talk06 练习与作业

目录

0.1 练习和作业说明

将相关代码填写入以 “‘{r} “‘ 标志的代码框中，运行并看到正确的结果；

完成后，用工具栏里的”Knit” 按键生成 PDF 文档；

将 PDF 文档改为：`姓名-学号-talk06 作业.pdf`，并提交到老师指定的平台/钉群。

0.2 Talk06 内容回顾

1. 3 个生信任务的 R 解决方案
2. factors 的更多应用 (forcats)
3. pipe

0.3 练习与作业：用户验证

请运行以下命令，验证你的用户名。

如你当前用户名不能体现你的真实姓名，请改为拼音后再运行本作业！

```
Sys.info()[["user"]]
```

```
## [1] "sicheng.wu"
```

```
Sys.getenv("HOME")
```

```
## [1] "/home/vkorpela"
```

0.4 练习与作业 1：作图

0.4.1 用下面的数据作图

1. 利用下面代码读取一个样本的宏基因组相对丰度数据

```
abu <-
  read_delim(
    file = "../data/talk06/relative_abundance_for_RUN_ERR1072629_taxonlevel_species.txt
    delim = "\t", quote = "", comment = "#");
```

2. 取前 5 个丰度最高的菌，将其它的相对丰度相加并归为一类 `Qita`；
3. 用得到的数据画如下的空心 pie chart:

```
## 代码写这里，并运行；
library(tidyverse)
```

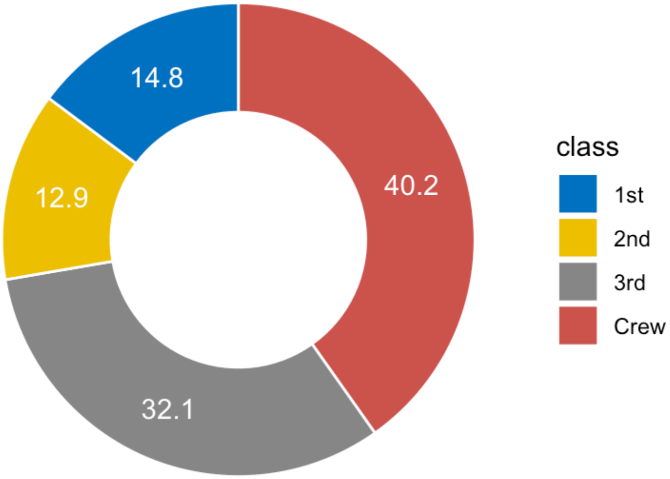


图 1: make a pie chart like this using the meteagenomics data

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

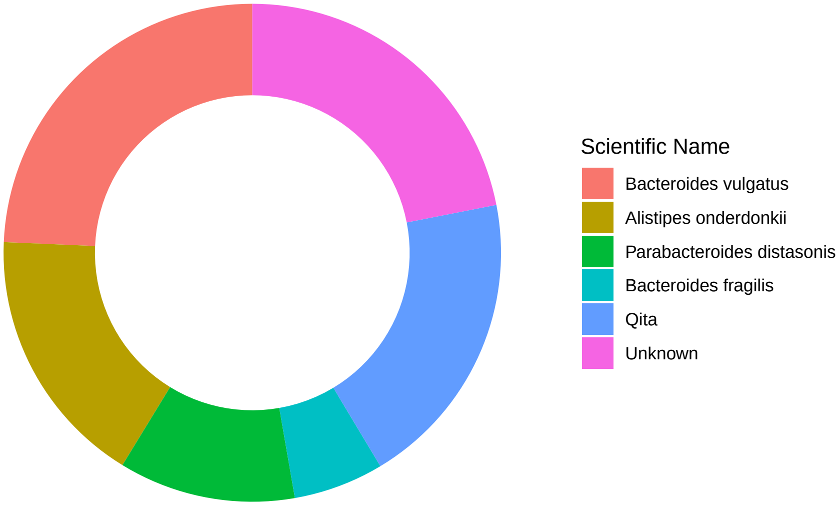
```
library(tidytidbits)

# 读取丰度数据
abu <- read_delim(
  file = "../data/talk06/relative_abundance_for_RUN_ERR1072629_taxonlevel_species.txt",
  delim = "\t",
  quote = "",
  comment = "#"
)
```

```
## Rows: 122 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# 取丰度前五高的微生物，其余归于 Qita 类
abu.filtered <- abu %>%
  arrange(desc(relative_abundance)) %>%
  lump_rows(
    scientific_name,
    relative_abundance,
    n = 5,
    other_level = "Qita"
  )
abu.filtered
```

```
## # A tibble: 6 x 3
##   ncbi_taxon_id relative_abundance scientific_name
##           <dbl>              <dbl> <chr>
## 1           821              24.3  Bacteroides vulgatus
## 2            -1              21.9  Unknown
## 3        328813              16.9  Alistipes onderdonkii
## 4           823              11.5  Parabacteroides distasonis
## 5           817               5.87 Bacteroides fragilis
## 6      31848070              19.5  Qita
```

```
# 绘制空心环状图
abu.filtered$scientific_name <-
  fct_relevel(
    fct_reorder(
      abu.filtered$scientific_name,
      abu.filtered$relative_abundance,
      .desc = TRUE
    ),
    "Qita", "Unknown",
    after = Inf
  )

plot1 <-
  ggplot(
    data = abu.filtered,
    aes(
      x = 3,
      y = relative_abundance,
      fill = scientific_name
    )
  ) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y", start = 0) +
  xlim(c(1, 3.5)) +
  labs(x = NULL, y = NULL, fill = "Scientific Name") +
  theme_void()
plot1
```

