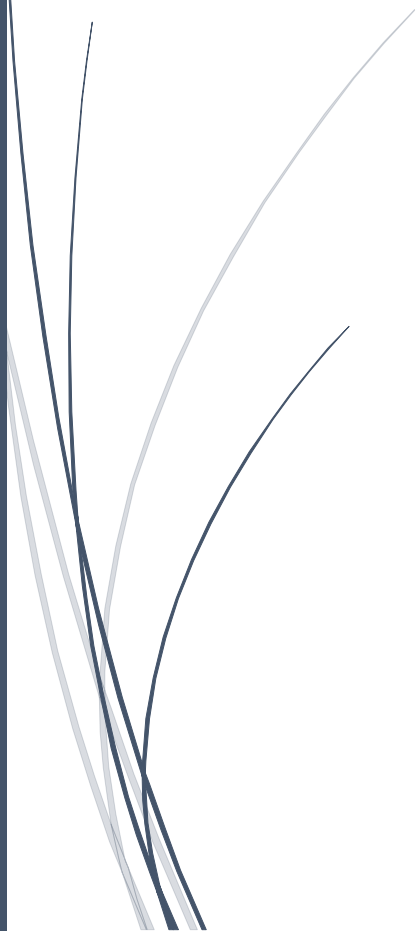
Elmshorn, December 3, 2021

Labrotation report

# Retrieving taxonomic information based on species names and incorporation into a heatmap

First examiner: Prof. Dr. Andrew Torda

Supervisor: Prof. Dr. Andrew Torda



ZBH - Center for Bioinformatics

By Daniel Waschestjuk

Matriculation number: 7386269

# Table of Contents

# 1. Introduction

Multiple sequence alignments carry a wealth of evolutionary information. At one level, they reflect the conservation of individual sites, but on the longer evolutionary time scale, they reflect domain structure. Unfortunately, large multiple sequence alignments can be almost impossible to interpret. Given 1000 columns or 1000 rows, there is no way to see any structure in the alignments. The aim of this work was to improve methods for summarising alignments with informative diagrams.

Summarized multiple sequence alignments play an important role in finding out functions and structures of proteins. These often lead to fundamental biological insights into the sequence-structure-function relationship of protein sequence families (Zhao, Guo, Sheng, & Shyr, 2014). Such a representation is possible in the form of a so-called heatmap. To do this, a FASTA file is imported (see Figure 1), in which all gaps within a sequence are converted to "0" and all amino acids are converted to "1". Next, specific colors are assigned to the two digits and then displayed in an image.

```
> XP_013146865.1 PREDICTED: ganglioside-induced differentiation-associated protein 1 [Papilio polytes]
------------------------------------------------------------
----------------------------------M---------H---YV----------
--------QKYLEK------Y---------------QLIKT-----------TN------
--N-K----M-----------S-----N-G--S-N-T----------NIYLYCN------
----YYSF-Y----S---Q---------K-------------------------------
-----------------------V---L-----M-TL-----------YEKNV-----DF
EPLIVD--IT----KGEQYSPWFLE-LNPRG-EVPVLKVRNEVIP--DSTRIIDYLEYHL
-DQ-----------------------------------------EL----------TP
----II-----NVSR-DSKV---------IK--NINR-----F----RDLLE----A-LP
AGVIT--------------------
> XP_033159913.1 ganglioside-induced differentiation-associated protein 1 [Drosophila mauritiana]
------------------------------------------------------------
----------------------------------M---------S---EQ----------
--------AKEI---------------------EALP-PTL-----------QD------
--F-K----A-----------P-----D-L--P-A-N----------KPVLFFH------
----PYNF-H----A---Q---------K-------------------------------
-----------------------V---L-----M-VF-----------YEKKI-----DF
FPYVVD--LC----NGEQYSNWFLN-LNPKG-DVPVLQDGALVIP--SSTHIINYVESKF
-RG-----------------------------------------D----------RS
----LK-----P-AH-NSKE---------FD--QMLV-----F----EQAMA----R-LP
VGTLSL--------------------
```

**Figure 1: Appearance of a FASTA file using the example of 2 sequences**

Partial section of a FASTA file of 2472 sequences. You can see the single-letter amino acid sequence for the ganglioside-induced differentiation-associated protein 1 for the two organisms *Papilio polytes* and *Drosophila mauritiana*. The numerous gaps are the result of aligning these sequences with another 2470 sequences. Above each sequence there is a header which contains a unique ID (identification number), a description of the sequence and a (unique) species name. It precedes the sequence data and starts with a greater-than sign (">").

A heatmap is used to see how a protein is encoded in different organisms. Among other things, it can be seen which regions overlap between organisms and which regions are additionally read for the protein in certain species. Figure 2 shows a heatmap for ganglioside-induced differentiation-associated protein 1 with 10 sequences. Here it is important to note that these are from a multiple sequence alignment of 2472 sequences. It is striking that we can see *Homo sapiens* several times. These are different isoforms of the protein. Furthermore, we see that all sequences from position 385 to 432 have the same amino acid pattern. However, this does not mean that the individual amino acid residues are the same. Rather, it shows that there are no insertions or deletions in this region. These could be conserved regions of the protein which form the active site for example, but 10 sequences are not sufficient to confirm that these are indeed conserved regions. The x-axis shows the positions of the individual gaps and amino acid residues. However, these positions have little significance because many gaps were added due to the multiple sequence alignment. For example, *Centrocerus urophas* encodes the protein with only 72 amino acid residues. Nevertheless, for the sequence alignment, more than 300

gaps had to be positioned before the start so that this sequence would fit on top of the other sequences. On the left side of the image a so-called dendrogram can be seen. This is a tree diagram that was created using the existing sequences by forming the individual branches through clustering observations. It is used to show the similarity levels of the sequences.

