Examples of data analysis projects

David Jansen

Archie Lab
Biological Sciences

david.awam.jansen@gmail.com

January 19th 2023

Background

- Research scientist in the Archie Lab at Notre Dame
  - Large coding projects for the ABRP
  - Manage large volumes of data and samples
  - Independent research
  - Help graduate students with analysis
  - 'Quick' coding projects

Background

- Amboseli Baboon Research Project
  - Long-term (40+ years) field site in Kenya
  - Studies wild habituated baboons
  - Data in a very large relational database
  - Collaboration between multiple universities

Type of coding and analysis

- Mainly R
  - Tidyverse
  - Rmarkdown
  - Writing functions
  - Parallel coding for cluster computing
- PostgreSQL
- LaTeX
- Some bash
- (Generalized) Linear mixed models
- Survival analysis
- Ordination

Example of some projects

1. **Elo ranking :** Develop an automated system of dominance rank assignment

2. **Large microbiome project :** (1) Coordinate project from start to finish, (2) Create database structure, and (3) Help with analysis

3. **Sociality indexes:** (1) Recreate code in R for a method described in a paper, (2) Create a version for juveniles, and (3) find unusual patterns.

Elo ranking

Elo ranking

- Linear dominance hierarchies structure in baboons
- Based on agonistic interactions
- Currently rank assignment is done manual through matrix optimization

Elo ranking

- Elo rating system is a method for calculating the relative skill levels of players in zero-sum games such as chess
  - Increasingly used in animal behavior
  - Can mostly be automated
  - Current R packages did not fit project data and needs
- Created new code from scratch
- Tested how well the Elo-based compared to the matrix-based
  - Rerun models from papers that have rank-related measures
  - Visualize some traits that are known to be correlated with rank.

