Holo omic sample problem

Carl Mathias Kobel (carl.mathias.kobel near nmbu.no) Wed Jan 17 17:51:55 2024

Technical setup and parameters for reproducibility

Technical note: We recommend that you use "docker://rocker/tidyverse:4.3.2" (https://rocker-project.org/images/versioned/rstudio.html) to ensure successful reproducibility.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## 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 become errors
```

```
R.version # v4.2.2
```

```
##                _
## platform       x86_64-pc-linux-gnu
## arch           x86_64
## os             linux-gnu
## system         x86_64, linux-gnu
## status         Patched
## major          4
## minor          2.2
## year           2022
## month          11
## day            10
## svn rev        83330
## language       R
## version.string R version 4.2.2 Patched (2022-11-10 r83330)
## nickname       Innocent and Trusting
```

```
packageVersion("tidyverse") # v2.0.0
```

```
## [1] '2.0.0'
```

```
knitr::opts_chunk$set(
    echo = T,
    results = 'show',
    message = F,
    warning = F
    )
```

Problem?

Let us consider a hypothetical holo-omics study, where we have measured the host transcriptome of the rumen wall in 100 cows (n = 100) and the meta-transcriptome of the rumen content in those same individuals (p = 20 000 host genes + average 3000 microbial genes x 200 microbial species = 620 000 features). Let us further assume that the experiment is set up to measure methane emission, and