Rethomics: an R framework to analyse high-throughput behavioural data

Quentin Geissmann[1*], Luis Garcia Rodriguez[2], Esteban J Beckwith[1], Giorgio F Gilestro[1*]

1 Department of Life Sciences, Imperial College London, London, United Kingdom
2 Affiliation Dept/Program/Center, Institution Name, City, State, Country

* qgeissmann@gmail.com, giorgio@gilest.ro

Abstract

Ethomics, a quantitative and high-throughput approach to animal behaviour, is a new and exciting field. The recent development of automatised methods that can score various behaviours on a large number of animals provides biologists with an unprecedented set of tools to decipher these complex phenotypes. Analysing ethomics data comes with many challenges that are largely shared across acquisition platform and paradigms. However, there is little effort in providing a generic framework to specifically analyse multiple and long behavioural time series. We developed the `rethomics` framework, a suite of `R` packages that altogether offer utilities to: import, store, visualise and analyse behavioural data. In this article, we describe it and show an example of its application to the blooming field of sleep and circadian rhythm in fruit fly. The `rethomics` framework is available and documented at https://rethomics.github.io.

Introduction

Animal behaviours are complex phenotypical manifestations of the interaction between nervous systems and external or internal environments. In the last few decades, our ability to record vast quantities of various phenotypical data has tremendously increased. Behaviour scoring is certainly not an exception to this trend. Indeed, many platforms (TODO citations) have been developed in order to allow biologists to continuously record behaviours such as activity, position and feeding of multiple animals over long durations (days or weeks).

The availability of large amounts of data is very exciting as it paves the way for in-depth quantitative analyses. Clearly, the multiplicity of model organisms, hypotheses and paradigms should be matched by a diverse range of recording tools. However, when it comes to the subsequent data analysis, there is no unified, programmatic, framework that could be used as a set of building blocks in a pipeline. Instead, tools tend to consist of graphical interfaces with rigid functionalities that only import data from a single platform. There are, at least, three issues with this approach. First of all, state-of-the-art analysis and visualisation requires a level of reproducibility, flexibility and scalability than only a programmatic interface can provide. Secondly, it favours replicated work as developers need to create their independent solution to similar problems. Lastly, it links analysis and visualisation to the target acquisition tool, which makes it very difficult to share cross-tool utilities and concepts.

Thankfully, behavioural data is conceptually largely agnostic of the acquisition platform and paradigm. Typically, the behaviour of each individual is described by a

long time series (possibly multivariate and heterogeneous). Importantly, individuals are labelled with arbitrary metadata defined by the experimenter (*e.g.* sex, treatment and genotype). Efficiently combining and manipulating metadata and data of hundreds of individuals, each recorded for weeks, is not trivial.

In the article herein, we describe `rethomics`, a framework that unifies analysis of behavioural dataset in an efficient and flexible manner. It offers an elegant computational solution to store, manipulate and visualise a large amount of data. We expect it to fill the gap between behavioural biology and data sciences, thus promoting collaboration between computational and behavioural scientists. `rethomics` comes with a extensive documentation and a set of both practical and theoretical demos and tutorials.

Design and Implementation

`rethomics` is implemented as a collection of small packages related to one another (Fig 1). This paradigm follows the model of modern frameworks such as the `tidyverse`, which results in increased testability and maintainability. In it, the different tasks of the analysis workflow (*i.e.* data import, manipulation and visualisation) are explicitly handled by different packages. At the core of `rethomics`, the `behavr` package offers a very flexible and efficient solution to store large amounts data (*e.g.* position and activity) as well as metadata (*e.g.* treatment, genotype and so on) in a single `data.table`-derived object. Any input package will import experimental data as a `behavr` table which can, in turn, be manipulated and visualised regardless of the original input platform. Results and plots integrate seamlessly within the `R` ecosystem, hence providing users with state-of-the-art visualisation and statistical tools.

Fig 1. The `rethomics` workflow. Diagram representing the interplay between, from left to right, the raw data, the `rethomics` packages (in blue) and the rest of the `R` ecosystem.

Internal data structure

We created `behavr` (Fig 2), a new data structure, based on the widely adopted `data.table` object, in order to address two challenges that are inherent to manipulating ethomics results.

Firstly, there could be very long (typically $k_i > 10^8, \forall i \in [1, n]$), multivariate (often, $q > 10$), time series for each individual. For instance, each series could represent variables that encode coordinates, orientation, dimensions, activity, colour intensity and so on, sampled several times per second, over multiple days. Therefore, data structure must be computationally efficient – both in term of memory footprint and processing speed.

Secondly, a large amount of individuals are often studied (typically $n > 100$). Each individual (i) is associated with metadata: a set of p "metavariables" that describe experimental conditions. For instance, metadata stores information regarding the date and location of the experiment, treatment, genotype, sex, *post hoc* observations and other arbitrary metavariables. It is good practice to record as many metavariables as possible so they can later be used as covariates. Therefore, typically $p > 10$.