▸ GitHub code for Elo based ordinal dominance rank

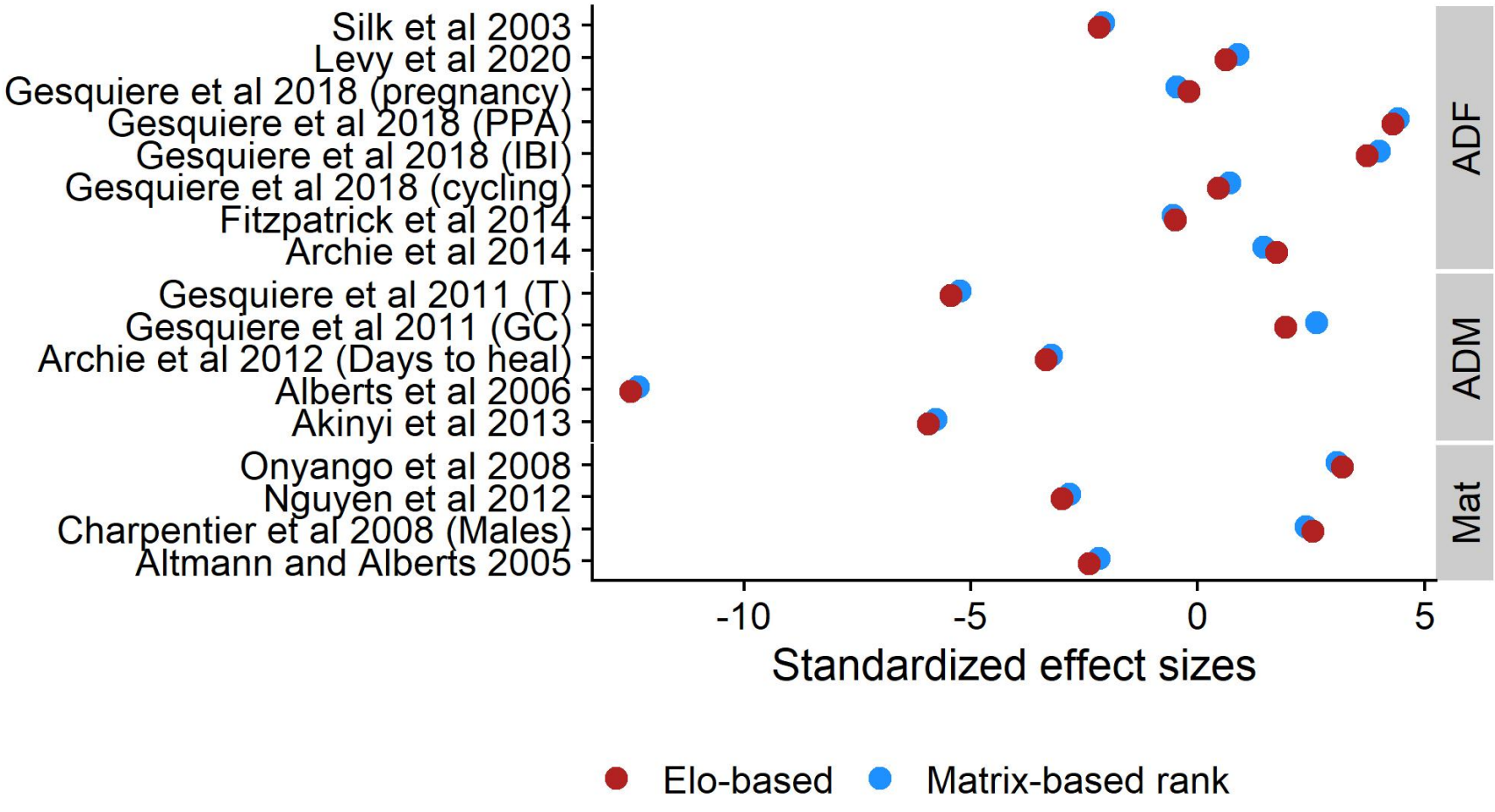
Elo ranking - compare models

```
model_compare_setup <-
  tibble(study_nr = c(1, 2, 2, .... 14),
         study =  c("Alberts et al 2006"
  mutate(Modfor = c(
    c("lm(formula = prophrs ~ rank_measure, data = df)")
  , c('glmmTMB(formula = logT~ season+temperature+age+rank_measure +(1|sname),
                data=df, family="gaussian")'))) %>%
   mutate(filename = c(c("./data/Alberts_et_al2006.csv")
                       , c("./data/Gesquiere_et_al_2011_T.csv"))) %>%
    mutate(dataset = map(filename,  ~ read.csv(file = .))) %>%
    mutate(models = map2(.x = ModFor, .y = dataset, .f = run_models)) %>%
    unnest(cols = c(models))

run_models <- function(model_formula, dataset) {
  df = dataset
  tibble(rank_measures = c("1", "adult_rank",
                           "overall_rank", "elo_rank", "elo_adult_rank")) %>%
    mutate(ModFor = str_replace(string = model_formula,
                                pattern = "rank_measure",
                                replacement = rank_measures))  %>%
    mutate(model_formula = str_replace_all(ModFor, "I\\(1\\^2\\)", "1")) %>%
    mutate(model = map(.x = model_formula, ~ eval(rlang::parse_expr(.x)))) %>%
    select(rank_measures, model)}
```

