Format Cetaceans Data

Contents

Creation - jeremy.andreoletti@ens.fr - 13/04/2020

Load occurrence dataset

```
##  occurrence_no     record_type    reid_no       flags         collection_no
##  Min.   :  68135   occ:4476    Min.   :11942   Mode:logical   Min.   :  4868
##  1st Qu.: 540497               1st Qu.:18135   NA's:4476      1st Qu.: 48884
##  Median : 725033               Median :21552                  Median : 67030
##  Mean   : 804995               Mean   :22595                  Mean   : 82570
##  3rd Qu.:1021108               3rd Qu.:27547                  3rd Qu.: 99584
##  Max.   :1396622               Max.   :34473                  Max.   :192401
##                                NA's   :4191
##                identified_name       identified_rank identified_no
##  Cetacea indet.        : 271   species       :2438   Min.   : 36652
##  Mysticeti indet.      : 140   genus         : 675   1st Qu.: 42951
##  Odontoceti indet.     : 124   family        : 648   Median : 63219
##  Delphinidae indet.    :  96   suborder      : 297   Mean   : 67516
##  Balaenopteridae indet.:  92   unranked clade: 287   3rd Qu.: 65078
##  Balaena mysticetus    :  79   superfamily   :  71   Max.   :367667
##  (Other)               :3674   (Other)       :  60
##                  difference              accepted_name        accepted_rank
##                       :3557   Cetacea           : 294   species       :2057
##  recombined as        : 302   Odontoceti        : 240   family        : 761
##  nomen dubium         : 301   Mysticeti         : 208   genus         : 741
##  subjective synonym of: 118   Balaenopteridae   : 156   suborder      : 471
##  invalid subgroup of  :  56   Delphinidae       : 107   unranked clade: 310
##  corrected to         :  30   Balaena mysticetus:  84   superfamily   :  74
##  (Other)              : 112   (Other)           :3387   (Other)       :  62
##   accepted_no          early_interval           late_interval
##  Min.   : 36652   Holocene    : 626                     :3967
##  1st Qu.: 42937   Langhian    : 355   Late Pliocene    :  80
##  Median : 62924   Burdigalian : 310   Langhian         :  60
##  Mean   : 64432   Zanclean    : 309   Messinian        :  59
##  3rd Qu.: 64626   Tortonian   : 267   Serravallian     :  58
##  Max.   :367667   Serravallian: 239   Early Pleistocene:  34
##                   (Other)     :2370   (Other)          : 218
##      max_ma           min_ma       reference_no
##  Min.   : 0.0117   Min.   : 0.000   Min.   :  289
##  1st Qu.: 3.6000   1st Qu.: 2.588   1st Qu.:12813
##  Median :11.6200   Median : 5.333   Median :23666
##  Mean   :13.9433   Mean   :10.117   Mean   :27322
##  3rd Qu.:20.4400   3rd Qu.:13.820   3rd Qu.:38452
##  Max.   :66.0000   Max.   :47.800   Max.   :65107
##

##    occurrence_no record_type reid_no flags collection_no
## 1          68135         occ      NA    NA          4868
## 2         137494         occ      NA    NA         11601
## 3         141404         occ      NA    NA         12121
## 4         147937         occ      NA    NA         13063
## 5         147938         occ      NA    NA         13064
## 6         148079         occ      NA    NA         13078
## 7         148335         occ      NA    NA         13090
## 8         148353         occ      NA    NA         13092
## 9         148356         occ      NA    NA         13096
## 10        148358         occ      NA    NA         13098
## 11        148360         occ      NA    NA         13100
## 12        148363         occ      NA    NA         13102
## 13        148364         occ      NA    NA         13102
## 14        148365         occ   19615    NA         13103
## 15        150826         occ      NA    NA         11596
## 16        150827         occ      NA    NA         13103
## 17        150828         occ      NA    NA         13402
## 18        150829         occ      NA    NA         13402
## 19        150830         occ      NA    NA         13403
## 20        150831         occ      NA    NA         13403
##                                  identified_name identified_rank identified_no
## 1        n. gen. Georgiacetus n. sp. vogtlensis         species         63123
## 2                       Argyrocetus joaquinensis         species         69897
## 3        n. gen. Kharthlidelphis n. sp. diceros         species         53161
## 4            n. gen. Pinocetus n. sp. polonicus         species         53140
## 5           n. gen. Basiloterus n. sp. hussaini         species         53165
## 6       n. gen. Sachalinocetus n. sp. cholmicus         species         63225
## 7          n. gen. Praekogia n. sp. cedrosensis         species         53139
## 8              Aulophyseter n. sp. rionegrensis         species         53106
## 9                    Microcetus n. sp. sharkovi         species         53137
## 10             n. gen. Mixocetus n. sp. elysius         species         64432
## 11 n. gen. Austrosqualodon n. sp. trirhizodonta         species         63212
## 12                            Basilosaurus isis         species         53287
## 13                                Dorudon atrox         species         53288
## 14                                Dorudon atrox         species         53288
## 15                Basilosaurus n. sp. drazindai         species         53163
## 16                            Basilosaurus isis         species         53287
## 17                            Basilosaurus isis         species         53287
## 18                                Dorudon atrox         species         53288
## 19                            Basilosaurus isis         species         53287
## 20                                Dorudon atrox         species         53288
##             difference                 accepted_name accepted_rank accepted_no
## 1                             Georgiacetus vogtlensis       species       63123
## 2                            Argyrocetus joaquinensis       species       69897
## 3         nomen dubium               Kharthlidelphis         genus       53160
## 4                                 Pinocetus polonicus       species       53140
## 5                                Basiloterus hussaini       species       53165
## 6                            Sachalinocetus cholmicus       species       63225
## 7                               Praekogia cedrosensis       species       53139
## 8  invalid subgroup of                 Physeteroidea   superfamily       53105
## 9                                 Microcetus sharkovi       species       53137
## 10                                  Mixocetus elysius       species       64432
## 11                      Austrosqualodon trirhizodonta       species       63212
## 12                                  Basilosaurus isis       species       62984
## 13                                      Dorudon atrox       species       62994
## 14                                      Dorudon atrox       species       62994
## 15                             Basilosaurus drazindai       species       53163
## 16                                  Basilosaurus isis       species       62984
## 17                                  Basilosaurus isis       species       62984
## 18                                      Dorudon atrox       species       62994
## 19                                  Basilosaurus isis       species       62984
## 20                                      Dorudon atrox       species       62994
##    early_interval  late_interval max_ma min_ma reference_no
## 1        Lutetian                47.800 41.300          289
## 2      Aquitanian                23.030 20.440         4175
## 3        Chattian                28.100 23.030         6018
## 4        Langhian                15.970 13.820         4344
## 5       Bartonian                41.300 38.000         6010
## 6   Early Miocene Middle Miocene 23.030 11.608         4357
## 7       Messinian                 7.246  5.333         4361
## 8       Messinian                 7.246  5.333         4362
## 9        Chattian                28.100 23.030         4365
## 10      Tortonian                11.620  7.246        10152
## 11    Duntroonian                27.300 25.200         4368
## 12     Priabonian                38.000 33.900         6026
## 13     Priabonian                38.000 33.900         6026
## 14     Priabonian                38.000 33.900        10457
## 15      Bartonian                41.300 38.000         6010
## 16     Priabonian                38.000 33.900         6026
## 17     Priabonian                38.000 33.900         6026
## 18     Priabonian                38.000 33.900         6026
## 19     Priabonian                38.000 33.900         6026
## 20     Priabonian                38.000 33.900         6026
```

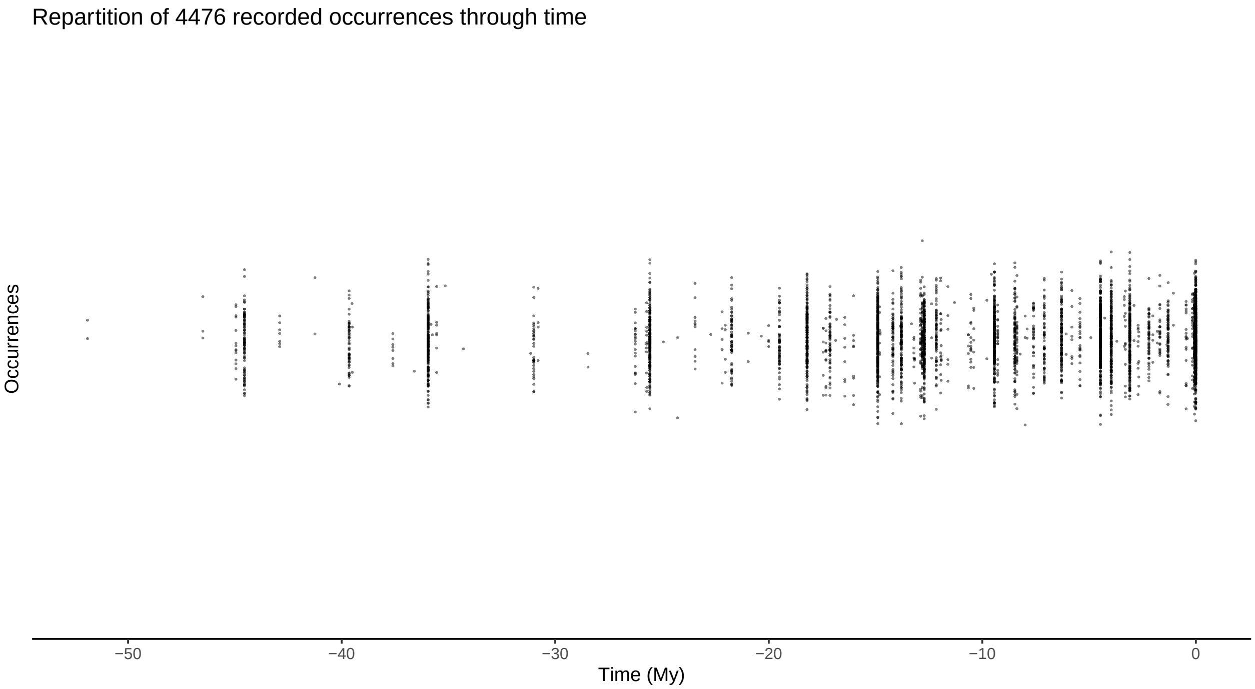
Reorder accepted ranks according to classification standard.

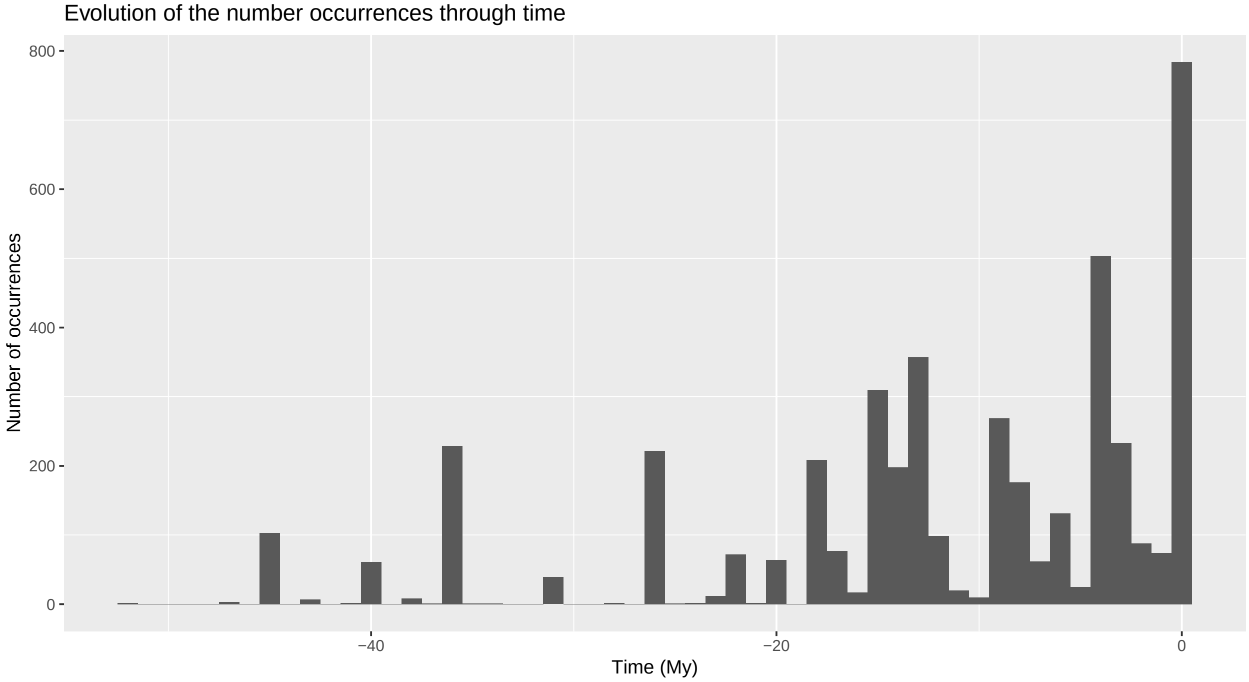
```
## [1] "family"         "genus"          "infraorder"     "species"
## [5] "subfamily"      "suborder"       "subspecies"     "superfamily"
## [9] "unranked clade"
```

Explore the dataset

Repartition through time

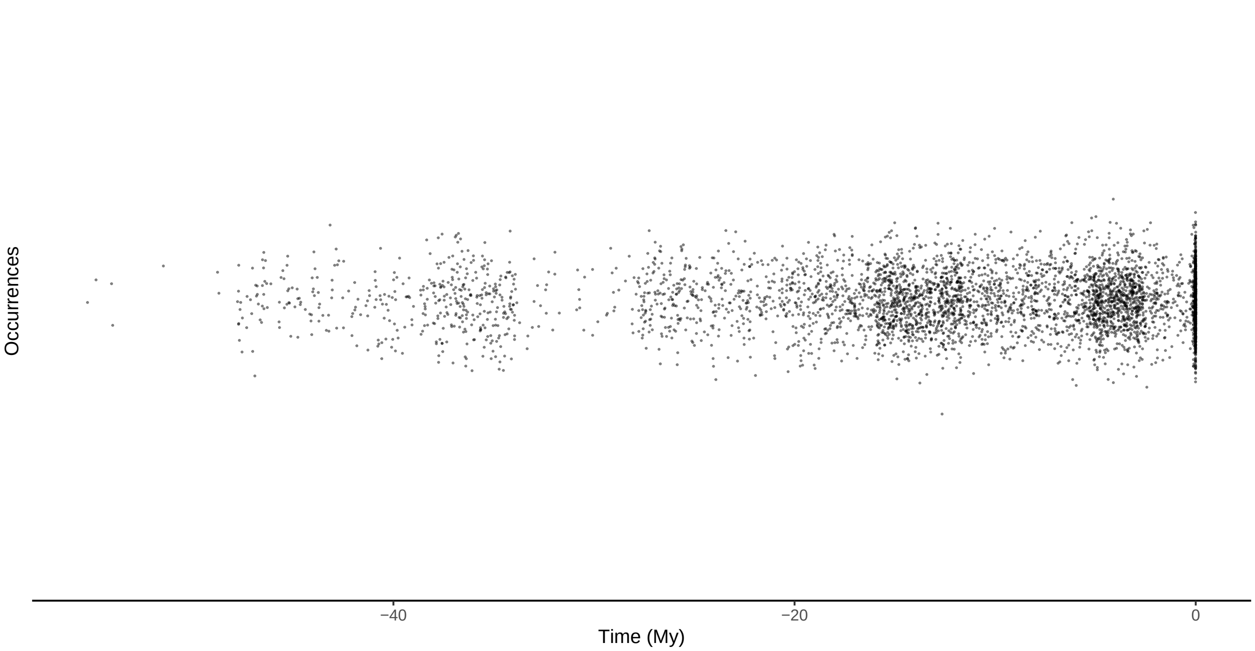
Full fossil record


Repartition of 4476 recorded occurrences through time

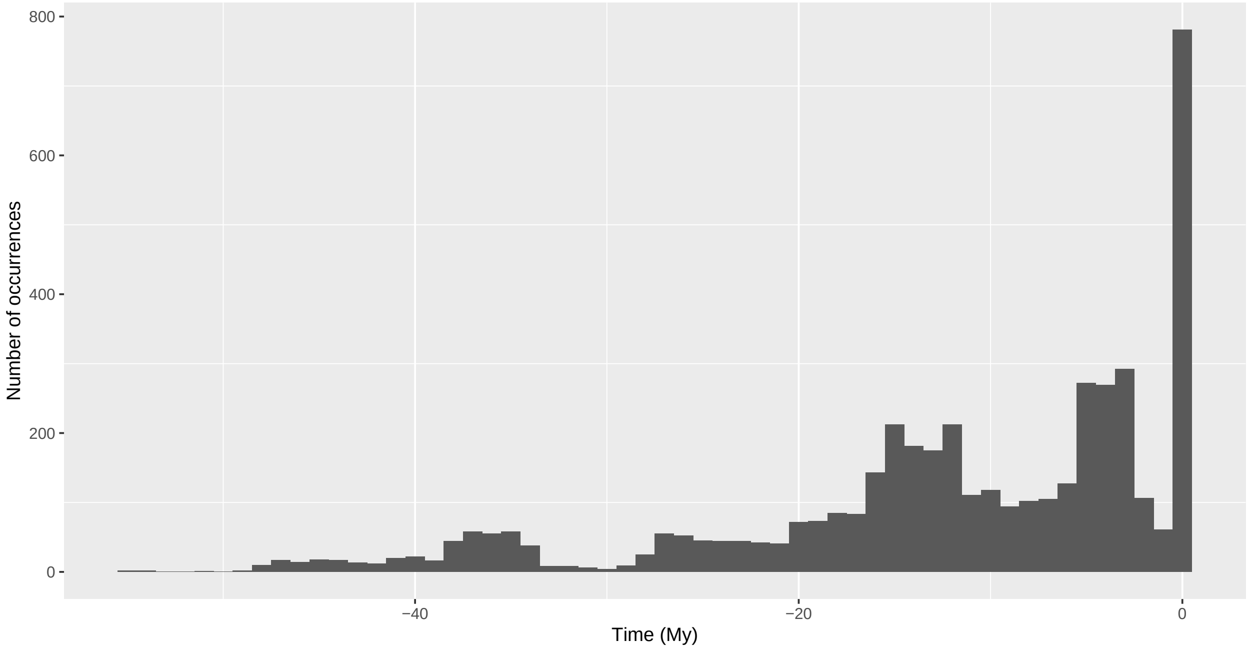

Evolution of the number occurrences through time

→ Numerous occurrences seem to have the same age interval so in order to avoid clusters let's draw them uniformly in their stratigraphic range rather than taking the mean.

Repartition of 4476 recorded occurrences through time



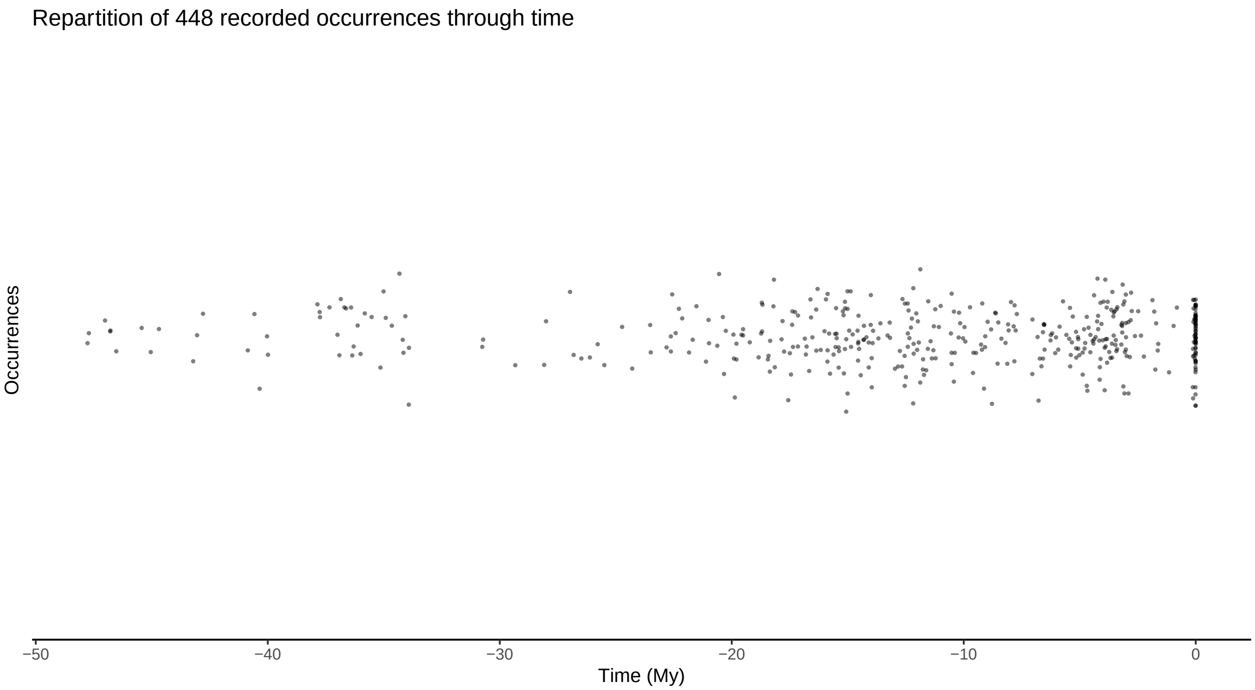
Evolution of the number occurrences through time

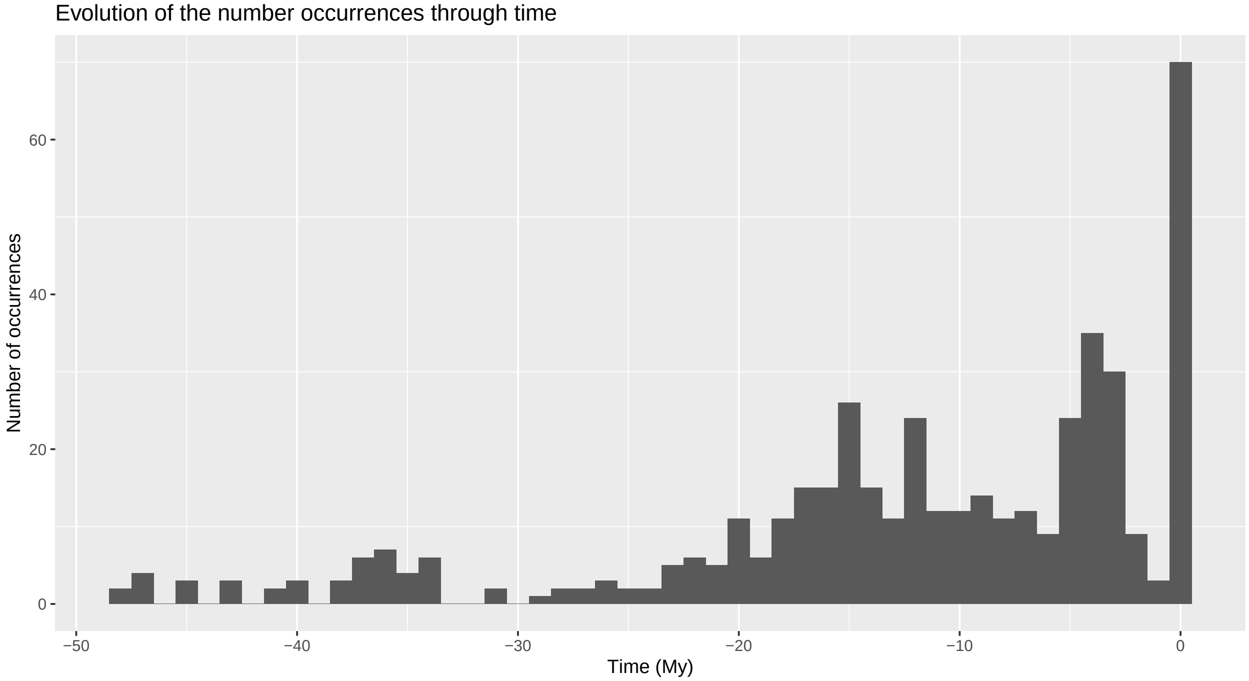


→ The repartition seems much smoother now.

