GCalignR. An R package for aligning Gas-Chromatography data

by Meinolf Ottensmann, Martin A. Stoffel, Joseph I. Hoffman

Abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract

Introduction

Chemical cues are arguably the most common mode of communication among animals (Wyatt, 2014). By exploring broad patterns in complex chemical signatures, researchers make inferences on kinship (Krause et al., 2012; Stoffel et al., 2015), genetic diversity (Charpentier et al., 2010; Leclaire et al., 2012), sexual maturation (Caspers et al., 2011) or species discrimination (de Meulemeester et al., 2011). One of the most common instruments to quantify the chemical composition of samples is gas-chromatography, a fast high-throughput method to detect individual chemicals and their abundances (McNair and Miller, 2011), while the additional implementation of mass-spectrometry (GC-MS) allows to identify specific substances (Caspers et al., 2011).

However, before similarity patterns across can be analysed, it is essential to align compounds among samples, thereby accounting for drifts in the retention times of peaks caused by subtle, random and often unavoidable variations of the chromatography machine parameters (Pierce et al., 2005). Many studies rely on manual alignment rather than (semi-)automated algorithms (citations necessary), but this approach bears three severe drawbacks: (1) In large scale studies this task becomes increasingly time consuming task and is impracticable. (2) Humans are prone to detect patterns in noise which is why the researcher may bias the alignment due to subjective experience and expectations. (3) The data analytic pipeline from the raw gas-chromatography data to the results of the statistical analysis is not reproducible. (citations for the first two points necessary) Several alignment algorithms have been proposed to overcome these issues, but these focus nearly exclusively on GC-MS data (Pierce et al., 2005; Robinson et al., 2007; Jiang et al., 2013) and only some a easily accessible as web-based tools (Hoffmann and Stoye, 2009; Wang et al., 2010) or independent software (Dellicour and Lecocq, 2013).

Here, we introduce **GCalignR**, a package that implements a simple and fast algorithm to align peaks from GC data and evaluate the resulting alignment using two data sets. **GCalignR** was specifically developed as a tool for pre-processing GC data from animal skin and preen glands prior to subsequent statistical analysis. In brief, the algorithm consists of two main steps: (1) Systematic shifts of chromatograms are corrected by applying appropriate linear shifts to whole chromatograms based on a single reference. (2) Retention times of individual peaks are grouped iteratively together with homologous peaks of other samples and aligned within the same row in a retention time matrix . The quality of this grouping procedure can be adjusted to specific datasets through three parameters that are described in detail below. Among several optional processing steps, the package allows to remove peaks belonging to contaminations, which are identified due to their presence in control samples. For an easy interpretation of the quality of an alignment we implemented several ways to plot the outcome (You can change this to something more specific). Furthermore, we demonstrate a complete workflow from chemical raw data to multivariate analyses with the popular and widely used **vegan** (Oksanen et al., 2016) package. This allows the integration of the full analysis into **RMarkdown** documents (Allaire et al., 2016) in order to meet the standards of reproducibility (Peng, 2011).

