Appendix 1: Data and Initial Filtering

Richard G. Stockey

March 2022

1. Data

Data Description

The data used in this study is from the Sedimentary Geochemistry and Paleoenvironments Project (SGP) Phase I product and was downloaded from the SGP website (sgp-search.io).

The API for the specific data download used in this study is:

{“type”:“nhhxrf”,“filters”:{},“show”:[“height_meters”,“section_name”,“fe”,“fe_hr_fe_t”,“fe_py_fe_hr”,“toc”, “alu”,“mo”,“u”,“fe_t_al”,“site_type”,“coord_lat”,“coord_long”,“basin_type”,“meta_bin”,“strat_name”, “strat_name_long”,“environment_bin”,“interpreted_age”,“max_age”,“min_age”,“lithology_name”]}

Without a specific API client (e.g. Postman), obtaining this data is achieved by navigating to the “Search” page of the SGP website and selecting the following options under their respective headers:

Type: No HHXRF. Show: Fe (wt%), FeHR/FeT, Fe-py/FeHR, FeT/Al, TOC (wt%), Al (wt%), Mo (ppm), U (ppm), height/depth, section name, site type, site latitude, site longitude, basin type, metamorphic bin, stratigraphy name, long stratigraphy name, environmental bin, lithology name, interpreted age, max age, min age.

The specific download date of this data was December 12, 2022. However, all data downloaded from the Phase I data product is identical regardless of download date (Farrell et al. 2021, Geobiology).

As all data downloaded from the SGP website is downloaded as “SGP.csv”, the download file used here is renamed as “SGP_iron_traces_no_hhxrf_20211210.csv”.

Data Import

The .csv filed downloaded from the SGP website is imported, assuming that the .csv file is in the current working directory (change current working directory if necessary using the setwd() function).

```
SGP.data <- read.csv("SGP_iron_traces_no_hhxrf_20230203.csv")
```

Following the data import, we can look at how many samples there are in the entire SGP Phase I data product.

```
nrow(SGP.data)
```

```
## [1] 82578
```

Initial data filtering

We apply filtering steps that are necessary to restrict our dataset to samples that will be useful for answering the research questions investigated in this study. In this section we apply these filtering steps and calculate the number of samples remaining after each filtering step. We call this filtered dataset “trace.toc.full”.

Load dplyr package for filtering.

```
library(dplyr)
```

Remove samples with no interpreted age. After each filtering step we calculate the number of remaining samples in our dataset.

```
trace.toc.full <- filter(SGP.data, !is.na(interpreted.age))
nrow(trace.toc.full)
```

```
## [1] 76643
```

Remove samples outside of the age range of our analyses (300-1000 Ma).

```
trace.toc.full <- filter(trace.toc.full, (interpreted.age <= 1000 & interpreted.ag
e >= 300))
nrow(trace.toc.full)
```

```
## [1] 33742
```

Only include shale and other fine-grained lithologies (and samples assigned no lithology, at least for now).

```
trace.toc.full <- filter(trace.toc.full, (lithology.name == "argillite"
                                          | lithology.name =="clay"
                                          | lithology.name == "claystone"
                                          | lithology.name == "dolomudstone"
                                          | lithology.name == "lime mudstone"
                                          | lithology.name == "meta-argillite"
                                          | lithology.name == "metapelite"
                                          | lithology.name == "metasiltstone"
                                          | lithology.name == "mud"
                                          | lithology.name == "mudstone"
                                          | lithology.name == "oil shale"
                                          | lithology.name == "pelite"
                                          | lithology.name == "phosphorite"
                                          | lithology.name == "shale"
                                          | lithology.name == "silt"
                                          | lithology.name == "siltite"
                                          | lithology.name == "siltstone"
                                          | lithology.name == "slate"
                                          | lithology.name == ""))
nrow(trace.toc.full)
```

```
## [1] 25684
```

Remove fluvial samples (keep only marine environmental bins).

```
trace.toc.full <- filter(trace.toc.full, (environmental.bin == "basinal"
                                          | environmental.bin == "inner shelf"
                                          | environmental.bin == "outer shelf"
                                          | environmental.bin == ""))
nrow(trace.toc.full)
```

```
## [1] 25681
```