Subsampling

These occurrences are too numerous for our current implementation, let’s subsample a fraction of them for now.

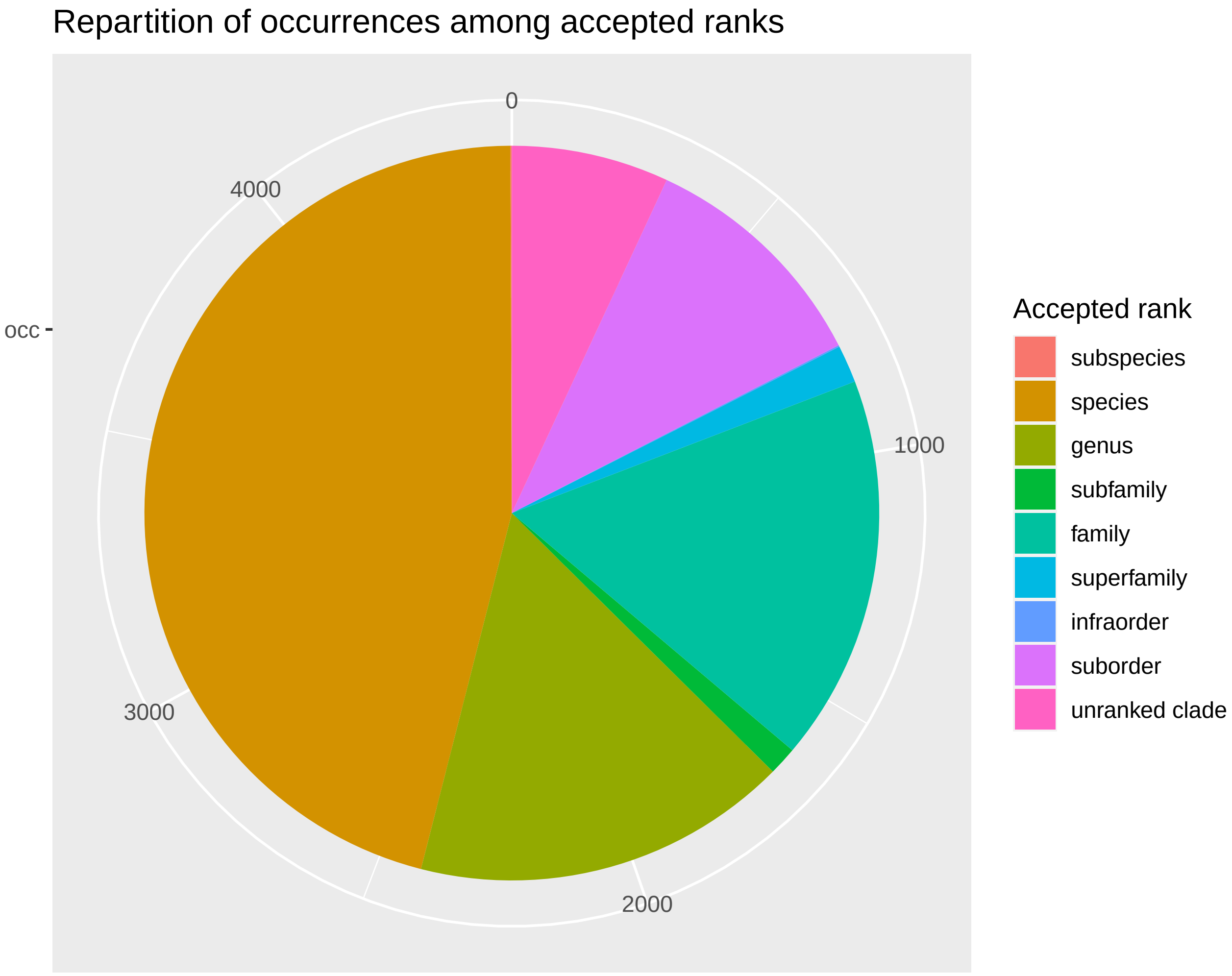




→ The distribution looks similar, with some noise due to higher variance with smaller sample.

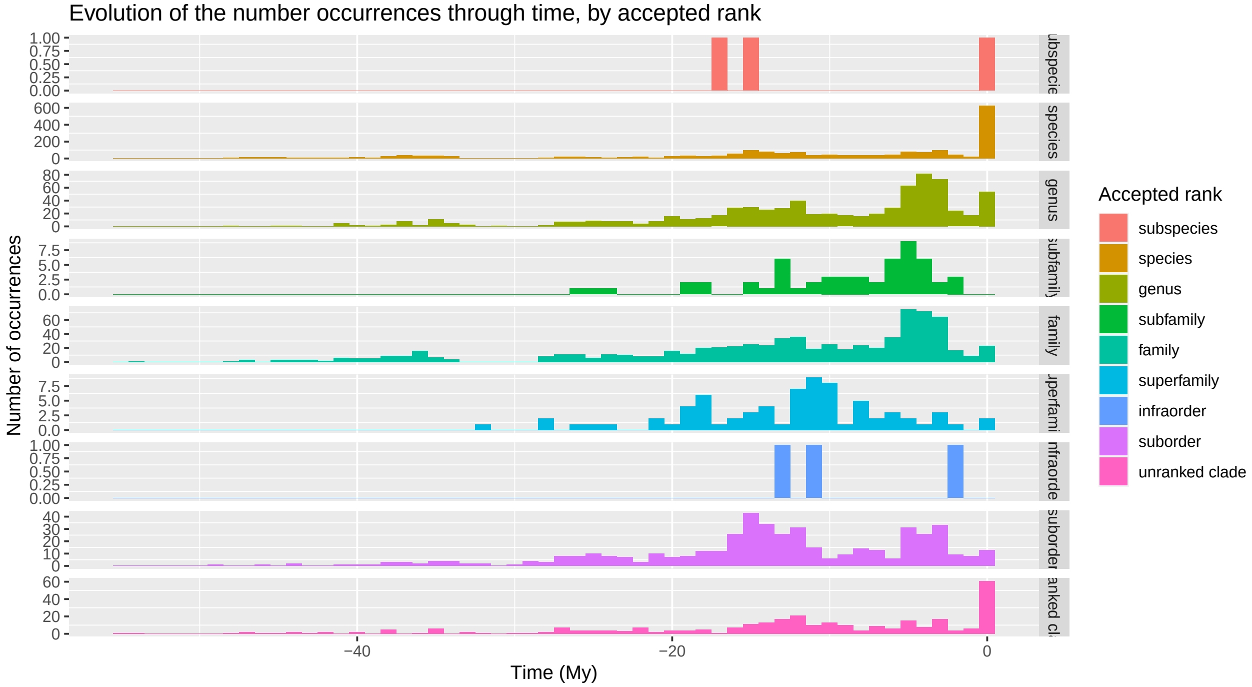
Repartition among accepted ranks

Pie chart



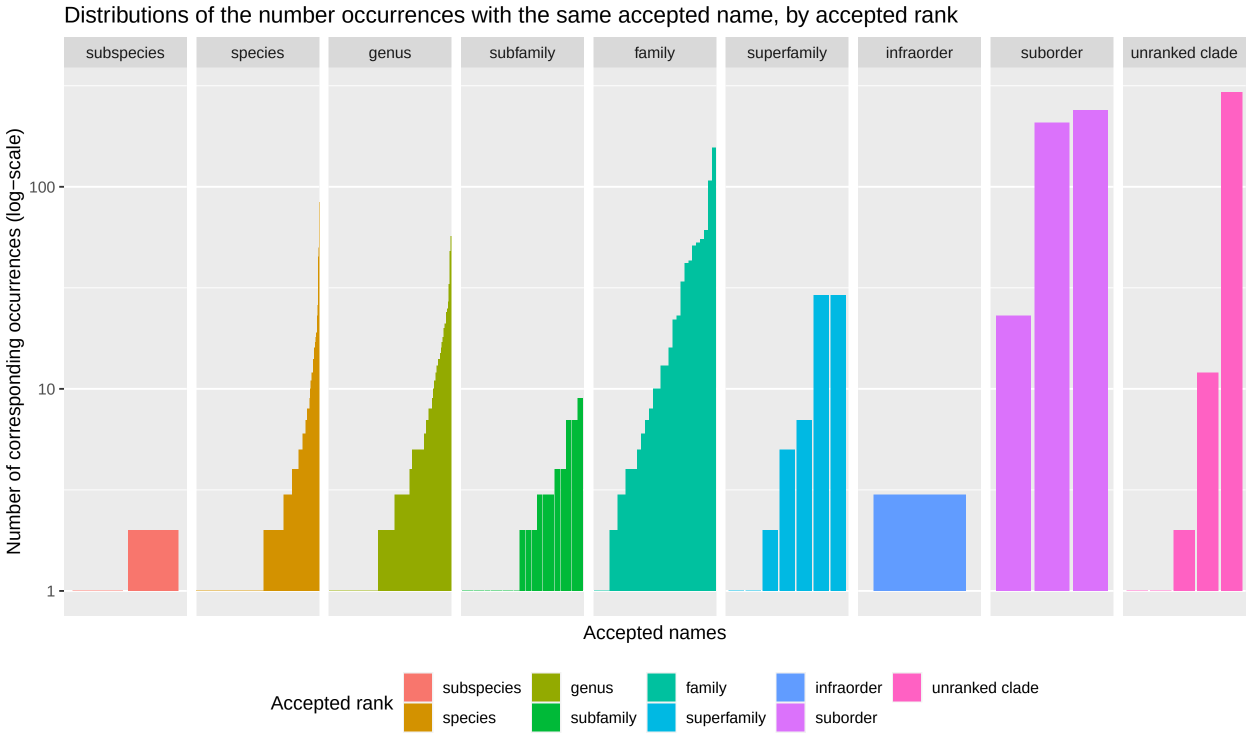
→ Half of the occurrences are identified at the level of the secies and 1/3 at the genus or family. Very few occurences for subspecies/infraorder, maybe fuse with species and suborder or remove ?

Time repartition by rank



→ Similar trends with peaks at ~15My and ~5My and a lot of them around 0 (artefact ?).

Redundancy of occurrences with the same accepted name

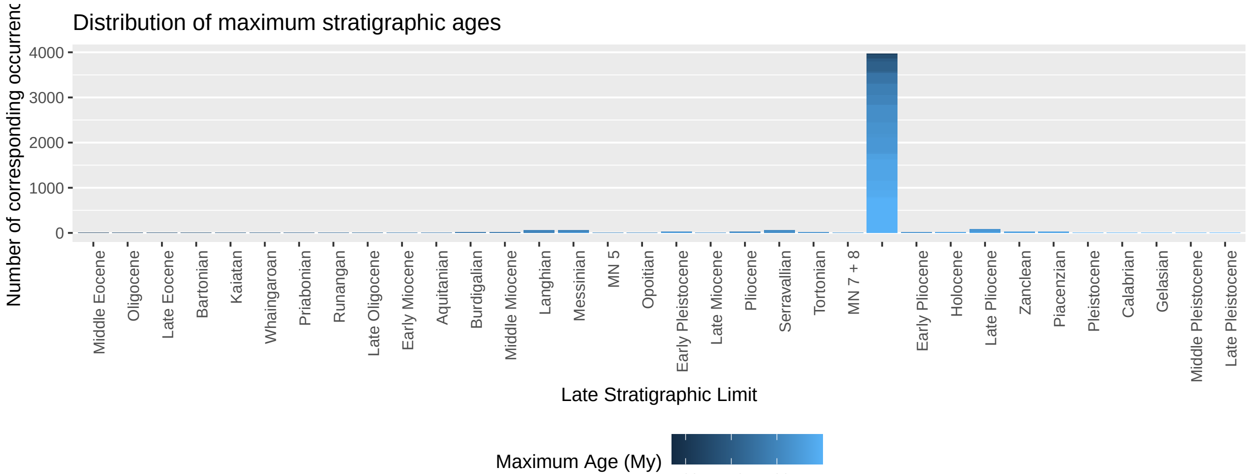


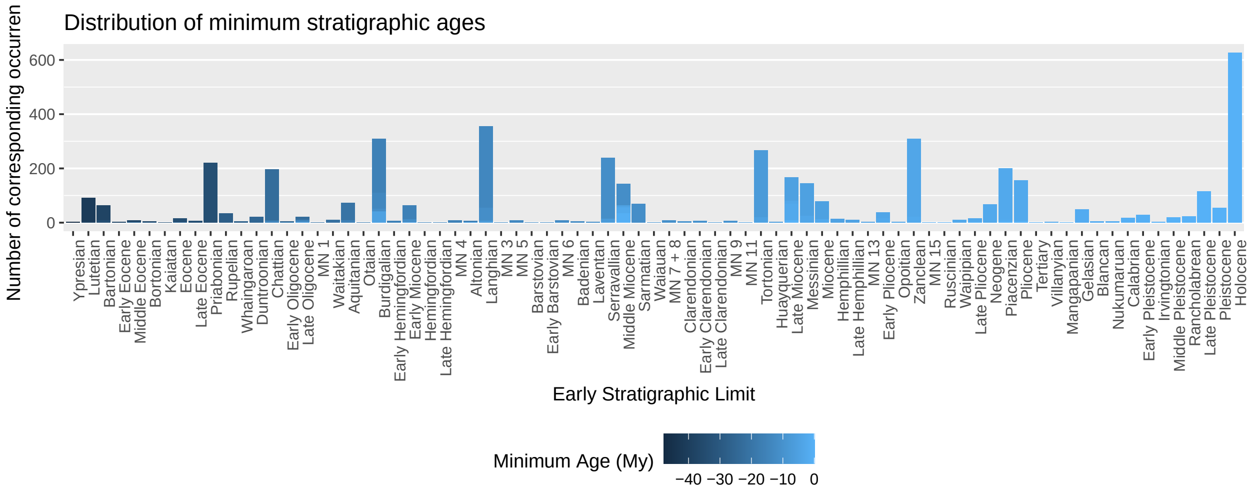
→ ~Half of species/genera/subfamilies have only one specimen by accepted name, but it could go up to ~50 within the same species and ~200 occurrences within the same suborder. **Those differences will have to be corrected because in our model all species are upposed to have the same abundance (identical sampling rates among branches).**

$\Longrightarrow$ Our goal now will be to correct this abundance bias.

Time intervals = stratigraphic age uncertainty

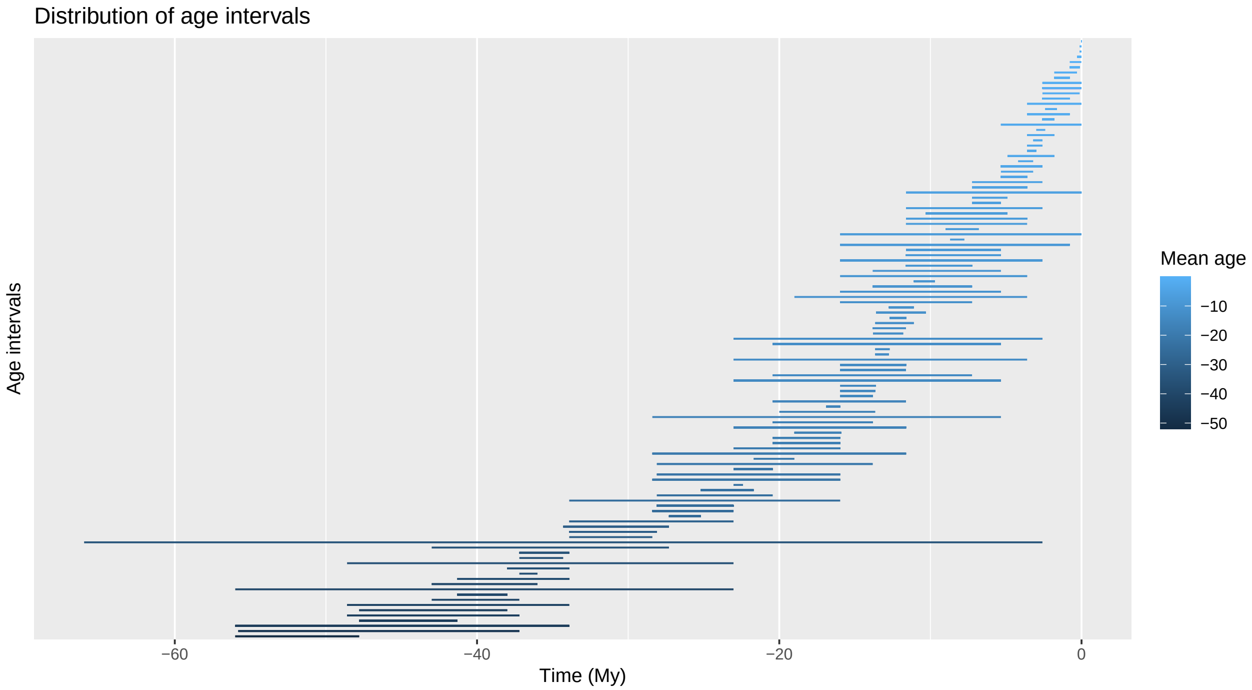
Minimum and maximum stratigraphic limits





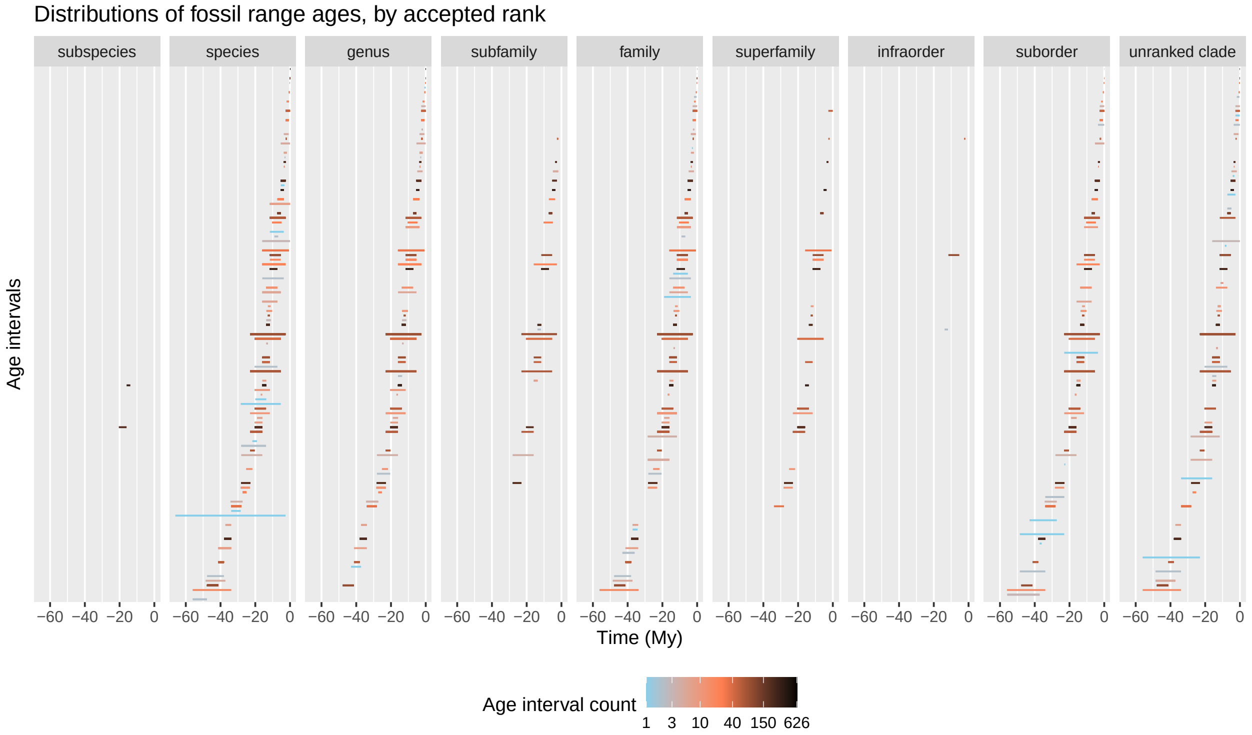
→ Most species have a early but not a late stratigraphic limit.