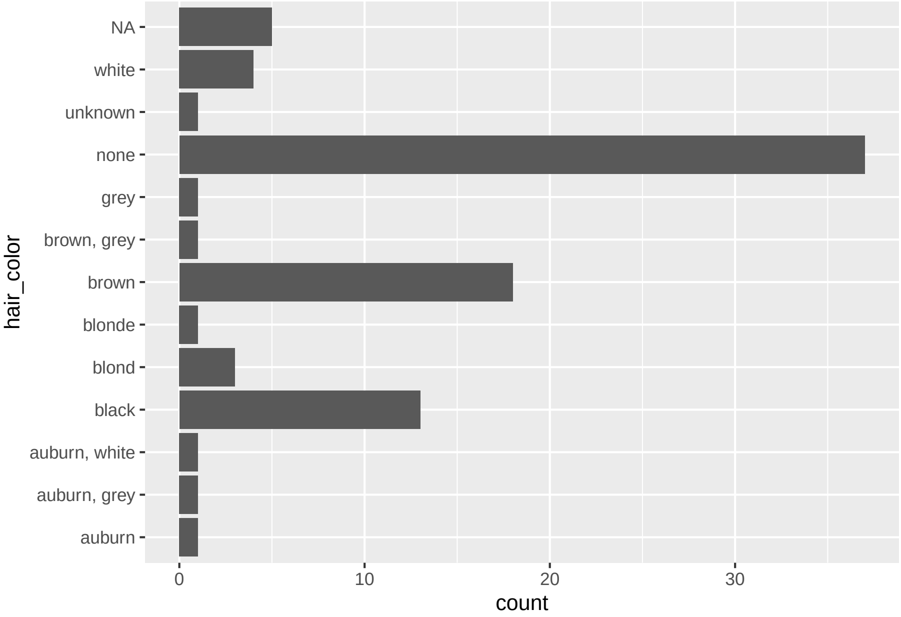


0.4.2 使用 starwars 变量做图

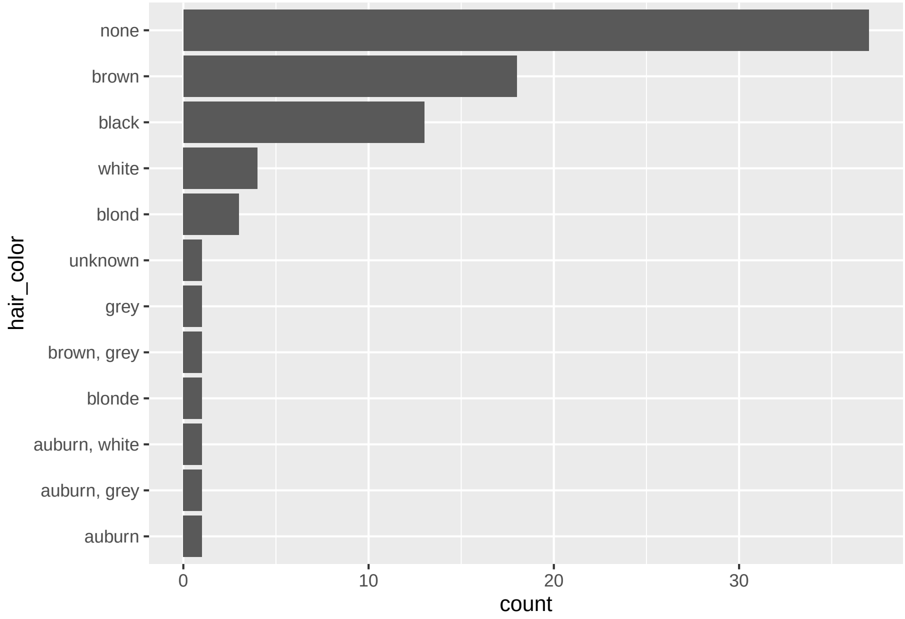
1. 统计 starwars 中 hair_color 的种类与人数时，可用下面的代码：

但是，怎么做到**按数量从小到大**排序？

```
library(dplyr)
library(ggplot2)
library(forcats)
ggplot(starwars, aes(x = hair_color)) +
  geom_bar() +
  coord_flip()
```

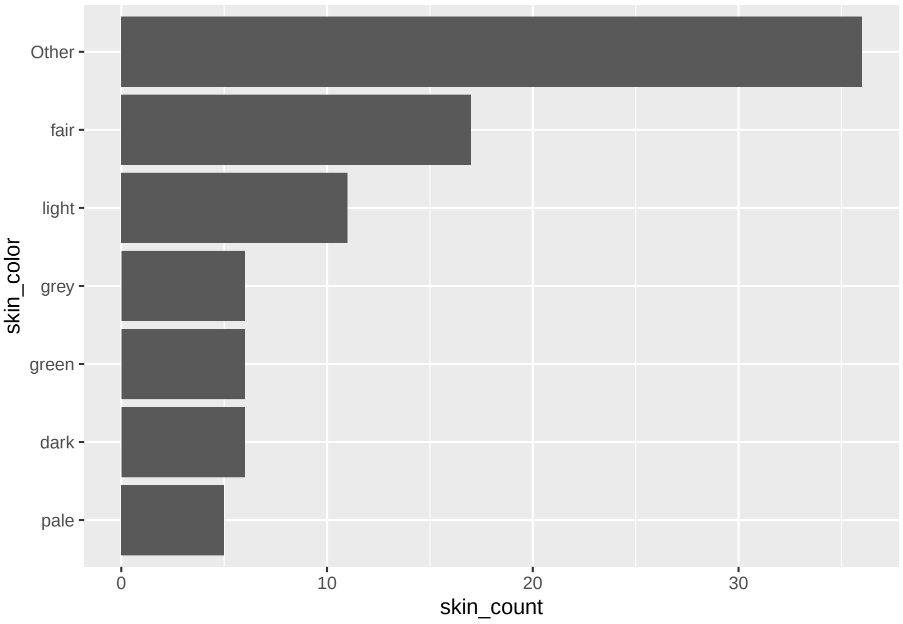


```
## 代码写这里，并运行；
sw.hair <- starwars %>%
  filter(!is.na(hair_color))

sw.hair$hair_color <-
  fct_reorder(
    sw.hair$hair_color,
    sw.hair %>%
      group_by(hair_color) %>%
      mutate(hair_count = n()) %>%
      .$hair_count,
    .desc = FALSE
  )

plot2 <-
  ggplot(sw.hair, aes(x = hair_color)) +
  geom_bar() +
  coord_flip()
plot2
```



