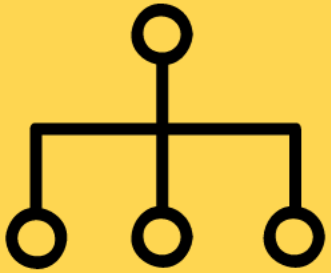


Applying principles of data science to breeding and genetics

Harly Durbin

Wednesday, February 24th

@harlyjaned

- Grew up near Fort Worth, TX
- Ag background
- Currently based in Columbia, MO, headed to Knoxville, TN
- Hobbies

- B.Sc., Animal Science 2016 – Texas A&M University
- Ph.D., Genetics 2020 – University of Missouri
- Genetics intern with the American Angus Association – largest beef cattle breeding cooperative in the world
- Currently – post-doctoral researcher, University of Missouri Division of Animal Sciences

Research at the intersection of quantitative genetics, population genetics, and data science

Projects in:

- Cattle & closely related wild species
- Ancient DNA
- Human populations
- Catfish

Using:

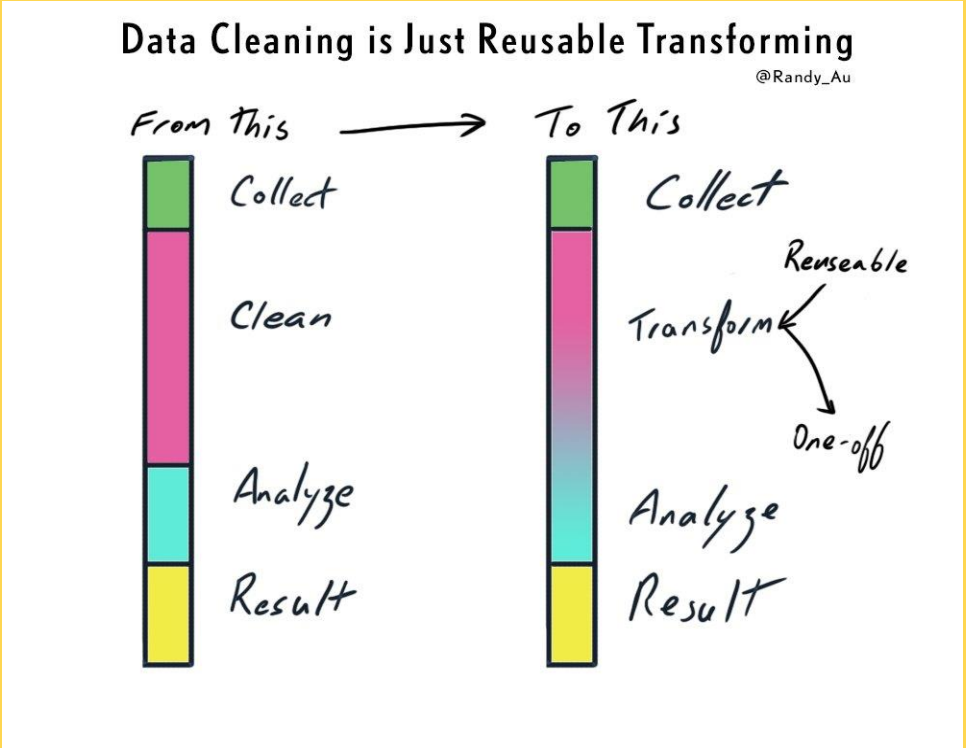
- SNP chip genotypes
- WGS data
- Farmer-sourced phenotypes
- Publicly-available environmental data

Data science philosophy

1. Data cleaning is part of data analysis
2. Domain-specific knowledge is essential
3. Reproducibility is more than leaving a paper trail

Data cleaning is part of analysis

- Can’t be completely automated and requires a degree of experience, intuition
- More than just “checking for errors”: data cleaning is an essential step in understanding the data
- Often drives hypothesis generation



Credit: Randy Au

Farmer-submitted “citizen science” data

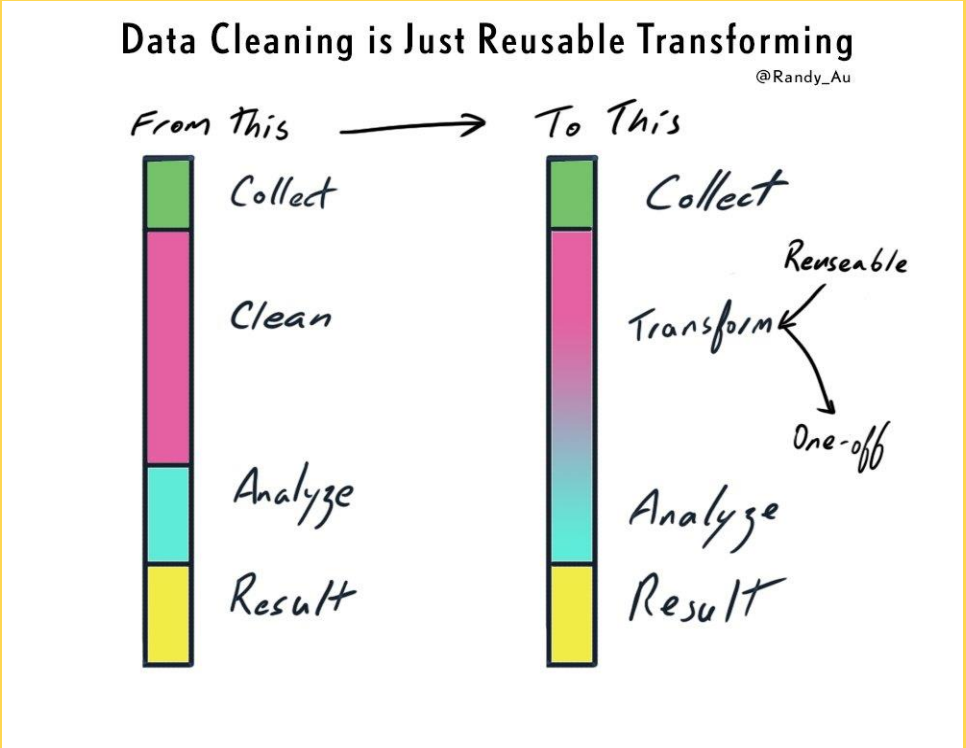
- **Goal:** Multi-year dataset with consistent IDs across years
- **Issues:** multiple raw file formats, registration with multiple breed associations, updated parentage information, movement of animals between herds, every possible human error or typo
- (Side note about hindsight)

Publicly available SRA meta-data

- **Goal:** Scrape as much species/subspecies, breed/ancestry, sex, geographic location metadata as possible for publicly available bovid WGS samples
- **Issues:** Many possible fields could be interpreted to mean the same thing, inconsistent coding of metadata

