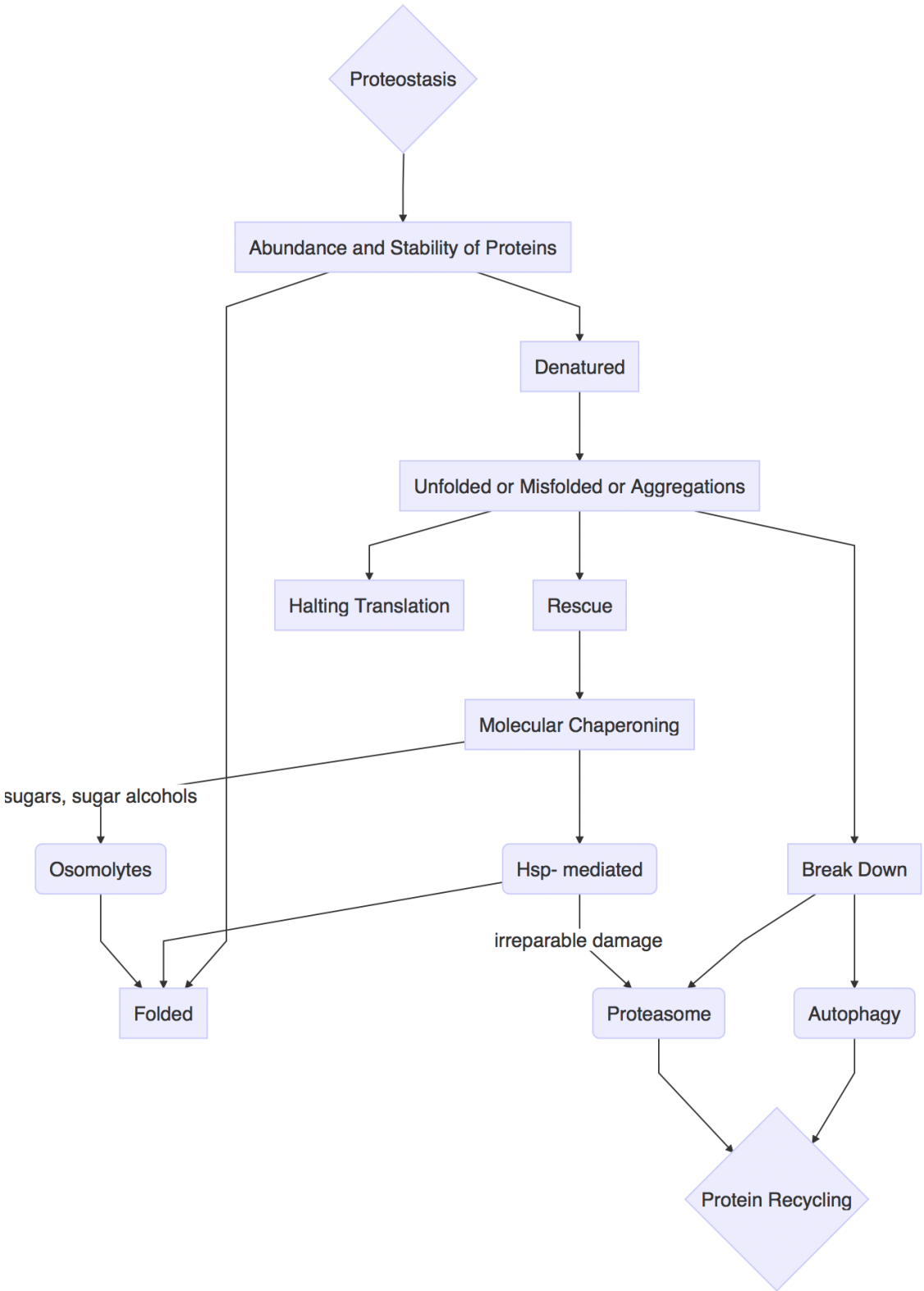
Hsps target labile proteins



Not sure what to expect from metabolics because there could be so many metabolites. Doing multiple linear regressions of Tm on metabolite concentration could be a start....and then test interaction with treatment. This would suggest that there are context dependencies

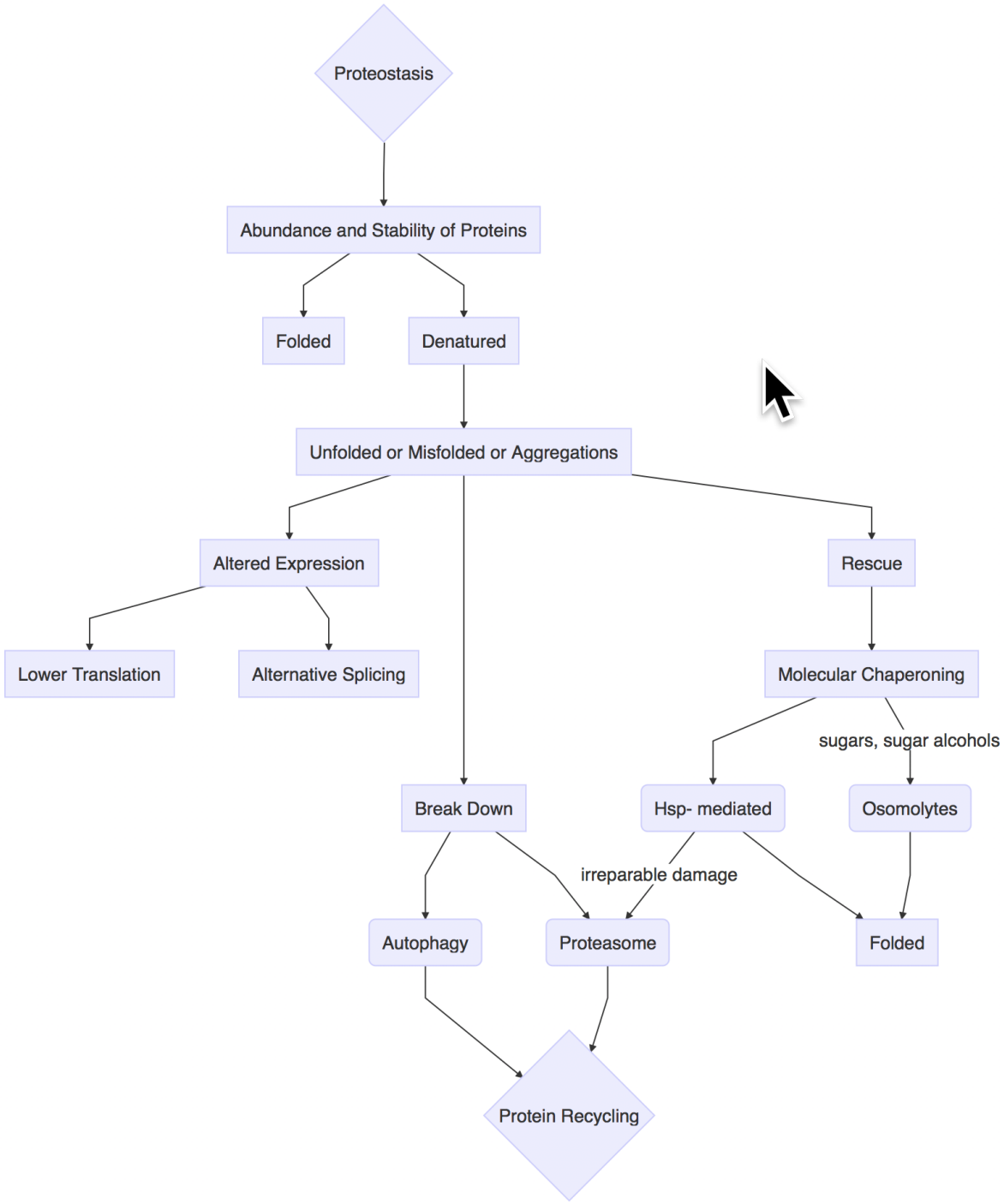
- **Hypothesis:** One way proteomes can become more stable or maintain stability is by expressing more stable splice variants (isoforms).
    - Approach - this would come out organically through multiplex mass spec
    - Expected outcome:

2018-03-12. Tweaking the flow diagram of proteostasis.

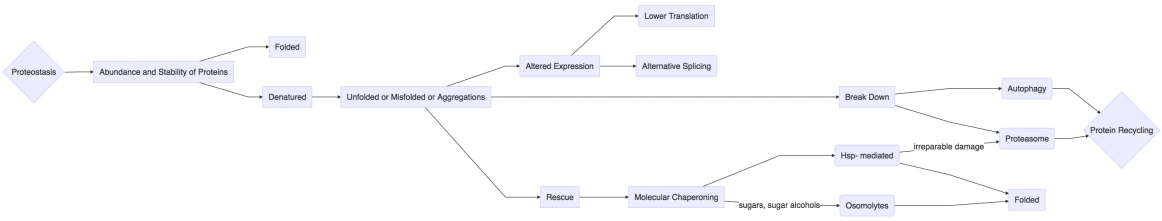


This one is not centered about the meaning of the different ways proteomes can be stable, but more about what are the types of fates or responses there are for stable and unstable proteins.

Ok, tweaked flow diagram some more:



horizontal version



Page 44: 2018-03-16. Flow of ideas for evolution talk

Got picked to be in evolutionary physiology symposium

S69:Organizers: Mathieu Buoro, Jacques Labonne, Matthew MacManes, Sylvie Oddou-Muratorio

Evolutionary physiology, or the study of an animal's or plant's physiological processes within an integrative evolutionary, ecological, or comparative framework has grown in popularity over the past several years. Particularly in the light of adaptation, understanding how organisms respond to their ever-changing biotic or abiotic environments, by altering physiological processes, is foundational yet often an overlooked discipline within evolutionary biology. This symposium aims to bring together a group of researchers interested in physiology across a wide variety of scales (e.g., from individual organisms to communities, including the exploration of feedback mechanisms between the two levels), physiological systems (e.g., neuro, endocrine, pulmonary, metabolic, photosynthetic), and approaches (e.g., modeling, genomic, molecular, population-level, field-based). This symposium will promote discussion and collaboration with symposium participants and the audience members, allowing for the formation of a deep and synergistic understanding of the mechanisms that underlie physiological responses to our changing world. This is particularly timely, as newer methodological approaches, including high-throughput phenotyping is becoming more commonplace, and as evolutionary biology faces the great challenge of producing operational predictions about biodiversity dynamics in a context of rapid environmental and climatic change.

I'll be presenting my hsp rxn norm work. Flow of ideas for talk:

- Varying environments pose a problem for populations and species to persist because living in each of them have different requirements. This is especially true for ectotherms where for example, environmental temperatures influence internal body temperatures and has consequences on how they perform. High temperatuers damage proteins, leading to loss in biological activity and negative fitness consequences.
- Illustrate the thermal performance curves
- How may we expect organisms to respond to high temperatures? And how have they historically adapted?
    - they need to resist or tolerate it
        - under resistance, proteins are more stable
        - under tolerance, protein function may be recovered, or the toxic ones are discarded
- So I'm asking two questions, how has selection operated on upper thermal limits and in what ways can organism resist and/or tolerate temperature stress? What are the adaptive ways species utilize their hsps to maintain proteostasis?
- Ants are a good system to approach these questions because:
    - They're everywhere, so they likely have adaptations that facilitate their diversificiations
    - They're long-lived, so they're accustomed to responding to variable environments within their lifetimes and we'd expect plastic mechanisms to be critical for their performance
    - common woodland ants
- Common woodland ants span a large lat gradient and
- In this system, we asked, what aspects of the local environment shapes their ability to take the heat?
    - we slow ramped heat shocked them , measured their ctmax
- We found that habitat type to be more critical for shaping variation in ctmax
    - Ancestral state reconstruction shows an evolutionary transition into open canopy forests
    - this represents an evolutionary innovation
- We then asked, ok, are ants resisting or tolerating the heat?
- Approach: we estimated tactics of resistance and tolerance by measuring the expression kinetics of critical proteins involved in the stress response. Go over each tactic.

- We found evidence for resistance and tolerance.
- Coordinated, adaptive shifts in the ways forest ants use hsps may have facilitated their divergence and diversification and lays out the expectations for the types of responses needed for a rapidly changing world.

Page 45: 2018-03-22. re-analysis of diapause exit in rhago

Tom sent me a data set where he refitted biphasic functions on a more accurate dataset. The problem with the last one was that it included metabolic rate measurements of adults. So this new dataset excluded those measurements. One additional concern is that there is error in the estimates of the function valued traits. I'll do some multiple linear regressions to test the effect of function valued trait parameters and their interaction with host on eclosion timing.

Uncorrected for estimate error

I'll scale the variables so that the betas are comparable for each continuous predictor.

multiple linear regression

```
uncor3<-data.frame(cbind(apply(uncor[,3:7],2,scale),uncor[,1:2],uncor[,8]))
full.mod3<-lm(Eclosion~host*b+host*c1+host*EX+host*plat+host*term,data=uncor3)
summary(full.mod3)
knitr::kable(summary(stepAIC(full.mod3,direction="both"))$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	65.6028022	0.3869879	169.5215640	0.0000000
hostH	-0.3691653	0.5814205	-0.6349368	0.5280102
c1	-1.0671138	0.2283625	-4.6728942	0.0000186
EX	2.0582639	0.5396524	3.8140551	0.0003383
plat	-0.5065150	0.4281648	-1.1829908	0.2417236
term	21.7178949	0.2734226	79.4297696	0.0000000
hostH:EX	1.0672529	0.6319020	1.6889531	0.0966923
hostH:plat	-1.2750041	0.6721143	-1.8970048	0.0628978

PCA parameters and then multiple linear regression

```
uncor<-dat%>%
  dplyr::select(-one_of(n))
uncor

ucor.pca.param<-princomp(uncor[,3:7])
summary(ucor.pca.param)
knitr::kable(ucor.pca.param$loadings[,1:3])
```

PC loadings

	Comp.1	Comp.2	Comp.3
b	-0.0001210	-0.0003812	0.0154918
plat	-0.0001587	-0.0019363	0.0285725
term	0.9998725	0.0159542	0.0006682
EX	0.0159609	-0.9998101	-0.0110578
c1	-0.0004855	-0.0110116	0.9994103

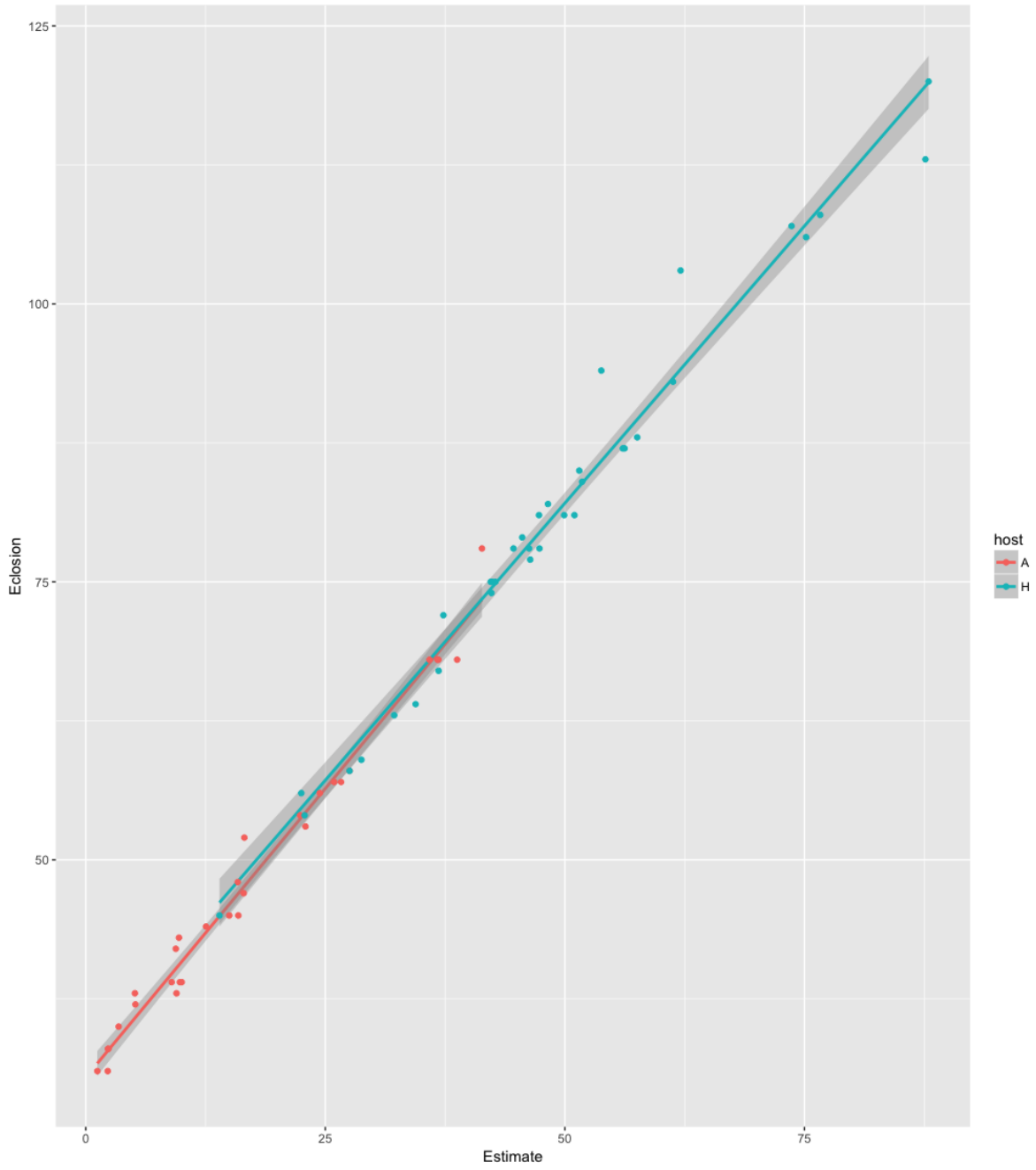
Termination timing dominates pc1, so test for it's independent effect.

