Albacore Review Diet Data Manipulation

Natasha Hardy

03/08/2021

About

This document contains code for manipulating diet data sets for albacore prey species from our global meta-analysis of albacore diet composition. Crucially, this is where we parse important decisions on data such as the known or likely size of albacore tuna prey taxa, calculate the maxillary or gape length for albacore tunas based on their reported or estimated fork lengths, and thereby associate relevant trait information.

```
#R documentation
library(pander)
# general
library(tidyverse)
library(readxl)
library(readr)
library(plyr)
library(dplyr)
library(devtools)
library(here)
"%notin%" = Negate('%in%')
here::here()
```

```
## [1] "/Users/tashhardy/Documents/GitHub/albacore-diet-global"
```

Outputs needed for these manips + load:

- Prey species' morphometric trait information (body shape, eye diameter to body, and standard to total length): *data/output_data/prey_morph_sum.csv*
- Prey species' nutritional composition traits (energy density, percent lipid and percent protein composition): *data/output_data/prey_qual_sum.csv*
- Prey species' trait values: (i) selected priority trait columns *data/output_data/prey_trait_select.csv* and (ii) trait values for adult life stages only *data/output_data/prey_trait_adult.csv* –> for downstream data joins and analyses.
- Prey species length ratios for max and min length: *data/output_data/prey_length_wide.csv*
- Prey species + filled in length ratios: *data/output_data/prey_length_ratiosp.csv*
- Diet data below.

```
prey_morph_sum = read.csv(here("data/output_data/prey_morph_sum.csv"))

prey_qual_sum = read.csv(here("data/output_data/prey_qual_sum.csv"))

prey_trait_select = read.csv(here("data/output_data/prey_trait_select.csv"))

prey_trait_adult = read.csv(here("data/output_data/prey_trait_adult.csv"))

prey_length_wide = read.csv(here("data/output_data/prey_length_wide.csv"))

prey_length_ratiosp = read.csv(here("data/output_data/prey_length_ratiosp.csv"))
```

Albacore diet review data

Here we need to load up the albacore global diet data to clean and select what we want. Tasks performed in these chunks:

- Merge diet data with processed size ratio data;
- Calculate probable prey lengths based on albacore gape length or reported prey lengths, then generate a prey probable life stage;
- Merge probable prey life stage with appropriate trait information below;
- Retain frequency of occurrence + stomach sampled information for traitglm.

Albacore diet raw data

Load diet data and merge with prey size trait information, and undertake necessary data selection and manipulation of variables.

```
#Load diet data spreadsheet
#need this: prey_length_ratiosp
#This contains species as rows for which we have useable species through to
#class-level ratio information on larval:adult and juvenile:adult lengths

#Also need prey_length_wide data for l_max for adult column
#Should be able to get this from prey_length_ratiosp

alb_global = as.data.frame(read_xlsx(here("data/input_data/albacore_diet_global.xlsx"),
                     sheet = "albacore_diet_global")) %>% #2134 rows and 77 vars
  dplyr::filter(TaxLev == "species") %>% #1107 rows and 77 vars
  #There are ~1105 observations when we filter by species
  dplyr::filter(Include %in% c("Yes", "Maybe") & !pred_life == "larvae") %>% #1028 rows
  #1028 when we remove larval albacore diet studies (there are 3)
  #dplyr::rename() %>%
  dplyr::select(StudyID, CiteAuth, CiteYear, OceanBasin, OceanBasinQ, LocatName,
                LocatLatitude:LocatLongitude, MonthRange, YearEnd, stomachs_used,
                pred_life:est_ref, stomachs_used, prey_class:prey_sp,
                prey_age_reported_1:prey_age_use, Include, FO,
                pFOuse, N, pN, `Total Mass (g)`, pM, maxl_use, maxl_type) %>%
  mutate(gape_lmax = pred_flmean*0.0823+1.758,
```

```
        gape_height_limit = pred_flmean*0.0246+4.603) %>%
  #1028 obsm 44 vars
  #calculates the gape length and height limits
  #Add reference paper for this calculation -- Menard et al. (2006) for yellowfin tuna
  left_join(prey_morph_sum, by="prey_sp")
  #added morph info --> 1028 obs, 44 vars

#Correct data type
alb_global$maxl_use <- as.numeric(alb_global$maxl_use)

#Cleaning metrics
alb_global = alb_global %>%
  mutate(maxtl_use = if_else(alb_global$maxl_type %in% c("SL", "ML"),
                             maxl_use/standard_total, maxl_use),
         prey_age_reported_1=case_when(
                             #prey_age_reported_1 == "NA" ~ "none",
                             prey_age_reported_1 == "larvae" ~ "larva",
                             prey_age_reported_1 == "juvenile" ~ "juvenile",
                             prey_age_reported_1 == "adult" ~ "adult")
         ) #%>%

#The NA's in the reported age for prey column cause problems later and need
#to be assigned a categorical value
alb_global$prey_age_reported_1 <- as.character(alb_global$prey_age_reported_1)
alb_global$prey_age_reported_1 <- replace_na(alb_global$prey_age_reported_1, "none")
#check this
unique(as.factor(alb_global$prey_age_reported_1))
```

```
## [1] none     adult    juvenile larva
## Levels: adult juvenile larva none
```

```
#Change vector type for YearEnd variable
alb_global$YearEnd <- as.integer(alb_global$YearEnd)
alb_global$stomachs_used <- as.integer(alb_global$stomachs_used)
alb_global$stomachs_used <- replace_na(alb_global$stomachs_used, "none reported")
alb_global$stomachs_used <- as.character(alb_global$stomachs_used)

#Save total df version
write.csv(alb_global, here("data/output_data/alb_global_diet_total.csv"),
          row.names = FALSE)
#1028 observations and 45 variables
```