Remove samples with Mo and U so high that they are ore grade metalliferous rocks (using a cutoff of 1000ppm for extremely high concentrations) - notably the vast majority of these are from USGS studies.

```
trace.toc.full <- filter(trace.toc.full, (Mo..ppm. < 1000 | is.na(Mo..ppm.)))
trace.toc.full <- filter(trace.toc.full, (U..ppm. < 1000 | is.na(U..ppm.)))

nrow(trace.toc.full)
```

```
## [1] 25600
```

Save filtered dataset for further filtering in spatiotemporal bootstrap and random forest scripts.

```
save(trace.toc.full, file = "Filtered.trace.toc.full.20230205.RData")
```

Data structure

We will view the general structure of the dataset with respect to geologic context variables in stacked histograms.

Load ggplot2 and deeptime packages for time series plotting.

```
library(ggplot2)
library(deeptime)
```

We adjust the date of the Tonian-Cryogenian boundary from the deeptime default age of 850Ma to the updated age of 720Ma (following the International Commission on Stratigraphy 2015 onwards and the Geological Society of America Geologic Timescale v5.0).

```
periods.edit <- deeptime::periods
periods.edit[14,2] <- 720
periods.edit[15,3] <- 720
```

Compile a composite figure illustrating the number of samples in the primary filtered dataset and the proportions of different lithologies, site types, basin types, environmental bins and metamorphic bins through the interval of study. The higher percentage of unknown context information in the Paleozoic relates to data from the USGS NGDB and CMIBS databases that were incorporated into SGP. These datasets lack the geological context information commonly coded by SGP team members.

For plotting, we assign missing categorical data NA values in a new dataframe specifically for this plot.

```
trace.toc.full.hist <- trace.toc.full

trace.toc.full.hist$lithology.name[trace.toc.full.hist$lithology.name == ""] <- "n
o data"
trace.toc.full.hist$site.type[trace.toc.full.hist$site.type == ""] <- "no data"
trace.toc.full.hist$basin.type[trace.toc.full.hist$basin.type == ""] <- "no data"
trace.toc.full.hist$environmental.bin[trace.toc.full.hist$environmental.bin == ""]
<- "no data"
trace.toc.full.hist$metamorphic.bin[trace.toc.full.hist$metamorphic.bin == ""] <-
"no data"
```

Order factors for plotting (with “no data” coming last for all cases…)

```
trace.toc.full.hist$lithology.name <- factor(trace.toc.full.hist$lithology.name, l
evels=c("argillite",
  "clay",
  "claystone",
  "dolomudstone",
  "lime mudstone",
  "meta-argillite",
  "metapelite",
  "metasiltstone",
  "mud",
  "mudstone",
  "oil shale",
  "pelite",
  "phosphorite",
  "shale",
  "silt",
  "siltite",
  "siltstone",
  "slate",
  "no data"))

trace.toc.full.hist$site.type <- factor(trace.toc.full.hist$site.type, levels=c(
  "core",
  "cuttings",
  "outcrop",
  "no data"
  ))

trace.toc.full.hist$basin.type <- factor(trace.toc.full.hist$basin.type, levels=c(
  "back-arc",
  "fore-arc",
  "intracratonic sag",
  "passive margin",
  "peripheral foreland",
  "retro-arc foreland",
  "rift",
  "wrench",
  "no data"
  ))
```

```
trace.toc.full.hist$environmental.bin <- factor(trace.toc.full.hist$environmental.
bin, levels=c(
  "basinal",
  "inner shelf",
  "outer shelf",
  "no data"
  ))

trace.toc.full.hist$metamorphic.bin <- factor(trace.toc.full.hist$metamorphic.bin,
levels=c(
  "Anchizone",
  "Diagenetic zone",
  "Epizone",
  "no data"
  ))
```

Rename metamorphic zone factor levels so lowercase is used for all legends.

```
levels(trace.toc.full.hist$metamorphic.bin) <- c(
  "anchizone",
  "diagenetic zone",
  "epizone",
  "no data"
  )
```