```
term.mod1<-lm(Eclosion~host*term,data=cdw.pc)
knitr::kable(summary(term.mod1)$coefficients)
```

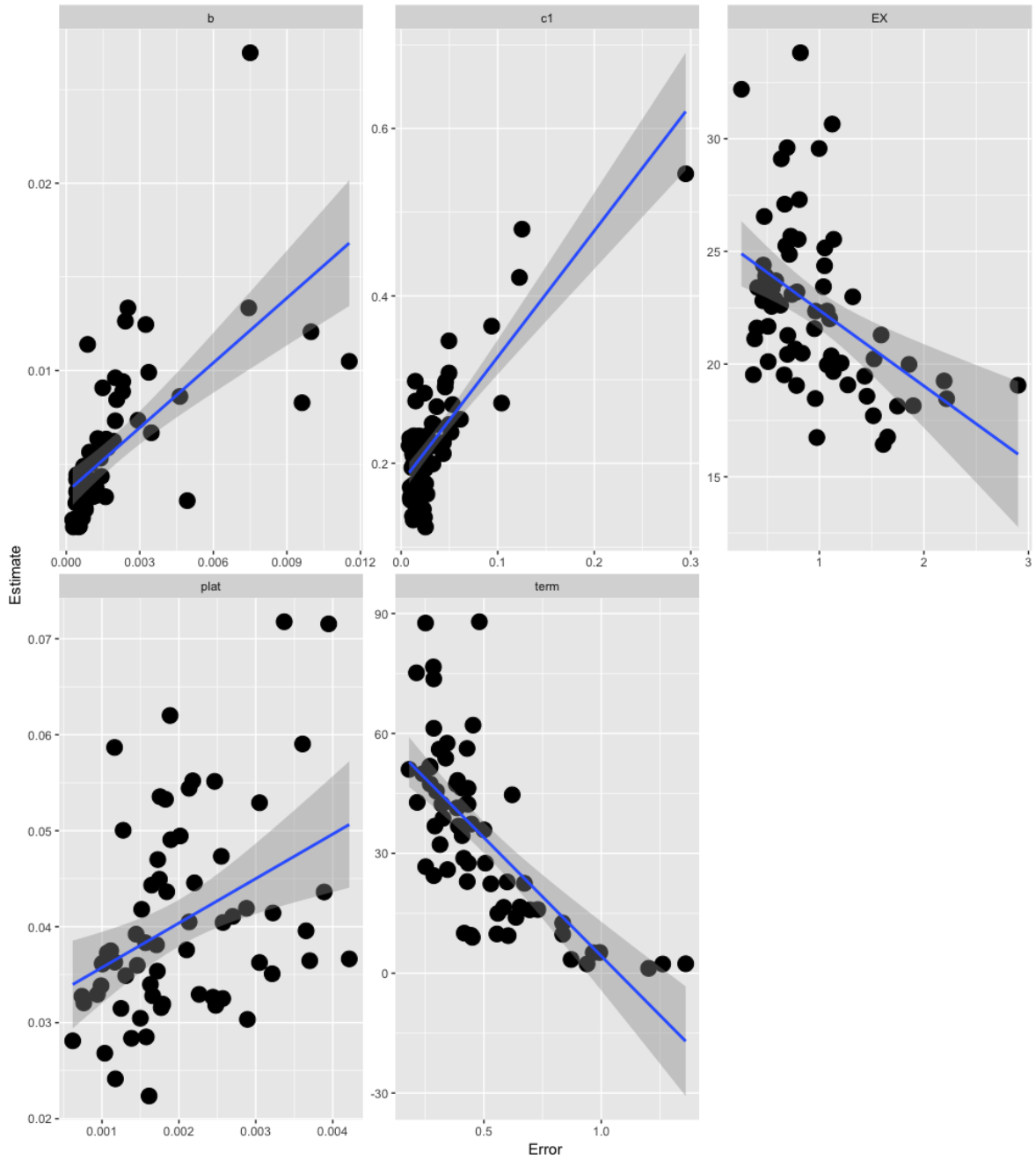
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.7100883	2.329147	20.4839354	0.0000000
hostH	25.4743780	3.031864	8.4022175	0.0000000
term	-0.7873064	2.544022	-0.3094731	0.7580163
hostH:term	17.2524922	3.140965	5.4927366	0.0000008

Associated figure

```
##grab only termination
ucterm<-comb.dat%>%
  filter(.,Param=="term")

str(ucterm)
ggplot(ucterm,aes(x=Estimate,y=Eclosion,colour=host))+stat_smooth(method="lm")+g
eom_point()
```



BUt...there is a relationship between parameter estimates and the error in those estimates. So our estimates may be biased.



some cod showing how to fit individual linear regression models to subsets of the data:

```
dat.mod<-comb.dat%>%
  group_by(Param)%>% # create different subsets , based on param
  do(res=scale(residuals(lm(Estimate~Error,data=.)))) # fit linear model and
scale the residuals

#checking order
#dat.mod
#comb.dat$Param
comb.dat$Estimate.corr<-unlist(dat.mod[[2]])
```

Error corrected

Multiple linear regression

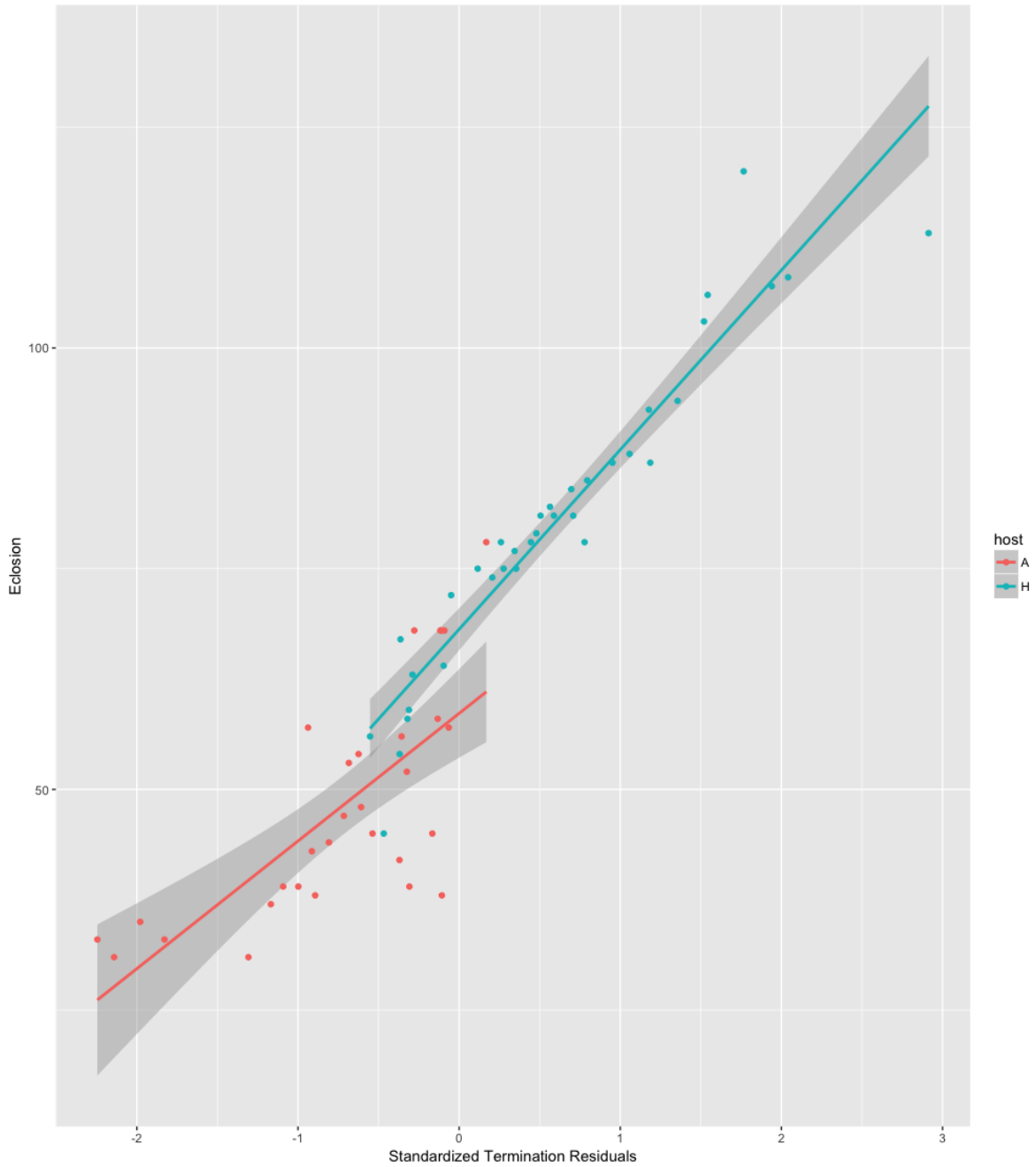
```
mod.sel1<-stepAIC(full.mod,direction="both")
summary(mod.sel1)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.544006	2.180970	21.7994752	0.0000000
hostH	26.359097	2.881963	9.1462285	0.0000000
b	-4.641483	1.583192	-2.9317242	0.0049345
c1	3.857267	2.143146	1.7998158	0.0774772
EX	3.339120	3.218459	1.0374905	0.3041309
plat	2.916959	2.527654	1.1540184	0.2535746
term	-1.750272	2.902379	-0.6030472	0.5490018
hostH:c1	-4.138989	3.108044	-1.3317020	0.1885522
hostH:EX	-6.483143	3.962527	-1.6361135	0.1076319
hostH:plat	6.175939	4.704094	1.3128860	0.1947736
hostH:term	18.416788	3.323757	5.5409553	0.0000009

**Figure showing the interaction between term and host on eclosion

standardized termination first:

```
comb.dat.wide$termres<-
scale(residuals(lm(Eclosion~b+c1+EX+plat,data=comb.dat.wide)))
ggplot(comb.dat.wide,aes(x=termres,y=Eclosion,colour=host))+stat_smooth(method="
lm")+geom_point()+xlab("Standardized Termination Residuals")
```



PCA and then multiple linear regression

PCA loadings:

```
pca.param<-princomp(comb.dat.wide[,6:10])
summary(pca.param)
knitr::kable(round(pca.param$loadings[,1:3],3))
```

	Comp.1	Comp.2	Comp.3
b	-0.535	0.052	0.137
c1	-0.333	0.478	-0.698
EX	-0.393	-0.713	0.043
plat	-0.596	-0.100	-0.115
term	0.305	-0.500	-0.692