Minimum and maximum ages

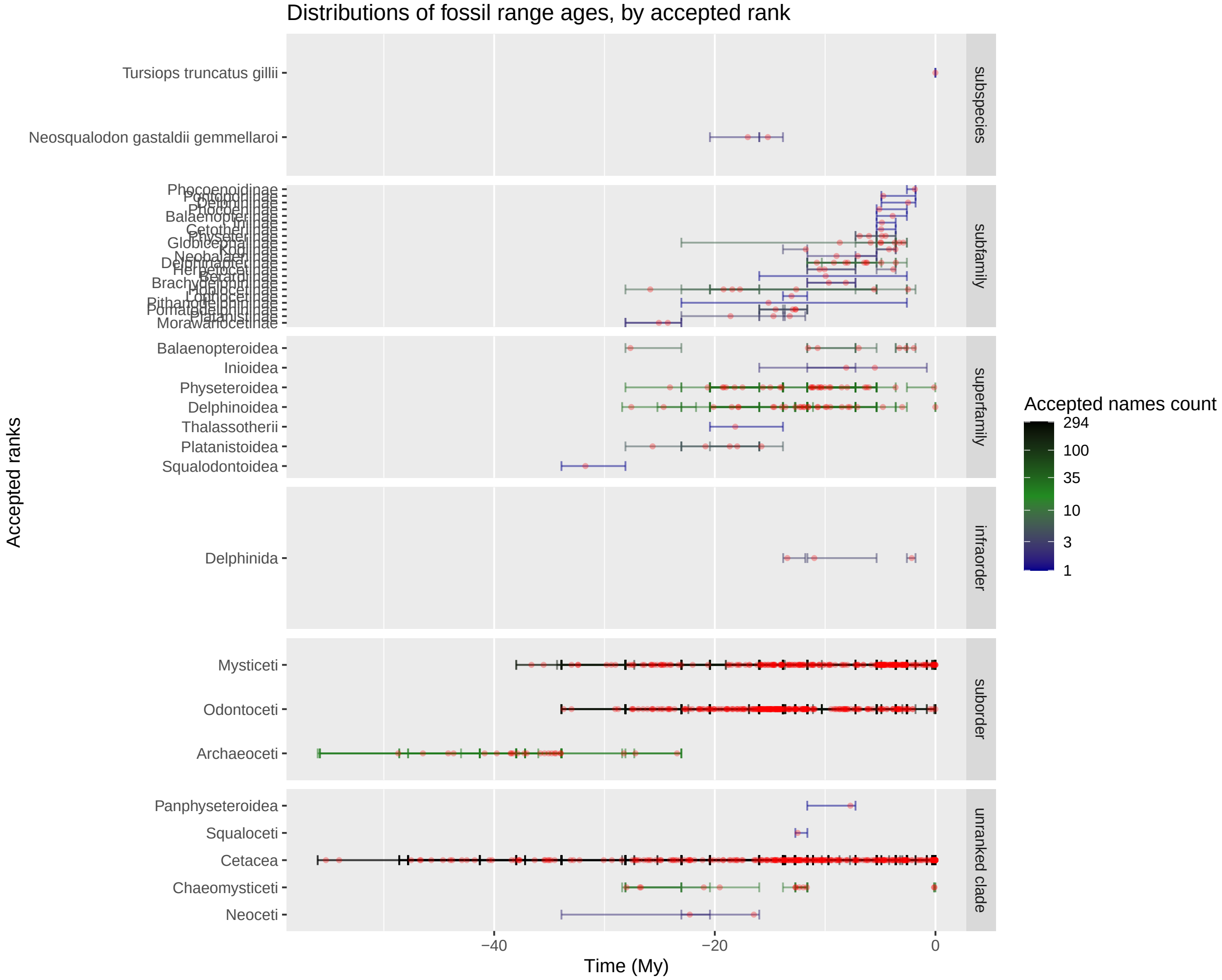


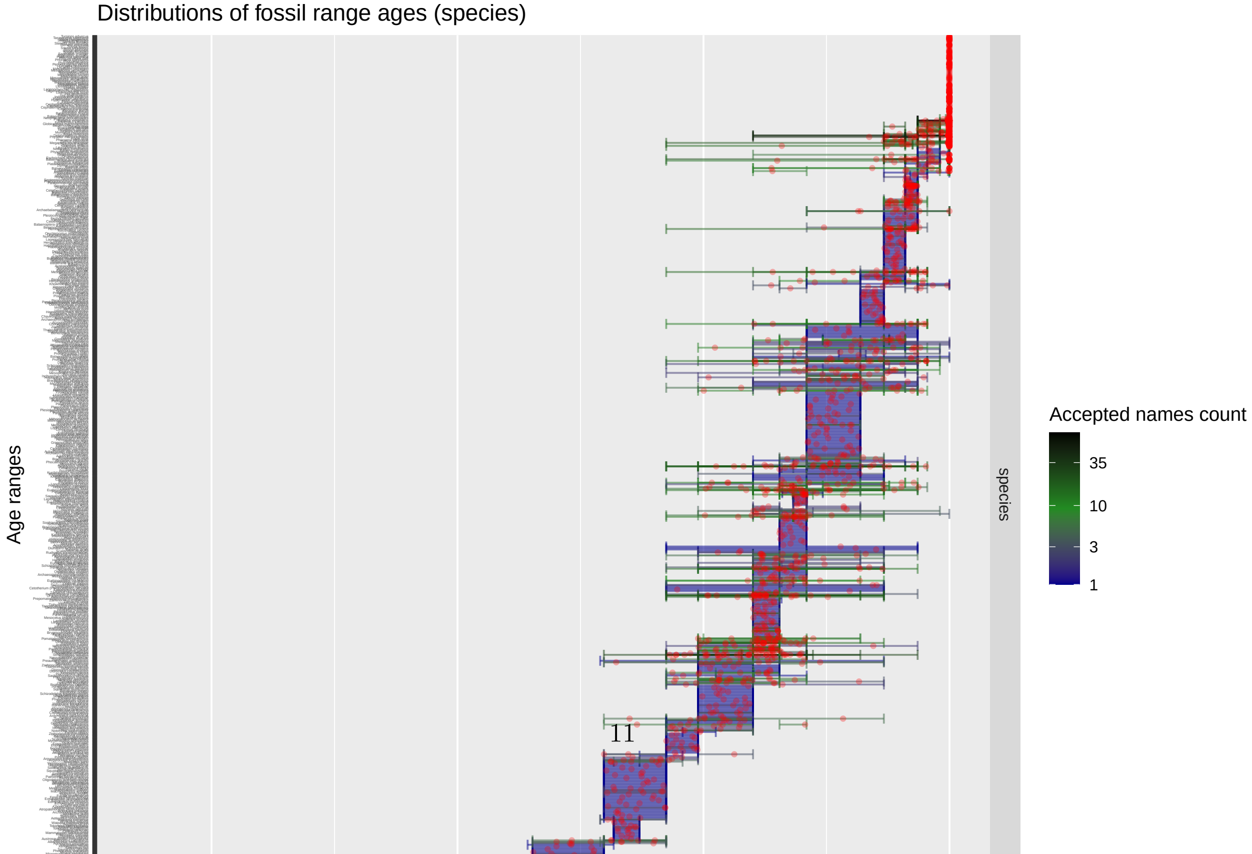
Time ranges = duration of the time intervals

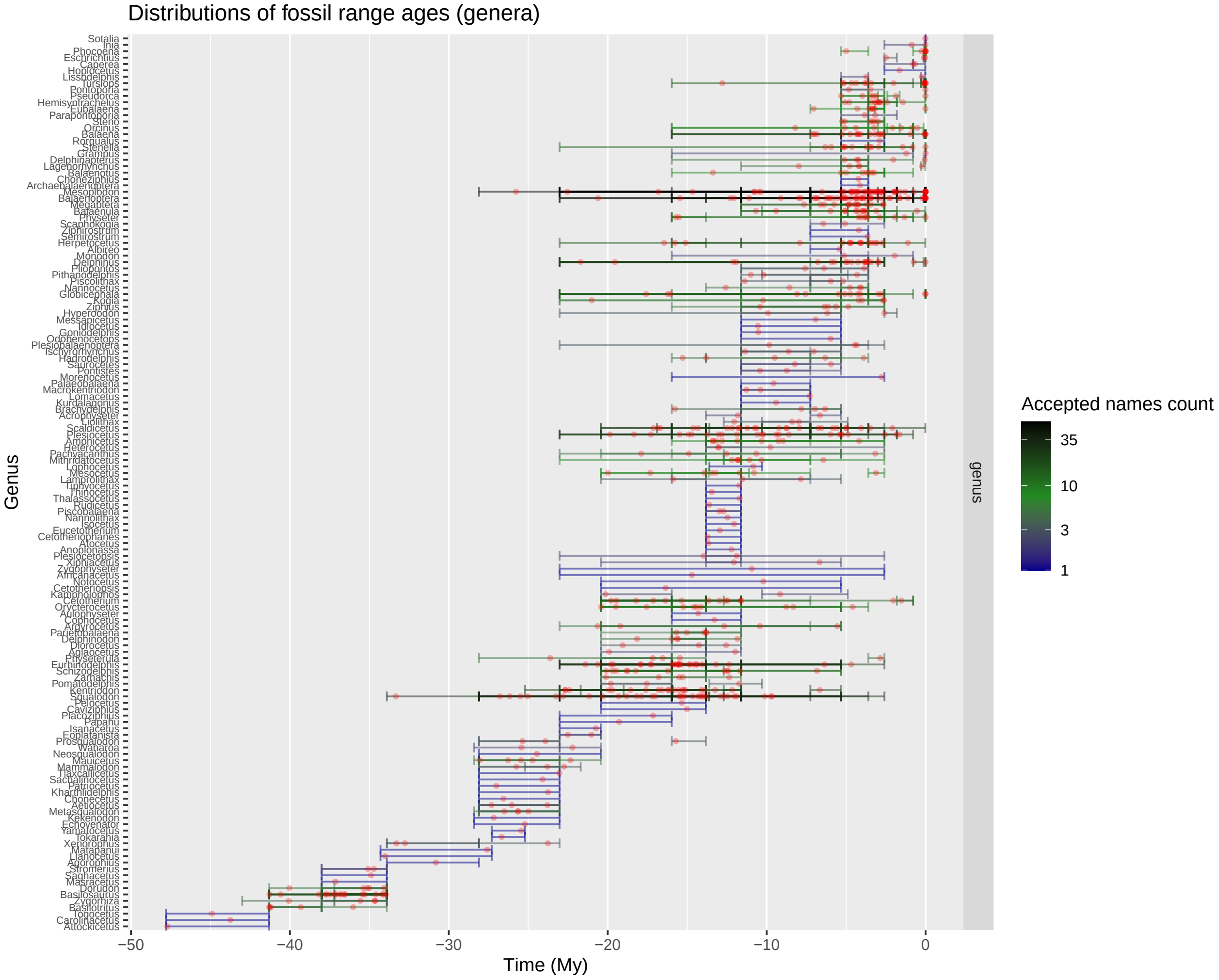
Count occurrences by age interval

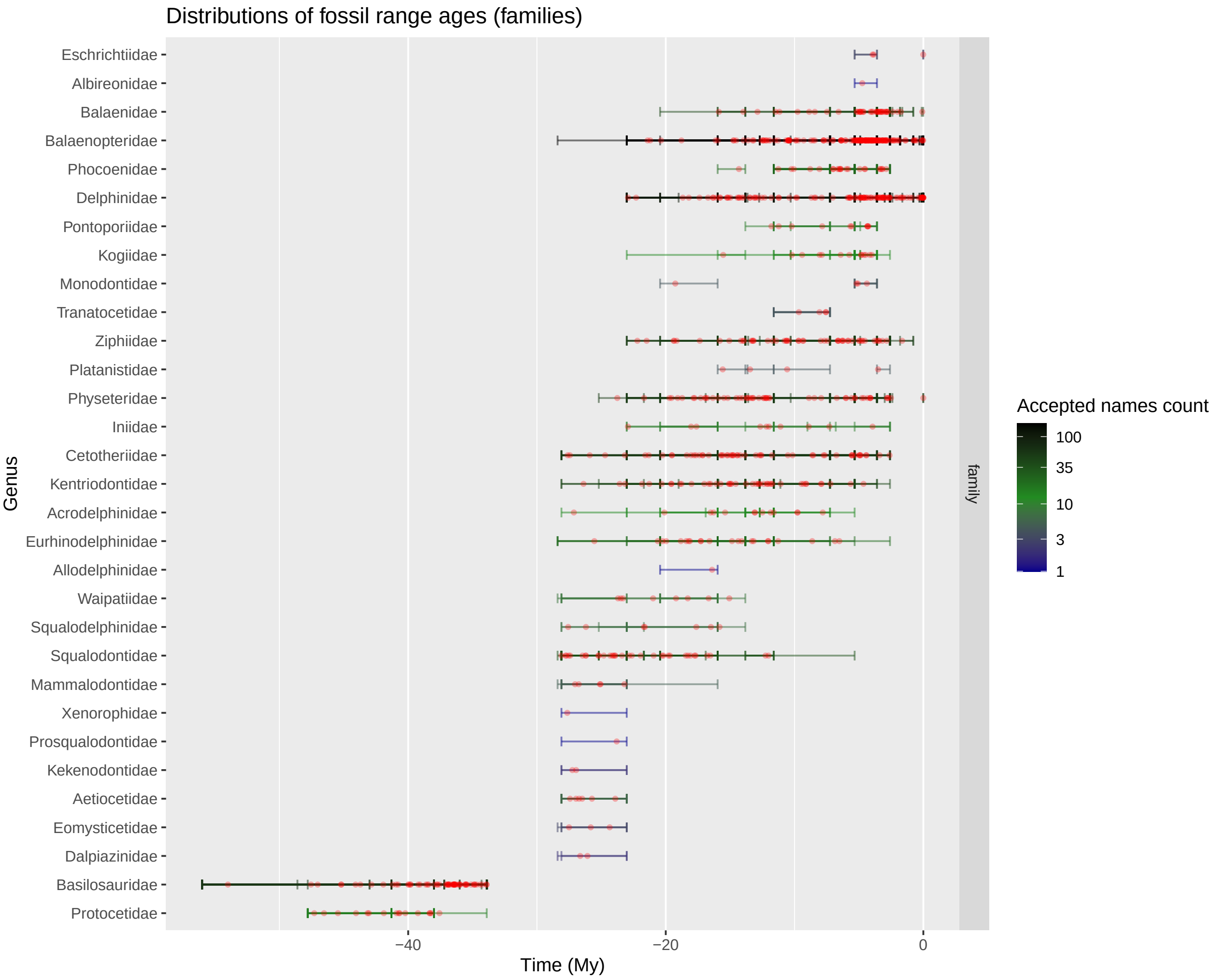


Count occurrences by accepted name





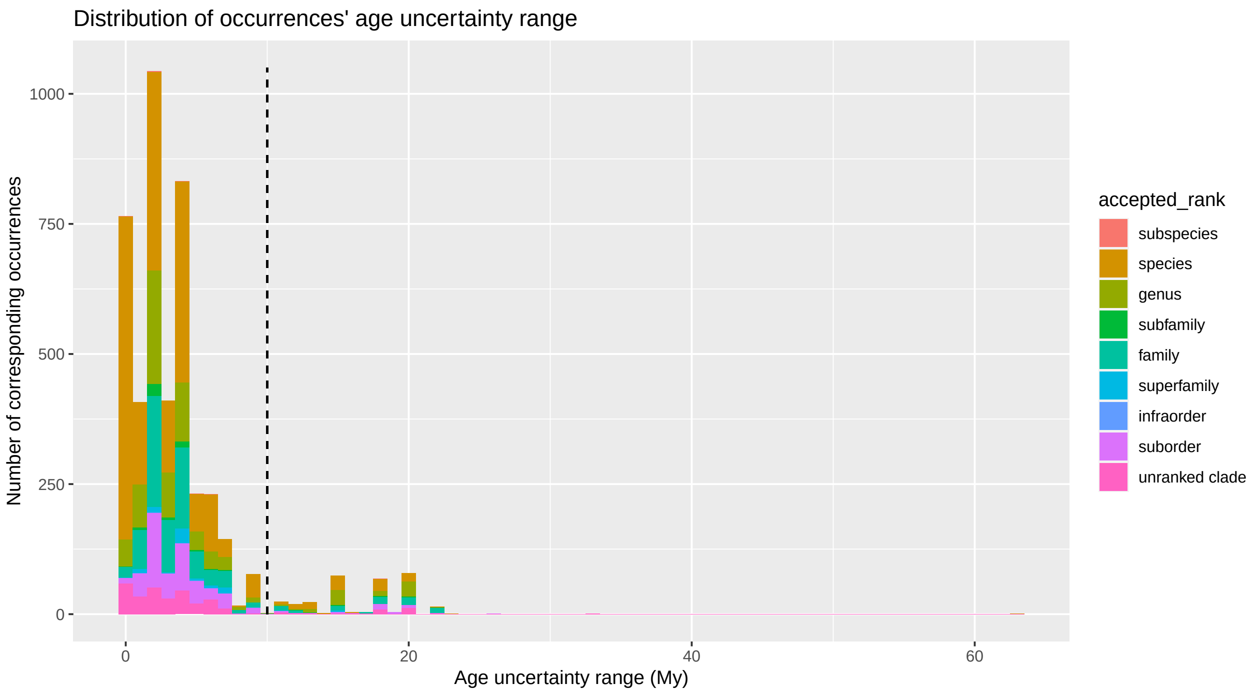

Distributions of fossil range ages (genera)
Genus
Sotalia
Inia
Phocoena
Eschrichtius
Caperea
Hoplocetus
Lissodelphis
Tursiops
Pontoporia
Pseudorca
Hemisyntrachelus
Eubalaena
Parapontoporia
Steno
Orcinus
Balaena
Rorqualus
Stenella
Grampus
Delphinapterus
Lagenorhynchus
Balaenotus
Choneziphius
Archaebalaenoptera
Mesoplodon
Balaenoptera
Megaptera
Balaenula
Physeter
Scaphokogia
Ziphirostrum
Semirostrum
Herpetocetus
Albireo
Monodon
Delphinus
Pliopontos
Pithanodelphis
Piscolithax
Nannocetus
Globicephala
Kogia
Ziphius
Hyperoodon
Messapicetus
Idiocetus
Goniodelphis
Odobenocetops
Plesiobalaenoptera
Ischyrorhynchus
Hadrodelphis
Saurocetes
Pontistes
Morenocetus
Palaeobalaena
Macrokentriodon
Lomacetus
Kurdalagonus
Brachydelphis
Acrophyseter
Liolithax
Scaldicetus
Plesiocetus
Amphicetus
Heterocetus
Pachyacanthus
Mithridatocetus
Lophocetus
Mesocetus
Lamprolithax
Tiphyocetus
Thinocetus
Thalassocetus
Rudicetus
Piscobalaena
Nannolithax
Isocetus
Eucetotherium
Cetotheriophanes
Atocetus
Anoplonassa
Plesiocetopsis
Xiphiacetus
Zygophyseter
Africanacetus
Notocetus
Cetotheriopsis
Kampholophos
Cetotherium
Orycterocetus
Aulophyseter
Cophocetus
Argyrocetus
Parietobalaena
Delphinodon
Diorocetus
Aglaocetus
Physeterula
Eurhinodelphis
Schizodelphis
Zarhachis
Pomatodelphis
Kentriodon
Squalodon
Pelocetus
Caviziphius
Placoziphius
Papahu
Isanacetus
Eoplatanista
Prosqualodon
Waharoa
Neosqualodon
Mauicetus
Mammalodon
Tlaxcallicetus
Sachalinocetus
Patriocetus
Kharthlidelphis
Chonecetus
Aetiocetus
Metasqualodon
Kekenodon
Echovenator
Yamatocetus
Tokarahia
Xenorophus
Matapanui
Llanocetus
Agorophius
Stromerius
Saghacetus
Masracetus
Dorudon
Basilosaurus
Zygorhiza
Basilotritus
Togocetus
Carolinacetus
Attockicetus
−50
−40
−30
−20
−10
0
Time (My)
genus
Accepted names count
35
10
3
1



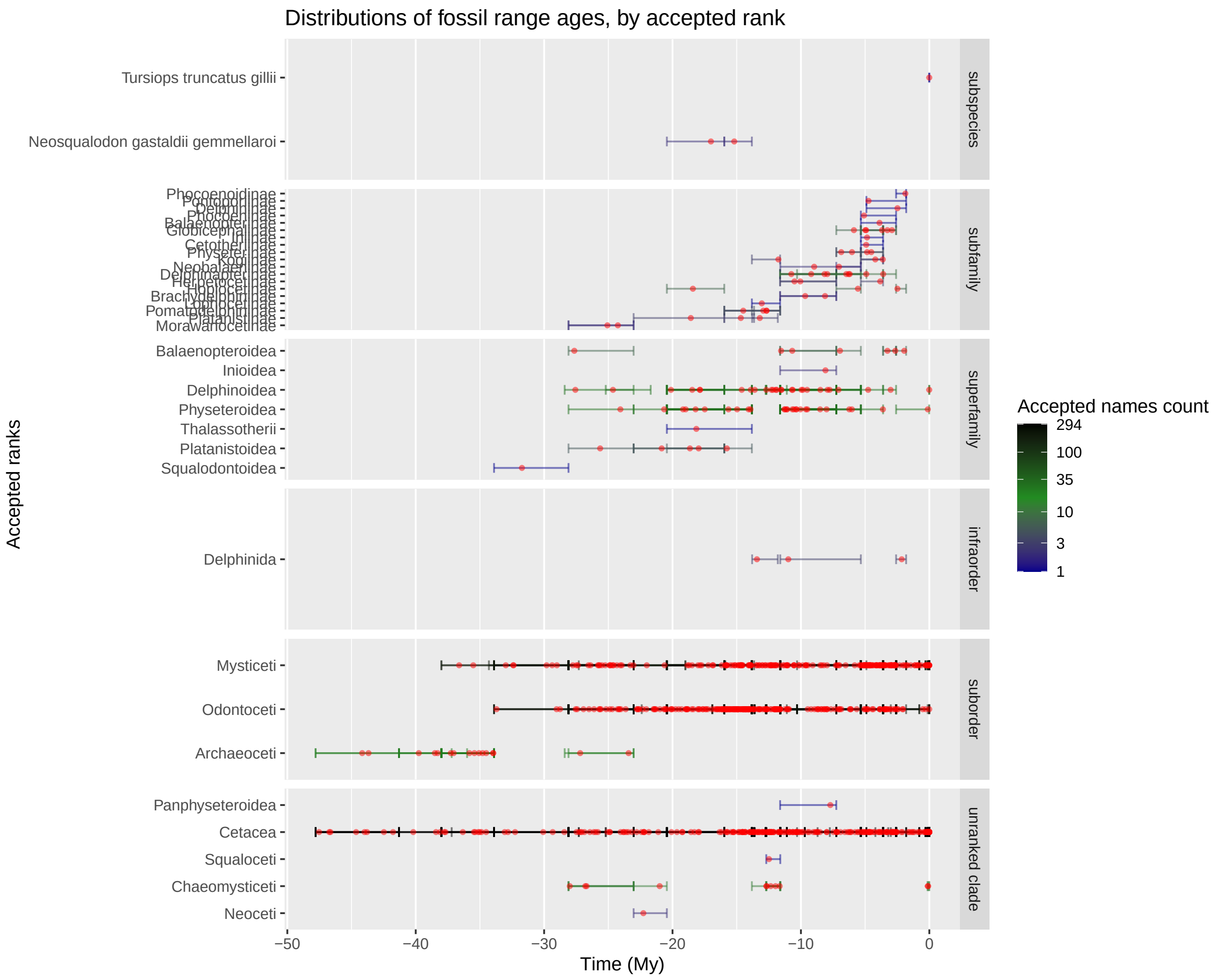
→ Some occurrences have too much age uncertainty, they could be removed because they are not very informative.