Update Neoproterozoic colours in deeptime.

```
periods.edit$color[13:15] <- c("#FED96A", "#FECC5C", "#FEBF4E")
```

Generate panel of histograms showing the lithologies, site types, basin types, environmental bins and metamorphic bins for the samples in our primary dataset.

```
lithology <- ggplot(trace.toc.full.hist, aes(x=interpreted.age))+
  geom_histogram(aes(fill=lithology.name), alpha=1, binwidth=25)+
  theme_bw()+
  ylab("Samples")+xlab("Time (Ma)")+
  scale_fill_manual(name = "Lithology", values = c("#99B575",
                                                   "#B3E095",
                                                   "#1A9D6F",
                                                   "#E5B75A",
                                                   "#F1E19D",
                                                   "#E36350",
                                                   "#FB9A85",
                                                   "#B051A5",
                                                   "#E3B9DB",
                                                   "#4EB3D3",
                                                   "#B3E3EE",
                                                   "#67A599",
                                                   "#BFD0C5",
                                                   "grey73"
                                                   ))+
  scale_x_reverse()+
  coord_geo(xlim=c(1001, 300), ylim=c(0,4300), expand=FALSE, # Geologic timescale
added for clarity
            pos = as.list(rep("bottom", 1)),
            abbrv=list(T),
            dat = list(periods.edit),
            height = list(unit(2, "lines")),
            size=list(7),
            bord=list(c("left", "bottom", "right")), lwd=as.list(1))+
  theme(panel.border = element_rect(fill=NA,color="black", size=2,linetype="solid"
),
        axis.ticks = element_line(size=1.1),
        axis.title = element_text(size=40),
        axis.text =  element_text(size=30, colour="black"),
        plot.margin = ggplot2::margin(10,30,10,10), # use ggplot2:: in case random
Forest package enabled
        legend.position=c(0.15, 0.62),
        legend.text = element_text(size=22),
        legend.title = element_text(size=26),
        axis.title.y = element_text(margin = ggplot2::margin(t = 0, r = 20, b = 0,
l = 0)),
        axis.title.x = element_text(margin = ggplot2::margin(t = 10, r = 0, b = 0,
l = 0)),
        panel.grid.major = element_blank(),panel.grid.minor = element_blank())
```