regression model with pcs:

```
knitr::kable(summary(full.mod3)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	45.565024	4.520428	10.0798034	0.0000000
hostH	40.266333	6.160305	6.5364190	0.0000000
Comp.1	-6.974462	9.467399	-0.7366819	0.4642351
Comp.2	-2.577335	7.277889	-0.3541321	0.7245024
hostH:Comp.1	19.754392	12.901890	1.5311239	0.1310834
hostH:Comp.2	9.558117	9.918091	0.9637052	0.3391271

No effect of PCs on eclosion.

Take home:

IN the error corrected analysis, we found a term * host interaction, but not in the uncorrected analysis. The PCA doesn't add much, other than the fact that the correlational structure among parameters does not account for any of the variation in eclosion (in the corrected analysis). This might be a good talking point, because these parameters may be evolving independently and aren't constraining one another (covariation isn't functionally relevant).

Page 46: 2018-03-23. Meeting with dan, plan of actions

Met with Dan.

1. Tom's re-analysis: write up results and methods, set figures
    - Settle on figures and send to Tom/Dan
2. Proteostasis project:
    - Choose which lines? DSPR? DGRP? Look into papers about these lines and the pros and cons of using them
    - For the lab meeting, just specify 1 project (prob proteome stability one) and send out paper.
        - Sent! April 2nd, Monday.
3. Range limits paragraph- GOOD to go!
    - Sent out new version of ms to co-authors!
    - Nick to give me feed back next week

Page 47: 2018-04-02. Redundancy analysis

Thinking about how to analyze datasets where there are not only multiple predictor variables(X's), but also to include multiple response variables (Y's). Redundancy analysis combines PCAs with multiple linear regressions with the goal of displaying and explaining variation in set of response varaibles CONSTRAINED by second set of predictor variables.

ex: What variables best explain variation in fungal communities sampled along a gradient of fungicide toxicity?

A lot of outputs are graphical, there a stats table?

Page 48: 2018-04-17. notes; rhagoletis brain transcriptome

Greg's group has analyzed data with edgeR and different clustering techniques to try to understand how host races differ in their transcriptional responses. They can have a common one which relates to just diapause. However, hosts can diverge in their transcriptional response and I'm still trying to understand what types of interactions are meaningful(or what they mean) in the context of race differences in diapause. They can start out diff early but then converge, or they can start out the same and differ in slopes.

Perhaps we need to think about this in the context of diapause modules: maintanence. If metabolic rate is characterized by suppression. Overall, are there more lower expressed genes? Think about other possibilities.

Meeting with Greg:

- PI's want me to work on cerasi data
- I need to be more familiar with the system - comparisons of low and high land
- Make the anlayses comparable with pomonella data. Are tehre similar modules?
- Try different clustering algorthms for diff expressed genes
- Use edgeR to test the effect of time x host interaction
- Greg to send me dataset and scripts; I'll make a private github repo

---

Page 49: 2018-04-18. update and biological rhythms notes

status updates on projects:

1. thermal niche ms- double checking data and then need to re-analyze
    - shc thinks ctmin values might shift if there are mis-IDs
2. hsp rxn norm ms- in shc's and njg's hands
    - they need to look over the results
    - i still need to finish the discussion, it is written, but needs more citations, and need to revisit
3. cerasi data analysis
    - Greg to get me dataset in 1 week
    - create a github repo with research plan and hypotheses, expected outcomes
4. biological rhythms
    - presenting a poster and datablitz 1 slide 1min talk at biological rhythms conference may 12-16. poster session is may 13 to 16.
    - data blitz talk is 8pm
    - posters are always up but I need to stand by mine on sunday
    - one thing i can test in my dataset is viability selection on MR. So does MR influence survival to adulthood?
5. evolution meeting - giving a talk and need to buy plane tickets, already registered

notes on biological rhythms. We were low on trikinetics monitor space, so we had to move monitor #6 around between entrainment and free-run. This is the circ repo script, but also adding in my nb. The below also refers to physical notebook, not electronic.

2018-04-18 additional notes:

We were running low on trikinetics space, so we had to switch some monitors around. First, #1 and 2 are in entrainment, while 3,4,5,6 are in free-run conditions. We took out #6 from free run to entrainment on 2018-03-23,20:00 or 8:00PM (nb#002,page 178, parse on 227487) and then switched it back on 2018-04-05(parse on count number 241326; nb#003,page14-15).

To deal with monitor #6 having a combination of free-run and entrainment data counts, we will need to subset out the data from when #6 was in entrainment. We can take this subset and call it a diff monitor altogether ("6.2") and then assign different experimental classes for #6 and #6.2.

Another complication: There are samples in the move from entrainment to free-run in monitor #6 that needed to be moved to monitors #1 and #2. So this means some flies have had entrainment that were in 2 monitors. See notebook #003, page 14-15. Here is the list:

uniqueID	trik_monitor_6	trik_monitor_6_position	new_trik_monitor	new_trik_position	time	date
2b15	6	1	2	1	20:30	2018-04-18
19r23	6	2	2	7	20:30	2018-04-18
4o72 or 12	6	4	2	9	20:30	2018-04-18
18w11	6	8	1	6	20:30	2018-04-18
13r12	6	9	1	9	20:30	2018-04-18
10w6	6	11	1	20	20:30	2018-04-18
11o16	6	14	1	21	20:30	2018-04-18
12o67	6	16	1	23	20:30	2018-04-18
11o40	6	22	1	28	20:30	2018-04-18
11w11	6	25	1	29	20:30	2018-04-18
18w22	6	27	1	30	20:30	2018-04-18
12w53	6	28	1	32	20:30	2018-04-18

Page 50: 2018-04-24. Reading *R. cerasi* papers

I'm getting more familiar with the system.

I'm encountering this word, Stenophagous, which refers to an animal that feeds on a limited variety of foods. Terrible word.

Decomposing the abstract-

No clear goals or objectives

Approach: Measure diapause incidence at 5 constant temperatures for 2 populations.

Result: short winters or warmer temperatures influence dormancy (the abstract doesn't have a direction of effect)

"Interestingly, extended chilling (longer than required for terminating diapause) ‘return’ pupae to another (facultative) cycle of dor- mancy enabling adults to emerge during the next appropriate ‘window of time’; a strategy first time reported for univoltine insects."

What does this even mean? ....

Looking through figures first.

Why are they showing climate for each site as a table. How about a graph?

Figure 1: They plot proportion of pupae in stacked bar pots for each chilling period or 3 conditions: adults, overlaying(surviving pupae), and dead. They do this for each site and constant temperature. It is hard to see if there is a site by chilling period by class interaction. It looks like pupae are more alive at higher temperatures for all sites and under shorter chilling periods.

Table 3 says there is a: population by temperature interaction, population by chilling period, temperature by chilling period interaction.

Figure 2: Warmer population has less overlaying pupae at shorter chilling periods (months at 5 C) than cooler populations

Going over intro:

Their first hypothesis was that diapause intensity varies... They predict pupae from warm regions need short exposure to cold to terminate diapause than pupae from colder regions.

3rd hypothesis- pupae cannot meet their chilling requirements for diapause termation under warmer winter temperatures and/or short cold exposures. This will cause prolonged dormancy.

Plain english: Populations that have evolved in warm or cold environments have different requirements for maintaining and terminating diapause. If conditions are unfavorable, these flies will continue to remain in dormancy. It probably takes less time in winter(cold temperatures) to maintain and terminate diapause in warm populations than cooler ones.

I wished they graphed the results better to see the population * temperature effects and other interactions.

ref:

Diapause termination of Rhagoletis cerasi pupae is regulated by local adaptation and phenotypic plasticity: escape in time through bet-hedging strategies

ref:

Physiological and biological patterns of a highland and a coastal population of the European cherry fruit fly during diapause

---

Page 51: 2018-04-25. Biological rhythms thoughts

One way to explore the data in the biological rhythms project is to test whether there is greater viability selection for flies that have high metabolic rate.

Ok, back up. The main question is this: What maintains the variation in metabolic rates prior to entering dormancy? We see a lot more variation in metabolic rates for early eclosers compared to late eclosers. Why? Is this a bet hedging strategy where ones with higher metabolic rates can be more readily active if future conditions are favorable? Is there a cost to having high metabolic rates? Feder suggests that early eclosers with high metabolic rates are "doomed", meaning they'll die when they encounter winter conditions.

Lets lay out the hypotheses:

1. Flies with high metabolic rates face greater mortality under simulated overwintering

We can compare this to flies that aren't forced to overwinter.

Page 52: 2018-05-01. project update

1. Biological rhythms proj
    - giving a poster May 7 for biodiv conf at UF and biological rhythms May12
    - giving a 1 min slide presentation data blitz at biological rhythms conference
        - need to upload the talk
        - some prelim shows eclosion diffs between flies and parasitoids and this pattern folows the dominant biological rhythms under free run but not entrainment
2. cerasi proj
    - Greg to send me dataset
3. diapause exit paper (tom)
    - tom sent me data and I need to redo figures
    - once I do that, I can clean up the results, methods, and work on discussions
4. hsp rxn norm paper
    - NJG and SHC hands
    - I should revisit it, especially the discussion
5. range limits ms
    - In review!!
6. thermal niche paper (lchick)
    - double checked data, results changed a little; main effect of local env temp, but no species effect in field ctmax and ctmin
        - however, the data are a bit non-linear for ctmin in the field, so we tried a breakpoint analysis and both species have breakpoints.
        - intepretation: "so at the cold end of the range, selection is acting more strong for rudis than picea. Picea may already be at it's cold limits at the warm edge of the range, we see the opposite, where selection is operating to a larger degree in picea than rudis"
7. lat var paper with chao
    - have to finish results section, analysis done

Page 53: 2018-05-18. status update

Projects

1. **Biological rhythms proj**
    - On-going, measuring eclosions and activity
    - need to think about collections for later this year
    - It looks like wasps emerge earlier under constant rearing conditions (cue for favorable environment) than flies and this is reversed under simulated overwintering. This suggests that the wasps are using environmental cues to shift their biological timing. **Is it possible that the wasps are multivoltine?**

2. **cerasi proj**
    - Greg to send me dataset
        - need to learn WGCNA, which script should I focus on ?
        - Should set up meeting with greg

In edgeR, I'm assuming they're testing for the interaction between time and population. What would the stats mean?

- Common responses : main effect of time
- Overall population differences: main effect of population
- Time by population interaction: populations time their gene expression differently

You can then do functional enrichment in the 3 of these categories.

3. **diapause exit paper (tom)**
    - Get on same page for analysis first,
        - then discuss figures and make them
        - need feedback on intro
4. hsp rxn norm paper
    - NJG and SHC hands
    - I should revisit it, especially the discussion
5. **range limits ms**
    - In review from 2018-04-03 to 2018-05-15
        - with AE & ED: decission process since 2018-05-15
6. **thermal niche paper (lchick)**
    - Lchick sent me edits 2018-05-16; need to look it over and then send it back; done (2018-05-18)
7. lat var paper with chao
    - have to finish results section, analysis done

Other items

- Need to set up project for Hchu and set up hours for undergrads.
- still need to buy tickets for evolution 2018

Meeting with Dan today:

1. Skype with tom 2:30PM
2. talk about Forbes et al. 2009 ms
3. Discuss project for Hchu

Meeting notes:

meeting with tom cancelled.

Sanderelli et al. 2007; bombos ; kinetics of timeless.

Talk to dopman. Are the lines pure breeding? If not, make them of period alleles.

angle for proteomics projects. Identification of temperature sensitive lethal alleles. Temperature sensitive alleles (david suzuki). May be important for non-models. Sterile technique strain optimization-- rescue with temperature sensitive alleles.

Dan action items:

1. Send dopman email about ECB period
2. Touch base with Al about proteomics projects
3.

Page 54: 2018-05-18. ECHO app, by Hannah de Los santos ; finding circadian rhythms with extended harmonic oscillators

youtube link: https://www.youtube.com/watch?v=TqQsUF7Yfkg

Taking notes:

Intro:

What are circadian rhythms - 24 hour endogenous cycles reinforced by external cues such as flights

Health problems associated with disrupting circadian cycles

People measure gene expression adn we observe circaidna rhythms in plants and animals

Data:

Gene expression for 24 time points- starting 2 hours, ending 48 hours, 2 hours resolution, 9624 genes, 3 replicates (6 missing time points)

they follow sinusoidal curves. But the models assume fixed amplitudes

Methods: Extended harmonic oscillators

the important part is gamma, which is the forcing coefficient. It is estimated with non-linear least squares. Minimize residuals. Use nlsLM algorithm.

For replicates, they add weights, where are inverse proportioin of the variance at each time point.

This is mainly for gxp. But she started with multiple cycles that dampen in amplitude to a dataset with only 2 cycles...what.

Page 55: 2018-05-21 Meeting with Tom

Diapause exit ms

Tom's concern is what do corrected residuals mean and their bearing on our biological hypotheses.

In the uncorrected data, scaling, exit, termination are significant, suggesting a common mechanism for apple and haw.

For corrected data, baseline and termination x host interaction are significant. So baseline is the common mechanism, but they diverge in their termination.

Need to consult Greg for the stats.

For the out of winter experiment, I will redo the analysis. The data has non-eclosers and I'll need to parse that out

Tom wants to keep experiment 1 (whole trajectory) and 2 (just baseline) in that order.

Dan- Action Items

1. Tom send me dataset with landmarks and I can re-analyze?
2. analyze experiment 2 and send tom an email
3. send Greg a report of our approaches
    - but coordinate with Tom first
4. Tom will send us comments on intro by end of weekly
5. Dan to send out Rhagoletis meeting for the whole family.

To do list to organize myself over the short term (this week) (not in any particular order in terms of priority)

1. Finish paying for evolution meeting travel. Do the paperwork
    - flight paid
    - need to pay for housing (going for dorms)
    - need to pay for train from orlando to miami
2. Understand edgeR, WGCNA and think about data analysis for cerasi dataset
3. Re-analyze Tom's dataset, use mixed effects model or repeated measures anova in R. Write up a small technical report and send to Tom.

Need to get these items done this week.

Page 56: 2018-05-22. Circadian rhythm talk by Katja Lamia, scripts research

She hosted a session at the biological rhythms conference. Her video link: https://www.youtube.com/watch?v=G7xvy7gPhoE&t=103s

Title: Saturday Science at Scripps Research: Biological Rhythms:From Sleep to Cancer and Metabolic Disease

She defines circadian rhythms as an internal clock in the body that keeps track of the time of day. Biological rhythms are under the broad framework of homeostasis, which was first talked about by Claude Bernard (1813-1878)

"The constancy of the internal environment is the condition for a free and indepedent life".

Maintaining a constant internal environment is beneficial for health. Changes in environment present a challenge to homeostasis!

Daily rhythms in light, temperature, moisture. They are a challenge, but they're also predictable. Organisms can optimize physiology to antipate changes.

First studies in circadian rhythms were first in plants. Jean Jacque D'ortous de mairan describes rhythmic movements of plant leaves that follow the movement of the sun.

But when plants were placed in constant darkness, the leaves still moved, so it suggests an internal mechanism, not solely driven by ligiht. Leaves also move before the sun comes up, so this is a signature of anticipation, driven by an internal clock. Darwin wrote a book about the movements of plants...wow.

Internal time keeping mechanism.

Woody Hastings (source of inspiration)- Took his course and studied a marine organism that produces bioluminesence at night. He showed that it was govern by a circadian rhythm. It is camo beacuse it makes the org look like moonlight to trick predators.

California Grunion spawns on the beach only at the times of high tide --- circalunar rhythm. Kristin Tessmar-Raible still trying to understand this mechanism (she was at the biological rhythms conference too)

She studies mammals and mice in particular. In mice, the circadian rhthm that is good to measure are running wheels. They're nocturnal and run in running wheels. And they can log counts.

1967 coined circa, circadian .

What is the location of time keeper in the body? In brain. They did lesion studies and disrupted different parts of the brain. 1 position above optic chiasm ---SCN, suprachiasmatic nucleus. And as a result, they disrupted behavioral rhythms. Animals have clocks and where they are located.

Behavior is genetically determined?

1971: Konopka and Seymore Benzer identified mutant flies that altered circadian rhythms. They foudn 3 mutants: 1 with no rhythms, 1 with short and 1 with long rhythms. Genes can influence behavior!

This work led to 1985 papers describing a single gene associated with diff in biological rhythms. Jeff Hall, Michael Rosbash, and Michael Young. Period gene.

Joe Takahashi found that you can disrupt clock and it influenced circadian rhythms in mice.

Negative feedback loop that underlies circadian rhythm, which produces occilations. You need a positive signal, then a 180 degree phase shift with a negative signal feedback loop.

The positive signal: Transcription factors The negative signal: Transcriptional repressors

Clock/Bmal increase transcription of per1/2/3 + cry1/2. Per/CRY repress Clock/Bmal.

The positive signal = Clock/biochemical The negative signal = Per/Cry

These clock sare everywhere - proboscus, antenna, leg, wing. IN mice, SCN, lung, liver.

Timing of when you eat matters. A shorter window = don't gain as much weight on a high fat diet.

Page 57: 2018-05-25. Meeting with Gragland

Agenda:

1. Discuss expectations and roles for cerasi data
    - Data analysis = middle author
    - data analysis + writing = lead author

Gragland thinks this is reasonable.

2. Data analysis

For pomonella, he wants to integrate pool-seq, RNA-seq, and brain development datasets. For cerasi, he wants gxp and functional enrichment analysis, but also integrated with the pomonella dataset. He needs to send me a global edgeR output.

GRagland suggestion for parsing data, look at pop x time interaction, but also just differences between low and highland for a given timepoint. He is finding a weird effect where sets of genes that come up early are the ones represented in highland. Counter intuitive, but could be termination supression genes (DHahn says).

Send Gragland github invite to collaborate on project.

GRagland's thoughts on correcting for error in the estimates of a parameters that come out of a NLS. A bayesian model could account for that error. (From what I'm perusing, it is similar to a metaanalysis). His stance is that correcting for error is probably the better way to go.

Page 58: 2018-05-29. Set of tasks for HChu

List of tasks for Hchu

1. European cornborer rearing (trained by Qinwen) and writing up protocols (through Rmarkdown)
2. Contribute to the Rhagoletis project:
    - Checking eclosions, deaths
    - Transferring entrainment samples to free run
    - documenting this in master spreadsheet and in notebook
3. Maintaining personal notebook
    - to do lists
    - reading papers
    - experimental notes
    - seminar notes
    - etc

We still need to discuss projects with Dan. Potential projects.

1. ECB: THe timing of protein protein interactions underlying seasonal changes in diapause
2. Proteome stability project: What are the timing of events for proteostasis
3. Rhagoletis data mining: What are the consequences of diapause depth for overwinter survival?

Page 59: 2018-06-04. meeting wth Gragland, cerasi data filtering and future analyses

Greg says the best evidence is the interaction between time and populations.

Problem with overall effect of population is that it may not be related to diapause at all.

Greg finds more common responses than pop * time interaction in pom. It may be due to the fact that pom are less divergent than cerasi .

Cerasi: Came from de novo assembly. Map reads to isoform to a gene model? Or the number is so big bc it includes a lot of diff isoforms. They amy be mapped to gene models. (longest transcript identified in de novo assembly). Could be redundancy, to get rid of it, you can annotate.

Down the line, I'll need to account for redudnancy with annotations.

WGCNA notes: authors want to use scale free. Greg has not used soft threshold. He's used 1, which is no transformation.

1. Look at lit, debate networks are scale free or not
2. it is harder to interpret if you're not dealing with raw correaltions. They lose meaning, what do the shapes look like?

Greg argues not to use soft thresholding in the past. Make sure I pick a good reason to use soft thresholding.