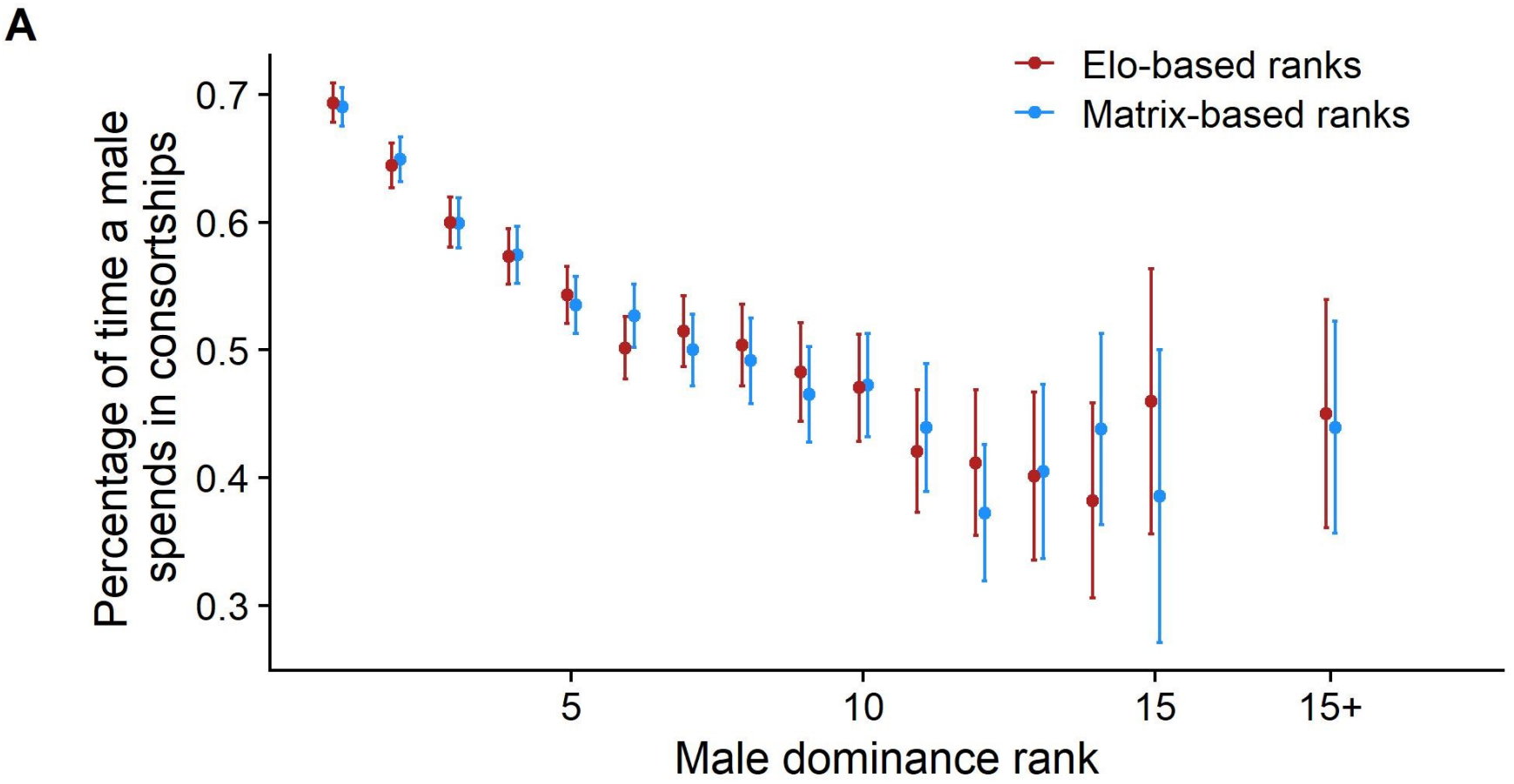
Elo ranking - compare models

```
get_estimates <- function(rank_measures, model) {
  broom.mixed::tidy(x = model) %>%
    filter(rank_measures == term)}

get_confidence <- function(rank_measures, model) {
  broom.mixed::tidy(x = model, conf.int = TRUE, conf.level = 0.95) %>%
    filter(rank_measures == term) %>%
    select(contains("conf"))}

model_compare <- model_compare_setup %>%
  mutate(AIC = map_dbl(.x = model, .f = get_AIC),
         estimate= map2(.x = rank_measures, .y =model, .f = get_estimates),
         tidy_results= map2(.x = rank_measures, .y =model, .f = get_confidence)
  unnest(cols = c(estimate, tidy_results)) %>%
   mutate(standard_estimate = estimate/std.error)
```

Elo ranking


Silk et al 2003
Levy et al 2020
Gesquiere et al 2018 (pregnancy)
Gesquiere et al 2018 (PPA)
Gesquiere et al 2018 (IBI)
Gesquiere et al 2018 (cycling)
Fitzpatrick et al 2014
Archie et al 2014
Gesquiere et al 2011 (T)
Gesquiere et al 2011 (GC)
Archie et al 2012 (Days to heal)
Alberts et al 2006
Akinyi et al 2013
Onyango et al 2008
Nguyen et al 2012
Charpentier et al 2008 (Males)
Altmann and Alberts 2005
ADF
ADM
Mat
-10
-5
0
5
Standardized effect sizes
Elo-based
Matrix-based rank