that half of the cows were given a methane inhibiting feed additive (treatment) that indeed reduced emissions. This dataset would pose a massive challenge for data analysis, and not primarily because it would require considerable computational resources to assemble and annotate metagenome assembled genomes (MAGs) and estimate expression (read mapping). The main challenge is related to the large number of features compared to samples. Naively one would think that this data set could be analysed using multivariate- or machine learning-based prediction methods, where the predictive model could be queried for features or combination of features that contributed significantly to the prediction, e.g. IF gene G on MAG5 is up and host gene H is down THEN low methane. However, with this many features there will be an enormous number of feature combinations that could separate low and high emitting cows, and with only 100 examples (cows) to constrain them, we would never be able to discern real biological feature-combinations from spurious ones (see supplement for R code showing perfect prediction of a randomly generated data set with n « p). This phenomenon is referred to as overfitting and is a consequence of the curse of dimensionality – the number of examples (cows) needed to identify the biologically meaningful features grows exponentially with the number of features.

Generate data

List n cow samples.

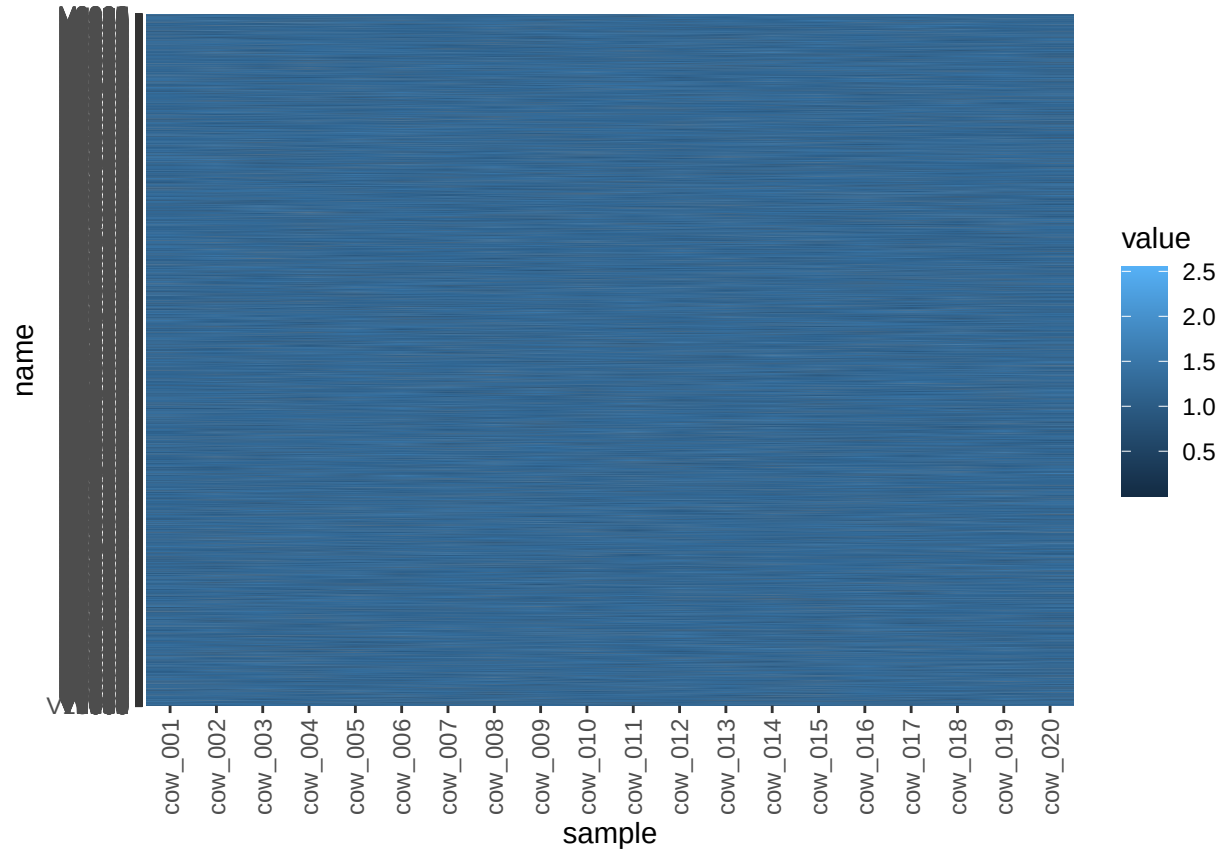
```
n = 20 # number of samples (cows)