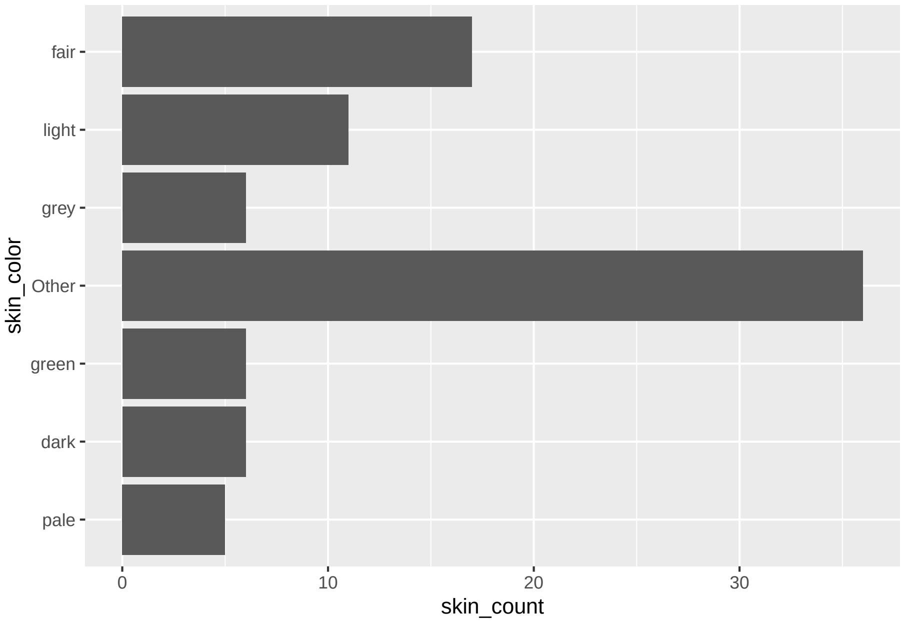
2. 统计 skin_color 时，将出现频率小于 0.05（即 5%）的颜色归为一类 Others，按出现次数排序后，做与上面类似的 barplot；

```
## 代码写这里，并运行；
sw.skin <- starwars %>%
  group_by(skin_color) %>%
  summarise(
    skin_rate = n() / count(starwars)[[1]],
    skin_count = n()
  ) %>%
  lump_rows(
    skin_color,
    skin_rate,
    prop = 0.05,
    other_level = "Other"
  )

sw.skin$skin_color <-
  fct_reorder(
    sw.skin$skin_color,
    sw.skin$skin_count,
    .desc = FALSE
  )

plot3 <-
  ggplot(sw.skin, aes(x = skin_color, y = skin_count)) +
  geom_bar(stat = "identity") +
  coord_flip()
plot3
```



3. 使用 2 的统计结果，但画图时，调整 bar 的顺序，使得 `Others` 处于第 4 的位置上。提示，可使用 `fct_relevel` 函数；

```
## 代码写这里，并运行；
sw.skin$skin_color <-
  fct_relevel(
    sw.skin$skin_color,
    "Other",
    after = 3
  )

plot4 <-
  ggplot(sw.skin, aes(x = skin_color, y = skin_count)) +
  geom_bar(stat = "identity") +
  coord_flip()
plot4
```



0.5 练习与作业 2：数据分析

0.5.1 使用 STRING PPI 数据分析并作图

1. 使用以下代码，装入 PPI 数据；

```
ppi <- read_delim( file = "../data/talk06/ppi900.txt.gz", col_names = T,
                   delim =  "\t", quote = "" );
```

2. **随机挑选**一个基因，得到类似于本章第一部分的互作网络图；

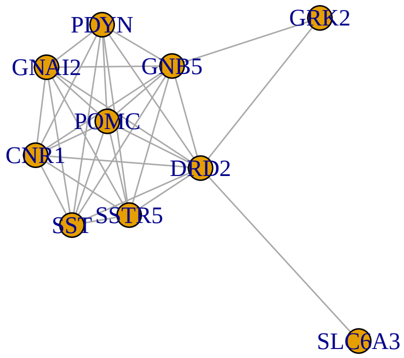
```
## 代码写这里，并运行；
library(igraph)
```

```
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
ppi <- read_delim(
    file = "../data/talk06/ppi900.txt.gz",
    col_names = TRUE,
    delim = "\t",
    quote = ""
  )
```

```
## Rows: 504436 Columns: 3

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): gene1, gene2
## dbl (1): score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
genelist <- ppi %>%
  filter(gene1 == "DRD2") %>%
  arrange(desc(score)) %>%
  slice(1:9) %>%
  .$gene2
genelist <- unique(c("DRD2", genelist))

ppi.drd2 <- ppi %>%
  filter(
    gene1 %in% genelist &
    gene2 %in% genelist
  ) %>%
  mutate(
    group = if_else(
      gene1 > gene2,
      paste(gene1, gene2, sep="-"),
      paste(gene2, gene1, sep="-"),
    )
  ) %>%
  group_by(group) %>%
  slice(1)

net.drd2 <- graph_from_data_frame(ppi.drd2, directed = FALSE)
plot(net.drd2)
```