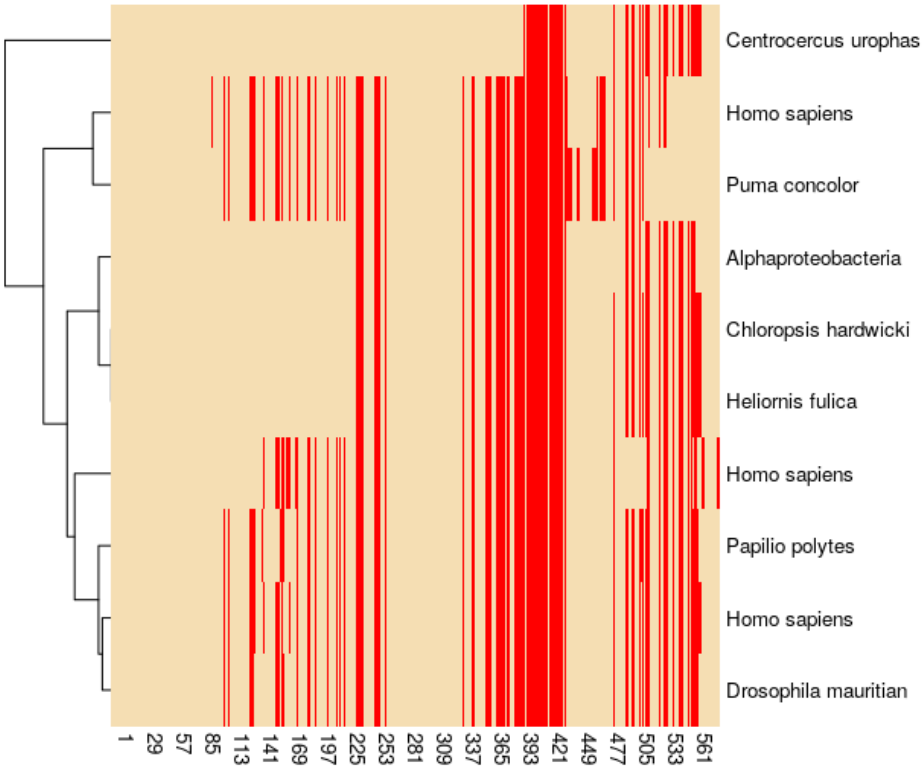



Figure 2: Typical heatmap of 10 sequences from a FASTA file

This is a heatmap for ganglioside-induced differentiation-associated protein 1 with 10 sequences from a multiple sequence alignment of 2472 sequences. Gaps are shown in wheat and amino acid residues in red. On the left side of the heatmap is a dendrogram whereas on the right side are the corresponding species names. The x-axis shows the positions from the FASTA file.

## 1.1 The problem

In the introduction, the heatmap was explained using 10 sequences. In practice, however, several thousands of sequences are used. In such cases, researchers are not interested in small details, but try to identify blocks of residues and possible domains. This leads to the question which sequences belong together to a certain taxonomic level. The hierarchy of biological classification includes a total of 7 levels: kingdom, phylum, class, order, family, genus, and species (see Figure 3). These range from a very general level (kingdom) to a very specific level (species). Since the individual species names provide only limited information (in the case that these are known) about their taxonomic information, it is a challenge to summarize the interesting aspects in a heatmap with several thousand sequences. Thus, researchers have to look up species name manually in the database to obtain taxonomic information and make statements. This problem can only be solved if it would be possible to return automatically information from a taxonomic level that a user prefers for all sequences to recognize exactly these blocks of residues and possible domains.

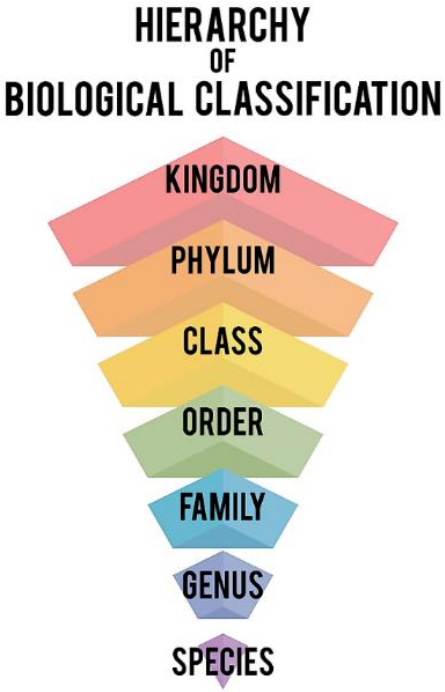



**Figure 3: Hierarchy of biological classification**

**Kingdom**: is the highest taxonomic level. This includes the kingdoms Animalia, Plantae, Fungi, Protista, Archaea, and Bacteria. However, the classification of the kingdoms is still controversial. It is currently undergoing further revision. **Phylum**: is the next rank after kingdom. There are 35 phylums in the kingdom Animalia for example. **Class**: was the most general ranking proposed by Linnaeus. In the kingdom Animalia there are 108 classes, including Mammalia (mammals), Aves (birds), and Reptilia (reptiles), among many others. **Order**: is the next highest level. For example, there are between 19-26 orders of mammals, depending on how the organisms are classified. Some orders of mammals are primates, cetaceans (whales, dolphins, and porpoises), Carnivora (large carnivores and omnivores), and Chiroptera (bats). **Family**: are more specific in their turn. There are a total of 12 families in the order Carnivora. **Genus**: is the first part of the scientific name of an organism in binomial nomenclature; the second part is the species name. Homo is the genus name, while sapiens is the species name. **Species**: are the most specific taxonomic category. There are an estimated 8.7 million different species of organisms on Earth, but the vast majority have yet to be discovered and categorized (Baron, 1996).

## 1.2 Design considerations

### 1.2.1 Concepts

First of all, the program must be able to extract the species names from the comment lines of a FASTA file. Then, these names are used to query a database for the taxonomic information. This must tolerate cases where there is no information, incomplete information, ambiguous information, or multiple different answers. After the database search is complete, the user should determine which rank (for heatmap3) or which ranks (for pheatmap) should be colored in the heatmap. Here, also typing errors should be checked. There may be hundreds of families for example in the case of several thousand sequences, so the user should be able to determine how many groups of the selected taxonomic level should be displayed, with the groups sorted by size and thus the largest groups are selected. Accordingly, the colors are chosen to match the number of groups. Next, all other groups of the taxonomic selection must be changed to "Others". After this is done, the user should be able to decide whether the group "Others" should be shown, in order to be able to see also unavailable data (are not marked), or hidden, so that they do not disturb, and the selected groups can be viewed more focused. Finally, this entire selection must be linked to the heatmap and displayed in a clearly visible manner, so without overlapping legends for example.

