

Format Cetaceans Data

Contents

Load occurrence dataset	1
Load the list of occurrences with morphological information	2
Explore the dataset	4
Repartition through time	4
Repartition among accepted ranks	8
Time intervals = stratigraphic age uncertainty (minimum and maximum ages)	11
Time ranges = duration of the time intervals (maximum age - minimum age)	11
Combined time ranges = unique time range for occurrences with the same name (without the biggest ones)	22
Occurrence density = number of occurrences by unit of time, in the stratigraphic interval of a taxon	24
Correlation between time range and age	25
Subsampling occurrences by homogenizing the combined occurrence density	26
Compare with a Poisson sampling process	29
Subsampling occurrences by aggregating similarly identified occurrences in each collection	32
Aggregate similarly identified occurrences in each collection	32
Aggregate similarly identified occurrences in each formation	33
Aggregate occurrences without formation by country + early interval	35
Aggregate occurrences without formation by geoplate + early interval	38
Check that the sampling methods do not introduce biases in the repartition between Odontoceti and Mysticeti	39
Faster genus-level analysis	44
Conclusions	46

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

This notebook aims at exploring the Cetacean occurrence dataset from the PaleoBiology DataBase (PBDB), looking for potential biases invalidating to our model assumptions and how to mitigate them. Subsampling this dataset will allow (1) to reduce the computational burden of our analysis and (2) to compensate for the biases mentioned above.

Load occurrence dataset

```
## [1] "occurrence_no"      "record_type"        "reid_no"
## [4] "collection_no"      "identified_name"     "identified_rank"
## [7] "identified_no"      "difference"          "accepted_name"
## [10] "accepted_rank"      "accepted_no"         "early_interval"
## [13] "late_interval"      "max_ma"              "min_ma"
## [16] "ref_author"         "ref_pubyr"           "reference_no"
## [19] "phylum"           "phylum_no"         "class"
## [22] "class_no"           "order"               "order_no"
```

```

## [25] "family"          "family_no"      "genus"
## [28] "genus_no"        "abund_value"    "abund_unit"
## [31] "lng"             "lat"            "occurrence_comments"
## [34] "collection_name" "collection_subset" "collection_aka"
## [37] "cc"              "state"          "county"
## [40] "latlng_basis"    "latlng_precision" "geogscale"
## [43] "geogcomments"    "paleomodel"     "paleolng"
## [46] "paleolat"        "geoplate"       "cc.1"
## [49] "protected"       "formation"      "stratgroup"
## [52] "member"          "stratscale"     "zone"
## [55] "localsection"    "localbed"       "localorder"
## [58] "regionalsection" "regionalbed"    "regionalorder"
## [61] "stratcomments"   "lithdescript"   "lithology1"
## [64] "lithadj1"        "lithification1" "minor_lithology1"
## [67] "fossilsfrom1"    "lithology2"     "lithadj2"
## [70] "lithification2"  "minor_lithology2" "fossilsfrom2"
## [73] "environment"     "tectonic_setting" "geology_comments"
## [76] "assembl_comps"   "articulated_parts" "associated_parts"
## [79] "common_body_parts" "rare_body_parts" "feed_pred_traces"
## [82] "artifacts"       "component_comments" "pres_mode"
## [85] "preservation_quality" "spatial_resolution" "temporal_resolution"
## [88] "lagerstatten"    "concentration"   "orientation"
## [91] "abund_in_sediment" "fragmentation"   "bioerosion"
## [94] "encrustation"    "preservation_comments" "collection_type"
## [97] "collection_methods" "museum"          "collection_coverage"
## [100] "collection_size"  "rock_censused"   "collectors"
## [103] "collection_dates" "collection_comments" "taxonomy_comments"
## [106] "taxon_environment" "motility"        "life_habit"
## [109] "diet"            "reproduction"    "ontogeny"
## [112] "composition"

```

Reorder accepted ranks according to classification standard.

```

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

```

Load the list of occurrences with morphological information

These are the specimens that are already included in the tree thanks to their morphological characters, and must therefore be removed from the occurrence dataset.

```

##           Taxon      composite..n.y.      occ.data.based.on
## Aetiocetus_cotylalveus : 1   Min.      :0.0000      :60
## Agorophius_pygmaeus    : 1   1st Qu.:0.0000   CMM-V-15 : 1
## Albertocetus_meffordorum: 1   Median :0.0000   MGGC 8548 : 1
## Albireo_whistleri      : 1   Mean    :0.3768   MNHN SAS 933: 1
## Archaeodelphis_patrius : 1   3rd Qu.:1.0000   UCMP 83790 : 1
## Ashleycetus_planicapitis: 1   Max.     :1.0000   USNM 10484 : 1
## (Other)                :63                (Other)   : 4
##           Specimen  pbdb_specimen_no pbdb_occurence.number
## AMNH 9485 : 1   Min.      : 25492   Min.      : 68135
## CASG 66660 : 1   1st Qu.: 25706   1st Qu.: 461107
## CCNHM-101 : 1   Median   : 25922   Median   : 487310

```

```
## ChM PV4256 : 1 Mean : 35167 Mean : 631596
## ChM PV4844 : 1 3rd Qu.: 26035 3rd Qu.: 763082
## CHM_PV_4253: 1 Max. :146940 Max. :1360382
## (Other) :63 NA's :56
##
##
## actually this info relates to the holotype, which is MGGC 8608, but it seems that MGGC 8599 has not
## It is unclear which specimen is in the pbdb, both are from the Pisco Formation. This one is the ref
## two specimen numbers in pbdb, but both relate to same specimen
## two specimens in database, but one comes from USA, so I selected the one from Peru based on collect
## two specimens in pbdb, unclear which is which because they belong to the same horizon
## (Other)
```

Combine those taxa with extant taxa to get all the species included in the tree.

```
##          taxon      min      max
## Balaena_mysticetus : 1 Min. :0 Min. :0
## Balaenoptera_acutorostrata: 1 1st Qu.:0 1st Qu.:0
## Balaenoptera_bonaerensis : 1 Median :0 Median :0
## Balaenoptera_borealis : 1 Mean :0 Mean :0
## Balaenoptera_brydei : 1 3rd Qu.:0 3rd Qu.:0
## Balaenoptera_edeni : 1 Max. :0 Max. :0
## (Other) :82

##          taxon      min      max
## Aetiocetus_cotylalveus : 1 Min. : 2.588 Min. : 3.60
## Agorophius_pygmaeus : 1 1st Qu.: 7.246 1st Qu.:11.62
## Albertocetus_meffordorum: 1 Median :13.820 Median :15.97
## Albireo_whistleri : 1 Mean :14.357 Mean :18.28
## Archaeodelphis_patrius : 1 3rd Qu.:23.030 3rd Qu.:28.10
## Ashleycetetus_planicapitis: 1 Max. :41.300 Max. :47.80
## (Other) :63

##          taxon      min      max
## Balaena_mysticetus : 1 Min. : 0.00 Min. : 0.000
## Balaenoptera_acutorostrata: 1 1st Qu.: 0.00 1st Qu.: 0.000
## Balaenoptera_bonaerensis : 1 Median : 0.00 Median : 0.000
## Balaenoptera_borealis : 1 Mean : 6.31 Mean : 8.035
## Balaenoptera_brydei : 1 3rd Qu.:11.62 3rd Qu.:13.820
## Balaenoptera_edeni : 1 Max. :41.30 Max. :47.800
## (Other) :151
```

Check that the names in the datasets are exactly the same as in the list above.

```
## [1] TRUE
```

Idem at the generic level for the genus analysis.

```
##          taxon      min      max
## Balaena_mysticetus : 1 Min. :0 Min. :0
## Balaenoptera_physalus : 1 1st Qu.:0 1st Qu.:0
## Berardius_bairdii : 1 Median :0 Median :0
## Caperea_marginata : 1 Mean :0 Mean :0
## Cephalorhynchus_heavisidii: 1 3rd Qu.:0 3rd Qu.:0
## Delphinapterus_leucas : 1 Max. :0 Max. :0
## (Other) :35

##          taxon      min      max
```

```
## Aetiocetus_cotylalveus : 1 Min. : 2.588 Min. : 3.60
## Agorophius_pygmaeus : 1 1st Qu.: 7.246 1st Qu.:11.62
## Albertocetus_meffordorum: 1 Median :13.820 Median :15.97
## Albireo_whistleri : 1 Mean :14.960 Mean :18.95
## Archaeodelphis_patrius : 1 3rd Qu.:23.030 3rd Qu.:28.10
## Ashleycetus_planicapitis: 1 Max. :41.300 Max. :47.80
## (Other) :56

##          taxon      min      max
## Balaena_mysticetus : 1 Min. : 0.000 Min. : 0.000
## Balaenoptera_physalus : 1 1st Qu.: 0.000 1st Qu.: 0.000
## Berardius_bairdii : 1 Median : 5.333 Median : 7.246
## Caperea_marginata : 1 Mean : 9.005 Mean :11.409
## Cephalorhynchus_heavisidii: 1 3rd Qu.:15.970 3rd Qu.:20.440
## Delphinapterus_leucas : 1 Max. :41.300 Max. :47.800
## (Other) :97
```

Check that the names in the datasets are exactly the same as in the list above.

```
## [1] TRUE
```

Remove those occurrences from our initial dataset to avoid redundancy.

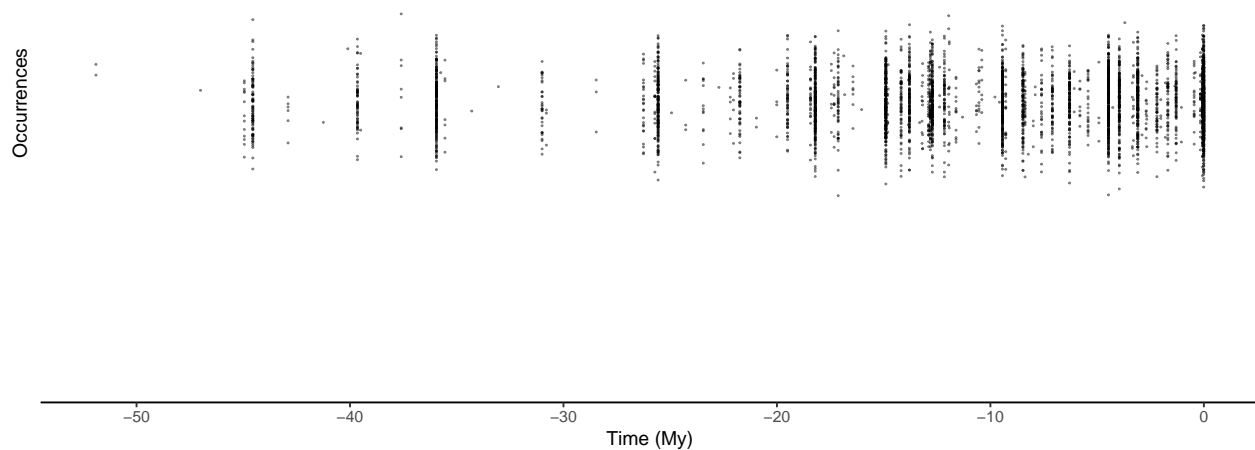
Explore the dataset

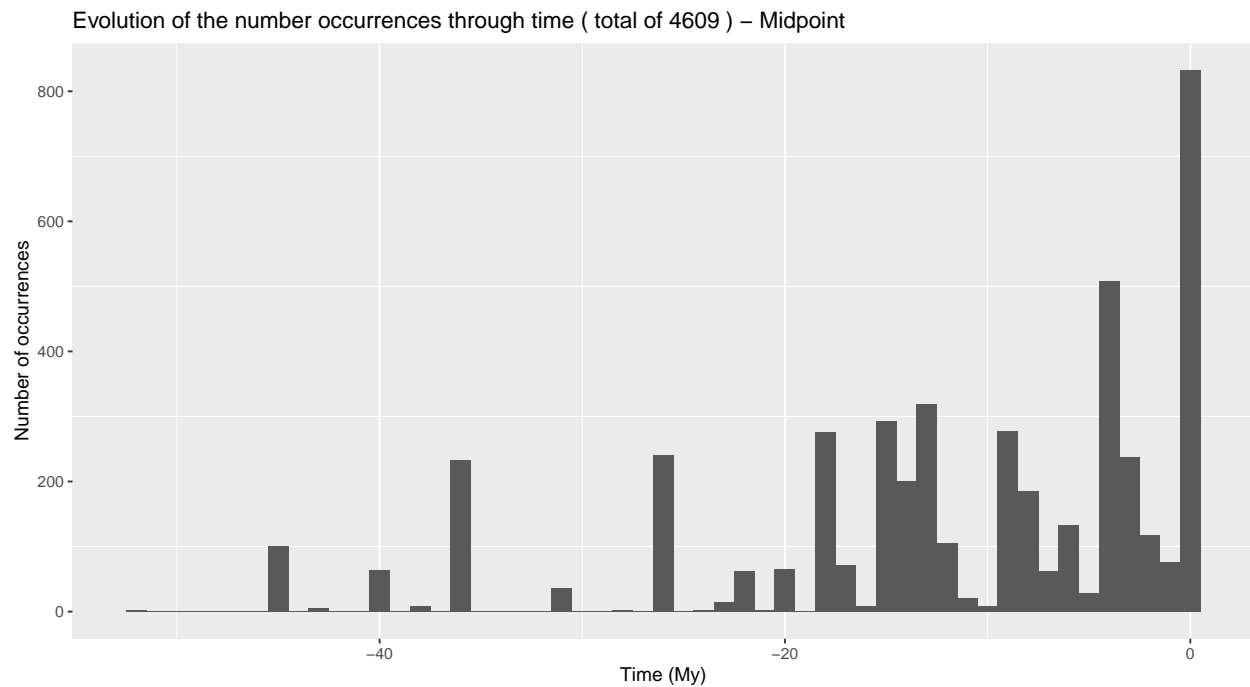
Repartition through time

Full fossil record

First, use the stratigraphic range's midpoint as the occurrence fixed age.

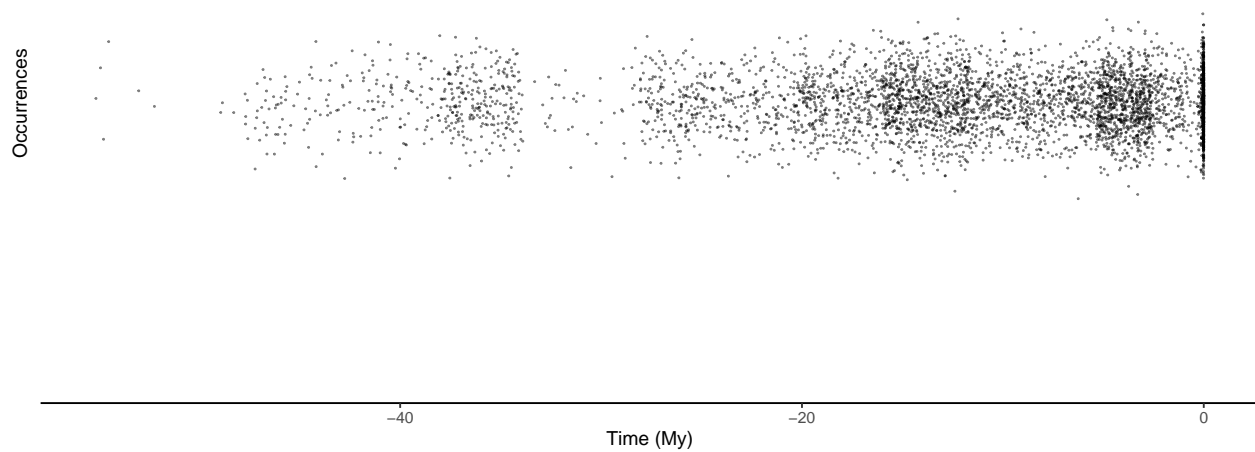
Repartition of 4609 recorded occurrences through time – Midpoint

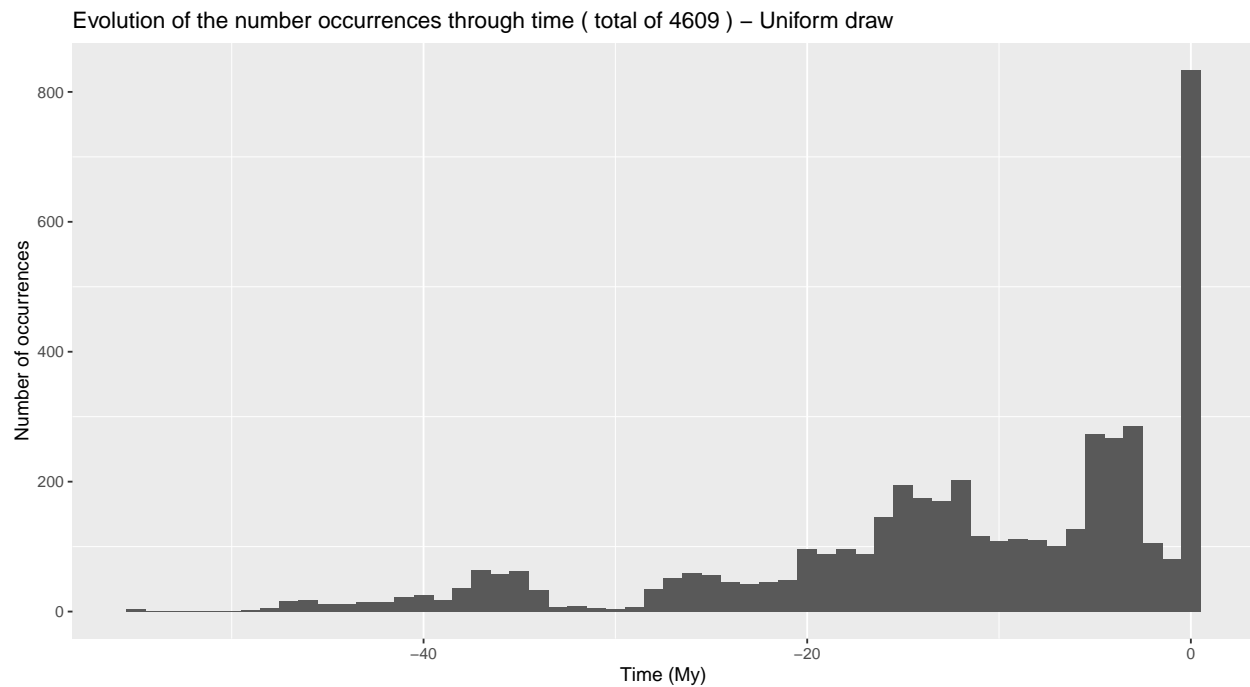




→ Numerous occurrences seem to have the same age interval so in order to avoid clusters let's draw them uniformly in their stratigraphic range rather than taking the midpoint.

Repartition of 4609 recorded occurrences through time – Uniform draw



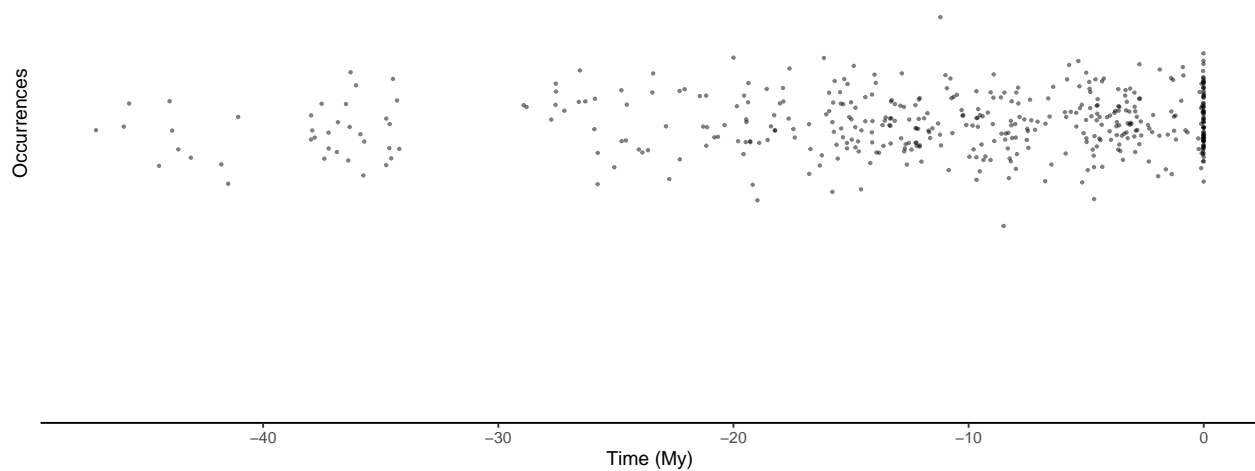


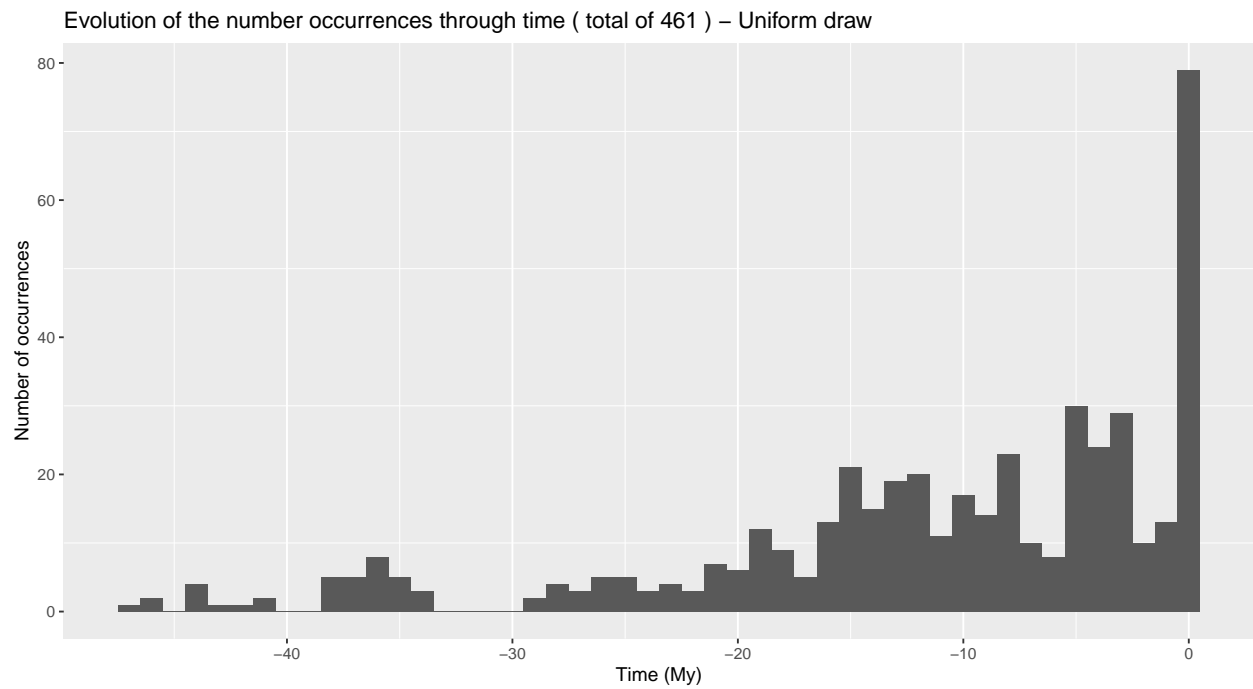
→ The repartition seems much smoother now.