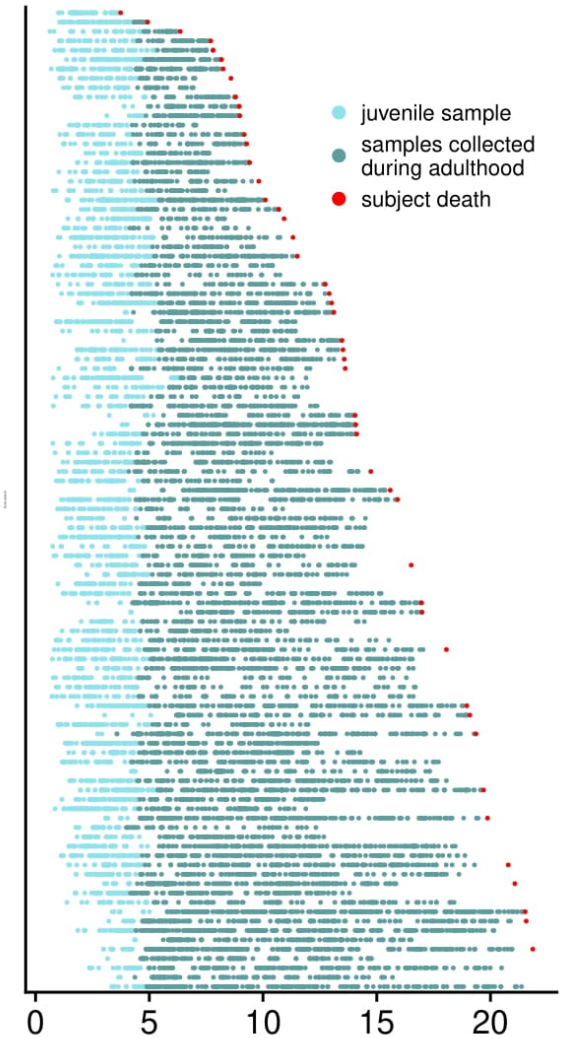
Elo ranking


A
Elo-based ranks
Matrix-based ranks
Percentage of time a male spends in consortships
0.7
0.6
0.5
0.4
0.3
5
10
15
15+
Male dominance rank

Large sequencing project

Gut microbe project

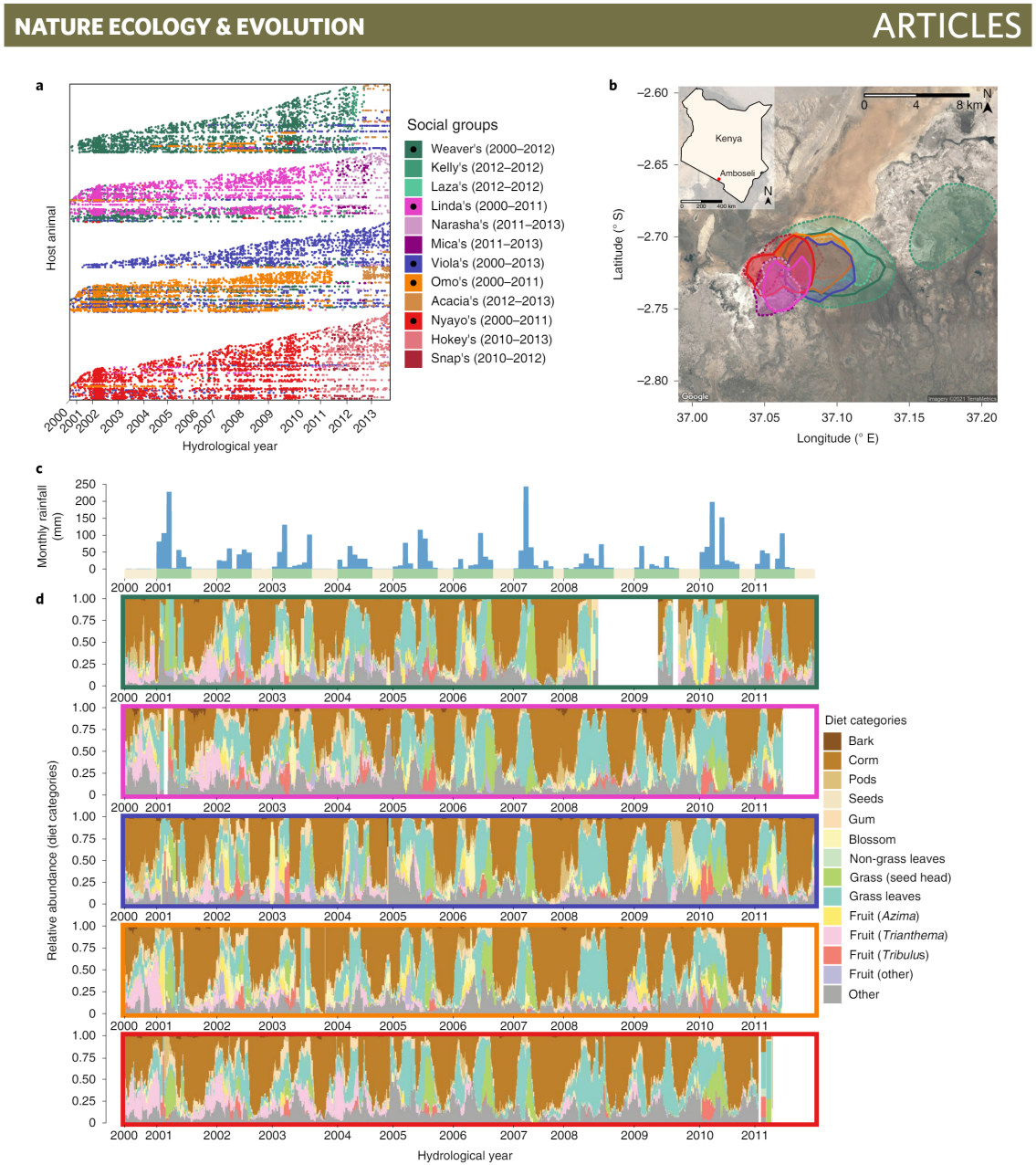
- Over 20.000 freeze-dried fecal samples
- Sequenced to get data on the gut microbiome of baboons