`behavr` tables link metadata and data within the same object, extending the syntax of `data.table` to manipulate, join and access metadata. This approach guaranties that any data point can be mapped correctly to its parent metadata. It also allows implicit update of metadata when data is altered. For instance, when is data filtered, only the

remaining individuals should be in the new metadata. It is also important that metadata and data can interoperate. For instance, when one wants to update variable according to the value of a metavariable (say, alter the variable x only for animals with the metavariable $sex =$ "*male*").

Fig 2. `behavr` table. A: Illustration of a `behavr` object, the core data structure in `rethomics`. The metadata holds a single row for each of the n individuals. Its columns, the p metavariables, are one of two kinds: either required – and defined by the acquisition platform (*i.e.* used to fetch the data) – or user-defined (*i.e.* arbitrary). In the data, each row is a "read" (*i.e.* information about one individual at one time point). It is formed of q variables and is expected to have a very large number of reads k for each individual i. Data and metadata are implicitly joined with the `id` field. Note that the names used in this for variables and metavariable in this example are only plausible cases that will likely differ in practice. B: Non exhaustive list of uses of a `behavr` table (refered as `dt`). In addition to operations on data, which are inherited from `data.table`, we provide utilities designed specifically to act on both metadata and data.

Data import

Data import package translate results from a recording platform (*e.g.* text files and databases) into a single `behavr` object. Currently, we provide a package to read Drosophila Activity Monitor (DAM2) data and another one for Ethoscope data. Although the structure of the raw results if very different, conceptually, loading data is very similar. In all cases the user is asked to generate a metadata table (one row per individual). In it, there will be both mandatory and optional columns. The mandatory ones are the necessary and sufficient information to fetch data (*e.g.* machine id, region of interest and date). The optional columns are user-defined arbitrary fields that translate experimental conditions (*e.g.* treatment, genotype and sex).

In this respect, the metadata file is a standardised an comprehensive data frame describing an experiment. Using such a structure comes with multiple advantages. For instance, it simplify collaboration and data exchange as all treatments and individuals are very explicit. Then, it promotes good experimental practices such as interspersion of treatments (indeed, without it, users are tempted to simplify their design by, for instance, confounding device/location and treatment). Furthermore, it streamlines the inclusion and analysis of further replicates in the same workflow. Indeed, additional replicates can simply be added as new rows – and the ID of the replicate later used, if needed, as a covariate.

Visualisation

Long time series often need to be preprocessed before visualisation. Typically, users are interested in understanding individual or population trends over time. To integrate visualisation in `rethomics`, we implemented `ggetho`, a package extending the widely adopted `ggplot2` by providing preprocessing tools as well as new layers and scales. Our tools make full use of the internal `behavr` structure to deliver efficient representations of temporal trends. It particularly applies to the visualisation of long experiments, with the ability to, for instance, annotate light and dark phases, wrap time over a circadian day, display "double-plotted actograms" and periodograms. Importantly, `ggetho` is fully compatible with `ggplot2`.

Circadian and sleep analysis

The packages `zeitgebr` and `sleepr` provide tools to analyse circadian behaviours and sleep, respectively. Together, they offer a suite of methods to compute periodograms and find their peaks, score sleep from inactivity (*e.g.* using the "five minute rule"), and characterise the architecture of sleep bouts (*e.g.* number, length and latency).

Results

TODO description of the dataset here
* experiment goal
* design
* source
* express the idea that shis is just simple example that could be scaled up

First of all, we load the necessary libraries (see availability section for installation instructions):

```
library(damr)       # input DAM2 data
library(zeitgebr)   # periodogram computation
library(sleepr)     # sleep analysis
library(ggetho)     # behaviour visualisation
```

Then, the metadata file is read and linked to the `.txt` result files.

```
metadata <- link_dam2_metadata("metadata.csv",".")     # linking
# print(metadata)                                      # check metadata
dt <- load_dam2(metadata)                              # loading
summary(dt)                                            # quick summary
```

```
## behavr table with:
##  58 individuals
##  8 metavariables
##  2 variables
##  1.58722e+05 measurements
##  1 key (id)
```

Preprocessing

We notice, from the metadata, that the two replicates do not have the same time in baseline. We would like to express the time relative to the important event: the transition to `LL`. To do so, we subtract the `baseline_days` metavariable from the `t` variable. This gives us an opportunity to illustrate the use `xmv()`, which expands metavariables as variables. In addition, we use the `data.table` syntax to create, in place, a `moving` variable. It is `TRUE` when and only when `activity` is greater than zero:

```
# baseline subtraction -- note the use of xmv
dt[,t := t - days(xmv(baseline_days))]
dt[, moving :=  activity > 0]
```

```
summary(dt)

## behavr table with:
##  58 individuals
##  8 metavariables
##  3 variables
##  1.58722e+05 measurements
##  1 key (id)
```

To simplify visualisation, we create our own `label` metavariable, as combination of a number and `genotype`. In the restricted context of this analysis, `label` acts a unique identifier. Importantly, we keep `id` , which is more rigorous and universal.

```
dt[, label := interaction(1:.N, genotype), meta=T]
print(dt)
```

Curation

It is important to visualise an overview of how each individual behaved and, if necessary, alter the data accordingly. For this, we generate a tile plot (Fig 3A).