Subsampling

These occurrences are too numerous for a faster preliminary analysis, let's randomly subsample a fraction of them.

Repartition of 461 recorded occurrences through time – Uniform draw



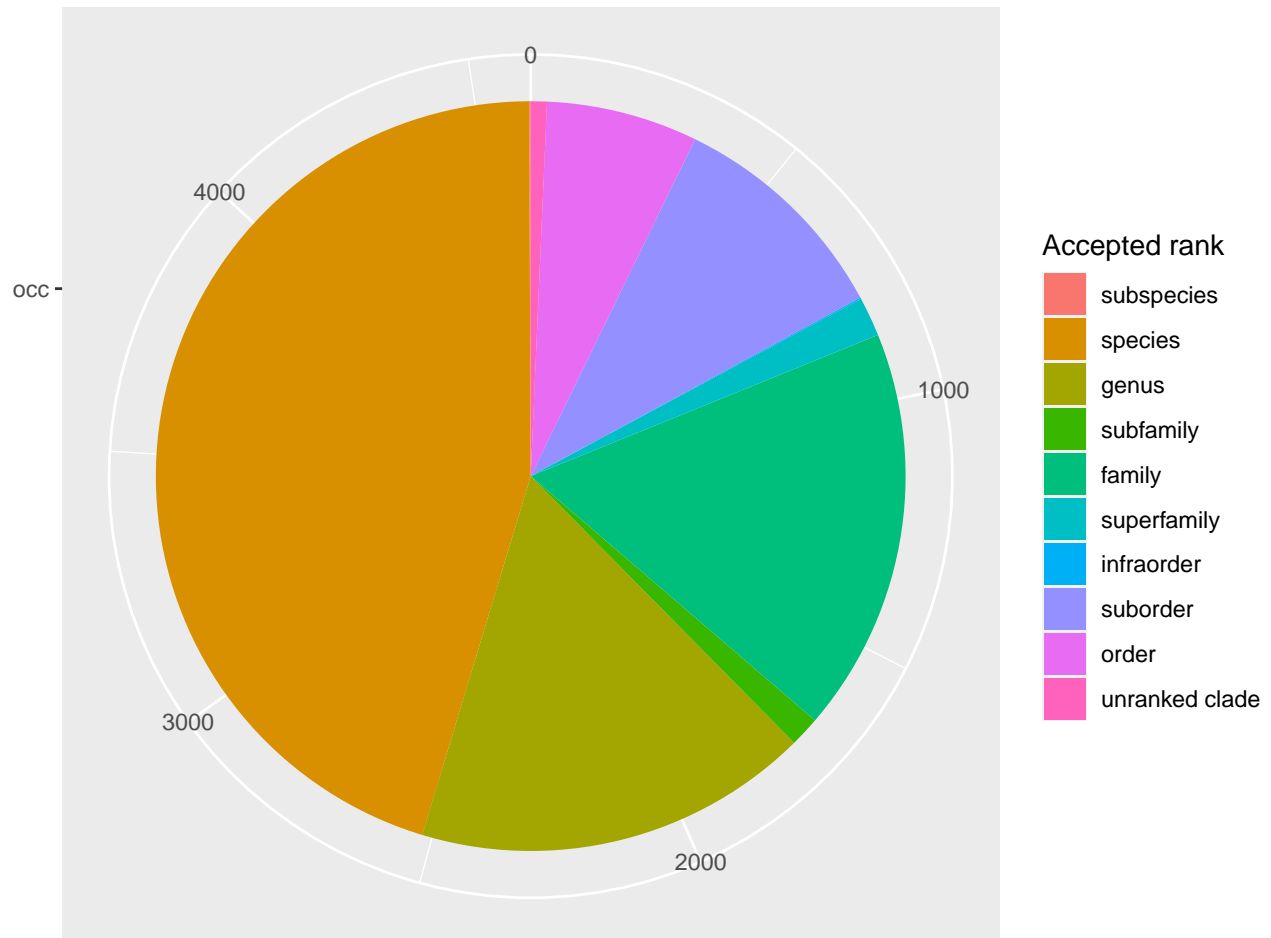


→ The distribution looks similar, with some noise due to higher variance with smaller sample.

Repartition among accepted ranks

Pie chart

Repartition of occurrences among accepted ranks

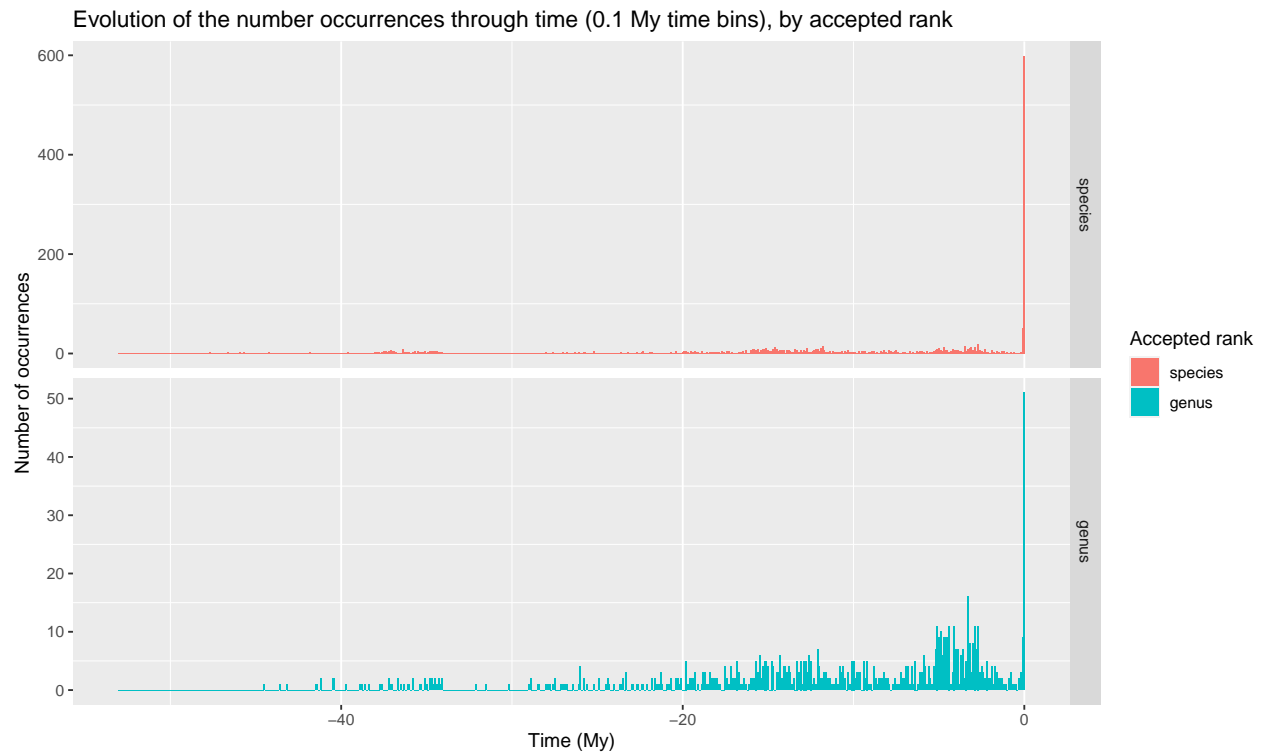


→ Half of the occurrences are identified at the level of the species and 1/3 at the genus or family.

Some clades are unranked :

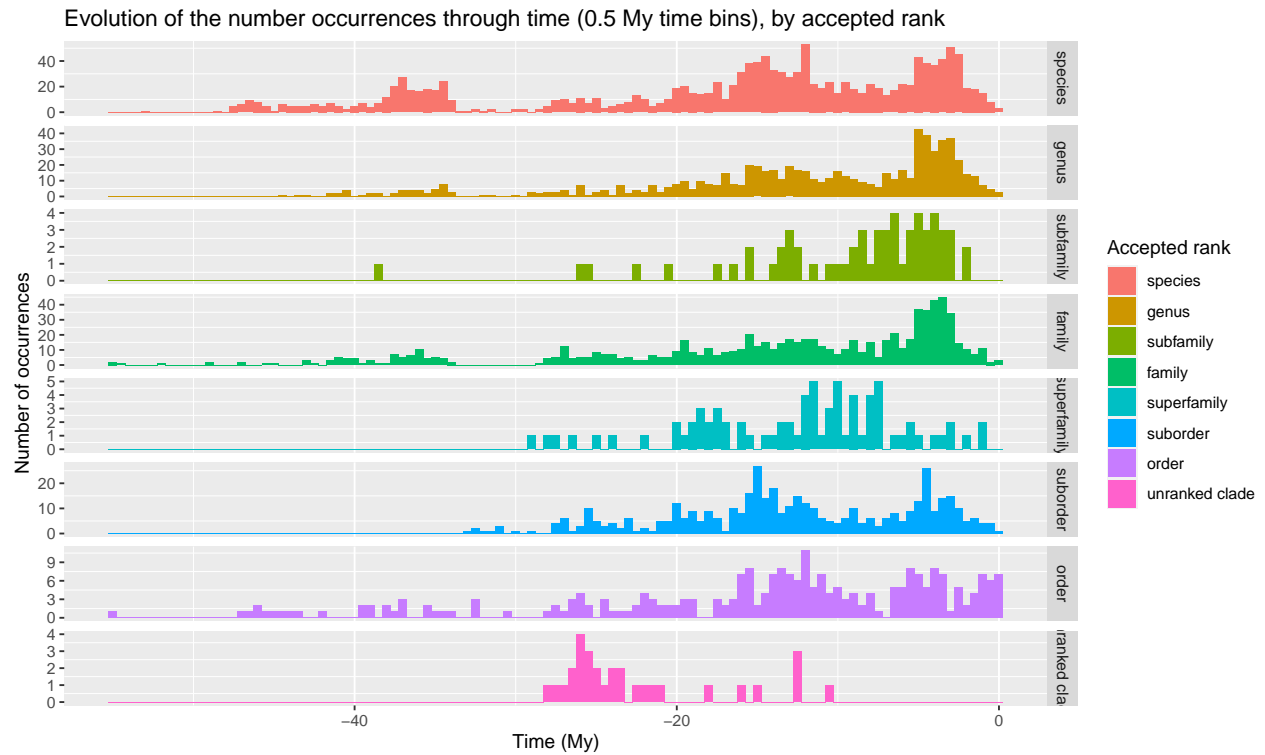
```
##
##   Chaemysticeti      Neoceti Panphyseteroidea  Pelagiceti
##           26           2           1           1
##   Platanidelphidi   Squaloceti
##           1           1
```


Time repartition by rank



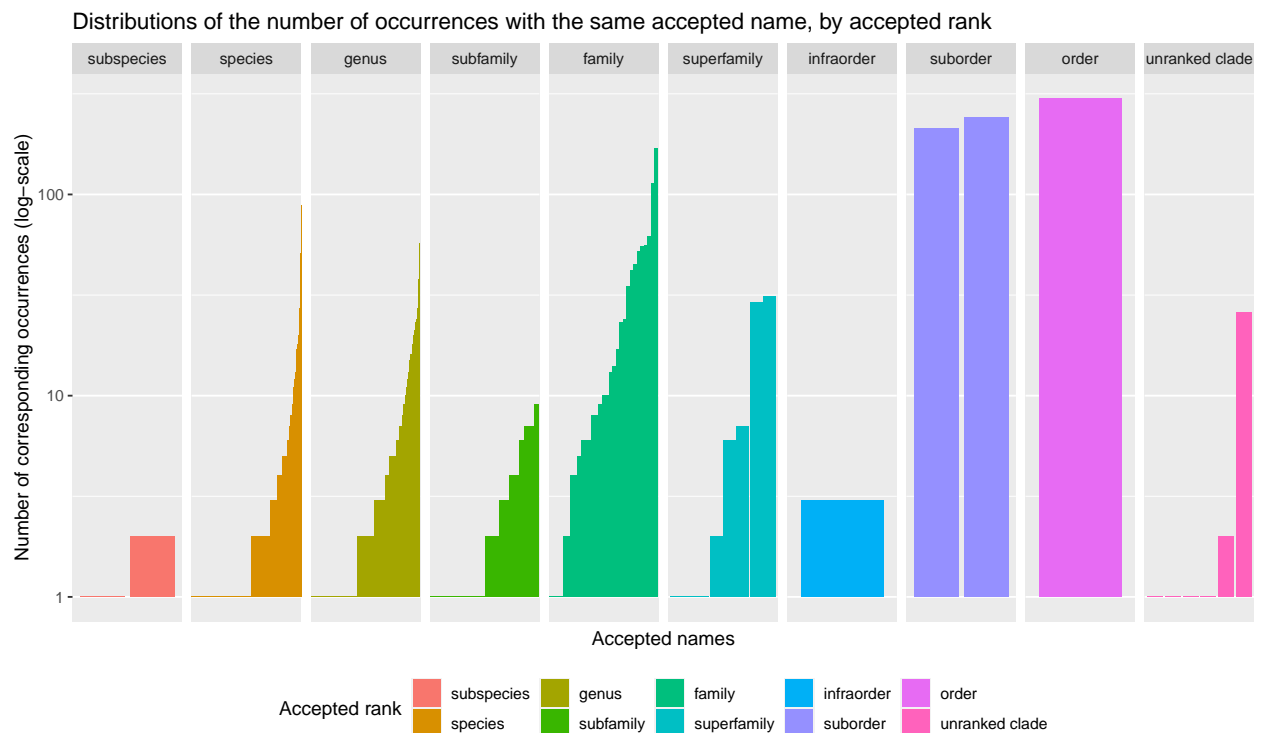
→ Apparent huge cluster of occurrences in recent times, with very precise dating = Artefact due to the “Pull of the Recent” effect ? → We decided to **remove all Late Pleistocene and Holocene occurrences** (thus setting the ω -sampling to 0) in order to avoid this bias.

```
##                               cetacea_pbdb Cetacea_occ removed
## Number of occurrences         4609         3807         802
```



→ We observe similar trends at each rank, with peaks at ~15My and ~5My.

Redundancy of occurrences with the same accepted name

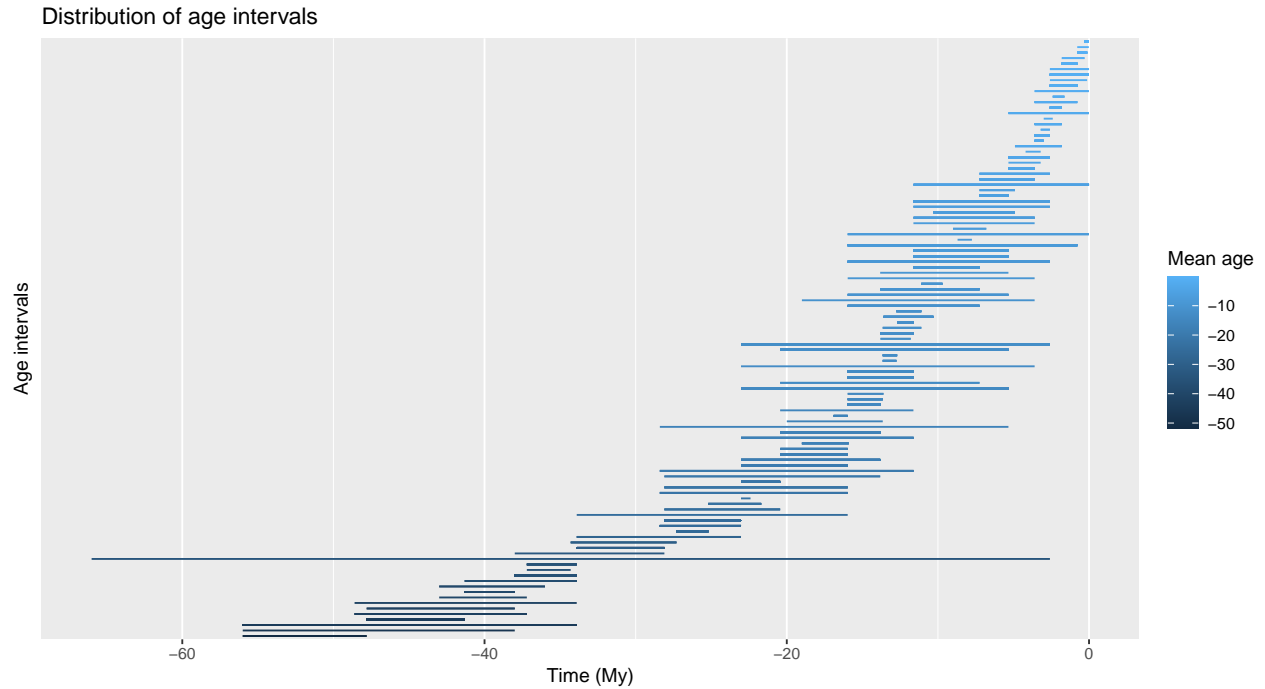


→ ~Half of species/genera/subfamilies have only one specimen by accepted name, but it could go up to ~50 within the same species and ~200 occurrences within the same suborder. **In our model all species**

are supposed to have the same abundance (identical sampling rates among branches), so those huge differences will have to be mitigated. See later for a quantification of this discrepancy.

⇒ Our goal now will be to correct this abundance bias.

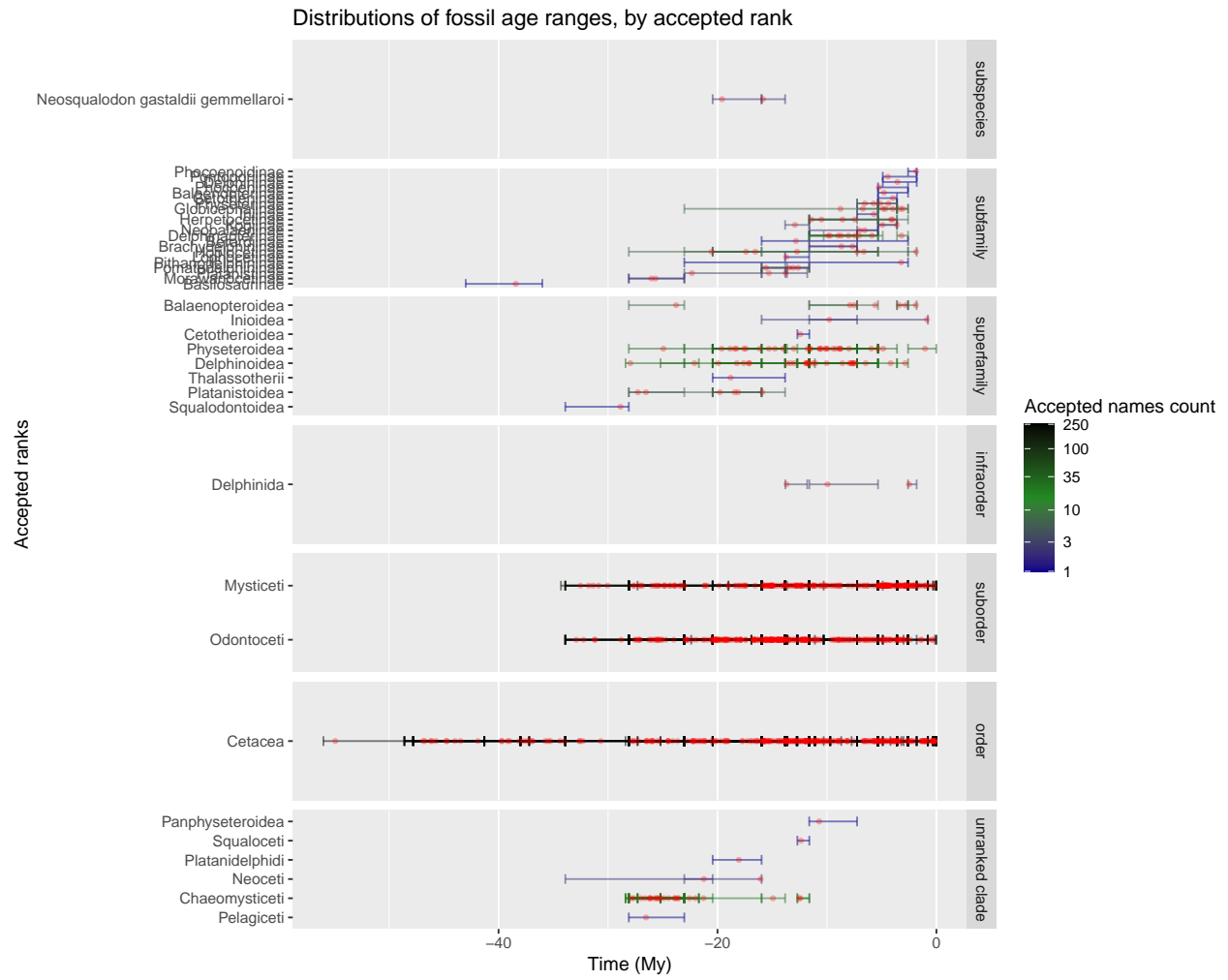
Time intervals = stratigraphic age uncertainty (minimum and maximum ages)