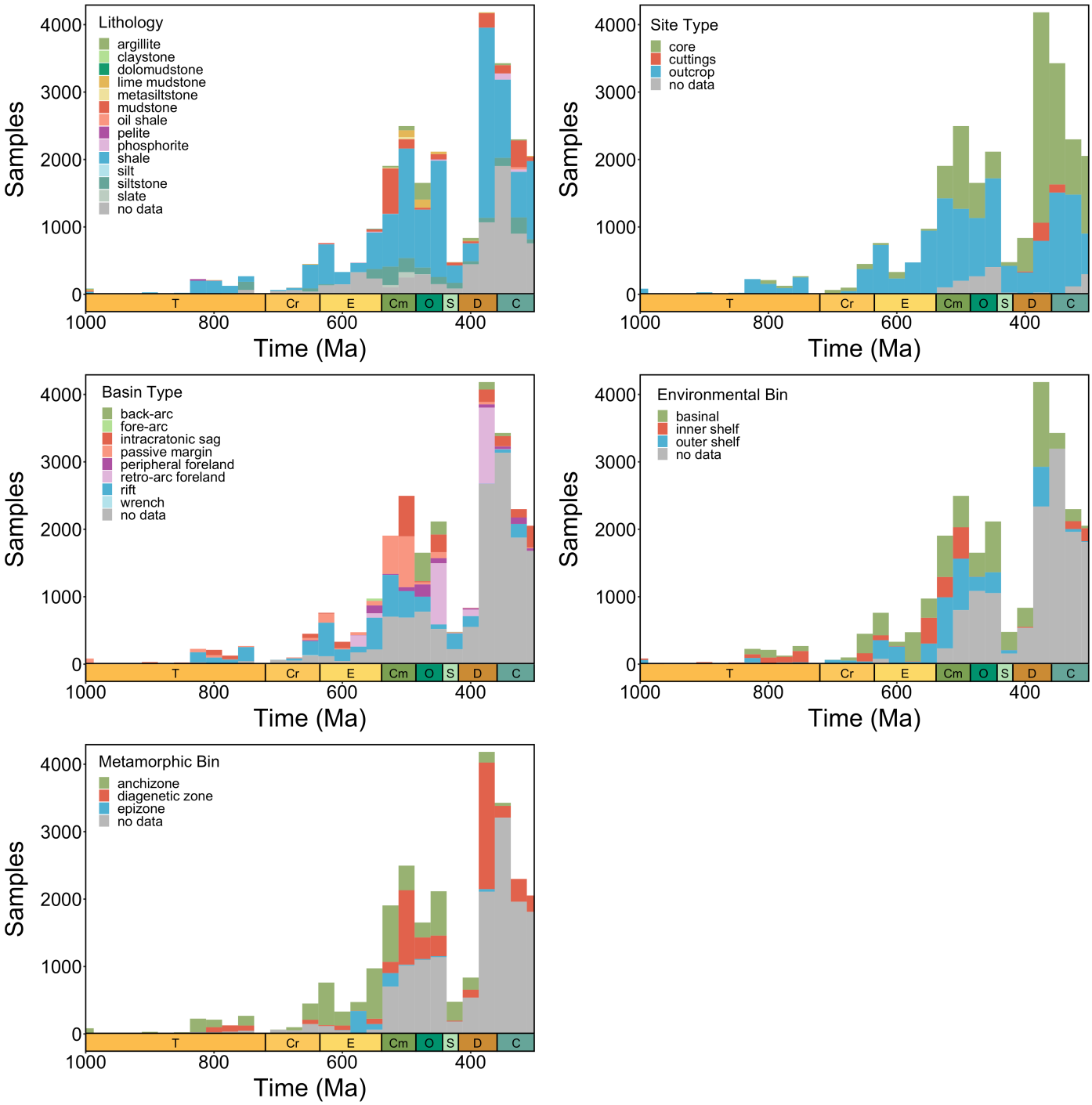
```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.
4.0.
## ℹ Please use the `linewidth` argument instead.
```

```
## Warning: The `size` argument of `element_rect()` is deprecated as of ggplot2 3.
4.0.
## ℹ Please use the `linewidth` argument instead.
```

```
site.type <- ggplot(trace.toc.full.hist, aes(x=interpreted.age))+
  geom_histogram(aes(fill=site.type), alpha=1, binwidth=25)+
  theme_bw()+
  ylab("Samples")+xlab("Time (Ma)")+
  scale_fill_manual(name = "Site Type", values = c("#99B575",
                                                   "#E36350",
                                                   "#4EB3D3",
                                                   "grey73"
                                                   ))+
  scale_x_reverse()+
  coord_geo(xlim=c(1001, 300), ylim=c(0,4300), expand=FALSE, # Geologic timescale
added for clarity
            pos = as.list(rep("bottom", 1)),
            abbrv=list(T),
            dat = list(periods.edit),
            height = list(unit(2, "lines")),
            size=list(7),
            bord=list(c("left", "bottom", "right")), lwd=as.list(1))+
  theme(panel.border = element_rect(fill=NA,color="black", size=2,linetype="solid"
),
        axis.ticks = element_line(size=1.1),
        axis.title = element_text(size=40),
        axis.text =  element_text(size=30, colour="black"),
        plot.margin = ggplot2::margin(10,30,10,10), # use ggplot2:: in case random
Forest package enabled
        legend.position=c(0.10, 0.83),
        legend.text = element_text(size=22),
        legend.title = element_text(size=26),
        axis.title.y = element_text(margin = ggplot2::margin(t = 0, r = 20, b = 0,
l = 0)),
        axis.title.x = element_text(margin = ggplot2::margin(t = 10, r = 0, b = 0,
l = 0)),
        panel.grid.major = element_blank(),panel.grid.minor = element_blank())

basin.type <- ggplot(trace.toc.full.hist, aes(x=interpreted.age))+
  geom_histogram(aes(fill=basin.type), alpha=1, binwidth=25)+
  theme_bw()+
  ylab("Samples")+xlab("Time (Ma)")+
  scale_fill_manual(name = "Basin Type",values = c("#99B575",
                                                   "#B3E095",
                                                   "#E36350",
                                                   "#FB9A85",
                                                   "#B051A5",
                                                   "#E3B9DB",
                                                   "#4EB3D3",
                                                   "#B3E3EE",
                                                   "grey73"
                                                   ))+
  scale_x_reverse()+
  coord_geo(xlim=c(1001, 300), ylim=c(0,4300), expand=FALSE, # Geologic timescale
added for clarity
            pos = as.list(rep("bottom", 1)),
            abbrv=list(T),
```

```
            dat = list(periods.edit),
            height = list(unit(2, "lines")),
            size=list(7),
            bord=list(c("left", "bottom", "right")), lwd=as.list(1))+
  theme(panel.border = element_rect(fill=NA,color="black", size=2,linetype="solid"
),
        axis.ticks = element_line(size=1.1),
        axis.title = element_text(size=40),
        axis.text =  element_text(size=30, colour="black"),
        plot.margin = ggplot2::margin(10,30,10,10), # use ggplot2:: in case random
Forest package enabled
        legend.position=c(0.185, 0.73),
        legend.text = element_text(size=22),
        legend.title = element_text(size=26),
        axis.title.y = element_text(margin = ggplot2::margin(t = 0, r = 20, b = 0,
l = 0)),