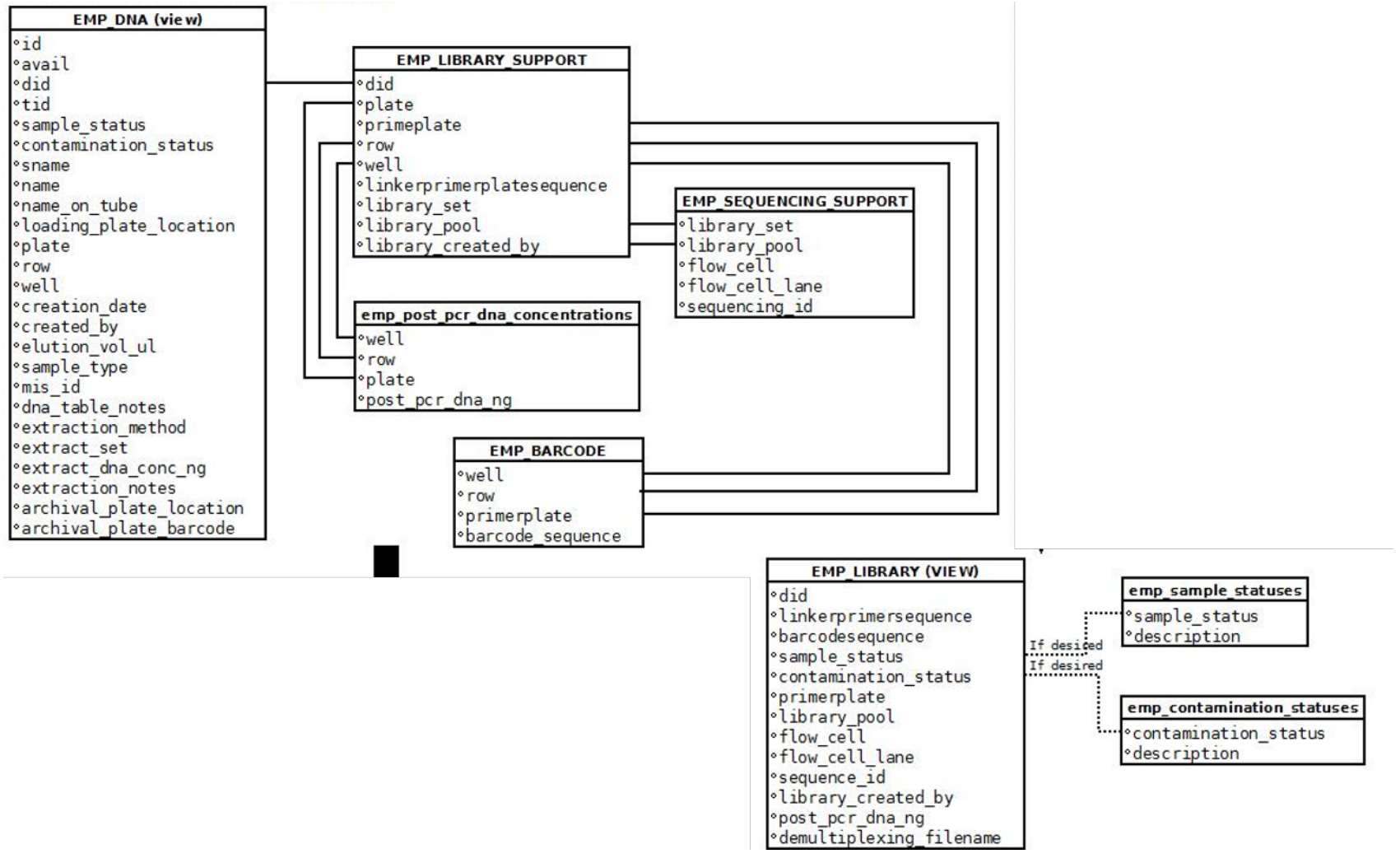
EMP sequencing project

- Responsible for the development of methods and coordinating lab work
- Created database structure and documentation
  - Tables for the various stages of the pipeline
  - Support tables for quality of samples
  - Links to the full project database as well as downstream analysis results
- Lots of work on cluster computers moving large volumes of data and linking datatype