0.5.2 对宏基因组相对丰度数据进行分析

1.`data/talk06` 目录下有 6 个文本文件，每个包含了一个宏基因组样本的分析结果：

```
relative_abundance_for_curated_sample_PRJEB6070-DE-073_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-074_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-075_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-076_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-077_at_taxonlevel_species.txt
```

2. 分别读取以上文件，提取 `scientific_name` 和 `relative_abundance` 两列；
3. 添加一列为样本名，比如 `PRJEB6070-DE-073`, `PRJEB6070-DE-074` ... ;
4. 以 `scientific_name` 为 `key`，将其内容合并为一个 `data.frame` 或 `tibble`，其中每行为一个样本，每列为样本的物种相对丰度。注意：用 `join` 或者 `spread` 都可以，只要能解决问题。

5. 将 NA 值改为 0。

```
## 代码写这里，并运行；
sample073 <-
  read_tsv("./data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-073_at_tax
  select(scientific_name, relative_abundance) %>%
  mutate(sample = "PRJEB6070-DE-073")
```

```
## Rows: 74 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): taxon_rank_level, scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sample074 <-
  read_tsv("./data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-074_at_tax
  select(scientific_name, relative_abundance) %>%
  mutate(sample = "PRJEB6070-DE-074")
```

```
## Rows: 78 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): taxon_rank_level, scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sample075 <-
  read_tsv("./data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-075_at_tax
  select(scientific_name, relative_abundance) %>%
  mutate(sample = "PRJEB6070-DE-075")
```

```
## Rows: 98 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): taxon_rank_level, scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sample076 <-
  read_tsv("./data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-076_at_tax
  select(scientific_name, relative_abundance) %>%
  mutate(sample = "PRJEB6070-DE-076")
```

```
## Rows: 90 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): taxon_rank_level, scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sample077 <-
  read_tsv("./data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-077_at_tax
  select(scientific_name, relative_abundance) %>%
  mutate(sample = "PRJEB6070-DE-077")
```

```
## Rows: 78 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): taxon_rank_level, scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sample.comb <-
  bind_rows(sample073, sample074, sample075,
            sample076, sample077) %>%
  pivot_wider(
    names_from = scientific_name,
    values_from = relative_abundance,
    values_fill = 0,
    values_fn = sum
  )
sample.comb
```

```
## # A tibble: 5 x 146
##   sample Faeca~1 [Euba~2 Bacte~3 Copro~4 Roseb~5 Bacte~6 Bacte~7 Rumin~8 Alist~9
##   <chr>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 PRJEB~    19.9    9.49    7.15     5.02 4.69     4.57     4.36  4.24     2.59
## 2 PRJEB~    13.8    0.221   3.74     0    0.00751  4.13     0.450 0.0152   0.0124
## 3 PRJEB~    18.6    0.322   0.450    1.08 2.87     0.362    0     4.80     2.43
## 4 PRJEB~    13.0    5.74    2.48     0    1.39     0.0140   0     0.00114  1.01
## 5 PRJEB~    10.0   25.5     1.47     0    2.35     2.33     0     6.58     2.48
## # ... with 136 more variables: `Bacteroides ovatus` <dbl>,
## #   `Bacteroides uniformis` <dbl>, `Roseburia intestinalis` <dbl>,
## #   `[Eubacterium] eligens` <dbl>, `Alistipes sp. HGB5` <dbl>,
## #   `Burkholderiales bacterium 1_1_47` <dbl>,
## #   `Barnesiella intestinihominis` <dbl>, `Ruminococcus lactaris` <dbl>,
## #   `Odoribacter splanchnicus` <dbl>, `Collinsella aerofaciens` <dbl>,
## #   `Adlercreutzia equolifaciens` <dbl>, `[Ruminococcus] torques` <dbl>, ...
```