### 1.2.2 Practical issues

R is a flexible programming language used for statistical data analysis and graph creation. R facilitates high-quality plotting and is committed to aesthetic and visually appealing graphs. This feature sets R apart from other programming languages. For this reason, it is necessary to work in R for heatmaps. On the other hand, R also has some disadvantages. These include, among other things, data processing. R stores objects in physical memory. Moreover, all the data is needed in a single place, the memory. For this reason, R is not ideally suited to handle very large amounts of data. However, this can be remedied with data management packages. In addition, R is not easy to learn, so it can be difficult for people who do not have any programming knowledge to learn R (DataFlair 2021).

### 1.2.3 Package choice

As researchers increasingly work with greater numbers and diversity of species, the challenge of ensuring consistent taxonomy has become acute. These problems have long been recognized in the literature (Boyle et al. 2013; Maldonado et al. 2015; Remsen 2016), and a growing number of databases and tools have emerged in recent decades (Roskov Y. 2018; Alvarez and Luebert 2018), although it remains difficult to obtain taxonomic names in a transparent, efficient and automatable manner. To address this problem, the R package taxadb was developed. Previously, there were packages which returned taxonomic information by constructing individual web queries at the R level. This has several major drawbacks. First, constant Internet access is required. Second, the queries are slow and inefficient to implement and execute. Thus, separate API calls are often required for each taxonomic name. Furthermore, the API design severely limits the types of queries that can be made, so it is usually not possible to make queries over the entire classification levels. Inconsistent query formats as well as responses between different name providers makes it difficult to apply scripts developed for one provider to scripts developed for other providers. Last, most queries are not reproducible due to the dependency of the results on the state of the central server. This is because many of the name providers update the server data either continuously or periodically, including revisions to existing names (in terms of spelling or changes in the accepted name designation) and the addition of new names (Boettiger C., Chamberlain S. and Norman K . 2020). Taxadb, in contrast, offers the advantages of not using existing web APIs, but instead automatically downloading a set of compressed text files, importing them, and storing them in a local database. By using a local database managed by R, this allows taxadb

R users to interact with large data files without consuming a large amount of memory. When querying via web APIs, a remote server must respond, execute the query, and serialize the response, which can take several seconds. Taxadb, on the other hand, by using a local database as a backend, can perform fast operations on millions of taxonomic names without large memory requirements. In addition, a local copy of the database need only be setup once. This allows offline use and reproducible queries. Taxadb can also use different databases. These include ITIS (Integrated Taxonomic Information System), NCBI (National Center for Biological Information's Taxonomy database), COL (Catalogue of Life), OTT (Open Tree Taxonomy) and many more (Boettiger et al. 2020).

# 2. Methods

The program was written in the programming language R version 4.1.1 on the interface RStudio. The following packages were used:

- taxadb version 0.1.3
- heatmap3 version 1.1.9
- pheatmap version 1.0.12
- seqinr version 4.2-8
- RColorBrewer version 1.1-2
- makeFlow version 1.0.2
- svDialogs version 1.0.3
- qpcR version 1.4-1
- data.table version 1.14.2

The package taxadb is started in the following environment:

```
> Sys.setenv(CONTENTID_REGISTRIES="https://hash-archive.carlboettiger.info")
```

# 3. Results

There are several possibilities for the color representation of the taxonomic names in the heatmap. These are limited by the R packages, as the marking only works in certain ways, depending on the function of the heatmap packages themselves. For these, 2 packages were examined more closely, the package pheatmap as well as heatmap3. In the following it will be described how the program works on the respective packages and which advantages and disadvantages the packages offer.

## 3.1 R package pheatmap

First of all, two small functions are defined. The first function is responsible for converting the gaps of each sequence to a "0" and all other characters to a "1". The second function is responsible for reading the FASTA file, extracting the headers of the sequences and to remove everything before and after the species name. After that the big mymain() function starts. Everything is programmed inside this function. This has the advantage that the individual objects are not stored after the function has been executed, and thus do not unnecessarily burden the memory space. After the environment for the package taxadb was started (see methods) a window is opened, in which the FASTA file can be selected. After applying the above three functions to the file, we get a dataset with all species names. Now a database search can be started by the function filter_name() from the package taxadb. The database NCBI is used because most of the names are found here. The ID of the names as well as all taxonomic

information for each biological class are taken and returned in a data frame. There are also sequences for which no or only incomplete information is found. Next, a window appears where the user can specify how many taxonomic levels should be displayed in the heatmap. Then another window appears where the user can enter which taxonomic levels he wants to choose. There are six levels to choose from: kingdom, phylum, class, order, family, and genus. After setting these parameters, exactly these taxonomic groups are extracted from the large data frame of taxadb. In some cases, there is a duplication of entries. This is because there are synonymous names for a species name. However, since for this package an exact mapping of the names must be done with the annotation column (column that colors the taxonomic data), the duplicate entries are removed. In the next step, all groups from the respective taxonomic levels are sorted by size. The groups are then arranged from top to bottom according to their frequency. After the groups have been sorted and made unique, another window opens and asks the user how many of these groups from a particular taxonomic class should be displayed. If not all classes are selected (20 being the maximum here), the remaining groups will be renamed to "Others". After this process has been completed for all selected taxonomic levels, the user can decide via another window whether the group "Others" should be displayed or not. If the user does not want to show this group, "Others" is replaced by the empty element NA (not available) so that it is not displayed. Now, a special formatting is needed for the heatmap, where the names are assigned to the respective group of all selected taxonomic levels. It is important that duplicate names (e.g. due to different isoforms for the same protein) are made unique for unambiguous assignment (see table 1).

| | class | order |
|---|---|---|
| **homo sapiens** | mammals | Primates Linnaeus, 1758 |
| **homo sapiens.1** | mammals | Primates Linnaeus, 1758 |
| **homo sapiens.2** | mammals | Primates Linnaeus, 1758 |
| **puma concolor** | mammals | carnivores |
| **chloropsis hardwickii** | birds | song birds |
| **centrocercus urophasianus** | birds | landfowls |
| **papilio polytes** | true insects | moths |
| **drosophila mauritiana** | true insects | *NA* |
| **heliornis fulica** | birds | *NA* |
| **alphaproteobacteria bacterium** | *NA* | *NA* |

Table 1) Special formatting for the assignment of names to the groups of the selected taxonomic level

Last, for the heatmap, the "0's" are assigned the color wheat and the "1's" are assigned the color red so that there is a good contrast. Then the heatmap includes the setting that names are only shown if there are less than 21 sequences, otherwise it would be too confusing. On the x-axis are the positions of the sequences. However, it was set that only every 20th position is displayed.