samples = tibble(
    sample = paste0("cow_", str_pad(1:n, 3, pad = 0)) # Formatting string prefixed with "cow_" and a nu
)

samples
```

```
## # A tibble: 20 x 1
##    sample
##    <chr>
##  1 cow_001
##  2 cow_002
##  3 cow_003
##  4 cow_004
##  5 cow_005
##  6 cow_006
##  7 cow_007
##  8 cow_008
##  9 cow_009
```

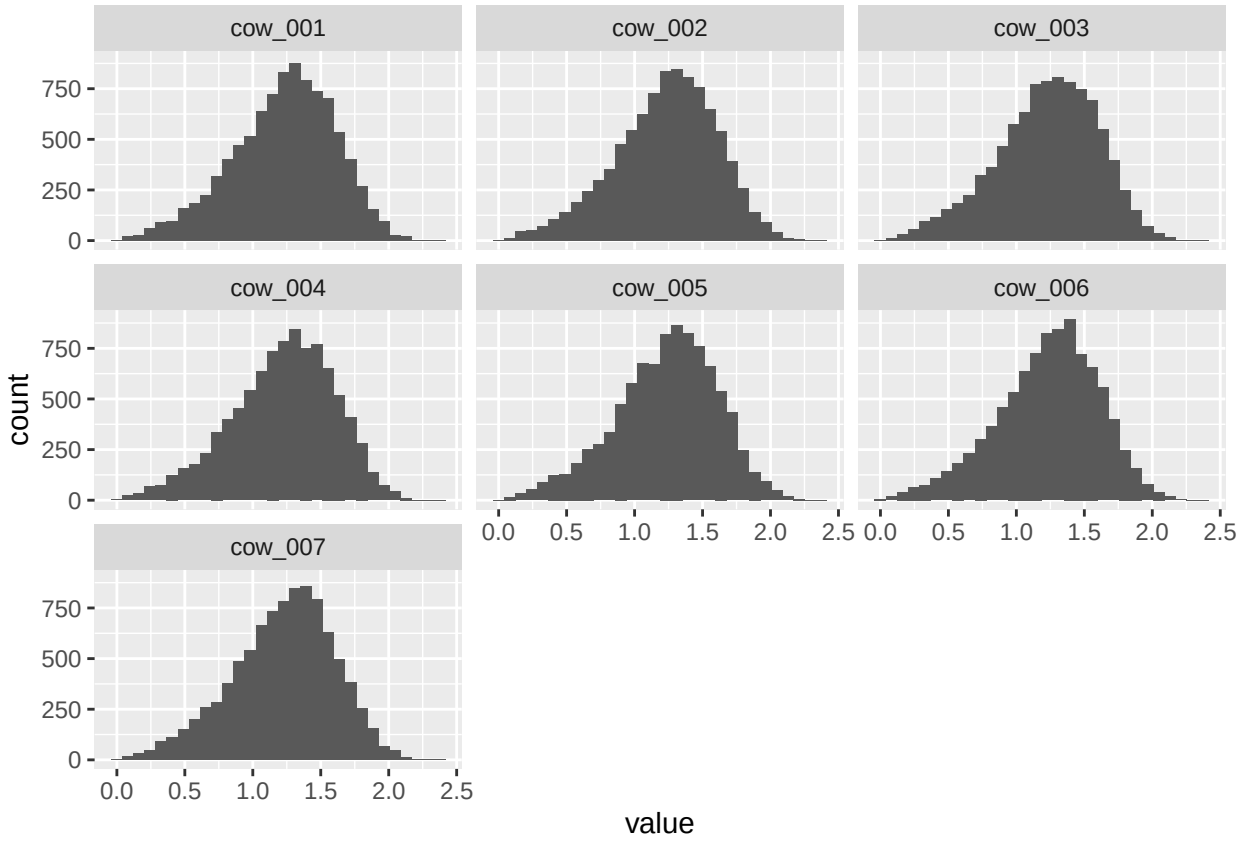
```
## 10 cow_010
## 11 cow_011
## 12 cow_012
## 13 cow_013
## 14 cow_014
## 15 cow_015
## 16 cow_016
## 17 cow_017
## 18 cow_018
## 19 cow_019
## 20 cow_020
```

Generate normal data from a random omic layer.