Data cleaning is part of analysis

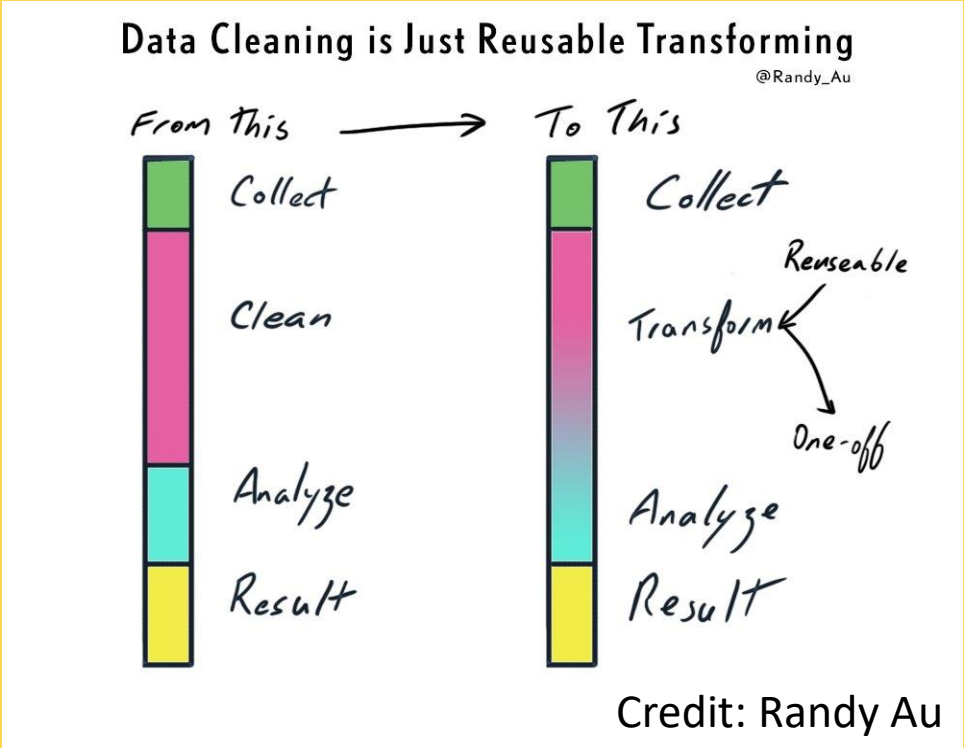
- Can’t be completely automated and requires a degree of experience, intuition
- More than just “checking for errors”: data cleaning is an essential step in understanding the data
- Often drives hypothesis generation



Credit: Randy Au

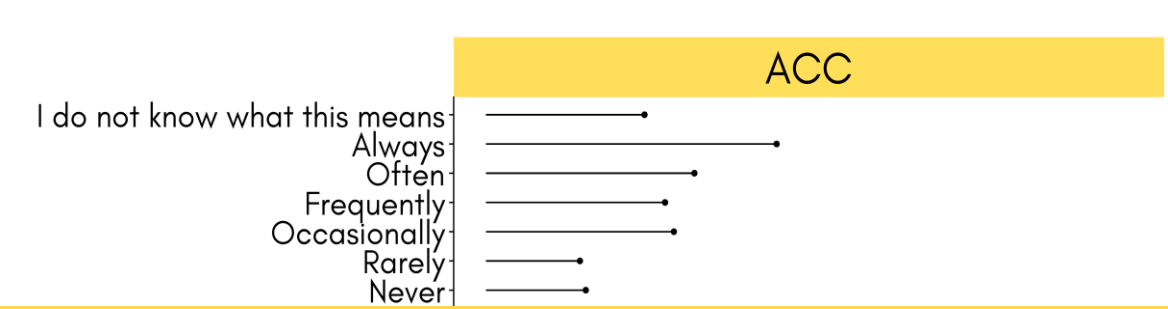
Data cleaning is part of analysis

- Can’t be completely automated and requires a degree of experience, intuition
- More than just “checking for errors”: data cleaning is an essential step in understanding the data
- Often drives hypothesis generation



Credit: Randy Au

Data cleaning drives hypothesis generation



Variable	*% Skipped question*
Accuracy of breeding value	11.5%
CEM	4.5%
REA	4.2%
Selection indexes	4.0%
CED	3.4%
MARB	1.7%
WW	1.1%
MILK	1.1%
BW	1.1%
YW	0.8%

- Written survey of beef producers' attitudes towards genetic technology
- During data cleaning, noticed disproportionate number of missing values for a particular question
- Changed conclusions, direction of future surveys & extension programming

Domain-specific knowledge is essential

Domain-specific knowledge is knowledge about the environment in which the data is processed and used

- Strong applied breeding & quantitative genetics background
- Knowing when to seek guidance

Reproducibility is more than leaving a paper trail

- Pipelined & flexible workflows enables easy integration of new data
- Easy integration of new data enables routine reporting
- Routine reporting shortens the path from analysis to actionable outcomes

Snakemake

Strategies for creating environment-aware predictions of genetic merit in animals

1. **Development of novel phenotypes**
2. Modification of existing methodologies

MU Thompson Research Farm

Troy Rowan

Oak Hollow Angus

Seasonal coat change & hair shedding scores

- Adaptive response to climactic variation, strongly tied to fitness in many species
- Conserved biological mechanism and timing across mammals
- In beef cattle: 5-point scale where 5 is 0% winter coat shed and 1 is 100% winter coat shed (Gray et al., 2011)

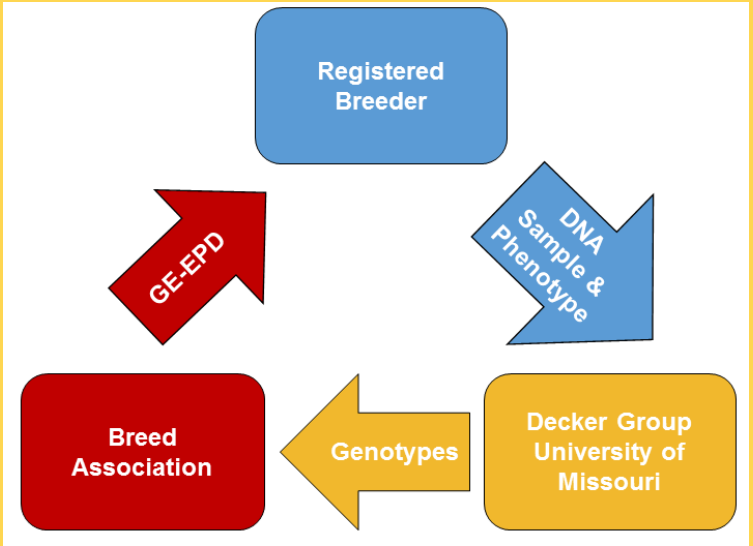
THE SIGNIFICANCE OF COAT TYPE IN CATTLE

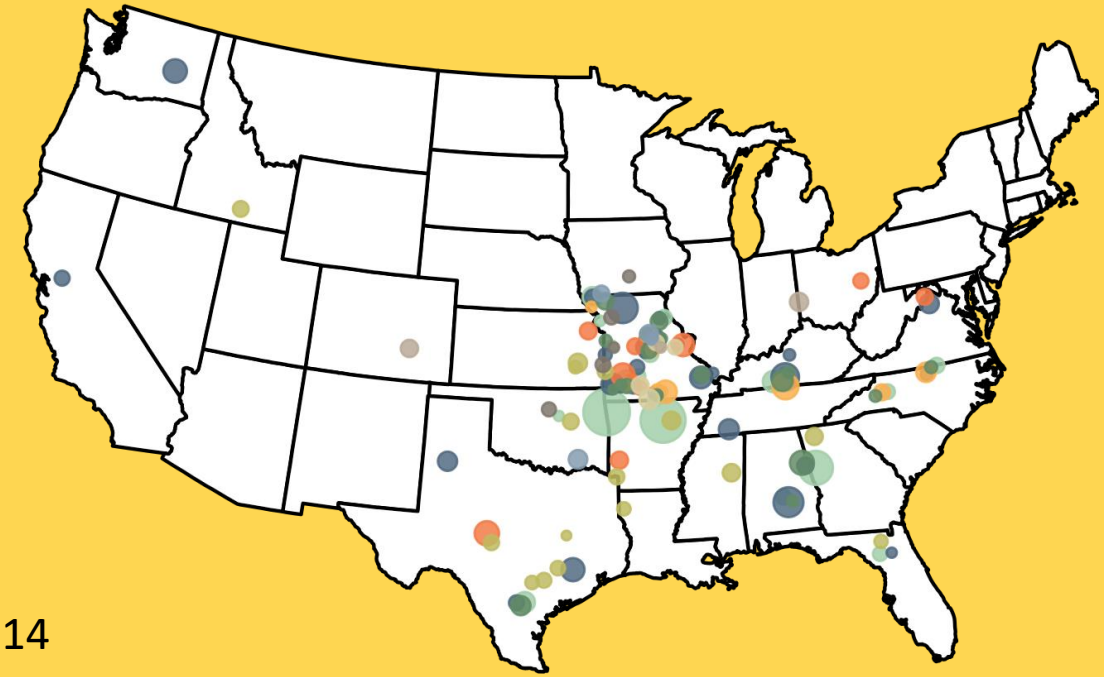
By H. G. Turner* and A. V. Schleger*

[*Manuscript received November* 16, 1959]

Mizzou Hair Shedding Project

- **36,899** total hair shedding scores on **13,364** cattle between 2016-2020
  - **77** breeders across the U.S.
- Large collaboration: data shared with **9** major beef cattle breed associations





American Gelbvieh Association

American Hereford Association

American Shorthorn Association
Performance with Purpose
est. 1872

SimGenetics
PROFIT THROUGH SCIENCE
American Simmental Association

ANGUS GENETICS

RED ANGUS
RANCH TESTED. RANCHER TRUSTED.