MOdules:

Enrichment: You need to compare it to something.

DAVID has flybase annotations GSEA might have uniprot annotations (double check)- presumes you have replication, but it might not work that well if you're using log2 fold changes. In order ot use replications, could use FPKM (flawed metric; worthwhile? vs log fold changes, more stat robust).

Greg cuts and paste flybase IDs directly and compared to the presumed background of the genome. An alternative is to use whole gene set, but Greg doesn't think they are much diff than background drosophila genome.

Workflow:

1. filtering ok
2. WGCNA analysis for time and then time * pop effects.
3. Enrichment analysis and compare them.
4. **Think about doing network statistics on any of these things. Certain genes to focus in on; hubness and connectness.** igraph package. Greg can send me some scripts on this. Make sure I can compare among modules for example. Early divergence genes may have more hubness than later diverging genes.

Where does the divergence occurs? Look at time points wise, how many transcripts are diff expressed at each time point. Pom = big at 2 months and small at 3 and 4 months. Good summary of where the variation lies. compare numbers to Eddy's output, they should b ethe same.

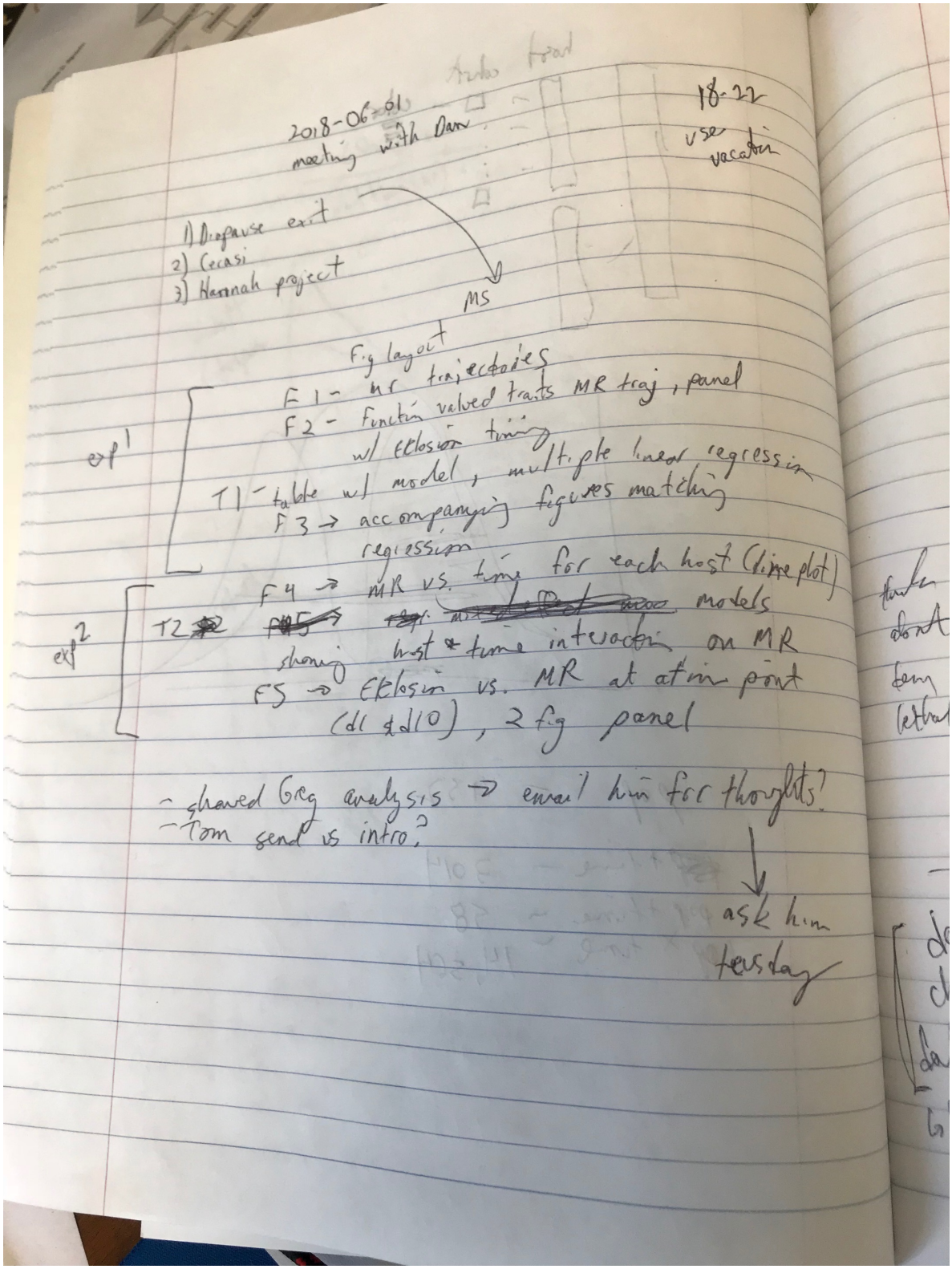
Page 60: 2018-06-04. Things Hchu is woring on:

1. Biological rhythms in ECB
    - learn rearing conditions
    - put different strains into trikinetics
2. mini coding proj: viability selection on MR
    - need to include pupae that dont make it to adults in adult lifespan column (=0)
    - need to find out whether our sampling of death is similar to when they die (filter out deaths in trikinetics dataset)
    - try survival analysis on top of negative binomial glm
3. proteome stability projs
    - phenotype , CTmax, KO time.
    - pick divergent strains, and start unfolding experiemnts

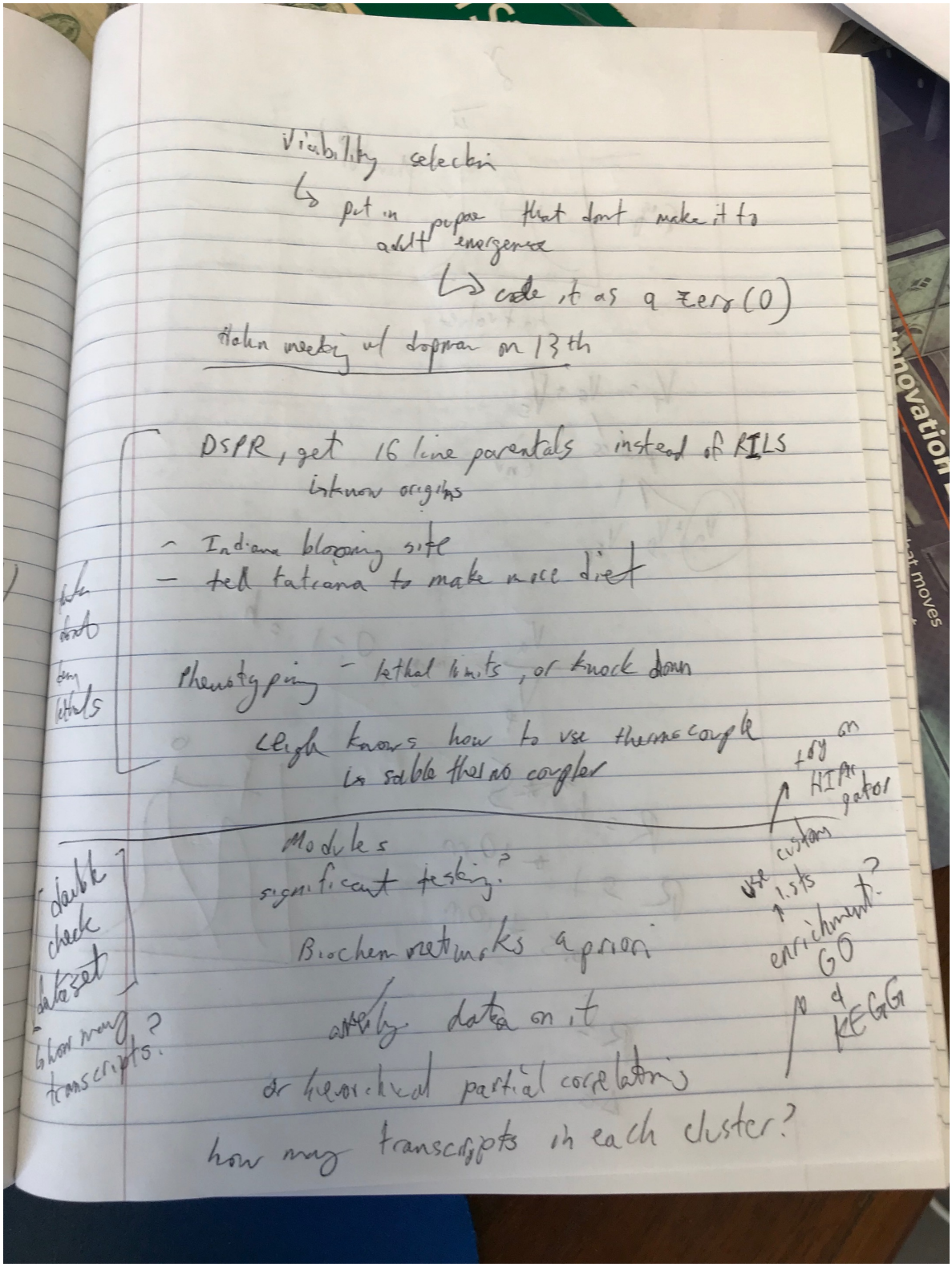
---

Page 61: 2018-06-01. meeting with Dan

Outline of the layout for diapause exit ms



Viability selection and DSPR for proteome stability project.



Page 62: 2018-06-05. kick starting proteome stability project in Hahn lab

Working on DSPR lines:

- 14 founder lines to start - need to phenotype metrics of thermal limits
    - CTmax under slow ramp
    - CTmax under fast ramp
    - KO under static HS
    - KO under hardening , static HS
- need to learn water bath. IT is a Lauda circulating waterbath.

    - automate ramping protocol
- let DSPR lines acclimate 2 weeks.
    - How should I sample adults? They should be all equal ages. 5 days?
- Additonal things to sample:
    - Eggs under HS and control. Measure proteome stability and adult survival
    - Larvae under HS and control. Measure proteome stability and adult survival
- We need to set up the ultracentrifuge. Do we have a small rotor that can hold 0.5mL sample tubes?

Page 63: 2018-06-07. thoughts on network analyses on cerasi dataset

So far, I've been constructing weighted co-expression networks with genes that have a population x time interaction. But the network created has both populations in them.

However, it may be more appropriate to construct population specific networks because it doesn't necessarily matter if genes in a population are correlated with genes in a another population.

So here is my thought process:

1. Construct separate weighted co-expression networks for high and low altitude populations with genes that display time x pop interaction
2. WGCNA will give a color code for each module for each population.
3. Determine trajectories of the module and compare whether the same genes are in different modules or the same one. Because of the interaction, we'd expect different modules to pop up in the same genes between populations.

Another approach if we want to understand position and influence in the network.

1. Construct a population specific weighted co expression network
2. Estimate different metrics of centrality for every node/gene
3. Determine whether genes that are significant(pop x time interaction ; time effect) are hub like genes are peripheral. We'd expect early divergent genes to be more hub like and then later expressing genes to be more peripheral in the co expression network.

Page 64: 2018-06-08. Notes on network analyses

Tutorial from Duke University, SSRI (Social Science Research Institute)

Fundamental concepts: **Networks are represnted as a graph-- which has a set of nodes that are connected through lines**

Key Elements

- Nodes, dots, vertices
- lines
- layout

A netowrk is a graph! This graph can be represented with categorical data (names represnted in the nodes) and continuous data (lines can be weighted)

Networks can differ in their layout: unimodal or multimodal (No clue what the difference is, the ppt doesn't specify)

Another property: **EDGES**

It provides the direction and work information.

- Dichtomous (0 or 1)
- Weighted (between 0 and 1)

These edges can be undirected or directed with an arrow and also can loop. Fundamentally, these types of ties represent different types of relations.

Layout! (my interpretation from the slide)

Hierarchical- goes in a direction such as top to bottom or side to side. The point is, that as you move across the network, the following nodes are nested within previous nodes.

Spring or Energy - nodes connected with no particular direction

Data structure

- Adjacency matrix

For example, if 4 connects to 1,2,3 but 1,2,3 don't connect with each other, the data is a mtrix of 0 and 1's indicating the connections

	1	**2**	**3**	**4**
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	1	1	1	0

- Edgelist, data with same structure as above:

Sender	**Target**	**Weight**
1	4	1
2	4	1
3	4	1

Ok, now that we know the properties of a network, what kind of things can we extract from it?

Two general approaches that can overlap

- Networks as variables - identify an individual's network position and other characteristics
    - These variables from the network can then be predictors in a statistical analysis.
- Networks as structures- performing network analysis to identify system level properties that influence of the network will chagne and function over time

No clue what this means, but you can analyze networks based on the types of **connections** and their **position**.

Types of questions for networks as variables:

- Are kids with smoking peers more likely to smoke themselves?
- Do positions in a social network influence the mental health of people occupying them?
- Do central actors control resources?

Types of questions for networks as structures:

- What generates hierarchy in social relations?
- What network patterns spread disease more quickly?
- How do roles such as being the "go-to-person" or "broker" in the network evolve out of consistent relational activity? (Um...what?)

Structural Features of networks:

- Density: proportion of connected ties over all possible connections. Dense networks are highly connected where information/diseases or resources move quickly across it
- average path length: average number of steps along the shortest paths for all possible pairs of network nodes.
    - How efficient information moves through network
- transitivity: extent to which nodes make up a dyad are also connected to a third node will both connect to that third node
    - can indicate mutualisms, shared costs, common enemy or threat, social sanction
- clustering: coefficient measure of degree which nodes in graph tend to cluster together
    - can indicate shared interests
    - can be calculated as proportion of closed triads over the total number of triads both closed and open.
    - isolated clusters are considered bottlenecks--can impeded disease spread to the whole network even if the cluster is infected
- degree distribution : refers to number of connections a node has to other nodes; prob distribution describing the likelihood of a given degree.

Positional features of a network: node level properties associated with occupying a particular place or filling a role in a network

- power and centrality (exhaustive list , see Wasserman and Faust 1994; Borgatti and Martin 2006)
    - degree -number of ties it has
        - in-degree centrality - number of ties node receives (needs directed network)
        - out-degree centrality - number of ties node sends (needs directed network)
    - closeness centrality - average length of shortest path between the node and all other nodes
        - The most central node is the closer it is to all other nodes
    - betweenness centrality - number of times a node acts as a bridge along the shortest path between two other nodes
        - calculated as, computing shortest paths between 2 nodes, for each pair of nodes, determine fraction of shortest paths through the node, sum this fraction over all pairs of nodes
    - Eigenvector centrality - a node's influence in terms of how connected the node is to other highly connected nodes (What..)

Page 65: 2018-06-08. Meeting with TPowell

We're pretty much in agreement with the intro of ms. Take out results section of the end of the intro. Tom doesn't like results there.

Agenda:

1. Analysis experiment 2
2. If the analysis is ok, lay out all of the figures we will want, which should not be substantially different from existing ones.
3. If we're in agreement, forge ahead and set a date to complete the results(figures + text)
4. If not, I'll perform any additional analyses and we can set a date for that.

Tom suggests to have a figure in the supplement of just the overwintering tiime point (0 time point). (Suppelmentary 4). Have a boxplot between host races there. Use Tom's, he has it.

Potential issue: Ones that break diapause over the later time points are outliers (3 flies). Day 10 interaciton pattern driven by those 3 flies, then it could be a potential problem.

A couple things:

- Between host race comparisons for each parameter, move to the supplement.
- Eclosion differences (histogram or proportion plot); histograms may be more intuitive. Cumulative plots are useful for the stats. Tom to come up with something he likes.
- agree with 2 tables in the main text,
- supplement tables - s1 = all of the actual model fitting from the function valued trait analyses (3 diff versions of the trajectory)
- S2 table = actual final model tables for everything (whole bunch of sub tables)
- S3 table = results of the pairwise t test between biphasic and triphasic model. Treat them as hierarchical models and not very different
- Results of the partial correatlion in the main experiment. Do the same with Day1 and Day2 in experiment 2. Have it in there to back up the fact with the diff error distribution in the multiple regression models. Qualitatively the same result with diff analysis.
- Tom will spend time to remake - individual mr trajectories with predictions and point data.
- I can make this figure in ggplot2

Altogether: 5 supp tables

Tom to send Greg an email.

Streamline discussion. Refocus it to modularity. Overall methods and results have superficial changes. Send to Greg, then Tom, Dan, and I will go over how to approach the discussion.

Deadline: Next Friday. Tom will work on supp stuff, I'll work on figures and make them look prettier.

Page 66: 2018-06-12. to do list for Hchu and ANBE

There has been a lot going on in the lab and we want to flesh out future plans with a to do list:

1. ECB project: we have an experimental approach outlined (SURF nb#001, page 32). Basically, we want to set up 5 cohorts for each strain (UZ and BE). For each cohort, measure larval post diapause termination(transition from larvae to pupae) under favorable and unfavorable conditions. And then, take the adults and measure their activity in the trikinetics

set up. In the trikinetics set up, have sugar water on 1 side and water on the other. (Hchu can you double check this?)

- learn rearing protocol and add upon it in the existing protocol (check readme for formatting and structure)
- create github repository and populate with documents, data, results
  - document experimental workflow in readme (also add a little background info on yourself and the system)
  - log the data in an excel spreadsheet + notebook
    - individual, strain, cohort, cohort date, pupation date, eclosion date, death date, trik enter date, trik exit date, trik position, trik monitor, (have free run version too) , sex

2. DSPR proteome stability project: Flies arrived today. WE obtained 13 strains (SURF nb#001, page 34). Hchu did test static HC at 39C.
   - let flies sit for 2 weeks. Coordinate a rearing schedule
   - document activities in notebook
   - Try static HS at 37 C
   - Try other heating protocols and summarise the data in a R markdown doc. Can try cold tolerance experiments but Hchu will need to change water bath.
   - Write up protocol in hahn lab protocol website github repo. Include info on vendor, pictures, and links for all materials.
   - when starting experiment; take adults from 5-7 days (they are fully mated)
   - check the line that may be missing (http://wfitch.bio.uci.edu/~dspr/index.html)
3. Viability selection coding proj:
   - Fixing pdf output to make it more presentable; send to ANBE by Wednesday
   - Create a direction of effect table to make sense of the data.
   - Show Dan Friday.
4. Calibration water bath
   - try calibration from 35-40C range.
   - write up technical report in Rmarkdown
   - clean water bath, it is cloudy. Dan suggests to put some ethanol in it. (also leave it open so it doesn't get as cloudy)
   - make a new rack, have tubes that can be inserted into the styrofoam for better handling

Notes:

- Date and title every page in SURF notebook

Page 67: 2018-06-26. Updated to do list

1. hsp rxn norm: in SHC and NJGs' hands
   - I should flesh out discussion and resend it to them (after range limits)
2. range limits: major revisions, address reviewer comments
3. biol rhythms: passively collecting data
   - need to check free- run to entrainment samples
4. thermal niche: major revisions from SHC's comments; start with abstracts

5. cerasi proj: need to do functional enrichment analysis and network Analyses
    - network construction fro 14,000 transcripts is overloading memory, need to run script on cluster
6. diapause exit ms: fixed changes suggested by DHahn, need all co-authors to agree on them; then decide next steps; need to set up meeting for this week.

Page 68: 2018-06-27. Range limits writing Notes

Putting older discussion here for reference:

Species may be limited by environmental conditions that surpass physiological capabilities and hinder population densities (Bridle and Vines 2007; Sexton et al. 2009). A classification and regression tree model of presence-absence data \textcolor{red}{suggests} that \textit{A. picea} requires both sufficiently high summer temperatures and an absence of severely low temperature extremes to persist (figure 4). Summer temperatures are critical for both foraging activity and brood development over the course of the growing season (Penick et al. 2017) and may impact preparation for surviving the subsequent winter. Consistent with a hard limit on low-temperature adaptation, we found a trade-off between basal cold-tolerance and thermal plasticity in a common garden experiment using a quantitative genetic framework (cooler-warmer variation, figures 3C, 3D, 5). Forest ants from colder habitats were more basally tolerant but less capable of increasing tolerance through hardening, indicating that cold adaptation in this ant species involves genetic assimilation of cold tolerance, a shift from plastic to constitutive mechanisms (Waddington 1953; Pigliucci et al. 2006). Taken together, these results suggest that trade-offs in cold performance likely contribute to the current northern range boundary of \textit{A. picea}.

Species ranges edges often contract and expand over ecological and evolutionary time (Sexton et al. 2009). Although \textit{A. picea} historically has expanded its northward range rapidly (DeMarco and Cognato 2015), the high accuracy of the CART model suggests that populations have encountered hard climatic limits that limit further expansion (figure 4). The nodes representing \textit{T}\textsubscript{max} and precipitation during the warmest quarter suggest that the conditions over the growing season are critical for overall colony performance. Because poleward populations have shorter growing season lengths, selection will favor faster developmental rates and developing brood may require high summer temperatures to complete growth (Yamahira and Conover 2002).

Winters in poleward regions become more seasonal and extremely cold (Sunday et al. 2011). \textcolor{red}{Temperature} Seasonality, which was highly correlated with \textit{T}\textsubscript{min}, represents a critical node in our regression tree model and suggests that selection will favor both cold tolerance and hardening ability (Teets et al. 2011). Basal cold tolerance offers protection and resistance when organisms first encounter cold stress, while hardening enhances the response upon repeated or sustained exposure (Chown and Terblanche 2006). Both basal and hardening ability varied clinally with local cold extreme temperatures, a pattern consistent with local adaptation. Models of local adaptation suggest that populations are able to expand if they are able adaptively track environmental gradients (Bridle and Vines 2007). However, expansion will halt if the environmental gradient is too steep and suggest that populations encounter a hard limit (Bridle and Vines 2007). Meta-analyses of invasive species support range limit models, because the extension of lower thermal limits across latitude corresponds with range expansion (Lancaster 2016). Therefore, lower thermal limits represent a critical functional trait for understanding evolutionary range limits in poleward boundaries.

The genetic architecture of cold performance suggests an evolutionary constraint at the northern range boundary for \textit{A. picea}. \textcolor{red}{We found no support for additive variation in cold performance} (figures 3A, 3B), which would promote range expansion because cold performance could respond adaptively to both variable temperature and cold extremes. Instead, we found evidence for a trade-off between basal cold tolerance and hardening ability (cooler-warmer variation, Kingsolver et al. 2001), which could limit the evolutionary response to selection at the northern edge of \textit{A. picea}'s distribution.

The trade-off between basal cold tolerance and hardening ability varied with \textit{T}\textsubscript{min}, suggesting that cold adaptation proceeds by investing in constitutive responses as opposed to induced ones (cf. Elkinton et al. 2017). The degree of correlation between constitutive and induced mechanisms will depend on the extent to which they share common molecular pathways and physiological modes of action (Saltz et al. 2017). Under genetic assimilation, a plastic response becomes fixed (constitutive) and suggests that plastic and constitutive responses share common mechanisms (Waddington 1953; Pigliucci 2006). Both structural and energetic changes are common responses to cold stress (Bale 2002). For example, metabolite profiles responding to cold stress change less and are maintained at a higher level under pre-stressed conditions in experimentally cold-adapted fruit-fly lines (Williams et al. 2014). In natural populations of fruit flies, basal and induced cold tolerances are somewhat negatively correlated and share some common functional pathways, but they also have a some degree of independence because specific genetic loci (SNPs) do not overlap (Gerken 2015). For example, basal cold tolerance and hardening were enriched in genes associated with reduced programmed cell death (apoptosis), but there were more genes associated with the recycling of macromolecules (autophagy) under hardening (Gerken 2015). Thus, basal cold tolerance may constrain the evolution of plasticity in fruit flies due to their negative correlation, which is pervasive across the whole clade (Nyamukondiwa et al. 2011). However, constraints in cold performance do not appear pervasive among other terrestrial ectothermic species (Gunderson and Stillman 2015), suggesting that distributions are under constant flux.

\section*{Conclusion}

To fully understand the range limits of species, it is critical to identify the direction and agents of selection (MacColl 2011) and assess whether populations can respond to those selective regimes (Wilson 2006). Previous models have shown that populations can expand if they are locally adapted in a single quantitative trait along an environmental gradient and assuming constant genetic variance (Kirkpatrick and Barton 1997; Bridle and Vines 2007). However, marginal populations will decline if environmental gradients are too steep for available genetic variation to overcome or through influx of maladaptive alleles (Brine and Vines 2007). In some instances, even ample amount of genetic variation in a single trait can still lead to local extirpations in the face of selection (Barton and Keighley 2002; Barton and Patridge 2000). Part of this conundrum is that selection can be indirect, and suites of functionally relevant traits can constrain one another (Arnold and Lande 1983), as may be the case for \textit{A. picea}. Future models for range limits should consider the full architecture of ecologically relevant functional and quantitative traits likely targeted by selection.

Page 69: 2018-06-28. notes on messing with hipergator

logging in

> ssh @hpg.rc.ufl.edu

Using cyberduck to transfer files.

To show what types of programs there are.

> module spider

Ok, I'm on hipergator. I need to use it to compute large weighted co-expression networks in the cerasi dataset.

How should I approach this?

I think just creating a script with a small network that works on my local is a good start, then estimate network properties on the larger dataset.

Right now, I have the WGCA analysis with the data parsing, Enrichment and other network approaches (building a graph and estimating centrality)

Separate out script into a set of scripts

1. parsing - grab datasets that are significant for different classes +WGCNA - module detection + Functional Enrichment assignments
    - Save and output datasets
2. Network analyses

Holding some WGCNA code for detecting soft power to create scale free networks