```
#set.seed(1324) # Use this to get equivalent results through space and time.
set.seed(26*65535)
p = 10000 # number of features for a hypothesized omic layer

omic_layer = samples %>%
    bind_cols(
        matrix( # Generating m vectors (biological features) with one value for each n samples.
            rnorm(n*p, mean = 1.5, sd = 1) %>% # Using a low mean, means that we will get some negative
                sqrt() %>%  # Trying to add some "noise" by squarerooting it .
                identity(),
            n,
            p
        ) %>%
            as_tibble()
    )




# Quick visualizations
omic_layer %>%
    pivot_longer(-sample) %>%
    ggplot(aes(sample, name, fill = value)) +
    geom_raster() +
    theme(
        axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)
    )
```




```
omic_layer %>%
    head(7) %>%
    pivot_longer(-sample) %>%

    ggplot(aes(value)) +
    geom_histogram() +
    facet_wrap(~sample)
```



Generate a random trait vector "methane".

```
trait = samples %>%
    mutate(methane = rnorm(n))

trait %>% glimpse()
```

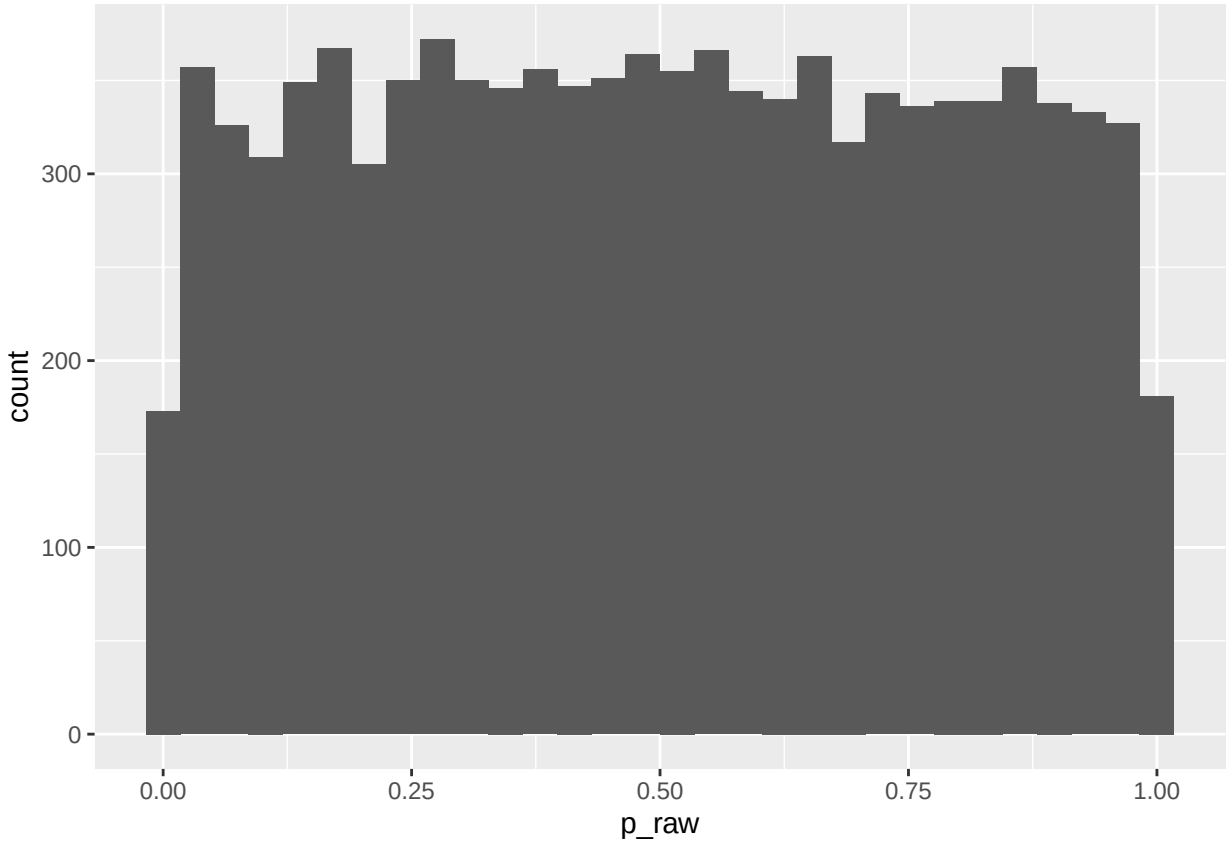
```
## Rows: 20
## Columns: 2
## $ sample  <chr> "cow_001", "cow_002", "cow_003", "cow_004", "cow_005", "cow_00~
## $ methane <dbl> 0.01976174, 0.08003521, 0.91722540, -0.99243056, -0.24581664, ~
```

Make models

Calculate linear models between omic layer features and methane trait

```
data = trait %>%
    left_join(omic_layer, by = "sample")  %>%
    column_to_rownames("sample")

# Perform multiple testing
tests = lapply(
    omic_layer[-1], # %>% as.data.frame(),
    function(x) {
        #x
```

```
        data = bind_cols(trait, feature = x) %>%
            column_to_rownames("sample")

        lm(
            formula = "methane ~ feature",
            data = data
        ) %>%
            summary() # The summary() function calculates F statistic and pvalue.
    }
)



# Extract pvalues alone.
pvals_raw = lapply(
    tests,
    function(x) {
        x$coefficients[2,4] # pvalue of the slope is at coordinate 2,4 in the table.
    }
) %>% unlist()   %>%
    as_tibble(rownames = "feature")  %>%
    rename(p_raw = value)