INTERNATIONAL BRANGUS
BREEDERS ASSOCIATION

CHAROLAIS

Hair shedding research breeding value at AAA

- Moderately heritable ($h^2 = 0.42$) and correlated with other economically relevant traits
- Worked with information systems, research, marketing teams to develop and launch
- Deployed to AAA February 2020
  - In production: breeders submitting their own data


Oak Hollow Angus


Oak Hollow Angus

ANGUS
THE BUSINESS BREED

Research Hair Shedding EPD Launched by Angus Genetics Inc.

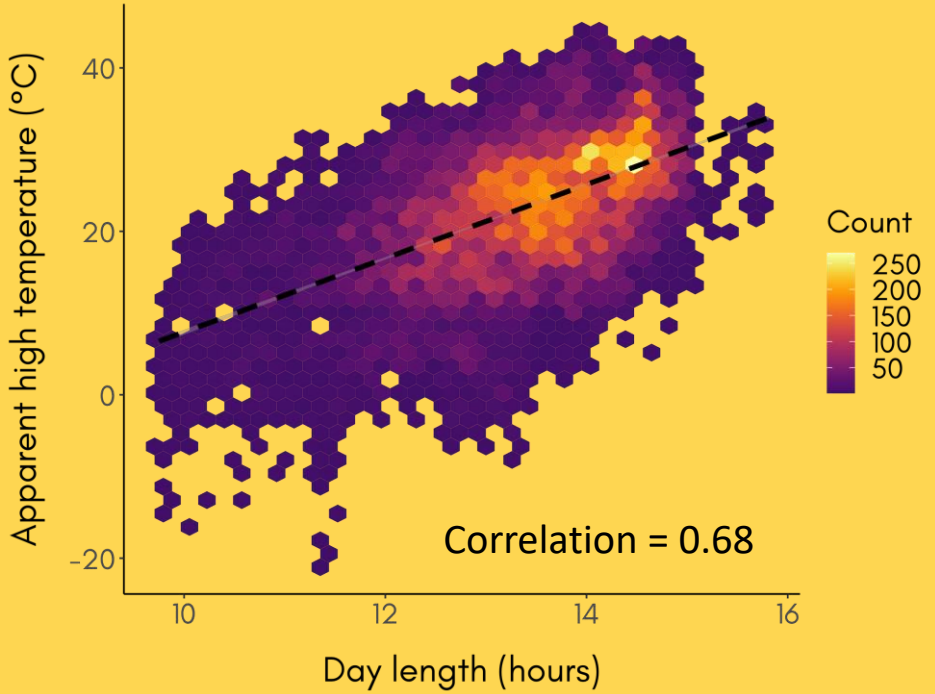
Development of a genetic evaluation for hair shedding in American Angus cattle to improve thermotolerance

Harly J. Durbin, Duc Lu, Helen Yampara-Iquise, Stephen P. Miller & Jared E. Decker

Genetics Selection Evolution **52**, Article number: 63 (2020) | Cite this article

Quantifying the effects of day length and temperature

- Infer geographic coordinates using address
- Gather day length and high apparent temperature for 30 days prior to score date at each location → average each
- Estimate effect of day length and/or temperature in 4 models, compare fit



Hair shedding score

$$y = \text{fixed effects} + \beta_T + \beta_{DL} + Z_1u + Z_2pe + e$$

Random residual

Year, calving season, age group, toxic fescue grazing status

Effect of temperature

Effect of day length

Genetic effect (breeding value)

Permanent environment effect (repeated records)

Plus: 1) DL only, 2) T only, 3) DL + T + DL*T

Quantifying the effects of day length and temperature

- **0.446** decrease in hair shedding score with each hour increase in average **day length** over past 30 days
- **0.072** decrease in hair shedding score with each 1°C decrease in average apparent high **temperature** over past 30 days
- **0.006** decrease in hair shedding score with each one unit increase in hours day length*°C standardized interaction units over past 30 days

Strategies for creating environment-aware predictions of genetic merit in animals

1. Development of novel phenotypes
2. **Modification of existing methodologies**

18
MU Thompson Research Farm

Troy Rowan

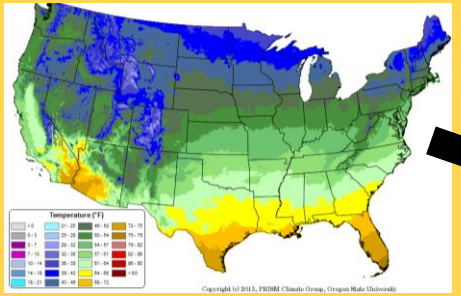
Oak Hollow Angus

Predicting genetic merit for weaning weaning weight

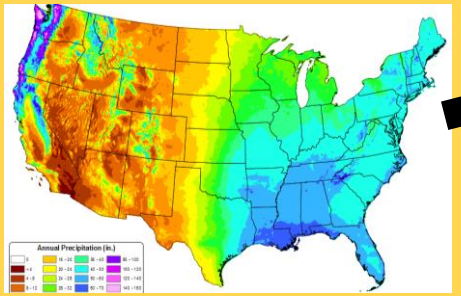
$$y = Xb + Z_1u + Z_2m + Z_3p + e$$

- y = weaning weight at ~7 months of age adjusted to 205 days
- b = fixed contemporary group effect (environment)
- u = random genetic effect of calf (direct)
- m = random genetic effect of dam (maternal)
- p = random maternal permanent environment effect

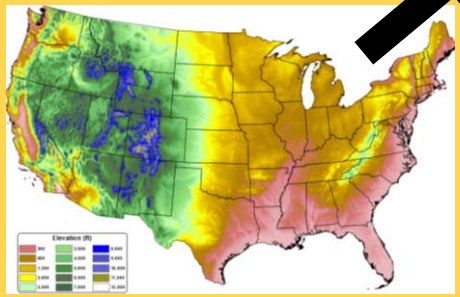
Continuous climate variables → K-means clustering → Discrete environmental regions