```
#allowWGCNAThreads(nThreads = 4)
#powers = c(c(1:10), seq(from = 12, to=20, by=2))
#sft = pickSoftThreshold(t(timeeff.sub[,-1]), powerVector = powers, verbose = 5)

#sizeGrWindow(9, 5)
#par(mfrow = c(1,2));
#cex1 = 0.9;
#plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
 #    xlab="Soft Threshold (power)",ylab="Scale Free Topology Model Fit,signed
R^2",type="n",
  #   main = paste("Scale independence"));
#text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
 #    labels=powers,cex=cex1,col="red");
#abline(h=0.90,col="red")
#20? in this case (softpower)
###################
# Mean connectivity as a function of the soft-thresholding power
#plot(sft$fitIndices[,1], sft$fitIndices[,5],
 #    xlab="Soft Threshold (power)",ylab="Mean Connectivity", type="n",
  #   main = paste("Mean connectivity"))
#text(sft$fitIndices[,1], sft$fitIndices[,5], labels=powers, cex=cex1,col="red")
##############################

#Scaling of Topological Overlap Matrices to make them comparable across sets
#softPower = 20; #change from above
```

Page 70: 2018-06-28. Meeting with Ruchir and training on ultracentrifuge

Ultracentrifuge is from Beckman Coulter: Optima TLX 120,000 RPM

Rotor = TLA 120.2 Beckman Coulter (http://www.laborgeraete-beranek.de/info/TLA-120.2.pdf)

Details: 120,000 rpm max speed. Can use 1-2mL tubes, usually polycarbonate and polyallomer that have to be thickwall.

- 1mL Polycarb thickwall - part number = 343778
- 1 mL polyallomer thickwall- part number= 347287

These tubes need to be at least half filled without caps

Temp limits = 2-25C ; thaw to 2 C before centrifuging

We need to calculate how many RPMs for 100,000g

Formula: RCF = 1.12r (RPM/1000)2

100,000g = 1.12r (RPM/1000)2

RPM= 52,987.99; just round to 53,000

How do we clean the tubes?

Here's an approach https://www.researchgate.net/post/Re-using_ultracentrifuge_tubes2

> We use beckman polyallomer tubes for lentiviral centrifugation and we often reuse them. For that, after you are done with one harvest, just add 10% bleach to the tube for 2-3 min and rinse it out. Then spray with 70% ethanol a few times and rinse it. Let the tubes dry. When you want to reuse the tubes, spray them with 70% ethanol twice. Let them dry. Then rinse them with PBS and DMEM and use them for the virus prep.

Critical Notes:

- Do not keep rotor in the ultracentrifuge.
- Keep rotor clean, if any liquid leaks, clean it up
- Make sure samples are balanced
- For 1 mL tubes, do not fill over 3/4's of the tube.

1. Turn on the machine
2. Set the spped(rpm), time (days:min), and temp (C).
3. Once the settings are set, press *display*
4. Then press start
5. Watch centrifuge until it goes to full RPM(speed).

Page 71: 2018-06-29. Working on hipergator

Sample script for submitting a job on the cluster (https://help.rc.ufl.edu/doc/Annotated_SLURM_Script):

04_cluster_script.sh

path: /home/andrew.nguyen/Cerasi_Networks/Script/

```
#!/bin/bash

#SBATCH --job-name=Test_R_script
##SBATCH --mail-user=andrew.nguyen@ufl.edu
#SBATCH --mail-type=ALL
#SBATCH --output my_job-%j.out
#SBATCH --nodes=4
```

```
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem=120gb
#SBATCH --time=72:00:00

date;hostname;pwd

module load R
```

It works on the home directory on the hipergator server.

it looks like the *sbatch* command is what we use to run scripts

```
sbatch 04_cluster_script.sh
```

error

> sbatch: Warning: can't run 1 processes on 4 nodes, setting nnodes to 1 sbatch: error: Batch job submission failed: Job violates accounting/QOS policy (job submit limit, user's size and/or time limits)

Sent in a ticket for help. My account was put under a group that has no resources, they will fix(re-assign me to right PI).

Ok, now I have acess. OK, what am I doing.

I have a cerasi project on the cluster:

path = /home/andrew.nguyen/Cerasi_Networks/Script

I have a few scripts

```
ls -al
total 12
drwxr-xr-x 2 andrew.nguyen dhahn    5 Jun 29 16:36 .
drwxr-xr-x 4 andrew.nguyen dhahn    4 Jun 28 12:39 ..
-rw-r--r-- 1 andrew.nguyen dhahn 1728 Jun 28 12:41 03_2018-06-
28_Network_analyses.Rmd
-rw-r--r-- 1 andrew.nguyen dhahn  247 Jun 29 16:31 03_test.R
-rwxrwxrwx 1 andrew.nguyen dhahn  354 Jun 29 16:35 04_cluster_script.sh
```

I'm testing to see whether my shell script(04_cluster_script.sh) can execute the R test script (03_test.R).

04_cluster_script.sh

```
#!/bin/bash
#SBATCH --job-name=Test_R_script
##SBATCH --mail-user=andrew.nguyen@ufl.edu
#SBATCH --mail-type=ALL
#SBATCH --output my_job-%j.out
#SBATCH --nodes=4
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem=120gb
#SBATCH --time=72:00:00
```

```
date;hostname;pwd

module load R


cd /home/andrew.nguyen/Cerasi_Networks/Script

Rscript 03_test.R
```

03_test.R

```
library(ggplot2)
library(qgraph)
library(data.table)
library(WGCNA)


#reading in a dataset
dat<-fread("../Data/02_script_01_2018-06-
28_overall_pop_level_diffs_no_time.csv")
str(dat)
fwrite(dat ,"../Data/test.csv")

# Session Info

sessionInfo()
```

So if I run this script, I'm wondering if the R on the cluster will have qgraph installed and wehther I'll get a new dataset in Data/.

Tried this code and it worked:

path =/home/andrew.nguyen

```
./Cerasi_Networks/Script/04_cluster_script.sh
```

Ok, submitting job:

```
srun ./Cerasi_Networks/Script/04_cluster_script.sh
```

Yep, works.

Ok, time to try a test script that estimates centrality for a smaller dataset (overall population level differences in cerasi)

cat 03_2018-06-28_Network_analyses.R

```
#Libraries
```

```
library(ggplot2)
library(qgraph)
library(data.table)
library(WGCNA)


# Network approaches/analyses

#Some goals:

#* Identify whether gene modules are more hub-like or distributed at the
periphery of a network.
#  * we'd have to have a comparable network for all modules
#  * within this large network, figure out the connectivity for each gene in the
module

### Overall population level differences dataset

popdiff<-fread("../Data/02_script_01_2018-06-
28_overall_pop_level_diffs_no_time.csv")

#create adjacency matrix

adj.pop<-adjacency(t(popdiff[,-1]), power =1)

# creating a network from adj matrix
graph_pop<-qgraph(adj.pop,layout="spring")

# grabbing measures of centrality
cent.pop<-centrality(graph_pop)

popdiff$cent<-cent.pop[[1]]
fwrite(popdiff,"2018-06-28_pop_diff_centrality.csv")



#sessionInfo()
```

Recoding the cluster script to execute '03_2018-06-28_Network_analyses.R'

cat 04_cluster_script.sh

```
#!/bin/bash
#SBATCH --job-name=Test_R_script
##SBATCH --mail-user=andrew.nguyen@ufl.edu
#SBATCH --mail-type=ALL
#SBATCH --output my_job-%j.out
#SBATCH --nodes=4
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem=120gb
#SBATCH --time=72:00:00


date;hostname;pwd
```

```
module load R


cd /home/andrew.nguyen/Cerasi_Networks/Script

Rscript 03_2018-06-28_Network_analyses.R
```

Ok, lets run it!

```
srun ./Cerasi_Networks/Script/04_cluster_script.sh
```

Page 72: 2018-07-02. Writing strategies for range limits ms

For the fucntion valued trait paragraph. address editor's comments. There can be local differences in G. And the ultimate response to selection depends on the additive genetic variance. YOu cant mate ants though.

reviewer 1, talk about growing degree days in discussion to cover alternative mechanisms

reviewer 2, lay off on causal sentences.

SHC suggests to have range limits as an end conclusion, not up front.

Maybe change title.

paragraphs so far:

> Species distributions often contract and expand over ecological and evolutionary time (Sexton et al. 2009). Although \textit{A. picea} historically has expanded its northward range (DeMarco and Cognato 2015), the categorical cutoffs identified in the CART model suggests that populations have encountered a hard limit due to temperature and temperature variation (figure 4). However, transplant experiments from the core and edge to beyond the range () are needed to test for limits in adaptation and to truly determine \textit{A. picea}'s species boundary (). Nonetheless, evaluating evolutionary constraints of critical functional traits can provide insights into how populations may respond to selection (). In order to respond to both cold temperature and temperature variation, ant colonies will need to be able to have high basal cold tolerance and greater cold hardening ability. Adapting a quantitative genetic analysis, we found a negative correlation between cold hardening and basal cold tolerance, a genetic architecture that would limit A. picea's response to selection due to cold temperature and temperature variation. Furthermore, this trade-off was clinally structured, suggesting that cold adaptation in A. picea proceeded by enhancing cold tolerance at the cost of phenotypic plasticity. Taken together, these results suggest that trade-offs in cold performance likely contribute to the current northern range boundary of \textit{A. picea}.

> Niche modeling identified potential and multiple agents of selection at the northern range boundary of A. picea. The nodes representing \textit{T}\textsubscript{max} and precipitation during the warmest quarter suggest that the conditions over the growing season are critical for overall colony performance. Because poleward populations have shorter growing season lengths, selection will favor faster developmental rates and developing brood may require high summer temperatures to complete growth (Yamahira and Conover 2002). \textcolor{red}{Temperature} seasonality, which was highly correlated

> with \textit{T}\textsubscript{min}, represents a critical node in our regression tree model and suggests that selection will favor both cold tolerance and hardening ability (Teets et al. 2011). Basal cold tolerance offers protection and resistance when organisms first encounter cold stress, while hardening enhances the response upon repeated or sustained exposure (Chown and Terblanche 2006).

In paragraph 1 i have an argument about hard limits, maybe put that in paragraph 2. In paragraph 1, play up the fact that cold tolerance is a critical trait for understanding range dynamics?

Page 73: 2018-07-03. Different strategies for constructing weighted co-expression networks in cerasi dataset

We have time series gene expression for 2 populations (lowland and highland) from 2 month - 4.5 months prior to when they typically eclose. Highland ecloses later than lowland.

We want to know how the structure and individual gene members in a network differ between populations in a way that is associated with their seasonal timing. So we have to exclude just overall differences between populations unrelated to time.

The data:

- We have log2(fold change) expression estimates for each time point for each population from an EDGER model (80k transcripts)

Genes that are signficant by category

Source	Significantly differentially expressed
Altitude	452
Time	3014
Altitude + Time	58
Altitude * Time	14564

Ways to construct networks (weighted co-expression, undirectional):

If we want to compare across sources of variation and determine their centrality: We'd expect genes in Alt * time effect to be more central?

- Build a network for high and lowland populations for the whole set of genes that are significant
    - This would reduce the time series into a single network though; we need replicates to estimate a network at a particular time point

If we want to compare within a source of variation: We'd expect modules showing divergence in earlier expressing genes to be more hublike/central than modules that diverge later in time.

- Build individual networks for high and lowland populations for each source of variation.
    - This would also reduce the time series into a single network ; we need replicates to estimate a network at a particular time point

Notes:

- If we want to build a network for each time point, we'd need a metric of gene expression with replicates for each time point

---

Page 74: 2018-07-03. Range limits ms: re-analysis with MAT

ANCOVA, testing effect of pretreatment temperature and its interaction with mean annual temperature on cold tolerance.

```
cold.mod1<-aov(treatment_recovery_s~MAT*pretreat_Temp+Colony,data=k.dat)
> #cold.mod1<-lm(treatment_recovery_s~Tmin*pretreat_Temp+Colony#,data=k.dat)
> summary(cold.mod1)
                  Df  Sum Sq Mean Sq F value   Pr(>F)
MAT                1  137239  137239   6.282 0.014879 *
pretreat_Temp      1  261310  261310  11.961 0.000996 ***
Colony            19  208333   10965   0.502 0.951730
MAT:pretreat_Temp  1  214946  214946   9.839 0.002630 **
Residuals         61 1332640   21847
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Relationship between basal and MAT

```
cold.mod101<-lm(treatment_recovery_s~MAT,data=subset(k.dat,pretreat_Temp=="25"))
> summary(cold.mod101)

Call:
lm(formula = treatment_recovery_s ~ MAT, data = subset(k.dat,
     pretreat_Temp == "25"))

Residuals:
        Min       1Q   Median       3Q      Max
-212.354  -81.225    7.096   64.495  307.747

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)  -257.76     228.56  -1.128 0.273464
MAT           154.75      38.11   4.061 0.000667 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 128.8 on 19 degrees of freedom
Multiple R-squared:  0.4647,    Adjusted R-squared:  0.4365
F-statistic: 16.49 on 1 and 19 DF,  p-value: 0.0006666
```

Relationship between induced at 0 and MAT

```
cold.mod102<-lm(hardening~MAT,data=assd)
> summary(cold.mod102)
```

```
Call:
lm(formula = hardening ~ MAT, data = assd)

Residuals:
    Min      1Q  Median      3Q     Max
-196.57 -119.52  -21.84   66.60  372.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -565.46     293.61  -1.926   0.0692 .
MAT           135.81      48.95   2.774   0.0121 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 165.5 on 19 degrees of freedom
Multiple R-squared:  0.2883,    Adjusted R-squared:  0.2509
F-statistic: 7.697 on 1 and 19 DF,  p-value: 0.01208
```

Page 75: 2018-07-12. Meeting with TPowell diapause exit ms

Agenda:

1. Go over figures
    - order ok? axes labels ok ?
    - diapause exit over termination timing?
    - figure 3
    - panel C, it is mass adjusted for baseline and p
        - exponential scaling and not exponential phase
        - put the parameter letters as italics.
2. Where are we publishing?

- Tom wants to get out.

Future directions on ms:

- Timing- Tom will finish the last version of the results. Stats light in methods and then stats in results. Early next week, he'll send us results. Discussion will need to be cut.
    - Tom will send dsicussion too.
- Andrew fix figures , tweaks

Page 76: 2018-07-16. Problem with trikinetics computer: lost data

There was a power outage on 2018-07-13, 16:29:04, (line 383617 on monitor 6), which turned off the trikinetics computer. And we turned the computer back on, on 2018-07-16, 11:39:00. We may want to get rid of this line of the data because it includes all the behavior counts from the past 3 days.

Page 77: 2018-07-17. Meeting with Gragland, networks

Use FPKM- normalized count of fragments base don the overall representation in the library (dual normalization)

EdgeR will produce FPKM; double check for library size.

Think about interpretation.

At what point does correlation become noise, vs signal.

Algorithm to threshold to account for shape of the correlation matrix. THINK about thresholding.

Be careful of the metrics. Some take into account edge weight. The ones that just count edges, you need thresholding.

Fiddle with EDGER to determine FPKM values.

Good idea? Yes!

Page 78: 2018-07-18. SHC range edge adaptation ms

SHC asking for another analysis:

> Could you quickly graph me the relationship between baseline and hardening? How tight is it really? What's the R2?

```
ggplot(ksub,aes(y=coldplot,x=hard.zero))+geom_point()+stat_smooth(method="lm")
> summary(lm(coldplot~hard.zero,data=ksub))

Call:
lm(formula = coldplot ~ hard.zero, data = ksub)

Residuals:
    Min      1Q  Median      3Q     Max
-186.60  -44.03   13.90   42.18  114.03

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 651.32917   27.07930  24.053 1.09e-15 ***
hard.zero    -0.81063    0.08839  -9.171 2.08e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.59 on 19 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.806
F-statistic:  84.1 on 1 and 19 DF,  p-value: 2.083e-08
```

Wow, the correlation is tighter than I expected.

Some more....