# Checking that the p-values are uniformly distributed which I think should be an assumption of multipl

pvals_raw %>%
    ggplot(aes(p_raw)) +
    geom_histogram()
```

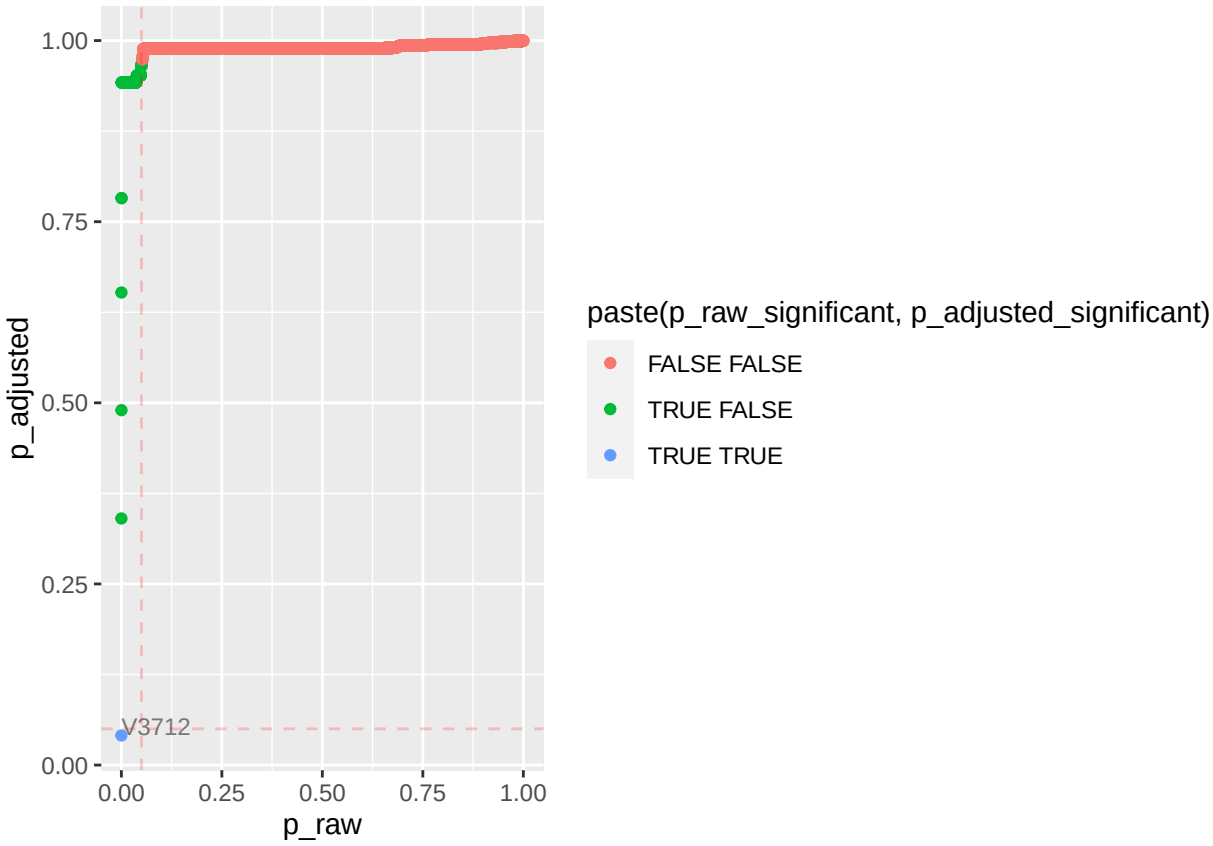


Adjust for multiple testing

Adjust for multiple testing by using benjamini hochberg correction.

```
pvals_adjusted = p.adjust(
    pvals_raw$p_raw,
    method = "BH"
)

alpha_raw = 0.05
alpha_adjusted = 0.05

pvals = bind_cols(
    pvals_raw, p_adjusted = pvals_adjusted) %>%
    mutate(
        p_raw_significant = p_raw < alpha_raw,
        p_adjusted_significant = p_adjusted < alpha_adjusted
    )


# Examine relationship between raw and adjusted pvalues.
pvals %>%

    mutate(label = case_when(
        p_adjusted_significant ~ feature,
```

```
        .default = ""
    )) %>%

    ggplot(aes(p_raw, p_adjusted, color = paste(p_raw_significant, p_adjusted_significant), label = labe
    geom_point() +
    geom_text(color = "black", alpha = 0.5, size = 3, hjust = 0, vjust = 0) + # Put labels on the signi

