文献复现

AUTHOR
max

PUBLISHED
May 23, 2025

分组数据分析实战

在这一部分，将以论文中的数据分析为例，展示分组数据分析和可视化的重复性研究。首先，我们简单介绍一下论文的研究背景、方法和主要结果。然后，使用原始数据进行可重复研究，通过复现论文中的图片，展示分组数据分析和可视化的重复性研究。

论文研究概述

研究背景

甲烷氧化与微生物群落： **厌氧甲烷氧化（AOM）**是一种由微生物驱动的共生过程，在全球海洋沉积物中广泛存在，通过将甲烷氧化与硫酸盐还原耦合来调节甲烷通量并促进自生矿物的产生，包括碳酸盐和硫化铁。 AOM通常由**厌氧甲烷氧化古菌（ANME）**和**硫酸盐还原细菌（SRB）**组成的多细胞群落介导。 硅质矿物的生物沉淀： 在甲烷渗漏沉积物中，ANME-SRB群落与粘土状硅酸盐矿物密切相关，但这些矿物的生物成因、地球化学组成及其在岩石记录中的保存潜力尚不清楚。 长期实验室培养的AOM富集培养物在硅不饱和的介质中产生了无定形硅酸盐颗粒，表明存在微生物介导的沉淀过程。

主要技术方法

样品采集与处理： 从加利福尼亚北部和南部以及哥斯达黎加边缘的三个海底甲烷渗漏点采集沉积物和自生碳酸盐样品。 使用密度梯度离心法从沉积物中分离ANME-SRB群落，并对其进行固定和染色处理以便后续显微镜观察。 显微镜观察与分析： 使用相关荧光原位杂交（FISH）、扫描电子显微镜（SEM）结合能量色散X射线光谱（EDS）以及纳米级二次离子质谱（nanoSIMS）等技术对ANME-SRB群落进行观察和分析。 通过透射电子显微镜（TEM）进一步观察ANME-SRB群落内部的硅质颗粒分布和形态。 地球化学分析： 使用电感耦合等离子体质谱（ICP-MS）测定培养基和沉积物样品中的硅浓度。 构建矿物稳定性图以预测在实验溶液组成下可能形成的硅酸盐矿物类型。

重要成果

ANME-SRB群落与硅质矿物的关联： 在甲烷渗漏沉积物、自生碳酸盐以及无沉积物的AOM富集培养物中均观察到ANME-SRB群落外被富含硅质的相所包裹。这些硅质颗粒形态上与在热泉蓝藻中观察到的无定形二氧化硅球体相似。 硅质矿物的生物成因证据： 在硅不饱和的介质中，**ANME-SRB群落**能够诱导产生新的硅质矿物相，这表明存在微生物介导的沉淀过程。 化学分析显示，与**ANME-SRB群落**相关的硅质相与沉积物中的硅质矿物在组成上存在差异，进一步支持了其生物成因的观点。 对微生物化石保存的意义： 硅质矿物的包裹可能增强了**ANME-SRB群落**在化石渗漏环境中的保存潜力，类似于在其他微生物生态系统中早期二氧化硅沉淀所提供的保护作用。 提出了在古渗漏碳酸盐中寻找富含硅质的环作为识别化石**ANME-SRB群落**的搜索图像的可能性。

结论

该研究首次证明 ANME-SRB 聚集体可在低硅浓度下主动诱导富硅相沉淀，揭示了一种新的微生物矿化机制。硅质包裹可能是微生物化石保存的关键因素，为古环境微生物化石识别和地外生命探索提供了矿物学

依据。未来研究可聚焦于具体生物化学途径（如 EPS 成分、膜表面电荷）及地质时间尺度下的保存效应。

数据准备

论文的原始数据及分析代码都在 GitHub 上。首先，使用 Git 命令将代码克隆到本地：

```
git clone https://github.com/daniosro/Si_biomineralization_ANME_SRB.git --depth 1
```