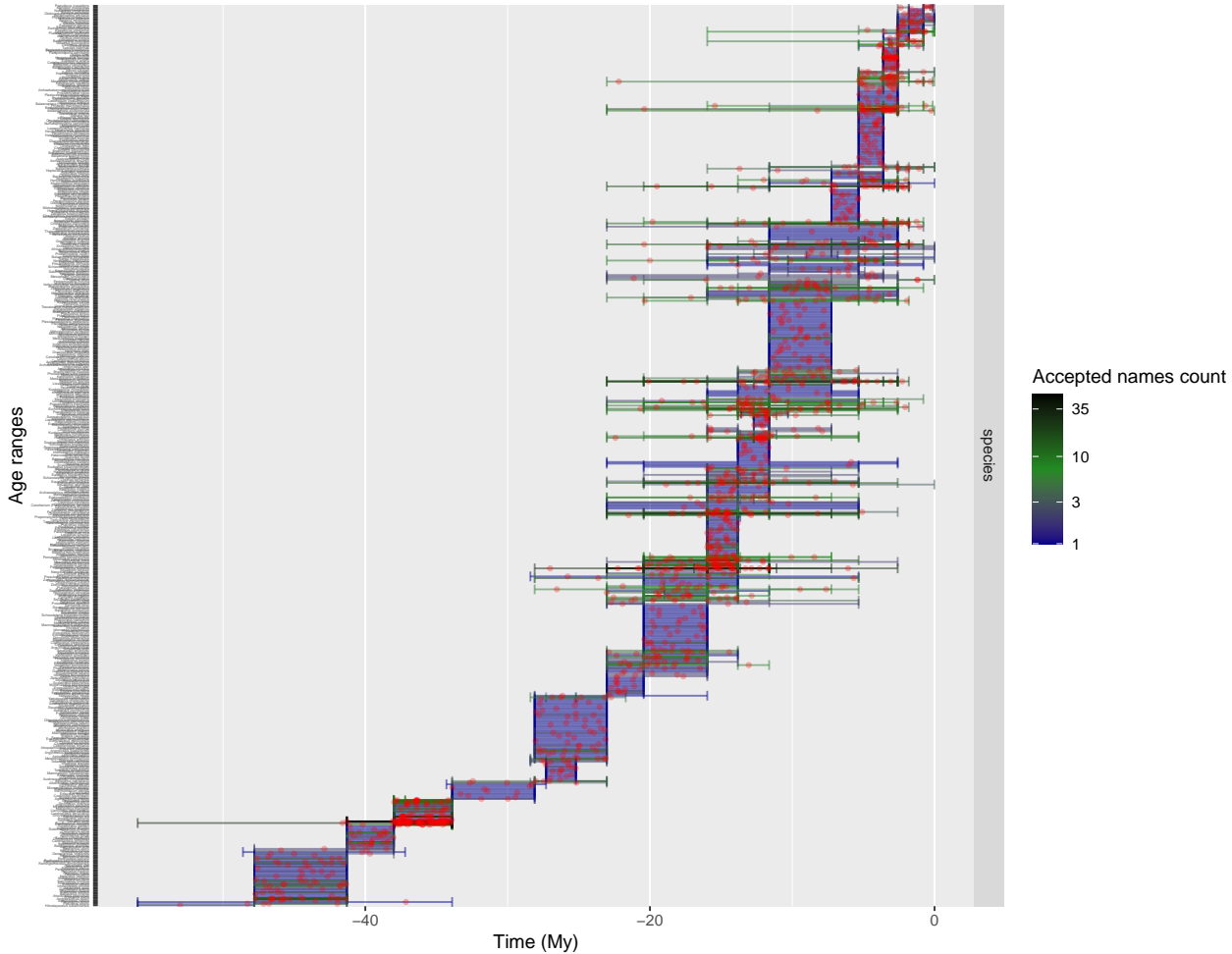
Time ranges = duration of the time intervals (maximum age - minimum age)

Count occurrences by accepted name

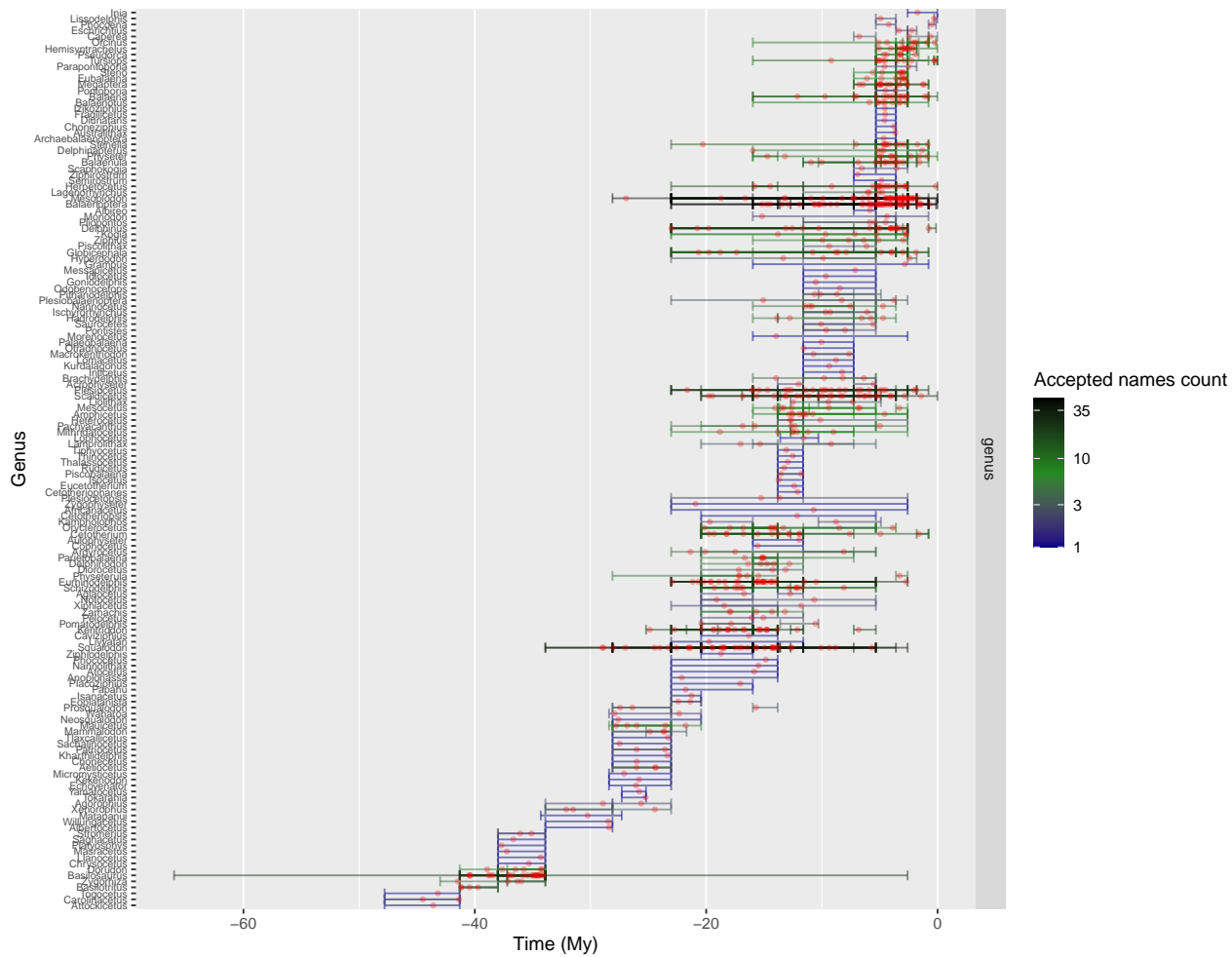
Count the number of occurrences with the same accepted name

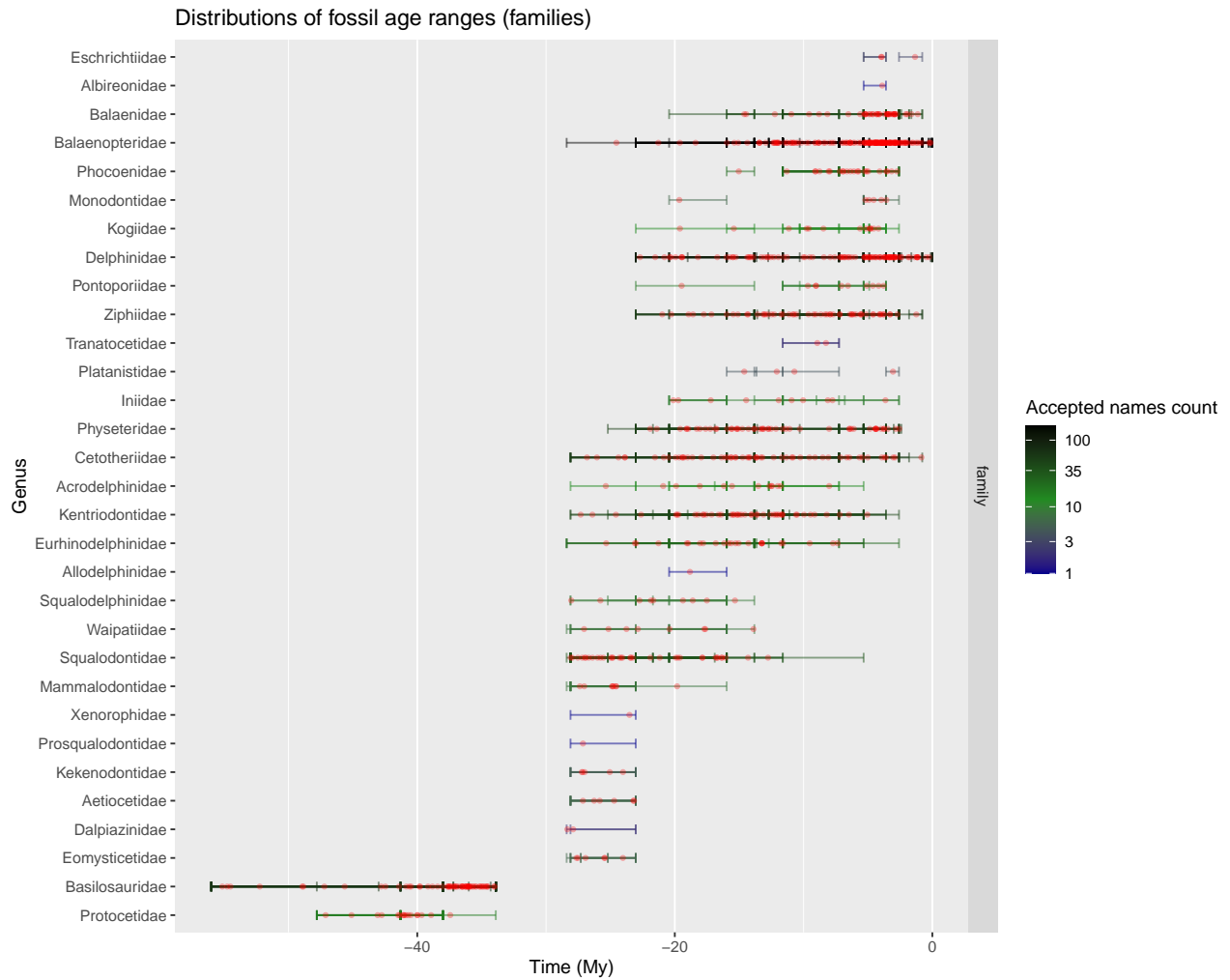


Distributions of fossil age ranges (species)



Distributions of fossil age ranges (genera)

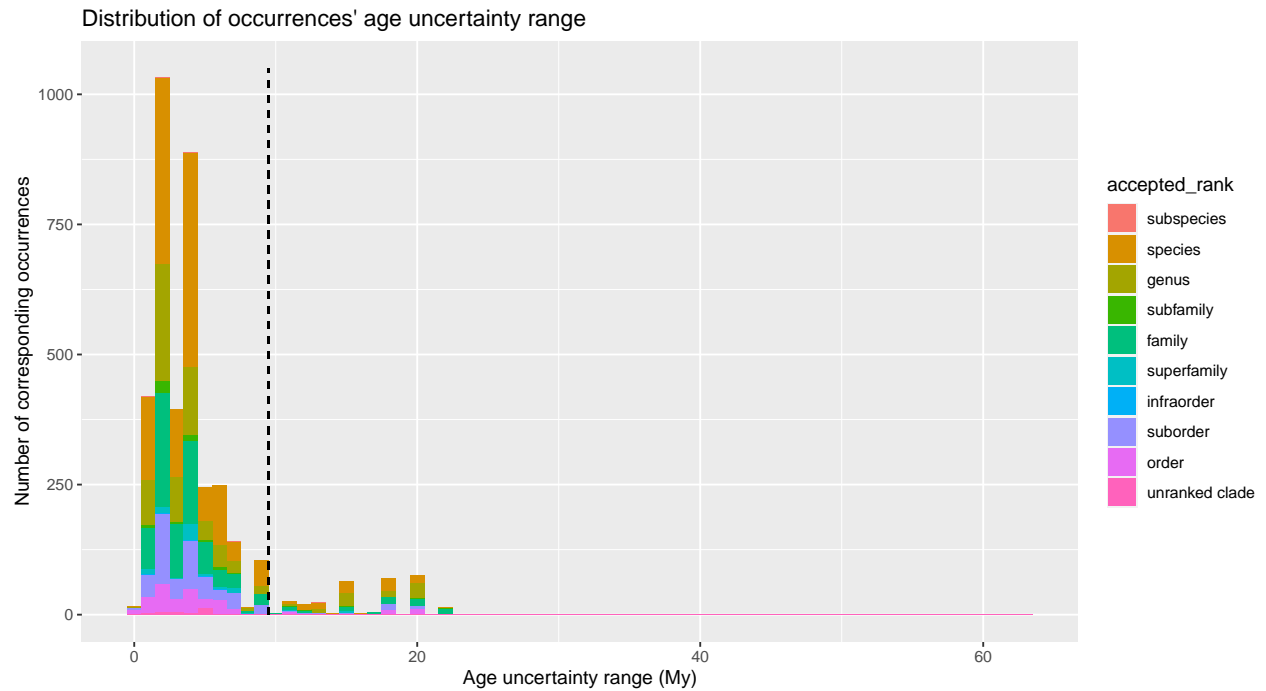




→ Some occurrences have too much age uncertainty (unrelated to species longevity in extreme cases like the one spanning 60My), they risk to artificially increase species durations.

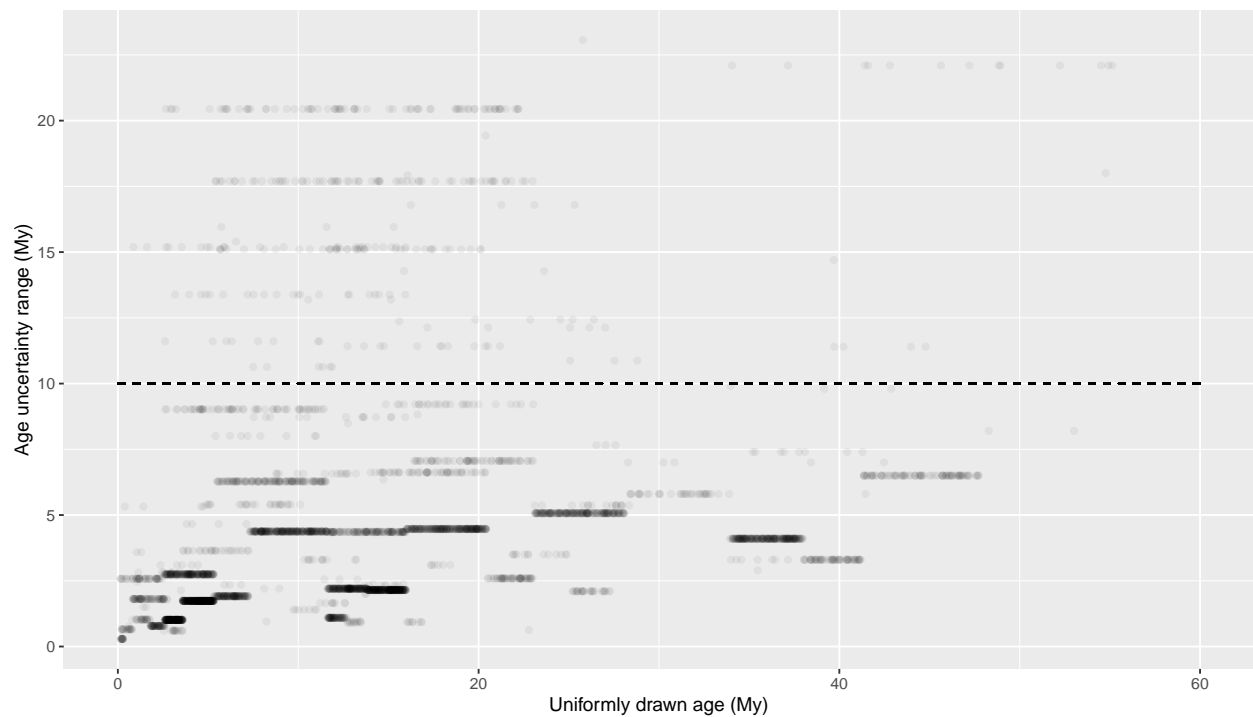
Remove occurrences with highly uncertain dating (range > 10My)

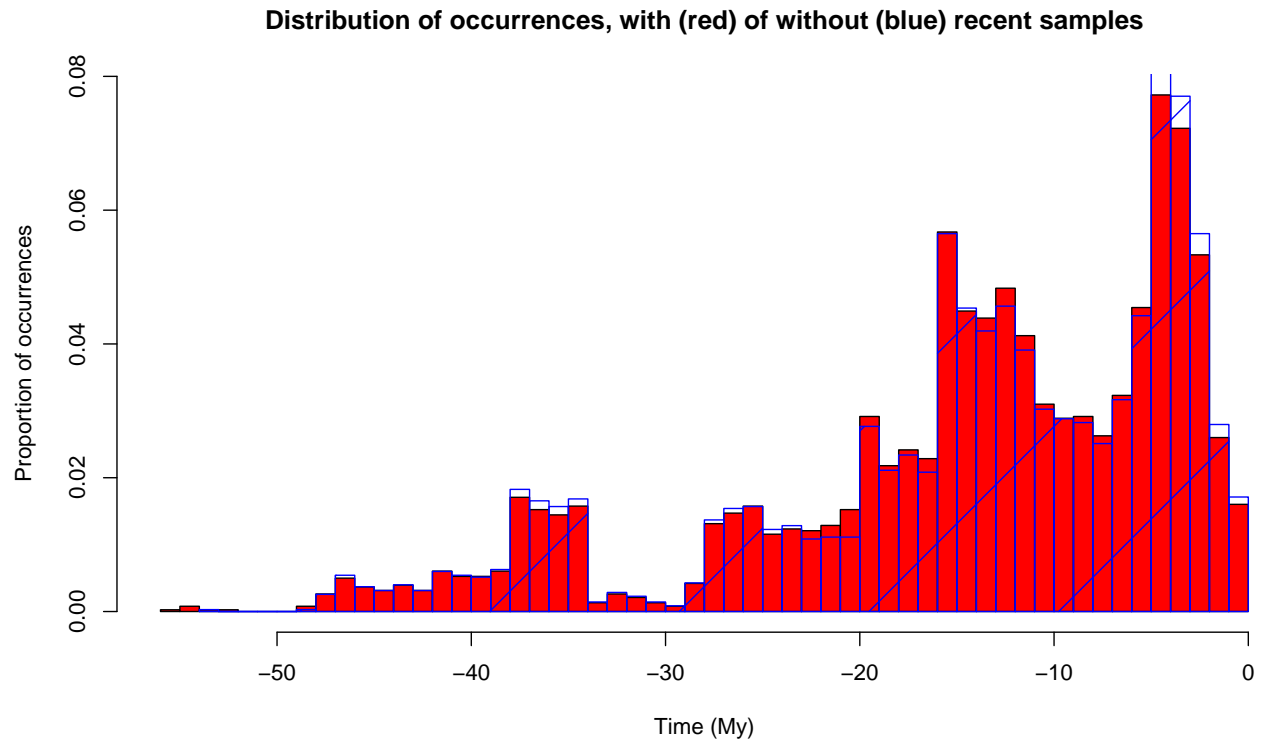
[1] 3505 117



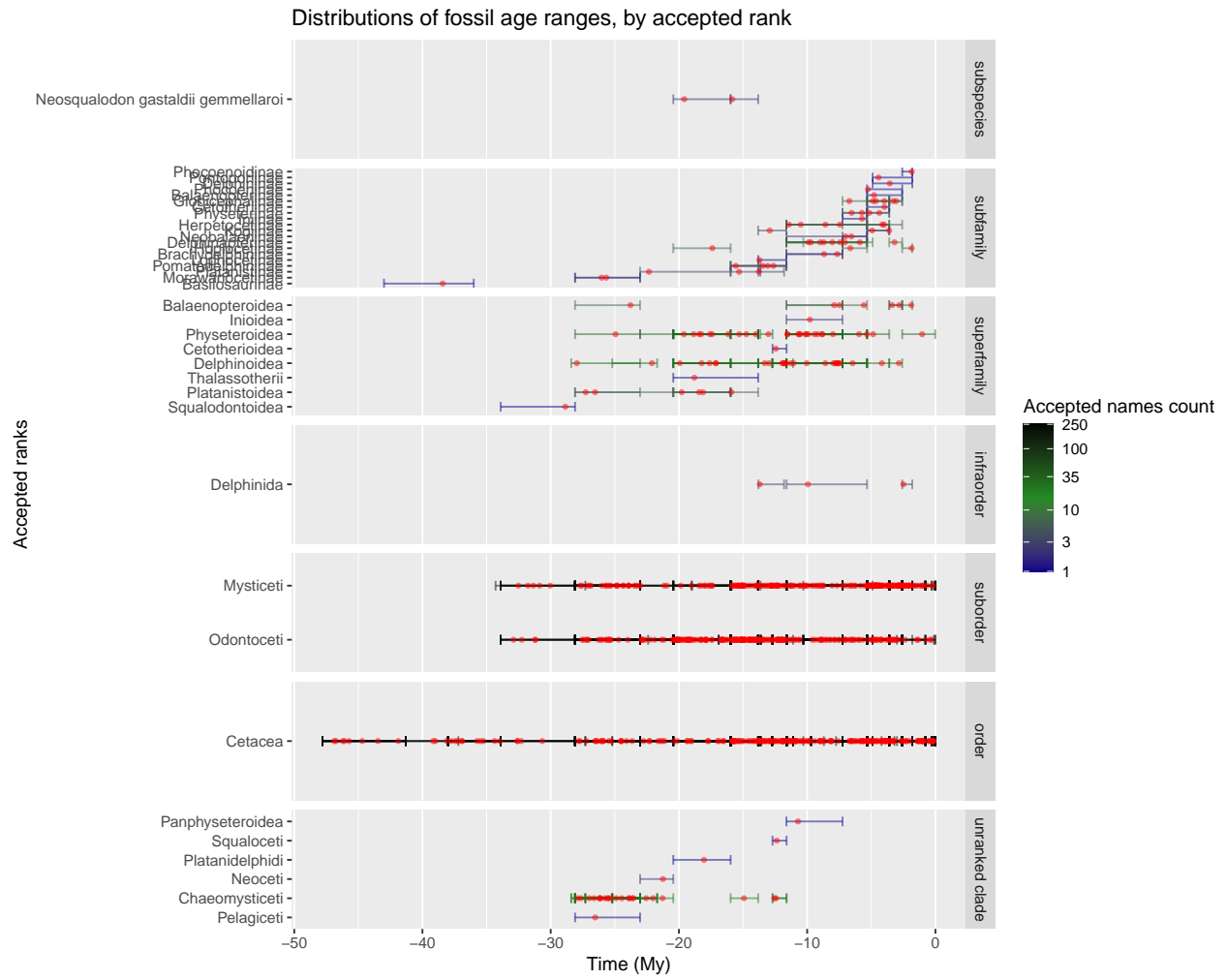
Most of occurrences show less than 10 My age uncertainty, let's keep only these ones.

```
##                               all_ranges smaller_10My removed
## Number of occurrences      3807      3505      302
Distribution of occurrences' age uncertainty range and removal limit (dashed line)
```