The Package

GCalignR consists of functions that allow the alignment of peaks from GC and GC-MS data based on retention times. The main aim of the package is to provide a simple tool that guides the user through the alignment of large data sets prior to hypothesis-testing of the multivariate data (Anderson, 2001). We summarise the underlying algorithm and workflow (figure 1) below and refer to the vignette that can be assessed via `browseVignettes('GCalignR')`.
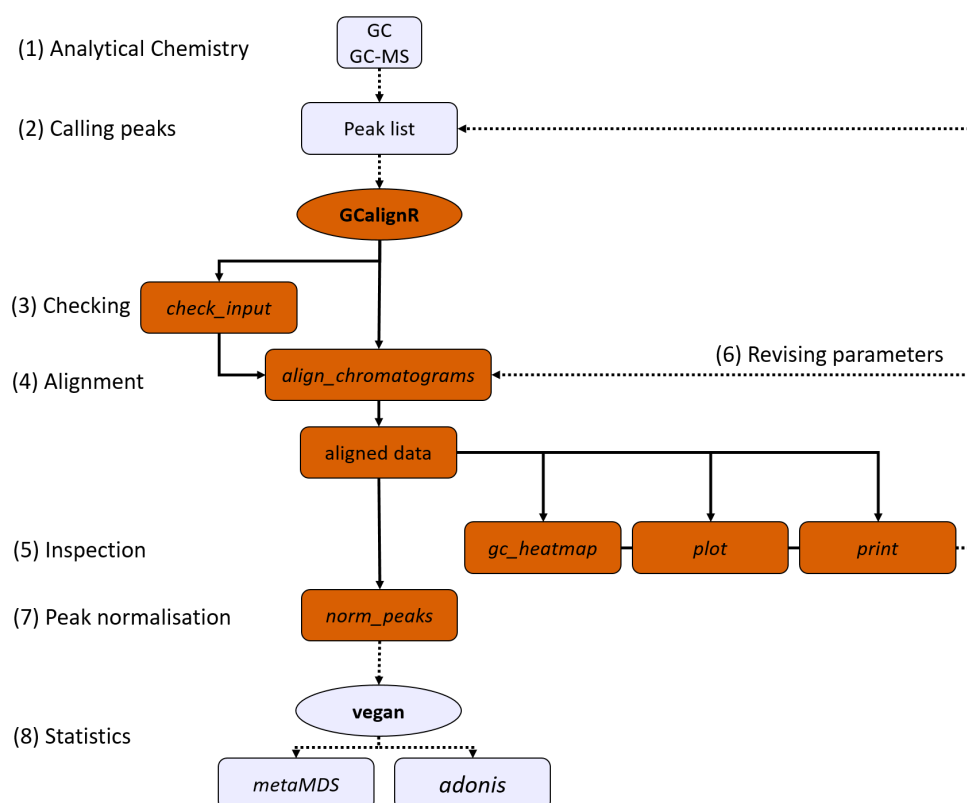
Figure 1: GCalignR workflow. In addition to the alignment of substances across samples, the package provides functions for checking and inspecting the data. The aligned data is ready to use for analyses in conjunction with other packages. Each function is explained within the text.

Example datasets

For demonstration purposes **GCalignR** includes data of chemical signatures that were obtained by sampling the skin of 82 Antarctic fur seals *Arctocephalus gazella*. It was previously shown that these signatures encode the membership to a breeding colony Stoffel et al. (2015). These data are available as a *List* with individual samples included as a *data.frame*. Two variables are available that represent the required retention time ("time") and concentration or peak abundance ("area") within a sample.

```
library(GCalignR)
# Seal scent data
data("peak_data")
# Data is organized in one list of data.frames
str(peak_data[1:2])

#> List of 2
#>  $ C3:'data.frame':  217 obs. of  2 variables:
#>   ..$ time: num [1:217] 4.53 4.55 4.62 4.68 4.71 4.79 4.83 4.87 5.01 5.14 ...
#>   ..$ area: num [1:217] 3331224 1462381 4834211 7754401 1267617 ...
#>  $ C2:'data.frame':  217 obs. of  2 variables:
#>   ..$ time: num [1:217] 4.52 4.55 4.57 4.67 4.69 4.73 4.75 4.8 4.83 4.85 ...
#>   ..$ area: num [1:217] 2695110 5926253 10406833 6805905 1672849 ...
```

The second data set is comprised of three bumble bee species *Bombus bimaculatus*, *B. ephippiatus* and *B. flavifrons* where signatures have been obtained from cephalic labial gland secretions of 24, 20 and 11 individuals respectively (Dellicour and Lecocq, 2013). Cephalic secretions haven been demonstrated to aid in species identification (de Meulemeester et al., 2011). These data are available in a text file (figure 2) in the format of typical output from proprietary peak detection software (e.g. XCalibur, Thermofischer Scientific or Labsolution, Shimadzu). Here, the names of samples and variables are given in the first two rows, while the peak data is included from below with single matrices concatenated horizontally.