加载必要的R包：

readxl：专门用于读取Excel文件，比常规方法更稳定 tidyverse：包含ggplot2(绘图)、dplyr(数据处理)等核心包 注意：使用library()而不是require()确保加载失败时报错

```
#| label: packages
library(readxl)       # 用于读取Excel文件
library(tidyverse)    # 包含ggplot2、dplyr等数据科学工具包
```

```
Warning: package 'ggplot2' was built under R version 4.4.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.4     ✓ tidyr     1.3.1
✓ purrr     1.0.4
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
## 设置ggplot2默认主题为黑白主题(theme_bw)
## 影响后续所有ggplot图形的外观：

theme_set(theme_bw())
```

```
# 读取数据集S3.xlsx文件，包含(Mg+Al+Fe)/Si比值数据
# 数据包括无沉积物的ANME-SRB联合体、沉积物中的ANME-SRB联合体以及不含ANME-SRB联合体的沉积物
# 使用xfun包的magic_path智能定位文件路径
# 优点：避免硬编码路径，增强代码可移植性

file = xfun::magic_path("Dataset S3.xlsx")  # 自动查找文件路径

# 读取Excel数据：
# 自动识别列类型
# 保留原始列名
# 存储在octtet_data数据框中
```

```
octtet_data <- read_excel(file)  # 读取Excel文件

# 打印数据框结构：
# - 检查列名是否正确
# - 查看前几行数据
# - 确认数据类型
octtet_data
```

```
# A tibble: 425 × 5
   mg_al_fe_to_si Source                          Basin source_order basin_order
            <dbl> <chr>                           <chr>        <dbl>       <dbl>
 1           0.07 Aggregate-attached, Sediment-f… Sant…            3           3
 2           0.35 Aggregate-attached, Sediment-f… Sant…            3           3
 3           0.29 Aggregate-attached, Sediment-f… Sant…            3           3
 4           0.29 Aggregate-attached, Sediment-f… Sant…            3           3
 5           0.04 Aggregate-attached, Sediment-f… Sant…            3           3
 6           0.41 Aggregate-attached, Sediment-f… Sant…            3           3
 7           0.4  Aggregate-attached, Sediment-f… Sant…            3           3
 8           0.12 Aggregate-attached, Sediment-f… Sant…            3           3
 9           0.34 Aggregate-attached, Sediment-f… Sant…            3           3
10           0.19 Aggregate-attached, Sediment-f… Sant…            3           3
# ℹ 415 more rows
```

```
# 按类别筛选数据
# 使用dplyr管道操作(%>%)进行数据清洗：
# 1. 从Jaco Scar获取的沉积物和ANME-SRB联合体附着的硅酸盐
octtet_data_Jaco <- octtet_data |>
  filter(Basin == "Jaco Scar") |>  # 筛选特定盆地
  select( # 选择需要的列：
    mg_al_fe_to_si,  # (Mg+Al+Fe)/Si比值
    Source,          # 样品来源
    Basin,           # 盆地名称
    source_order,    # 来源排序指标
    basin_order      # 盆地排序指标
  ) |>
  mutate(Source = as_factor(Source))  # 将来源转为因子，便于分组分析

# Case 2-6: 其他盆地的类似处理
# (代码结构相同，仅筛选条件不同)

# 2. 从Santa Monica盆地获取的沉积物和ANME-SRB联合体附着的硅酸盐
octtet_data_SMB <- subset(octtet_data, Basin == "Santa Monica",
                          select = c("mg_al_fe_to_si","Source","Basin","source_order","basin_or

# 3. 从Santa Monica盆地培养实验中无沉积物的ANME-SRB联合体附着的富含硅相
octtet_data_SMB_sedfree <- subset(octtet_data_SMB,
                                  Source == "Aggregate-attached, Sediment-free" | Source == "Se
                                  select = c("mg_al_fe_to_si","Source","Basin","source_order","

# 4. Santa Monica盆地沉积物中ANME-SRB联合体附着的硅酸盐
octtet_data_SMB_fromsed <- subset(octtet_data_SMB,
                                  Source == "Aggregate-attached" | Source == "Sediment",
                                  select = c("mg_al_fe_to_si","Source","Basin","source_order","
```

```r
# 5. Santa Monica盆地沉积物中ANME-SRB联合体附着的硅酸盐与培养实验中无沉积物的ANME-SRB联合体附着
octtet_data_SMB_freevsfromsed<- subset(octtet_data_SMB,
                                       Source == "Aggregate-attached" | Source == "Aggregate-at
                                       select = c("mg_al_fe_to_si","Source","Basin","source_ord

# 6. Eel River盆地的沉积物和ANME-SRB联合体附着的硅酸盐
octtet_data_ERB <- subset(octtet_data, Basin == "Eel River",
                          select = c("mg_al_fe_to_si","Source","Basin","source_order","basin_o
```

```r
# 对Jaco Scar沉积物和ANME-SRB联合体附着的硅酸盐进行单因素方差分析
# 使用aov()进行单因素方差分析：
# 模型公式：mg_al_fe_to_si ~ Source
# 即检验不同来源样品的元素比值是否存在显著差异
resot1.aov <- aov(mg_al_fe_to_si ~ Source, data = octtet_data_Jaco)  # 执行ANOVA

# 结果摘要输出：
# - Df: 自由度
# - Sum Sq: 平方和
# - Mean Sq: 均方
# - F value: F统计量
# - Pr(>F): p值
summary(resot1.aov)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Source        1  2.242  2.2416   34.29 5.95e-08 ***
Residuals   101  6.602  0.0654
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 对其他分组重复相同分析流程

# 对Santa Monica盆地培养实验中无沉积物的ANME-SRB联合体附着的富含硅相进行单因素方差分析
resot2.aov <- aov(mg_al_fe_to_si ~ Source, data = octtet_data_SMB_sedfree)
summary(resot2.aov)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Source        1   9.34   9.337   41.29 1.23e-09 ***
Residuals   173  39.12   0.226
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 对Santa Monica盆地沉积物中ANME-SRB联合体附着的硅酸盐进行单因素方差分析
resot3.aov <- aov(mg_al_fe_to_si ~ Source, data = octtet_data_SMB_fromsed)
summary(resot3.aov)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Source        1   5.22   5.218   18.62 2.67e-05 ***
Residuals   174  48.77   0.280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 对Santa Monica盆地沉积物中ANME-SRB联合体附着的硅酸盐与培养实验中无沉积物的ANME-SRB联合体附着的
resot4.aov <- aov(mg_al_fe_to_si ~ Source, data = octtet_data_SMB_freevsfromsed)
summary(resot4.aov)
```

