Tidy env info

Load the csv file and tidy the data

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x purrr::%||%()   masks base::%||%()
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
# Load the csv file
env_info <- read_csv("C:\\Users\\DuYih\\Desktop\\sequence-PVC.csv")
```

```
Rows: 13754 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (6): Locus, Accession, Version, Project, Isolation Source, Isolation Sou...

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.
```

```r
env_info
```

```
# A tibble: 13,754 x 6
   Locus    Accession Version  Project `Isolation Source` Isolation Source Sim~1
   <chr>    <chr>     <chr>    <chr>   <chr>              <chr>
 1 MT193413 MT193413  MT19341~ <NA>    "patina on cave q~ patina on cave quartz~
 2 MT193412 MT193412  MT19341~ <NA>    "patina on cave q~ patina on cave quartz~
 3 KT122326 KT122326  KT12232~ <NA>    "inundated soil o~ water
 4 KT122322 KT122322  KT12232~ <NA>    "inundated soil o~ water
 5 KT122301 KT122301  KT12230~ <NA>    "inundated soil o~ water
 6 KT122296 KT122296  KT12229~ <NA>    "inundated soil o~ water
 7 KT122291 KT122291  KT12229~ <NA>    "inundated soil o~ water
 8 KT122210 KT122210  KT12221~ <NA>    "sediment of 155m~ water
 9 KT122209 KT122209  KT12220~ <NA>    "sediment of 155m~ water
10 KT122196 KT122196  KT12219~ <NA>    "sediment of 155m~ water
# i 13,744 more rows
# i abbreviated name: 1: `Isolation Source Simplified`
```

```r
# Tidy the data
env_info %>% group_by(Project, `Isolation Source`) %>%
  summarise(count=n())
```

```
`summarise()` has grouped output by 'Project'. You can override using the
`.groups` argument.
```

```
# A tibble: 1,447 x 3
# Groups:   Project [12]
   Project     `Isolation Source`                                          count
   <chr>       <chr>                                                       <int>
 1 PRJNA171131 "interface from Hypersaline Lake Medee,\n                  ~     5
 2 PRJNA33175  "Algal-bacterial consortia"                                     2
 3 PRJNA33175  "Hirudo medicinalis"                                            1
 4 PRJNA33175  "Sphagnum peat bog"                                             2
 5 PRJNA33175  "UASB granular sludge"                                          1
 6 PRJNA33175  "acidic geothermal spring"                                      1
 7 PRJNA33175  "acidic hotspring"                                              1
 8 PRJNA33175  "acidic soil from the Solfatara crater"                         3
 9 PRJNA33175  "algae"                                                         1
10 PRJNA33175  "anoxic bulk soil of a flooded rice\n                     ~     1
# i 1,437 more rows
```

```r
env_info %>% group_by(Project) %>%
  summarise(count=n())
```

```
# A tibble: 12 x 2
   Project     count
   <chr>       <int>
 1 PRJNA171131     5
 2 PRJNA33175     79
 3 PRJNA33823     16
 4 PRJNA34525     54
 5 PRJNA38465     61
 6 PRJNA39207    152
 7 PRJNA46435      7
 8 PRJNA49615      4
 9 PRJNA555798     1
10 PRJNA71063      1
11 PRJNA76619      3
12 <NA>        13371
```

```r
env_info %>% filter(Project == "PRJNA38465") %>%
  group_by(Project, `Isolation Source`) %>%
  summarise(count=n())
```

```
`summarise()` has grouped output by 'Project'. You can override using the
`.groups` argument.
```

```
# A tibble: 11 x 3
# Groups:   Project [1]
   Project    `Isolation Source`                                           count
   <chr>      <chr>                                                        <int>
 1 PRJNA38465 "biofilm in 1m deep hydrothermal vent in\n"                  ~    3
 2 PRJNA38465 "biomat 11m deep in cenote La Palita"                             7
 3 PRJNA38465 "biomat 30m deep in cenote La Palita"                            19
 4 PRJNA38465 "biomat 80m deep in cenote La Palita"                             1
 5 PRJNA38465 "biomat in a rock outcrop in cenote La\n"                    ~    2
 6 PRJNA38465 "biomat in the sediment of cenote La\n"                      ~    7
 7 PRJNA38465 "green biomat sample from 8m deep in\n"                      ~    4
 8 PRJNA38465 "orange biomat sample from 8m deep in\n"                     ~    6
 9 PRJNA38465 "red biomat sample from 12m deep in\n"                      c~    3
10 PRJNA38465 "water column sample from 32m deep in\n"                     ~    8
11 PRJNA38465 "water column sample from 53m deep in\n"                     ~    1
```