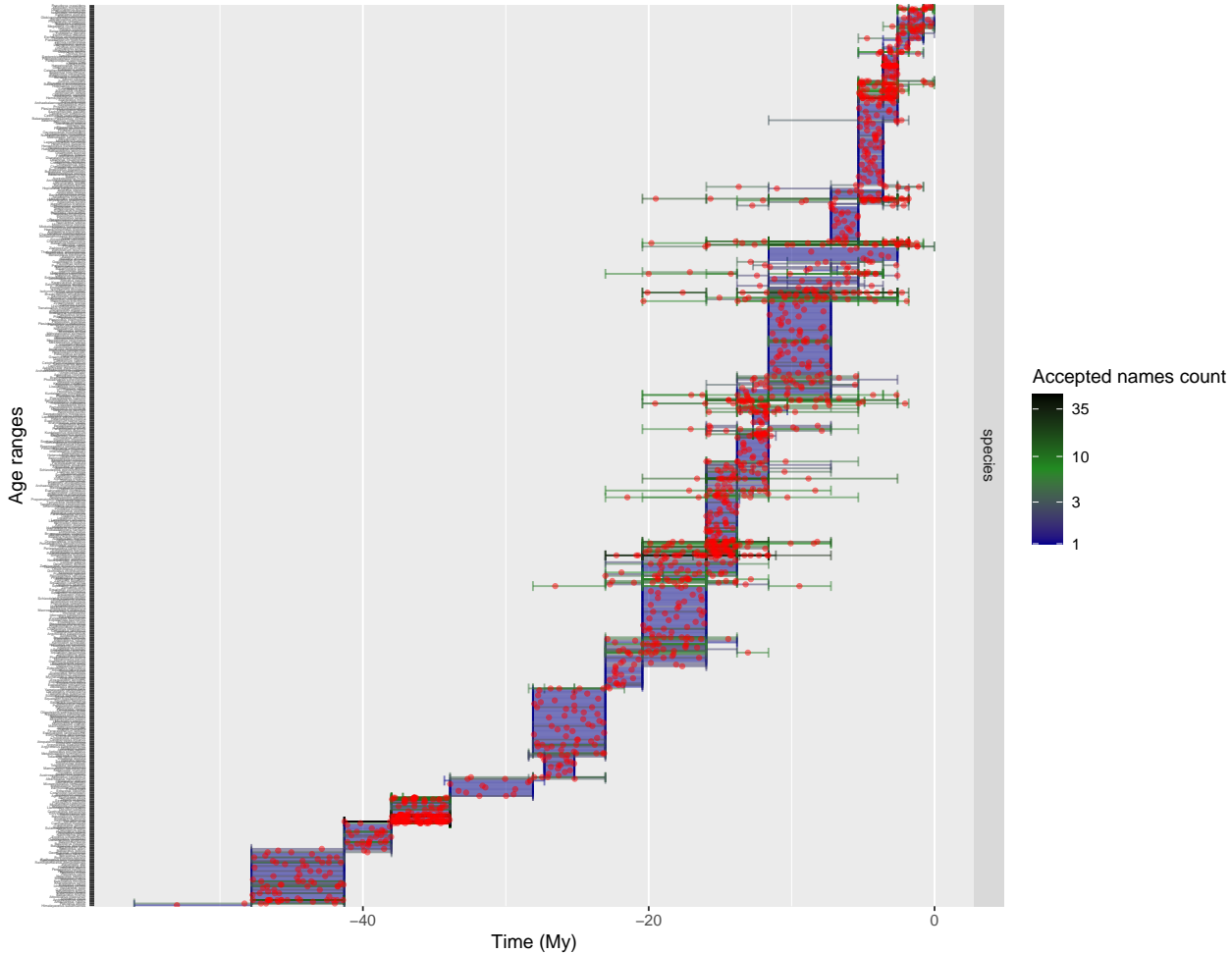




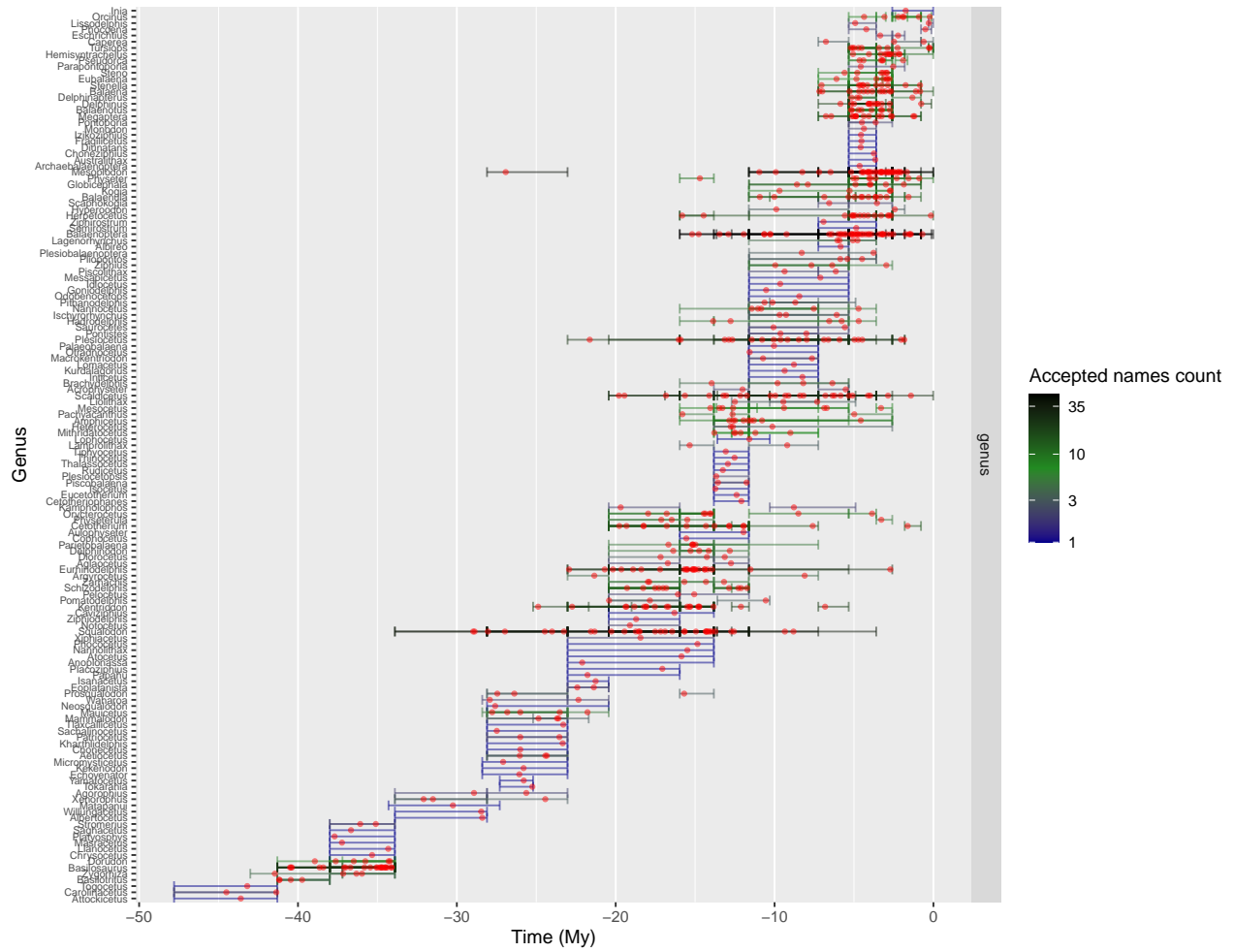
→ The removal of highly uncertain occurrences seems to be only a little biased, even if uncertainty globally increases with age.

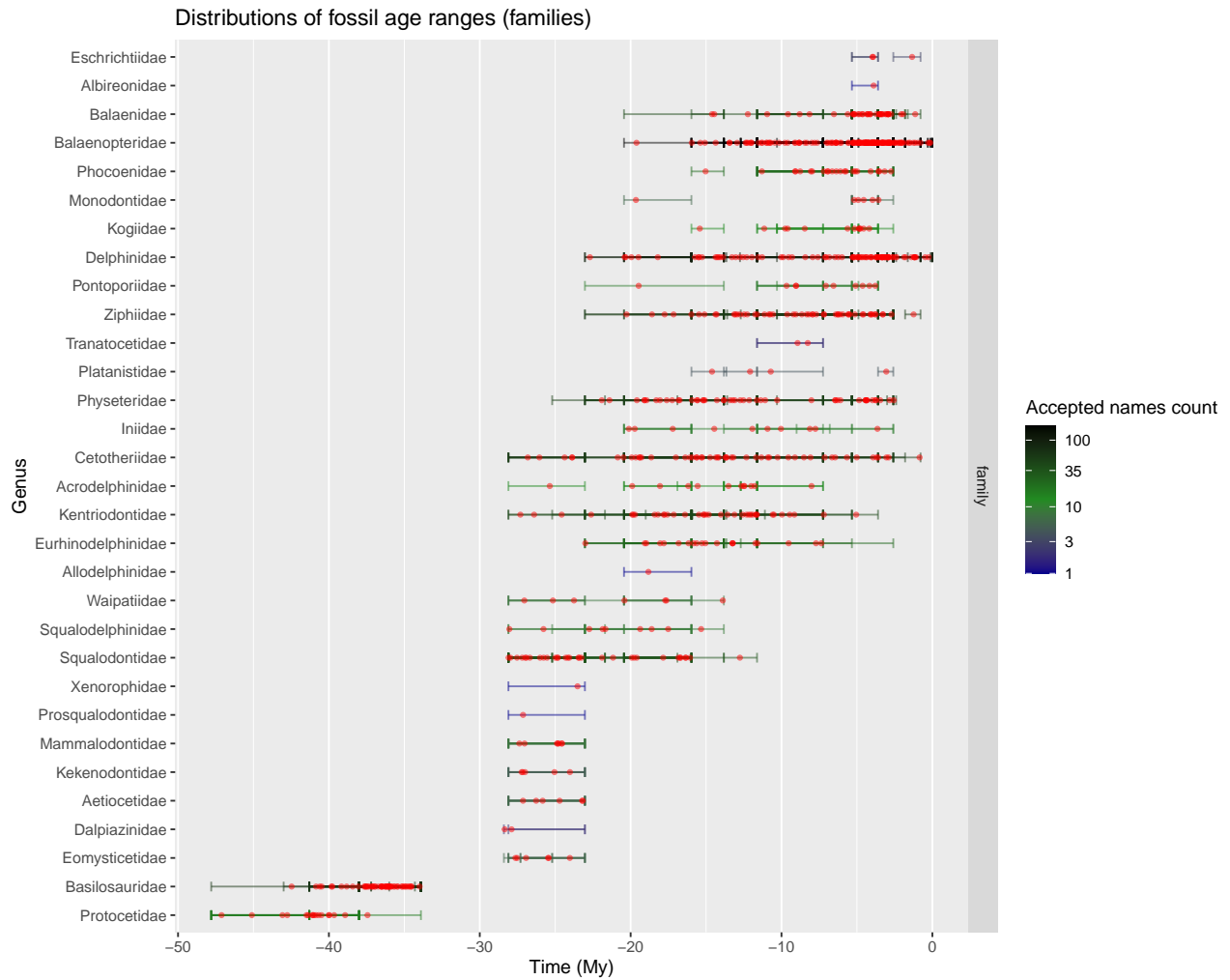


Distributions of fossil age ranges (species)



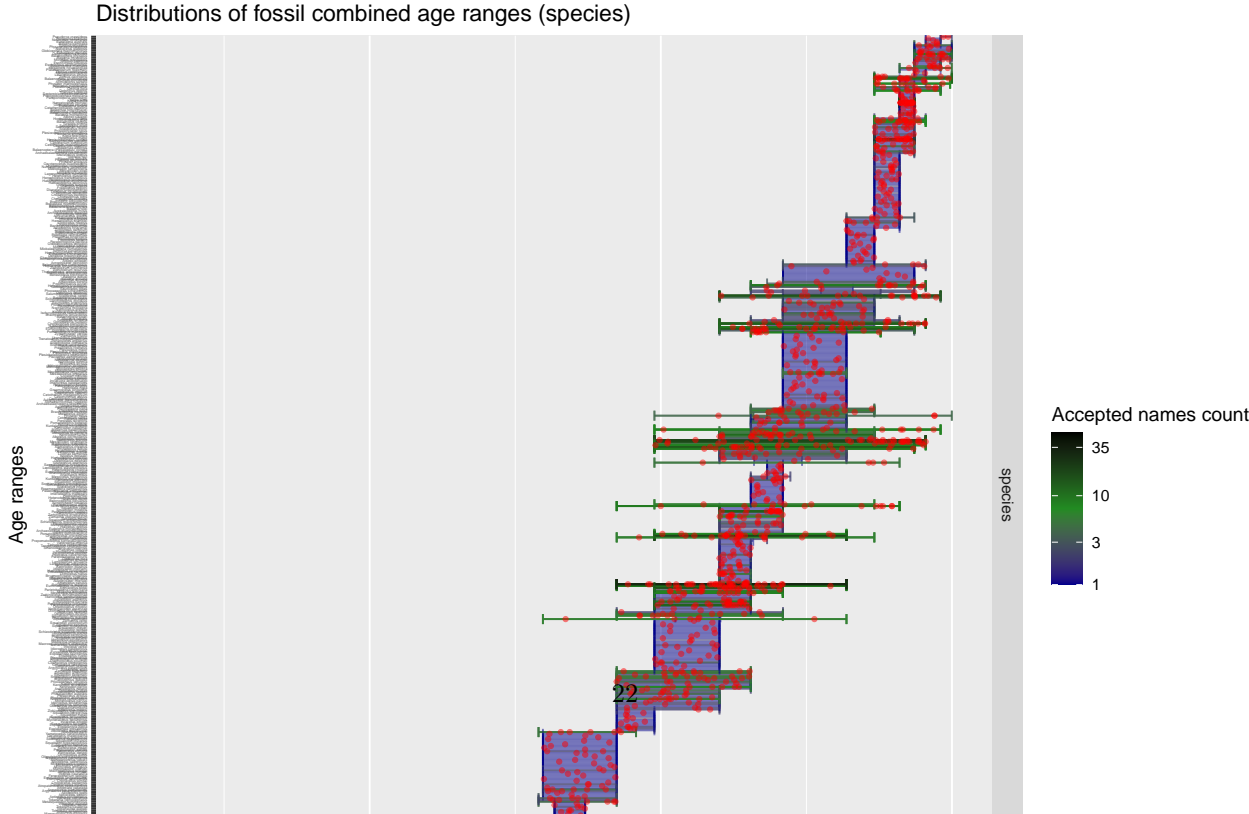
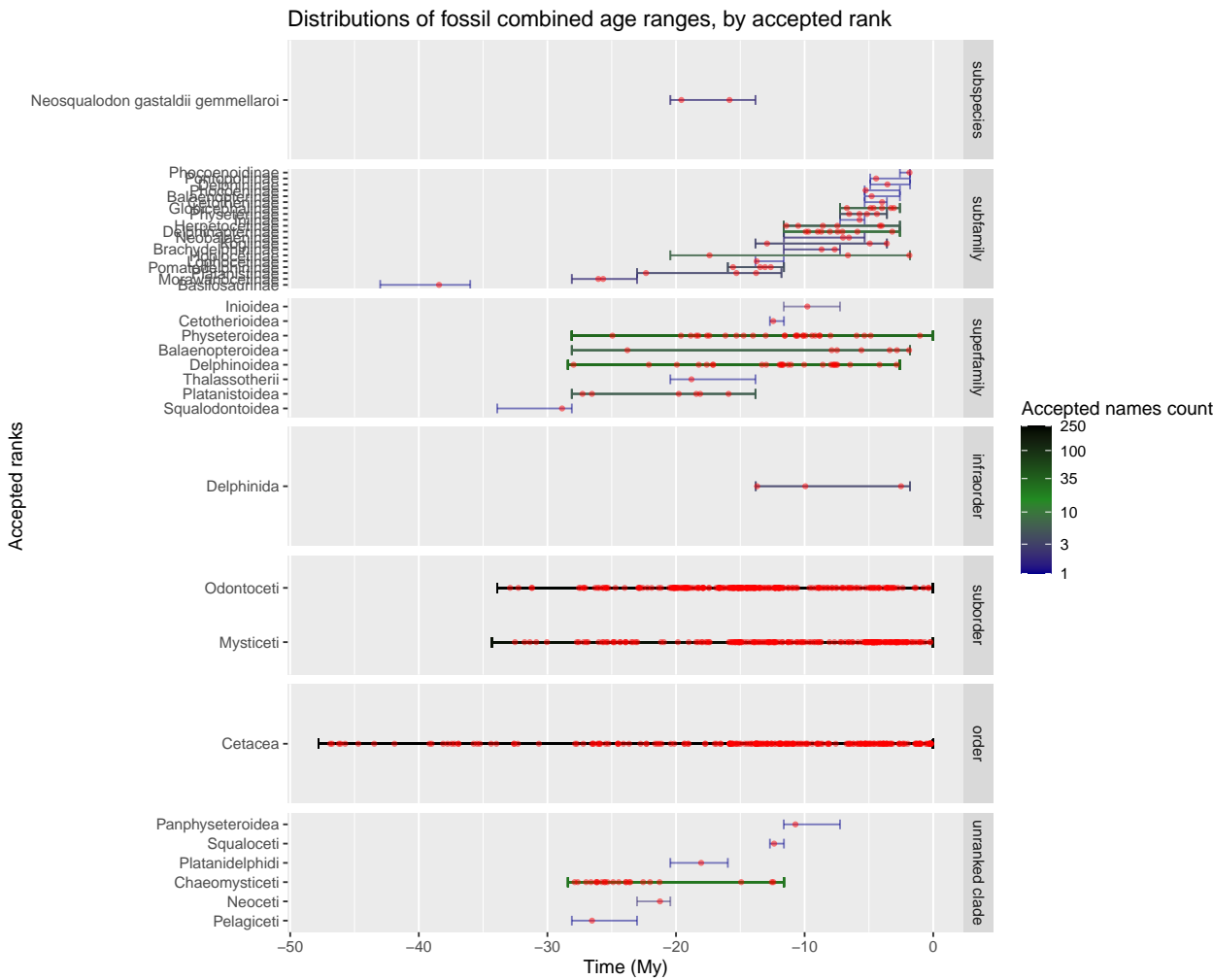
Distributions of fossil age ranges (genera)



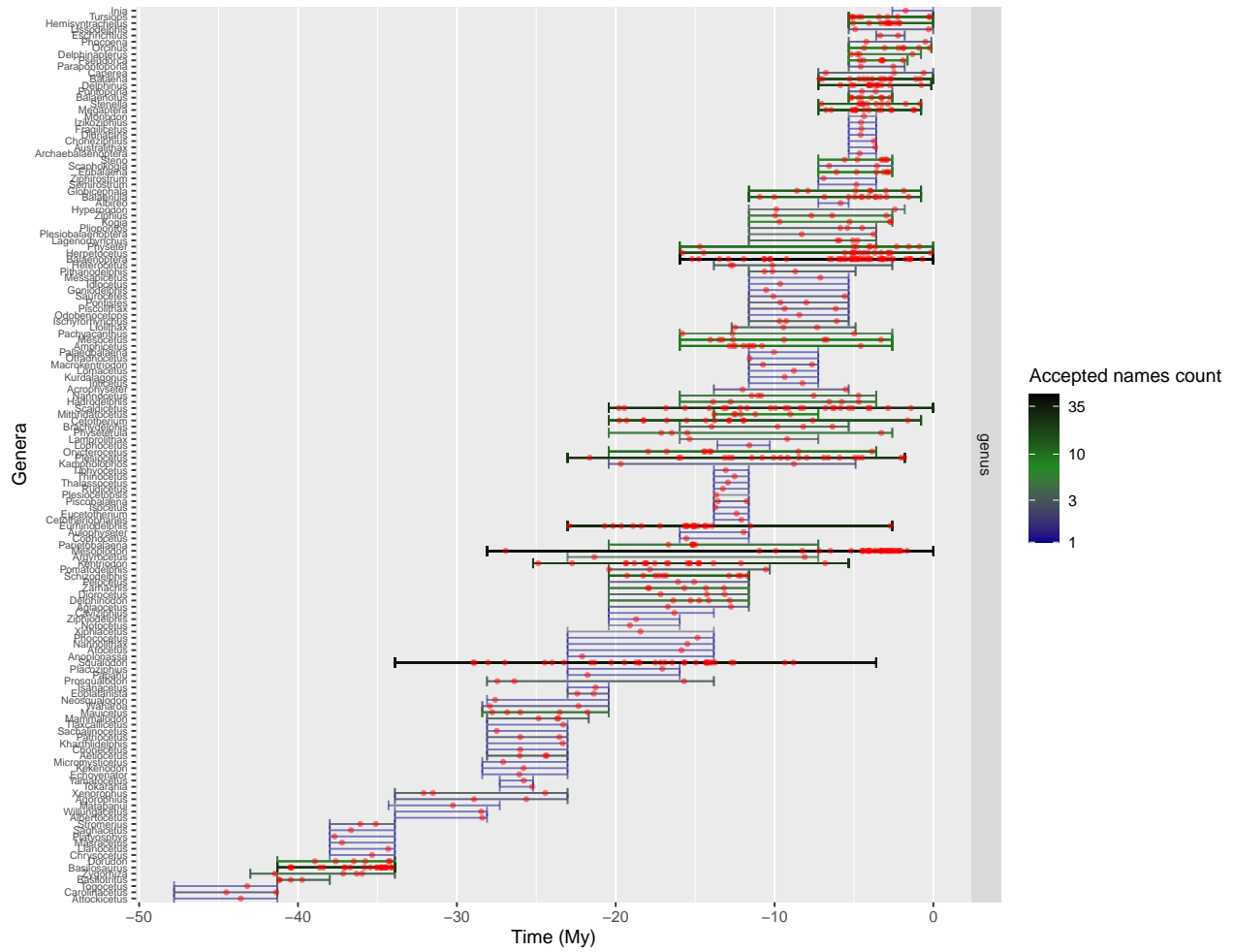


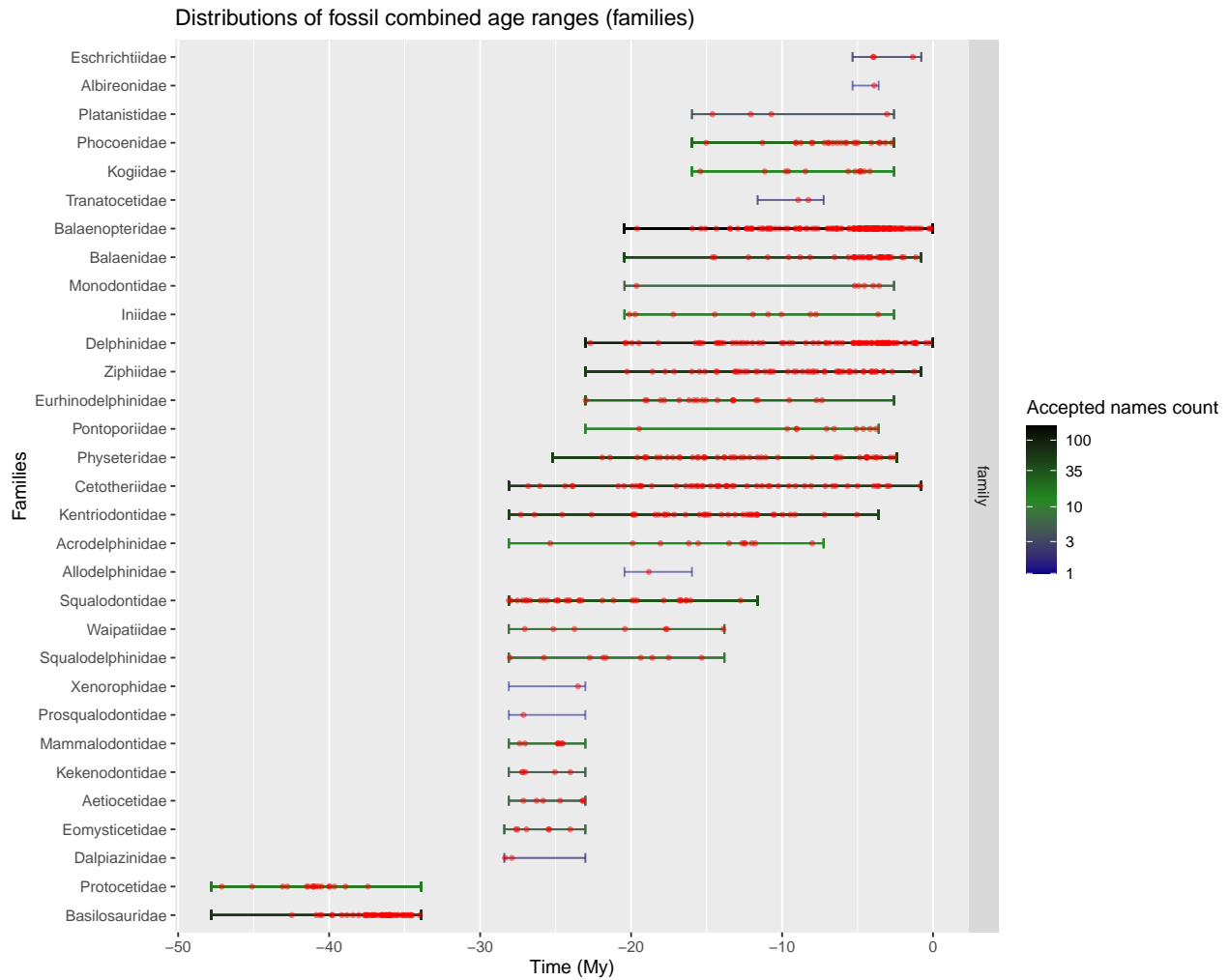
→ Some species (or other ranks) have several occurrences with several time ranges, **let's combine them into a unique range covering all the others.**