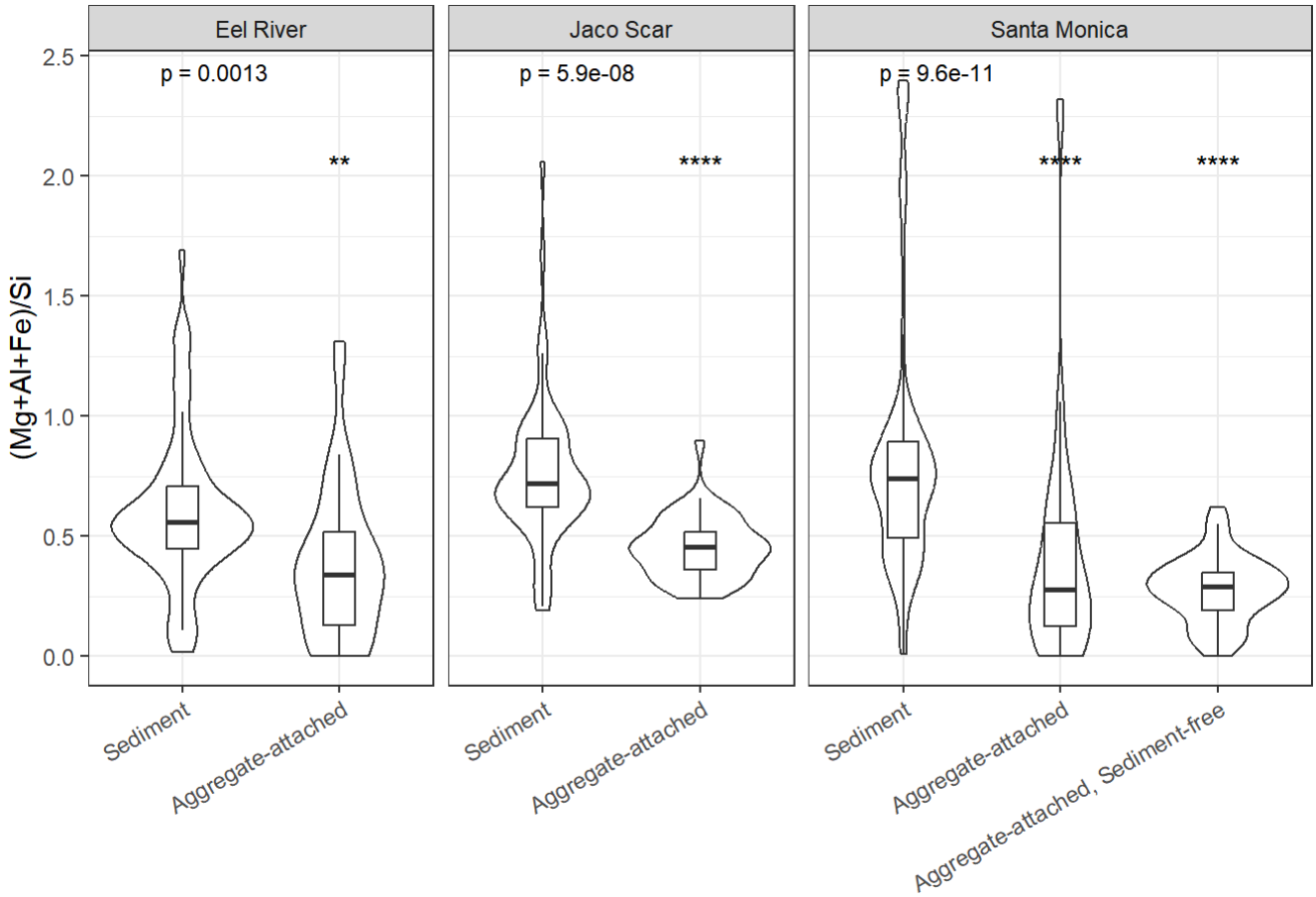
```
            Df Sum Sq Mean Sq F value Pr(>F)
Source       1  0.415  0.4148       3 0.0871 .
Residuals   81 11.200  0.1383
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 对Eel River盆地的沉积物和ANME-SRB联合体附着的硅酸盐进行单因素方差分析
resot5.aov <- aov(mg_al_fe_to_si ~ Source, data = octtet_data_ERB)
summary(resot5.aov)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
Source        1  1.126  1.1259   10.95 0.00129 **
Residuals   103 10.590  0.1028
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 加载ggpubr包用于统计可视化
# 提供专业出版级图形功能
library(ggpubr)

# 数据重组：
# - 按原始定义的顺序重新排列因子水平
# - 确保图表中分组显示顺序正确
octtet_data <- octtet_data |>
  mutate(
    new_source_order = reorder(Source, source_order),
    new_basin_order = reorder(Basin, basin_order)
  )

# 创建小提琴图
# 构建ggplot图形对象：
# - x轴：样品来源(按预定顺序)
# - y轴：(Mg+Al+Fe)/Si比值
ot = ggplot(octtet_data, aes(new_source_order, mg_al_fe_to_si))  # 设置基础图形

# 图形叠加和美化
ot +
  geom_violin() +  # 添加小提琴图
  geom_boxplot(width=0.2, outliers = FALSE) +  # 添加窄箱线图，不显示异常值



  # 统计标注：
  stat_compare_means(method = "aov", label = "p", size = 3) +  # 添加ANOVA p值
  stat_compare_means(
    method = 't.test',
    ref.group = "Sediment",  # 以沉积物组为参照
```

```