The following summarises the max observed FO, N and M for each species in the dataset. This information can be useful when investigating the relative importance of species in albacore diets.

```
alb_global_ad_metrics = alb_global %>%
  #unique(c(alb_global_prob$prey_sp, alb_global_prob$prey_select))
  dplyr::select(prey_class:prey_sp, pFOuse, pN, pM, maxtl_use) %>%
  group_by(prey_sp) %>% #, life_stage , pFOuse, pN, pM
  dplyr::summarize(maxFO = max(pFOuse, na.rm=TRUE),
                   maxN = max(pN, na.rm=TRUE),
                   maxM = max(pM, na.rm=TRUE),
```

```
                  maxTL = max(maxtl_use, na.rm=TRUE))

write.csv(alb_global_ad_metrics, here("data/output_data/alb_global_ad_metrics.csv"),
          row.names = FALSE)
```

Albacore diet + probable life stage estimation

Here we want to estimate the probable life stage of the species consumed using their selected size ratio information and relationship to albacore tuna gape limits.

```
alb_global_prob = alb_global %>%
  left_join(prey_length_ratiosp, by=c("prey_sp", "prey_family", "prey_order", "prey_class")) %>% #previ
  #In using an inner_join I am selecting only rows for which we have all length-based data
  mutate(prey_length = if_else(maxtl_use > 0, maxtl_use/10,
                               if_else(l_adult < gape_lmax, l_adult, gape_lmax))) %>%
  #Note that we recorded any reported prey lengths in cm's and we need them in mm's now.
  mutate(prey_frac = if_else(l_adult < gape_lmax, 100, 100*prey_length/l_adult)) %>%
  mutate(prey_proba = if_else(prey_frac < larva_adult_use, "larva", if_else(prey_frac<juve_adult_use, "
  mutate(life_stage = if_else(prey_age_reported_1 == "none", prey_proba, prey_age_reported_1)) #985 obs

#And remove troublesome NA's in dataset
#Because we need to give these a life stage, I am assigning them as "adult" because of their l_max typi
#These were all small species
alb_global_prob$life_stage[is.na(alb_global_prob$life_stage)] <- "adult"

## SAVE DATA
write.csv(alb_global_prob, here("data/output_data/alb_global_diet_prob.csv"), row.names = FALSE)
#This include data for which we are able to make an assessment of probable life stage --> therefore wil

#Note that the inner join created an issue where we had missing data --
#985 observations 55 variables
#alb_missing = alb_global %>%
#  anti_join(alb_global_prob)
```

The following summarises the max observed FO, N and M for each species in the dataset AND for each life stage likely consumed. This also generates **Table S7, for Appendix B, Supplementary Information**.

Note that there are 28 taxa for which two life stages were consumed. In the chunk below this one we will select the prey species and life stage combination that was most common. Most of the time this wouldn't be a large problem because typically one life stage was consumed much more frequently than the other. However, for *Cololabis saira*, the %FO of the adult and juvenile life stages consumed were very similar in summary with the adult life stage slightly higher.

```
# All combinations of life stages and species consumed
alb_global_prob_metrics = alb_global_prob %>%
    dplyr::select(prey_class:prey_sp, prey_age_reported_1, life_stage, pFOuse, pN,
        pM, maxtl_use, l_adult, gape_lmax, prey_length, prey_frac, larva_adult_use,
        juve_adult_use) %>%
    group_by(prey_sp, prey_age_reported_1, life_stage, l_adult, larva_adult_use,
        juve_adult_use) %>%
    dplyr::summarize(maxFO = max(pFOuse, na.rm = TRUE), maxN = max(pN, na.rm = TRUE),
        maxM = max(pM, na.rm = TRUE), maxTL = max(maxtl_use, na.rm = TRUE)/10, meanGAPE = mean(gape_lma
            na.rm = TRUE), meanLENGTH = mean(prey_length, na.rm = TRUE), meanFRAC = mean(prey_frac,
```

```
            na.rm = TRUE)) %>%
    dplyr::select(prey_sp, prey_age_reported_1, maxTL, meanGAPE, l_adult, meanLENGTH,
        meanFRAC, larva_adult_use, juve_adult_use, life_stage, maxFO, maxN, maxM)
```

```
## `summarise()` has grouped output by 'prey_sp', 'prey_age_reported_1', 'life_stage', 'l_adult', 'larva
```

```
# maxGAPE = max(gape_lmax, na.rm=TRUE), maxLENGTH = max(prey_length,
# na.rm=TRUE)

## SAVE DATA BELOW
write.csv(alb_global_prob_metrics, here("data/output_data/alb_global_prob_metrics_alllife.csv"),
    row.names = FALSE)
```

