# Format Cetaceans Data

## Contents



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

## Load occurrence dataset

```
##  occurrence_no  record_type    reid_no     collection_no
##  1002162:   1   occ:4678    11942  :   1   52582  :  62
##  1002163:   1               11943  :   1   47465  :  52
##  1002167:   1               12051  :   1   48093  :  42
##  1002169:   1               12057  :   1   75152  :  41
##  1002202:   1               12058  :   1   47428  :  29
##  1002203:   1               (Other): 286   48887  :  27
##  (Other):4672               NA's   :4387   (Other):4425
##                identified_name    identified_rank identified_no
##  Cetacea indet.        : 280   species    :2525   36652  : 256
##  Mysticeti indet.      : 147   genus      : 726   42971  : 156
##  Odontoceti indet.     : 128   family     : 689   42937  : 128
```

```
##  Delphinidae indet.    : 102   suborder   : 287   42953  : 110
##  Balaenopteridae indet.: 101   order      : 285   42977  : 105
##  Balaena mysticetus    :  82   superfamily:  74   64683  :  87
##  (Other)               :3838   (Other)    :  92   (Other):3836
##                  difference              accepted_name      accepted_rank
##                       :3757   Cetacea           : 301   species    :2155
##  recombined as        : 312   Odontoceti        : 242   family     : 804
##  nomen dubium         : 296   Mysticeti         : 213   genus      : 787
##  subjective synonym of: 120   Balaenopteridae   : 169   suborder   : 455
##  invalid subgroup of  :  50   Delphinidae       : 113   order      : 301
##  corrected to         :  32   Balaena mysticetus:  88   superfamily:  78
##  (Other)              : 111   (Other)           :3552   (Other)    :  98
##   accepted_no       early_interval       late_interval      max_ma
##  36652  : 301   Holocene   : 659               :4128   Min.   : 0.0117
##  42937  : 242   Langhian   : 353   Langhian     :  81   1st Qu.: 3.6000
##  42971  : 213   Burdigalian: 340   Late Pliocene:  80   Median :11.6200
##  42977  : 169   Zanclean   : 317   Messinian    :  76   Mean   :13.8392
##  42953  : 113   Tortonian  : 307   Serravallian :  52   3rd Qu.:20.4400
##  64683  :  88   Priabonian : 227   Tortonian    :  29   Max.   :66.0000
##  (Other):3552   (Other)    :2475   (Other)      : 232
##      min_ma                    ref_author     ref_pubyr     reference_no
##  Min.   : 0.000   Uhen              : 230   Min.   :1818   12148  :  98
##  1st Qu.: 2.588   Bianucci et al.   : 122   1st Qu.:1975   24251  :  95
##  Median : 5.333   Pilleri           : 122   Median :1998   13672  :  74
##  Mean   :10.058   Oishi and Hasegawa: 104   Mean   :1983   19106  :  68
##  3rd Qu.:13.820   Fitzgerald        :  98   3rd Qu.:2008   32630  :  61
##  Max.   :47.800   Barnes            :  87   Max.   :2500   12161  :  60
##                   (Other)           :3915                  (Other):4222
##       phylum     phylum_no         class       class_no        order
##  Chordata:4678   33815:4678   Mammalia:4678   36651:4678   Cetacea:4678
##
##
##
##
##
##
##   order_no                    family       family_no             genus
##  36652:4678                      : 837          : 837               :1701
##                Delphinidae       : 544   42953  : 544   Balaenoptera: 133
##                Balaenopteridae   : 399   42977  : 399   Balaena     : 116
##                NO_FAMILY_SPECIFIED: 382   NF     : 382   Mesoplodon  : 101
##                Ziphiidae         : 333   42951  : 333   Squalodon   :  93
##                Basilosauridae    : 279   42936  : 279   Basilosaurus:  81
##                (Other)           :1904   (Other):1904   (Other)     :2453
##     genus_no     abund_value            abund_unit        lng
##         :1701   Min.   :  1.000             :2788   Min.   :-179.983
##  36678  : 133   1st Qu.:  1.000   elements  :   5   1st Qu.: -76.800
##  36676  : 116   Median :  1.000   fragments :   1   Median :   4.433
##  36763  : 101   Mean   :  2.527   individuals:  52   Mean   :  -4.137
##  36852  :  93   3rd Qu.:  1.000   specimens :1832   3rd Qu.:  26.459
##  36681  :  81   Max.   :537.000                     Max.   : 179.533
##  (Other):2453   NA's   :2786
##       lat                                  collection_name collection_subset
##  Min.   :-70.75   Whaling Survey                   : 256   Min.   :   20
```

```
##  1st Qu.: 25.18   Lee Creek Mine, Yorktown Formation:  62   1st Qu.:18577
##  Median : 37.08   Anvers                            :  52   Median :55621
##  Mean   : 26.34   Felixstow                         :  42   Mean   :39910
##  3rd Qu.: 45.28   Zwilbroek                         :  41   3rd Qu.:59059
##  Max.   : 82.27   Orciano Pisano                    :  29   Max.   :78523
##                   (Other)                           :4196   NA's   :4650
##                              collection_aka       cc
##                                     :3285   US     :1105
##  Crag d'Anvers                      :  52   IT     : 380
##  Felixstowe                         :  43   JP     : 365
##  Wiegerink, Zwillbrock              :  41   BE     : 241
##  Oberschwaben, Wurttemberg          :  25   NZ     : 167
##  Fort 4, Mortsel, Oude God, Oude-God:  15   (Other):2419
##  (Other)                            :1217   NA's   :   1
##             state               county                          latlng_basis
##                :1909               :3358                                : 211
##  California    : 271   Calvert     : 108   based on nearby landmark: 197
##  Maryland      : 165   Beaufort    :  93   based on political unit : 174
##  Antwerpen     : 158   Westmoreland:  80   estimated from map      :3104
##  Virginia      : 141   San Diego   :  60   stated in text          : 947
##  North Carolina: 130   Suffolk     :  52   unpublished field data  :  45
##  (Other)       :1904   (Other)     : 927
##  latlng_precision            geogscale     paleomodel      paleolng
##  seconds:3372                     :3661         : 215   Min.   :-178.760
##  minutes: 471     basin           :  47   gp_mid:4463   1st Qu.: -69.820
##  6      : 286     hand sample     :  13                 Median :   5.560
##  1      : 122     local area      : 388                 Mean   :  -3.457
##  5      : 107     outcrop         : 469                 3rd Qu.:  22.328
##  3      : 103     small collection: 100                 Max.   : 179.610
##  (Other): 217                                           NA's   :614
##     paleolat                                             geoplate
##  Min.   :-70.75   109                                         : 651
##  1st Qu.: 18.98   315                                         : 537
##  Median : 38.41   coordinates not computable using this model: 399
##  Mean   : 25.77   307                                         : 360
##  3rd Qu.: 44.53   610                                         : 335
##  Max.   : 82.27                                               : 215
##  NA's   :614      (Other)                                     :2181
##       cc.1      protected     formation         stratgroup
##  US     :1064      :4586           :2219             :4056
##  IT     : 379   FED:  78   Calvert : 240   Chesapeake: 365
##  JP     : 361   NPS:  14   Pisco   :  85   Jackson   :  58
##         : 247              Yorktown:  81   Hawthorn  :  42
##  BE     : 238              Berchem :  48   Cooper    :  17
##  (Other):2388              Red Crag:  48   Ashiya    :  15
##  NA's   :   1              (Other) :1957   (Other)   : 125
##              member             stratscale            zone
##                 :3823                :3062              :4516
##  Plum Point     : 127   bed          : 250   MPL4b/MPL5a:  34
##  Calvert Beach  :  79   formation    : 604   N18-N19    :  12
##  Eibergen       :  43   group        :  21   MN 17      :   9
##  Bone Valley    :  42   group of beds: 324   MN 3       :   8
##  Antwerpen Sands:  34   member       : 417   MN 14      :   7
##  (Other)        : 530                        (Other)    :  92
```

```
##        localsection     localbed                   localorder
##               :4584           :4593                       :4594
##  Pisco        :  39   2       :  17   bottom to top      :  39
##  Krkwd        :  11   AGL     :  14   no particular order:  43
##  SDgFm        :   7   SAO     :  12   top to bottom      :   2
##  Fan Delta    :   4   1       :   7
##  Langebaanweg:   4   MTM     :   7
##  (Other)      :  29   (Other):  28
##       regionalsection    regionalbed         regionalorder
##                :4434            :4439                 :4441
##  Blankenhorn  :   2   14       :  45   bottom to top: 237
##  Panama Canal :   1   12       :  36
##  Santa Barbara:   1   10       :  20
##  Shattuck     : 237   13       :  16
##  Siwa Oasis   :   1   d\x8ec-13:  15
##  Waihao River :   2   (Other)  : 107
##
##
##  bluish clayey sand\\r\\n"The Yorktown Formation at Lee Creek... is generally a very muddy quartz sa
##  shelf mudstone
##  no lithological description given
##  upper marine molasse
##  brownish sandy clay
##  (Other)
##         lithology1           lithadj1             lithification1
##  not reported:1628               :3853                    :4320
##               : 904   blue       : 119   lithified       :  88
##  sandstone    : 888   glauconitic :  75   poorly lithified:  95
##  claystone    : 340   diatomaceous:  52   unlithified     : 175
##  "limestone"  : 186   phosphatic  :  40
##  siltstone    : 157   brown       :  32
##  (Other)      : 575   (Other)     : 507
##            minor_lithology1 fossilsfrom1        lithology2
##                     :4238     :2706                  :4473
##  sandy              : 171    Y:1972       sandstone   :  68
##  argillaceous,muddy:  62                 siltstone   :  62
##  silty              :  56                "limestone" :  17
##  argillaceous       :  48                conglomerate:  14
##  calcareous         :  42                "shale"     :  12
##  (Other)            :  61                (Other)     :  32
##                               lithadj2             lithification2
##                                   :4620                    :4647
##  tuffaceous,gray                  :  15   lithified       :  11
##  yellow                           :  14   poorly lithified:  11
##  black,brown                      :   5   unlithified     :   9
##  fine                             :   5
##  lenticular,"cross stratification":   2
##  (Other)                          :  17
##                 minor_lithology2 fossilsfrom2         environment
##                         :4659     :4617       marine indet. :2005
##  calcareous             :   2    Y:  61                     : 596
##  calcareous,carbonaceous:   5                coastal indet.: 515
##  sandy                  :   7                estuary/bay   : 208
##  silty                  :   2                foreshore     : 131
```

```
##  silty,sandy            :   3                 offshore shelf: 124
##                                               (Other)       :1099
##         tectonic_setting                                  assembl_comps
##                 :4630                                            :1269
##  cratonic basin  :   7    macrofossils                            :3220
##  foreland basin  :   7    macrofossils,mesofossils                :  69
##  passive margin  :  32    macrofossils,mesofossils,microfossils: 101
##  pull-apart basin:   2    macrofossils,microfossils               :  12
##                           mesofossils                             :   4
##                           mesofossils,microfossils                :   3
##  articulated_parts associated_parts
##      :4583             :4623
##  many:  14         many:   5
##  none:  62         none:  11
##  some:  19         some:  39
##
##
##
##                                                        common_body_parts
##                                                                 :4659
##  partial skeletons,teeth,vertebrae,limb elements,plant debris:   4
##  partial skeletons                                              :   3
##  mandibles                                                      :   2
##  partial skulls                                                 :   2
##  shells                                                         :   2
##  (Other)                                                        :   6
##  rare_body_parts            feed_pred_traces
##        :4674                       :4634
##  valves:   4     drill holes       :   1
##                  gastric dissolution:   4
##                  tooth marks       :  39
##
##
##
##                                                          artifacts
##                                                               :4657
##  stone points,stone tools                                     :   8
##  stone points,stone tools,bone tools,textiles                 :   6
##  stone tools,ceramics,structural remains,historical artifacts:   2
##  charcoal/hearths                                             :   1
##  historical artifacts                                         :   1
##  (Other)                                                      :   3
##                                  pres_mode    preservation_quality
##  body                                 :1810           :4424
##  body,original phosphate              :1446   excellent:  89
##                                       :1254   good     :  92
##  body,original phosphate,anthropogenic:  26   medium   :  42
##  body,anthropogenic                   :  25   poor     :  17
##  body,soft parts,original phosphate   :  18   variable :  14
##  (Other)                              :  99
##         spatial_resolution    temporal_resolution       lagerstatten
##                 :4647                 :4651                 :4672
##  allochthonous   :   3     condensed   :   1     conservation:   6
##  autochthonous   :  20     snapshot    :   9
```

```
##  parautochthonous:   8     time-averaged:  17
##
##
##
##    concentration         orientation   abund_in_sediment    fragmentation
##           :4675                :4669           :4645                :4661
##  #NOM?    :   2   life position:   1   abundant:   6     frequent  :   7
##  dispersed:   1   preferred    :   6   common  :  22     occasional:  10
##                   random       :   2   rare    :   5
##
##
##
##       bioerosion       encrustation               collection_type
##            :4675             :4674                        : 750
##  none      :   1   frequent  :   2   archaeological       :  70
##  occasional:   2   occasional:   2   biostratigraphic     :  55
##                                      general faunal/floral:2656
##                                      paleoecologic        :  25
##                                      taphonomic           :  38
##                                      taxonomic            :1084
##                                           collection_methods       museum
##                                                     :3911              :4447
##  surface (in situ),field collection                : 157     UCMP     :  69
##  field collection                                   : 138     USNM     :  53
##  bulk,salvage,surface (float),sieve                 :  62     LACM     :  49
##  surface (float),surface (in situ),field collection:  53     LACM,UCMP:   8
##  surface (float),field collection                   :  38     FLMNH    :   7
##  (Other)                                            : 319     (Other)  :  45
##                           collection_coverage      collection_size
##                                      :4501                  :4603
##  some genera                         :  45     1 specimens  :  36
##  some macrofossils                   :  40     0            :  11
##  some macrofossils,some microfossils:  20     1 individuals:   6
##  some microfossils                   :  16     185 specimens:   3
##  all macrofossils,some genera        :  14     2 specimens  :   3
##  (Other)                             :  42     (Other)      :  16
##  rock_censused                                          collectors
##         :4674                                                :4642
##  2000 kg:   3   Bowman                                      :   6
##  5 kg   :   1   Dubalen                                     :   5
##                 H. Lodge                                    :   4
##                 Fordyce, Rust, A. Grebneff, and S. Wilson,:   3
##                 Marasti                                     :   2
##                 (Other)                                     :  16
##       collection_dates              taxon_environment            motility
##              :4661                          : 774                   : 774
##  2014-2016   :   5    freshwater            :  17    actively mobile:3904
##  December 1998:   3    freshwater,terrestrial:   5
##  1899        :   1    marine                :1749
##  1969        :   1    marine,freshwater     :2067
##  1987        :   1    oceanic               :  66
##  (Other)     :   6
##                     life_habit                              diet
##                          :3424                                : 774
```

```
##  amphibious, depth=surface:   5   carnivore                   :1604
##  aquatic                   : 666   carnivore, suspension feeder: 593
##  aquatic, depth=surface    : 583   piscivore                   : 221
##                                    piscivore, carnivore        : 903
##                                    suspension feeder           : 583
##
##      reproduction                    ontogeny            composition
##            : 774                         : 774                 : 774
##  viviparous:3904   modification of parts:3904   hydroxyapatite:3899
##                                                phosphatic    :   5
##
##
##
##
```