    label = "p.signif",      # 显示*号标记显著性
    label.y = 2              # 标注位置调整
  ) +

  # 分面显示：
  facet_grid(~new_basin_order, scales = "free", space = "free") +  # 按盆地分面，自由调整比例和
  # 分面显示：
  theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +  # 调整x轴标签角度
  labs(x = "", y = "(Mg+Al+Fe)/Si")  # 设置轴标签
```




```
# 读取Dataset S2.xlsx文件
# 使用xfun::magic_path()自动定位文件路径，避免绝对路径依赖
AlSi_data <- read_excel(xfun::magic_path('Dataset S2.xlsx'))
AlSi_data
```

```
# A tibble: 425 × 5
   Al.per.Si Source                          Basin `Source order` `Basin order`
       <dbl> <chr>                           <chr>          <dbl>         <dbl>
 1    0      Aggregate-attached, Sediment-fr… Sant…              3             3
 2    0.196  Aggregate-attached, Sediment-fr… Sant…              3             3
 3    0.181  Aggregate-attached, Sediment-fr… Sant…              3             3
 4    0.182  Aggregate-attached, Sediment-fr… Sant…              3             3
 5    0.0285 Aggregate-attached, Sediment-fr… Sant…              3             3
 6    0.315  Aggregate-attached, Sediment-fr… Sant…              3             3
 7    0.238  Aggregate-attached, Sediment-fr… Sant…              3             3
```