If we as a user would read in a FASTA file with 1000 sequences and then determine that we have 2 taxonomic levels, namely "class" with 4 out of possible 8 groups and "order" with 10 out of 108 possible groups, we would see as output figure 4 and figure 5 respectively (depending on whether we want the "Others" group displayed or not).

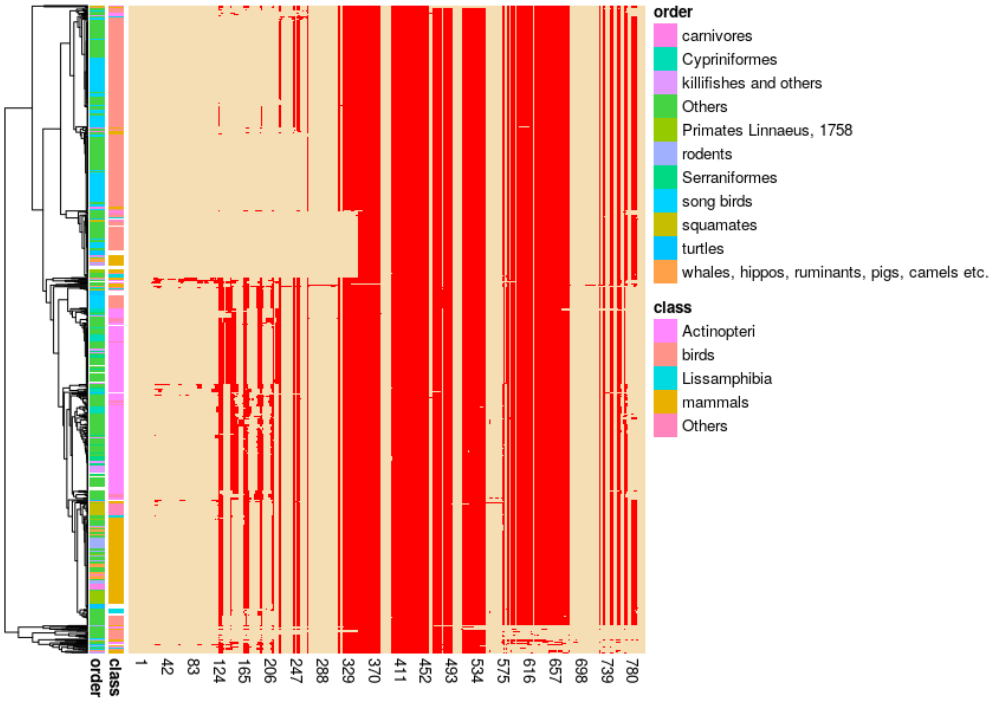



Figure 4) pheatmap with 1000 sequences including 2 annotation columns (with the group "Others")

This is a heatmap for ganglioside-induced differentiation-associated protein 1 with 1000 sequences created with the R package pheatmap. Gaps are shown in wheat and amino acid residues in red. The x-axis shows the positions from the FASTA file. On the left side of the heatmap is a dendrogram and two annotation columns. These columns represent the order and class of the sequences, while the corresponding colors to the groups are shown on the right side. Here, the 10 largest groups (out of 108) of the rank "order" and the 4 largest groups (out of 8) of the rank "class" were colored.

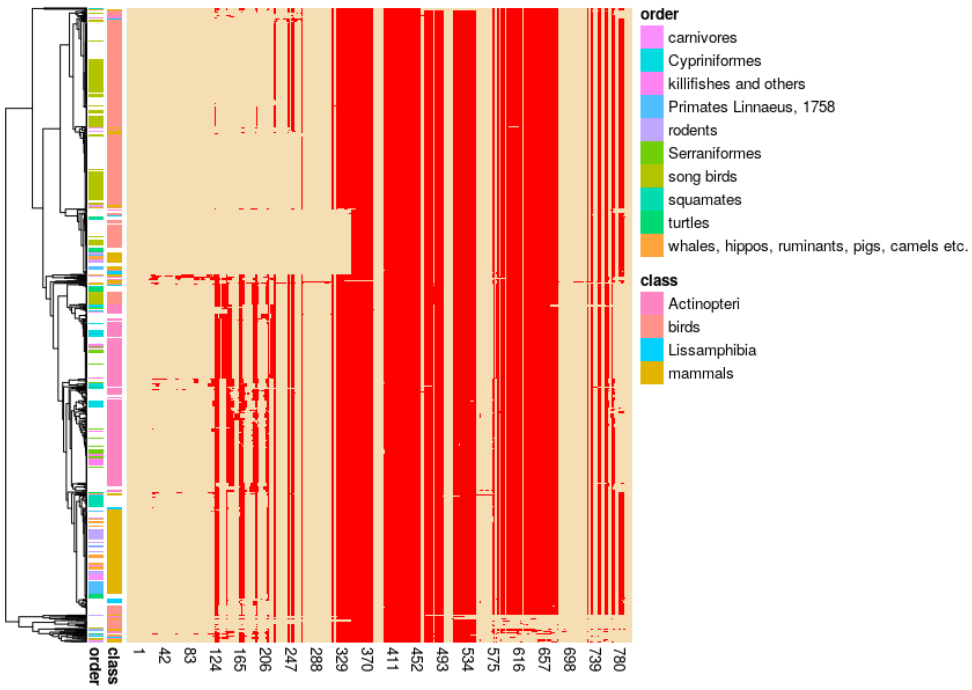



Figure 5) pheatmap with 1000 sequences including 2 annotation columns (without the group "Others")

This is a heatmap for ganglioside-induced differentiation-associated protein 1 with 1000 sequences created with the R package pheatmap. Gaps are shown in wheat and amino acid residues in red. The x-axis shows the positions from the FASTA file. On the left side of the heatmap is a dendrogram and two annotation columns. These columns represent the order and class of the sequences, while the corresponding colors to the groups are shown on the right side. Here, the 10 largest groups (out of 108) of the rank "order" and the 4 largest groups (out of 8) of the rank "class" were colored.