> Interesting – could you do me a favor? Flip the axes (since baseline is the one expected to be driving the loss of hardening), and get rid of Vermont (I assume that is the far-right point).

```
ksub.maine<-ksub%>%filter(State=="Maine")
ggplot(ksub.maine,aes(x=coldplot,y=hard.zero))+geom_point()+stat_smooth(method="
lm")
summary(lm(hard.zero~coldplot,data=ksub.maine))
all:
lm(formula = hard.zero ~ coldplot, data = ksub.maine)

Residuals:
    Min      1Q  Median      3Q     Max
-95.476 -51.133  -6.479  54.165 126.942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 713.1477    72.2246   9.874  5.9e-08 ***
coldplot     -1.0230     0.1396  -7.326  2.5e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.37 on 15 degrees of freedom
Multiple R-squared:  0.7816,    Adjusted R-squared:  0.767
F-statistic: 53.67 on 1 and 15 DF,  p-value: 2.501e-06
```

Page 79: 2018-07-23. Paper notes: Salachan and Sørensen 2017, JEB

doi: 10.1242/jeb.165308

Intro/background

Organisms face temperature changes that influence their performance and activity over space and time. Temp fluctuations differ depending on

Big Question

How does mean temperature and variable temperature influences temperature responses?

Hypothesis

- Outcomes depend on experimental protocol

Experimental approach

Fluctuate temperatures around the mean at +/-4 and +/- 8.

THermal assays measured during heating phase of fluctauting regimes Important note

...Under developmental Acclimation:

They grew eggs under 19 or 23 degree C under 0, 4, 8 temperature fluctuations and measured adult thermal hardiness.

Under adult acclimation

They grew eggs under 19 or 23 degree C under 0, 4, 8 temperature fluctuations, transferred to constant or fluctuating conditoins, and then measured adult thermal hardiness. Important note: 19 and 23 degree C were separated and samples were not crossed over for this experiment

Under adult acclimation with temperature change

They grew eggs under 19 or 23 degree C under 0, 4, 8 temperature fluctuations, transferred to constant or fluctuating conditons for both 19 and 23 temperature, and then measured adult thermal hardiness.

Results

1. Main effect of treatment (temperature fluctuation 0, 4, 8) and main effect of temperature for both CTmmax and CTmin
2. Interaction between dev treatment and adult treatment for Ctmin and not CTmax under 4 temp fluctuations and not 8 temp fluctuations.

Conclusion

IT looks like CTmin is more sensitive to these temperature treatments than CTmax. CTmin may have more plasticity.

Thoughts:

Graphs obscure patterns and there are no post hoc tests. It is hard to know what is different from each other. I'd plot either CTmax or CTmin vs temperature fluct (0,4,8) and have different transfer (Dev-Adult condition;C-F, F-C, C-C, F-F) combinatios as separate lines.

Page 80: 2018-07-25. Proteome stability idea dump

Knowledge Gap: What is the actual effect of having misfolded/aggregated proteins. Is it a matter of less function or do they hinder function?

Hypothesis:

1. Misfolded/aggregated proteins hinder function
2. Misfolded/aggregated lose function

Experimental Approach: Compare enzyme activity in the presence/absence of misfolded/aggregated proteins

1. Create 2 unfolding curves
2. Measure enzyme activity in one ---Aggregate group or U + F group; treat 25 and 70C
3. In the other, ultracentrifugate, and then measure enzyme activity ---Non-aggregate group or U group; treat 25 and 70 C

Expected Outcomes:

1. Hinder result: interaction between treatment(U+F vs F) and temperature on enzyme activity
    - enzyme activity is lower in U + F under 70C compared to 25 C .
2. Loss result : additive effect of temperature on enzyme activity
    - enzyme activity is higher in 25 than 70 C

Big picture:

Pitch in terms of drop in performance under temperature threats. Is the drop in performance related to how aggregates hinder enzyme activity?

Page 81: 2018-07-31. Updated to do list

1. Hsp rxn norm ms- on hold

2. range limits ms- revisions
    - I will redo figures that takes out VT samples; I also need to change the text ; including new stats and new colony numbers
    - SHC will take a crack at the results
3. Thermal niche ms-
    - Lchick to work on major revisions
    - I sent a version of abstract to NJG and he sent back edits. Fill in blanks and then send to Lchick
4. Started a literature spreadsheet for my post doc ; fill in as I go
    - It'll track questions/hypotheses/main results/limitations and paper details in the spreadsheet
5. Proteome stability project
    - We found variation in thermal traits in the DSPR lines
    - Want to test whether differences in thermal traits are reflected in differences in the stability of proteomes
        - problem with protein quantification - talk to Dan
6. Diapause exit ms
    - Wrote up abstract and results - in DHahn's hands
7. Cerasi data set & analysis
    - Build time series networks for each population
    - Fill in network literature in the tab in the spreadsheet under #4 on this list
    - I should be using FPKM
8. Tweak talk for evolution conference
    - Practice for Dan
    - Clean up figures to make them consistent, especially the difference in habitat types one.
9. Circadian rhythm proj
    - STill measuring, no entrainment samples right now.
    - Dylan is hopping in on trik monitors 1 and 2 to do his experiment. I moved out both datasets so he is collecting data in new files.

---

range limits stats

SHC wants pca of climate variables only for maine sites

```
clim.pca<-princomp(scale(maine[,23:29]))
summary(clim.pca)
clim.pca$loadings[,1:2]
```

	Comp.1	Comp.2
MAT	0.3434587	0.4656631
MDR	-0.4178970	0.3184612
ISO	-0.2746063	0.4465854
SD	-0.4394715	0.0623454
Tmax	0.1567641	0.6253765
Tmin	0.4517551	0.2360573
TAR	-0.4584722	0.1777525

new gmax loadings

```
knitr::kable(gload)
```

	loadings	temp
pretreat_Temp-5	0.2543283	-5
pretreat_Temp0	-0.4586843	0
pretreat_Temp25	0.6246196	25
pretreat_Temp5	0.5785984	5

regression models

filtering out VT first

```
maine<-ksub%>%
  filter(State!="Vermont")
```

basal

```
summary(lm(coldplot~MAT,data=maine))

Call:
lm(formula = coldplot ~ MAT, data = maine)

Residuals:
    Min      1Q  Median      3Q     Max
-172.45  -89.53  -33.33   90.05  178.51

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1213.31     250.70   4.840 0.000216 ***
MAT          -124.59      43.49  -2.865 0.011803 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 115 on 15 degrees of freedom
```

```
Multiple R-squared:  0.3537,    Adjusted R-squared:  0.3106
F-statistic: 8.208 on 1 and 15 DF,  p-value: 0.0118
```

hardening

without quadratic

```
summary(lm(hard.zero~MAT,data=maine))
Call:
lm(formula = hard.zero ~ MAT, data = maine)

Residuals:
     Min       1Q   Median       3Q      Max
-188.569 -135.468   -8.519   56.382  294.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -405.13     324.51  -1.248   0.2310
MAT           106.00      56.29   1.883   0.0792 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148.9 on 15 degrees of freedom
Multiple R-squared:  0.1912,    Adjusted R-squared:  0.1373
F-statistic: 3.546 on 1 and 15 DF,  p-value: 0.07921
```

With quadratic

```
summary(lm(hard.zero~MAT+I(MAT^2),data=maine))
lm(formula = hard.zero ~ MAT + I(MAT^2), data = maine)

Residuals:
    Min      1Q  Median      3Q     Max
-222.55  -81.84  -20.87  102.42  215.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6751.92    2634.89  -2.563   0.0226 *
MAT          2365.02     933.73   2.533   0.0239 *
I(MAT^2)     -198.45      81.91  -2.423   0.0296 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 129.4 on 14 degrees of freedom
Multiple R-squared:  0.4301,    Adjusted R-squared:  0.3487
F-statistic: 5.283 on 2 and 14 DF,  p-value: 0.01952
```

Correlation between hardening at 0 and basal cold tolerance

```
summary(lm(hard.zero~coldplot2,data=maine))

Call:
lm(formula = hard.zero ~ coldplot2, data = maine)

Residuals:
```

```
    Min      1Q  Median      3Q     Max 
-95.476 -51.133  -6.479  54.165 126.942 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 435.1594    36.9233  11.785 5.53e-09 ***
coldplot2    -1.0230     0.1396  -7.326 2.50e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.37 on 15 degrees of freedom
Multiple R-squared:  0.7816,    Adjusted R-squared:  0.767 
F-statistic: 53.67 on 1 and 15 DF,  p-value: 2.501e-06
```

splitting basal cold tolerance at the median and evaluating correlation

lower than 261

```
maine.low<-maine%>%
+   filter(coldplot2<median(coldplot2))
> summary(lm(hard.zero~coldplot2,data=maine.low))

Call:
lm(formula = hard.zero ~ coldplot2, data = maine.low)

Residuals:
        Min       1Q   Median       3Q      Max 
-119.031  -53.738    2.432   36.239  151.116 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept) 366.5633    56.4144   6.498 0.000632 ***
coldplot2    -0.2834     0.4464  -0.635 0.548913    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90.43 on 6 degrees of freedom
Multiple R-squared:  0.06296,   Adjusted R-squared:  -0.09322 
F-statistic: 0.4031 on 1 and 6 DF,  p-value: 0.5489
```

higher than 261

```
maine.hi<-maine%>%
+   filter(coldplot2>median(coldplot2))
> summary(lm(hard.zero~coldplot2,data=maine.hi))

Call:
lm(formula = hard.zero ~ coldplot2, data = maine.hi)

Residuals:
     Min      1Q  Median      3Q     Max 
-57.606 -18.338   1.451  27.219  38.604 
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept) 506.8081    72.6225   6.979 0.000431 ***
coldplot2    -1.2813     0.2061  -6.216 0.000801 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.76 on 6 degrees of freedom
Multiple R-squared:  0.8656,    Adjusted R-squared:  0.8432
F-statistic: 38.63 on 1 and 6 DF,  p-value: 0.0008008
```

figure 5 stats ; without vt samples

pre-treatment temp as a factor

```
> summary(aov(treatment_recovery_s~factor(pretreat_Temp)+Colony,new.dat3))
                     Df Sum Sq Mean Sq F value   Pr(>F)
factor(pretreat_Temp)  3 420808  140269   9.517 5.53e-05 ***
Colony                15 182778   12185   0.827    0.644
Residuals             45 663215   14738
```

Post hocs

```
Fit: aov(formula = treatment_recovery_s ~ factor(pretreat_Temp) + Colony, data =
new.dat3)

$`factor(pretreat_Temp)`
            diff        lwr       upr     p adj
0--5  -183.63021 -298.13214 -69.12827 0.0005460
5--5   -29.62500 -144.12694  84.87694 0.9003008
25--5   24.56771  -89.93423 139.06964 0.9398036
5-0    154.00521   39.50327 268.50714 0.0043922
25-0   208.19792   93.69598 322.69985 0.0000868
25-5    54.19271  -60.30923 168.69464 0.5910534
```

We really have a repeated measures anova design and having colony as a factor doesn't make sense because colonies will covary across pre treatment tempos

Here is the mixed effects model and colony is a random effect

```
summary(lmer(treatment_recovery_s~factor(pretreat_Temp)+
(1|Colony),data=new.dat3))
Fixed effects:
                      Estimate Std. Error      df t value Pr(>|t|)
(Intercept)             600.33      29.69   60.00  20.223  < 2e-16 ***
factor(pretreat_Temp)0  -183.63      41.98   60.00  -4.374 4.94e-05 ***
factor(pretreat_Temp)5   -29.62      41.98   60.00  -0.706    0.483
factor(pretreat_Temp)25   24.57      41.98   60.00   0.585    0.561
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

So, -5 pretreatment temp is the baseline comparison in this output. 0 is diff from -5,5,25.

Page 82: 2018-08-08 Circadian rhythm thoughts, data analysis

Some QC issues: Double check dates. One way to id whether the dates have been correctly inputted is to construct a table with number of days for free-run, entrainment, for every unique ID

For example, h4o4 has 3 days in entrainment which doesn't seem right. Double check

Some thoughts on data analysis.

- As for measuring biological rhythms. Try out different programs such as Lomb-Scargle method.
- So far, I have done a spectral density analysis and wavelet analysis(discrete). For the discrete wavelet, I could analyze this separate and get higher estimates of biological rhtyhms because the time series is longer than the entrainment. This way, we can compare rhythms that are on the order of days instead of hours.
    - An idea to bin the data: by 1 day
- I can also gather more accurate estimates of death days in the trikinetics data.
- I still need to figure out a way to get amplitude and phase parameters from these analyses.

Page 83: 2018-08-09. Detecting rhythms in R

A few notes, showing packages and code for each different type of method.

A good ref:

Leise, T. L. (2013, July 1). Wavelet analysis of circadian and ultradian behavioral rhythms. Journal of Circadian Rhythms. Ubiquity Press. https://doi.org/10.1186/1740-3391-11-5

Analysis:

1. Fourier Periodogram (TSA package - estimates dominant frequencies in a stationary time series)

```
library(TSA)

pergram<-periodogram(h4o13$counts15,xlim=c(0,.05))
str(pergram)
List of 16
 $ freq     : num [1:4000] 0.000125 0.00025 0.000375 0.0005 0.000625 ...
 $ spec     : num [1:4000] 2482 340 2169 535 692 ...
 $ coh      : NULL
 $ phase    : NULL
 $ kernel   : NULL
 $ df       : num 1.95
 $ bandwidth: num 3.61e-05
 $ n.used   : int 8000
 $ orig.n   : int 7801
 $ series   : chr "x"
```

```
 $ snames   : NULL
 $ method   : chr "Raw Periodogram"
 $ taper    : num 0
 $ pad      : num 0
 $ detrend  : logi FALSE
 $ demean   : logi TRUE
```

2. Maximum entropy spectral analysis - MESA : Estimates dominant frequency in a time series function, built in stats package

```
spec.ar()
```

I went with a spectral density analysis, which is similar : It grabs the highest power + frequency, but there isn't a way to tell whether the periods are significant.

```
sa.an<-function(ts=counts15$counts15){
  sa1<-
spectrum(ts,method=c("pgram","ar"),plot=FALSE,demean=TRUE,detrend=TRUE,tape=.2)
  spx<-sa1$freq
  spy<-2*sa1$spec
  pw<-data.frame(spx,spy)
  cc1<-pw[order(pw$spy,decreasing=TRUE),]
  cc2<-subset(cc1,spx<0.05)
  cc2$density<-density(cc2$spy,n=length(cc2$spy))$y
  cc2$t<-density(cc2$spy,n=length(cc2$spy))$x
  #cc<-findpeaks(cc2[,3],minpeakheight=1E-6)
  cc<-findpeaks(cc2[,1])
  cc2[order(cc2$density,decreasing=TRUE),]
  out<-1/cc2[cc[,2],][,1]/4
  #out<-1/cc2[cc[,2],][,1]/4/24
  return(out[1:4])
  ## hours


}
```

3. Discrete wavelet analysis:(wmtsa package) Evaluates the amplitude and period/frequency at different levels of time(freqnecy bands). More technically, the analysis produces different details, whose sum, gives the original time series. More simply, you can get amplitude and period as a function of time. This approach is also good for non-stationary time series, meaning that amplitude and period can change over time!

Lumping continuous in here, because its the same package and similar approach.

Note: usually translation-invariant DWT with Daubechies least assymmetric filter of 12.

I wrote a function to grab different filters

```
globdwf<-function(vec=test$counts15,Jcirc=6){
  DJt_circ <- wavMRDSum(vec,levels=Jcirc ,keep.smooth=FALSE,
keep.details=TRUE,reflect=TRUE,wavelet="s12",xform="modwt")
  IBL<-data.frame(findpeaks(DJt_circ))
  names(IBL)<-c("Height","mid_time","initial_time","final_time")
  return(mean(diff(IBL$mid_time)/4))
}

globdwf()
```

4. Lomb-scargle periodogram: (lomb package) similar to other periodograms but this analysis can handle unevenly sampled data. The cool thing about this analysis is that in R, it can set a threshold for significant power and can objectively evaluate period estimates.

```
library(lomb)
a<-lsp(h4o13$counts15,times=h4o13$time/4,type="period",from=5,to=100)
summary(a)
a$peak
a$peak.at[1]
```

There is also this online program: http://132.187.25.13/actogramj/review.html

```
A brief review of automatic period estimation methods
The period estimation by eye fit line was the classically used method, whereas
today the automatic calculations are preferred because of their objectivity.
However, the automatic methods should not be overestimated because most of
biological data contain noise, which prevents or influences precise period
estimation. Therefore the eye fit line is still useful and sometimes necessary
to confirm the automatic periodogram analyses. Also, the period estimation by
eye fitted line is still useful when the rhythms are composed of multiple
components, the periods of which are often difficult to find automatically.