AND we need to select the max observed –> this is also where we essentially select a single life stage that any prey could be, even though there are several occurrences of prey from different life stages being consumed in different areas of the world. There were 328 individual species and life stage combinations above, but we cannot account for all of these unless we treat these as different species. So 50 cases were simplified and merged. We can check in the file above which taxa this affected and their frequency of occurrence.

```
# Need to summarise prey life stages and attach trait data
# unique(as.factor(alb_global_prob_metrics$prey_sp)) #there are 303 unique
# species So we should obtain 300 species in the following code.

# The most common life stage and species consumed!!
alb_global_prob_metrics2 = alb_global_prob_metrics %>%
    group_by(prey_sp) %>%
    top_n(1, maxFO) %>%
    sample_n(1)

write.csv(alb_global_prob_metrics2, here("data/output_data/alb_global_prob_metrics.csv"),
    row.names = FALSE)
```

*Note: the maxFO, maxN or maxM are related to the study and differ between life stages if a study reported different prey life stages.

Albacore prey size when reported

Here we include the raw prey size data from the trait database, including the maximum and minimum reported lengths for prey for adult, juvenile and larval life stages. Data collected from FishBase (2020), SeaLifeBase (2020) and from extensive literature searches on Google Scholar. This generated **Table S8, for Appendix B, Supplementary Information**.

```
alb_prey_raw_size = alb_global %>%
    filter(maxtl_use > 0) %>%
    dplyr::select(StudyID, OceanBasin, LocatName, LocatLatitude, LocatLongitude,
        YearEnd, pred_life, pred_flmean, pred_flmin, pred_flmax, gape_lmax, prey_sp,
        maxl_use, maxl_type, standard_total, maxtl_use, pFOuse, pN, pM) %>%
    mutate(maxtl_use = maxtl_use/10, maxl_use = maxl_use/10, presence = 1)

write.csv(alb_prey_raw_size, here("data/output_data/alb_prey_raw_size.csv"), row.names = FALSE)
```

Data Joins

Traits

```
# Mixed categorical, and other traits

prey_adults_merge = prey_trait_adult %>%
    left_join(prey_morph_sum, by = "prey_sp") %>%
    left_join(prey_qual_sum, by = "prey_sp") %>%
    left_join(alb_global_ad_metrics, by = "prey_sp") %>%
    dplyr::select(prey_class:season_cat, gregarious_primary, l_max, body_shape, b_shape_r:standard_total
        phys_defense:countershade, energy_density:maxTL, trophic_level, fisheries_status)

# 308 spp and ~31 traits
glimpse(prey_adults_merge)
```

Prey traits - adult life stage

```
## Rows: 308
## Columns: 32
## $ prey_class         <chr> "Hydrozoa", "Malacostraca", "Malacostraca", "Malaco~
## $ prey_order         <chr> "Siphonophorae", "Decapoda", "Decapoda", "Decapoda"~
## $ prey_family        <chr> "Diphyidae", "Acanthephyridae", "Oplophpridae", "Po~
## $ prey_sp            <chr> "Chelophyes appendiculata", "Acanthephyra pelagica"~
## $ vert_habitat       <chr> "epipelagic", "mesopelagic", "mesopelagic", "epipel~
## $ horz_habitat       <chr> "continental shelf", "oceanic", "oceanic", "contine~
## $ diel_migrant       <int> 1, 1, 1, NA, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0,~
## $ diel_migrant_cat   <chr> "diel_yes", "diel_yes", "diel_yes", "diel_UN", "die~
## $ refuge             <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, ~
## $ refuge_cat         <chr> "refuge_no", "refuge_no", "refuge_no", "refuge_yes"~
## $ season_migrant     <int> 1, 1, 1, 0, 1, 1, 1, NA, 1, 0, 1, 0, NA, 1, 1, 0, N~
## $ season_cat         <chr> "season_yes", "season_yes", "season_yes", "season_n~
## $ gregarious_primary <chr> "schooling", "schooling", "schooling", "shoaling", ~
## $ l_max              <dbl> 0.2, 27.0, 1.7, 4.4, 110.0, 60.0, 37.0, 8.0, 31.0, ~
## $ body_shape         <chr> "fusiform", "elongated", "elongated", "depressiform~
## $ b_shape_r          <dbl> 4.105854, 6.237058, 8.425900, NA, 3.517372, 5.59883~
## $ eye_body_r         <dbl> 0.00000000, 0.02214046, 0.02530975, NA, 0.03949694,~
## $ standard_total     <dbl> 1.0000000, 0.8005450, 0.8591838, NA, 0.9093951, 0.8~
## $ phys_defense       <int> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ transparent        <int> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ col_disrupt        <int> 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, ~
## $ silver             <int> 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ countershade       <int> 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ energy_density     <dbl> NA, NA, 3.156000, 4.200000, 6.129560, 7.823333, 5.4~
## $ percent_protein    <dbl> NA, NA, 9.30000, 14.10000, 23.79125, 14.90000, 12.8~
## $ percent_lipid      <dbl> NA, 6.53750, 2.59000, 1.40000, 1.69875, 11.85000, N~
## $ maxFO              <dbl> 16.00000, 38.88889, 8.00000, 51.42857, 0.60000, 1.1~
## $ maxN               <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.273390, 0~
## $ maxM               <dbl> 0.000000, 0.000000, 0.000000, 1.000000, 0.000000, 0~
## $ maxTL              <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 8.0~
## $ trophic_level      <dbl> NA, NA, 3.32, 3.46, 4.43, 3.63, 4.50, NA, 3.67, 3.8~
## $ fisheries_status   <chr> "none", "none", "none", "none", "highly commercial"~
```

```
# Save
write.csv(prey_adults_merge, here("data/output_data/prey_traits_adults.csv"), row.names = FALSE)

# Reload prey_adults_merge =
# read.csv(here('data/output_data/prey_adults_traits.csv'))
```

```
prey_probable_merge = alb_global_prob_metrics %>%
    left_join(prey_trait_select, by = c("prey_sp", "life_stage")) %>%
    left_join(prey_morph_sum, by = "prey_sp") %>%
    left_join(prey_qual_sum, by = "prey_sp") %>%
    dplyr::select(prey_class:prey_family, prey_sp, prey_age_reported_1, life_stage,
        vert_habitat:season_cat, gregarious_primary, l_max, body_shape, b_shape_r:standard_total,
        phys_defense:countershade, energy_density:percent_lipid, maxFO:maxM, maxTL,
        trophic_level, fisheries_status)
```