## 3.2 R package heatmap3

The program works the same as in the pheatmap package until the extraction of the required columns from the large data frame of taxadb, except that here the user cannot display multiple taxonomic levels but must choose one. Sorting works a bit differently in this package. Since we only have one taxonomic level, in this case we don't have to use a for loop for all the chosen taxonomic levels but can directly determine the frequency of the groups within the sequences, sort by them, and then isolate the groups and make them unique. Now the user can decide via a window how many groups should be kept in the heatmap. Thereby, 50 groups are the maximum, because otherwise it becomes too messy and up to 11 groups plus the group "Others" colors can be used for color blind people and must not be taken from the function rainbow(). If not all groups are selected for the heatmap, the remaining groups are changed to "Others" as in the pheatmap. Next, the user can also decide if the group "Others" should be displayed. If yes, it will be colored gray and if not, all sequences with the corresponding group "Others" will be overwritten with NA. After this is done, the program decides on some parameters for the heatmap. Unlike the R package pheatmap, the legend can be customized. So first the parameter for the transparency of the colors is set, which depends on the sequences. The more sequences are present, the more transparent the colors must be, so that the heatmap is still well recognizable. Also, the size of the species names on the right side of the heatmap is defined. The more sequences are present, the smaller the names should be to show as many species names as possible. In the next step the groups are sorted again, because the user has decided how many groups he wants to keep and thus changed the number of groups. Depending on this, further parameters are set for the legend. This includes the formation of the colors, the maximum name length of the groups, the size of the font of the groups, the width of the individual colors on the legend, as well as the number of columns for the legend, so that the legend can be displayed legibly above the heatmap and does not touch it. After that, a special sorting of the sequences by the groups is performed (see table 2). In this process, all species names must be listed for each group. Unlike the pheatmap, there does not have to be a unique assignment here. If we have sorted the name *Homo sapiens* under the group mammals, all *Homo sapiens* are automatically assigned to the group mammals by color.

| | mammals | birds | true insects | Purple bacteria, alpha subdivision |
|---|---|---|---|---|
| 1 | homo sapiens | chloropsis hardwickii | papilio polytes | alphaproteobacteria bacterium |
| 2 | puma concolor | centrocercus urophasianus | drosophila mauritiana | *NA* |
| 3 | *NA* | heliornis fulica | *NA* | *NA* |

Table 2) Special formatting for the assignment of the colors to the species names

After this, the maximum name length is set for the sequences that appear on the right side of the heatmap and the “Others” group is assigned the color gray (if present). Next, two closed for loops are started, which call each sequence of the special sort and assign a color for each position. The color changes only when a new group is entered. This process is shown schematically in table 3.

| | as.vector.sapply.bc_list..rep..ncol.presence... | rep.seq_len.ncol.presence....length.bc_list.. | rep.colour..length.bc_list....ncol.presence.. | rep.1..length.bc_list....ncol.presence.. |
|---|---|---|---|---|
| 554 | 1 | 554 | #1B9E774D | 1 |
| 555 | 1 | 555 | #1B9E774D | 1 |
| 556 | 1 | 556 | #1B9E774D | 1 |
| 557 | 1 | 557 | #1B9E774D | 1 |
| 558 | 1 | 558 | #1B9E774D | 1 |
| 559 | 1 | 559 | #1B9E774D | 1 |
| 560 | 1 | 560 | #1B9E774D | 1 |
| 561 | 1 | 561 | #1B9E774D | 1 |
| 562 | 1 | 562 | #1B9E774D | 1 |
| 563 | 1 | 563 | #1B9E774D | 1 |
| 564 | 1 | 564 | #1B9E774D | 1 |
| 565 | 1 | 565 | #1B9E774D | 1 |
| 566 | 2 | 1 | #1B9E774D | 1 |
| 567 | 2 | 2 | #1B9E774D | 1 |
| 568 | 2 | 3 | #1B9E774D | 1 |
| 569 | 2 | 4 | #1B9E774D | 1 |
| 570 | 2 | 5 | #1B9E774D | 1 |
| 571 | 2 | 6 | #1B9E774D | 1 |
| 572 | 2 | 7 | #1B9E774D | 1 |
| 573 | 2 | 8 | #1B9E774D | 1 |
| 574 | 2 | 9 | #1B9E774D | 1 |

...

| | | | | |
|---|---|---|---|---|
| 5630 | 10 | 545 | #E7298A4D | 1 |
| 5631 | 10 | 546 | #E7298A4D | 1 |
| 5632 | 10 | 547 | #E7298A4D | 1 |
| 5633 | 10 | 548 | #E7298A4D | 1 |
| 5634 | 10 | 549 | #E7298A4D | 1 |
| 5635 | 10 | 550 | #E7298A4D | 1 |
| 5636 | 10 | 551 | #E7298A4D | 1 |
| 5637 | 10 | 552 | #E7298A4D | 1 |
| 5638 | 10 | 553 | #E7298A4D | 1 |
| 5639 | 10 | 554 | #E7298A4D | 1 |
| 5640 | 10 | 555 | #E7298A4D | 1 |
| 5641 | 10 | 556 | #E7298A4D | 1 |
| 5642 | 10 | 557 | #E7298A4D | 1 |
| 5643 | 10 | 558 | #E7298A4D | 1 |
| 5644 | 10 | 559 | #E7298A4D | 1 |
| 5645 | 10 | 560 | #E7298A4D | 1 |
| 5646 | 10 | 561 | #E7298A4D | 1 |
| 5647 | 10 | 562 | #E7298A4D | 1 |
| 5648 | 10 | 563 | #E7298A4D | 1 |
| 5649 | 10 | 564 | #E7298A4D | 1 |
| 5650 | 10 | 565 | #E7298A4D | 1 |

Showing 5,629 to 5,650 of 5,650 entries, 4 total columns

**Table 3) Created Data frame for the assignment of the colors to the species names**

For the assignment of the colors in the package heatmap3, each position is annotated with a color for each sequence from figure 7. Here, the color is changed only in the case that the species names of another group are accessed. This creates a huge data frame. In this case there are 10 species names, where the sequences of the species consist of 565 positions. Thus, a Date frame with the length of 5650 rows is created. Here, the first column contains the sequence number, the second column the position and the third column the color for the position.

Finally, the R package heatmap3 is executed with all the previously determined parameters. If we would import a FASTA file with 1000 sequences like in pheatmap and then select from the taxonomic level "class" 5 out of 8 groups, the output would be figure 6 with the group "Others" and figure 7 when the group "Others" is hidden.