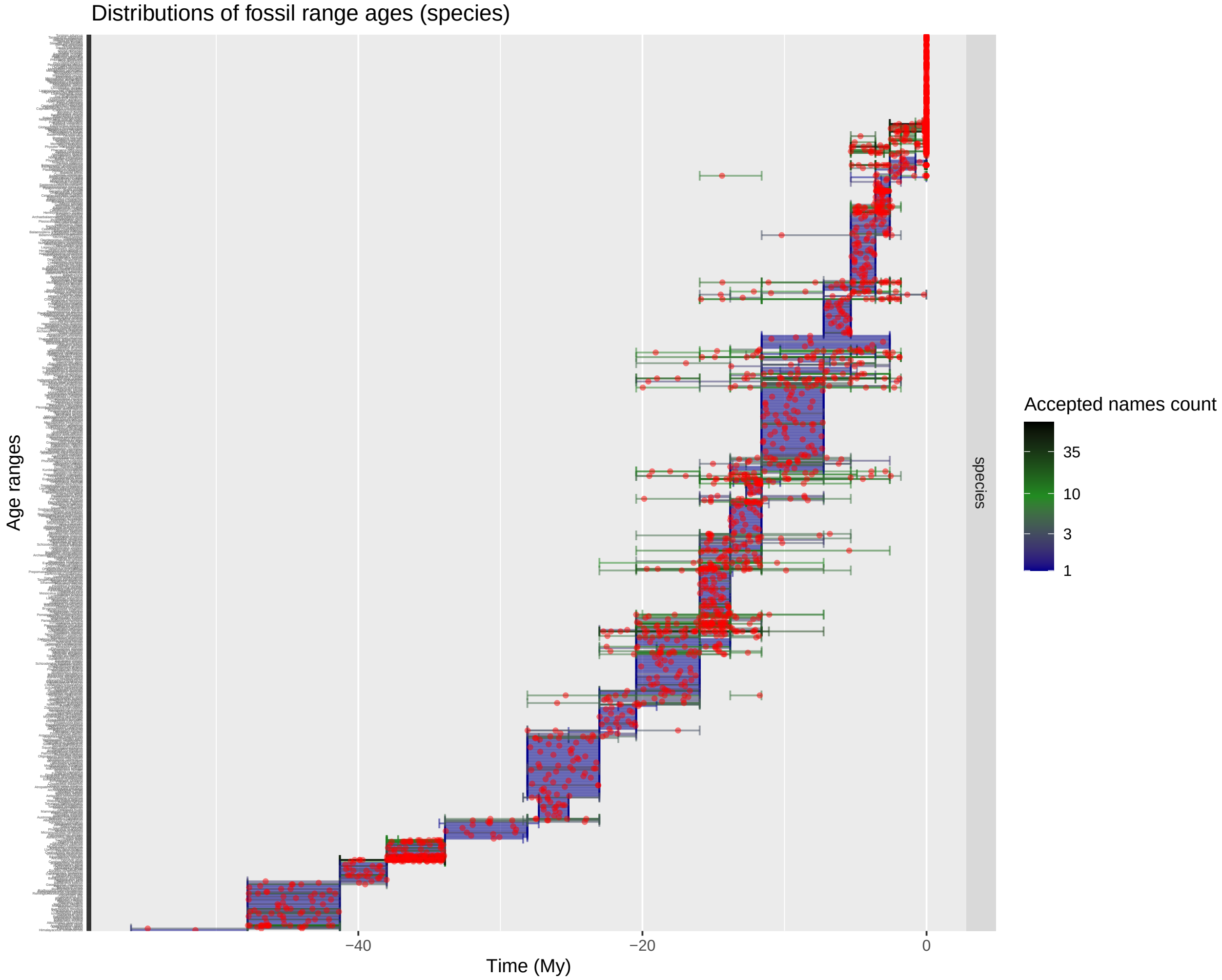
Remove occurrences with highly uncertain datation (range > 10My)

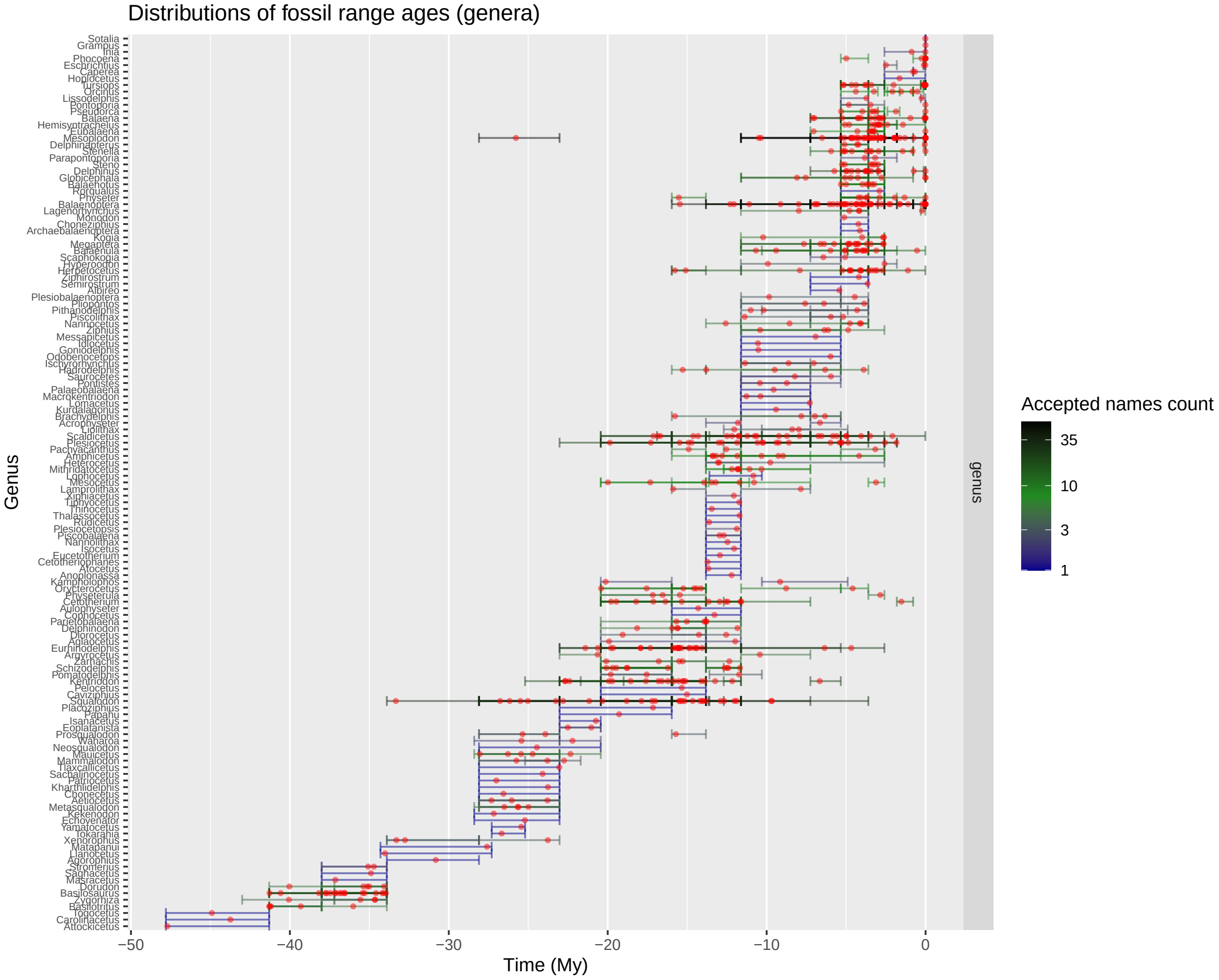
```
## [1] 4157   23
```

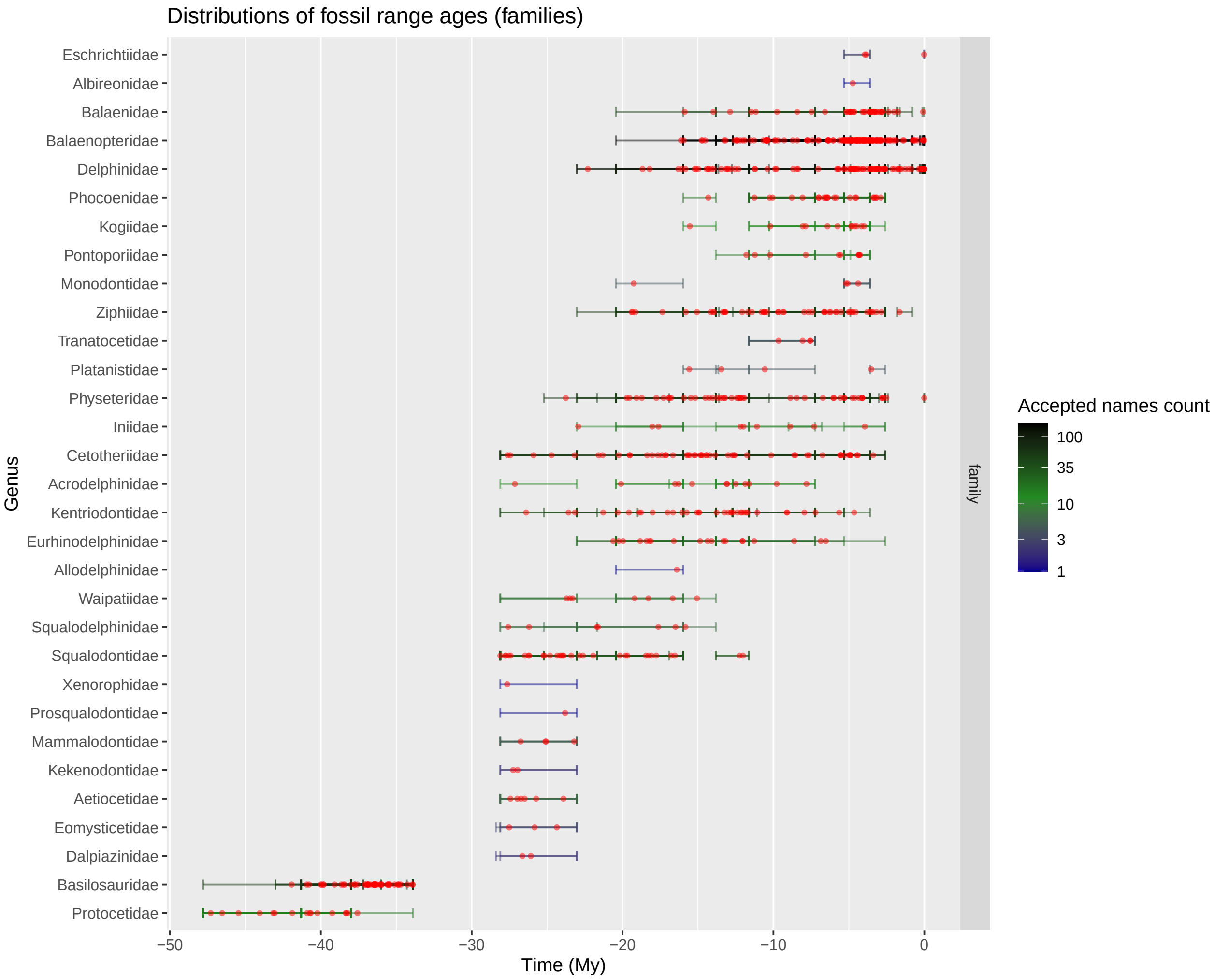


Most of occurrences (4157) show less than 10 My age uncertainty, let's keep only these ones.


Distributions of fossil range ages, by accepted rank
Accepted ranks
Tursiops truncatus gillii
Neosqualodon gastaldii gemmellaroi
subspecies
Phocoenoidinae
Pontoporiinae
Delphininae
Phocoeninae
Balaenopterinae
Globicephalinae
Iniinae
Cetotheriinae
Physeterinae
Koginae
Neobalaeninae
Delphinapterinae
Herpetocetinae
Hoplocetinae
Brachydelphininae
Lophocetinae
Pomatodelphininae
Platanistinae
Morawanocetinae
subfamily
Balaenopteroidea
Inioidea
Delphinoidea
Physeteroidea
Thalassotherii
Platanistoidea
Squalodontoidea
superfamily
Delphinida
infraorder
Mysticeti
Odontoceti
Archaeoceti
suborder
Panphyseteroidea
Cetacea
Squaloceti
Chaeomysticeti
Neoceti
unranked clade
−50
−40
−30
−20
−10
0
Time (My)
Accepted names count
294
100
35
10
3
1

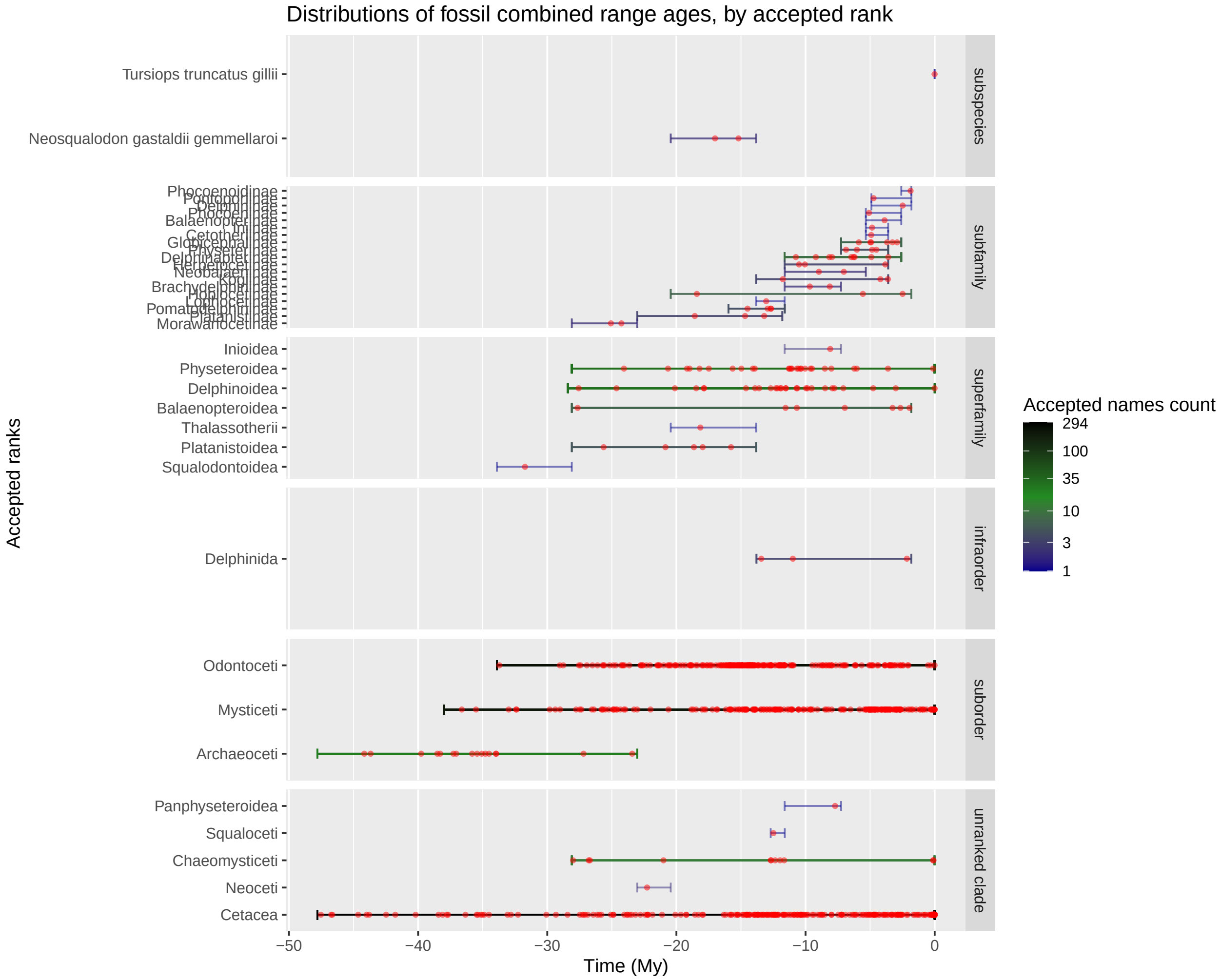

Distributions of fossil range ages (species)
Age ranges
Time (My)
−40
−20
0
species
Accepted names count
35
10
3
1

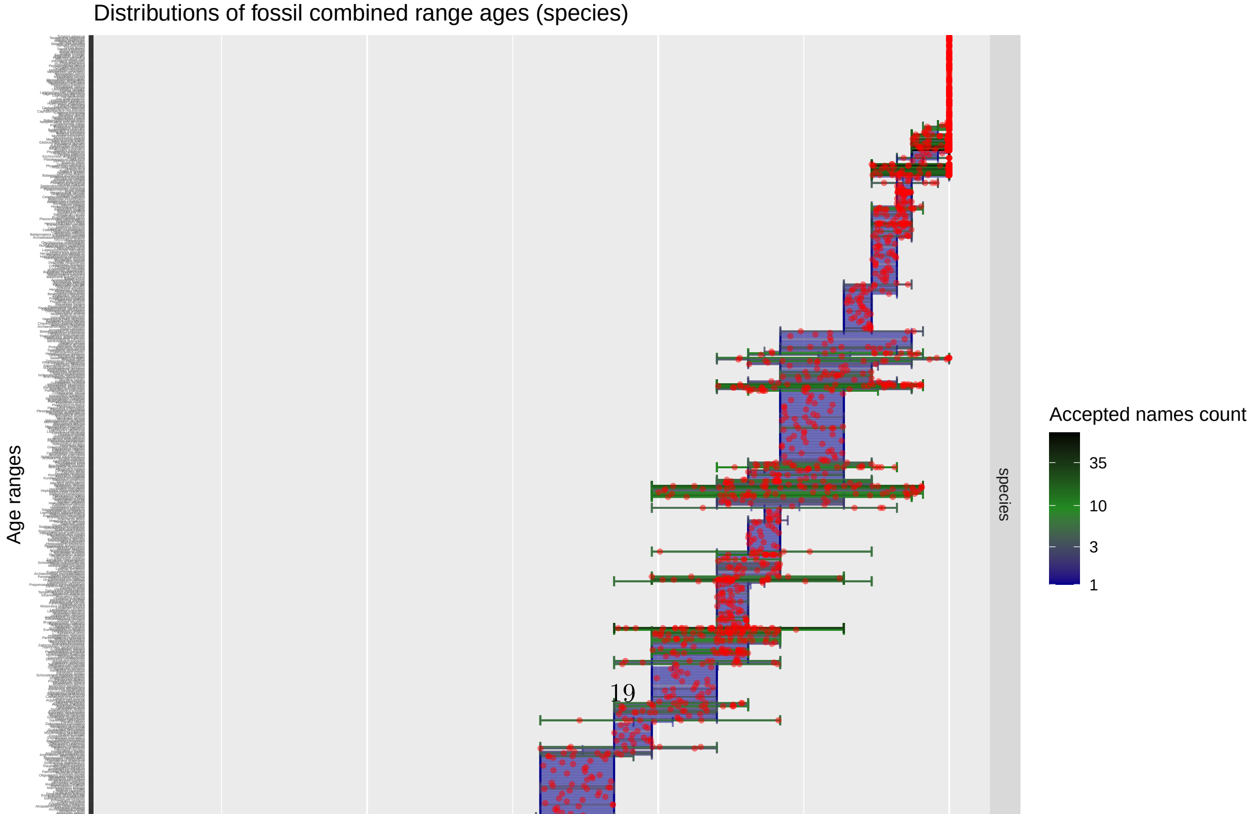

Distributions of fossil range ages (genera)
Genus
Sotalia
Grampus
Inia
Phocoena
Eschrichtius
Caperea
Hoplocetus
Tursiops
Orcinus
Lissodelphis
Pontoporia
Pseudorca
Balaena
Hemisyntrachelus
Eubalaena
Mesoplodon
Delphinapterus
Stenella
Parapontoporia
Steno
Delphinus
Globicephala
Balaenotus
Rorqualus
Physeter
Balaenoptera
Lagenorhynchus
Monodon
Choneziphius
Archaebalaenoptera
Kogia
Megaptera
Balaenula
Scaphokogia
Hyperoodon
Herpetocetus
Ziphirostrum
Semirostrum
Albireo
Plesiobalaenoptera
Pliopontos
Pithanodelphis
Piscolithax
Nannocetus
Ziphius
Messapicetus
Idiocetus
Goniodelphis
Odobenocetops
Ischyrorhynchus
Hadrodelphis
Saurocetes
Pontistes
Palaeobalaena
Macrokentriodon
Lomacetus
Kurdalagonus
Brachydelphis
Acrophyseter
Liolithax
Scaldicetus
Plesiocetus
Pachyacanthus
Amphicetus
Heterocetus
Mithridatocetus
Lophocetus
Mesocetus
Lamprolithax
Xiphiacetus
Tiphyocetus
Thinocetus
Thalassocetus
Rudicetus
Plesiocetopsis
Piscobalaena
Nannolithax
Isocetus
Eucetotherium
Cetotheriophanes
Atocetus
Anoplonassa
Kampholophos
Orycterocetus
Physeterula
Cetotherium
Aulophyseter
Cophocetus
Parietobalaena
Delphinodon
Diorocetus
Aglaocetus
Eurhinodelphis
Argyrocetus
Zarhachis
Schizodelphis
Pomatodelphis
Kentriodon
Pelocetus
Caviziphius
Squalodon
Placoziphius
Papahu
Isanacetus
Eoplatanista
Prosqualodon
Waharoa
Neosqualodon
Mauicetus
Mammalodon
Tlaxcallicetus
Sachalinocetus
Patriocetus
Kharthlidelphis
Chonecetus
Aetiocetus
Metasqualodon
Kekenodon
Echovenator
Yamatocetus
Tokarahia
Xenorophus
Matapanui
Llanocetus
Agorophius
Stromerius
Saghacetus
Masracetus
Dorudon
Basilosaurus
Zygorhiza
Basilotritus
Togocetus
Carolinacetus
Attockicetus
−50
−40
−30
−20
−10
0
Time (My)
genus
Accepted names count
35
10
3
1



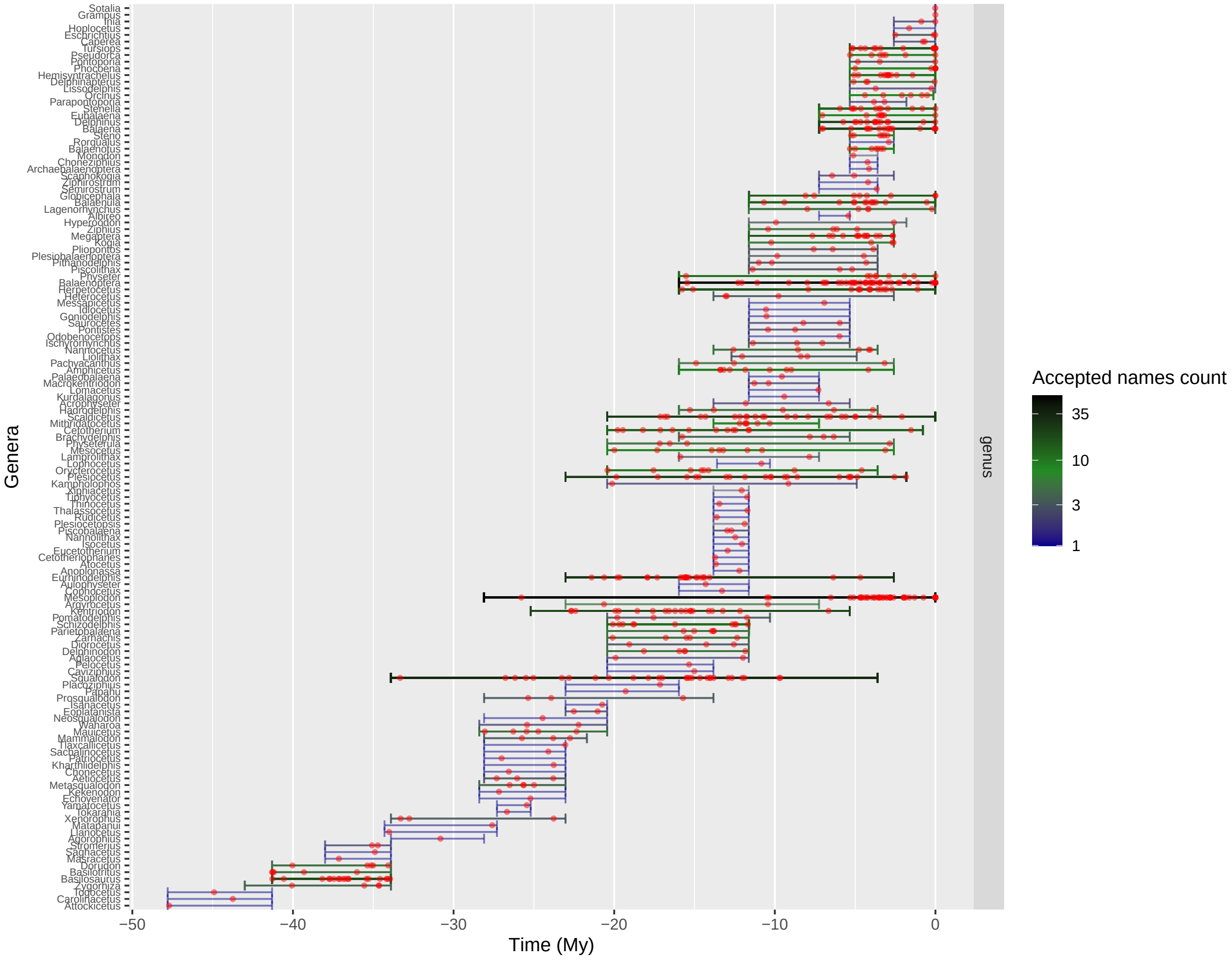
→ Some species (or other ranks) have several occurrences with several time ranges, let's combine them into a unique range covering all the other.