        axis.title.x = element_text(margin = ggplot2::margin(t = 10, r = 0, b = 0,
l = 0)),
        panel.grid.major = element_blank(),panel.grid.minor = element_blank())

environmental.bin <- ggplot(trace.toc.full.hist, aes(x=interpreted.age))+
  geom_histogram(aes(fill=environmental.bin), alpha=1, binwidth=25)+
  theme_bw()+
  ylab("Samples")+xlab("Time (Ma)")+
  scale_fill_manual(name = "Environmental Bin", values = c("#99B575",
                                                           "#E36350",
                                                           "#4EB3D3",
                                                           "grey73"
                                                           ))+
  scale_x_reverse()+
  coord_geo(xlim=c(1001, 300), ylim=c(0,4300), expand=FALSE, # Geologic timescale
added for clarity
            pos = as.list(rep("bottom", 1)),
            abbrv=list(T),
            dat = list(periods.edit),
            height = list(unit(2, "lines")),
            size=list(7),
            bord=list(c("left", "bottom", "right")), lwd=as.list(1))+
  theme(panel.border = element_rect(fill=NA,color="black", size=2,linetype="solid"
),
        axis.ticks = element_line(size=1.1),
        axis.title = element_text(size=40),
        axis.text =  element_text(size=30, colour="black"),
        plot.margin = ggplot2::margin(10,30,10,10), # use ggplot2:: in case random
Forest package enabled
        legend.position=c(0.185, 0.83),
        legend.text = element_text(size=22),
        legend.title = element_text(size=26),
        axis.title.y = element_text(margin = ggplot2::margin(t = 0, r = 20, b = 0,
l = 0)),
        axis.title.x = element_text(margin = ggplot2::margin(t = 10, r = 0, b = 0,
l = 0)),
        panel.grid.major = element_blank(),panel.grid.minor = element_blank())
```

```
metamorphic.bin <- ggplot(trace.toc.full.hist, aes(x=interpreted.age))+
  geom_histogram(aes(fill=metamorphic.bin), alpha=1, binwidth=25)+
  theme_bw()+
  ylab("Samples")+xlab("Time (Ma)")+
  scale_fill_manual(name = "Metamorphic Bin", values = c("#99B575",
                                                         "#E36350",
                                                         "#4EB3D3",
                                                         "grey73"
                                                         ))+
  scale_x_reverse()+
  coord_geo(xlim=c(1001, 300), ylim=c(0,4300), expand=FALSE, # Geologic timescale
added for clarity
            pos = as.list(rep("bottom", 1)),
            abbrv=list(T),
            dat = list(periods.edit),
            height = list(unit(2, "lines")),
            size=list(7),
            bord=list(c("left", "bottom", "right")), lwd=as.list(1))+
  theme(panel.border = element_rect(fill=NA,color="black", size=2,linetype="solid"
),
        axis.ticks = element_line(size=1.1),
        axis.title = element_text(size=40),
        axis.text =  element_text(size=30, colour="black"),
        plot.margin = ggplot2::margin(10,30,10,10), # use ggplot2:: in case random
Forest package enabled
        legend.position=c(0.165, 0.84),
        legend.text = element_text(size=22),
        legend.title = element_text(size=26),
        axis.title.y = element_text(margin = ggplot2::margin(t = 0, r = 20, b = 0,
l = 0)),
        axis.title.x = element_text(margin = ggplot2::margin(t = 10, r = 0, b = 0,
l = 0)),
        panel.grid.major = element_blank(),panel.grid.minor = element_blank())

context.hists <- ggarrange2(lithology, site.type, basin.type, environmental.bin, m
etamorphic.bin, ncol=2)
```




```
ggsave("Figure Sx Histograms of geologic context data for primary dataset.pdf", co
ntext.hists, height=25, width=23)
```