- Assisted with analyses and visualizations of multiple papers

EMP sequencing project



EMP sequencing project



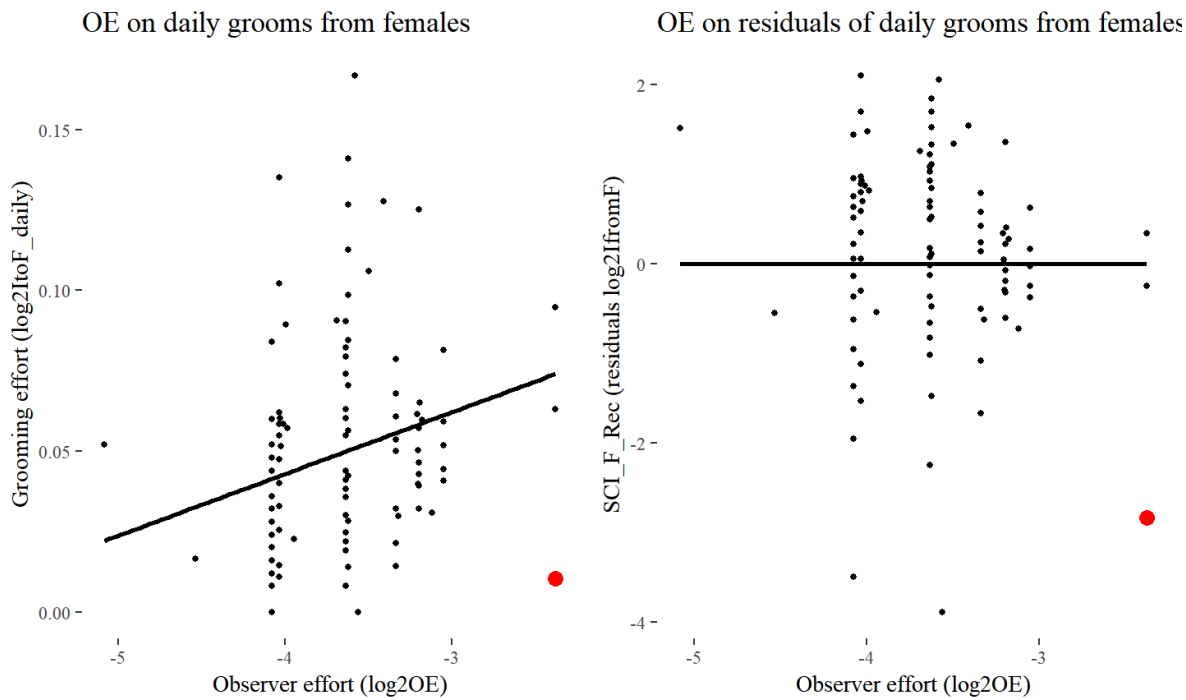
Sociality indexes

Social connectedness index

- Measures of how socially connected individuals are
- Being socially isolated is bad for your survival/fitness
- Based on grooming behaviors

Social connectedness index

- ‘composite’ indices of social connectedness
- Problem of observer effort and social group sizes.