```
##    occurrence_no record_type reid_no collection_no
## 1          68135         occ    <NA>          4868
## 2         137494         occ    <NA>         11601
## 3         141404         occ    <NA>         12121
## 4         147937         occ    <NA>         13063
## 5         147938         occ    <NA>         13064
## 6         148079         occ    <NA>         13078
## 7         148335         occ    <NA>         13090
## 8         148353         occ    <NA>         13092
## 9         148356         occ    <NA>         13096
## 10        148358         occ    <NA>         13098
##                            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
##             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
##    early_interval  late_interval max_ma min_ma             ref_author
## 1        Lutetian                47.800 41.300          Hulbert et al.
## 2        Chattian                28.100 23.030                 Barnes
## 3        Chattian                28.100 23.030  Mchedlidze and Pilleri
## 4        Langhian                15.970 13.820 Czyzewska and Ryziewicz
## 5       Bartonian                41.300 38.000        Gingerich et al.
```

```
## 6   Early Miocene Middle Miocene 23.030 11.608                Dubrovo
## 7       Messinian                 7.246  5.333                 Barnes
## 8       Messinian                 7.246  5.333                 Gondar
## 9        Chattian                28.100 23.030     Dubrovo and Sharkov
## 10      Tortonian                11.620  7.246                 Kellogg
##    ref_pubyr reference_no   phylum phylum_no    class class_no   order order_no
## 1       1998          289 Chordata     33815 Mammalia    36651 Cetacea    36652
## 2       1979         4175 Chordata     33815 Mammalia    36651 Cetacea    36652
## 3       1988         6018 Chordata     33815 Mammalia    36651 Cetacea    36652
## 4       1976         4344 Chordata     33815 Mammalia    36651 Cetacea    36652
## 5       1997         6010 Chordata     33815 Mammalia    36651 Cetacea    36652
## 6       1971         4357 Chordata     33815 Mammalia    36651 Cetacea    36652
## 7       1973         4361 Chordata     33815 Mammalia    36651 Cetacea    36652
## 8       1974         4362 Chordata     33815 Mammalia    36651 Cetacea    36652
## 9       1971         4365 Chordata     33815 Mammalia    36651 Cetacea    36652
## 10      1934        10152 Chordata     33815 Mammalia    36651 Cetacea    36652
##                 family family_no           genus genus_no abund_value
## 1         Protocetidae     42934    Georgiacetus    36720          NA
## 2  NO_FAMILY_SPECIFIED        NF     Argyrocetus    36668           1
## 3  NO_FAMILY_SPECIFIED        NF Kharthlidelphis    53160           2
## 4  NO_FAMILY_SPECIFIED        NF       Pinocetus    36810          NA
## 5       Basilosauridae     42936     Basiloterus    53164          NA
## 6  NO_FAMILY_SPECIFIED        NF  Sachalinocetus    36843          NA
## 7             Kogiidae     53256       Praekogia    36824          NA
## 8                                                                  NA
## 9  NO_FAMILY_SPECIFIED        NF      Microcetus    36767           1
## 10      Tranatocetidae    328006       Mixocetus    36774          NA
##    abund_unit        lng       lat occurrence_comments
## 1              -81.76056  33.14333
## 2   specimens -118.84834  35.49278
## 3   specimens   43.54833  42.53278
## 4               20.52639  50.52028
## 5               70.44056  30.78778
## 6              142.05000  47.05000
## 7             -115.18333  28.36667
## 8              -64.93333 -40.73333
## 9   specimens   51.43195  43.81944
## 10            -118.19945  34.08361
##                       collection_name collection_subset
## 1    Vogtle Electric Generating Plant               NA
## 2  Pyramid Hill Sand Member grit zone               NA
## 3                      Cedisi Village               NA
## 4                           Nowa Wies               NA
## 5                         Bari Nadi 3               NA
## 6                            Sakhalin               NA
## 7                       Arroyo Delphin               NA
## 8                           Rio Negro               NA
## 9                            Karagiya               NA
## 10                     Lincoln Heights               NA
##                   collection_aka cc           state        county
## 1                                US         Georgia         Burke
## 2  LACMVP Loc. 1603, 1626, 1627
## 3                                GE        Caucasus Gori District
## 4                       Pinczow PL
```