We will also generate a map of samples, color-coded by geologic age.

Load packages for map plotting

```
library(maps)
library(ggthemes)
```

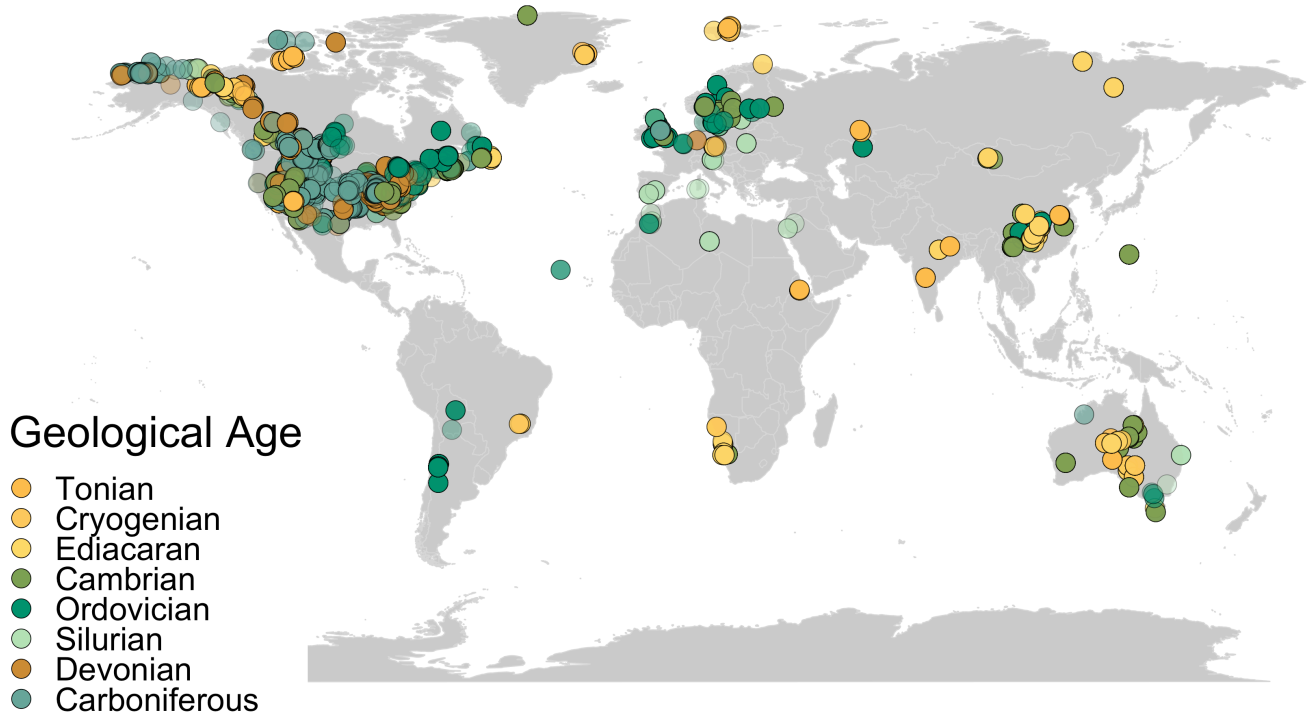
Before generating the map figure, we need to categorize the interpreted ages of our samples into geologic ages. To do this we generate a duplicate dataframe specifically for this plot and add an additional column for geologic age. Ages used here are from the Geological Society of America Geologic Timescale v5.0.

```
trace.toc.full.map.data <- trace.toc.full
trace.toc.full.map.data$Geological.Age <- NA # initiate geologic age column with a
ll NAs

# Assign samples ages based on interpreted age
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
1000
                                       & trace.toc.full.map.data$interpreted.age >
720] <- "Tonian"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
720
                                       & trace.toc.full.map.data$interpreted.age >
635] <- "Cryogenian"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
635
                                       & trace.toc.full.map.data$interpreted.age >
541] <- "Ediacaran"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
541
                                       & trace.toc.full.map.data$interpreted.age >
485.4] <- "Cambrian"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
485.4
                                       & trace.toc.full.map.data$interpreted.age >
443.8] <- "Ordovician"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
443.8
                                       & trace.toc.full.map.data$interpreted.age >
419.2] <- "Silurian"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
419.2
                                       & trace.toc.full.map.data$interpreted.age >
358.9] <- "Devonian"
trace.toc.full.map.data$Geological.Age[trace.toc.full.map.data$interpreted.age <=
358.9
                                       & trace.toc.full.map.data$interpreted.age >
298.9] <- "Carboniferous"

# Make Geological.Age a factor and relevel for plotting
trace.toc.full.map.data$Geological.Age <- factor(trace.toc.full.map.data$Geologica
l.Age,
                                                 levels=c("Tonian",
                                                          "Cryogenian",
                                                          "Ediacaran",
                                                          "Cambrian",
                                                          "Ordovician",
                                                          "Silurian",
                                                          "Devonian",
                                                          "Carboniferous"
                                                          ))
```