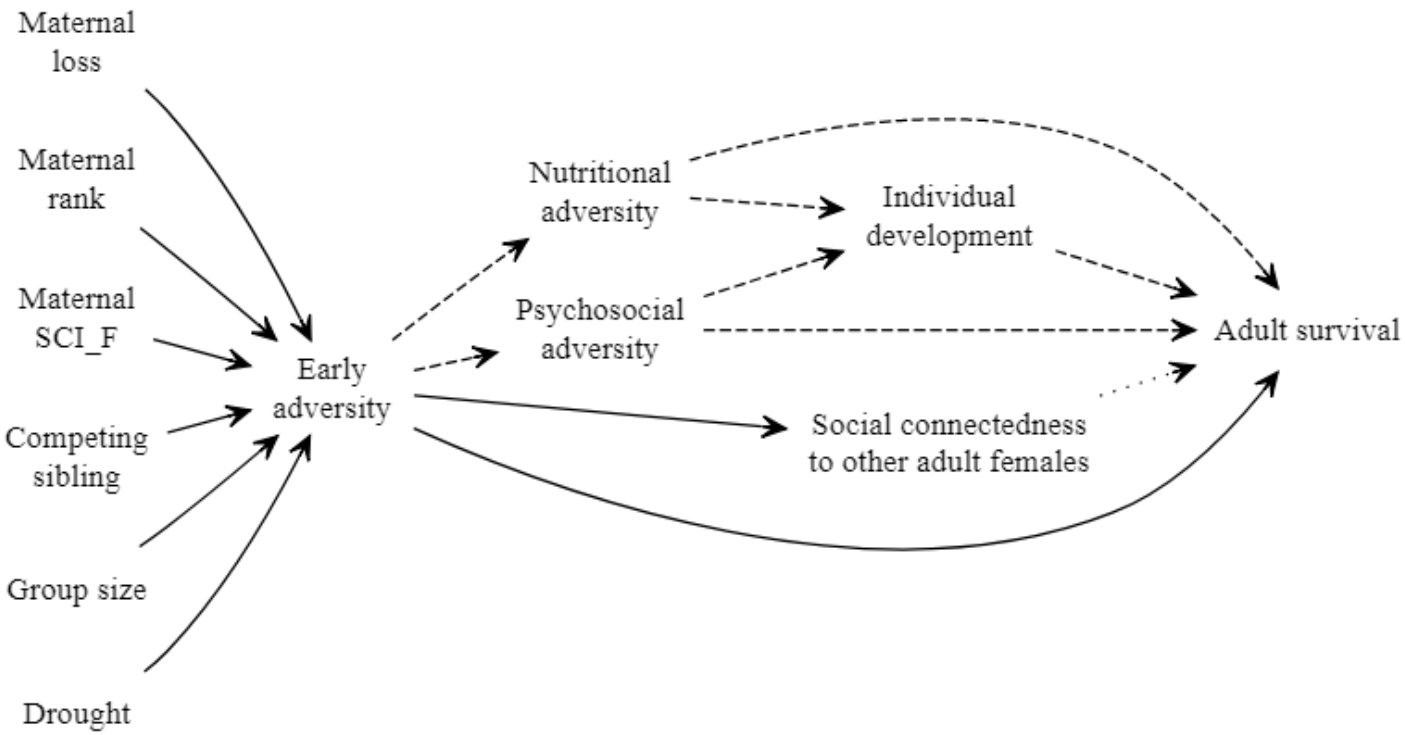


Social connectedness index

- Original code was mostly 'lost'.
- Reconstruct code from methods in paper
- Test code by comparing results with results from paper
  - Lots of base R and the datatable and reshape packages
- In collaboration with a post-doc this was later turned into an R package
  - Started to use Tidyverse and parallel coding
  - Added dyadic interaction (on-on-one) strengths