```
## 5                          Ro2 PK          Punjab
## 6                              RU
## 7                  UCR RV-7315 MX Baja California
## 8                              AR
## 9                              KZ
## 10                             US      California   Los Angeles
##                latlng_basis latlng_precision        geogscale
## 1  based on nearby landmark          seconds small collection
## 2
## 3  based on nearby landmark          seconds          outcrop
## 4        estimated from map          seconds          outcrop
## 5            stated in text          seconds          outcrop
## 6   based on political unit                2       local area
## 7                                     minutes          outcrop
## 8   based on political unit          seconds       local area
## 9        estimated from map          seconds       local area
## 10       estimated from map          seconds          outcrop
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7  on the ridge west of "Arroyo Delphin" the first prominent drainage system to reach the shoreline r
## 8                                                                                  "El Sotano", Estancia de l
## 9        Western Kazakhstan, Mangyshlak Peninsula, western flank of the Karagiye depression\\r\\nminc
## 10
##    paleomodel paleolng paleolat                                      geoplate
## 1      gp_mid   -61.49    35.55                                           109
## 2                  NA       NA
## 3      gp_mid    44.11    37.53                                           511
## 4      gp_mid    22.03    47.50                                           305
## 5      gp_mid    72.15    11.74                                           501
## 6      gp_mid   138.33    47.14                                           610
## 7      gp_mid       NA       NA coordinates not computable using this model
## 8      gp_mid   -63.84   -40.48                                           291
## 9      gp_mid    51.39    38.98                                           402
## 10     gp_mid  -114.35    35.74                                           105
##    cc.1 protected         formation stratgroup                member
## 1    US                  Blue Bluff
## 2
## 3    GE
## 4    PL           Pińczów Limestone
## 5    PK                    Drazinda                      upper middle
## 6    RU
## 7    MX                     Almejas                             lower
## 8    AR
## 9    KZ                Karaginskaya                          Segendyk
## 10   US                      Modelo            Elysian Park Sandstone
##       stratscale  zone localsection localbed localorder regionalsection
## 1            bed NP 16
## 2
## 3
```

```
## 4                   M4
## 5  group of beds   P14
## 6
## 7      formation
## 8
## 9
## 10        member
##    regionalbed regionalorder
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
##
## 1
## 2
## 3
## 4  no formation or group listed, Badenian M4\\r\\n\\r\\nThe Pińczów Formation has been assigned to t
## 5
## 6
## 7
## 8
## 9
## 10                                                                                  It is not crystal clea
##                     lithdescript    lithology1           lithadj1
## 1  gray argillaceous calcilutite lime mudstone concretionary,gray
## 2
## 3
## 4                                  "limestone"
## 5
## 6                                 not reported
## 7            an ocre-yellow sand     sandstone             yellow
## 8                                 not reported
## 9                                  phosphorite
## 10                                   sandstone
##      lithification1 minor_lithology1 fossilsfrom1   lithology2 lithadj2
## 1  poorly lithified     argillaceous            Y
## 2
## 3
## 4                                               Y
## 5
## 6
## 7  poorly lithified                             Y conglomerate   pebbly
## 8
## 9                                               Y    siltstone
## 10                                              Y
##    lithification2 minor_lithology2 fossilsfrom2           environment
## 1                                                 deep subtidal shelf
## 2
```