Fourier analysis is a classical periodogram analysis technique and is still
widely used today. This method is often employed for short-term rhythms such as
bioluminescence of transgenic organisms ([8]). Chi-square periodogram analysis
is the most widely used method today. However, since at least 10 cycles of the
rhythms are required to apply the statistical test ([3]), we alternatively
implemented the Lomb-Scargle periodogram method in ActogramJ. The Lomb-Scargle
periodogram is especially suited to analyze unequally spaced time series data,
but it also shows an outstanding performance for equally spaced data ([1, 2]).
The Lomb-Scargle periodogram is based on Fourier analysis ([7]), so that both
methods show almost the same result when the data points are equally spaced.
Recently the Lomb-Scargle periodogram was even satisfactorily applied to gene
expression data by DNA-array studies ([6]), suggesting the reliability of the
statistical evaluation for short term data. Furthermore the Lomb-Scargle
periodogram is noise tolerant in comparison to the chi-square periodogram ([3]).
Thus there are several superior features in Lomb-Scargle periodogram compared to
chi-square periodogram. Van Dongen et al. [1] even recommended using it as a
default method. However, the Lomb-Scargle periodogram is not the best method for
all data: because the periodogram is based on the least-square fitting of sine
waves to the data, the period estimation for non-sinusoidal rhythms such as
bimodal rhythms is less precise ([9]). Therefore the chi-square and Lomb-Scargle
periodograms would be complementary. These periodogram analyses can be applied
readily with a few mouse clicks in ActogramJ. One can quickly compare the
results obtained by the different methods, which provides an additional check
for their correctness.
```

Missing gaps: Phase shifts? Amplitude?

Page 84: 2018-08-13. Rhagoletis summer collections

What are the numbers so far?

Host	treatment	organism	counts
Apple	RT	fly	230
Apple	SO	fly	160
Apple	SO	wasp	2
Haw	RT	fly	40
Haw	RT	wasp	16
Haw	SO	fly	12

Ideal numbers for 2018:

Experiment	Host	Flies	Wasps
RT	Apple	100	100
SO	Apple	100	100
RT	Haw	100	100
SO	Haw	100	100

Page 85: 2018-08-14. Descrepancies on rhagoletis sampling for biological rhythms

I am double checking trikinetics dates and positions on the master spreadsheet.

Sample: h404

Problem: The exit date from entrainment (2017-10-18) was not consistent with entry into free run (2017-10-24). I double checked the free run date it was 2017-10-24,

Solution: I changed the entrainment exit date to 2017-10-24 (nb#001, page 164).

Sample : 2b23

Problem: The free run date does not match entrainment dates.

Entrainment entry- 2017-10-20 Entrainment exit- 2017-10-29

Free run entry- 2017-10-16 (this is before it eclosed) Free run exit - 2018-01-20

1. There is no entry for 2b23 for free run on 2017-10-29, nb #001, page 164.
2. On nb#001 page 152, 2b23 (trik -position 2-16) was listed as frozen and it is also logged in the notes_2 column (AJ)

Solution: Remove all free run dates because the sample died.
Free_run_trik_monitor,Free_run_trik_position,Free_run_entry_date,Free_run_entry_time,Free_run_exit_date,Free_run_exit_time,notes_3,Adult_death_date 4,8,2017-10-16,20:17,2018-01-20,20:03,changed water 2017-11-02 22:54; disturbed to check death 2017-11-07 21:32; changed water 2017-11-17 22:39; changed water 2017-12-10 20:00,2018-01-20

Sample : 14b9

Problem: death date(exit free run date (2017-01-04)) is before the free run entry date (2017-11-04). But the logged death date in nb#002, page 59 is 2018-01-04.

Solution: Change free run exit date to 2018-01-04.

Sample: 13o45

Problem: death date(free run exit date) is note a date (43028). Death logged in nb#001, page 135. The sample was also frozen.

Solution : Changed free run exit date to 2017-10-20

Sample: 13o17

Problem: The free run exit time is a date (4/2/18) and not a time. Time(20:05) is in nb#004, page 28.

Solution. Changed free run exit time to 20:05

Sample: 13r12

problem: No free run trik monitor,position, time entry, and date entry. Position is 6-10. Time is 20:40. Date is 2018-04-11. Logged nb#003, page 10.

SolutioN: Entered in data.

Sample: 12b73

problem: entrainment exit (2018-04-05) different than free run entyr date (2018-04-06).Checked nb#002, page 190 and entry date should be 2018-04-06

SolutioN: Changed free run entry date to 2018-04-06.

Sample: 11o24,2r38 , h4o2 ,4o25

problem: free run exit date ()2017-01-04) is earlier than free run entry date (2017-11-16).But the logged death date in nb#002, page 59 is 2018-01-01.

Solution: Change free run exit date to 2018-01-04.

Page 86: 2018-09-03. Reading Murren et al. 2015; Constraints on the evolution of phenotypic plasticity: limits and costs of phenotype and plasticity

Two types of restrictions for performance in different environments:

1. Costs - reduction in fitness when a trait is produced via plasticity rather than constitutively
2. limits - inability to produce the optimal trait value

There should be a distinction between cost of plasticity and costs of phenotypic production- shouldn't be conflated.

"A cost of plasticity refers to the fitness decrement a highly plastic genotype pays relative to a less plastic genotype".

I think this is a good way of thinking about it, comparing among genotypes. I've been thinking about plasticity within a genotype and comparisons of performance across environment. But...if you look at figure 1....the cost of plasticity isn't consistent with Angiletta's definition (slope of the reaction norm) and they're not comparing the slope of the reaction norm. Is it even appropriate to ask about what the cost of plasticity is? Or is it better just to describe the phenomenon as- performance differences across an environment among genotypes (ie home site advantage in a reciprocal transplant experiment).

"A cost of phenotype refers in part to the fitness tradeoffs inherent in allocating resources to one trait vs another as well as the costs of obtaining information on the environment (Callahan et al., 2008)."

This distinction is a little bit difficult to agree on because you can imagine that the aforementioned cost to plasticity can be brought about by a cost in producing a phenotype in a given environment. So why the distinction? Figure 1 still looks like a cost of the phenotype....where they label cost of plasticity.

Conditions that favor plasticity:

Environmental Heterogeneity; but it should be within a lifetime.

When species encounter novel environments, plasticity can allow populations to move closer to the new phenotypic optimum and uncover phenotypic variation (cryptic genetic variation). But...these two items are intertwined, and they listed it as two unique things.

Constraints on plasticity

Slope and curvature can readily evolve in populations and species.

Page 87: 2018-09-13. Rhagoletis field collection notes 2018-08-29

Went to field and collected Apple fruit 2018-08-29. These notes have coordinates of where to find fruit. Even though we collected Apple, we have some coordinates for Hawthorn trees. We also collected crab apples.

2018-08-29

coordinates, trip w/ Dan Hahn, Hinal

Site

Lansing - Clinical Center, MSU campus Apple site

42° 42' 58" N
84° 27' 39" W

3 apple trees, 36 ins, large

* picking apples that are less than 50% damaged & not too damaged

crab apple collection
42° 43' 8" N
84° 28' 13" W 860 ft elevation

3 trees, close together, apples mostly down, very little on tree branches

- most were > 50% rotten, Dan found apple w/ 2 maggots

East Hebner Haws
42° 43' 34" N
84° 27' 53" W

looking at Haws
No fruit on branches

2 trees

Masted last year
One thinks this is why there
are 2 year diapause classes

Apple - Farm & Wilson road

42° 43' 33" N
84° 28' 19" W

3 trees 2 tubs

checking New Mitford

42° 43' 55" N
84° 28° 9" W

not very fruit on tree, mostly on
floor

Albrecht & Durant Apple site

42° 44' 4" N
84° 28' 17" W

on corner & 1 tree, good infestations in the past

many fruit on tree branches. Dan says to leave fruit collection & wait ~2 weeks when he comes back

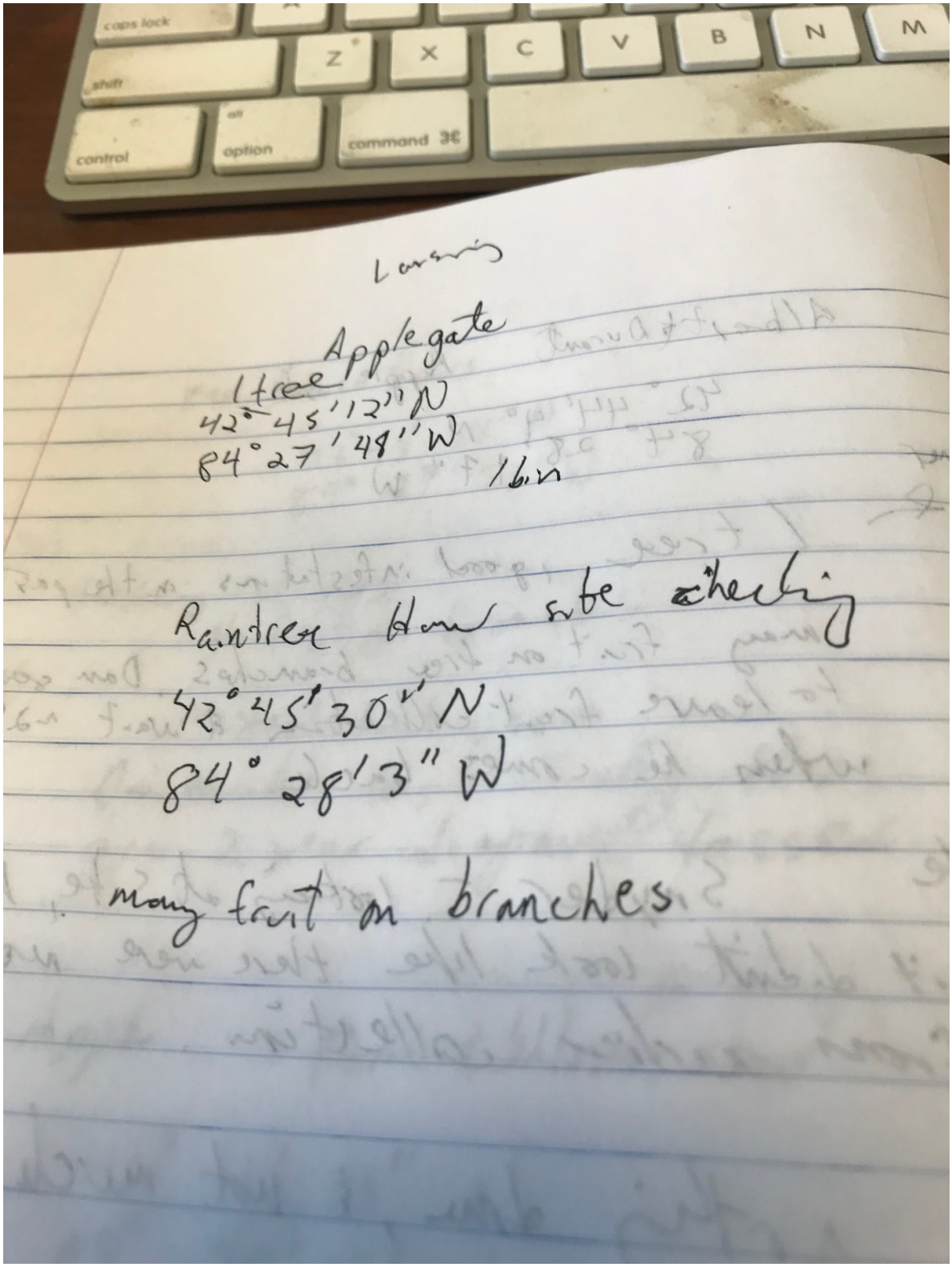
Apple site Snyder – looking at site, Dan says it didn't look like there were many fruit from earlier collection

nothing down, & not much on tree

42° 44' 33" N
84° 27' 45 W

Jim Smith Apple site behind his house
no fruit on trees, bad yield

42° 44' 36" N
84° 28' 4" W


caps lock
shift
control
option
command
Z
X
C
V
B
N
M
Lansing
Applegate
1 tree
42° 45' 12" N
84° 27' 48" W
16m
Rantree Home site checking
42° 45' 30" N
84° 28' 3" W
many fruit on branches

Grant Site

Fruit ridge - Haws
43° 13' 23" N
85° 46' 15" W
780 elevation

No Haws this year
but further down the street

Ferris road - has haw & Apples

haw (1 tree) 43° 18' 43" N
85° 50' 25" W

close to van Dyke Farm

1 apple tree 43° 18 42 N
85° 50' 26" W

Cant collect, not too many apples

1 apple tree Leonard's house
43° 18' 53" N
85° 50' 26" W

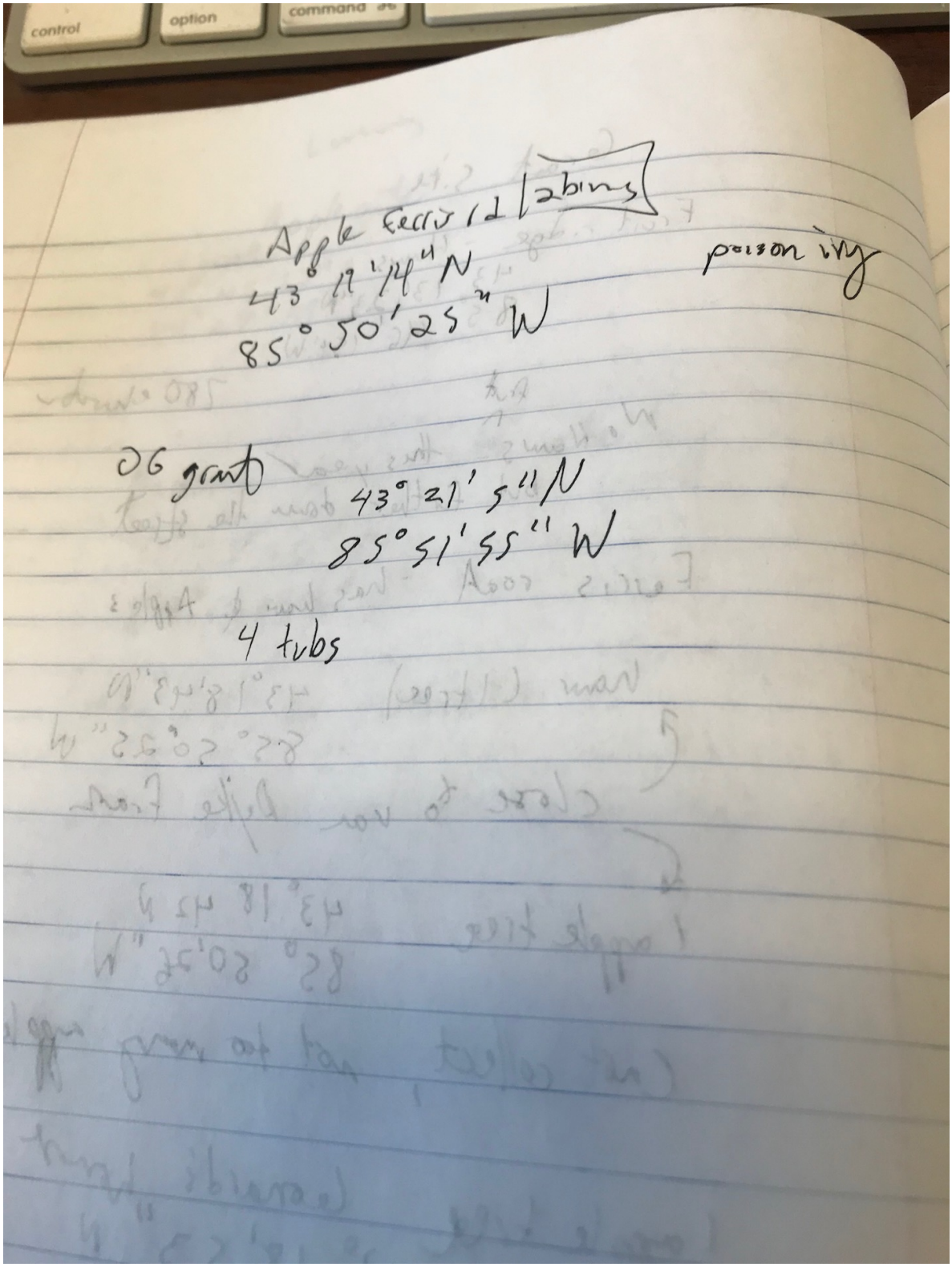
1 tub

---

1 apple tree 43° 18' 59" N
85° 50' 28" W

---

row of haw trees
43° 19' 12" N
85° 50' 25" W



Overall, Grant OG has yielded the most maggots. Infestation of crab apples is low.

Page 88: 2018-10-26. Thoughts on amnnat revisions

In the discussion, we have to make more changes to make the narrative more consistent with our data(so the reviewers and editors think). We have modelling of a species range and physiology at the range margin.

An approach to write about the study:

Paragraph 1- It summarizes the results and places it in a slightly broader content. We can add caveats here such as :

- we need reciprocal transplants and beyond the range transplants to determine local adaptation, and a cline is more of a signature of adaptation
- When we do the transplants, then do we see reduced organismal performance due to their cold tolerance.
- For the genetic constraints part, we cannot completely rule out the effect of maternal acclimation, and worker acclimation through a season. In fact, workers previous in the season may have different cold tolerances than ones later in the season.

Paragraph 2 - It goes into the occupancy model and and explains the meaning of each climate variable. We can do this more briefly/quickly and then discuss Angert's 2018 paper on climate modeling that tests the role of fitness and dispersal limitation. Dispersal limitation is when they don't fill in their range and there is suitable habitat. Fitness limitation is when their suitable habitat and occupancy align. These data suggest that they may be dispersal limited instead of fitness limitation.

Paragraph 3 - This paragraph goes leads the modleling into the physiology

When dispersal is limited, genetic drift can become more efficient and it actually is associated with increased trait independence at the species range. In contrast to these results, we identified a trade off in cold tolerance that varies across temperature, suggesting that selection is acting on these populations

---

Page 89: 2018-10-30. more thoughts on amnat discussion tweaks

last entry, it was not well thought out and implementing those ideas was difficult. SHC suggests creating a lit list so we can better place our arguments into a broader picture. So the lit list will have modeling and experimental papers.

From the modeling aspect, people do make claims about range limits. The models usually are logistic regressions to find variables that best explain occurrence. Generally, there are two broad forces that limit species: fitness and dispersal. If a species is dispersal limited, then this means they fail to fill in their niche, whereas, fitness limitation means that they've totally filled in their niche and they can't go outside of that space.

The problems with these models is that they're correlative and assume that a species range is static. Species ranges may not be static or at equilibrium, especially if they are dispersal limited. For our data, it may also assume that the colonies we found are edge and marginal populations that are not sink populations. Meaning, their fitness is low and cannot maintain populations; instead, their populations are totally sustained by migrants. The influence of dispesral limitation and fitness needs to be experimentally tested . In fact...Hargreaves study 2014, Amnat and get into the nuances. Also mention Angert et al. 2018, and how dispersal limitation could be more prevalent in poleward populations and that dispersal and fitness limitation can act together. Many of these studies measure survival at different life stages and fitness, which is really difficult for mobile animal species. Instead, proxies of fitness can be measured. Dispersal limitation can decrease genetic variation.

These ants postglacially expanded their range. And the evidence for dispersal limitation is mixed (Svennig et al. 2008), but generally, climate and dispersal limitation are both important. What could slow them down is expansion load, when a deleterious mutation becomes fixed , and it is more likely to be come fixed due to low effective population sizes. Increased mutation load may

influence the genetic architecture and subsequent response to selection.

Whether populations can respond to selective pressures at the range margin will depend on the genetic architecture of ecologically functional traits. Genetic variation decreases in the species range edges (paccard) and have been shown to make traits more independent, which may on the surface ehnahce the response to selection because there are more trait combinations for selection to operate upon. Yet, this contrasts with our results where traits are less independent. Infact the leading PC explained the majority of the variation and it is this axis that adaptation proceeds along. The genetic architecture of the leading PC (berger 2013 evolution) does change across latitude. And for poleward populations, they shift towards cooler-warmer variation. Cooler-warmer variation seems to be the prevailing thing for performance curves (Kingsolver 2015)

Most species may be at equilibrium (Hargraves, Lee-Yaw) because the majority of studies find that models match up with niche limits.

Willi 2018 and Paccard 2016, there amy be a link between expansion load and the quantitative genetic architecture. They find expansion load in exons/introns, but also increased trait independence at the range edges.

niche models that show range limits are niche limits does not account for whether populations are locally adapted.

however which ecological traits are important at the range margin is unclear

---

Page 90: 2018-11-05. running stuff on hipergator

New script,

```
#!/bin/bash
#SBATCH --job-name=Test_R_script
##SBATCH --mail-user=andrew.nguyen@ufl.edu
##SBATCH --mail-type=ALL
#SBATCH --output=my_job-%j.out
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem=110gb
#SBATCH --time=72:00:00
#SBATCH --partition=bigmem
#SBATCH --account=dhahn
#SBATCH --qos=dhahn-b
date;hostname;pwd

module load R

export ALLOW_WGCNA_THREADS=4

cd /ufrc/dhahn/andrew.nguyen/Cerasi_Networks/Script

Rscript 03_2018-10-31_Network_centrality_subsets_each_timepoint_population.R

date
```

notes:

Run with sbatch

```
sbatch run.sh
```

I need to be in the /ufrc/dhahn/andrew.nguyen/ directory, not home directory

Some help from help desk :

```
Oleksandr Moskalenko 2018-11-05 13:49:40 EST
I did a short test run. Here's what I saw:

==========================================================================
*
*  Package WGCNA 1.66 loaded.
*
*    Important note: It appears that your system supports multi-threading,
*    but it is not enabled within WGCNA in R.
*    To allow multi-threading within WGCNA with all available cores, use
*
*          allowWGCNAThreads()
*
*    within R. Use disableWGCNAThreads() to disable threading if necessary.
*    Alternatively, set the following environment variable on your system:
*
*          ALLOW_WGCNA_THREADS=<number_of_processors>
*
*    for example
*
*          ALLOW_WGCNA_THREADS=32
*
*    To set the environment variable in linux bash shell, type
*
*           export ALLOW_WGCNA_THREADS=32
*
*     before running R. Other operating systems or shells will
*     have a similar command to achieve the same aim.
*
==========================================================================


It looks like wgcna can use multiple threads, so you should set

export ALLOW_WGCNA_THREADS=4

in the job script. Then, you can use

#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4

in the job resource request if wgcna constitutes a significant portion of the
analysis.