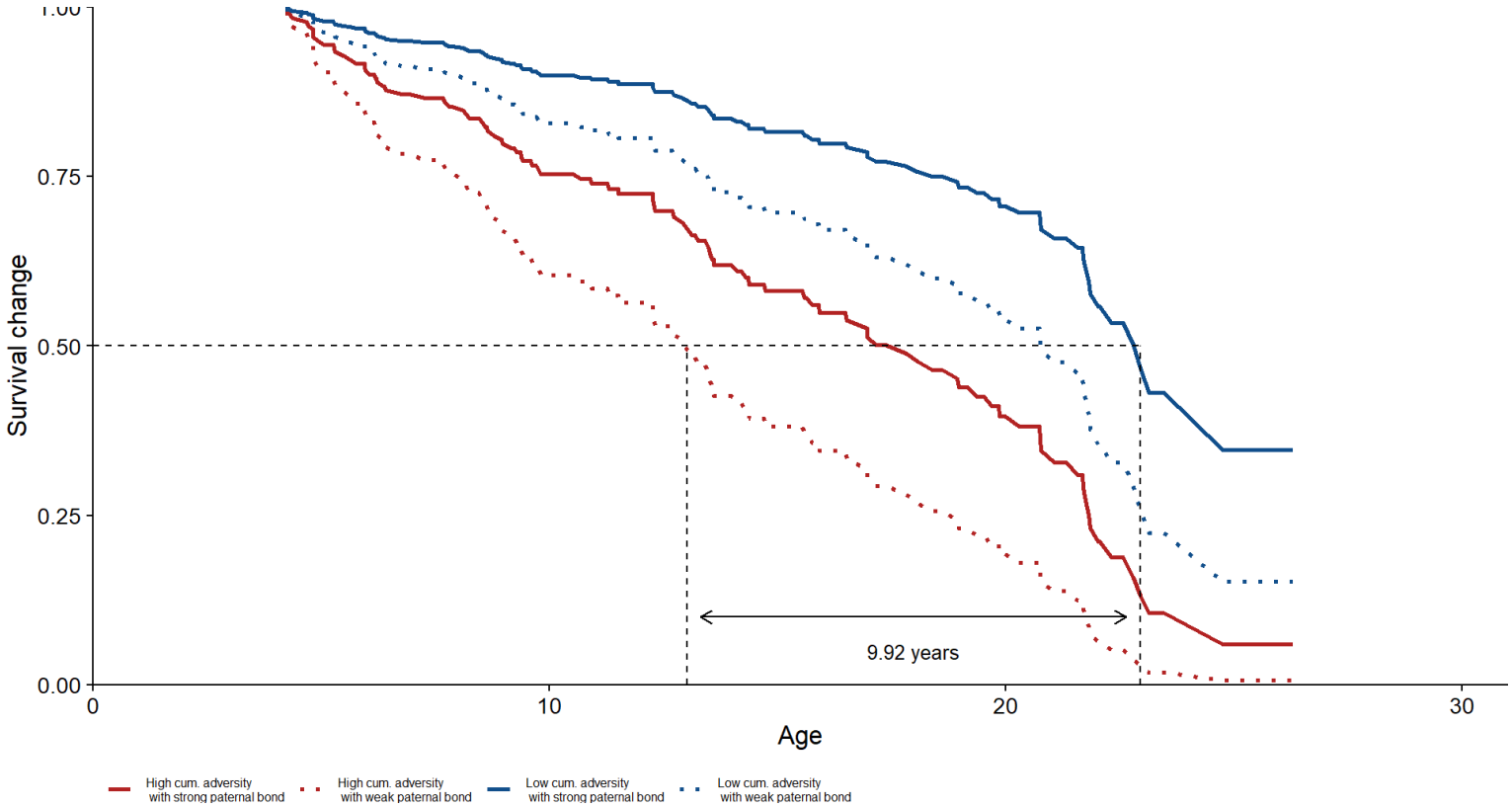
Social connectedness index

- Initially only for adults
- I alter code to create indexes for juveniles



Effect of male interactions with juvenile females

- Found that strong bonds with fathers increase adult survival
- Strong bonds with dads may mitigate effects of early adversities



Effect of male interactions with juvenile females

	Hazard ratio (95% CI)		Model parameters		
model	Cum. adversity of juvenile	jDSI with dad	AICc	ΔAICc	coxzph
C	1.386*** (1.088-1.767)	0.668* (0.483-0.922)	811.48	0.00	0.48
B	1.273* (1.012-1.601)		815.23	3.76	0.34
A		0.75 (0.547-1.03)	816.29	4.81	0.84

There are 216 juvenile famales in this analysis

A = Adult female survival - juvenile dyadic strength with dad (jDSI with dad)

B = Adult female survival - cumulative adversity

C = Adult female survival - cumulative adversity + juvenile dyadic strength with dad (jDSI with dad)