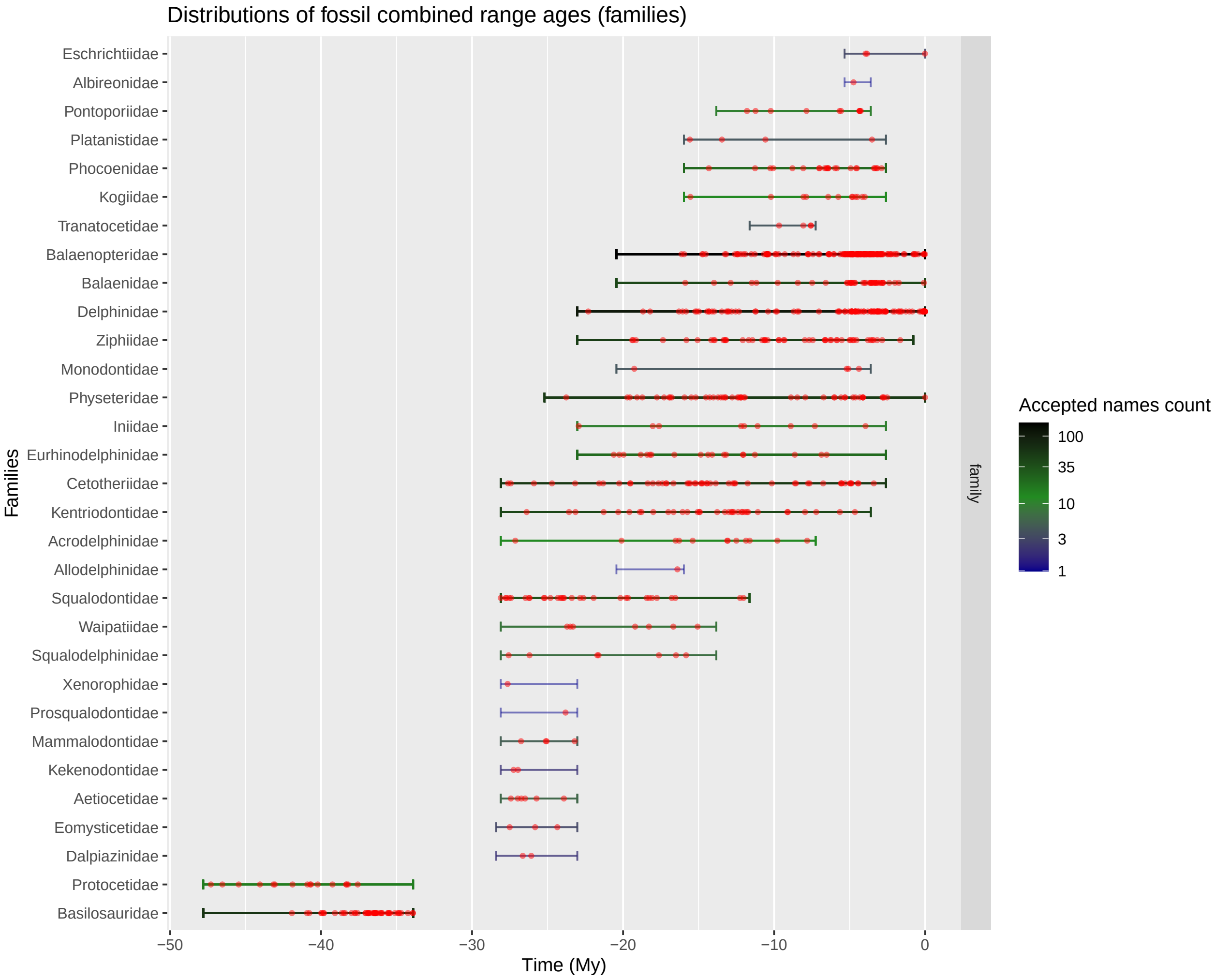
Combined time ranges = unique time range for occurrences with the same name (without the biggest ones)


Distributions of fossil combined range ages, by accepted rank


Distributions of fossil combined range ages (species)

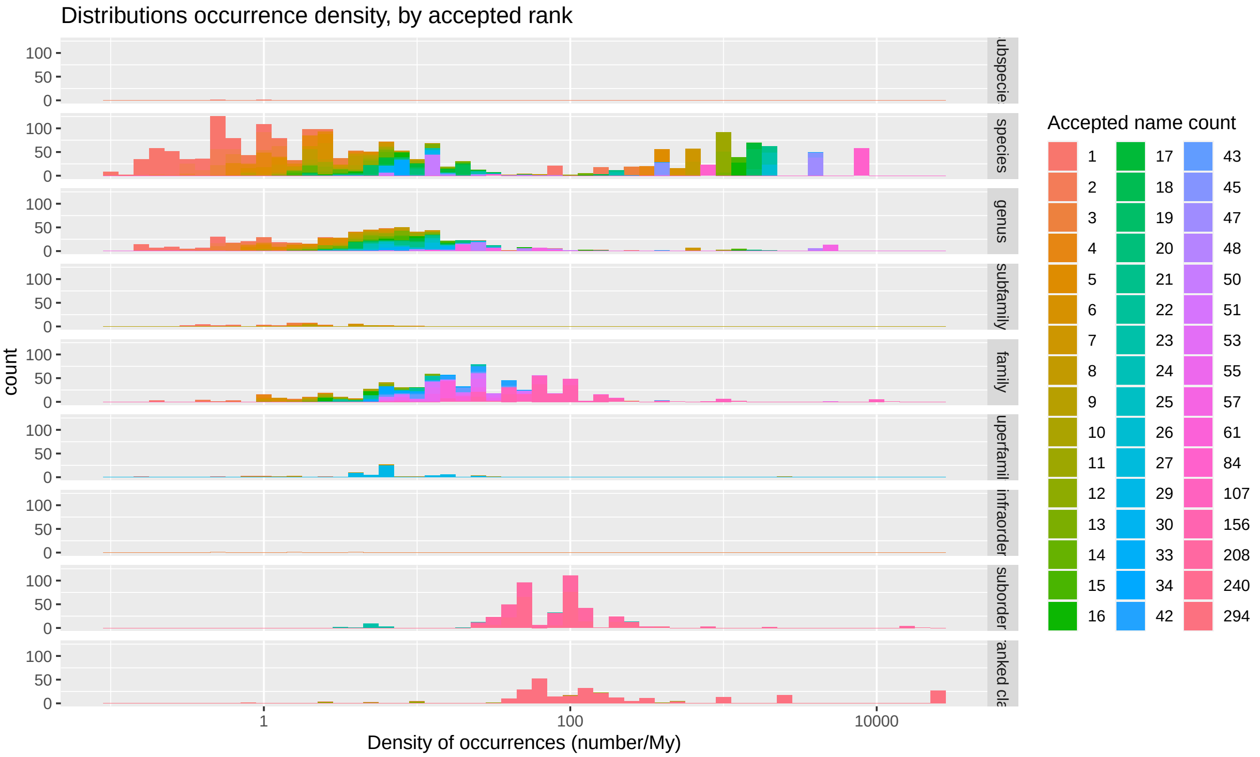
Distributions of fossil combined range ages (genera)



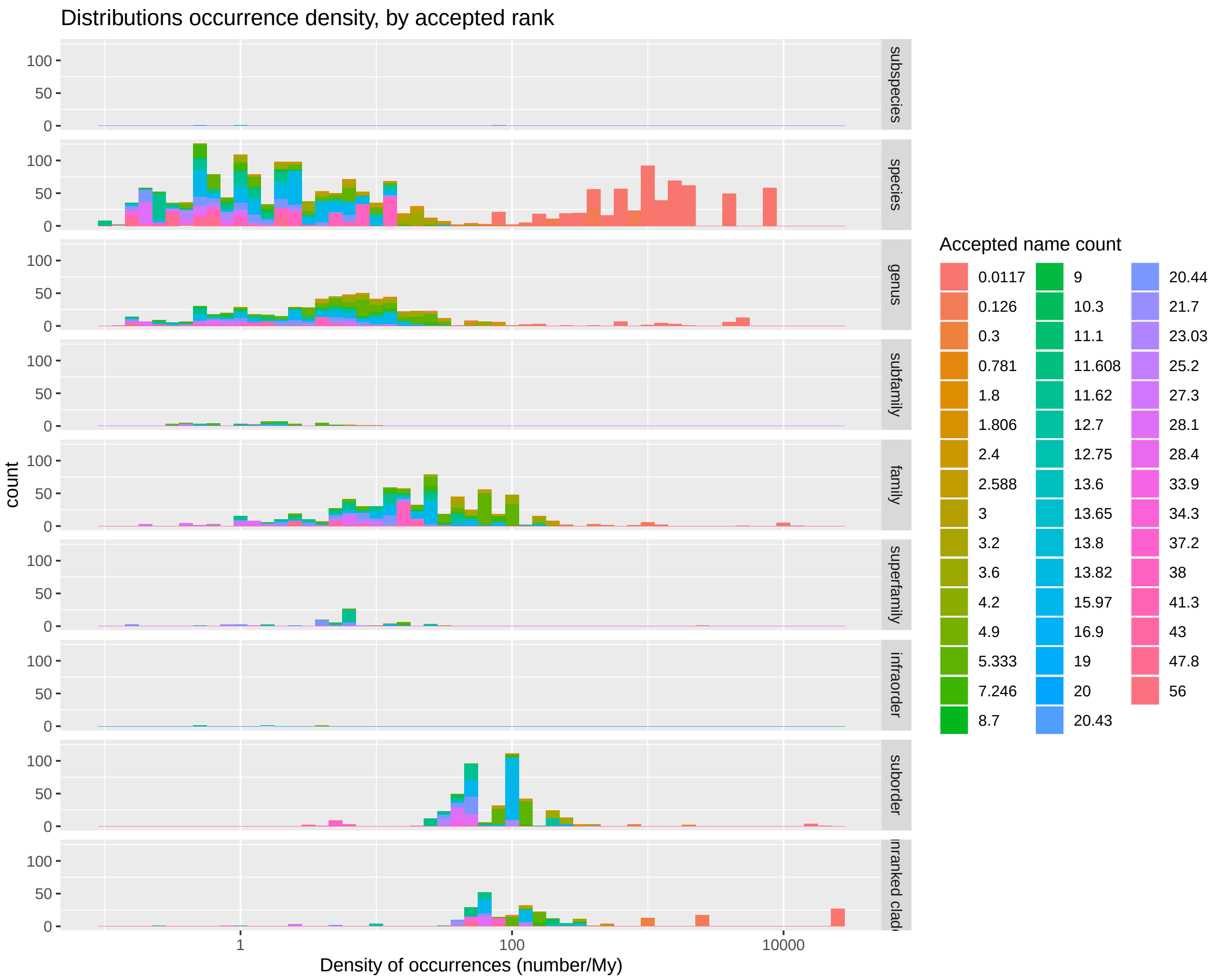

Distributions of fossil combined range ages (families)
Eschrichtiidae
Albireonidae
Pontoporiidae
Platanistidae
Phocoenidae
Kogiidae
Tranatocetidae
Balaenopteridae
Balaenidae
Delphinidae
Ziphiidae
Monodontidae
Physeteridae
Iniidae
Eurhinodelphinidae
Cetotheriidae
Kentriodontidae
Acrodelphinidae
Allodelphinidae
Squalodontidae
Waipatiidae
Squalodelphinidae
Xenorophidae
Prosqualodontidae
Mammalodontidae
Kekenodontidae
Aetiocetidae
Eomysticetidae
Dalpiazinidae
Protocetidae
Basilosauridae
Families
−50
−40
−30
−20
−10
0
Time (My)
family
Accepted names count
100
35
10
3
1

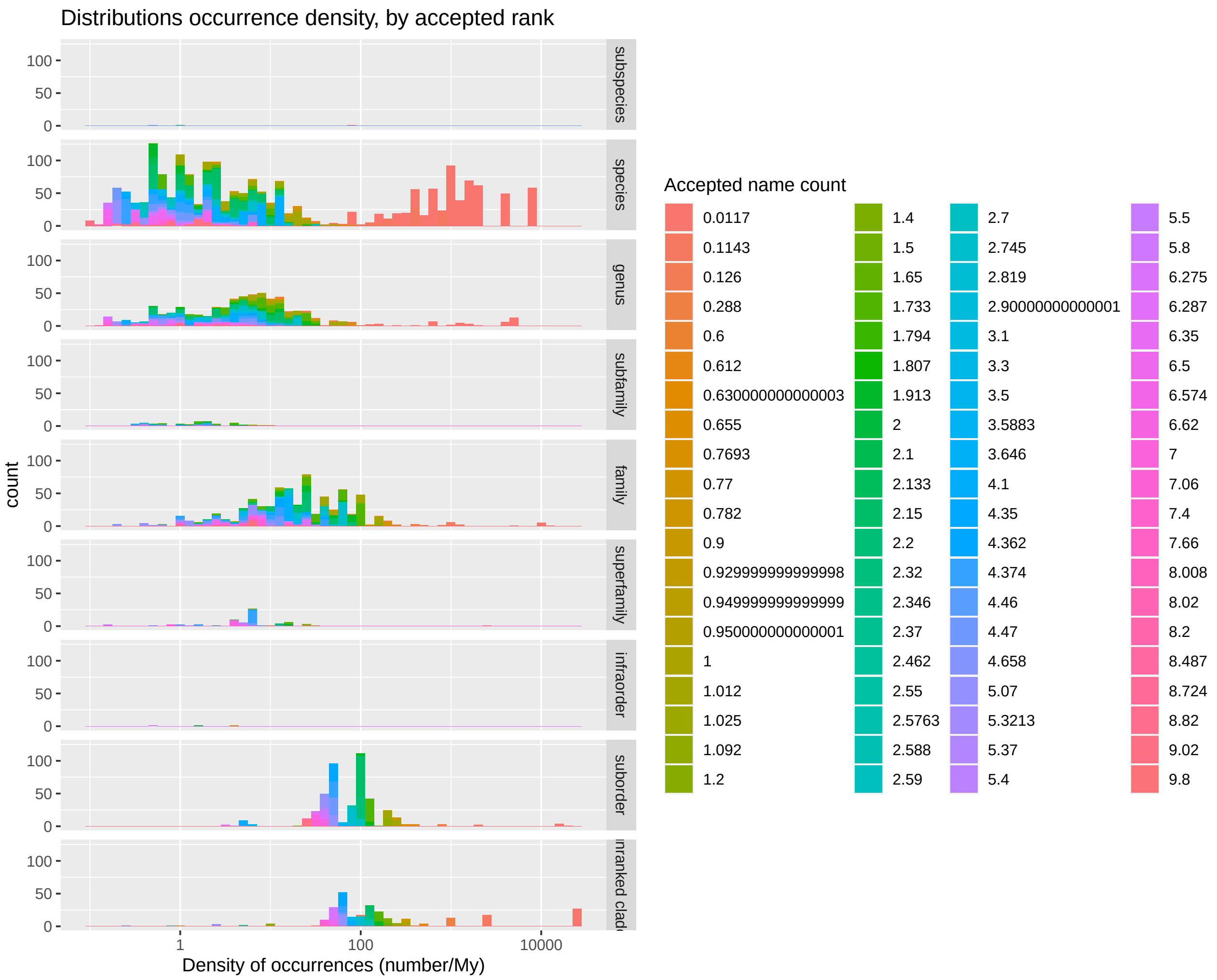
Occurrence density

Density distributions



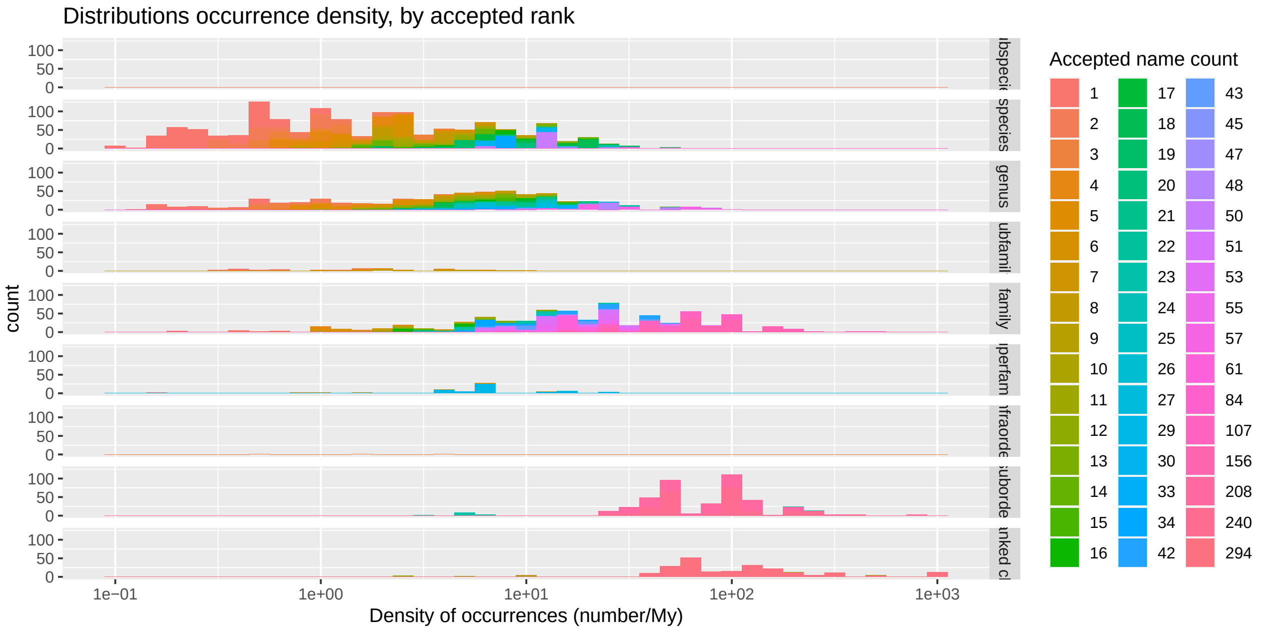
→ Density logically increases as taxa ranks increase, but there is a cluster of high density within the species. Let's identify its origin :


Distributions occurrence density, by accepted rank
count
100
50
0
subspecies
species
genus
subfamily
family
superfamily
infraorder
suborder
nranked clad
1
100
10000
Density of occurrences (number/My)
Accepted name count
0.0117
0.126
0.3
0.781
1.8
1.806
2.4
2.588
3
3.2
3.6
4.2
4.9
5.333
7.246
8.7
9
10.3
11.1
11.608
11.62
12.7
12.75
13.6
13.65
13.8
13.82
15.97
16.9
19
20
20.43
20.44
21.7
23.03
25.2
27.3
28.1
28.4
33.9
34.3
37.2
38
41.3
43
47.8
56



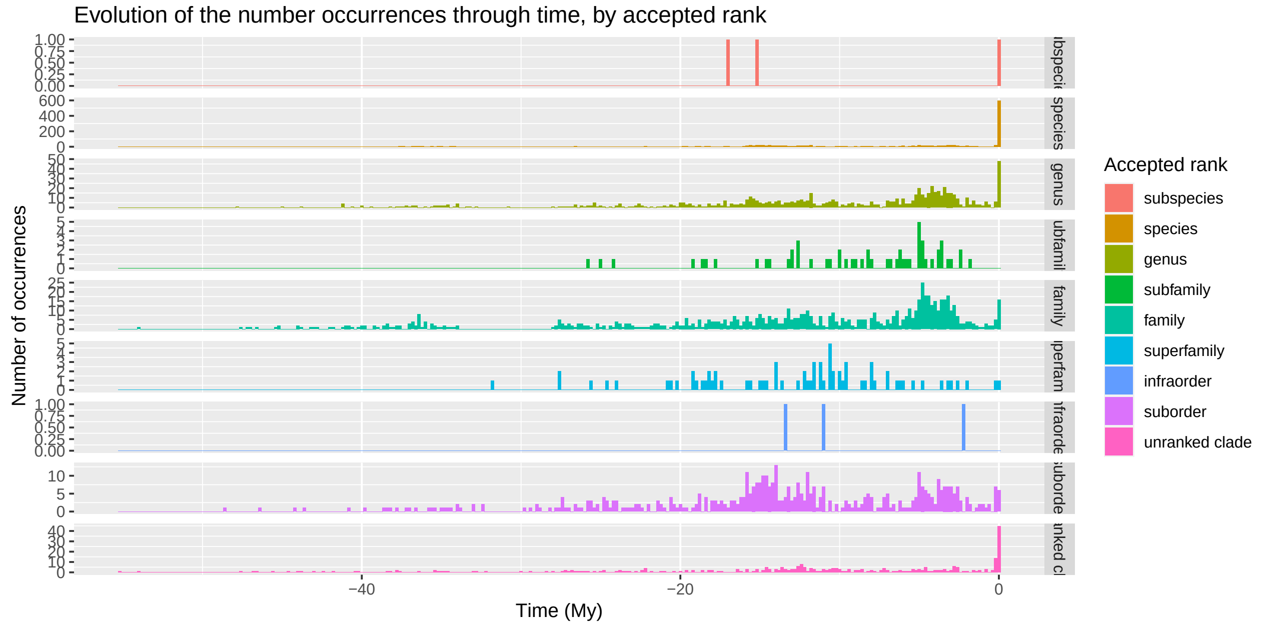
→ This cluster corresponds to recent samples (< 150.000 years) with very precise datation (different technique ?). Let's try to remove it but later it could be interesting to subsample instead.