Prey traits - probable life stage - all life stages

```
## Adding missing grouping variables: `l_adult`, `larva_adult_use`
```

```
# 303 spp and ~28 traits
glimpse(prey_probable_merge)
```

```
## Rows: 363
## Columns: 36
## Groups: prey_sp, prey_age_reported_1, life_stage, l_adult, larva_adult_use [363]
## $ l_adult             <dbl> 6.834282, 9.410809, 17.455708, 12.072722, 10.01898~
## $ larva_adult_use     <dbl> 21.883929, 21.883929, 6.625000, 21.883929, 37.1428~
## $ prey_class          <chr> "Cephalopoda", "Cephalopoda", "Cephalopoda", "Ceph~
## $ prey_order          <chr> "Oegopsida", "Oegopsida", "Oegopsida", "Oegopsida"~
## $ prey_family         <chr> "Enoploteuthidae", "Enoploteuthidae", "Enoploteuth~
## $ prey_sp             <chr> "Abralia redfieldi", "Abraliopsis affinis", "Abral~
## $ prey_age_reported_1 <chr> "none", "none", "none", "none", "none", "adult", "~
## $ life_stage          <chr> "adult", "adult", "adult", "adult", "adult", "adul~
## $ vert_habitat        <chr> "mesopelagic", "mesopelagic", "mesopelagic", "meso~
## $ horz_habitat        <chr> "oceanic", "oceanic", "oceanic", "continental slop~
## $ diel_migrant        <int> 1, 1, 1, 1, 1, 1, 1, NA, 1, NA, 1, 1, 1, 0, NA, NA~
## $ diel_migrant_cat    <chr> "diel_yes", "diel_yes", "diel_yes", "diel_yes", "d~
## $ refuge              <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ refuge_cat          <chr> "refuge_no", "refuge_no", "refuge_no", "refuge_no"~
## $ season_migrant      <int> NA, NA, NA, NA, NA, 1, NA, 1, 0, NA, 1, 1, 1, 1, 1~
## $ season_cat          <chr> "season_NA", "season_NA", "season_NA", "season_NA"~
## $ gregarious_primary  <chr> NA, NA, NA, NA, NA, "schooling", "schooling", "sch~
## $ l_max               <dbl> 3.60, 4.30, 8.00, 5.10, 3.50, 27.00, NA, 7.80, 0.4~
## $ body_shape          <chr> "fusiform", "fusiform", "fusiform", "fusiform", "f~
## $ b_shape_r           <dbl> 3.914416, 5.345385, 5.280579, 5.630847, 8.389906, ~
## $ eye_body_r          <dbl> 0.08487754, 0.06979495, 0.05075453, 0.05010374, 0.~
## $ standard_total      <dbl> 0.5267561, 0.4569214, 0.4583028, 0.4224399, 0.3493~
## $ phys_defense        <int> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1,~
## $ transparent         <int> 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ col_disrupt         <int> 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1,~
```

```
## $ silver              <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,~
## $ countershade        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,~
## $ energy_density      <dbl> NA, NA, 4.40000, NA, NA, NA, 1.70000, 3.90000, NA,~
## $ percent_protein     <dbl> NA, NA, 17.40000, NA, NA, NA, NA, 16.40000, NA, NA~
## $ percent_lipid       <dbl> NA, NA, NA, NA, NA, 6.5375, NA, 1.3000, NA, NA, NA~
## $ maxFO               <dbl> 5.000000, 1.162791, 36.400000, 0.700000, 4.900000,~
## $ maxN                <dbl> 0.00000000, 0.07651109, 11.30000000, 0.20000000, 3~
## $ maxM                <dbl> 0.000000000, 0.006092434, 2.132352045, 0.100000000~
## $ maxTL               <dbl> 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 0.000~
## $ trophic_level       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ fisheries_status    <chr> "none", "none", "none", "none", "none", "none", "n~
```

```
write.csv(prey_probable_merge, here("data/output_data/prey_trait_probable_all.csv"),
    row.names = FALSE)

# prey_probable_merge = read.csv(here('data/prey_probable_traits.csv'))
```

```
prey_probable_merge2 = alb_global_prob_metrics2 %>%
    left_join(prey_trait_select, by = c("prey_sp", "life_stage")) %>%
    left_join(prey_morph_sum, by = "prey_sp") %>%
    left_join(prey_qual_sum, by = "prey_sp") %>%
    dplyr::select(prey_class:prey_family, prey_sp, prey_age_reported_1, life_stage,
        vert_habitat:season_cat, gregarious_primary, l_max, body_shape, b_shape_r:standard_total,
        phys_defense:countershade, energy_density:percent_lipid, maxFO:maxM, maxTL,
        trophic_level, fisheries_status)  #meanGAPE, meanLENGTH, meanFRAC,

# 308 spp and ~28 traits
glimpse(prey_probable_merge2)
```