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



Distributions of fossil combined age ranges (genera)

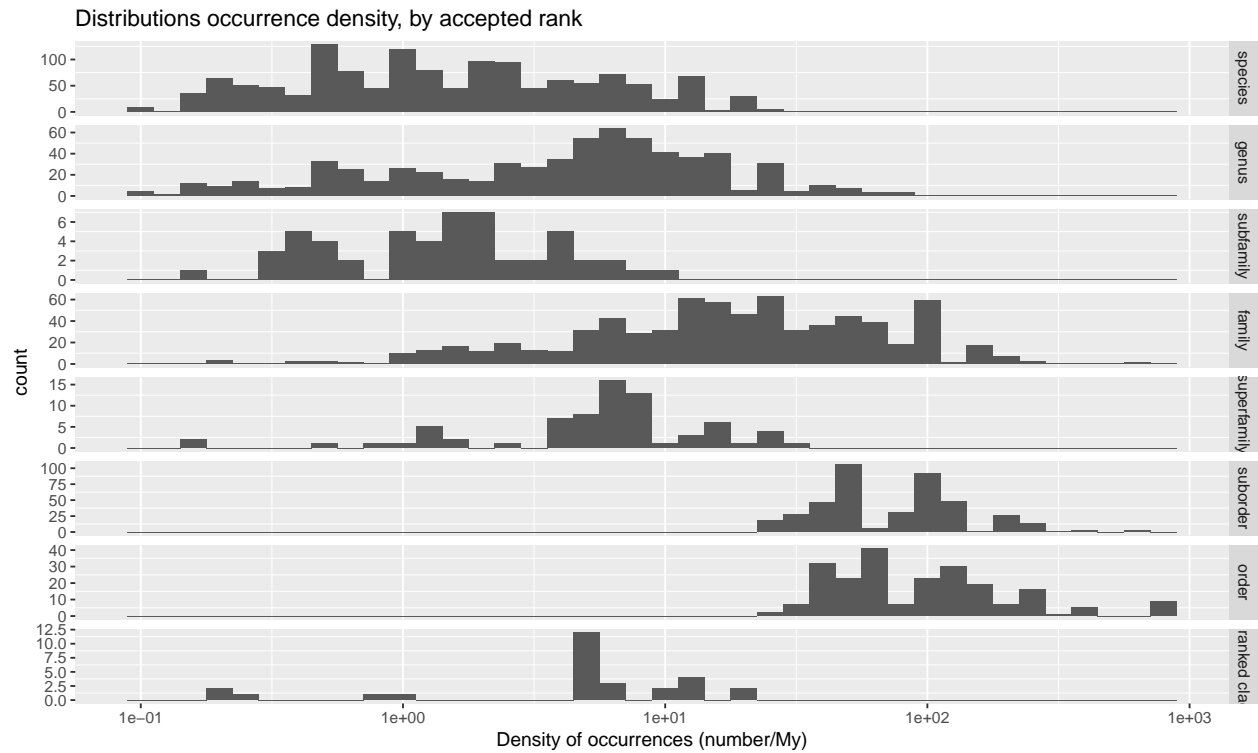




Occurrence density = number of occurrences by unit of time, in the stratigraphic interval of a taxon

Density distributions

Plot the density of occurrences for each taxonomic rank, and highlight the role

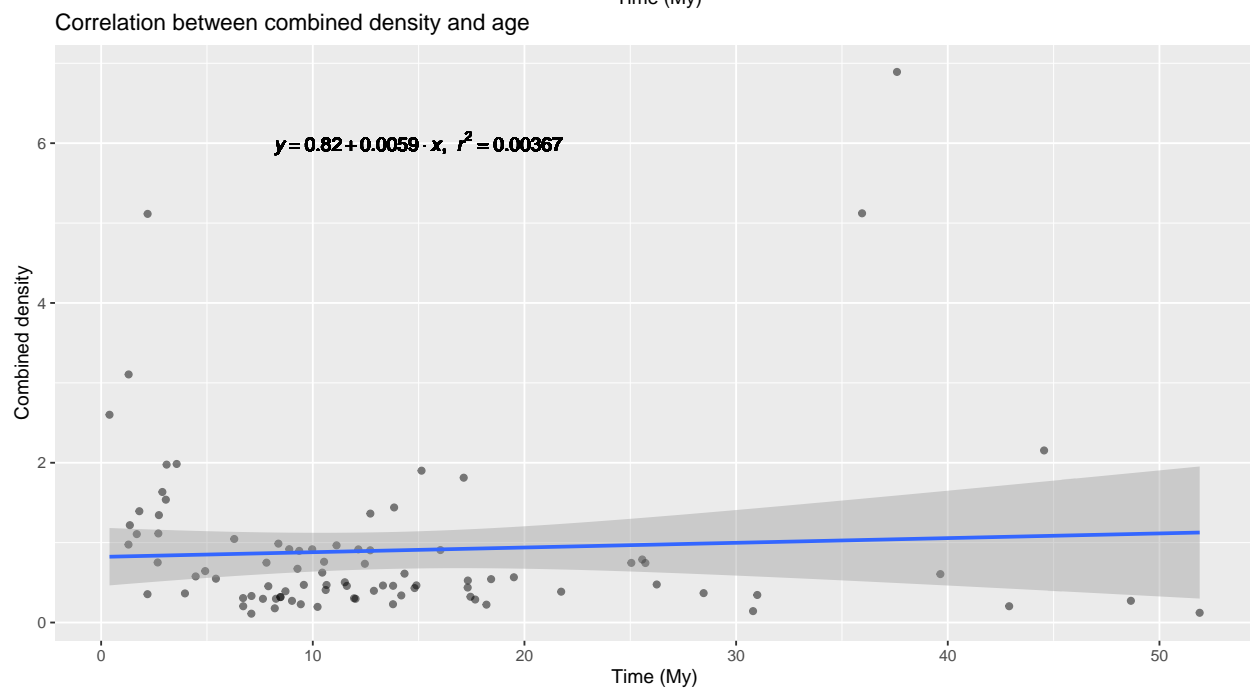
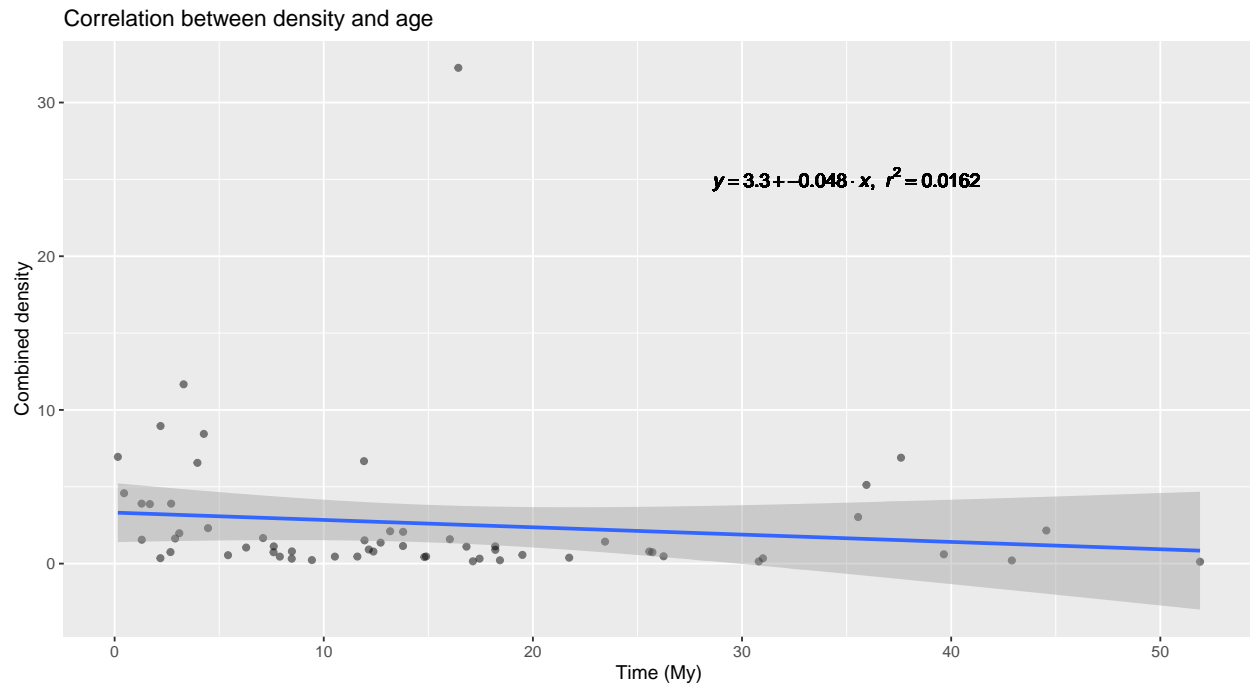


→ Density logically increases with taxa ranks, but the distributions for species and genera dramatically spread over two orders of magnitude.

Correlation between time range and age

If we want to correct species abundance differences based on the density of occurrences in the age interval, those factors should not depend on time in order to avoid penalizing periods with higher densities.

Let's compare the occurrence densities computed with the initial time ranges and the *combined densities* computed with the combined ranges.



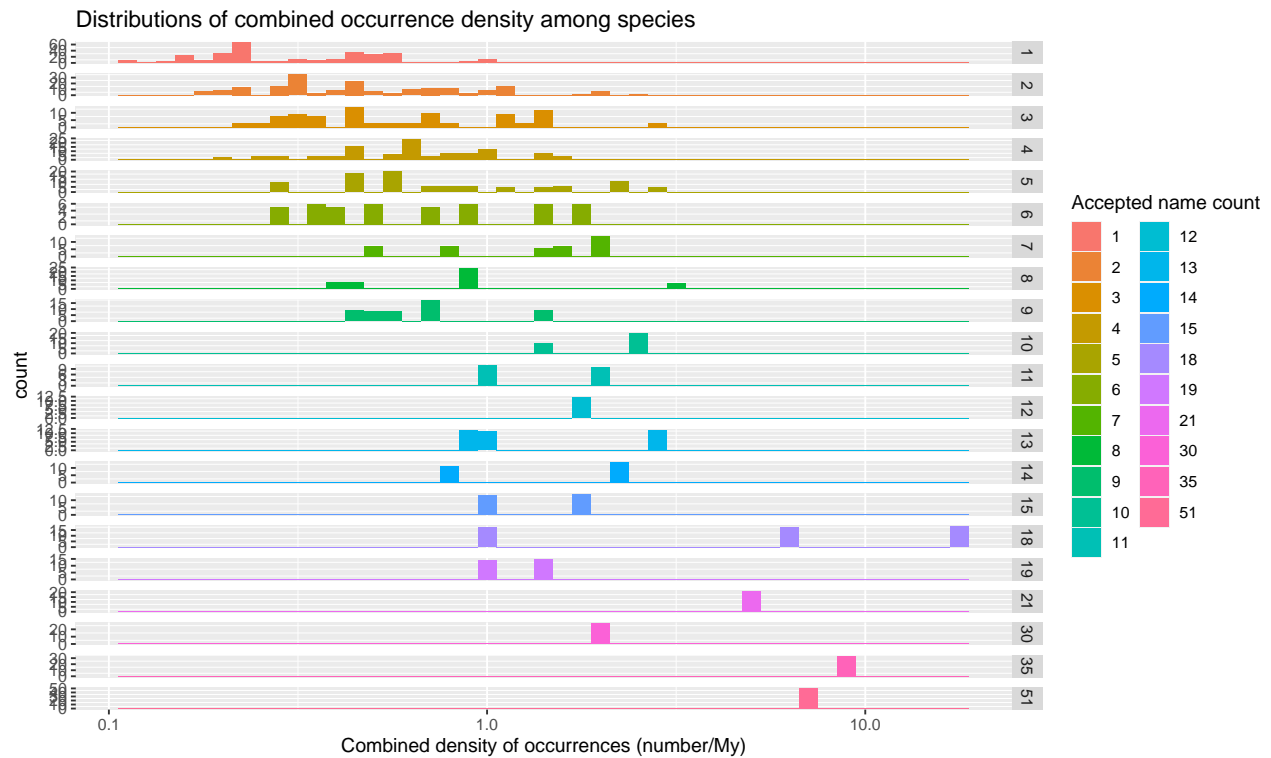
→ The density based on combined ranges is much less time-dependant than the density based on initial time ranges. We will therefore use the combined density for our corrections.

Subsampling occurrences by homogenizing the combined occurrence density

Subsampling by homogenization = reducing the discrepancy between high and low occurrence densities among taxa by subsampling preferentially the densest ones.

Compare densities by accepted name count (species only)

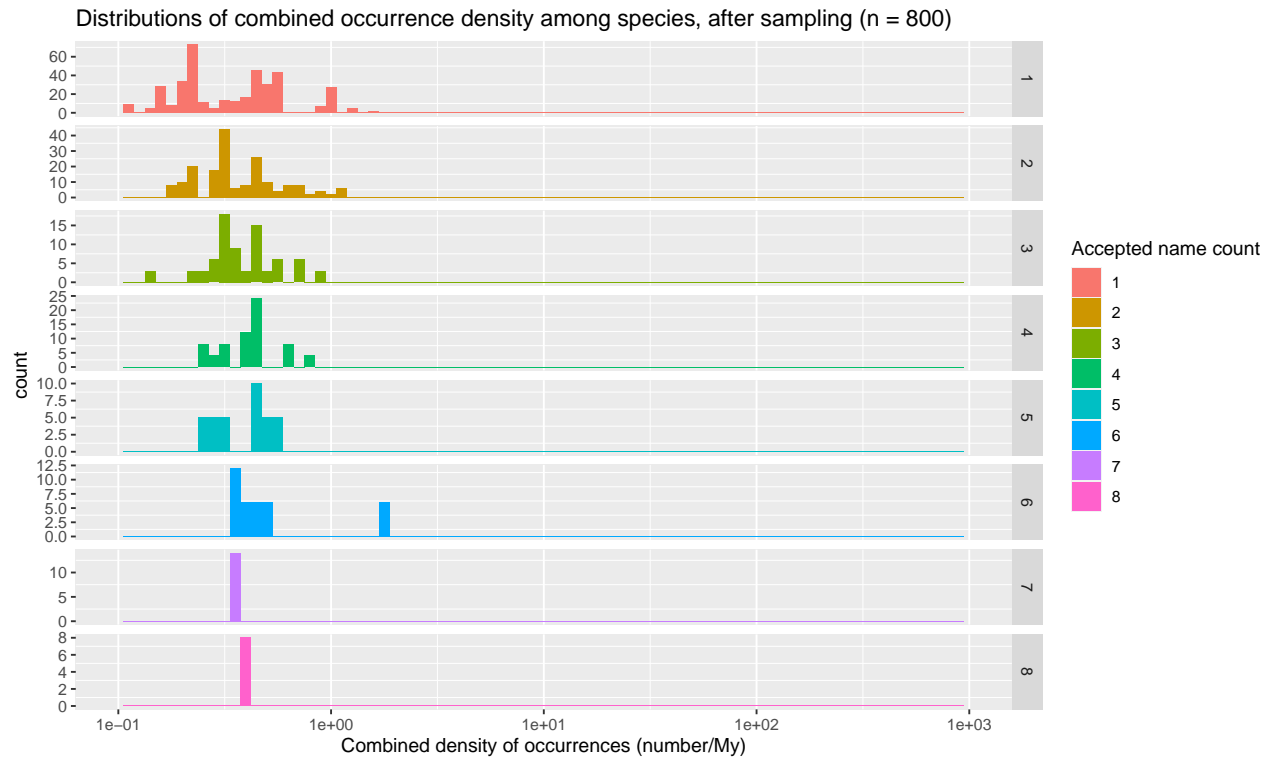
Let's focus now on the occurrences accepted at the species level because they are the one for which we can correct the abundance bias by subsampling the most concentrated combined intervals.



→ There 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)

```
## Warning: Removed 16 rows containing missing values (geom_bar).
```

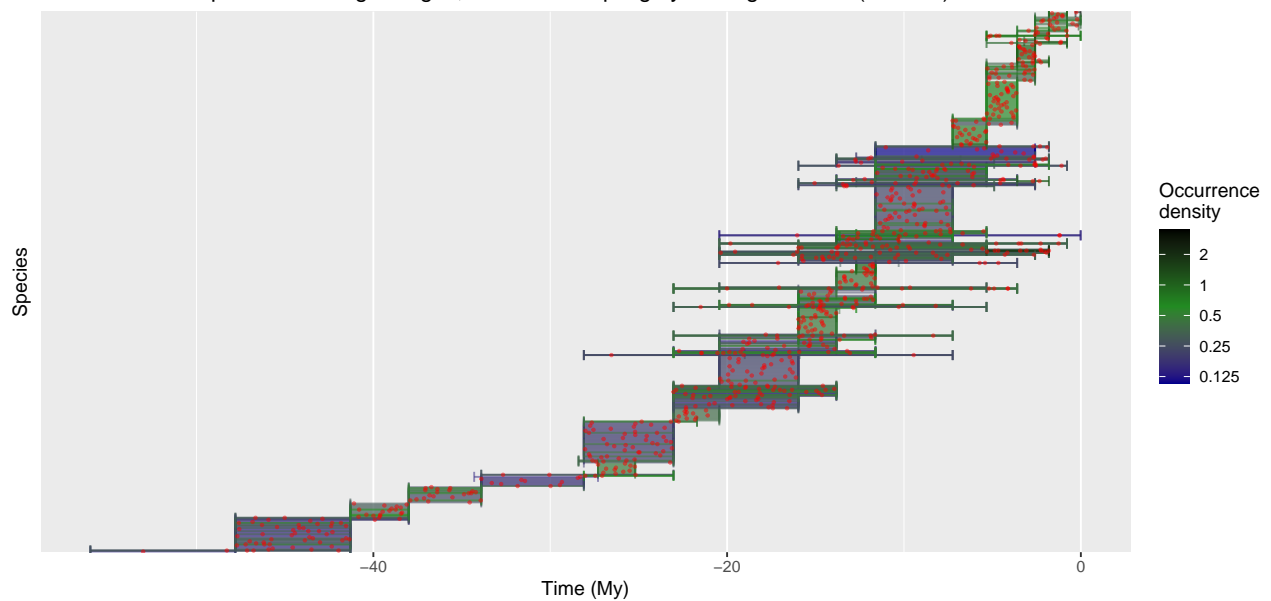


→ Subsampling by homogenization successfully reduces the density span from 2 to 1 order of magnitude.