Fig 3. Experiment quality control. Tile plot representing the fraction of time spent moving as a colour intensity. Each individual is represented by a row and time, on the x axis, is binned in 30 minute.

```
# make a ggplot object with label on the y and moving on the z axis
fig3A <- ggetho(dt, aes(y = label, z = moving)) +
  # show data as a tile plot
  # that is z is a pixel whose intensity maps moving
  stat_tile_etho() +
  # add layers to draw annotations to show L and D phases
  # as white and black, respectivelly
  # the first layer is for the baseline (until t = 0)
  stat_ld_annotations(x_limits = c(dt[,min(t)], 0)) +
  # in the 2nd one, we start at 0 and use grey
  # instead of black as we work in LL
  stat_ld_annotations(x_limits = c(0, dt[, max(t)]),
                      ld_colours = c("white", "grey"))
```

Activity of dead or escaped animals is falsely scored as long series of zeros (see, for instance, individual 30 and 18 in Fig 3A). Our `sleepr` package offer a tool to detect and remove such artefactual data. The updated version can be visualised in Fig 3B.

```
# remove data after death
dt <- sleepr::curate_dead_animals(dt, moving)
# same as above
fig3B <- ggetho(dt, aes(y = label, z = moving)) +
    stat_tile_etho() +
    stat_ld_annotations(x_limits = c(dt[, min(t)], 0)) +
    stat_ld_annotations(x_limits = c(0, dt[, max(t)]),
                        ld_colours = c("white", "grey"))
```

For the purpose of this example, we keep only individuals that have *at least five days in LL*.

```
# for each id, we check for validity
valid_dt <- dt[, .(valid = max(t) > days(5)), by = id]
# a vector of all valid ids
valid_ids <- valid_dt[valid == T, id]
# filter dt with the valid ids
dt <- dt[id %in% valid_ids]
summary(dt)

## behavr table with:
##  52 individuals
##  9 metavariables
##  3 variables
##  1.40609e+05 measurements
##  1 key (id)
```

Note that as a result, we now have 52 "valid"individuals.

Double plotted actograms

"Double-plotted actograms" are a common visualisation of periodicity and rhythmicity in circadian experiments. In S1 FigA, we show the double-plotted actograms of each animals. A representative sample of eight individuals is shown in Fig 4A.

```
# we also show a subset of this figure in 4A
figS1A <- ggetho(dt, aes(z = moving), multiplot = 2) +
            # bars n the z axis could
            #  one could also use stat_tile_etho
            stat_bar_tile_etho() +
            # divide plot by individual
            facet_wrap( ~ label, ncol = 4) +
            # rename the y axis
            scale_y_discrete(name="Day")
```

Periodograms

Ultimately, in order to quantify periodicity and rhythmicity, we compute periodograms. Several methods are implemented in `zeitbebr`. In this example, we generate χ^2 periodrograms and lay them out in a grid. A subset of eight animals is shown in Fig 4B (see S1 FigB for all individuals).

```
dt_ll <- dt[t > days(1)]
```

```
per_dt <- periodogram(moving,
                      dt_ll,
                      resample_rate = 1/mins(10),
                      FUN=chi_sq_periodogram)

per_dt <- find_peaks(per_dt)
# we also show a subset of this figure in supplementary materials
figS1B <- ggperio(per_dt, aes(y = power, peak=peak)) +
                  # periodogram drawn as a line
                  geom_line() +
                  # the significance line in red
                  geom_line(aes(y = signif_threshold), colour="red") +
                  # point and text at the peak
                  geom_peak() +
                  # divide plot by individual
                  facet_wrap( ~ label, ncol = 4)
```

Fig 4. Visualisation of the periodicity in activity of eight representative animals. A: Double plotted actograms showing activity over the experiment. Transition to LL happens at day 0. B: χ^2 periodograms over the LL part of the experiment matching the animals in A. The blue annotation represents the first peak (if present) above the significance threshold (red line). Titles on top of each facet refer to the label allocated to each individual. A version of this figure with all animals is available S1 Fig.