Prey traits - probable life stage - select life stages

```
## Rows: 308
## Columns: 34
## Groups: prey_sp [308]
## $ prey_class          <chr> "Cephalopoda", "Cephalopoda", "Cephalopoda", "Ceph~
## $ prey_order          <chr> "Oegopsida", "Oegopsida", "Oegopsida", "Oegopsida"~
## $ prey_family         <chr> "Enoploteuthidae", "Enoploteuthidae", "Enoploteuth~
## $ prey_sp             <chr> "Abralia redfieldi", "Abraliopsis affinis", "Abral~
## $ prey_age_reported_1 <chr> "none", "none", "none", "none", "none", "adult", "~
## $ life_stage          <chr> "adult", "adult", "adult", "adult", "adult", "adul~
## $ vert_habitat        <chr> "mesopelagic", "mesopelagic", "mesopelagic", "meso~
## $ horz_habitat        <chr> "oceanic", "oceanic", "oceanic", "continental slop~
## $ diel_migrant        <int> 1, 1, 1, 1, 1, 1, 1, NA, 1, NA, 1, 1, 1, 0, NA, NA~
## $ diel_migrant_cat    <chr> "diel_yes", "diel_yes", "diel_yes", "diel_yes", "d~
## $ refuge              <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ refuge_cat          <chr> "refuge_no", "refuge_no", "refuge_no", "refuge_no"~
## $ season_migrant      <int> NA, NA, NA, NA, NA, 1, NA, 1, 0, NA, 1, 1, 1, 1, 1~
## $ season_cat          <chr> "season_NA", "season_NA", "season_NA", "season_NA"~
## $ gregarious_primary  <chr> NA, NA, NA, NA, NA, "schooling", "schooling", "sch~
## $ l_max               <dbl> 3.60, 4.30, 8.00, 5.10, 3.50, 27.00, NA, 7.80, 0.4~
```

```
## $ body_shape          <chr> "fusiform", "fusiform", "fusiform", "fusiform", "f~
## $ b_shape_r           <dbl> 3.914416, 5.345385, 5.280579, 5.630847, 8.389906, ~
## $ eye_body_r          <dbl> 0.08487754, 0.06979495, 0.05075453, 0.05010374, 0.~
## $ standard_total      <dbl> 0.5267561, 0.4569214, 0.4583028, 0.4224399, 0.3493~
## $ phys_defense        <int> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1,~
## $ transparent         <int> 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ col_disrupt         <int> 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1,~
## $ silver              <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,~
## $ countershade        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,~
## $ energy_density      <dbl> NA, NA, 4.40000, NA, NA, NA, 1.70000, 3.90000, NA,~
## $ percent_protein     <dbl> NA, NA, 17.40000, NA, NA, NA, NA, 16.40000, NA, NA~
## $ percent_lipid       <dbl> NA, NA, NA, NA, NA, 6.5375, NA, 1.3000, NA, NA, NA~
## $ maxFO               <dbl> 5.000000, 1.162791, 36.400000, 0.700000, 4.900000,~
## $ maxN                <dbl> 0.00000000, 0.07651109, 11.30000000, 0.20000000, 3~
## $ maxM                <dbl> 0.000000000, 0.006092434, 2.132352045, 0.100000000~
## $ maxTL               <dbl> 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 0.000~
## $ trophic_level       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ fisheries_status    <chr> "none", "none", "none", "none", "none", "none", "n~
```

```
write.csv(prey_probable_merge2, here("data/output_data/prey_trait_probable_select.csv"),
    row.names = FALSE)

# prey_probable_merge = read.csv(here('data/prey_probable_traits.csv'))
```