S1	S2	S3	S4				
time	area						
8.30	0.47	7.58	0.03	7.59	0.04	9.00	0.15
9.00	0.13	8.31	0.35	8.32	0.26	8.30	0.18
11.46	0.27	11.47	0.42	8.99	0.46	11.19	0.20
18.11	0.13	12.19	0.10	11.48	0.23	11.48	0.38
18.20	0.15					15.30	0.09

Figure 2: Text files are the standard input of **GCalignR**

Checking the input

The package provides the function **check_input** checks the input for typical formatting errors and incomplete data. We encourage to use unique names for samples that consist only of letters, numbers and underscores. If the data fails the test, indicative warnings are returned which guide in correcting the errors. In addition, the argument *list_peaks* allows to create a barplot of the initial peak distribution to get an overview of the data.

```
check_input(peak_data)

#> All checks passed!
#> Ready for processing with align_chromatograms

#> All checks passed!
#> Ready for processing with align_chromatograms
```
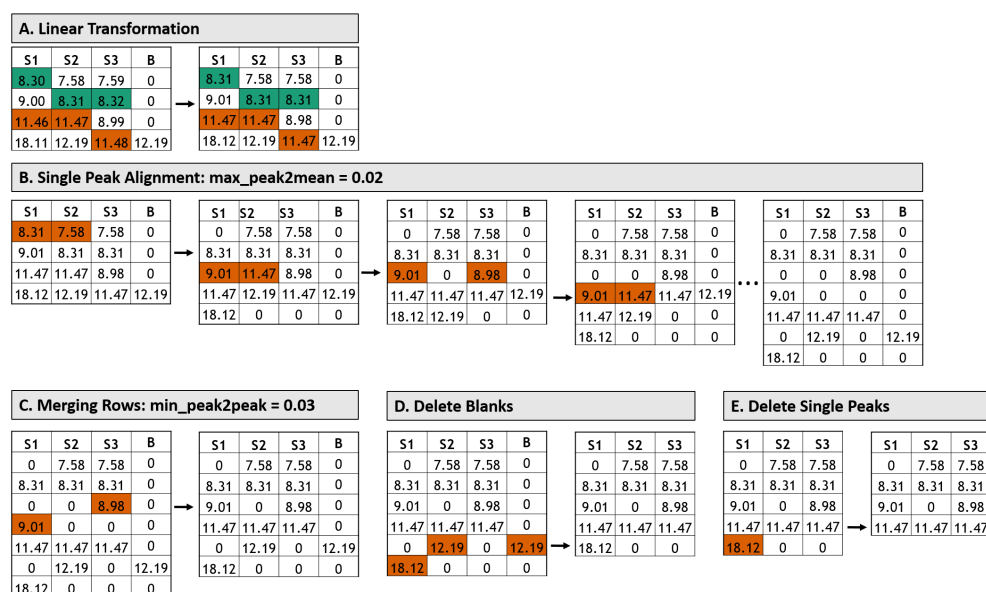
Aligning substances among samples

The alignment procedure is divided into five steps (figure 3). All steps are executed by the main function *align_chromatograms* and will be explained in in the next sections.

Linear adjustments of chromatograms

At first, chromatograms are linearly shifted with respect to a reference sample to account for systematic shifts in retention times among homologous chemicals shared by samples. Therefore, the same small linear adjustments are applied to the entire set of peaks in a chromatogram (figure 3 A), such that the number of peaks that are shared between the sample and the reference at a given threshold (i.e. 0.6 seconds) is maximised.The parameter *max_linear_shift* defines the maximum of linear shifts that are considered by the program.
Note: This method relies on the occurrence of substances that are shared among most substances to produce efficient adjustments. If those are absent, it is unlikely to find a suitable shift and chromatograms remain untransformed.

A reference may be selected automatically by searching for the sample with the highest average similarity to all other samples based on the number of shared peaks prior to alignment. Alternatively, a chromatogram may be included that contains peaks of an internal standard which peaks are *a-priori* known to occur in all samples. In this case, the sample should be named "reference" and will be removed after the alignment was conducted.