Population statistics

Both double-plotted actograms and periodograms suggest that NKCCOX/+ flies are mostly arhythmic in LL whilst Tim/NKCCOX appear to have a consistent, long-period rhythm. To visualise this at the population scale, we can plot an average periodogram (see Fig 5A):

```
fig5A <- ggperio(per_dt, aes(y = power - signif_threshold,
                             colour = genotype)) +
          # periodogram shown as a line for population mean
          # and bootstrap error bars
          stat_pop_etho(method = ggplot2::mean_cl_boot) +
          # rename x and y axis
          scale_y_continuous(name="Relative power") +
          scale_x_hours("Period")
```

To further quantify this difference, we can show the value of the peak for both groups (see Fig 5B). First of all, we compute a summary per individual (`by=id`):

```
summary_dt <- per_dt[,.(
```

Fig 5. Population statistics on circadian phenotype. A: Average periodograms. The aggregated relative power of the periodogram of all animals. The solid lines and the shaded areas show population means and their 95% bootstrap confidence interval, respectively. B: Peak periodicity power and average. Values of the peak period for animals with a significant peak. Individual animals are shown by dots whose size represent power of the peak period. The error bars are 95% bootstrap confidence interval on the population mean. C: Frequencies of rhythmic animals. Number of rhythmic animals (*i.e.* with a significant peak) in each genotypes. Dark and clear fillings indicate rhythmic and arhythmic animals, respectively.

```
                        first_peak_period = period[peak==1],
                        first_peak_power = power[peak==1],
                        is_rhythmic = any(peak==1)),
                      by=id]


# rejoin metadata
summary_dt <- rejoin(summary_dt)
```

`summary_dt` is just a regular data frame with one row per individual, containing both metadata and our summary statistics. It can therefore be used directly by `ggplot` and other tools:

```
# standard ggplot
fig5B <- ggplot(summary_dt, aes(y = first_peak_period,
                                x = genotype)) +
              # draw the mean of each genotype group
              stat_summary(fun.y = mean, geom = "point", shape=3) +
              # draw bootstrap confidence intervals
              stat_summary(fun.data = mean_cl_boot, geom = "errorbar") +
              # shows all individuals as points.
              # the size of the point expresses the power of the peak
              geom_jitter(aes(colour= genotype,
                              size=first_peak_power),
                          alpha=.67) +
              # We would like to convert time in hour
              scale_y_hours("Period")
```

To convey that Tim/NKCCOX mutants are more rhythmic, we could represent the proportion of rhythmic flies (see Fig 5C):

```
# standard ggplot
fig5C <- ggplot(summary_dt, aes(x = genotype,
                                fill = genotype,
                                alpha = is_rhythmic
                                )) +
              geom_bar(colour="black")
```

`R` provides one of the richest statistical toolbox available, which allows users to go deeper in he analysis of the extracted variables. One could, for instance, perform a χ^2 test on the number of rhythmic *vs* arhythmic flies in both genotypes. To address the

same question, we fit a binomial generalised linear model, which shows a strong positive effect of genotype Tim/NKCCOX on the probability of being rhythmic (p-value $< 10^{-5}$):

```
fit <- glm(is_rhythmic ~ genotype, summary_dt, family="binomial")

summary(fit)$coefficients

##                     Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)        -1.504077  0.5527708 -2.720978 6.508902e-03
## genotypeTim/NKCCOX  4.143135  0.9172057  4.517127 6.268433e-06
```

Lastly, we can generate a table that compute arbitrary population statistics for each genotype:

```
result_dt <- summary_dt[,
          .(
            mean_period = mean(first_peak_period, na.rm = T) / hours(1),
            sd_period = sd(first_peak_period, na.rm = T) / hours(1),
            n_rhythmic = sum(is_rhythmic),
            n = .N
            ),
          by = genotype
          ]
result_dt

##      genotype mean_period sd_period n_rhythmic  n
## 1:   NKCCOX/+    25.92500  4.133098          4 22
## 2: Tim/NKCCOX    26.36071  1.834513         28 30
```

Some conclusion here, TODO:
* high scalability
* reproducibility
* from raw data to publication-quality figures
* flexibility

Availability and Future Directions

All packages in the `rethomics` framework are available under the terms of the GPLv3 license and listed at https://github.com/rethomics/. Extensive installation instructions as well as reproducible demos and tutorials are available at https://rethomics.github.io/. All packages are continuously integrated and unit tested on several version of `R` to minimise the risk of present and future issues.

* Other inputs
* Position analysis
* GUI
* ...

Supporting information

S1 Fig. Complete version of Fig 4. See Fig 4 for legend.

Acknowledgements

TODO:
Han Kim
Maite
Hannah, Alice, Diana
Markus, Patrick Krätschmer Rob

References

1. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008 Dec;9(12):938–950.

2. Ohno S. Evolution by gene duplication. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.; 1970.

3. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet. 2011 Oct;7(10):e1002337.