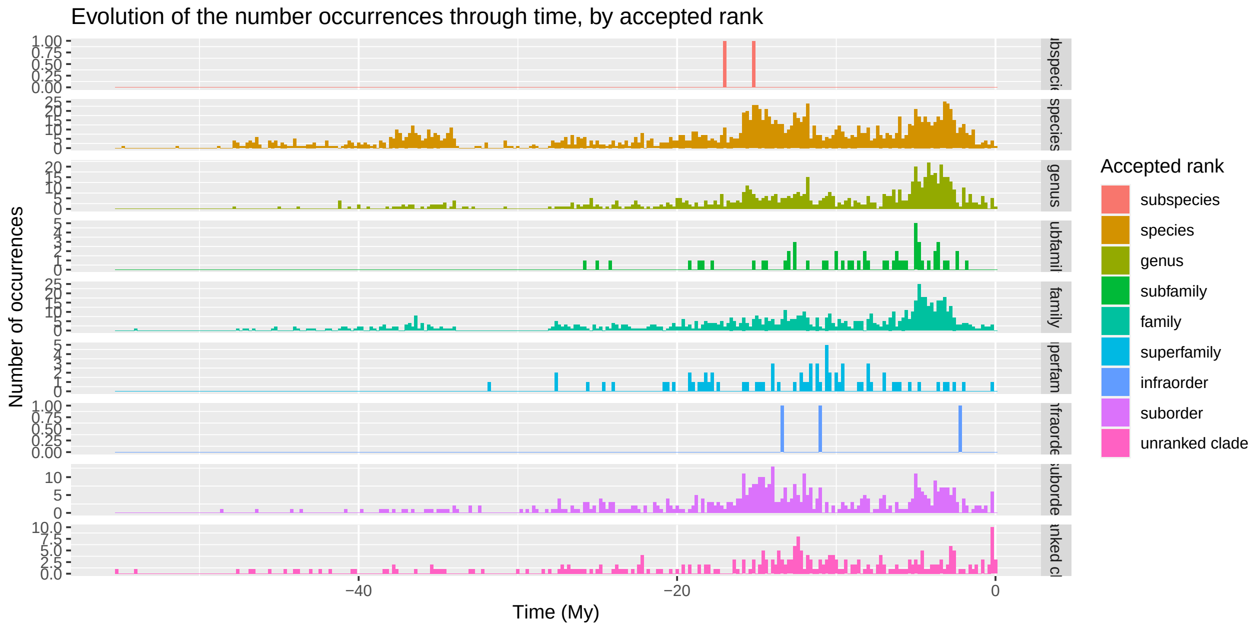
Remove highly concentred occurrences at present



→ The distributions look normal again.

Compare the number of occurrences through time with or without the concentrated ones at present

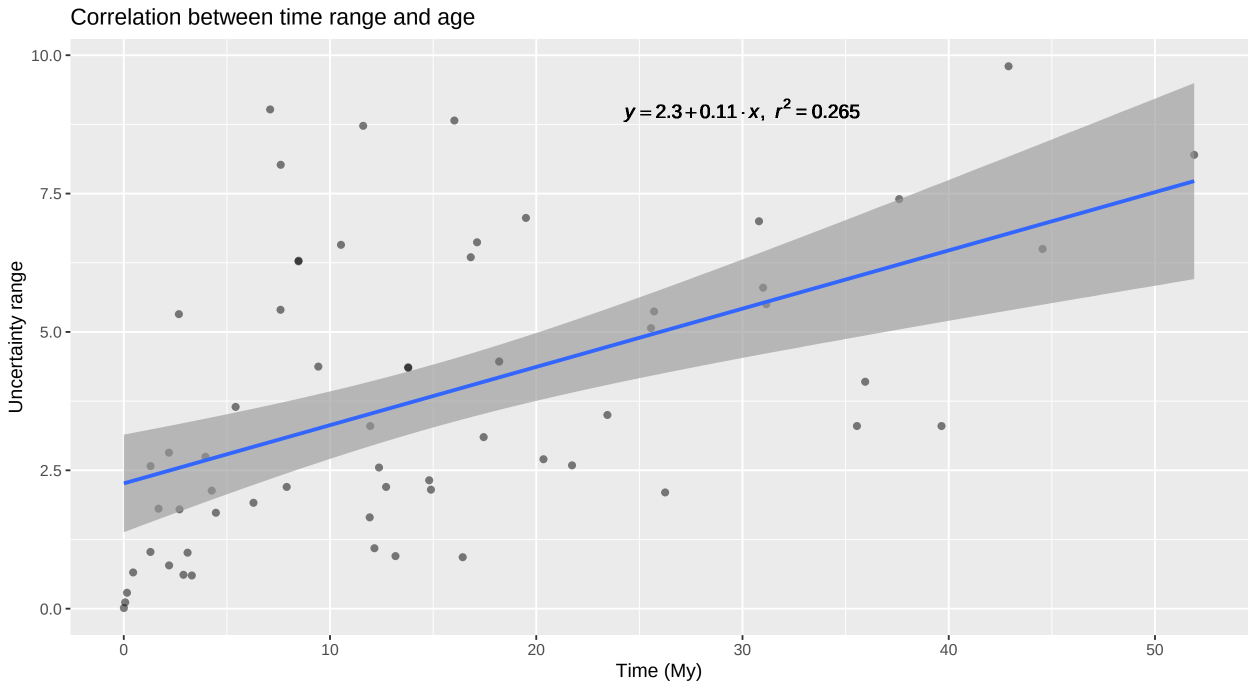


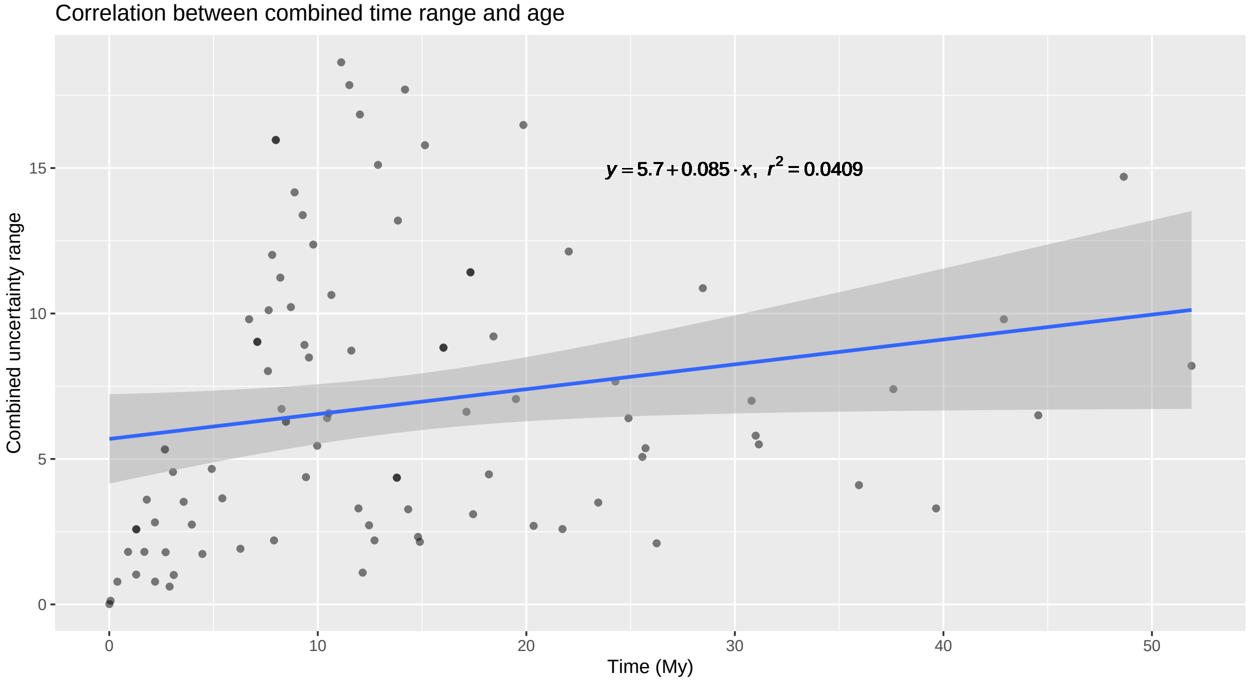


→ The removal gives much clearer occurrence distributions at the species and genera levels.

Correlation between time range and age

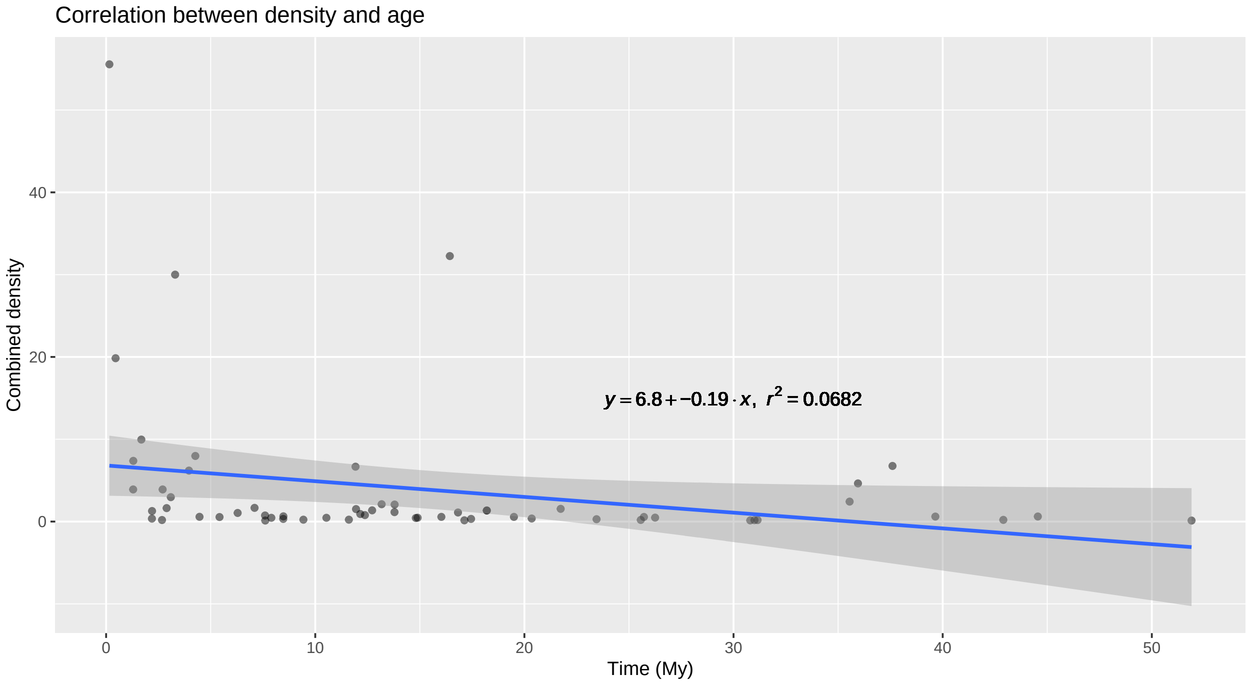
If we want to correct species abundance differences based on the number of occurrences in the time range ("density"), those factors should not depend on time in order to avoid penalizing periods with bigger ranges.

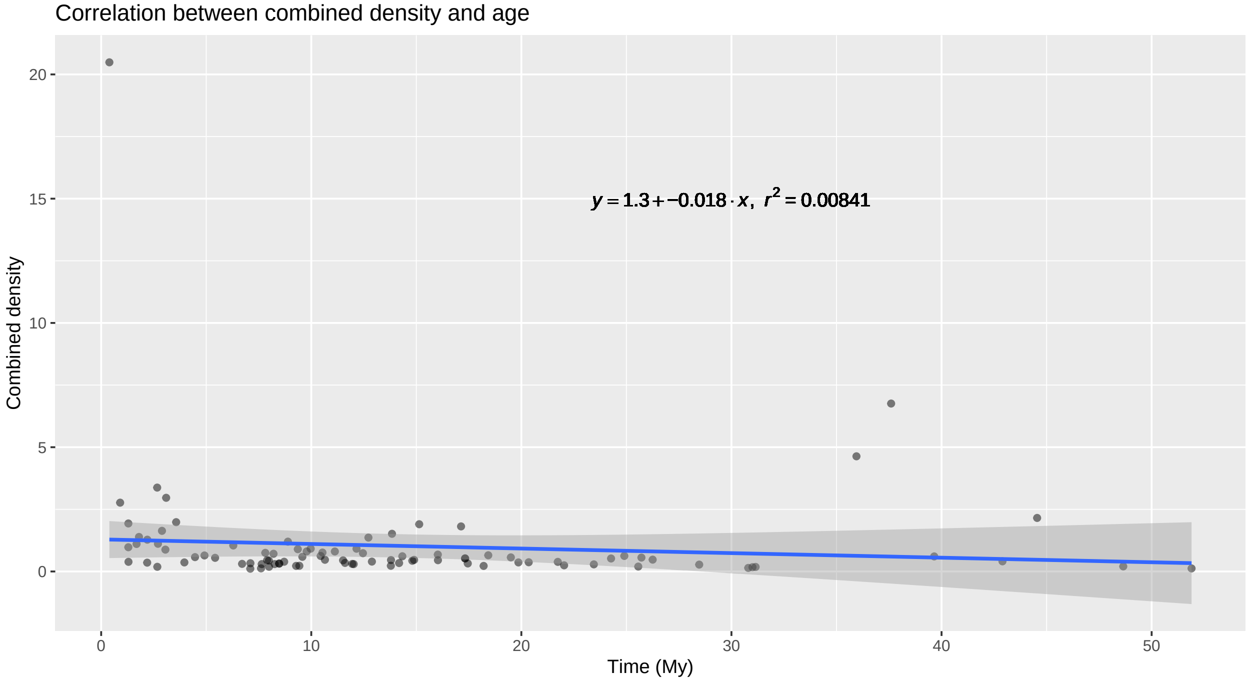




→ It seems that age range in much less correlated with time when we take the full combined range into account. However this plot is quite ugly (“triangle” instead of a nice point cloud) so this correlation may not be very meaningful. Nevertheless, we will use those ranges for normalising the occurrence density because we don’t want to penalize older specimens.

Let’s look at the density directly, because this is what is interesting us directly.

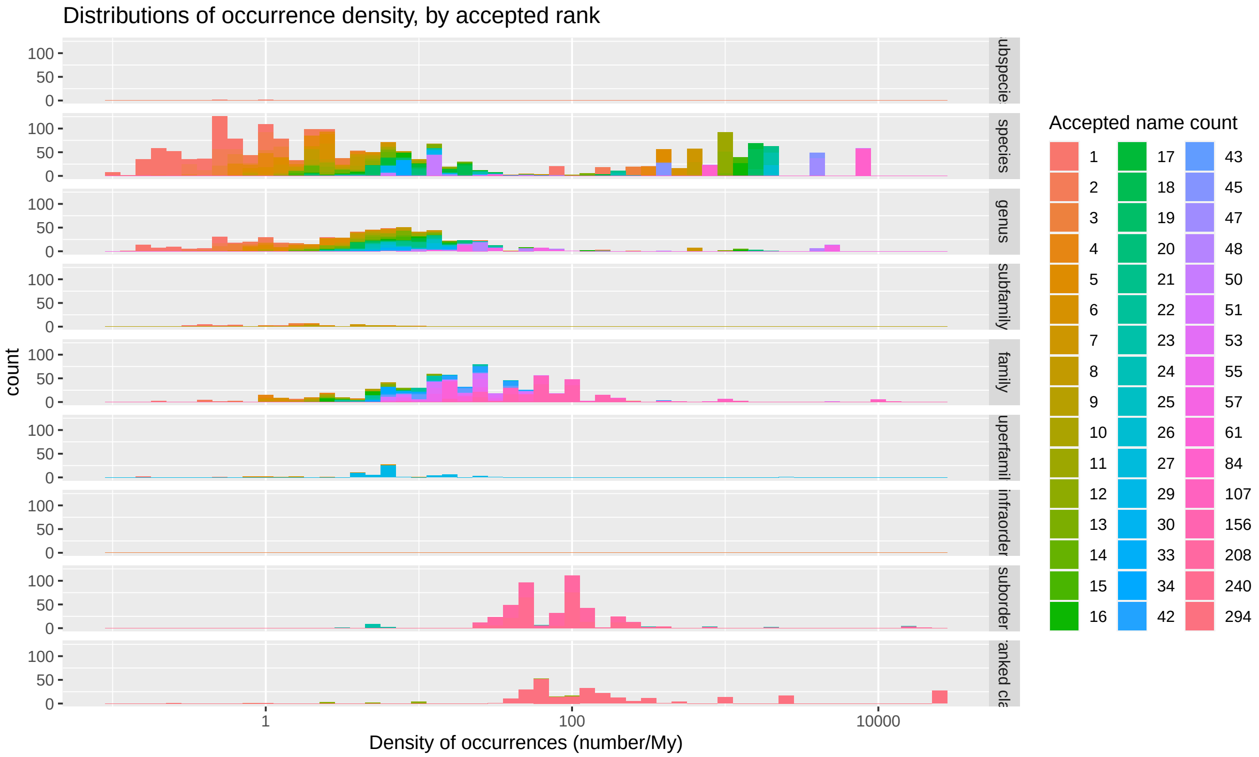


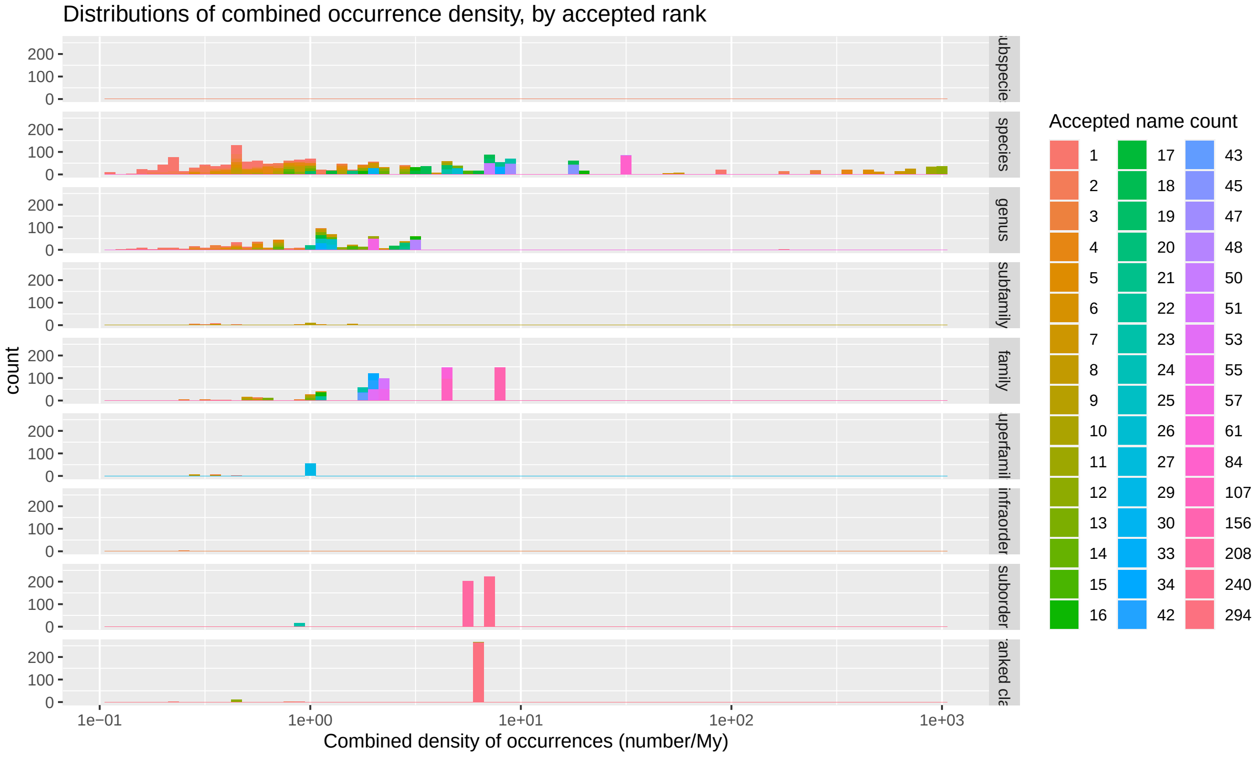


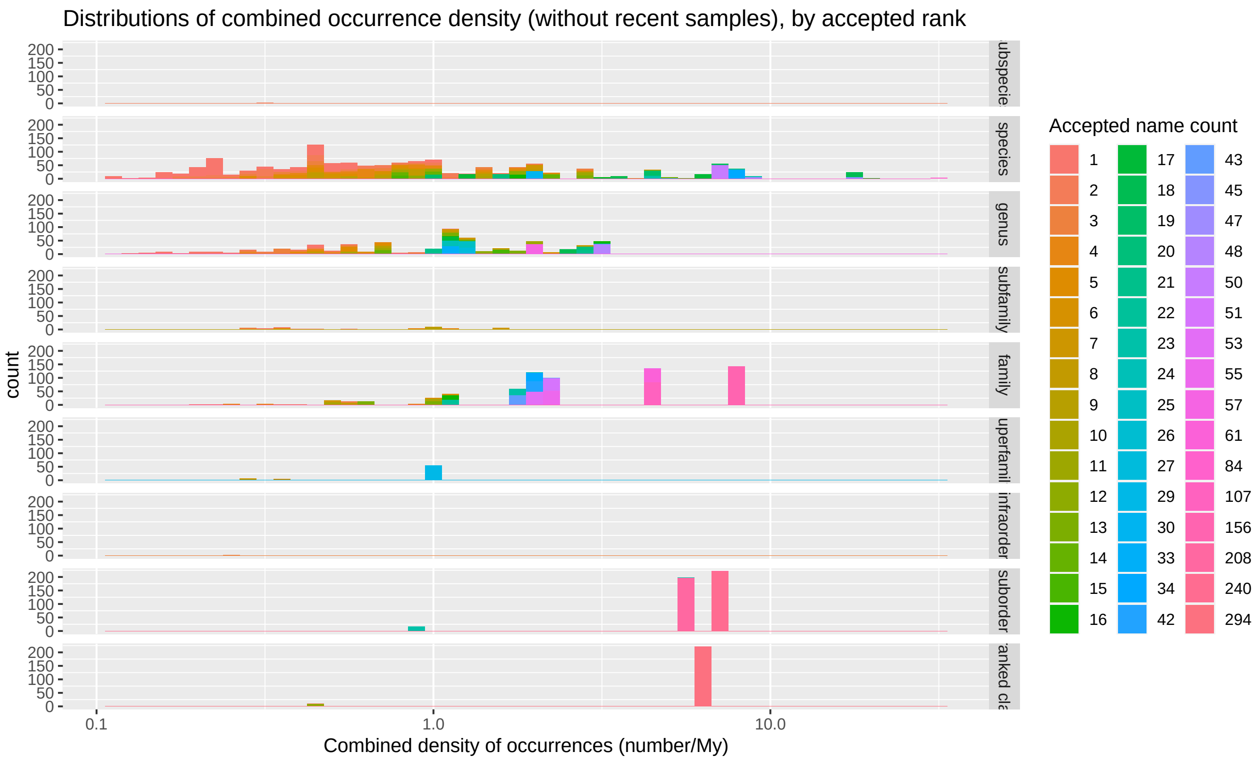
→ Again the combined version is less correlated with time.