I modified run.sh accordingly.
```

```
I see that you just moved the directory tree in question to /ufrc. Let's create
a shortcut to make it easier to change into that directory tree:

[andrew.nguyen@login1 ~]$ ln -s /ufrc/dhahn/andrew.nguyen/ ufrc
[andrew.nguyen@login1 ~]$ cd ufrc/Cerasi_Networks/Script

Let's submit a test job.

[andrew.nguyen@login1 Script]$ sbatch run.sh
Submitted batch job 27514270

   PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND

363261 andrew.+  20   0 8699748   6.7g  10924 R  99.7  0.4   1:36.23 R

So far it's only using 1 core, but perhaps this is the parsing phase.

my_job-27514399.out is the job log. The job is running.
```

Page 91: 2018-11-06. Comparing qgraph and igraph packages for estimating networks

Comparing qgraph and igraph function from an adjacency matrix. I made a small dataset with 4 genes.

```
library(microbenchmark)
microbenchmark(qgraph(adj,layout="spring"),graph_from_adjacency_matrix(adj,mode=
"undirected",weighted=TRUE))
Unit: microseconds
                                                                        expr
 min          lq        mean
                                        qgraph(adj, layout = "spring")
225917.868 304599.3010 417874.8655
 graph_from_adjacency_matrix(adj, mode = "undirected", weighted = TRUE)
 208.302    224.8995    301.2776
     median          uq         max neval
 324930.956 559629.7600 1490447.354   100
    254.335    362.0175     741.082   100
```

It looks like qgraph is so slow, which may be why it is taking up too much memory on hipergator (computer cluster)s

Ok, lets compare how fast it takes to estimate centrality. One caveat is that centrality() in qgraph does 8 calculations, while eigen_centrality() in igraph calculates 1.

```
microbenchmark(centrality(network),eigen_centrality(net))
Unit: microseconds
                  expr      min        lq      mean   median       uq       max
neval
   centrality(network) 2636.858 2861.3810 3512.7286 3103.009 3404.803 35386.829
  100
 eigen_centrality(net)  384.781  411.1595  491.9623  446.119  540.057   851.331
  100
```

Page 92: 2018-11-09. git version control on hipergator (computer cluster)

I'm working on the UF hipgator computer cluster and doing all of this stuff without version control and without annotating any of the files.

So I need to learn git from the command line and I'll be following this tutorial : https://git-scm.com/book/en/v2/Getting-Started-First-Time-Git-Setup

I should also be annotating my project. File layout so far when I log in:

```
[andrew.nguyen@login4 ~]$ pwd
/home/andrew.nguyen

[andrew.nguyen@login4 ~]$ls
R  ufrc
```

One of the mistakes I made initially was that I read/write data on my home/andrew.nguyen directory which is not somethign I should do because it is inefficient to do here. Something about parallel processing not supported on home directory, but it is on the ufrc/ directory. So the help desk helped me set up a directory (symbolic link?) right on my home.

```
[andrew.nguyen@login1 ~]$ ln -s /ufrc/dhahn/andrew.nguyen/ ufrc
```

Ok, so I should be running scripts on /ufrc/dhahn/andrew.nguyen/

Ok, what is in ufrc?

```
[andrew.nguyen@login4 ~]$ cd ufrc
[andrew.nguyen@login4 ufrc]$ ls
Cerasi_Networks
```

Cerasi_Networks is the main transcriptome project I've been working on. So I want to basically construct networks for each population (2) and each time point (4) and estimate structural and node properties.

Layout of Cerasi_Networks

```
ls
Data  Script
```

All of the files are in Script:

```
[andrew.nguyen@login4 Script]$ ls
03_2018-10-
31_Network_centrality_subsets_each_timepoint_population_igraph_cluster.R
04_cluster_script.sh  test_cluster
03_data_set_2018-10-31_wide_filtered_sig_genes.csv
older_scripts
```

older_scripts folder has old scripts from past runs and test_cluster has files from initial test scripts.

03_data_set_2018-10-31_wide_filtered_sig_genes.csv is the dataset of all the significant genes found with edgeR.

03_2018-10-31_Network_centrality_subsets_each_timepoint_population_igraph_cluster.R is the R script that operates on the dataset.

04_cluster_script.sh is the job script (use sbatch)

Initial git set up:

```
[andrew.nguyen@login4 Script]$ git config --global user.name "Andrew D. Nguyen"
[andrew.nguyen@login4 Script]$ git config --global user.email
"anbe642@gmail.com"
[andrew.nguyen@login4 Script]$ git config --global core.editor vim
[andrew.nguyen@login4 Script]$ git config --list
user.name=Andrew D. Nguyen
user.email=anbe642@gmail.com
core.editor=vim
```

Starting new git repo.

```
[andrew.nguyen@login4 Script]$ git init
Initialized empty Git repository in
/ufrc/dhahn/andrew.nguyen/Cerasi_Networks/Script/.git/
```

*This creates a new subdirectory named .git that contains all of your necessary repository files — a Git repository skeleton. At this point, nothing in your project is tracked yet. (See Git Internals for more information about exactly what files are contained in the .git directory you just created.)

If you want to start version-controlling existing files (as opposed to an empty directory), you should probably begin tracking those files and do an initial commit. You can accomplish that with a few git add commands that specify the files you want to track, followed by a git commit:*

```
[andrew.nguyen@login4 Script]$ git status
On branch master

No commits yet
```

```
Untracked files:
  (use "git add <file>..." to include in what will be committed)

    03_2018-10-
31_Network_centrality_subsets_each_timepoint_population_igraph_cluster.R
    03_data_set_2018-10-31_wide_filtered_sig_genes.csv
    04_cluster_script.sh
    my_job-27691268.out
    older_scripts/
    test_cluster/

nothing added to commit but untracked files present (use "git add" to track)
```

ok, let's git add

```
[andrew.nguyen@login4 Script]$ git status
On branch master

No commits yet

Changes to be committed:
  (use "git rm --cached <file>..." to unstage)

    new file:   03_2018-10-
31_Network_centrality_subsets_each_timepoint_population_igraph_cluster.R
    new file:   03_data_set_2018-10-31_wide_filtered_sig_genes.csv
    new file:   04_cluster_script.sh
    new file:   my_job-27691268.out
    new file:   older_scripts/03_2018-10-
31_Network_centrality_subsets_each_timepoint_population_igraph.R
    new file:   older_scripts/run.sh
    new file:   test_cluster/03_2018-06-28_Network_analyses.R
    new file:   test_cluster/03_2018-10-
31_Network_centrality_subsets_each_timepoint_population.R
    new file:   test_cluster/03_cluster_test.sh
    new file:   test_cluster/03_test.R
    new file:   test_cluster/2018-06-28_pop_diff_centrality.csv
    new file:   test_cluster/Rplots.pdf
```

git commit

```
[andrew.nguyen@login4 Script]$ git commit
[master (root-commit) 8453a1c] First initial commit. Adding mainly the folders
in the script folder. It has older scripts and new ones
 12 files changed, 19114 insertions(+)
 create mode 100644 03_2018-10-
31_Network_centrality_subsets_each_timepoint_population_igraph_cluster.R
 create mode 100644 03_data_set_2018-10-31_wide_filtered_sig_genes.csv
 create mode 100755 04_cluster_script.sh
 create mode 100644 my_job-27691268.out
 create mode 100644 older_scripts/03_2018-10-
31_Network_centrality_subsets_each_timepoint_population_igraph.R
 create mode 100755 older_scripts/run.sh
```

```
 create mode 100644 test_cluster/03_2018-06-28_Network_analyses.R
 create mode 100644 test_cluster/03_2018-10-
31_Network_centrality_subsets_each_timepoint_population.R
 create mode 100755 test_cluster/03_cluster_test.sh
 create mode 100644 test_cluster/03_test.R
 create mode 100644 test_cluster/2018-06-28_pop_diff_centrality.csv
 create mode 100644 test_cluster/Rplots.pdf
```

Now I can just mess with one script and play with settings instead of making duplicates of it.

Page 93: 2018-11-27. Montanucci et al. 2011, MBE

Title: Molecular Evolution and Network-level analysis of the N-Glycosylation Metabolic Pathway Across Primates

Background:

N-glycosylation is decorating a nitrogen atom (amide nitrogen on asparagine) with a glycan or sugars. Mostly eukaryotes and archae do this and some bacteria.

To attach glycans (oligosaccharides) to the protein:

- glycosidic bonds bring sugar moieties together. The bonds are fromed between carbons 1 and 4 of the sugar molecules. It reuires ATP hydrolysis

Big picture ideas:

How pathogens recognize and invade our body is through our tissue and cellular structures. Glyco-proteins are diverse set of structural components of the cell that pathogens can recognize. These authors were interested in whether the pathway of how glycoproteins are made has diversified and evolved by natural selection within primates.

The n-glycosylation pathway: They class them

1. Substrate Donor
2. Precursor biosynthesis
3. Attachment of precursor to peptide
4. Quality Control
5. Glycan Extension

It looks linear for the most part. Glycan extension looks like the most integrated network. There are so different gene sets for each class

Question:

What shapes the evolutionary trajectory of the pathway producing glycoproteins?

Hypotheses:

Length, amino acid position, how genes interact (network) may limit or enhance evolutionary rates.

Or

Overall, individual protein properties and/or the position of the protein within the functional network may limit or enhance evolutionary rates.

Experimental Approach:

They picked a tractable network: N-glycosylation pathway. They measured:

Gene list of 52 genes for 4 primates : chimps, gorilla, orangutan, and macaque

1. evolutionary rates, omega, dn, ds
    - omega = dn/ds
    - dn = nonsynonmous amino acid substitution, - functional change in amino acid sequence
    - ds = synonymous amino acid substitution - null change

They use PAML. So they test an evolving model with a neutral model with LRT.

2. Created a functional network and measured centrality: THey got the graph from KEGG pathway
    - betweenness -fraaction of shortest paths that pass trhough that node
    - closeness - reciprocal of average distance to all nodes
    - degree - number of links/ fraction of nodes it is connected to
3. Protein properties
    - Codons used
    - Codon bias (effective number of codons) = ENC
    - Length

They used path analysis to try to determine the direct and indirect effects of protein properties and network properties on evolutionary rates.

Results:

1. Table 1. No signature of positive selection. In fact, omega was below 0, evidence for purifying selection
2. Table 2 and 3. Comparing omega for each functinal class : No sig differences among funcional classes for ds
    - Tehy evaluated different components of evo rates
        - omega was diff between precursor and oST /Glycan extension
3. Strength of purifying selection and pathway structure- Figure 2 and table 4 and figure 3
    - Negative correlation between pathway position and omega. UPstream genes are under relaxed selection than downstream ones. Driven by Dn and not ds. Why?
    - Negative correlation between degree/closeness and omega/dn.
    - mantel test of pairwise distance between genes sig for dn but not ds. Neighboring genes share similar evolutionary constraints
4. Path analysis

Conclusion:

Take home: Where genes are in the network and how many neighbors they have decreases the evolutionary rate of that gene.

Issues: * The analysis assumes perfect knowledge of the protein fucntional network. * Biased taxanomic sampling : what if they included more species?

Big idea: Evolutionary dispensibility of a gene. extent of given protein can have amino acid changes .

Where you are on the network can constrain you evolutionarily.

Page 94: 2018-11-30. Basic WGNCA code

```
### Load data
dat.wide<-fread("../Data/CerasiCountsIsos/03_data_set_2018-10-
31_wide_filtered_sig_genes.csv")
glimpse(dat.wide)

### WGNCA

adjacency = adjacency(t(dat.wide[,4:7]), power =1); #adjacency matrix
#topologicla overlap matrix
TOM = TOMsimilarity(adjacency);
dissTOM = 1-TOM
geneTree = hclust(as.dist(dissTOM), method = "average");

#identifying modules using dynamic tree cut
dynamicMods = cutreeDynamic(dendro = geneTree, distM = dissTOM,
                            deepSplit = 2, pamRespectsDendro = FALSE,
                            minClusterSize = 37);

dynamicColors = labels2colors(dynamicMods)
table(dynamicColors)
#calculate eigengenes and then cluster modules eigengenes that are similar
MEList = moduleEigengenes(t(dat.wide[,4:7]), colors = dynamicColors)
MEs = MEList$eigengenes
MEs=MEs[,-6]

MEDiss = 1-cor(MEs);
METree = hclust(as.dist(MEDiss), method = "average");
#sizeGrWindow(7, 6)
#plot(METree, main = "Clustering of module eigengenes",
 #    xlab = "", sub = "")
# merge modules?
merge = mergeCloseModules(t(dat.wide[,4:7]), dynamicColors, cutHeight =.25,
verbose = 3)
mergedColors = merge$colors;
#mergedMEs = merge$newMEs;
```

I want to construct a network for each time point in a time course experiment. Then compare the modules across all time points. One way to get at robustness of networks or how they can change over time is to determine which genes are consistently modular or correlated in their expression and ones that switch between modules.

Some code to determine how gene sets overlap across modules :

```
x<-c("a","b","c")
y<-c("e","f","c")

z<-c("a","f","c")
b<-c("q","f","c")
c<-c("p","f","c")

intersectSeveral <- function(...) { Reduce(intersect, list(...)) }
intersectSeveral(x,y,z,b,c)
```

This will identify genes shared among modules.

Page 95: 2018-12-06. Meeting with Dhahn

job application for UF

- differentiate myself from Dan
- teach2 classes: insect ecology and something else I wanted *

Tom's paper:

- dan focus on intro
- andrew focus on results and execute on 2 items
    - remake box plots and make them prettier
    - straight up correlation matrix

Cerasi data analysis :

- Detect modules for each population and each time point.
- A priori: maintenance modules assemble and then disperse through time and termination modules do the opposite
- Dan wants to bootstrap the adjacency matrix to determine robustness.

reviewing differnt metrics of gxp

- FPKM - Fragments per kilobase of transcript per million mapped reads
- RPMK - reads per kilobase of transcript per million mapped reads

---

Page 96: 2018-12-07. Lab meeting reading : Johnson et al. 2018

Title: Network Architecture and Mutational Sensitivity of the C elegans metabolome

1. Background

construct networks in c elegans for 29 metabolites.

2. Big picture ideas of the paper

Gap in knowledge: We have a poor understanding of how mutation impacts networks and subsequent evolutionary trajectory

3. Questions

What is the influence of mutation and the accumulation of mutations on metabolite networks?

Are there or what positional properties of the metabolite network that impact fitness?

4. Hypotheses

Mutations in central genes negatively impact fitness.

Predict system level property negatively related to fitness

5. Experimental Approach

measures: centrality (in and out degree, closeness, betweenness), mutational heritability , shortest path length , Core number (k -core is the largest subgraph that contains nodes of at least degree K.), mutational bias,

System: They used mutation accumulation lines (43). Figure1 - variance will increase , mutation bais = slope of Fitness vs generations . Genetic variance will increase over time

Used canoical correlation analysis to find significant asosciations with mutational parameters and network parameters.

6. Key Results

The graph itself: 656 metabolites, 1203 reactions connecting metabolites.

Figure 2: example figure of K-core

Figure 4: shows a positive correlation to me (no line), but stats don't support it. Darn.

Figure5: randomization for correlations between Rm(genetic correlation between two traits in MA lines) and shortest path length in directed network; absolute value of Rm and shortest path network; Rm and and shortest path length for undirected network.

7. Conclusion and big picture

It looks like the total higher number of degrees can impact the heritability of mutation (short term response to selection on mutational variance)