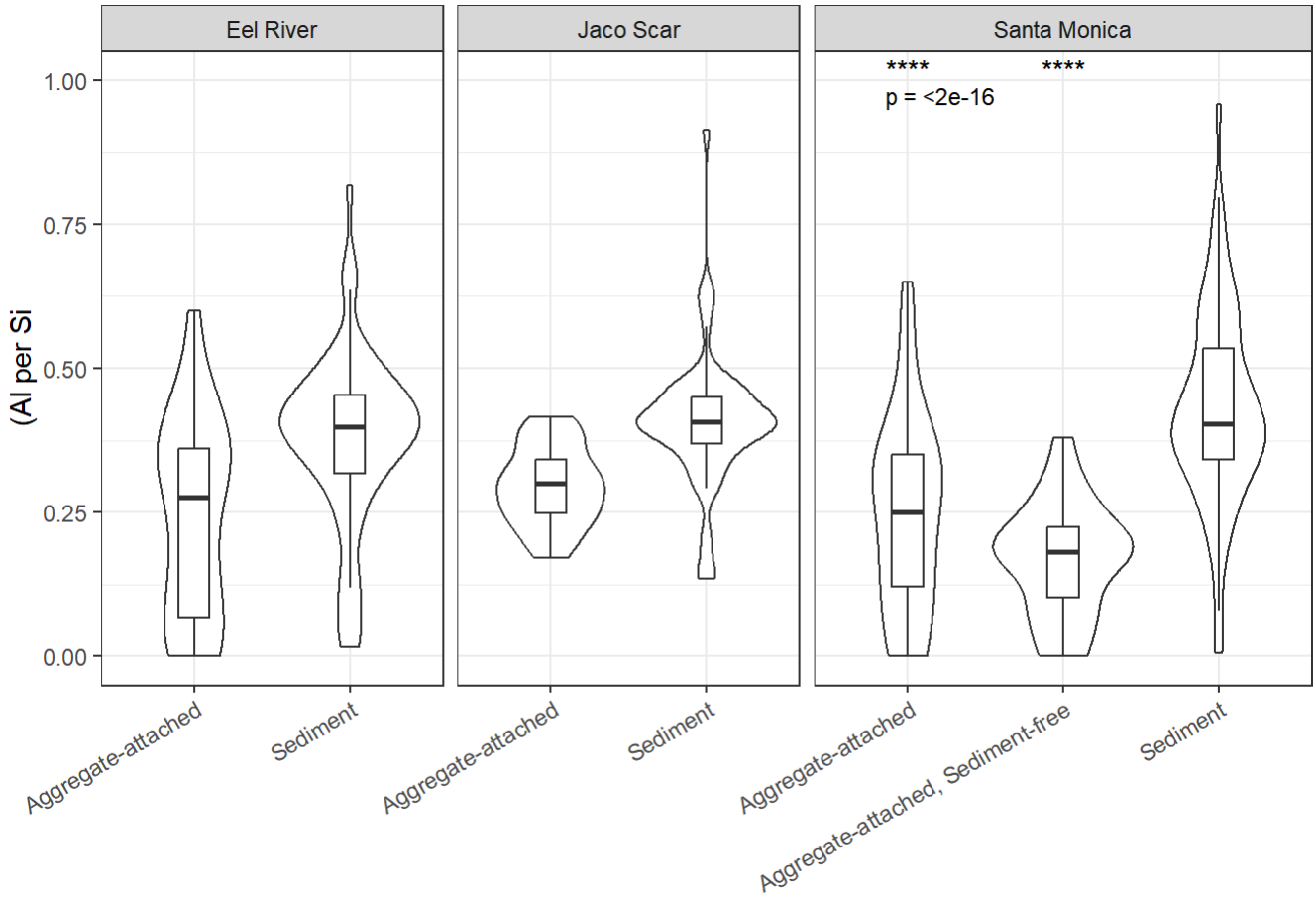
```
 8    0.0657 Aggregate-attached, Sediment-fr… Sant…              3             3
 9    0.194  Aggregate-attached, Sediment-fr… Sant…              3             3
10    0.146  Aggregate-attached, Sediment-fr… Sant…              3             3
# ℹ 415 more rows
```

```r
# 按类别筛选
#来自雅可疤的沉积物和附着在ANME-SRB共生体上的硅酸盐
AlSi_data_Jaco <- subset(AlSi_data, Basin == "Jaco Scar", select = c("Al.per.Si","Source","Bas

#来自圣莫尼卡盆地的沉积物和附着在ANME-SRB共生体上的硅酸盐
AlSi_data_SMB <- subset(AlSi_data, Basin == "Santa Monica", select = c("Al.per.Si","Source","B

#来自圣莫尼卡盆地在培养中附着在无沉积物ANME-SRB共生体上的硅酸盐
AlSi_data_SMB_sedfree <- subset(AlSi_data_SMB,Source ==  "Aggregate-attached, Sediment-free" |

#来自圣莫尼卡盆地在沉积物中附着在ANME-SRB共生体上的硅酸盐
AlSi_data_SMB_fromsed <- subset(AlSi_data_SMB,Source ==  "Aggregate-attached" | Source ==  "Se

#来自圣莫尼卡盆地在沉积物中附着在ANME-SRB共生体上的硅酸盐和来自圣莫尼卡盆地在培养中附着在无沉积
AlSi_data_SMB_freevsfromsed<- subset(AlSi_data_SMB,Source ==  "Aggregate-attached" | Source ==

#来自伊尔河盆地的沉积物和附着在ANME-SRB共生体上的硅酸盐
AlSi_data_ERB <- subset(AlSi_data, Basin == "Eel River", select = c("Al.per.Si","Source","Basi
```

```r
#创建小提琴图

ot = ggplot(AlSi_data, aes(Source, Al.per.Si))
# 使用ggplot2包创建基础绘图对象，指定数据框AlSi_data，以及x轴为Source，y轴为Al.per.Si

ot +
  geom_violin() +
  geom_boxplot(width=0.2, outliers = FALSE) +
  stat_compare_means(method = "aov", label = "p", size = 3) +
  stat_compare_means(method = 't.test', label.y = 1,
                     ref.group = "Sediment", label = "p.signif") +