Covariates

```
alb_covars = alb_global %>%
  dplyr::select(StudyID, CiteYear, OceanBasin, OceanBasinQ, LocatName, LocatLatitude:pred_life_est) %>%
  distinct(StudyID, CiteYear, OceanBasin, OceanBasinQ, LocatName, LocatLatitude,
           LocatLongitude, YearEnd, MonthRange, stomachs_used, pred_life) %>%
  #, stomachs_used
  mutate(lat_cat = if_else(abs(LocatLatitude) > 23.43662, "temperate", "tropical"),
         pred_life_adjust=case_when(pred_life == "juvenile" ~ "juvenile",
                                    pred_life == "juvenile, adult" ~ "mixed",
                                    pred_life == "none" ~ "mixed",
                                    pred_life == "adult" ~ "adult"))

alb_covars$rep_ID = paste(alb_covars$StudyID, row.names(alb_covars), sep = "_")

#Merge these back to the diet df
alb_global_diet = alb_global %>%
  left_join(alb_covars)
```

```
## Joining, by = c("StudyID", "CiteYear", "OceanBasin", "OceanBasinQ", "LocatName", "LocatLatitude", "L
```

```
write.csv(alb_covars, here("data/output_data/alb_covars.csv"), row.names = FALSE)
```

GO TO 'longhurst' work then back here

Extracting co-variates Could edit the Longhurst doc to separate biomes and codes as per province definitions: https://en.wikipedia.org/wiki/Longhurst_code#:~:text=Longhurst%20code%20refers%20to%20a,unique%20set%20of
Note that we are having issues with this code and need to re-run it / troubleshoot.

```
# Covars + Longhurst

alb_covars_geog <- read.csv(here("data/output_data/alb_covars_longhurst.csv"))  #%>%
# dplyr::rename(MonthRange2 = `MonthRange`) dplyr::select(-MonthRange)
# dplyr::select(StudyID, lat_cat:code) #%>% group_by(StudyID) %>% sample_n(1)
# names(alb_covars_geog)

str(alb_covars_geog)
```

```
## 'data.frame':    75 obs. of  17 variables:
##  $ StudyID         : chr  "Aloncle1973" "Bello1999" "Bello1999" "Bernard1985" ...
##  $ CiteYear        : chr  "1973.0" "1999.0" "1999.0" "1985.0" ...
##  $ OceanBasin      : chr  "N Atlantic" "Mediterranean" "Mediterranean" "N Pacific" ...
##  $ OceanBasinQ     : chr  "NE Atlantic" "Mediterranean" "Mediterranean" "NE Pacific" ...
##  $ LocatName       : chr  "NE Atlantic, Spain, France, Azores" "SW Adriatic" "SW Adriatic" "Southern
##  $ LocatLatitude   : num  44.6 41.1 41.2 35.4 48 ...
##  $ LocatLongitude  : num  -5.5 17.5 17.5 -121.9 -11.7 ...
##  $ MonthRange      : chr  "May-Sep" "Sep-Oct" "Sep-Oct" "Aug-Sep" ...
##  $ YearEnd         : int  1971 1992 1994 1983 1929 1930 1930 1931 1932 1932 ...
##  $ stomachs_used   : chr  "1754" "35" "21" "94" ...
##  $ pred_life       : chr  "juvenile" "juvenile" "juvenile" "juvenile" ...
##  $ lat_cat         : chr  "temperate" "temperate" "temperate" "temperate" ...
##  $ pred_life_adjust: chr  "juvenile" "juvenile" "juvenile" "juvenile" ...
##  $ rep_ID          : chr  "Aloncle1973_1" "Bello1999_2" "Bello1999_3" "Bernard1985_4" ...
##  $ province        : chr  "Westerlies - N. Atlantic Drift Province (WWDR)" "Westerlies - Mediterranea
##  $ code            : chr  "NADR" "MEDI" "MEDI" "CCAL" ...
##  $ optional        : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```
str(alb_covars)
```

```
## 'data.frame':    75 obs. of  14 variables:
##  $ StudyID         : chr  "Aloncle1973" "Bello1999" "Bello1999" "Bernard1985" ...
##  $ CiteYear        : chr  "1973.0" "1999.0" "1999.0" "1985.0" ...
##  $ OceanBasin      : chr  "N Atlantic" "Mediterranean" "Mediterranean" "N Pacific" ...
##  $ OceanBasinQ     : chr  "NE Atlantic" "Mediterranean" "Mediterranean" "NE Pacific" ...
##  $ LocatName       : chr  "NE Atlantic, Spain, France, Azores" "SW Adriatic" "SW Adriatic" "Southern
##  $ LocatLatitude   : num  44.6 41.1 41.2 35.4 48 ...
##  $ LocatLongitude  : num  -5.5 17.5 17.5 -121.9 -11.7 ...
##  $ MonthRange      : chr  "May-Sep" "Sep-Oct" "Sep-Oct" "Aug-Sep" ...
##  $ YearEnd         : int  1971 1992 1994 1983 1929 1930 1930 1931 1932 1932 ...
##  $ stomachs_used   : chr  "1754" "35" "21" "94" ...
##  $ pred_life       : chr  "juvenile" "juvenile" "juvenile" "juvenile" ...
##  $ lat_cat         : chr  "temperate" "temperate" "temperate" "temperate" ...
##  $ pred_life_adjust: chr  "juvenile" "juvenile" "juvenile" "juvenile" ...
##  $ rep_ID          : chr  "Aloncle1973_1" "Bello1999_2" "Bello1999_3" "Bernard1985_4" ...
```

```
# Need to merge this with diet data first overall data then probable prey
# consumed
unique(alb_covars_geog$code)
```

```
##  [1] "NADR" "MEDI" "CCAL" "SATL" "CNRY" "NPTG" "NASW" "NECS" "MONS" "ISSG"
## [11] "EAFR" "NPPF" "ARCH" "SANT" "NEWZ" "SPSG"
```

```
# check names match unique(sort(alb_covars_geog$StudyID)) ==
# unique(sort(alb_global$StudyID)) #they do
# unique(sort(alb_covars_geog$CiteYear)) == unique(sort(alb_global$CiteYear))
# #they do unique(sort(alb_covars_geog$YearEnd)) ==
# unique(sort(alb_global$YearEnd)) #they do unique(sort(alb_covars_geog$code))

# check names match
unique(sort(alb_covars_geog$StudyID)) == unique(sort(alb_covars$StudyID))  #they do
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
unique(sort(alb_covars_geog$CiteYear)) == unique(sort(alb_covars$CiteYear))  #they do
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
unique(sort(alb_covars_geog$YearEnd)) == unique(sort(alb_covars$YearEnd))  #they do
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE
```

```
unique(sort(alb_covars_geog$MonthRange)) == unique(sort(alb_covars$MonthRange))  #they do NOT
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE
```

```
# Potential problem with MonthRange

alb_covars_use = alb_covars_geog %>%
    dplyr::select(-optional, -MonthRange)
```

```
alb_global_diet = alb_global_prob %>%
  left_join(alb_covars_use) %>% #looks good #, by = "StudyID"
  dplyr::select(StudyID:YearEnd, rep_ID, province, code, lat_cat, pred_life_adjust,
                stomachs_used:est_ref, prey_class:maxtl_use, prey_length:life_stage)
```

```
## Joining, by = c("StudyID", "CiteYear", "OceanBasin", "OceanBasinQ", "LocatName", "LocatLatitude", "L
```

```
#Write over previous output from first data chunk
write.csv(alb_global_diet, here("data/output_data/alb_global_diet_total.csv"),
          row.names = FALSE)
```