Generate map figure. The 269 rows that are flagged as removed in the warning message are 269 samples without longitude and latitude data.

```
map <- ggplot() +
  borders("world", colour = "gray85", fill = "gray80") +
  theme_map()  +
  geom_point(aes(x = site.longitude, y = site.latitude,
                 fill = Geological.Age, group=interpreted.age),
             data = trace.toc.full.map.data,
             alpha = .4,
             shape=21, size=8) +
  scale_fill_manual(values = c( rgb(254, 191, 78, maxColorValue = 255),
                                rgb(254, 204, 92, maxColorValue = 255),
                                rgb(254, 217, 106, maxColorValue = 255),
                                rgb(127, 160, 86, maxColorValue = 255),
                                rgb(0, 146, 112, maxColorValue = 255),
                                rgb(179, 225, 182, maxColorValue = 255),
                                rgb(203, 140, 55, maxColorValue = 255),
                                rgb(103, 165, 153, maxColorValue = 255)

  ))+
  labs(fill = 'Geological Age')+
  theme(plot.margin = ggplot2::margin(1,1,1,1,"cm"),
        panel.border = element_rect(fill=NA,color=NA, size=2,linetype="solid"),
        legend.title = element_text(size=40),
        legend.text = element_text(size=30))+
  guides(fill = guide_legend(override.aes = list(alpha = 1)))

map
```

```
## Warning: Removed 269 rows containing missing values (`geom_point()`).
```



```
ggsave("Figure Sx Map of samples in primary dataset color-coded by geologic age 22 0230403.pdf", map, height=10.3, width=18.3)
```

```
## Warning: Removed 269 rows containing missing values (`geom_point()`).
```