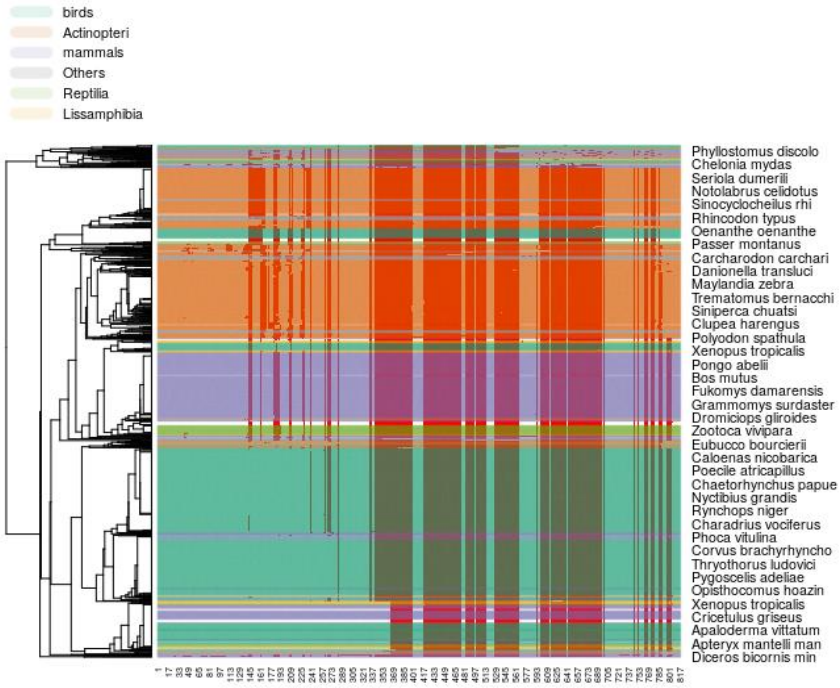



**Figure 6) heatmap3 with 1000 sequences including a coloring (with the group "Others")**

This is a heatmap for ganglioside-induced differentiation-associated protein 1 with 1000 sequences created with the R package heatmap3. Gaps are shown in wheat and amino acid residues in red. The x-axis shows the positions from the FASTA file. On the left side of the heatmap is a dendrogram. Here, the 5 largest groups (out of 8) of the rank "class" were colored, while the corresponding colors to the groups are shown on top of the heatmap.

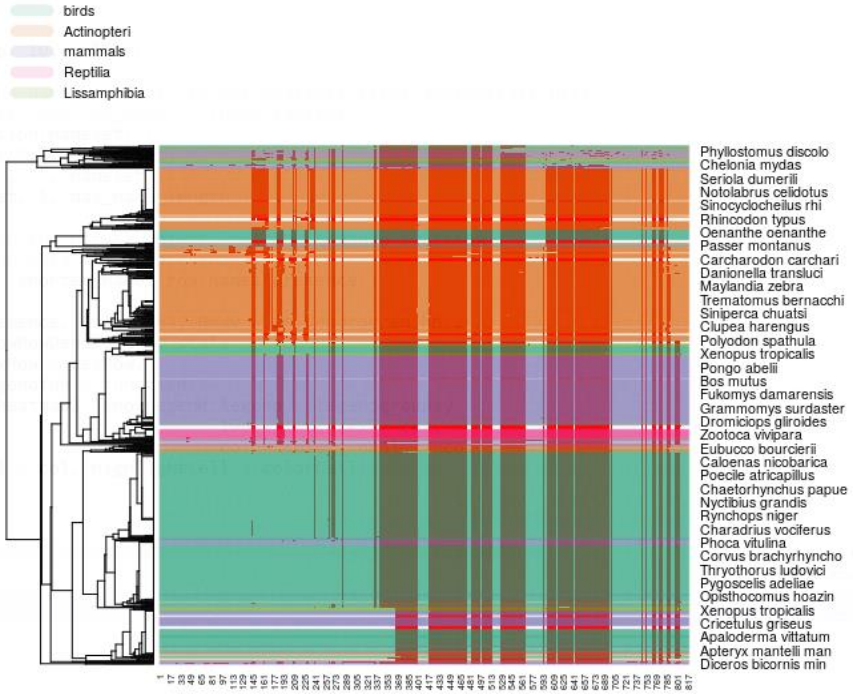



**Figure 7) heatmap3 with 1000 sequences including a coloring (without the group "Others")**

This is a heatmap for ganglioside-induced differentiation-associated protein 1 with 1000 sequences created with the R package heatmap3. Gaps are shown in wheat and amino acid residues in red. The x-axis shows the positions from the FASTA file. On the left side of the heatmap is a dendrogram. Here, the 5 largest groups (out of 8) of the rank "class" were colored, while the corresponding colors to the groups are shown on top of the heatmap.

## 3.3 Computational time

Several factors were considered to determine the computation time (see table 4). It is clear that the computing time depends on the number of sequences in the FASTA file. For this reason, files with 10, 100, 1000, 2000 and 5000 sequences were used for comparison. Next, it is important to know how long it takes to create a heatmap without making color markings in order to compare these times with those from my program. Then there is the question how long the searching in the local database through taxadb takes. Finally, it has to be determined if the calculation time depends on how many groups the user wants to display and if the group "Others" plays a role.

| | | Sequences | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 100 | 1000 | 2000 | 5000 |
| Creation of a heatmap without using my program | pheatmap | < 1 s | < 1 s | 14 s | 38 s | 128 s |
| | heatmap3 | < 1 s | < 1 s | 13 s | 59 s | 104 s |
| | taxadb search | 31 s | 30 s | 30 s | 30 s | 30 s |
| Creation of a heatmap after taxadb search without "Others" group | pheatmap with 5 groups on 1 taxonomic level | < 1 s | 1 s | 14 s | 38 s | 130 s |
| | pheatmap with 10 groups on 3 taxonomic levels | / | 1 s | 15 s | 38 s | 131 s |
| Creation of a heatmap after taxadb search with "Others" group | pheatmap with 5 groups on 1 taxonomic level | < 1 s | 1 s | 15 s | 38 s | 131 s |
| | pheatmap with 10 groups on 3 taxonomic levels | / | 1 s | 15 s | 38 s | 130 s |
| Creation of a heatmap after taxadb search without "Others" group | heatmap3 with 5 groups | < 1 s | 1 s | 15 s | 30 s | 88 s |
| | heatmap3 with 30 groups | / | 3 s | 79 s | 175 s | 424 s |
| Creation of a heatmap after taxadb search with "Others" group | heatmap3 with 5 groups | < 1 s | 3 s | 78 s | 171 s | 931 s |
| | heatmap3 with 30 groups | / | 3 s | 122 s | 286 s | 1454 s |

**Table 4) Computational time**

In this table the time was measured, which the program needs to execute certain operations. FASTA files with 10, 100, 1000, 2000 and 5000 sequences were used for this purpose. In the package pheatmap the taxonomic ranks "class" and "order" were selected. In the package heatmap3 the groups were always used in the taxonomic rank "order".