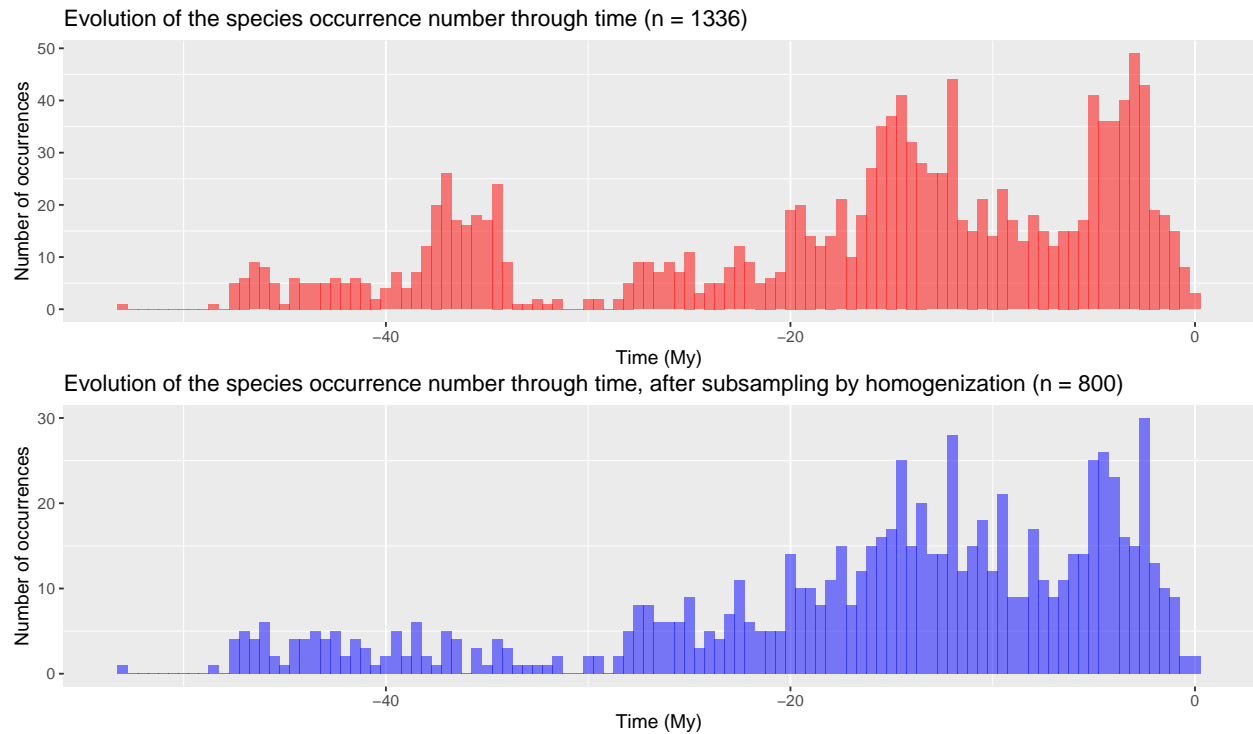
Impact of subsampling on occurrences repartition (species only)

See what our distributions look like after subsampling :

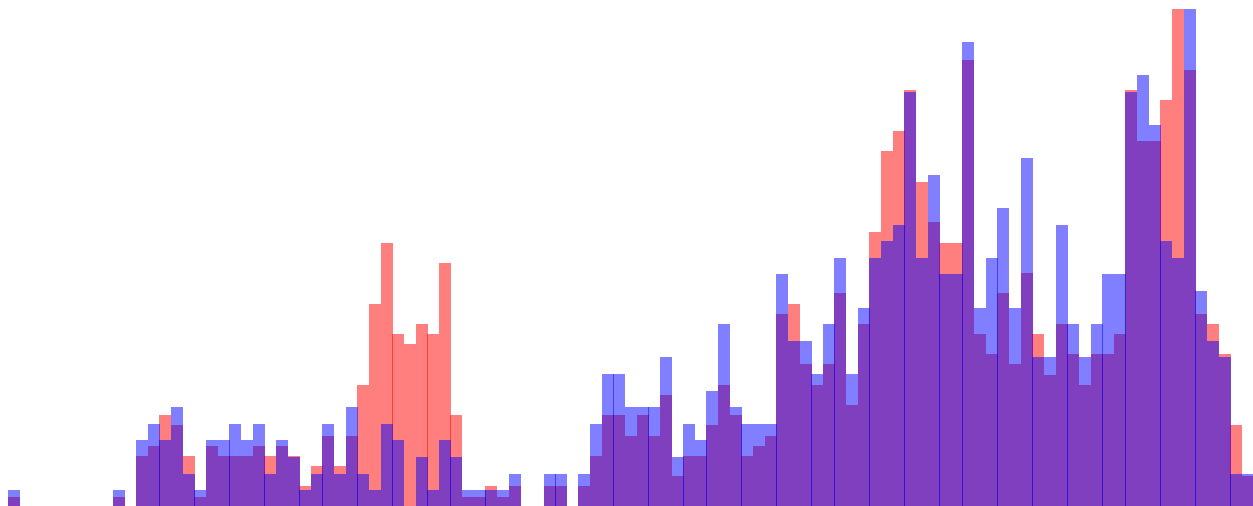
Distributions of species fossil age ranges, after subsampling by homogenization (n = 800)



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



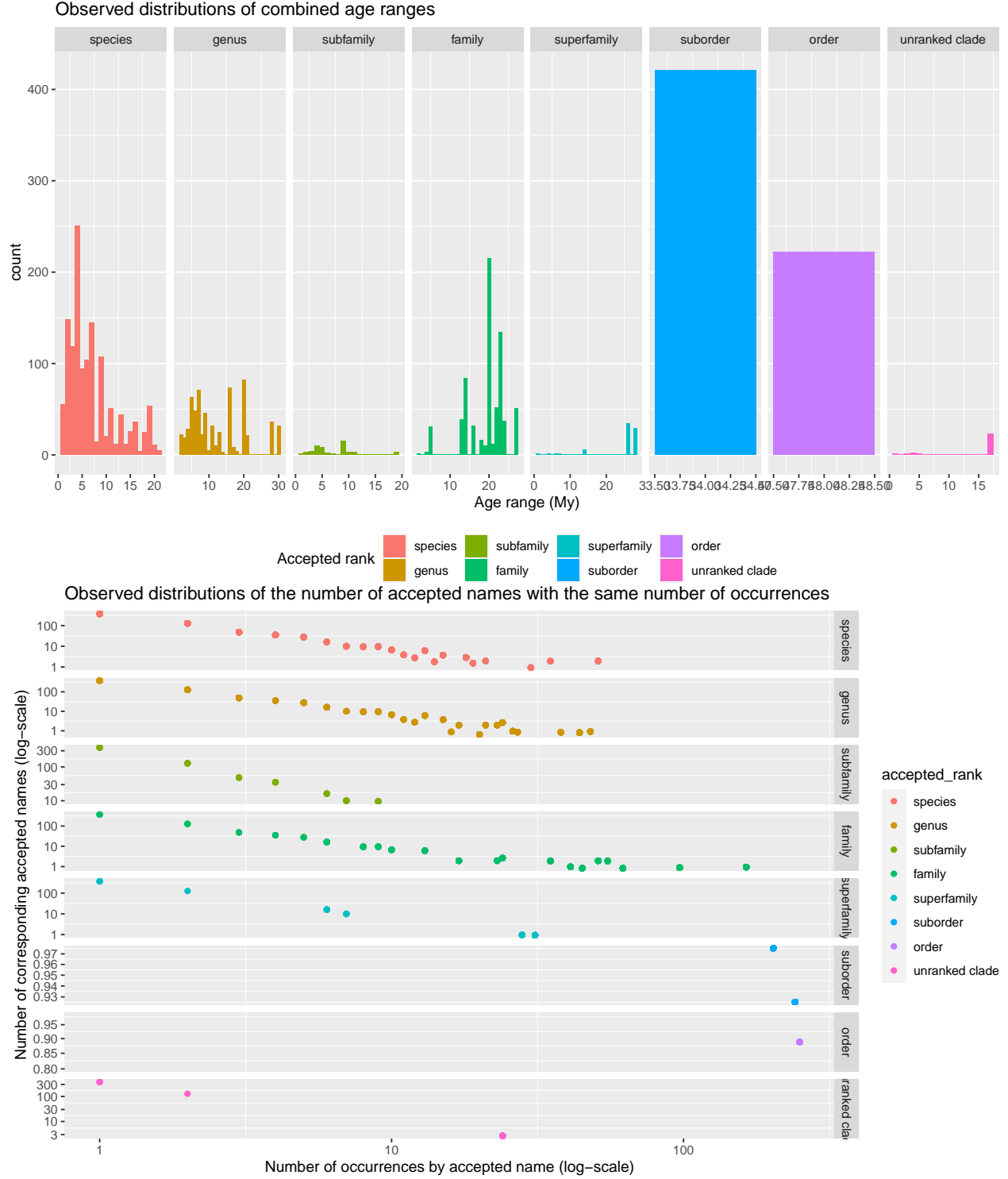
If we superpose these 2 plots :



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

Compare with a Poisson sampling process

In order to check if the data fit our assumptions of constant-fossilisation-rate Poisson sampling we compare the observed occurrences distributions with the expected ones. Specifically, we will look at the number of taxa represented by 1, 2, 3, ... occurrences and the one that we would expect for a given distribution of combined age ranges (seen here as a proxy for species duration).



In a Poisson process with occurrence sampling rate ω and for a given time interval of length t , the probability of observing $N_t = k$ occurrences is given by the Poisson distribution of mean of parameter $\omega \times t$:

$$\mathbb{P}(N_t = k) = e^{-\omega t} \frac{(\omega t)^k}{k!}$$

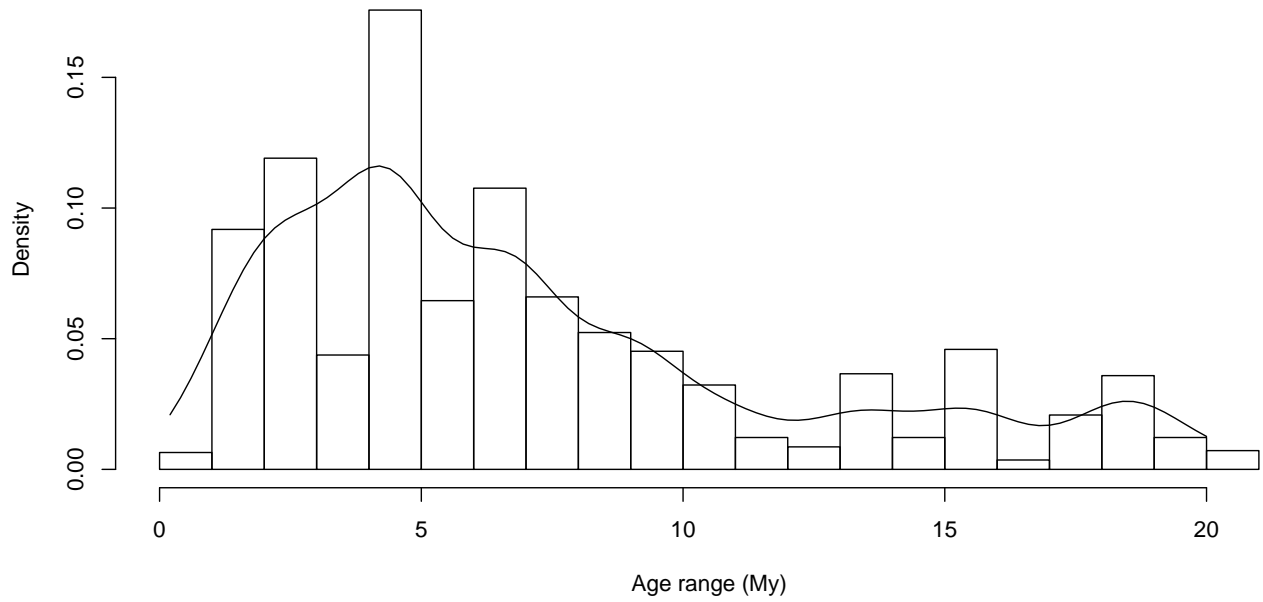
So in order to have the absolute probability of observing $N_0 = n$ occurrences we have to integrate over the

full distribution of age ranges t , called $f(t)$:

$$\mathbb{P}(N_0 = n) = \int_t P(N_t = n) f(t) dt = \int_t e^{-\omega t} \frac{(\omega t)^n}{n!} f(t) dt$$

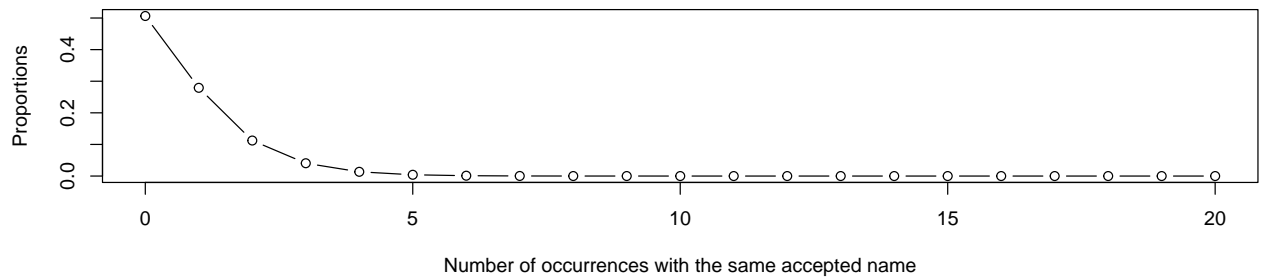
First, approximate this distribution :

Density approximation of the empirical range distribution

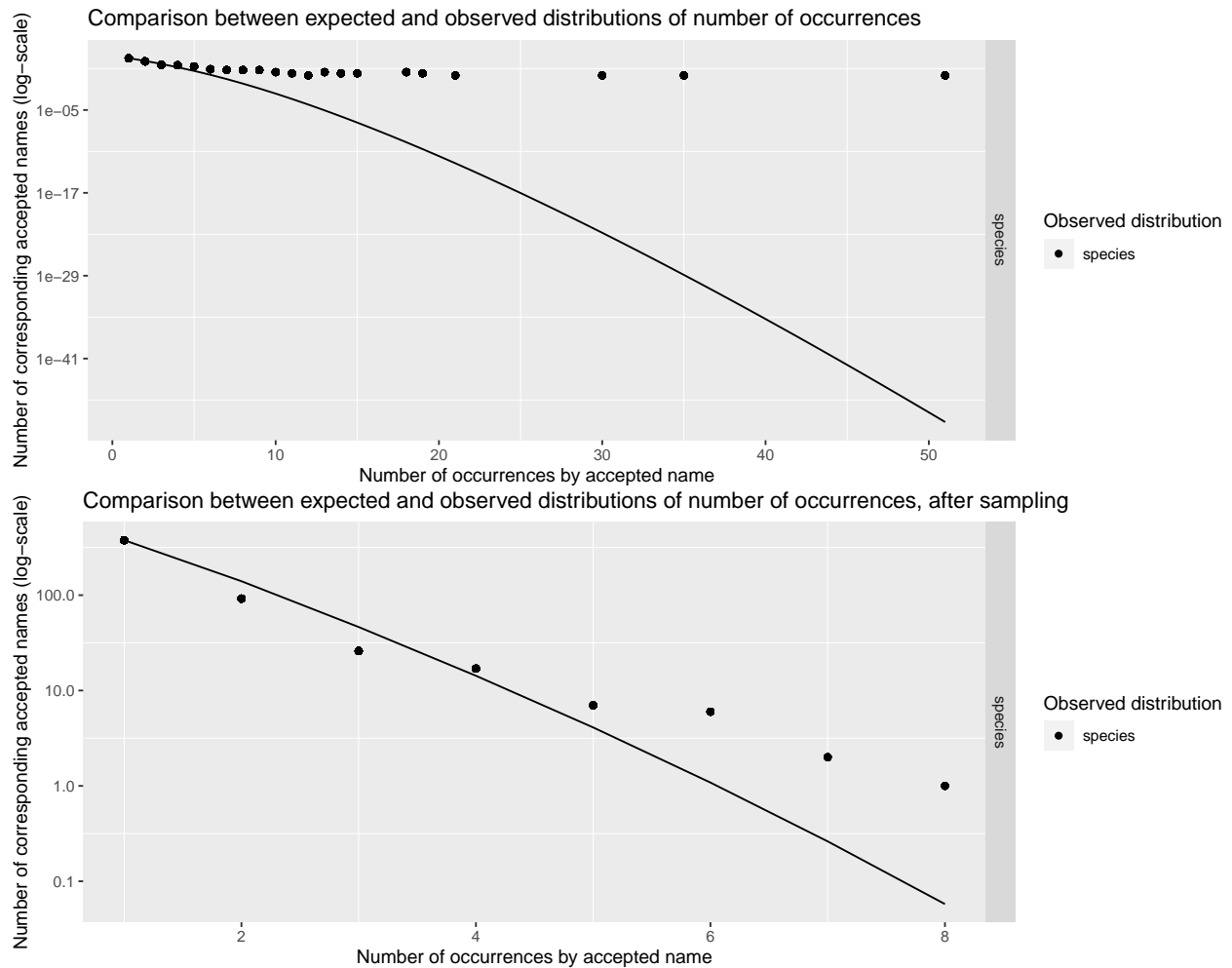


Then integrate and plot the expected distribution for a given omega :

Expected distribution of the number of accepted names with the same number of occurrences



Finally, try to find an ω value that approximately fits the first points (the least affected by oversampling biases) and check if the other points follow the expected curve :



⇒ Initial observations really do not fit the expectations, **species with more than 5 occurrences must remain very rare !** But our subsampling seems to correct most of this bias.

However, this method requires to make several arbitrary choices that may introduce new biases so we will instead subsample at other levels (palaeontological collection, geological formation). In each case only one occurrence will be sampled for the similarly identified, a process we will refer to as **aggregating** these occurrences according to the chosen factor.

Subsampling occurrences by aggregating similarly identified occurrences in each collection

Subsampling by aggregation = aggregate all the occurrences of a given taxon according to a given criterium (collection, geological formation, country, geological plate).

Aggregate similarly identified occurrences in each collection

In order to reduce the abundance bias, we may keep only one occurrence for each collection :

```
##                               Cetacea_occ Cetacea_occ_aggreg removed
## Number of occurrences                3807                3559    248
## Number of occurrences (species only)    1436                1363     73
```


→ Not enough occurrences are removed to make a sufficient difference. If we look at the collection with the highest number of occurrences :

##		
##	Aprixokogia kelloggi	Balaena
##	1	2
##	Balaenoptera acutorostrata	Balaenopteridae
##	2	2
##	Balaenula	Bohaskaia monodontoides
##	2	1
##	Cetotheriinae	Delphinapterus
##	1	3
##	Delphinidae	Delphinus
##	2	2
##	Globicephala	Gricetoides aurorae
##	2	1
##	Herpetocetinae	Herpetocetus sendaicus
##	1	1
##	Herpetocetus transatlanticus	Kogia breviceps
##	1	1
##	Kogiidae	Kogiinae
##	1	2
##	Kogiopsis floridana	Lagenorhynchus
##	1	2
##	Lagenorhynchus harmatuki	Megaptera
##	1	2
##	Mesoplodon longirostris	Monodon
##	2	1
##	Ninziphius platyrostris	Orycterocetus
##	3	1
##	Physeter macrocephalus	Physeteridae
##	1	1
##	Physeterinae	Physeterula dubusi
##	2	1
##	Plesiocetus	Pliopontos littoralis
##	1	1
##	Pontoporia	Pontoporiidae
##	1	2
##	Pseudorca	Scaldicetus
##	2	1
##	Stenella	Stenella rayi
##	2	1
##	Tursiops	Ziphius cavirostris
##	2	2

→ There are very few redundancies among the accepted names in collections so aggregating those won't reduce the abundance bias.

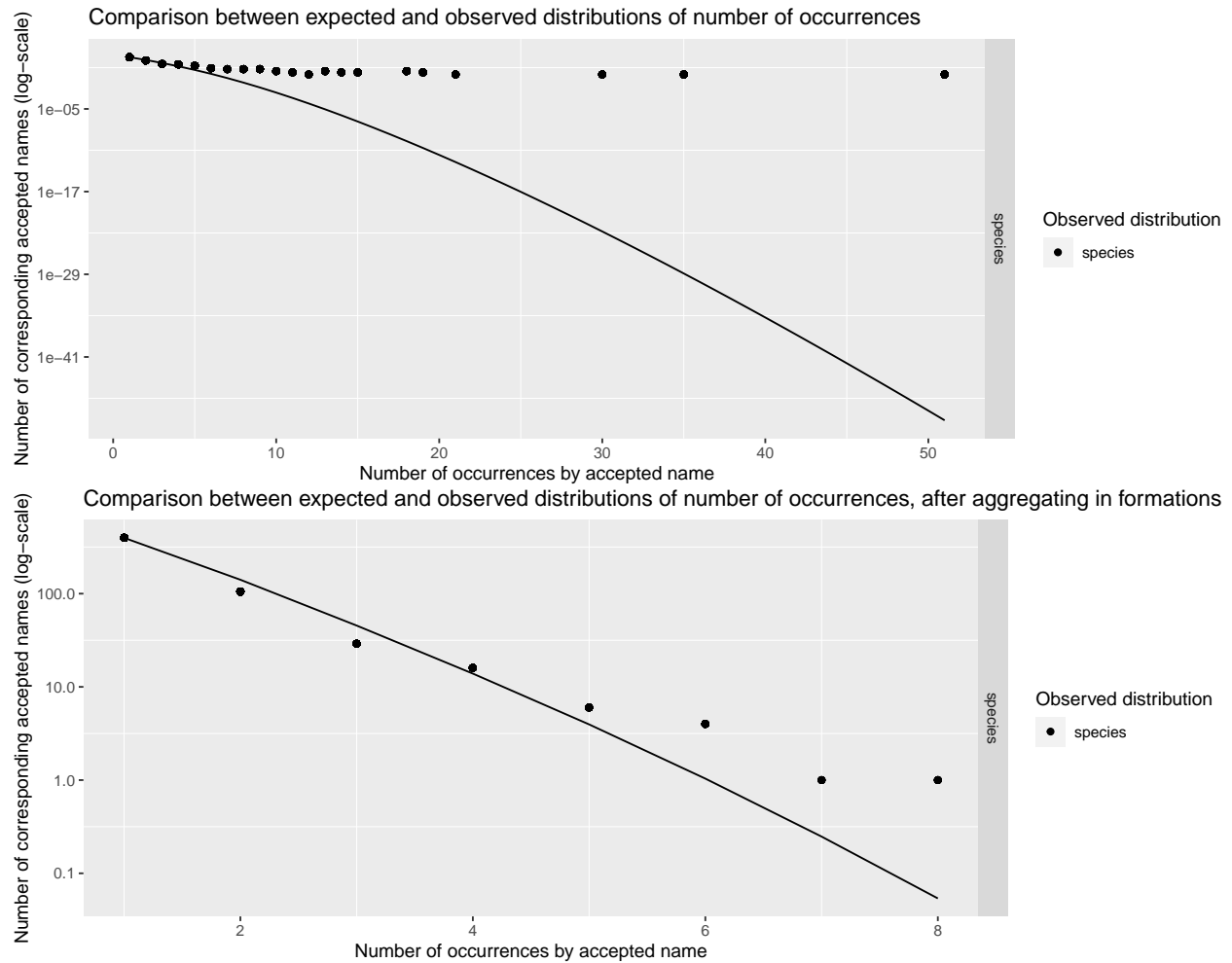
Instead, we may try to **aggregate occurrences with the same accepted name at the level of the geological formation** (ie subsample only one for each).

Aggregate similarly identified occurrences in each formation

##	Cetacea_occ	Cetacea_occ_aggreg	removed
## Number of occurrences	3807	1981	1826

Number of occurrences (species only) 1436 827 609

→ In that case the sub-sampling is big enough to hope correcting our bias.



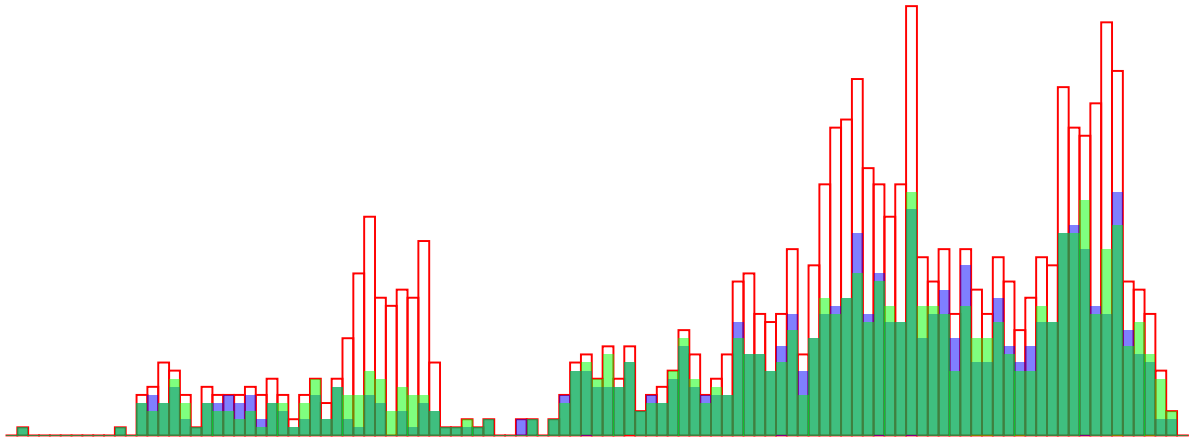
→ The initial bias is mostly corrected.