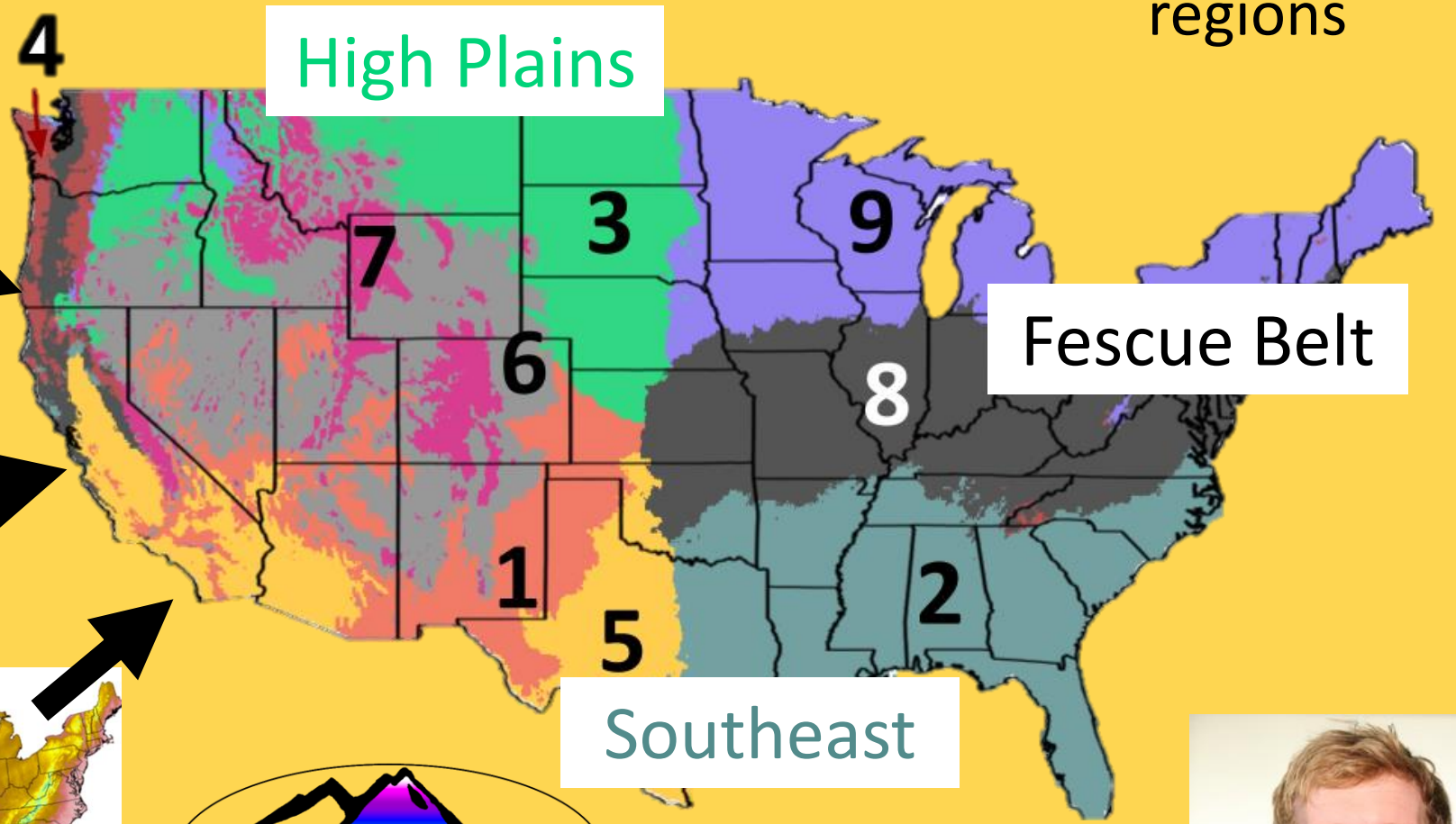
Temperature



Precipitation



Elevation



PRISM GROUP

Troy Rowan

Estimating the effect of environment: contemporary groups

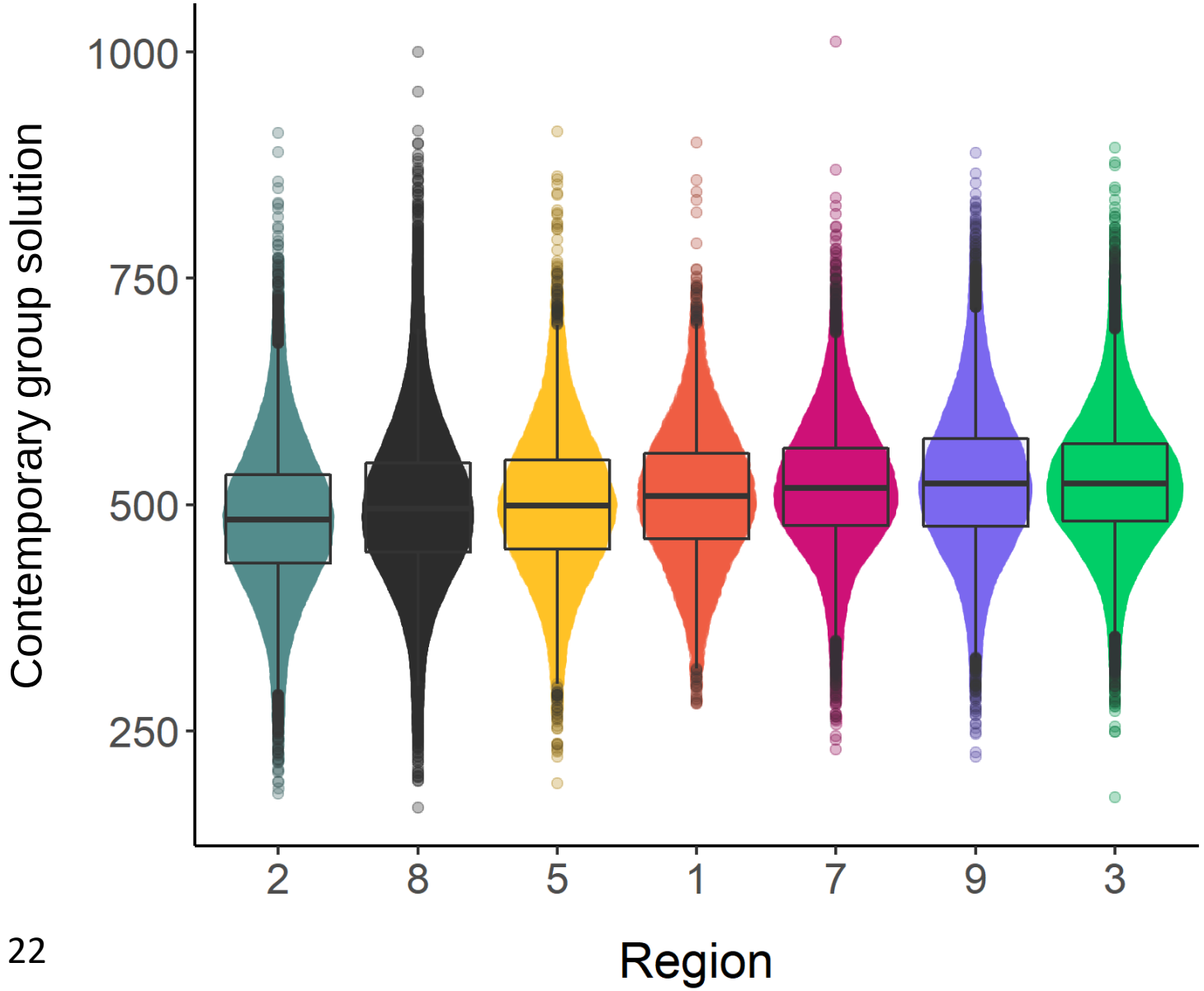
Set of animals who had equal opportunity to perform:

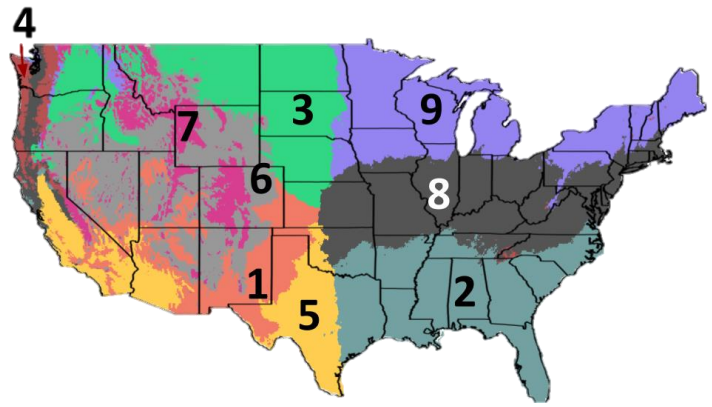
- Managed alike at same farm
- Same sex, close in age
- Exposed to the same environmental conditions & feed resources

$$y = X\mathbf{b} + Z_1u + Z_2m + Z_3p + e$$

Contemporary group solution (BLUE)