First, it is noticeable that the local database search by taxadb is apparently independent of the number of sequences. In the R package pheatmap, the creation of the annotation columns (independent of the number of groups displayed) does not take longer than the creation of the heatmap itself. It is noticeable that the color marking in the package heatmap3 depends on the number of groups as well as on the group "Others". The more groups are to be marked, the longer the process takes. It can be seen that as soon as the group "Others" is hidden, the computing time is almost halved. This is related to the data frame that is created for the colors. Especially if there are many sequences, there are always more entries in "Others", so that these do not have to be considered and accelerate the computing time.

## 3.4 Comparison between pheatmap and heatmap3

The two R packages for heatmap carry some advantages and disadvantages. On the one hand, it can be advantageous that the pheatmap package can display several taxonomic ranks at the same time. For this purpose, there is the possibility to display several separate columns on the left side of the heatmap. However, on the other hand the columns make it difficult to track the individual sequences through the heatmap, especially if there are many sequences. So, you cannot tell exactly which sequence it is. Heatmap3 offers a better alternative. The marker runs through the whole heatmap, so that the sequences can be assigned exactly. In this package, however, only one taxonomic class can be represented at a time. The package pheatmap offers a further advantage with the colored representation. In heatmap3, the colors make it difficult to recognize the actual heatmap, even if the colors are transparent. Thus, when analyzing the image, the background, i.e. the heatmap itself, must be considered first and only then should the colored markings be taken into account. Moreover, it is a challenge to choose the transparency in such a way that the heatmap is still recognizable, but also the colors are clearly distinguishable from each other. In contrast, with the R package pheatmap, the colors do not interfere with the heatmap. However, with this display option it is problematic to keep the color overview, especially if several annotation columns are used. Another color problem is that currently the colors in pheatmap cannot be selected independently yet, because so far, no connection with the heatmap could be established (which is potentially possible). Thus, the colors could not be selected as in heatmap3, meaning that if the number of groups is less than 12, they are still visible to colorblind people. Additionally, the group "Others" is marked in gray, so that a better distinction is possible. However, with a larger number of groups (> 12), the colors are taken from the rainbow() function, which makes it more difficult to distinguish between the individual groups. Furthermore, the color highlighting is only possible if the names are assigned to the annotation column. As a result, only all names can be displayed at the same time on the right side of the heatmap or they are omitted, to keep the highlighting. If only some names are selected, only these will be marked. In comparison, in heatmap3 the color marking is independent of the species names on the right side of the heatmap. Thus, specific names can also be displayed, while the color marking exists for the whole heatmap. In addition, the legend at pheatmap is sorted alphabetically for each column. Thus, it is not possible to see directly how the size ratios of the groups are distributed. In comparison, with the package heatmap3, the legend is sorted by the size of each group, making it possible to directly see the ratios of the groups in the heatmap. Likewise, the package heatmap3 is in the comparison to pheatmap with the organization options advantageous. Here the possibility exists to adapt the legend and to format it thus depending upon number of the groups suitably over the Heatmap. However, for the sake of clarity, this is limited to a number of 50 groups. In the package pheatmap no adjustment of the legend is possible. The groups are only displayed to the right of the heatmap. If there is a large number of groups, this display method will cause many groups to disappear below the image. Furthermore, with this R package, it is possible to make inputs that are unfavorable. For example, the taxonomic groups "class" can be entered twice in the input window, which would then result in the same display twice in the legend of the heatmap. Nevertheless, pheatmap is much faster than heatmap3, because here it is not necessary to assign a color in a data frame to each position of a sequence for each unique species name. Overall, both packages offer individual advantages and disadvantages. Especially advantageous is that the groups "Others" can be hidden, so that on the one hand the computing time in heatmap3 is reduced and on the other hand it does not disturb in the heatmap. Both packages have the disadvantage that the legend is limited. Not too many groups of a taxonomic rank can be displayed, otherwise the legend simply disappears to the right (heatmap3) or downwards (pheatmap). Furthermore, the order of the sequences varies in both packages because different clustering algorithms are used.

# 4. Discussion

## 4.1 Exceptions in the heatmap

When looking at a typical heatmap as shown in figure 7, it quickly becomes apparent that the taxonomic markers of the sequences do not only occur in groups despite the clustering. There are also some exceptions, for example, sequences that occur sporadically outside their group. Next to the organism *Phoca vitulina* in figure 7 there are some sequences of the taxonomic group mammals. Such outliers could have the reason that many proteins have isoforms and thus it is possible that this isoform has similarities with an isoform of the protein from the birds group.

When the organism *Passer montanus* is viewed, it can be seen that there has been no labeling in isolated places. These gaps mean that no or incomplete database entries were found for these unlabeled sequences. This can have several reasons. On the one hand, it may be due to the taxadb function. In some cases, matches are found with multiple accepted names (with different database IDs). In this case, the user would have to manually set which ID the program should use to get the taxonomic data. Then it could be that the sequences cannot be solved due to typos or wrong formatting in the database. Another reason is that different providers use synonymous names. While this does not apply to this program because the search is done in the NCBI database, taxadb can access different databases. If the user does a BLAST search and then decides to search in the fb (Fishbase), this error could occur. Another reason for the gaps could be that the organisms have not yet been taxonomically categorized. Thus, there is an ID for the organism, but it is not yet known to which class it belongs, for example.

Another interesting feature can be seen in figure 7 at the position 190 to 210. Here, at the level of the organism *Passer montanus*, there is a distortion of the sequences. This distortion most likely originates from sequencing errors.