```
## 3
## 4                                                 open shallow subtidal
## 5
## 6                                                         marine indet.
## 7                                                        coastal indet.
## 8                                                         marine indet.
## 9                                                         marine indet.
## 10                                                       coastal indet.
##    tectonic_setting
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
##
## 1  upper part of an 8 m bed that "was soft and poorly indurated" with "small concretions""relatively
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
##    assembl_comps articulated_parts associated_parts common_body_parts
## 1   macrofossils                                 many
## 2
## 3
## 4   macrofossils
## 5
## 6   macrofossils
## 7   macrofossils
## 8   macrofossils
## 9   macrofossils
## 10  macrofossils
##    rare_body_parts feed_pred_traces artifacts component_comments
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
##                       pres_mode preservation_quality spatial_resolution
## 1         body,original calcite
```

```
## 2
## 3
## 4       body,original phosphate                medium
## 5
## 6                           body
## 7       body,original phosphate                  good
## 8                           body
## 9   body,replaced with phosphate                medium
## 10      body,original phosphate
##    temporal_resolution lagerstatten concentration orientation abund_in_sediment
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
##    fragmentation bioerosion encrustation preservation_comments
## 1
## 2
## 3
## 4     occasional
## 5
## 6
## 7
## 8
## 9
## 10
##          collection_type
## 1              taxonomic
## 2
## 3
## 4              taxonomic
## 5
## 6   general faunal/floral
## 7              taxonomic
## 8   general faunal/floral
## 9              taxonomic
## 10 general faunal/floral
##                                                    collection_methods museum
## 1                                                    field collection
## 2
## 3
## 4                                                    field collection
## 5
## 6
## 7   selective quarrying,surface (in situ),mechanical,field collection
## 8
## 9
## 10                             selective quarrying,field collection
##       collection_coverage collection_size rock_censused collectors
```

```
## 1  difficult macrofossils
## 2
## 3
## 4                            1 specimens
## 5
## 6
## 7       some macrofossils
## 8
## 9       some macrofossils
## 10
##    collection_dates collection_comments taxonomy_comments taxon_environment
## 1                                                                    marine
## 2
## 3
## 4                                                                    marine
## 5
## 6                                                          marine,freshwater
## 7                                                          marine,freshwater
## 8                                                          marine,freshwater
## 9                                                          marine,freshwater
## 10                                                                   marine
##           motility             life_habit                       diet
## 