Sub-sampling of occurrences with a normalized density along the combined ranges

Compare densities for single vs. combined ranges.





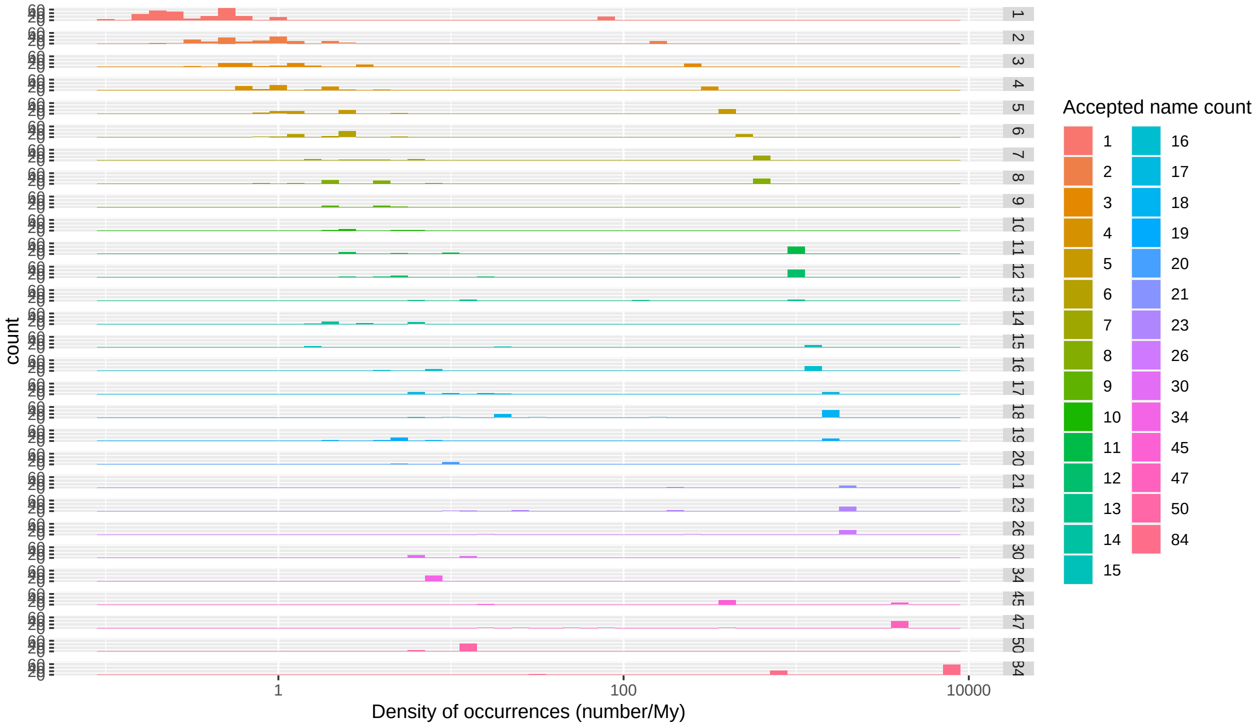


→ Densities are smaller and more concentrated with the combined ranges (larger time span + less ranges in total because of combination).

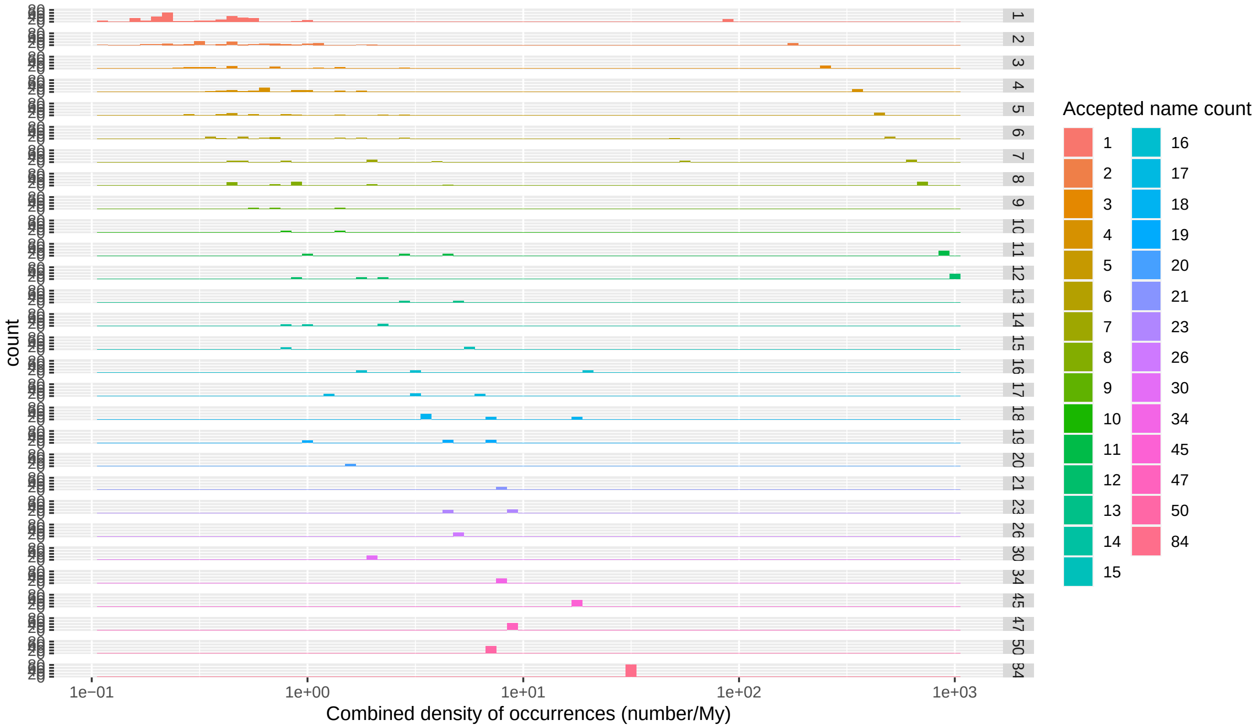
Compare densities by accepted name count (species only)

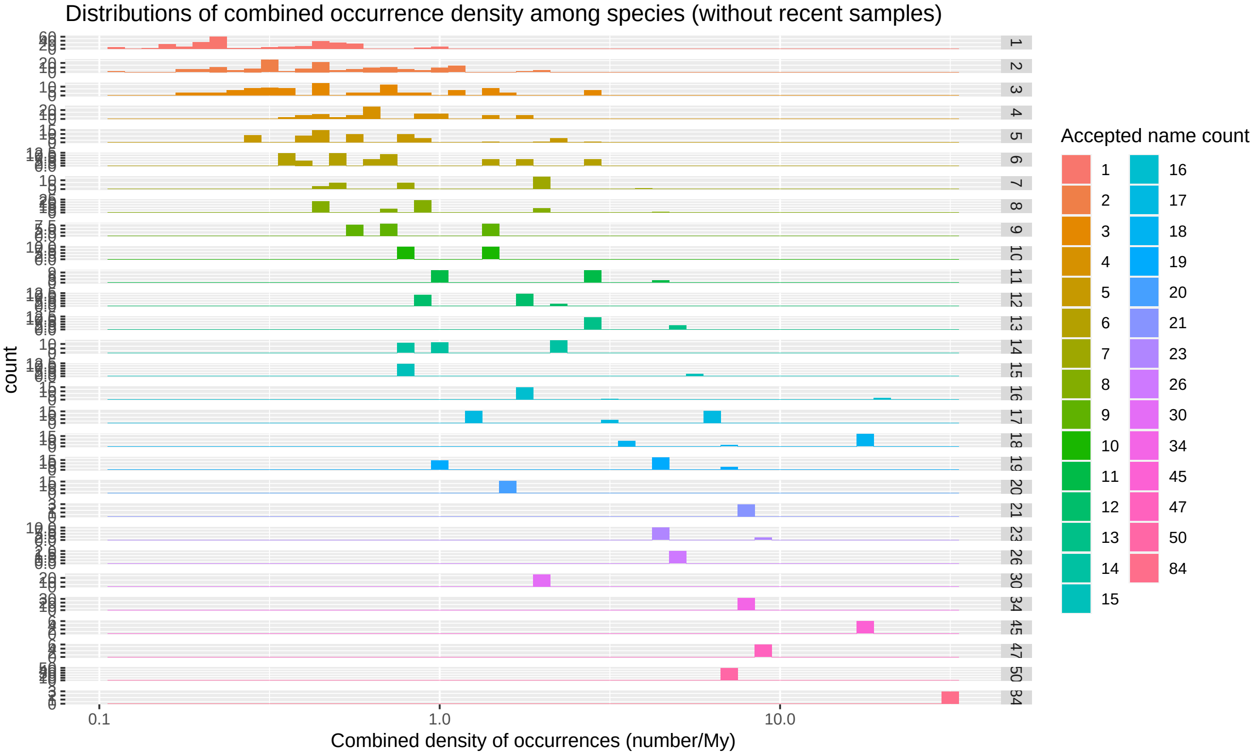
Let's focus now on the occurrences accepted at the species level because they are the one for which we can correct the abundance bias by subsampling the most concentrated combined intervals.

Distributions of occurrence density among species



Distributions of combined occurrence density among species

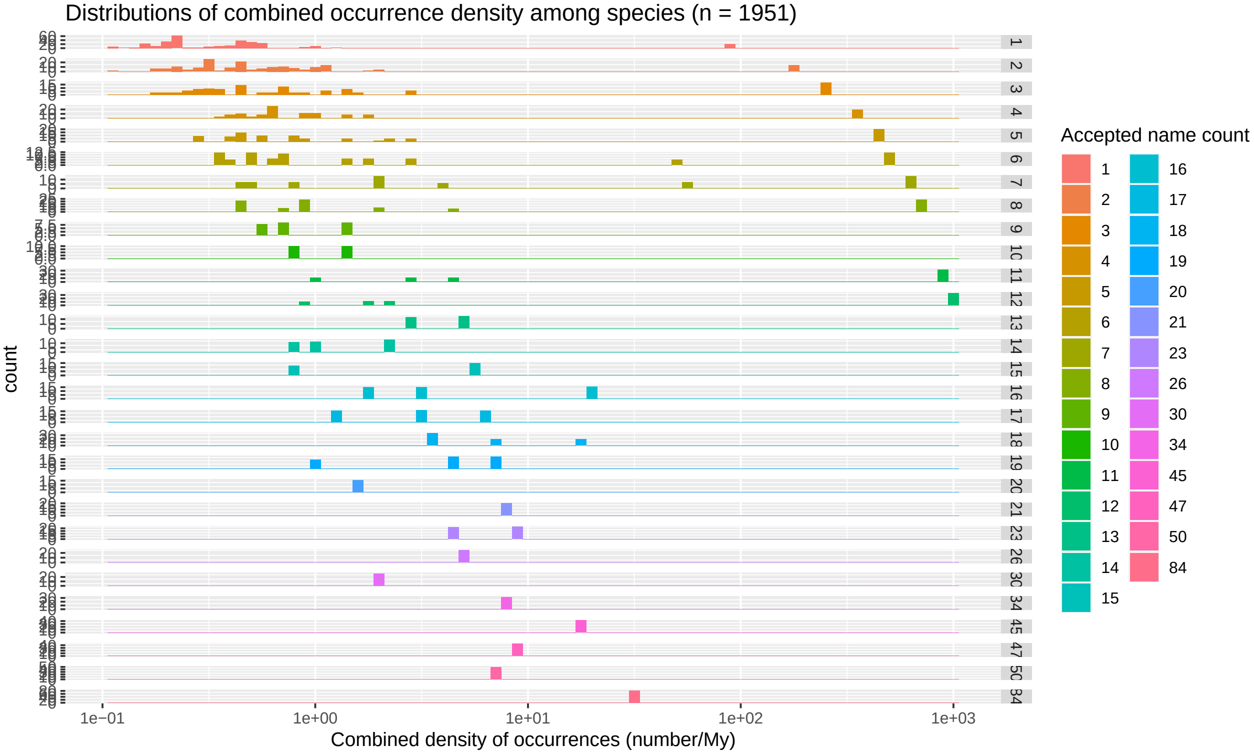




Distributions of combined occurrence density among species (without recent samples)

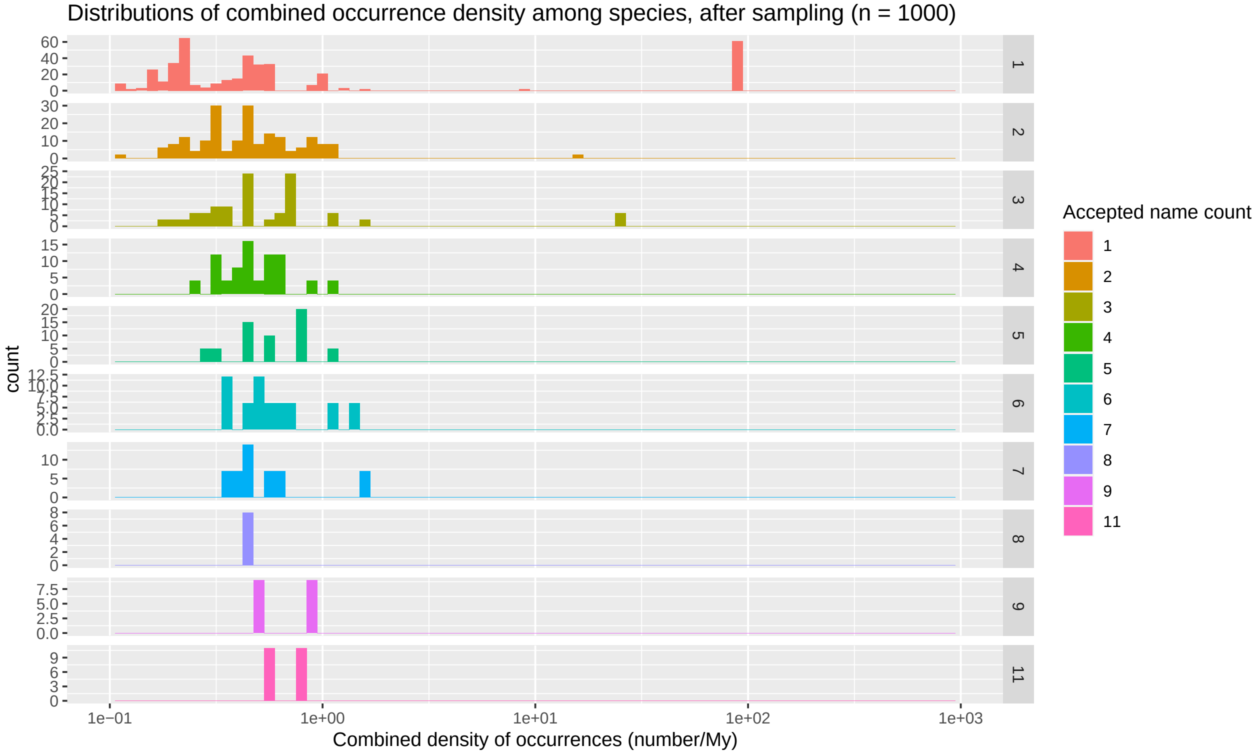
→ Their is a huge span of densities driven by the number of occurrences for the same species that we can reduce by subsampling the most concentrated intervals.

Impact of correcting subsampling on density distributions (species only)

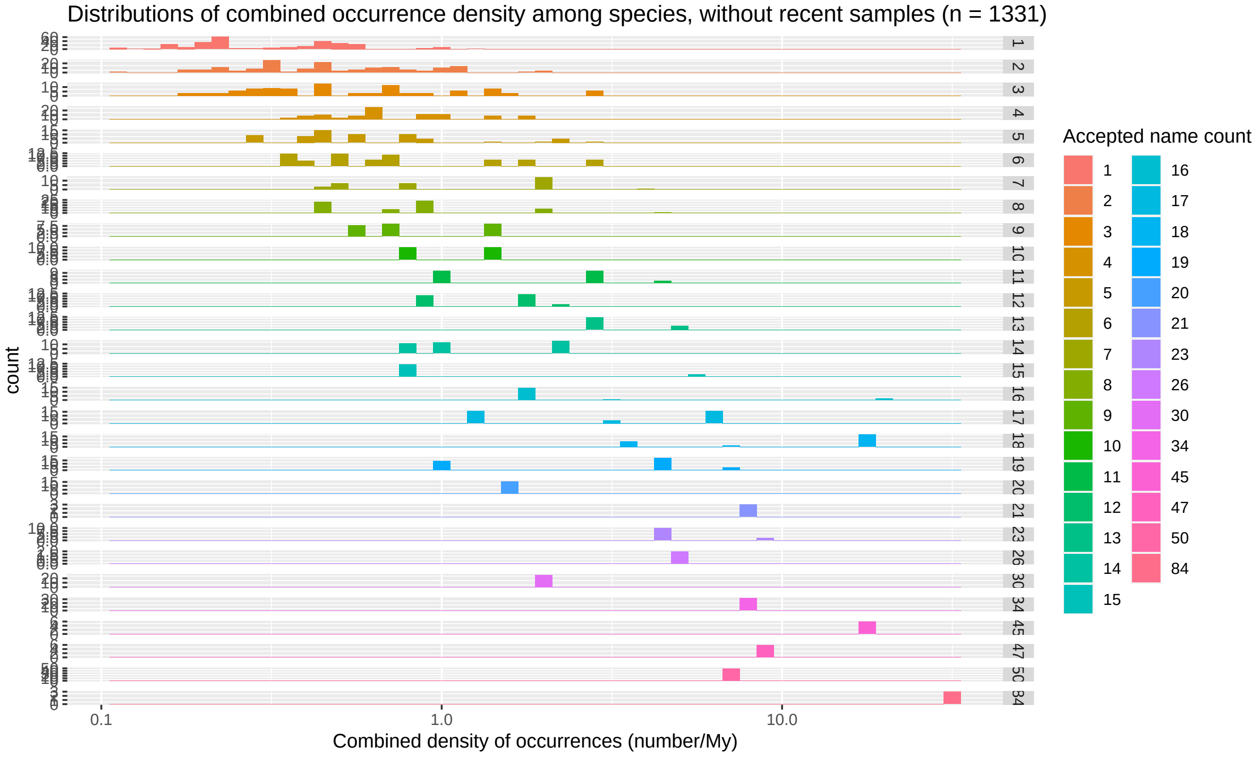


Distributions of combined occurrence density among species (n = 1951)

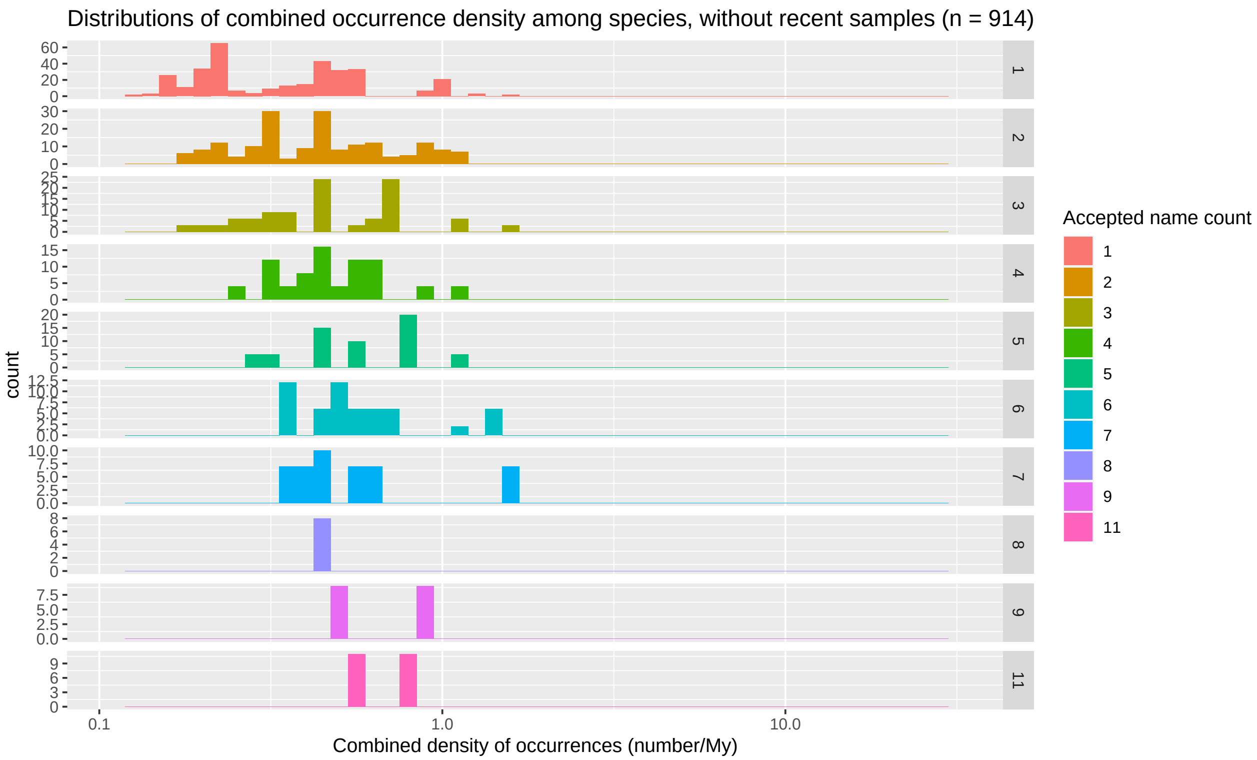
```
## Warning: Removed 20 rows containing missing values (geom_bar).
```



→ Subsampling successfully reduces the density span from 2 to 1 order of magnitude, apart from the artefactual recent samples that we can hide :