Warning: Removed 2 rows containing missing values (geom_bar).

Warning: Removed 2 rows containing missing values (geom_bar).

Warning: Removed 2 rows containing missing values (geom_bar).

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in formations (green)



→ Comparing with the initial occurrences distribution and with the distribution after our first sub-sampling it appears that both methods lead to very similar distributions. This comforts us about the robustness of those approaches.

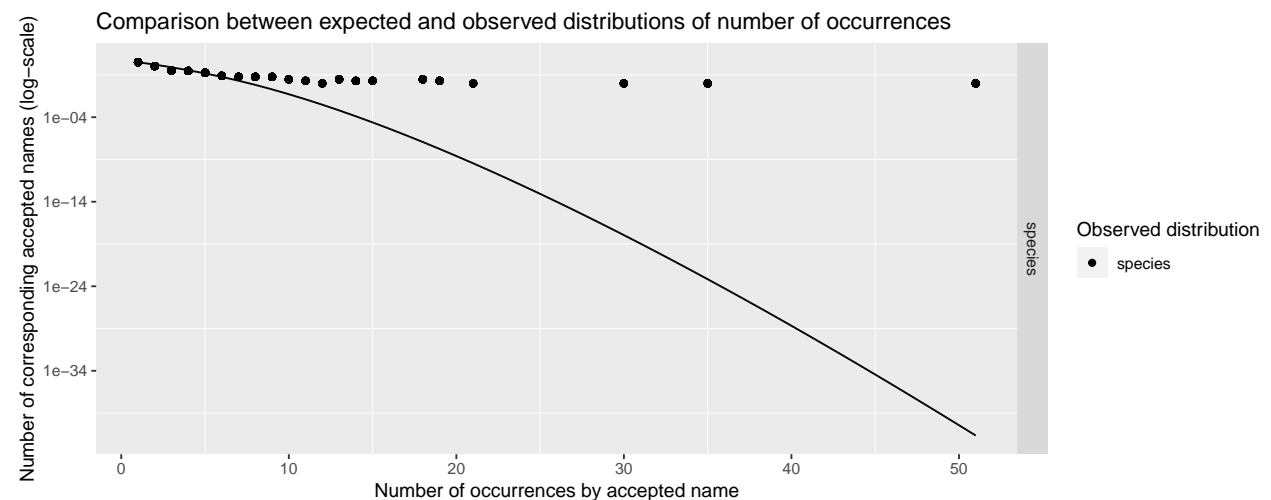
However, we have to take into account the occurrences that do not have any indicated geological formation to subsample them separately.

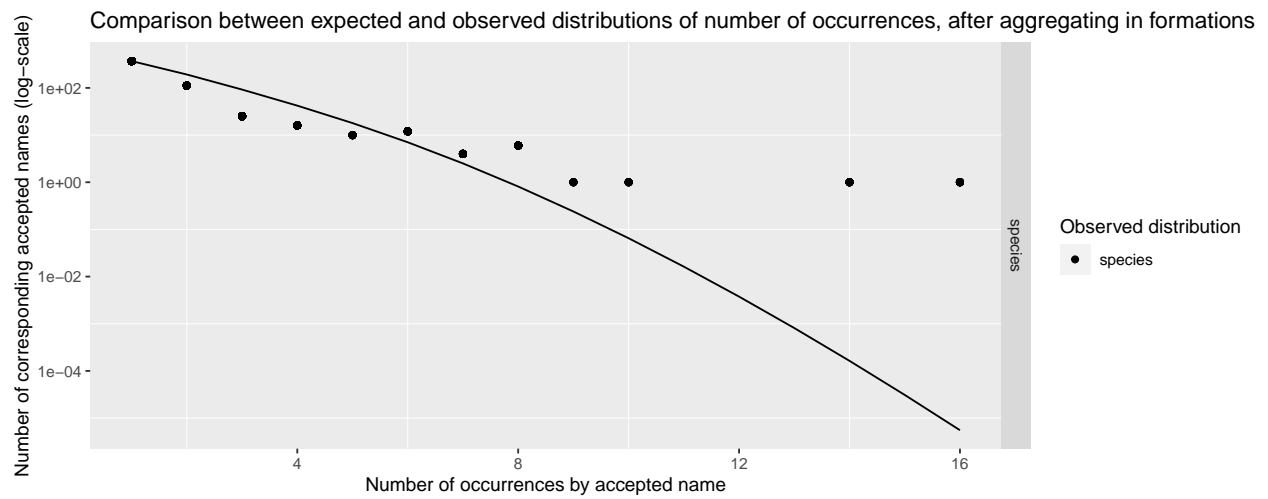
Aggregate occurrences without formation by country + early interval

To approximate geological formation we chose to proceed to the aggregation based on the combination of the country and the early stratigraphic interval.

##	Cetacea_occ	Cetacea_occ_aggreg	removed
## Number of occurrences	3807	2642	1165
## Number of occurrences (species only)	1436	980	456

More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution :

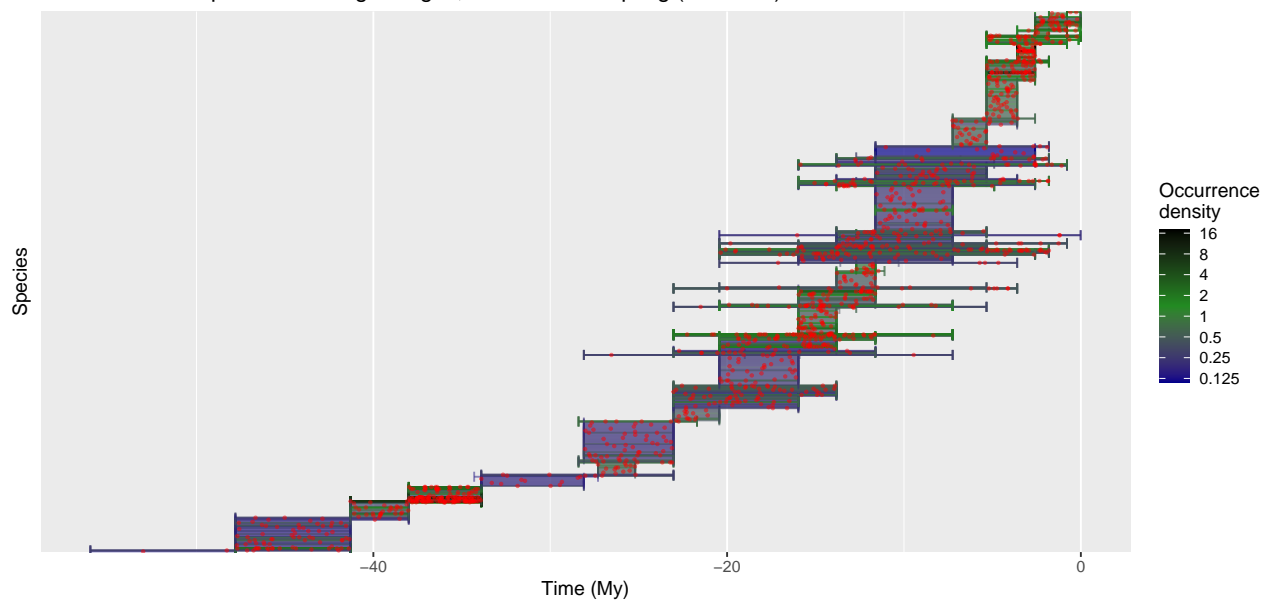




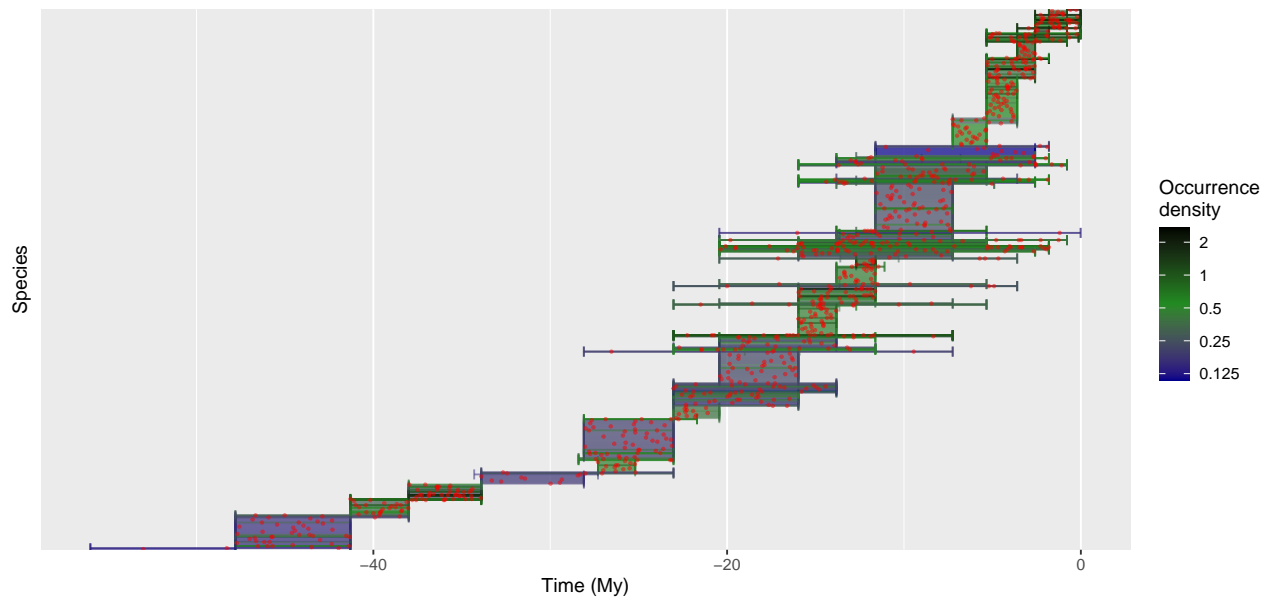
→ The correspondance is still good, except for two taxa :

```
##
## Scaldicetus grandis Schizodelphis sulcatus
##          14          16
```

Distributions of species fossil age ranges, before subsampling (n = 1336)



Distributions of species fossil age ranges, after subsampling by aggregation (n = 800)

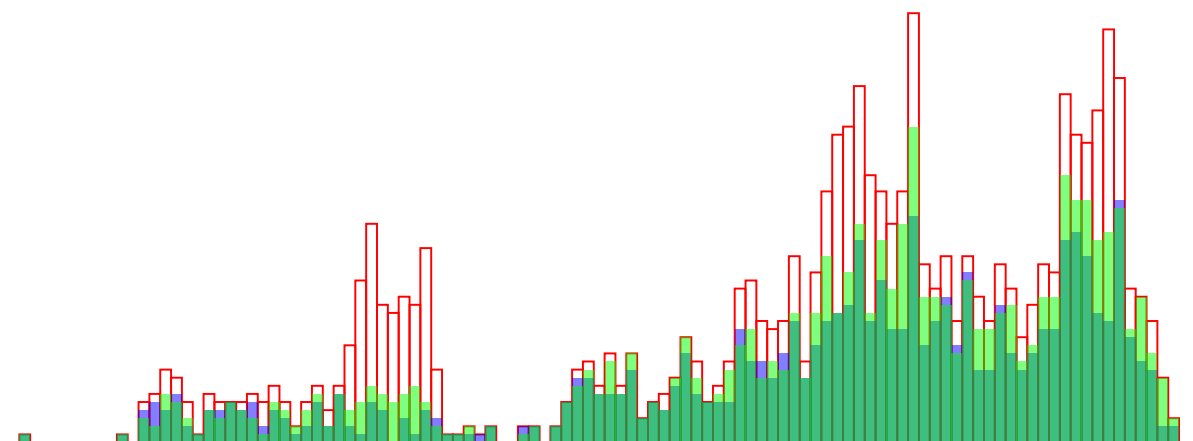


Warning: Removed 2 rows containing missing values (geom_bar).

Warning: Removed 2 rows containing missing values (geom_bar).

Warning: Removed 2 rows containing missing values (geom_bar).

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in formations (green)



Saving 10 x 5 in image

Warning: Removed 2 rows containing missing values (geom_bar).

→ Comparing with the initial occurrences distribution and with the distribution after our first sub-sampling it appears that both methods lead to very similar distributions. This confornts us about the robustness of those approaches.

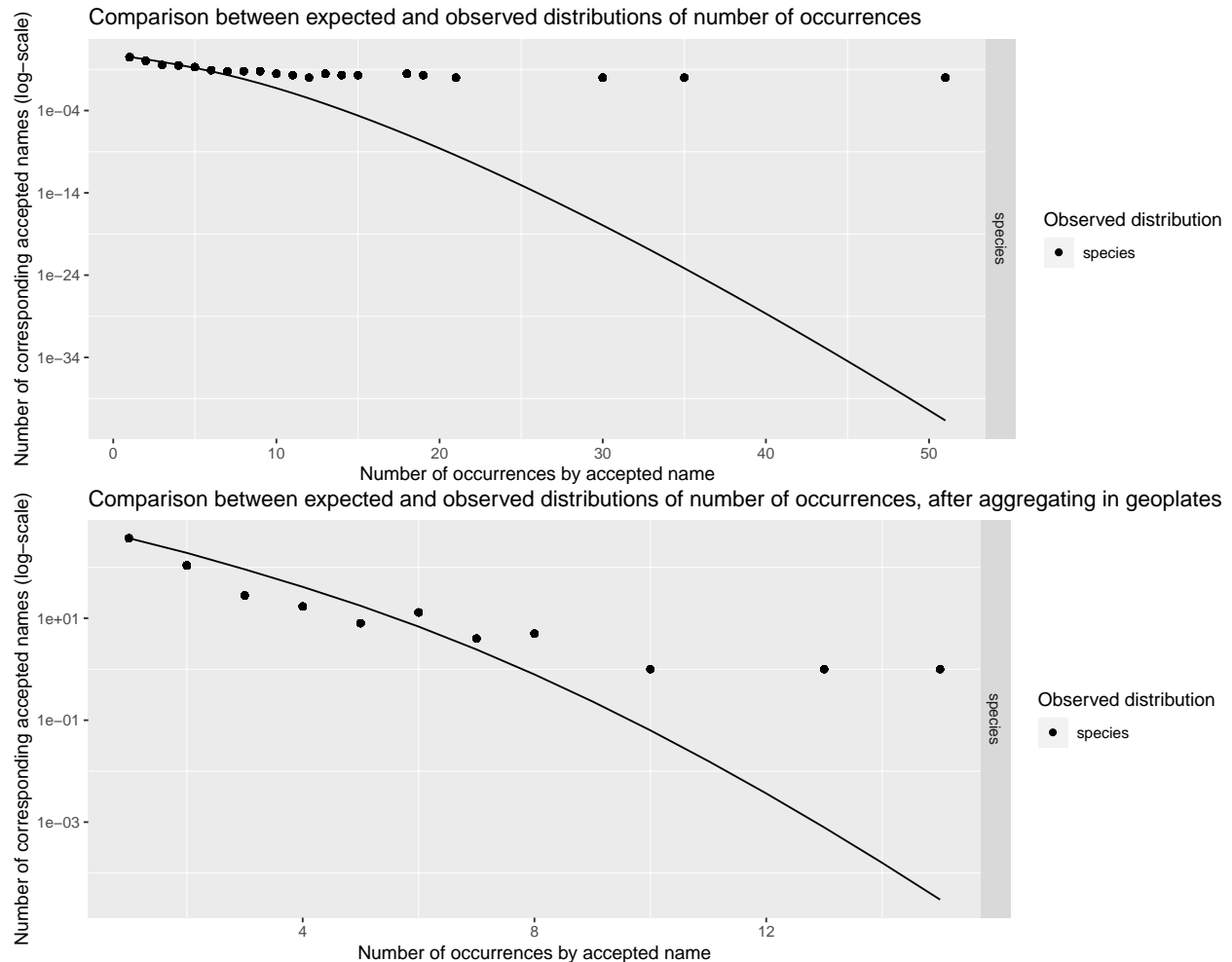
Replacing coutries by geological plates seems to make more sens from a palaeontological perspective, so let's

try it.

Aggregate occurrences without formation by geoplate + early interval

```
##                               Cetacea_occ Cetacea_occ_aggreg removed
## Number of occurrences                3807                2606    1201
## Number of occurrences (species only)   1436                966    470
```

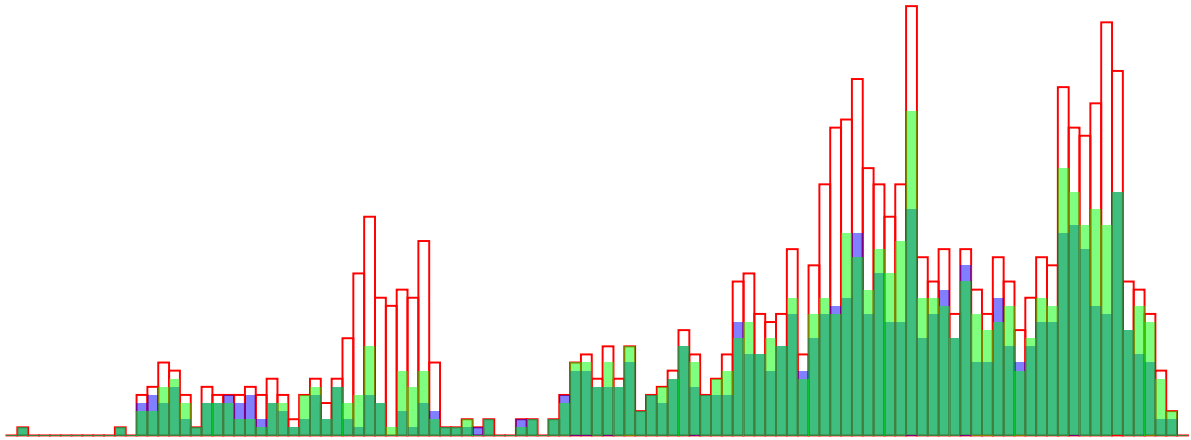
More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution :



→ The correspondance is still good, except for two taxa :

```
##
## Scaldicetus grandis Schizodelphis sulcatus
##                13                15
## Warning: Removed 2 rows containing missing values (geom_bar).
## Warning: Removed 2 rows containing missing values (geom_bar).
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in geoplates (green)

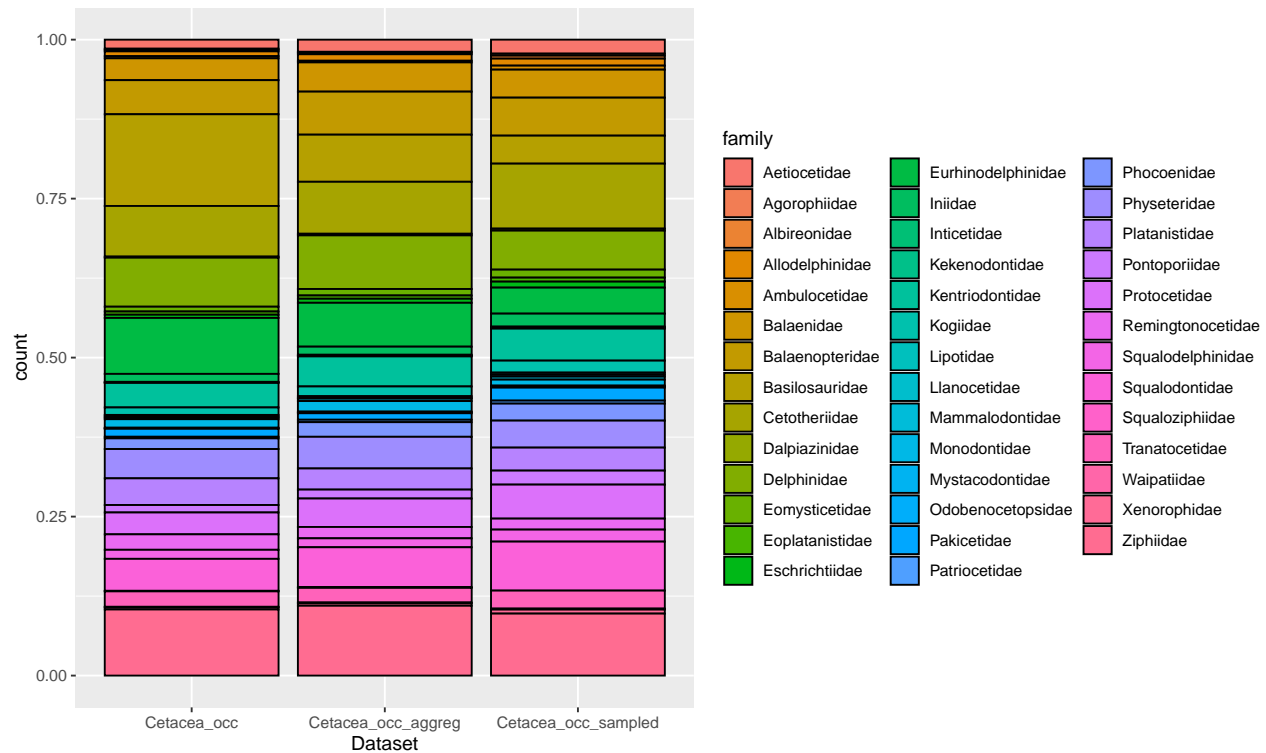


→ Delimiting by geological plates (+ age) instead of countries (+ age) leads to similar distributions, so we will keep it.

Check that the sampling methods do not introduce biases in the repartition between Odontoceti and Mysticeti

Mystecetes are usually larger than odontocetes, and size is associated with a wider geographic range so since we are subsampling occurrences according to geological formation we may be biasing our data towards more widespread species, therefore towards mystecetes.

Look at the families first :



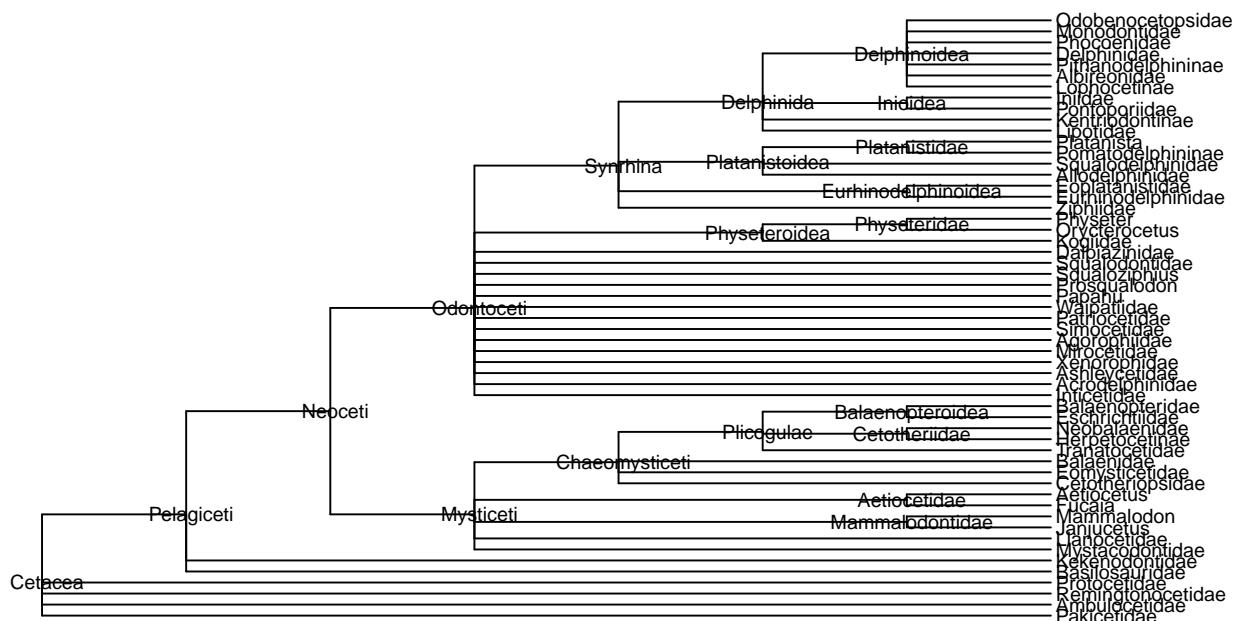
→ The family repartitions vary a bit after subsampling, but because of the limited number of species by family the fact that we corrected the oversampling of some species could have a disproportionate effect.

Let's look rather at a higher taxonomic rank, by importing the topology of cetacean families (from Marx et al. 2016) :

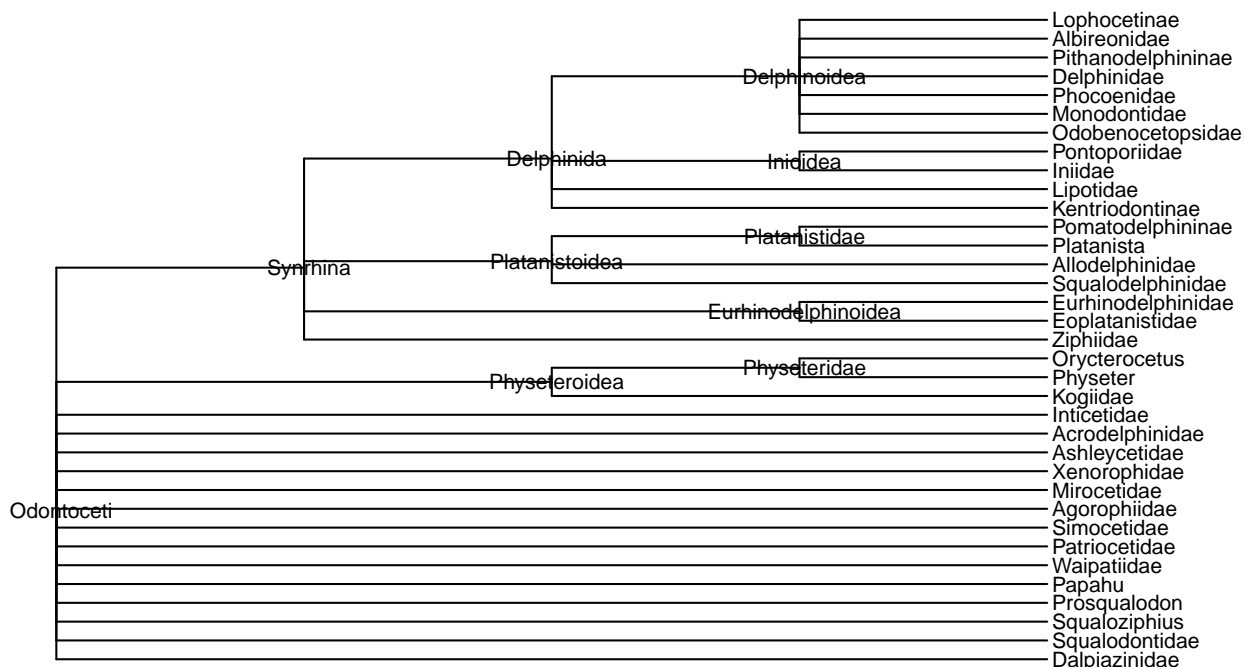
```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## Warning: `mutate_()` is deprecated as of dplyr 0.7.0.
## Please use `mutate()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