 # stat_compare_means(method = "t.test",
#                      comparisons = list(
#                        c("Sediment", "Aggregate-attached"),
#                        c("Aggregate-attached, Sediment-free", "Aggregate-attached"),
#                        c("Sediment", "Aggregate-attached, Sediment-free")
 #                     )) +
  # geom_jitter(width = 0.1)
   # 被注释掉的散点抖动图层，用于展示原始数据点，避免重叠，抖动宽度为0.1
  facet_grid(~Basin, scales = "free", space = "free") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, vjust = 1)) +
  labs(x = "", y = "(Al per Si")
```

```
Warning: Computation failed in `stat_compare_means()`.
Caused by error in `mutate()`:
ℹ In argument: `p = purrr::map(...)`.
Caused by error in `purrr::map()`:
ℹ In index: 1.
```

```
ℹ With name: x.1.
Caused by error in `contrasts<-`:
! contrasts can be applied only to factors with 2 or more levels
Warning: Computation failed in `stat_compare_means()`.
Caused by error in `mutate()`:
ℹ In argument: `p = purrr::map(...)`.
Caused by error in `purrr::map()`:
ℹ In index: 1.
ℹ With name: x.2.
Caused by error in `contrasts<-`:
! contrasts can be applied only to factors with 2 or more levels

Warning: Computation failed in `stat_compare_means()`.
Caused by error:
! Can't find specified reference group: 2. Allowed values include one of: 1, 3

Warning: Computation failed in `stat_compare_means()`.
Caused by error:
! Can't find specified reference group: 2. Allowed values include one of: 3, 1
```




```
# 数据可视化：
# 使用小提琴图和箱线图展示不同Source（来源）的Al.per.Si（铝与硅的比值）分布。
# 统计分析：
#通过stat_compare_means函数，结合ANOVA和t检验，对不同组别进行统计比较，并在图中显示显著性。
# 分面绘图：
# 按Basin（盆地）变量分面，每个面显示对应盆地的数据分布。
```

```
# 自定义样式：
# 调整x轴标签的角度，使其更易读；并可以添加散点抖动图层来展示原始数据点。
```