Weaning weight contemporary group solutions (1990-present)





- 2: Southeast
- 8: Fescue Belt
- 5: Arid Prairie
- 1: Desert
- 7: Forested Mountains
- 9: Upper Midwest & Northeast
- 3: High Plains

$P = G + E + G^* E$

Estimating the extent of GxE interaction

Random regression (reaction norm) models

- Genotypes regressed on continuous variable (i.e., temperature, disease incidence, etc.)
  - Each genotype has a unique curve
- Difficult to capture stressors that aren't explicitly measured

Multivariate models

- Observations on the same variable made under different conditions treated as separate **(potentially correlated)** traits
- Genetic correlation (r_g) interpreted as degree of re-ranking across environments
  - 0.8 threshold (Falconer, 1952)
- **May be better for capturing local stressors & interactions**

$P = G + E + G*E$

Possible GxE in weaning weight maternal

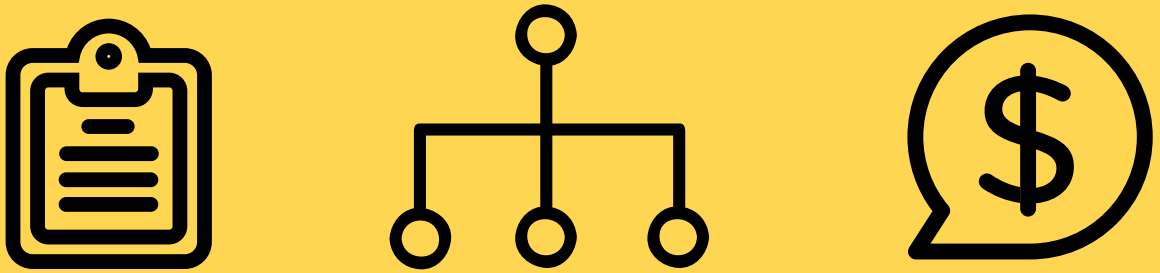
Region	N records	Min. r_M	Mean r_M	Max. r_M
1: Desert	165,057	0.80	0.86	0.95
2: Southeast	**508,565**	**0.67**	**0.77**	**0.89**
3: High Plains	2,075,979	-	-	-
5: Arid Prairie	208,689	0.78	0.86	0.91
7: Forested Mountains	696,033	0.78	0.85	0.93
8: Fescue Belt	**1,462,959**	**0.66**	**0.82**	**0.95**
9: Upper Midwest & Northeast	600,051	0.72	0.84	0.95

$P = G + E + G*E$

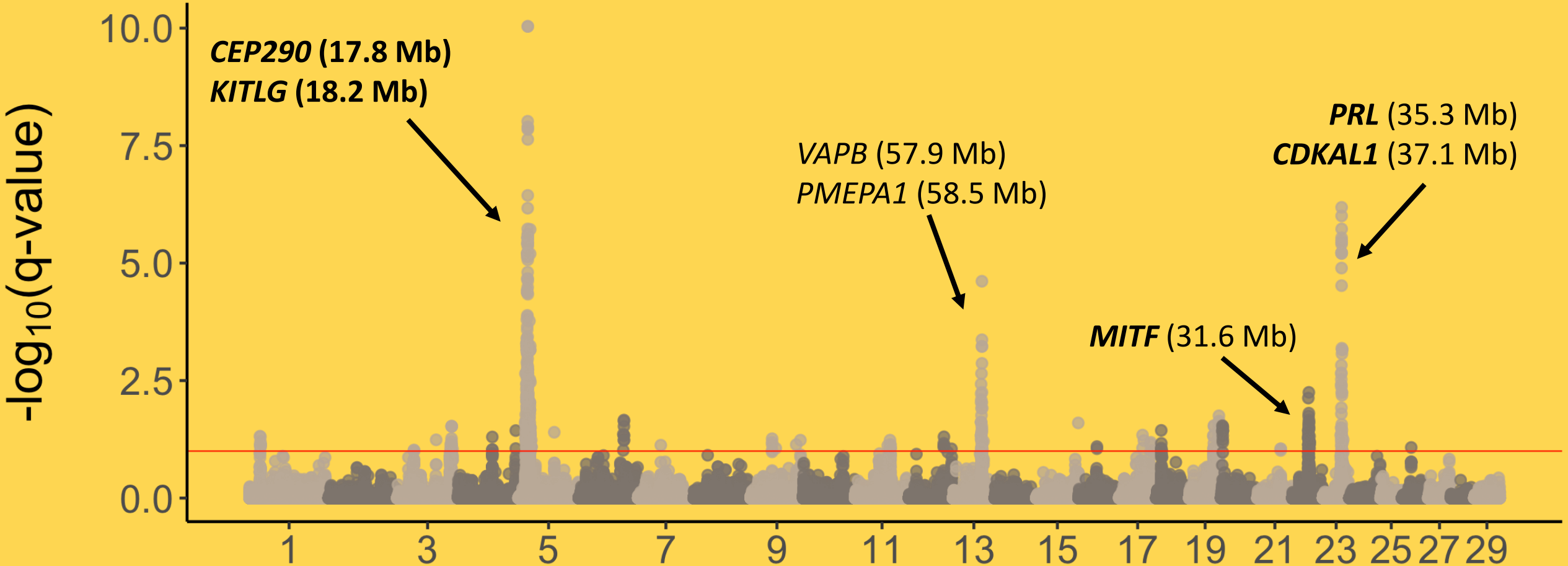
What excites me about this role?

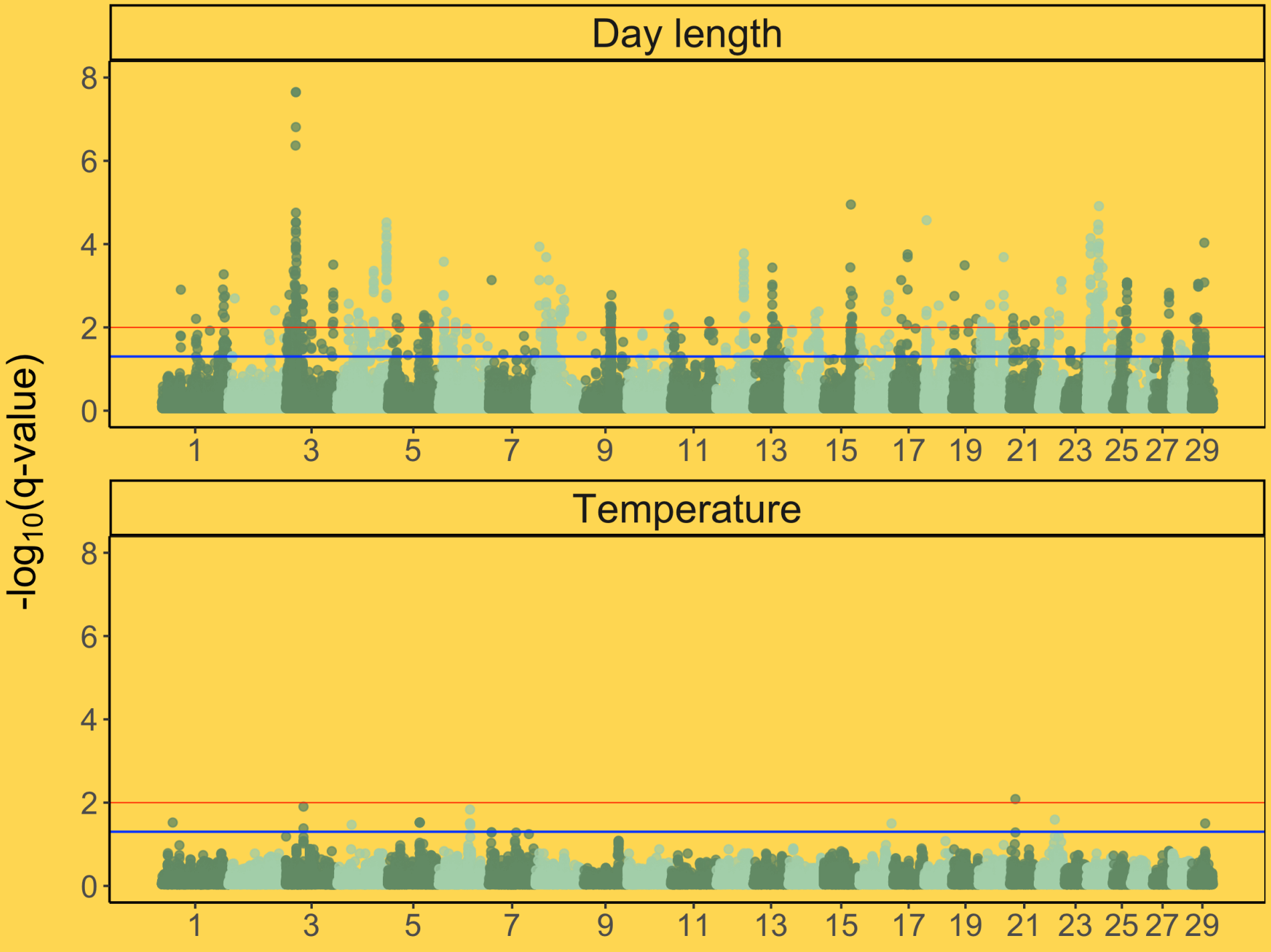
- Opportunity to work with multiple teams to “shorten the path” from analysis to actionable outcomes
- Data consolidation and organization – creating connectivity between orphaned data
  - Looking at existing data with new and novel perspectives
- More driven by piecing together the intermediate steps than the final answer