## 4.2 Problems

At the beginning, there were many problems in creating the program. First of all, the question of how to extract taxonomic data from a sequence name had to be solved. The first idea was to communicate directly with the databases using Web APIs (Application Programming Interfaces). However, the R package taxadb was chosen because of its use of a local database. After installation and isolation of the desired taxonomic groups, the question arose how to connect the obtained data to the heatmap and what formatting the data must have for this purpose. As a result of much research, trial and errors, a link to the heatmap packages was established by formatting the data appropriately (by reordering it in the data frame (see table 1 and table 2)). Later it was determined that the heatmap with all groups of a class becomes very confusing, so the question arose how the markers and legends can be optimized. For this it was decided to leave the selection of the displayed groups to the user, so that all remaining groups can be renamed to "Others". Thus, the full spectrum of a taxonomic class is not displayed for all sequences, but always only the largest groups.

Despite the individual fixes, it is possible that the entire system is fragile because of the taxadb package. After taxadb was installed, it did not work as expected, but produced error messages for many commands. This is because the current version of taxadb's resolver does not recognize some locations. The resolver is responsible for initiating and sequencing the query to translate the searched resources.

For this reason, the package had to be started under a specific environment variable (see methods), which registers the data in the hash-archive, so that the old resolver is used. Hashing refers to the conversion of a string into a numeric key. This is then used to retrieve elements in the database. By the environment variable taxadb has hardwired addresses and if anything breaks, the entire system will break down.

Furthermore, the program has the weakness that the displayed dendrogram is useful because it reflects the similarity relationship of each sequence, but each time other similar or complemented data is clustered, the entries are sorted differently.

Last, there is the question of whether the program provides a long-term solution for identifying taxonomic groups. There is no concrete answer for this, but if the developers of taxadb are no longer active at some point, the package will no longer be updated, so that possible errors will not be fixed. At the moment, however, this is not likely to happen since the package has only been released since 2020.

## 4.3 Potential improvements

Although the program shows the taxonomic groups of the sequences in color in the heatmap, there is still room for improvement in many places. First, it is noticeable that the heatmap3 package takes a long time to generate the heatmap. This is due to the data frame that is created for the colors. If there are 5000 unique species names, each with a sequence of 500 positions, a data frame with a length of 2.500.000 entries has to be created. This size can only be reduced by the user hiding the "Others" group, as this simply omits many sequences and does not take them into account when creating the Data frame. Nevertheless, it should be tried to fix exactly this problem by trying an alternative color assignment. Furthermore, in both packages the legends can be displayed in a limited way, so that the user can only display a certain number of groups without losing the overview (in heatmap3 maximum 50 groups and in pheatmap maximum 20 groups per taxonomic rank). Much worse, in the pheatmap package the legend of the annotation columns cannot be customized at all. For this reason, the maximum per taxonomic rank was set to 20. However, if the user wants to display 4 taxonomic ranks with 20 groups each, this will cause the legend to disappear under the image, since the legends for the individual annotation columns are displayed one below the other. In addition, individual species names cannot be shown in this package, because otherwise the colored marking is shown only for exactly these names and disappears for the rest of the sequences. For this reason, either all names must be faded in, which makes them unrecognizable even if there are 100 species names, or no names must be shown, so that at least the highlighting remains for the entire heatmap.

## 4.4 Future

As described in the upper part, the program still has weaknesses in many places and although some of these disadvantages cannot be fixed, there are still many possibilities to improve the program. Since in the package heatmap3 the colors can be displayed independently of the species names, it should be possible to show only certain names to get a better overview. Additionally, more parameter settings could be made possible for the user. This includes for example the independent setting of the transparency of the colors or the choice of a favored color palette for the marking in the heatmap. It is also important to include a feature that saves the heatmap immediately after execution. Furthermore,

there is currently a script for the pheatmap package and a script for the heatmap3 package. It would be beneficial to combine them to prevent code duplication.

In fact, there are related problems where marking the groups of a taxonomic level would be beneficial, phylogenetic trees. Here, individual branches could be color-marked, making it easier to visualize evolutionary development. Perhaps in the future it would even be possible to develop a more general method to display taxonomic information. This can be imagined as a box, which acts as a sequence identifier and thus, when a sequence or species name is entered, it displays the number of the matches as well as a list of these matches. Thus, by only one input all taxonomic information for the sequence as well as relationships would be reproduced. This box could then have an interface to R or a phylogeny manipulation program.

# References


Alvarez, Miguel, and Federico Luebert. 2018. "The Taxlist Package: Managing Plant Taxonomic Lists in R." Biodiversity Data Journal, no. 6 (May). https://doi.org/10.3897/BDJ.6.e23635.

Baron, E. J. (1996). Classification. In th & S. Baron (Eds.), *Medical Microbiology*. Galveston (TX).

Boettiger, C., Chamberlain, S., Norman, K. (2020). taxadb: A high-performance local taxonomicdatabase interface. *British Ecological Society*, doi: 10.1111/2041-210X.13440

Boyle, Brad, Nicole Hopkins, Zhenyuan Lu, Juan Antonio Raygoza Garay, Dmitry Mozzherin, Tony Rees, Naim Matasci, et al. 2013. "The Taxonomic Name Resolution Service: An Online Tool for Automated 221 Standardization of Plant Names." BMC Bioinformatics 14 (1): 16. https://doi.org/10.1186/1471-2105-14-16.

DataFlair 2021. Pros and Cons R Programming Language – Unveil the Essential Aspects! https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/

Maldonado, Carla, Carlos I. Molina, Alexander Zizka, Claes Persson, Charlotte M. Taylor, Joaquina Albán, Eder Chilquillo, Nina Rønsted, and Alexandre Antonelli. 2015. "Estimating Species Diversity and 249 Distribution in the Era of Big Data: To What Extent Can We Trust Public Databases?" Global Ecology and 250 Biogeography 24 (8): 973–84. https://doi.org/10.1111/geb.12326.

Remsen, David. 2016. "The Use and Limits of Scientific Names in Biological Informatics." ZooKeys 550 (July): 207–23. https://doi.org/10.3897/zookeys.550.9546

Roskov Y., Orrell T., Abucay L. 2018. "Species 2000 & ITIS Catalogue of Life, 2018 Annual Checklist." Leiden, the Netherlands.: Species 2000: Naturalis. www.catalogueoflife.org/annual-checklist/2018

Zhao, S., Guo, Y., Sheng, Q., & Shyr, Y. (2014). Advanced heat map and clustering analysis using heatmap3. *Biomed Res Int, 2014*, 986048. doi:10.1155/2014/986048

Zhao S, Guo Y, Sheng Q, Shyr Y. (2014) Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics*.;15(Suppl 10):P16. Published 2014 Sep 29. doi:10.1186/1471-2105-15-S10-P16