对不同数据集（不同来源的硅酸盐数据）进行单因素方差分析（One-way ANOVA），每个数据集都对应特定的地理区域和样本来源，用于比较不同来源（Source）对铝与硅的比值（Al.per.Si）的影响。

```
# 对来自Jaco的沉积物和附着在ANME-SRB 共生体上的硅酸盐进行单因素方差分析（One-way ANOVA）测试
# 输出方差分析的结果摘要，包括 F 值、显著性水平（p 值）等信息
resot6.aov <- aov(Al.per.Si ~ Source, data = AlSi_data_Jaco)
summary(resot6.aov)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Source        1 0.3026 0.30257   29.56 3.78e-07 ***
Residuals   101 1.0339 0.01024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#对来自圣莫尼卡盆地在培养中附着在无沉积物 A N M E - S R B 共生体上的硅酸盐进行单因素方差分析（C
resot7.aov <- aov(Al.per.Si ~ Source, data = AlSi_data_SMB_sedfree)
summary(resot7.aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Source        1  2.130  2.1299   86.96 <2e-16 ***
Residuals   173  4.237  0.0245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 对来自圣莫尼卡盆地在沉积物中附着在 A N M E - S R B 共生体上的硅酸盐进行单因素方差分析（One-wa
resot8.aov <- aov(Al.per.Si ~ Source, data = AlSi_data_SMB_fromsed)
summary(resot8.aov)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Source        1  1.058  1.0584   36.61 8.61e-09 ***
Residuals   174  5.030  0.0289
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 对来自Santa Monica Basin的沉积物中附着于ANME-SRB联合体的硅酸盐，以及在培养实验中无沉积物的 AN
resot9.aov <- aov(Al.per.Si ~ Source, data = AlSi_data_SMB_freevsfromsed)
summary(resot9.aov)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Source       1  0.128 0.12798   6.676 0.0116 *
Residuals   81  1.553 0.01917
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 对来自伊尔河盆地的沉积物和附着在 ANME-SRB 共生体上的硅酸盐进行单因素方差分析（One-way ANOVA）
resot10.aov <- aov(Al.per.Si ~ Source, data = AlSi_data_ERB)
summary(resot10.aov)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
Source        1 0.4429  0.4429   17.88 5.1e-05 ***
Residuals   103 2.5510  0.0248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```