A. Linear Transformation

S1	S2	S3	B
8.30	7.58	7.59	0
9.00	8.31	8.32	0
11.46	11.47	8.99	0
18.11	12.19	11.48	12.19

→

S1	S2	S3	B
8.31	7.58	7.58	0
9.01	8.31	8.31	0
11.47	11.47	8.98	0
18.12	12.19	11.47	12.19

B. Single Peak Alignment: max_peak2mean = 0.02

S1	S2	S3	B
8.31	7.58	7.58	0
9.01	8.31	8.31	0
11.47	11.47	8.98	0
18.12	12.19	11.47	12.19

→

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
9.01	11.47	8.98	0
11.47	12.19	11.47	12.19
18.12	0	0	0

→

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
9.01	0	8.98	0
11.47	11.47	11.47	12.19
18.12	12.19	0	0

→

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
0	0	8.98	0
9.01	11.47	11.47	12.19
11.47	12.19	0	0
18.12	0	0	0

...

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
0	0	8.98	0
9.01	0	0	0
11.47	11.47	11.47	0
0	12.19	0	12.19
18.12	0	0	0

C. Merging Rows: min_peak2peak = 0.03

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
0	0	8.98	0
9.01	0	0	0
11.47	11.47	11.47	0
0	12.19	0	12.19
18.12	0	0	0

→

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
9.01	0	8.98	0
11.47	11.47	11.47	0
0	12.19	0	12.19
18.12	0	0	0

D. Delete Blanks

S1	S2	S3	B
0	7.58	7.58	0
8.31	8.31	8.31	0
9.01	0	8.98	0
11.47	11.47	11.47	0
0	12.19	0	12.19
18.12	0	0	0

→

S1	S2	S3
0	7.58	7.58
8.31	8.31	8.31
9.01	0	8.98
11.47	11.47	11.47
18.12	0	0

E. Delete Single Peaks

S1	S2	S3
0	7.58	7.58
8.31	8.31	8.31
9.01	0	8.98
11.47	11.47	11.47
18.12	0	0

→

S1	S2	S3
0	7.58	7.58
8.31	8.31	8.31
9.01	0	8.98
11.47	11.47	11.47

Figure 3: Overview of the algorithm performed by GCalignR. Row of matrices correspond to substances, columns are samples. Zeros indicate absence of peaks and are ignored in calculations. **A**. Chromatograms are linearly shifted with respect to a reference (S2). StrongB. From left to right the first four steps from the input matrix to the final alignment are shown. Peaks are aligned row by row. Initially, always the second sample is compared to the first. Then the next sample is compared to all samples in previous columns until the last column is reached. Coloured cells represent conflicting retention times using *max_peak2mean = 0.02*. **C**. After all peaks have been aligned, rows are merged depending on *min_peak2peak*, which defines the minimum difference that is expected between substances. If merging does not result in the loss of any data, rows are merged. **D**. If specified, all peaks found in one or more blanks are removed as well as the blank itself. **E**. Unique peaks which are present in only a single individual are not of interest for similarity analyses and can be removed as well.

Piece-wise alignment of substances

The core of the alignment procedure is based on clustering of individual peaks among samples. This is performed by examining retention times within single rows, where samples are compared consecutively with all previous samples starting with the second column (figure 3 B):

$$rt_m > \left(\frac{\sum_{i=1}^{m-1} rt_i}{m - 1} \right) + max\_peak2mean \qquad (1)$$

If the examined peak is moved into the next row, whereas all previous samples are moved

$$rt_m < \left(\frac{\sum_{i=1}^{m-1} rt_i}{m - 1} \right) - max\_peak2mean \qquad (2)$$

with *rt* = retention time; *m* = current column and *max_peak2mean* defining the maximal deviation of the mean retention time.

By considering the mean retention time among all previous samples the algorithm accounts for substance specific variations, such that less variable retention times are treated more stringent than chemicals exhibiting higher variability. Once the last retention time of a row was evaluated the whole procedure is repeated with the next row until the end of the retention time matrix was reached. Afterwards, rows with similar mean retention times are assessed for redundancy (figure 3 C), which applies whenever a merging does not cause any loss of any information (i.e. no sample exists that contains substances in both rows). The similarity threshold is given by *min_peak2peak* defining the minimal difference between peaks that is expected.

In combination these two fairly simple algorithms align compounds by considering compound-specific variation in both steps.