Questions?



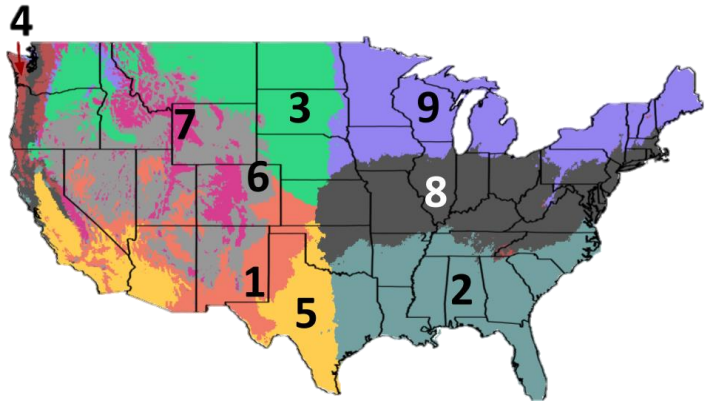
- Pseudo-phenotypes in GWAS were EBVs de-regressed with Garrick et al., 2009 method & parent average removed then weighted by 1/reliability
  - Only animals with known parents included (n = 9,865)

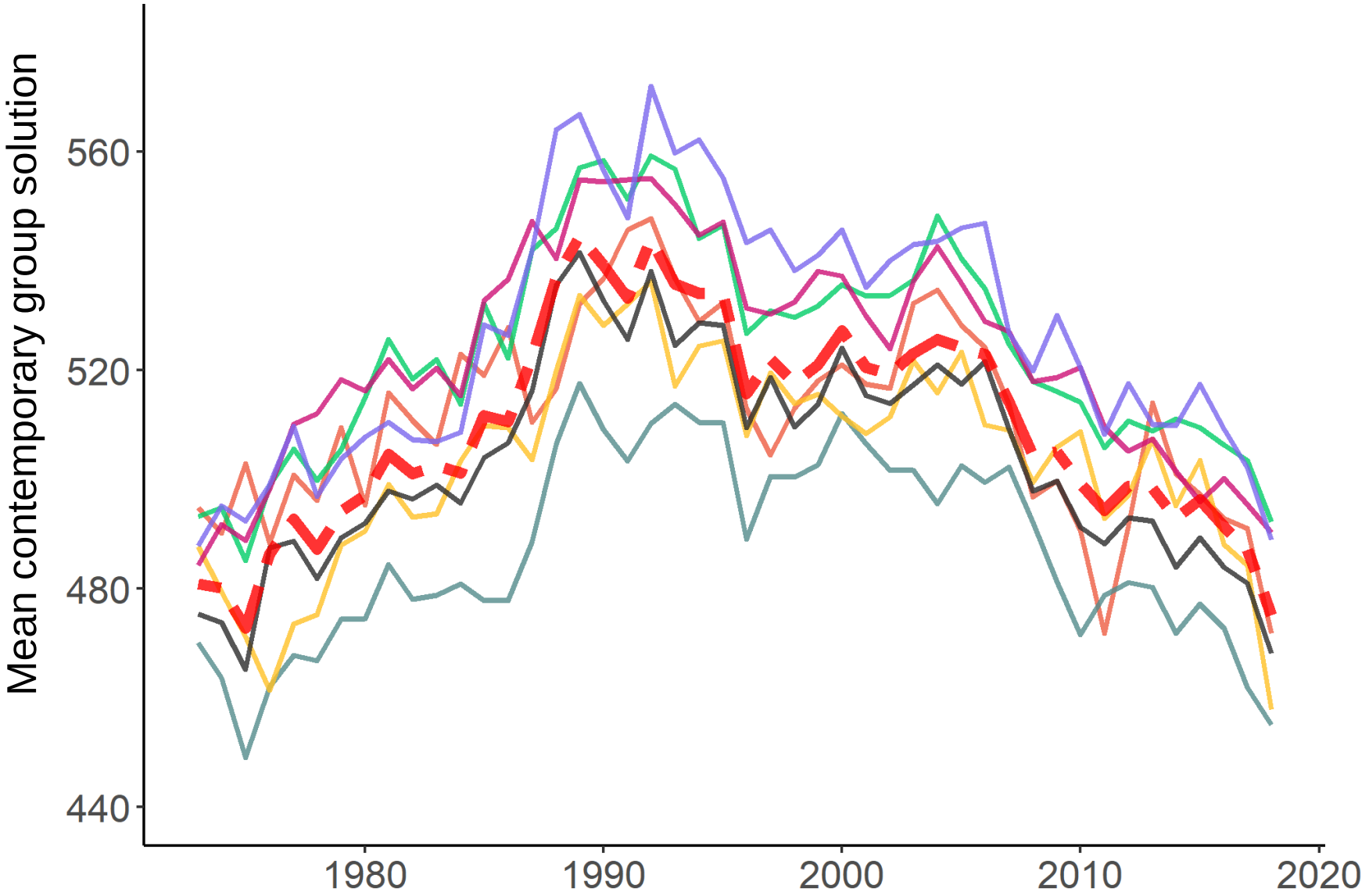




- GxE GWAS of each variable (day length or temperature) within each year (2016-2019)
- Meta-analysis of results

Weaning weight contemporary group solutions reflect year-to-year environmental trends





- 1: Desert
- 2: Southeast
- 3: High Plains
- 5: Arid Prairie
- 7: Forested Mountains
- 8: Fescue Belt
- 9: Upper Midwest & Northeast
- National average