NOTE Below we include the adult cluster number but need to use the cluster number from probable life stage if we stick to this for final analysis.

Diet datasets

Note that there are a handful of cases where the same prey species was consumed and reported for two different life stages. Because this affects three species and for only one study, the simplest solution to be able to move forward, we need to just select the most commonly consumed species/life stage combination.

```
# Previous problem rows View(alb_global_diet[c(192, 226, 111, 118, 204, 247,
# 99, 117, 227, 248),]) select 192, 111, 247, 99, 248 alb_global_diet_use =
# alb_global_diet %>% slice(., -c(226, 118, 117, 204, 227))

# Updated data was downloaded with different order of the data so these rows
# have moved
View(alb_global_diet[c(58, 71, 70, 77, 64, 81, 69, 76, 72, 82), ])

# We select to keep the most common one (typically these were the same species
# appearing in the same study for two different life stages and this only
# affected the Bouxin & Legendre papers)
alb_global_diet_use = alb_global_diet %>%
    slice(., -c(71, 77, 64, 76, 72))
```

Presence/absence Here we will join our additional covariates (inc. Longhurst province data) with the diet data, and generate a wide format df with species as columns and values for species' presence absence (PA). These data will be used for traitglms.

```
#Extract data --> filter for only species that we have adult traits + cluster info for
# --> transform to wide data for MV GLMs + trait GLMs
alb_global_dietpa = alb_global_diet_use %>% #_use
  dplyr::select(#StudyID:stomachs_used, lat_cat:code, pred_life_adjust, pred_flmean, prey_age_reported_
    StudyID, rep_ID, YearEnd, OceanBasin, OceanBasinQ, province, code, lat_cat,
    pred_life_adjust, prey_sp) %>%
  mutate(pa="1") %>%
  group_by(prey_sp) %>%
  #dplyr::mutate(grouped_id = row_number()) %>%
  spread(prey_sp, pa, fill = 0, convert = FALSE)

#Check covariate representation
summary(as.factor(alb_global_dietpa$lat_cat))
```

```
## temperate  tropical
##        69         6
```

```
summary(as.factor(alb_global_dietpa$code))
```

```
## ARCH CCAL CNRY EAFR ISSG MEDI MONS NADR NASW NECS NEWZ NPPF NPTG SANT SATL SPSG
##    3   24    3    2    1    9    1   20    1    1    2    3    1    2    1    1
```

```
summary(as.factor(alb_global_dietpa$OceanBasin))
```

```
##        Indian Mediterranean    N Atlantic     N Pacific    S Atlantic
##             3             9            25            28             2
##     S Pacific
##             8
```

```
summary(as.factor(alb_global_dietpa$OceanBasinQ))
```

```
## Mediterranean   NE Atlantic    NE Pacific   NW Atlantic     NW Indian
##             9            24            25             1             1
##    NW Pacific   SE Atlantic   SW Atlantic     SW Indian    SW Pacific
##             3             1             1             2             8
```

```
write.csv(alb_global_dietpa, here("data/output_data/alb_global_wide_dietpa.csv"),
          row.names = FALSE)
```