    geom_vline(xintercept = alpha_raw, linetype = "dashed", color = "red", alpha = 0.2) +
    geom_hline(yintercept = alpha_adjusted, linetype = "dashed", color = "red", alpha = 0.2)
```



Investigating "significant" results.

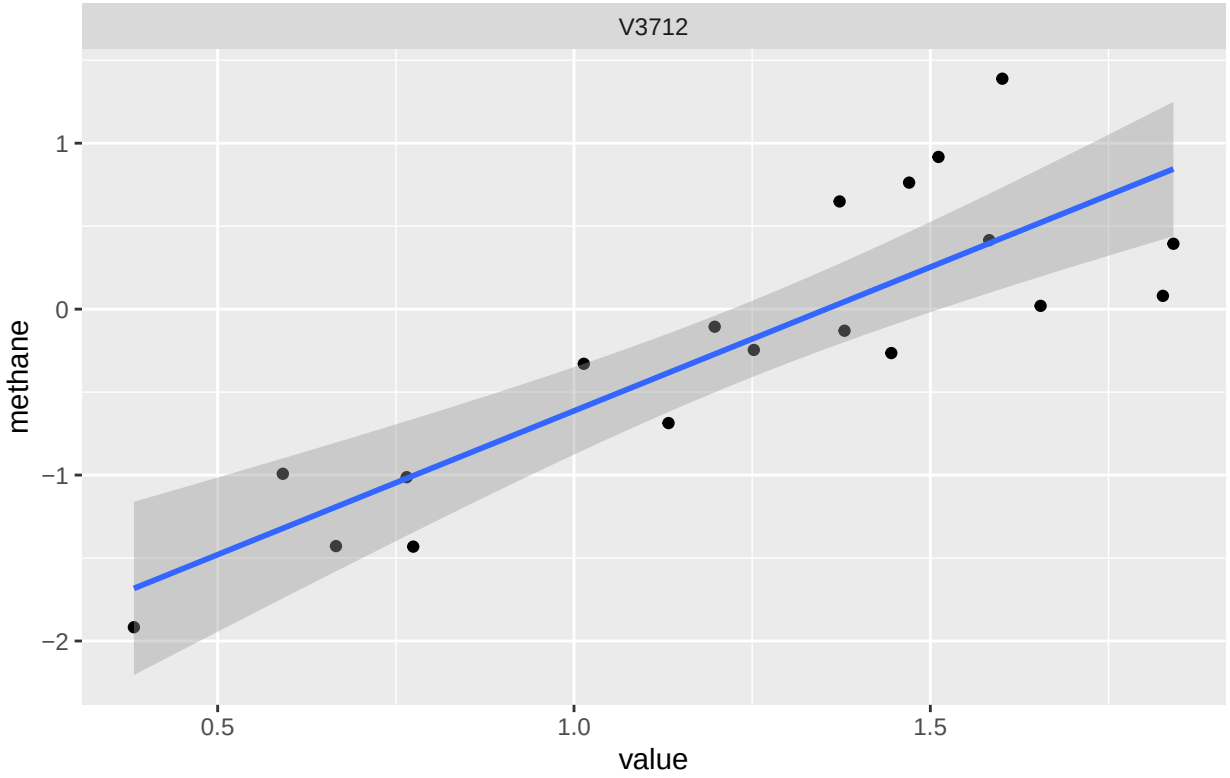
Depending on your seed and setting, this might contain a spuriously correlated gene that "explains" methane. Investigate the statistically significant results (which are in fact random and spurious)

```
# Stop this script if there are no significant results.
stopifnot(
    (pvals %>%
    filter(p_adjusted_significant) %>%
    nrow()) >= 1
)


significant_features = pvals %>%
```

```
    filter(p_adjusted_significant)

significant_features
```

```
## # A tibble: 1 x 5
##   feature      p_raw p_adjusted p_raw_significant p_adjusted_significant
##   <chr>        <dbl>      <dbl> <lgl>             <lgl>
## 1 V3712   0.00000409     0.0409 TRUE              TRUE
```

```
omic_layer %>%
    select(sample, significant_features %>% pull(feature)) %>%
    left_join(trait, by = "sample") %>%

    pivot_longer(-c(sample, methane)) %>%

    ggplot(aes(value, methane)) +
    facet_wrap(~name) +
    geom_point() +
    geom_smooth(method = "lm") +
    labs(
        caption = paste("These spurious results exist even though we did adjustment for multiple testing
    )
```



These spurious results exist even though we did adjustment for multiple testing.
20 samples (cows), and 10000 biological features in the hypothetical omic layer.

Just to collect html documents. Be aware that this chunk has a circular dependency, because it requires that this document is saved and knitted as both html and pdf.

```
zip portable_documents.zip *.html *.pdf *.Rmd
```

```
## updating: holo-omic-review-R-example.html (deflated 54%)
## updating: holo omic review R example.nb.html (deflated 54%)
## updating: holo-omic-review-R-example.pdf (deflated 8%)
## updating: holo omic review R example.Rmd (deflated 58%)
## updating: holo-omic-review-R-example.Rmd (deflated 58%)
```