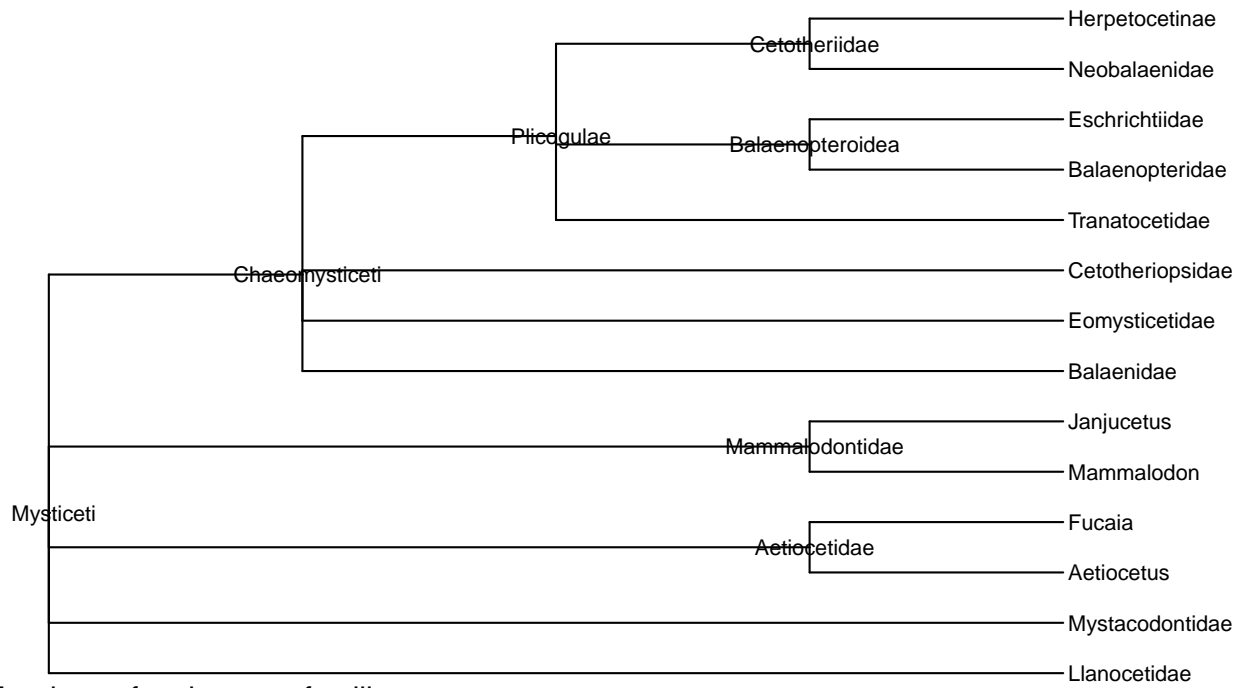

Topology of cetacean families



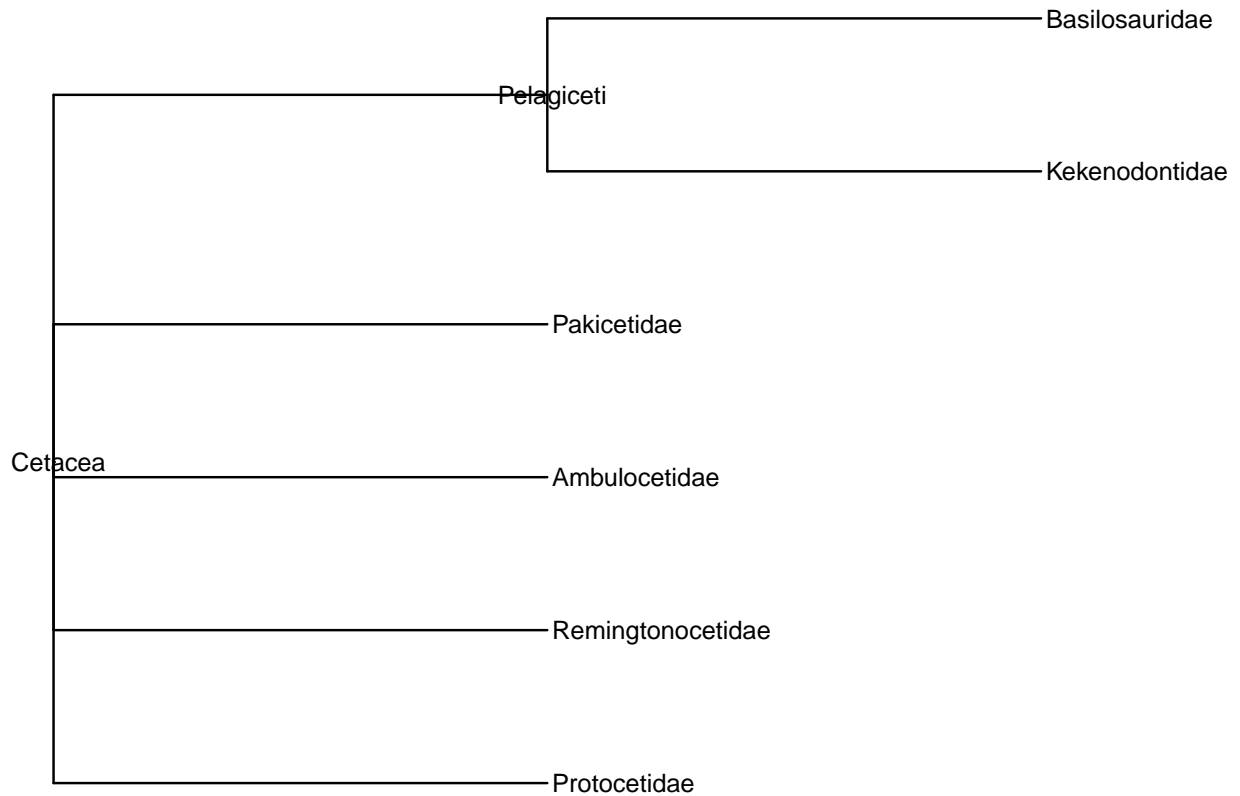
Topology of odontocete families

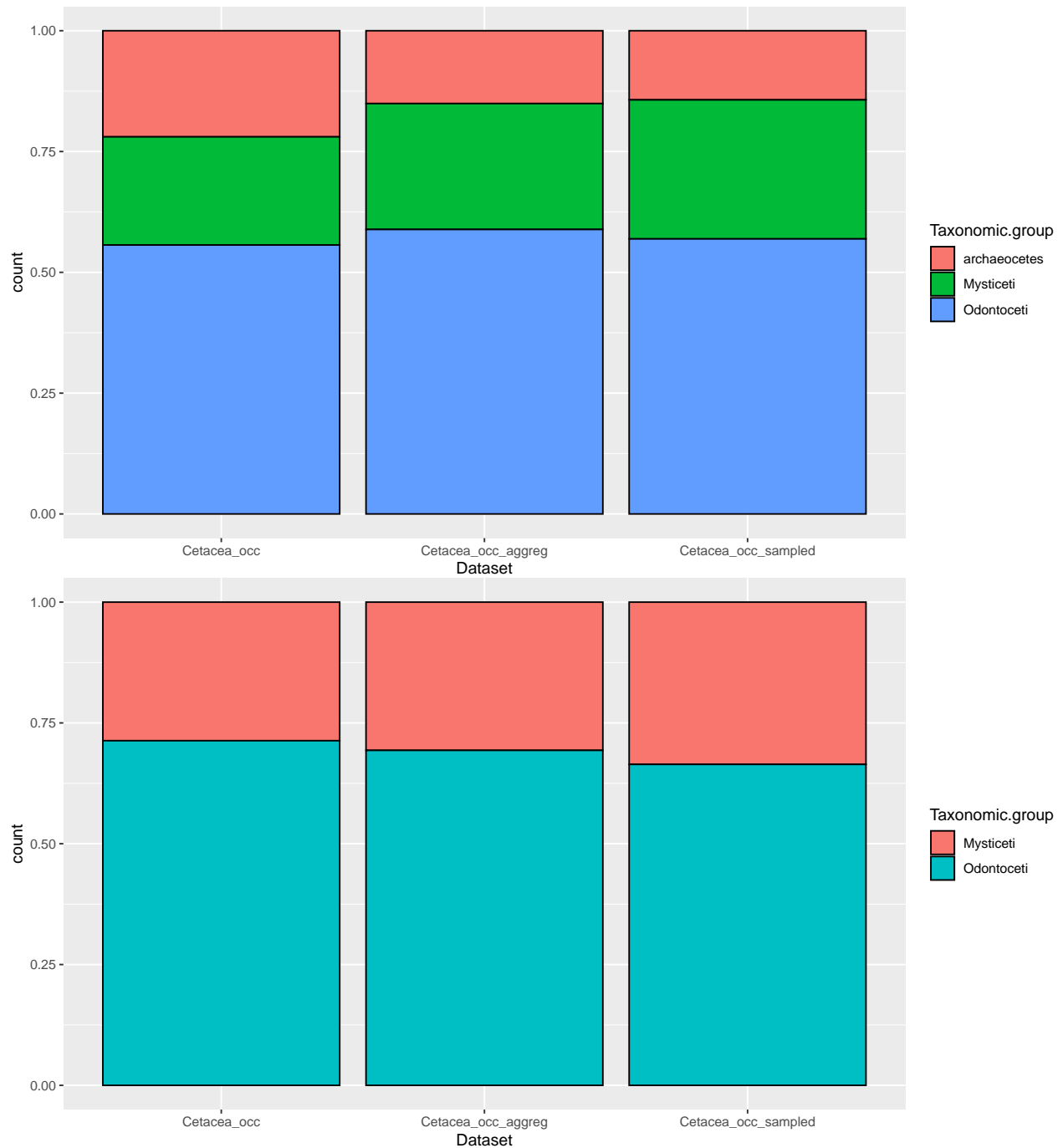


Topology of mysticete families



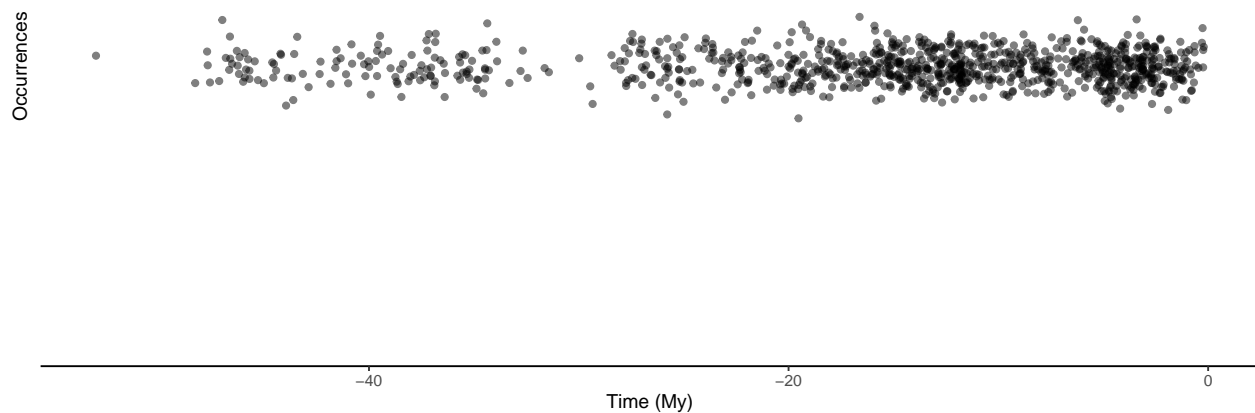
Topology of archeocete families





→ There is a smaller proportion of archeocete occurrences after either sampling, because a huge cluster is subsampled around 35 My ago, but this effect is expected. However, for the Mysticeti vs. Odontoceti there is no huge apparent bias, especially with the aggregating method.

Repartition of 4609 recorded occurrences through time

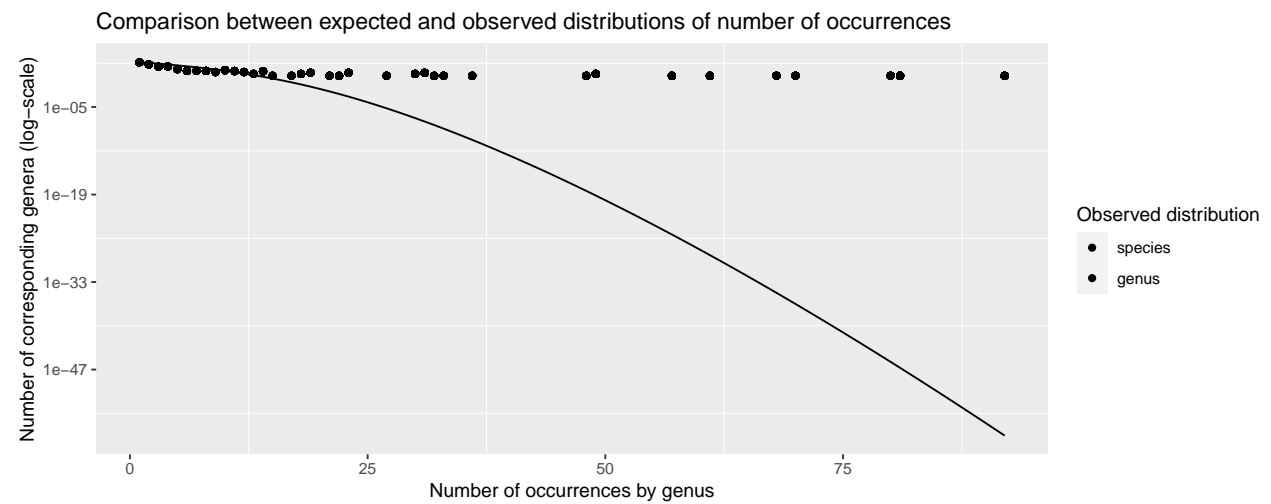


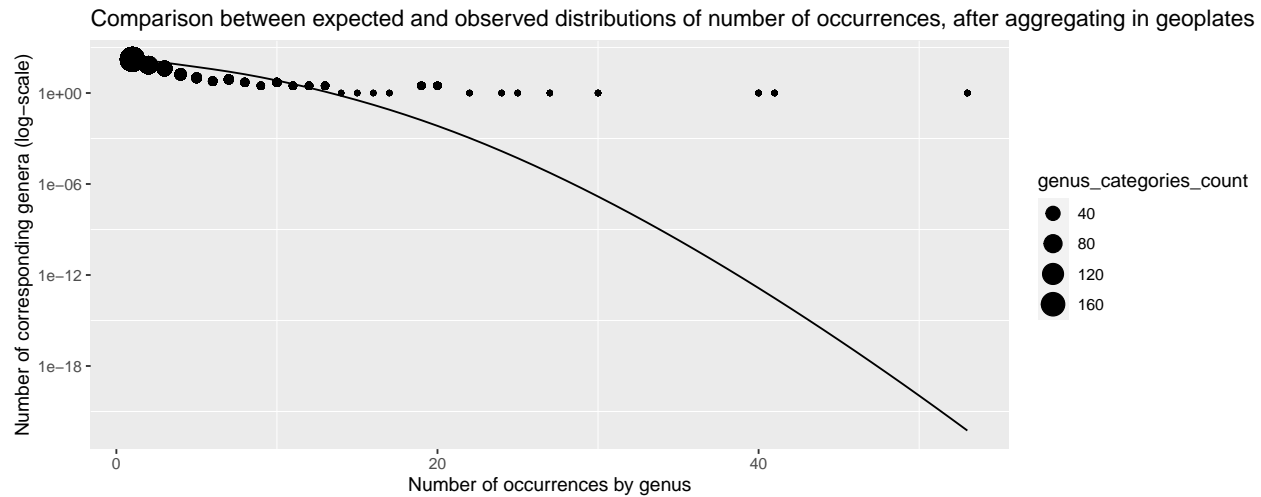
Faster genus-level analysis

Apply the same transformations at the generic level to perform the analysis faster.

```
##                               Cetacea_occ
## Number of occurrences          3807
## Number of occurrences (species and genera only) 2163
##                               Cetacea_occ_aggreg_gen removed
## Number of occurrences          1839    1968
## Number of occurrences (species and genera only) 1301    862
```

More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution :





→ The correspondance is much less improved than with species aggregation because some genera have a lot of occurrences due to their high number of species species :

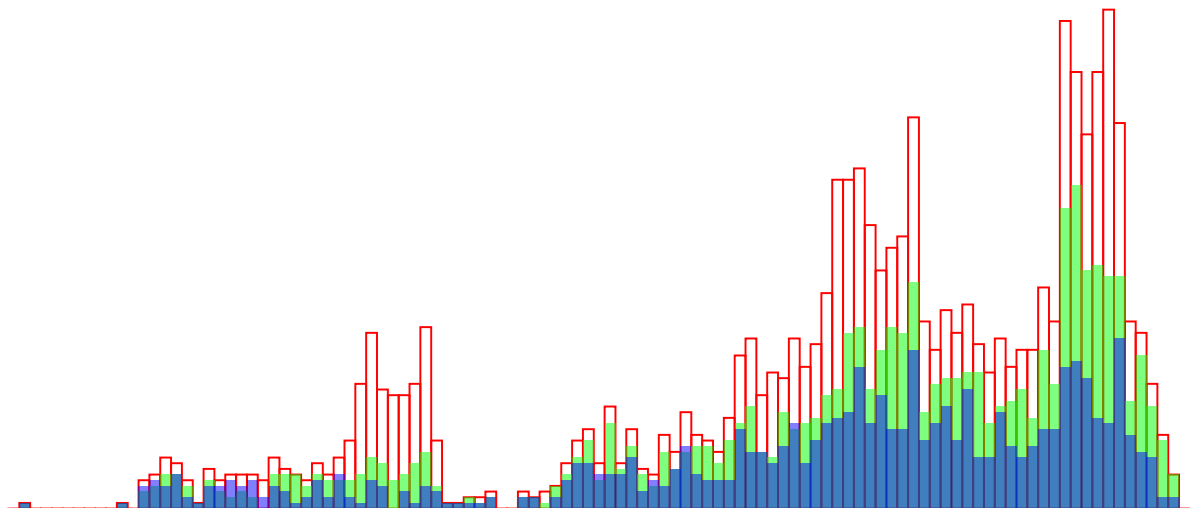
```
##
## Balaenoptera Kentriodon Mesoplodon Scaldicetus Squalodon
##          53          27          30          41          40
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in geoplates (green)



→ Delimiting by geological plates (+ age) instead of countries (+ age) leads to similar distributions, with a bit more occurrences.

Conclusions

- It seems possible to adequately reduce the abundance bias by subsampling the most concentrated intervals, based on a metrics of occurrence density (number of occurrences / time range)
- An alternative sampling method consisting in aggregating the occurrences of the same taxa found in the same geological formation (or the same geoplate if the geological formation is not indicated) seems to work as well, and is more paleontologically relevant
- Using combined ranges by species appears to be more robust
- 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