Normalisation

Many multivariate analysis techniques, like those available in **vegan**, require a data frame of independent variables as input format. Moreover is generally advisable to normalise abundancies prior to statistical analysis to correct for variations in the total concentration of samples. This is utilised in **GCalignR** function `normalise_peaks` which normalises peak abundancies by calculating realtive abundancies for each sample.

Workflow

Here, we demonstrate the typical workflow using our seal data. This is done using the function `align_chromatograms`. A list of all parameters and their description can be assessed from the documentation in the helpfile by typing `?align_chromatograms`:

```
seal_aligned <- align_chromatograms(data = peak_data,
                    conc_col_name = "area",
                    max_diff_peak2mean = 0.02,
                    min_diff_peak2peak = 0.05,
                    max_linear_shift = 0.05,
                    rt_col_name = "time",
                    delete_single_peak = T,
                    blanks = c("C2","C3"))
```

Now, we can inspect the results by retrieving summaries of the alignment process. The printing method summarises the function call including defaults that have not been explicitly specified during the function call. We also get the relevant information to retrace every step in the alignment:

```
print(seal_aligned)

#>    Summary of Peak Alignment running align_chromatograms from package GCalignR
#>    Input: peak_data    Start:  2016-08-16 17:21:23     Finished:  2016-08-16 17:29:59
#>
#> Call:
#>    GCalignR::align_chromatograms(data=peak_data, conc_col_name=area,
#>    rt_col_name=time, rt_cutoff_low=8, reference=M29, max_linear_shift=0.05,
#>    max_diff_peak2mean=0.02, min_diff_peak2peak=0.03, blanks=(C2,
#>    C3), delete_single_peak=T, sep=\t, rt_cutoff_high=NULL, n_iter=1,
#>    merge_rare_peaks=FALSE)
#>
#> Summary of scored substances:
#>
#>     Peaks  In_Blanks  Singular  Retained
#>      416       119        70       227
#>
#>    In total 416 substances were identified among all samples. NA substances were
#>    present in blanks. The corresponding peaks as well as the blanks were removed
#>    from the data set. 70 substances were present in just one single sample and were
#>    removed. 227 substances are retained after all filtering steps.
#>
#> Sample Overview  The following 84 Samples were aligned to the reference 'M29':
#>    M2, M3, M4, M5, M6, M7, M8, M9, M10, M12, M14, M15, M16, M17, M18, M19, M20,
#>    M21, M23, M24, M25, M26, M27, M28, M29, M30, M31, M33, M35, M36, M37, M38, M39,
#>    M40, M41, M43, M44, M45, M46, M47, M48, P2, P3, P4, P5, P6, P7, P8, P9, P10,
#>    P12, P14, P15, P16, P17, P18, P19, P20, P21, P23, P24, P25, P26, P27, P28, P29,
#>    P30, P31, P33, P35, P36, P37, P38, P39, P40, P41, P43, P44, P45, P46, P47, P48
#>
#> For further details:
#>    Type 'gc_heatmap(seal_aligned)' to retrieve a heatmap for the alignment accuracy
#>    Type 'plot(seal_aligned)' to retrieve further diagnostic plots
```

The quality of an alignment will depend on sensible parameters that facilitate the (i) correction of linear shifts that might fall in a larger range with increasing sample size and (ii) and the variability of retention times. Optimally, linear shifts do not exhaust the range given by `max_linear_shift` completely, which would in turn indicate that not all uncertainties haven been fully compensated for. This can be assessed by some diagnostic plots:

Bibliography

J. J. Allaire, J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, H. Wickham, A. Atkins, and R. Hyndman. rmarkdown: Dynamic documents for r, 2016. URL https://CRAN.R-project.org/package=rmarkdown. [p1]

M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001. ISSN 14429985. doi: 10.1111/j.1442-9993.2001.01070.pp.x. [p1]

B. A. Caspers, F. C. Schroeder, S. Franke, and C. C. Voigt. Scents of adolescence: the maturation of the olfactory phenotype in a free-ranging mammal. *PloS one*, 6(6):e21162, 2011. [p1]

M. J. Charpentier, J. C. Crawford, M. Boulet, and C. M. Drea. Message 'scent': Lemurs detect the genetic relatedness and quality of conspecifics via olfactory cues. *Animal Behaviour*, 80(1):101–108, 2010. ISSN 00033472. doi: 10.1016/j.anbehav.2010.04.005. [p1]