Diet data - frequency of occurrence Here we will join our additional covariates (inc. Longhurst province data) with the diet data, and generate a wide format df with species as columns and values for reported frequency of occurrence (FO). Because these data may be used for traitglms and converted to presence/absence, we filter for very low occurrence species ($< 1\%$).

```
#Extract %FO data --> filter for only species that we have adult traits +
# cluster info for --> transform to wide data for MV GLMs + trait GLMs

alb_global_dietfo = alb_global_diet_use %>%
  dplyr::select(StudyID, rep_ID, YearEnd, OceanBasin, OceanBasinQ, province, code,
                lat_cat, pred_life_adjust, prey_sp, pFOuse) %>%
  filter(pFOuse >= 1) %>% #set this value to include or exclude species #goes down to 62 observations
  group_by(prey_sp) %>%
  #dplyr::mutate(grouped_id = row_number()) %>%
  spread(prey_sp, pFOuse, fill = 0, convert = FALSE)

#Check these data
which(rowSums(alb_global_dietfo[,10:ncol(alb_global_dietfo)])<1)
```

```
## integer(0)
```

```
#Before spreading data - range(alb_global_dietfo$pFOuse)
#0 to 100 need to exclude zeros as they are not true zeros --> 0.1 to 100%

#Check covariate representation
summary(as.factor(alb_global_dietfo$lat_cat))
```

```
## temperate  tropical
##        56         6
```

```
summary(as.factor(alb_global_dietfo$code))
```

```
## ARCH CCAL CNRY EAFR ISSG MEDI MONS NADR NASW NECS NEWZ NPPF NPTG SANT SATL SPSG
##    3   14    3    2    1    7    1   19    1    1    2    3    1    2    1    1
```

```
summary(as.factor(alb_global_dietfo$OceanBasin))
```

```
##        Indian Mediterranean    N Atlantic     N Pacific    S Atlantic
##             3             7            24            18             2
##     S Pacific
##             8
```

```
summary(as.factor(alb_global_dietfo$OceanBasinQ))
```

```
## Mediterranean   NE Atlantic    NE Pacific   NW Atlantic     NW Indian
##             7            23            15             1             1
##    NW Pacific   SE Atlantic   SW Atlantic     SW Indian    SW Pacific
##             3             1             1             2             8
```

```
write.csv(alb_global_dietfo, here("data/output_data/alb_global_wide_dietfo.csv"),
          row.names = FALSE)
```

```
# Extract for %N data
alb_global_dietn = alb_global_diet_use %>%
    dplyr::select(StudyID, rep_ID, YearEnd, OceanBasin, OceanBasinQ, province, code,
        lat_cat, pred_life_adjust, prey_sp, pN) %>%
    filter(pN > 0) %>%
    # filter(prey_sp %in% alb_adult_traits$prey_sp) %>%
group_by(prey_sp) %>%
    spread(prey_sp, pN, fill = 0, convert = FALSE)

# Check covariate representation
summary(as.factor(alb_global_dietn$lat_cat))
```

Diet data - abundance

```
## temperate  tropical
##        19         4
```

```
summary(as.factor(alb_global_dietn$code))
```

```
## CCAL EAFR ISSG MEDI MONS NADR NECS NPPF NPTG
##    8    2    1    5    1    1    1    3    1
```

```
summary(as.factor(alb_global_dietn$OceanBasin))
```

```
##        Indian Mediterranean    N Atlantic     N Pacific    S Atlantic
##             3             5             2            12             1
```

```
summary(as.factor(alb_global_dietn$OceanBasinQ))
```

```
## Mediterranean   NE Atlantic    NE Pacific     NW Indian    NW Pacific
##             5             2             9             1             3
##   SE Atlantic     SW Indian
##             1             2
```

```
write.csv(alb_global_dietn, here("data/output_data/alb_global_wide_dietn.csv"), row.names = FALSE)
```

```
# Extract for %M data range(alb_global_dietm$pM) #0 to 95.2, need to exclude
# zeros as not true zeros

alb_global_dietm = alb_global_diet %>%
    dplyr::select(StudyID, rep_ID, YearEnd, OceanBasin, OceanBasinQ, province, code,
        lat_cat, pred_life_adjust, prey_sp, pM) %>%
    filter(pM > 0) %>%
    # filter(prey_sp %in% alb_adult_traits$prey_sp) %>%
group_by(prey_sp) %>%
    spread(prey_sp, pM, fill = 0, convert = FALSE)

# Check covariate representation
summary(as.factor(alb_global_dietm$lat_cat))
```

Diet data - biomass

```
## temperate  tropical
##        28         5
```

```
summary(as.factor(alb_global_dietm$code))
```

```
## ARCH CCAL CNRY EAFR ISSG MEDI MONS NADR NECS NEWZ NPPF SANT SPSG
##    3    3    3    2    1    4    1    7    1    2    3    2    1
```

```
summary(as.factor(alb_global_dietm$OceanBasin))
```

```
##        Indian Mediterranean    N Atlantic     N Pacific    S Atlantic
##             3             4            11             6             1
##     S Pacific
##             8
```

```
summary(as.factor(alb_global_dietm$OceanBasinQ))
```

```
## Mediterranean   NE Atlantic    NE Pacific     NW Indian    NW Pacific
##             4            11             3             1             3
##   SE Atlantic     SW Indian    SW Pacific
##             1             2             8
```

```
write.csv(alb_global_dietm, here("data/output_data/alb_global_wide_dietm.csv"), row.names = FALSE)
```