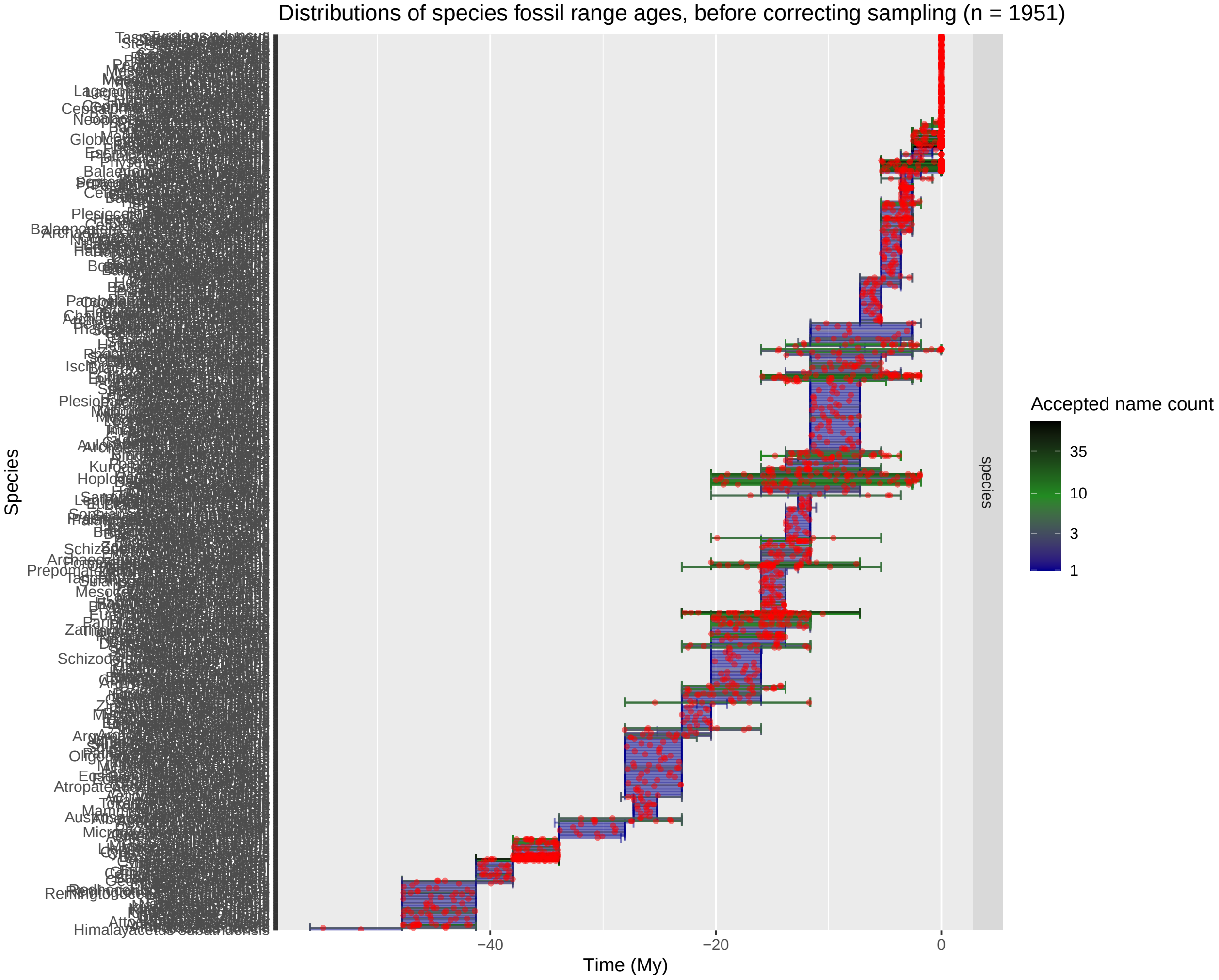


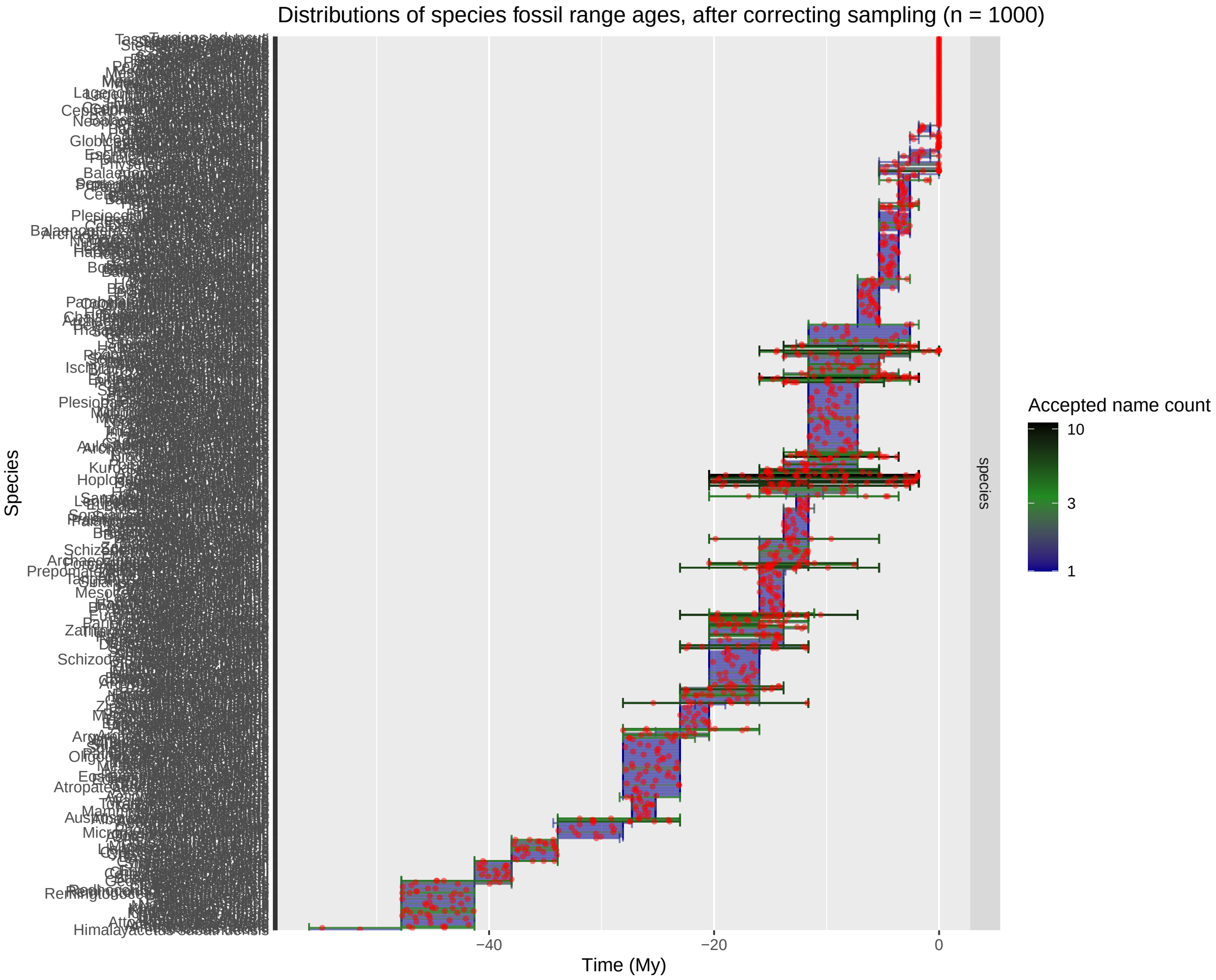
```
## Warning: Removed 20 rows containing missing values (geom_bar).
```


Distributions of combined occurrence density among species, without recent samples (n = 914)

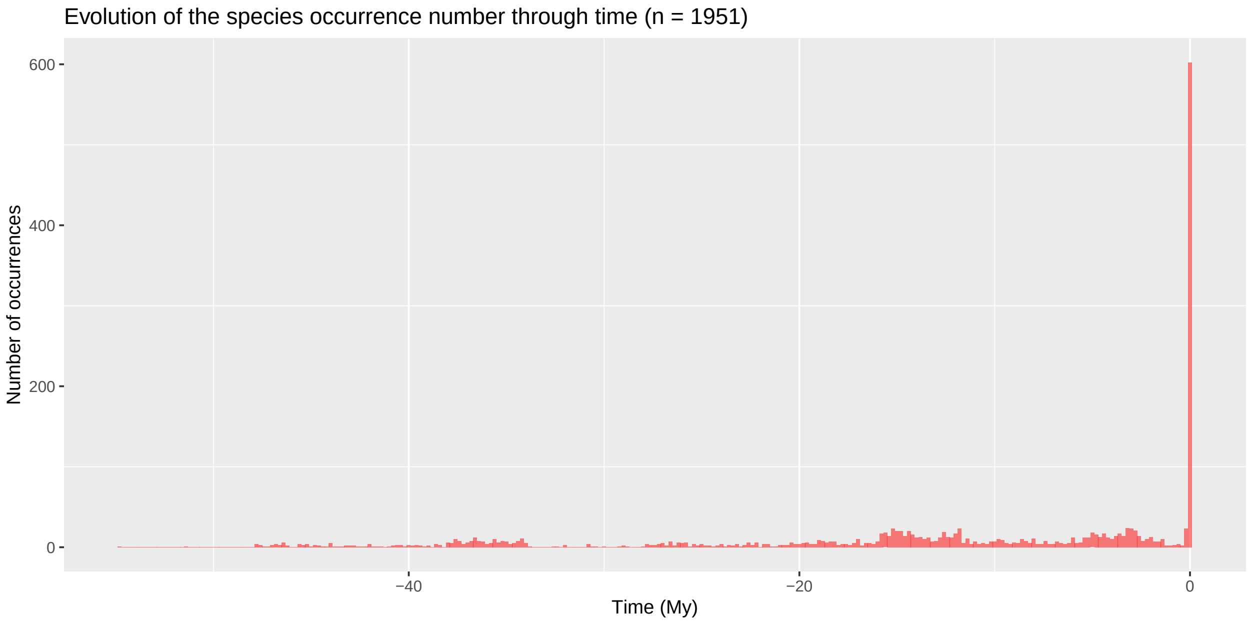
Impact of subsampling on occurrences repartition (species only)

See what our distributions

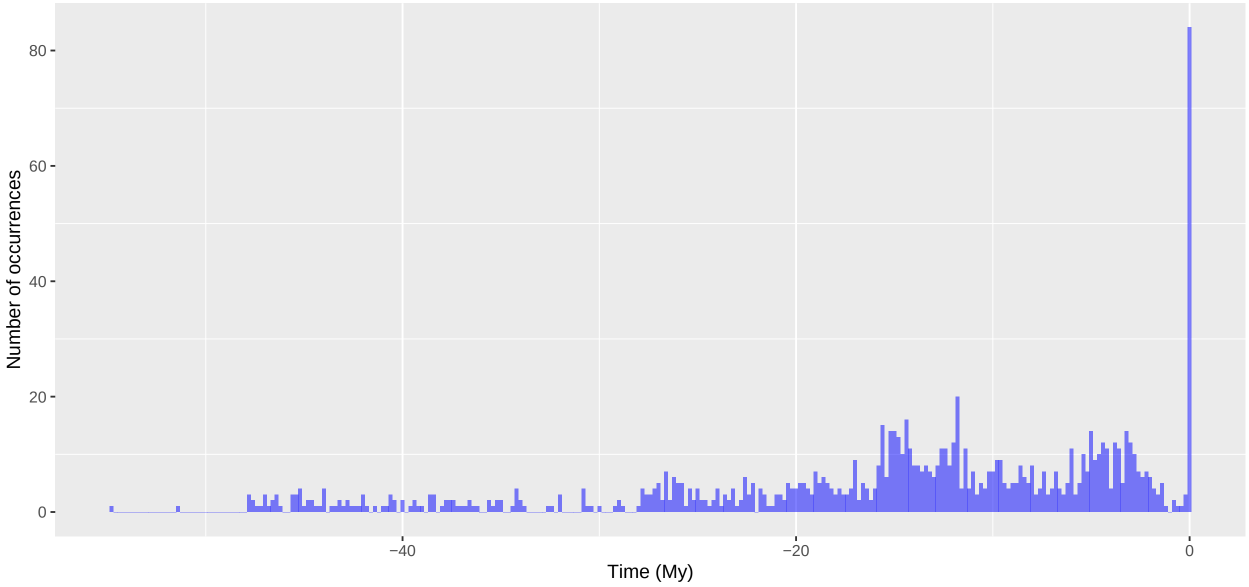

Distributions of species fossil range ages, before correcting sampling (n = 1951)
Species
Time (My)
−40
−20
0
species
Accepted name count
35
10
3
1
Himalayacetus subathuensis



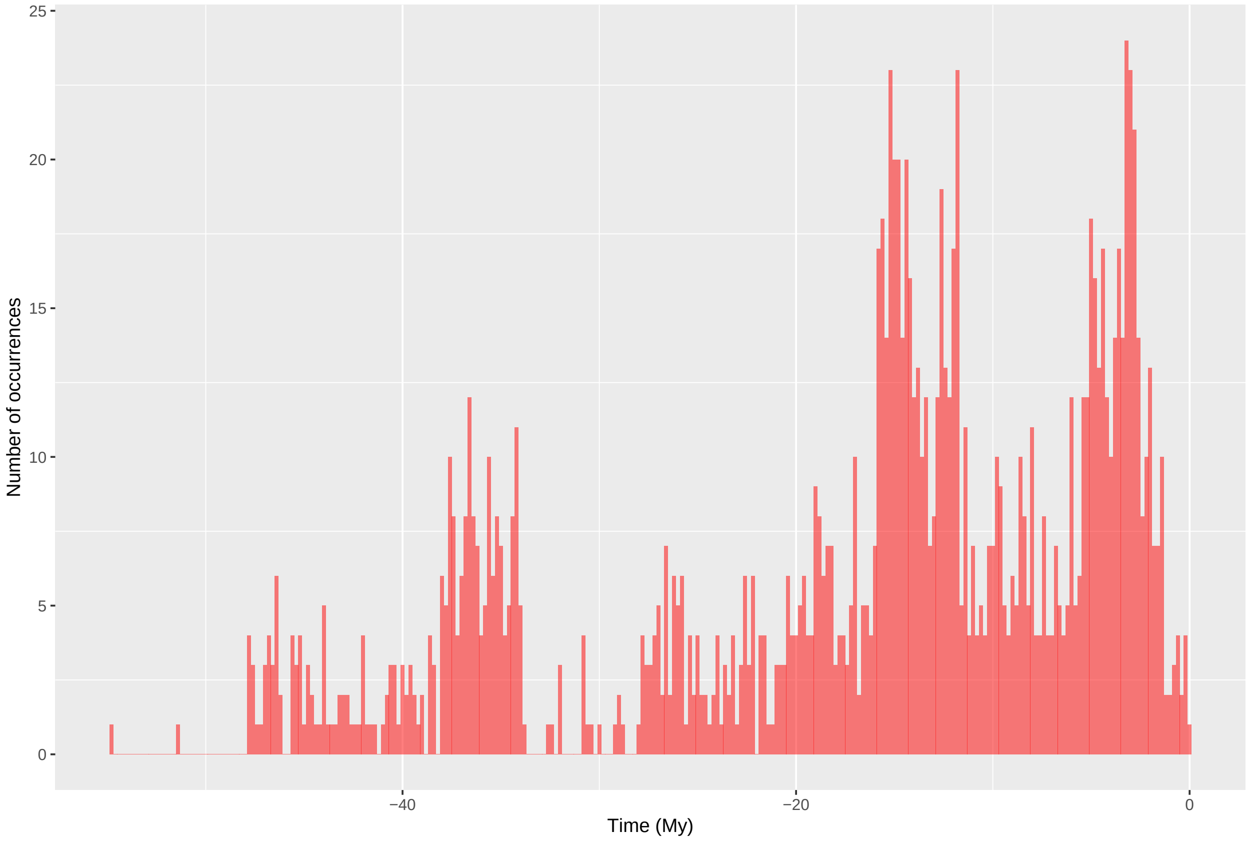
→ Some highly dense cluster became much more similar to the others.

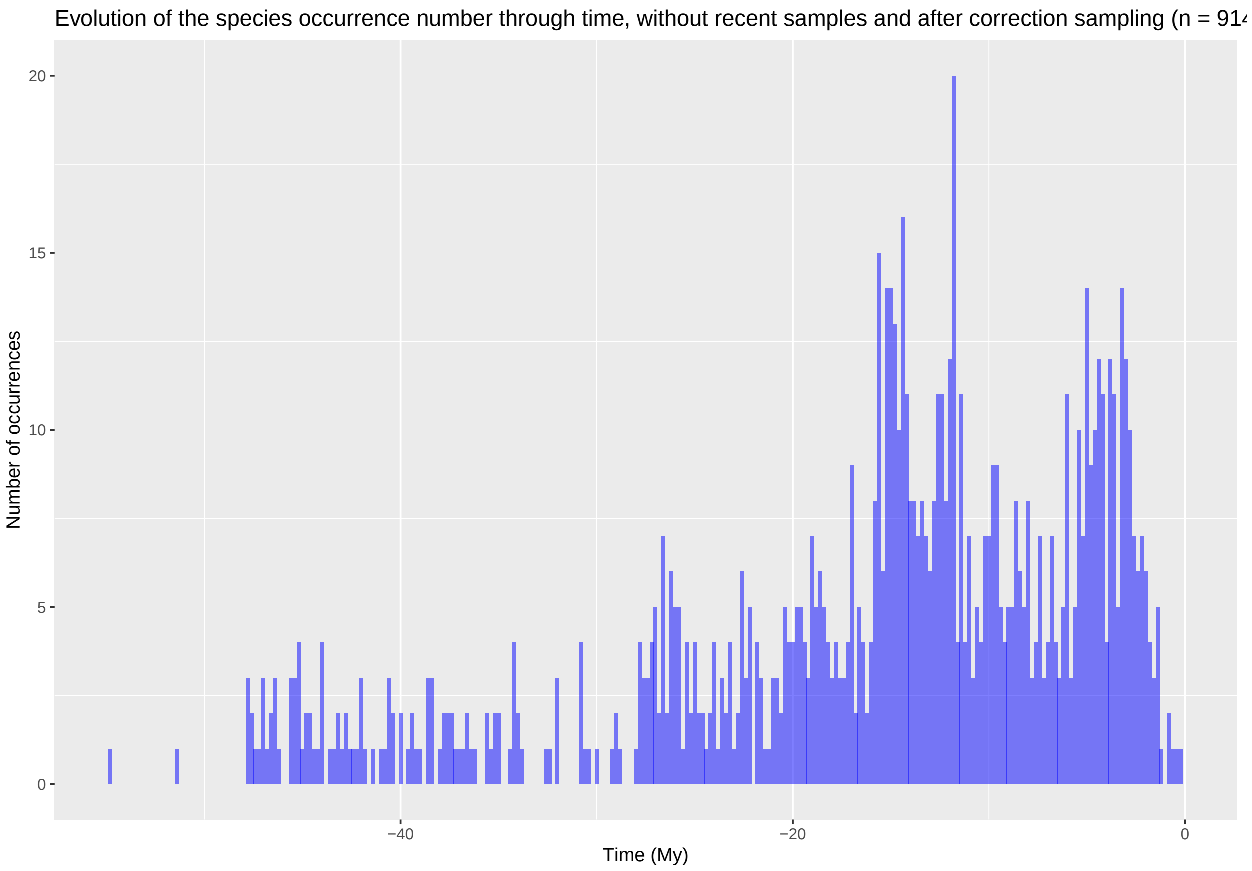


Evolution of the species occurrence number through time, after correction sampling (n = 1000)

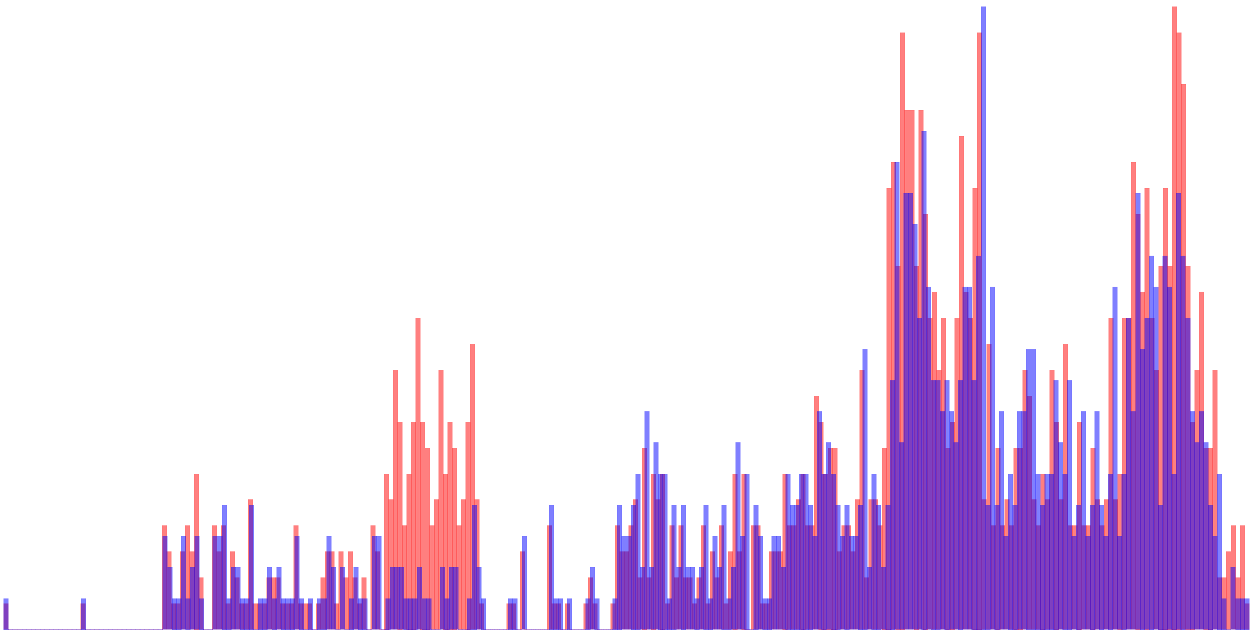


Evolution of the species occurrence number through time, without recent samples (n = 1331)




Evolution of the species occurrence number through time, without recent samples and after correction sampling (n = 914

If we superpose these 2 plots :



→ We get the new species occurrence repartition after subsampling correction, that could be used for doing inference with the occurrence birth-death model.

Conclusions

Achievements :

- It seems possible to adequately reduce the abundance bias by subsampling the most concentrated intervals → **species only**
- Using combined ranges by species appears to be more robust → **to be confirmed**
- Very recent samples may have been dated with a more precise method and contain much more fossils, so they should be removed or treated separatadely → **additional information needed**

Open questions :

- What about other accepted ranks ?
  1. The problem is that differences in the number of occurrences at higher ranks could be due to differences in individual abundances inside species or due to differences in the number of species inside that group.
  2. A solution could be to look at the number of species by group based on the indicated species, and include it in the bias correction : homogeneizing the *number of occurrences / time unit / number sf species in the group* → **additional data required** (ranks classification)
- Why do most occurrences miss a late stratigraphic limit ?
- Some occurrences have very huge time intervals → **Was is a good idea to remove those >10My ? Should we remove more of them (>5My) ?**