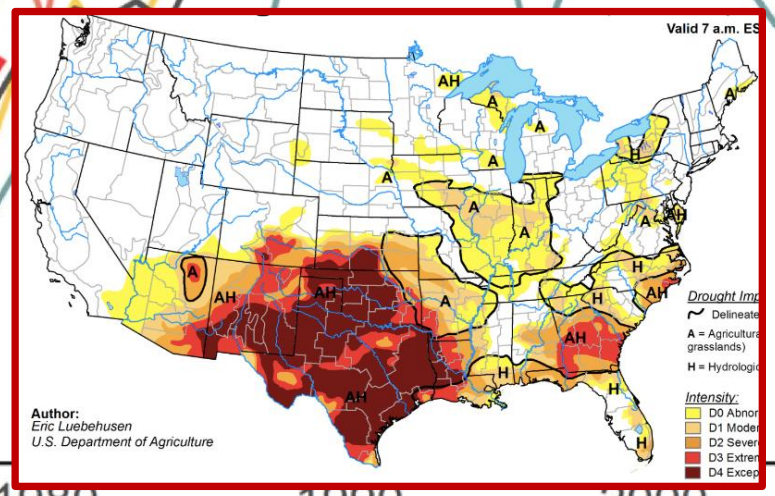
$P = E + G + G*E$

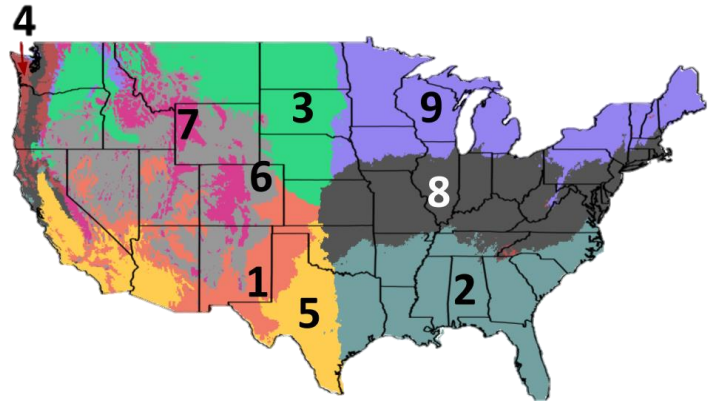
Weaning weight contemporary group solutions reflect year-to-year environmental trends

Mean contemporary group solution

560
520
480
440

1980
1990
2000
2010
2020

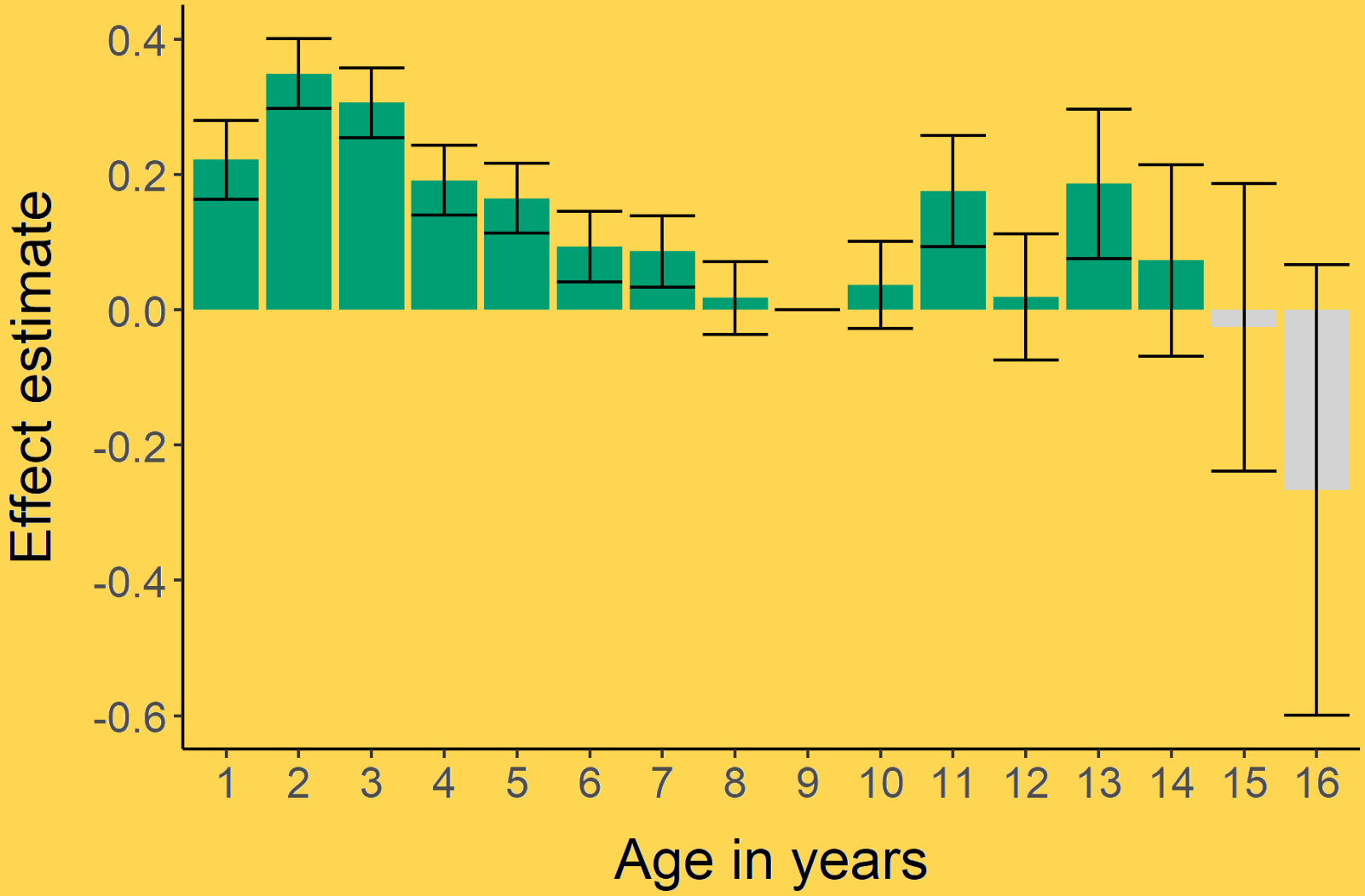




1: Desert
2: Southeast
3: High Plains
5: Arid Prairie
7: Forested Mountains
8: Fescue Belt
9: Upper Midwest & Northeast
National average