T. de Meulemeester, P. Gerbaux, M. Boulvin, A. Coppée, and P. Rasmont. A simplified protocol for bumble bee species identification by cephalic secretion analysis. *Insectes Sociaux*, 58(2):227–236, 2011. ISSN 0020-1812. doi: 10.1007/s00040-011-0146-1. [p1, 2]

S. Dellicour and T. Lecocq. Gcaligner 1.0: an alignment program to compute a multiple sample comparison data matrix from large eco-chemical datasets obtained by gc. *Journal of separation science*, 36(19):3206–3209, 2013. ISSN 1615-9306. doi: 10.1002/jssc.201300388. URL http://onlinelibrary.wiley.com/doi/10.1002/jssc.201300388/full. [p1, 2]

N. Hoffmann and J. Stoye. Chroma: signal-based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics (Oxford, England)*, 25(16):2080–2081, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp343. [p1]

W. Jiang, Z.-M. Zhang, Y. Yun, D.-J. Zhan, Y.-B. Zheng, Y.-Z. Liang, Z. Y. Yang, and L. Yu. Comparisons of five algorithms for chromatogram alignment. *Chromatographia*, 76(17-18):1067–1078, 2013. ISSN 0009-5893. doi: 10.1007/s10337-013-2513-8. [p1]

E. T. Krause, O. Krüger, P. Kohlmeier, and B. A. Caspers. Olfactory kin recognition in a songbird. *Biology letters*, 8(3):327–329, 2012. ISSN 1744-9561. [p1]

S. Leclaire, T. Merkling, C. Raynaud, H. Mulard, J.-M. Bessière, É. Lhuillier, S. A. Hatch, and É. Danchin. Semiochemical compounds of preen secretion reflect genetic make-up in a seabird species. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1731):1185–1193, 2012. [p1]

H. M. McNair and J. M. Miller. *Basic gas chromatography*. John Wiley & Sons, 2011. ISBN 1118211200. [p1]

J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. vegan: Community ecology package, 2016. URL https://CRAN.R-project.org/package=vegan. [p1]

R. D. Peng. Reproducible research in computational science. *Science (New York, N.Y.)*, 334(6060): 1226–1227, 2011. ISSN 0036-8075. doi: 10.1126/science.1213847. [p1]

K. M. Pierce, J. L. Hope, K. J. Johnson, B. W. Wright, and R. E. Synovec. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *Journal of Chromatography A*, 1096(1):101–110, 2005. [p1]

M. D. Robinson, D. P. de Souza, W. W. Keen, E. C. Saunders, M. J. McConville, T. P. Speed, and V. A. Likić. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC bioinformatics*, 8(1):419, 2007. ISSN 1471-2105. [p1]

M. A. Stoffel, B. A. Caspers, J. Forcada, A. Giannakara, M. Baier, L. Eberhart-Phillips, C. Müller, and J. I. Hoffman. Chemical fingerprints encode mother–offspring similarity, colony membership, relatedness, and genetic quality in fur seals. *Proceedings of the National Academy of Sciences*, 112(36): E5005–E5012, 2015. [p1, 2]

S.-Y. Wang, T.-J. Ho, C.-H. Kuo, and Y. J. Tseng. Chromaligner: a web server for chromatogram alignment. *Bioinformatics (Oxford, England)*, 26(18):2338–2339, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq337. [p1]

T. D. Wyatt. *Pheromones and animal behavior: chemical signals and signatures.* Cambridge University Press, 2014. ISBN 1107647495. [p1]

Meinolf Ottensmann
Department of Animal Behaviour
Bielefeld University
Morgenbreede 45
33615 Bielefeld
Meinolf.Ottensmann@web.de

Martin A. Stoffel
Department of Animal Behaviour
Bielefeld University
Morgenbreede 45
33615 Bielefeld
Martin.Adam.Stoffel@gmail.com

Joseph I. Hoffman
Department of Animal Behaviour
Bielefeld University
Morgenbreede 45
33615 Bielefeld
j_i_hoffman@hotmail.com