$P = E + G + G*E$

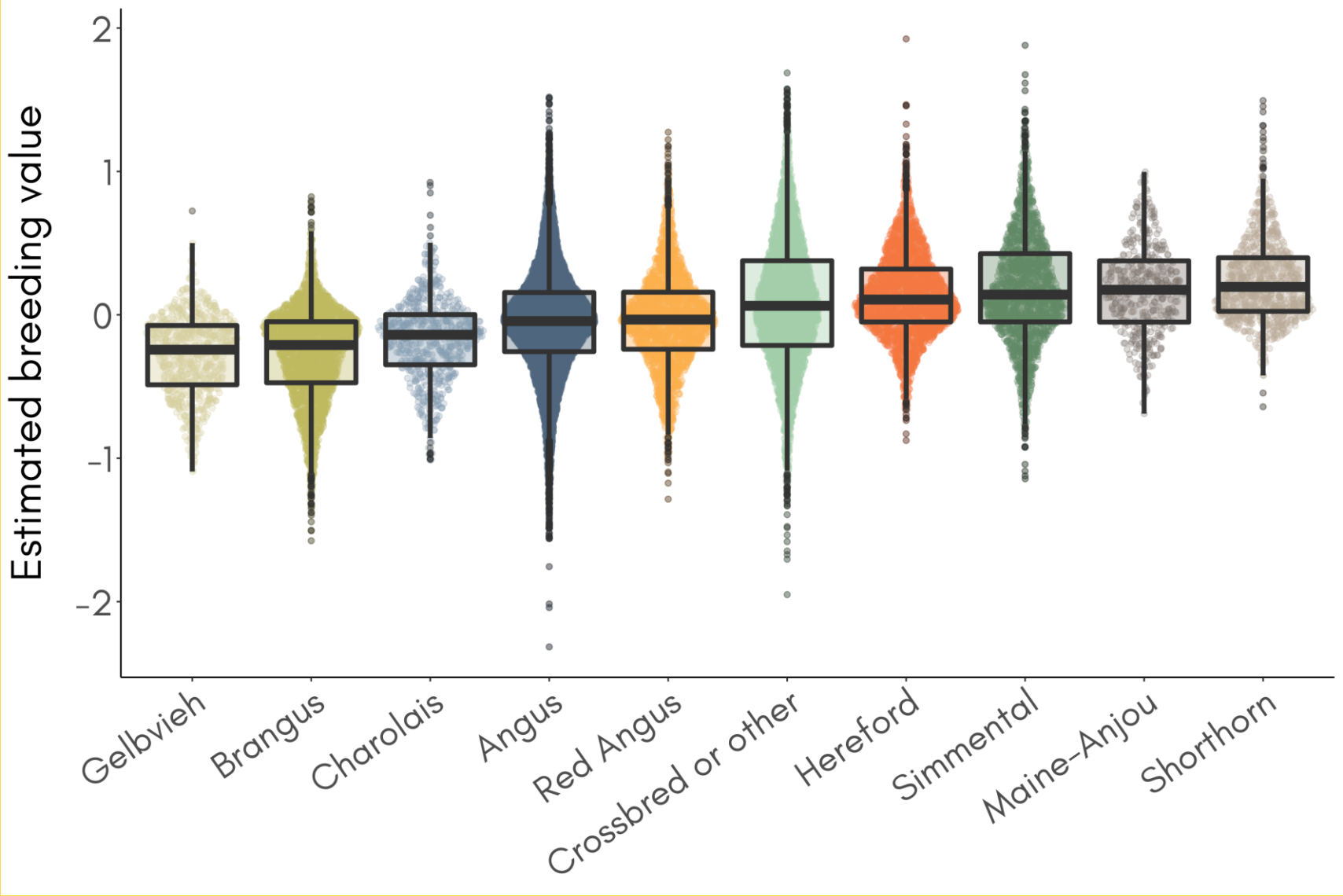
The effect of age on hair shedding score is “U-shaped”



- Youngest and oldest animals tend to shed later
  - Stress of growth
  - Senescence
- Old cows likely represent selected sample (Schons et al., 1985)
- Similar patterns observed in other species (Déry et al., 2019)

AGI data

- Variation largely overlaps between breeds
- Breeds recently selected for the show ring (Shorthorn, Maine-Anjou) tend to have less desirable EBVs
- Breeds with *Bos indicus* influence (Brangus, Charolais) tend to have more desirable EBVs
- Gelbvieh?



Mizzou data

Hair shedding score is moderately heritable

Dataset	N scores	N animals	Avg. scores per animal	h^2	r
AGI	14,465	8,642	1.67	0.40	0.44
Full Mizzou	36,899	13,364	2.76	0.37	0.45
Angus Mizzou	8,674	3,953	2.19	0.37	0.42
Brangus Mizzou	1,829	984	1.92	0.40	0.40
Hereford Mizzou	2,857	1,235	2.31	0.32	0.40
IGS breeds Mizzou	10,996	4,713	2.33	0.41	0.48

- Turner & Schleger (1960) h^2 using 7-point scoring system: 0.63
- Gray et al. (2011) h^2 using same scoring system but pedigree only: 0.35

Repeated records animal model

Hair shedding score

$$\mathbf{y = contemporary\ group + Z_1u + Z_2pe + e}$$

Random residual

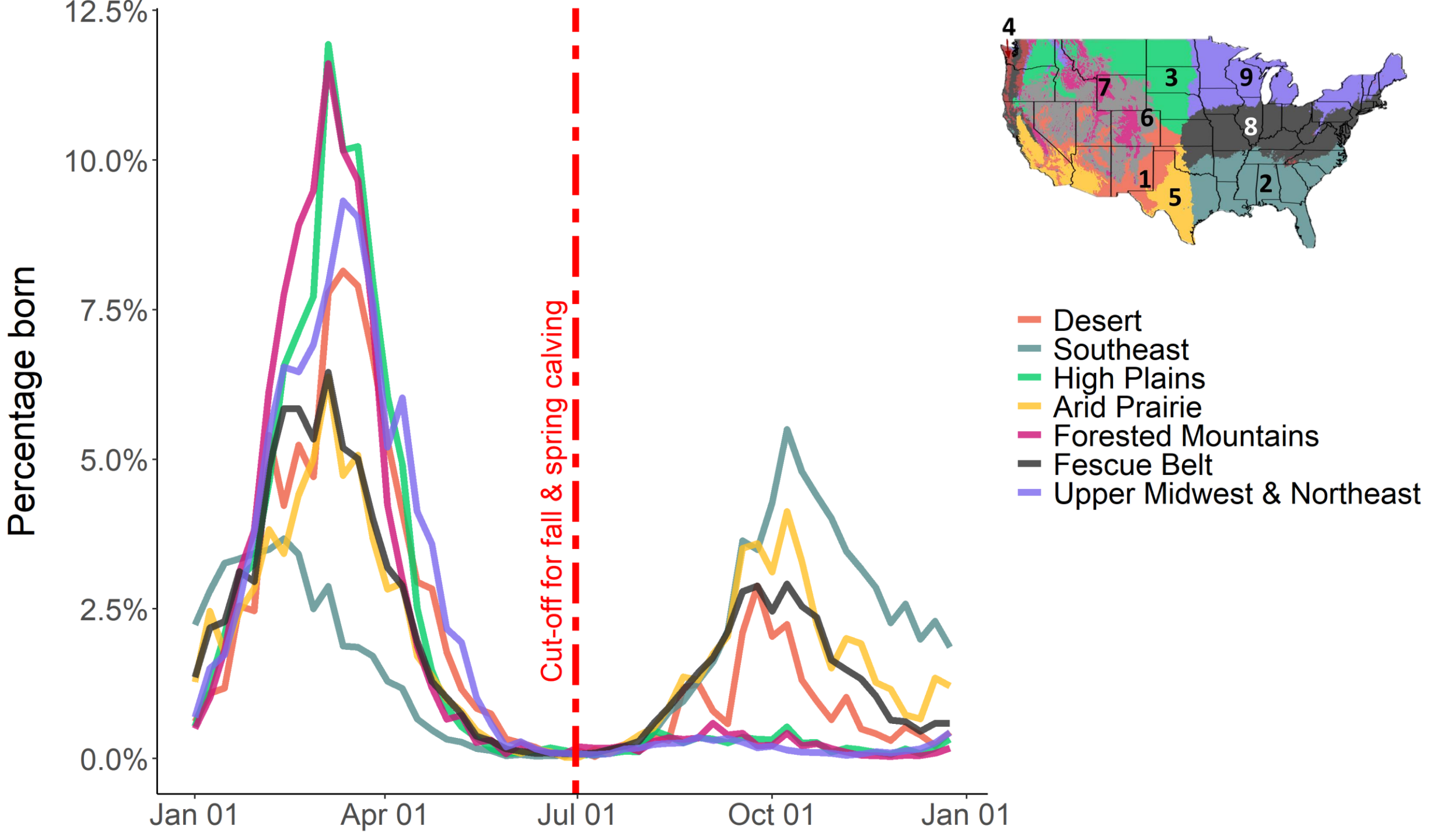
Combination of year, farm calving season, age group, toxic fescue grazing status, "score group" within farm

Additive genetic effect (breeding value)

Permanent environment effect (repeated records)

Single-step BLUP: represent relationships between animals in relationship matrix H^{-1} (Aguilar et al., 2011) which "blends" pedigree relatedness and genomic relatedness information

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{H}\sigma^2_a);\ \mathbf{pe} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{pe});\ \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_e)$$


Percentage born
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
Jan 01
Apr 01
Jul 01
Oct 01
Jan 01
Cut-off for fall & spring calving
4
7
3
9
6
8
1
5
2
Desert
Southeast
High Plains
Arid Prairie
Forested Mountains
Fescue Belt
Upper Midwest & Northeast