1  actively mobile aquatic, depth=surface                  carnivore
## 2
## 3
## 4  actively mobile               aquatic carnivore, suspension feeder
## 5
## 6  actively mobile                                          carnivore
## 7  actively mobile                                          carnivore
## 8  actively mobile                                          carnivore
## 9  actively mobile                                          carnivore
## 10 actively mobile               aquatic carnivore, suspension feeder
##    reproduction              ontogeny    composition
## 1    viviparous modification of parts hydroxyapatite
## 2
## 3
## 4    viviparous modification of parts hydroxyapatite
## 5
## 6    viviparous modification of parts hydroxyapatite
## 7    viviparous modification of parts hydroxyapatite
## 8    viviparous modification of parts hydroxyapatite
## 9    viviparous modification of parts hydroxyapatite
## 10   viviparous modification of parts hydroxyapatite
```

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

```
##                      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 refe
##  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 collecti
##  two specimens in pbdb, unlcear 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
##  Ashleycetus_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
```

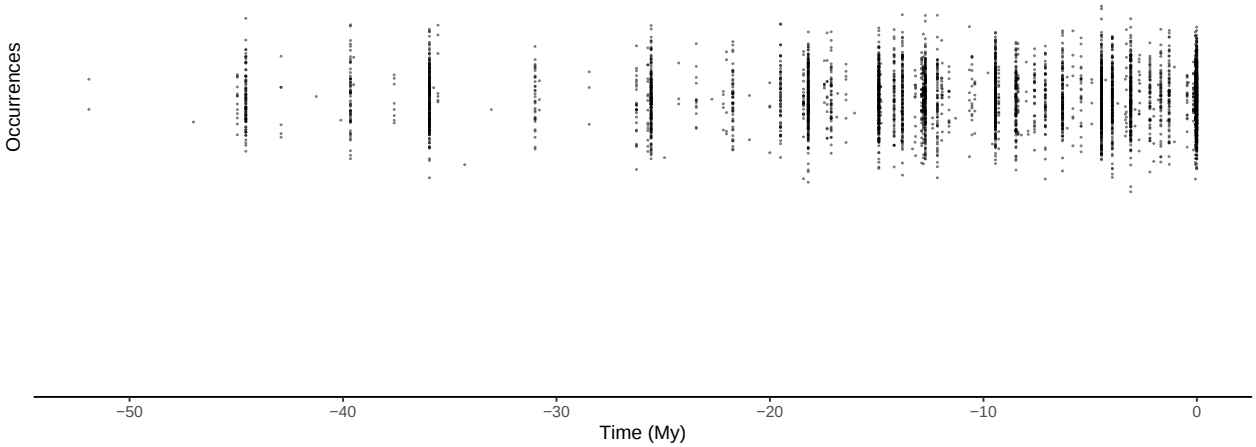
Remove those occurrences from our initial dataset to avoid redundancy.

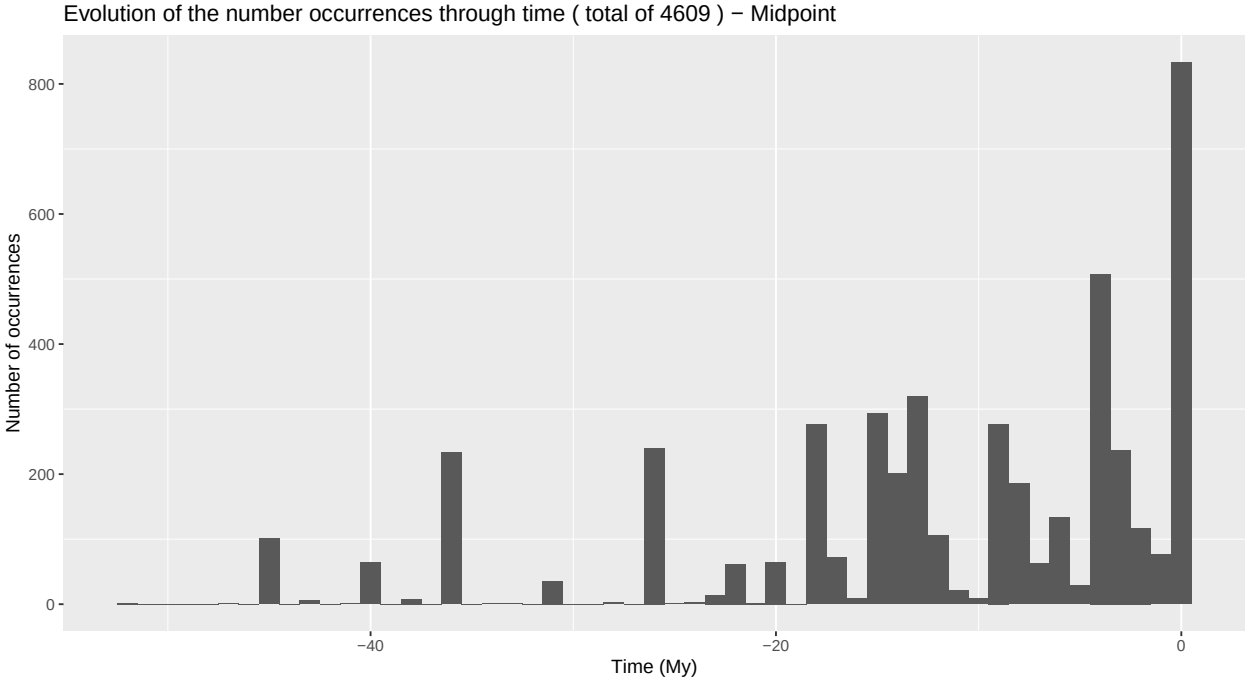
# Explore the dataset

## Repartition through time

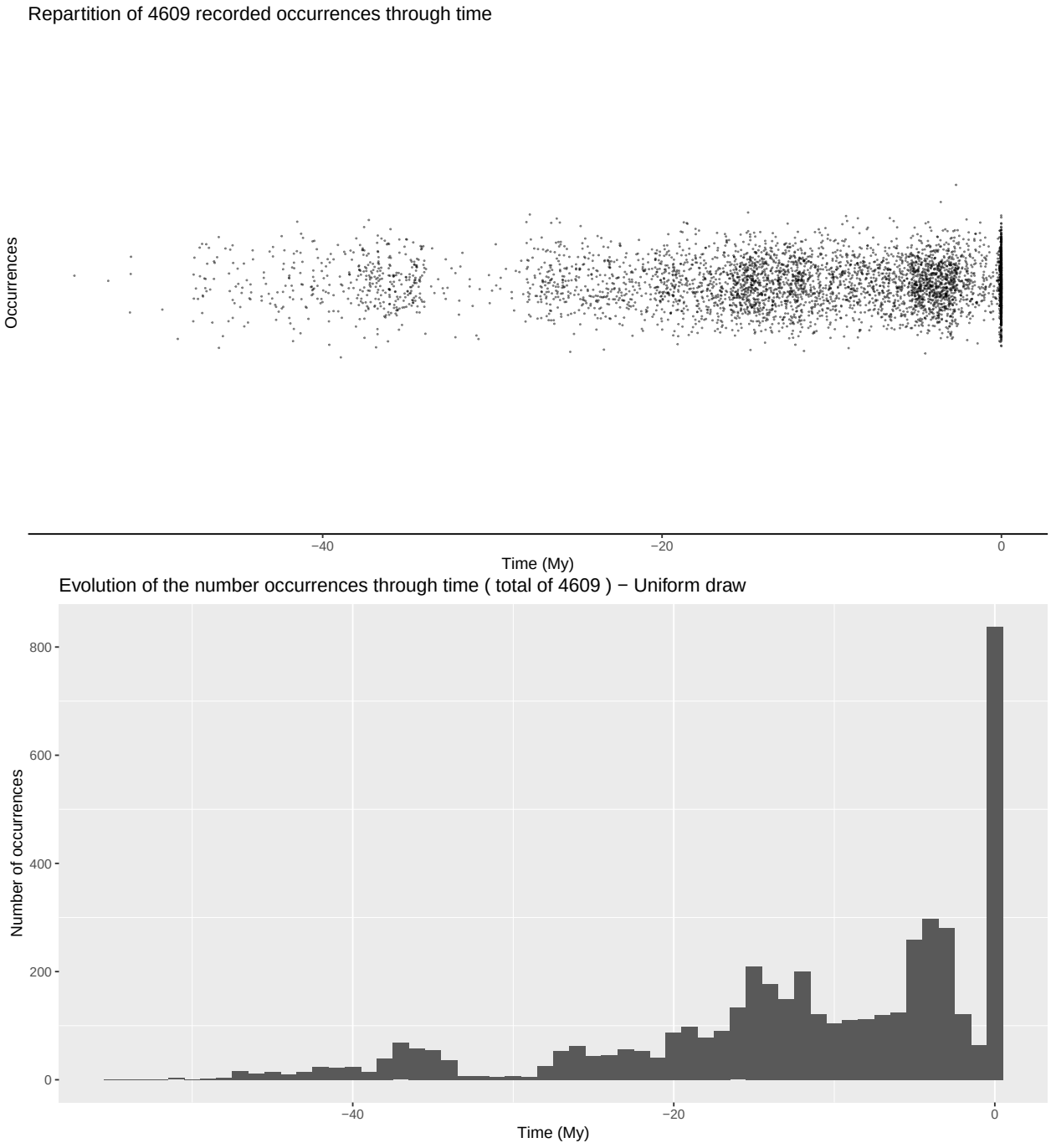
### Full fossil record

Repartition of 4609 recorded occurrences through time



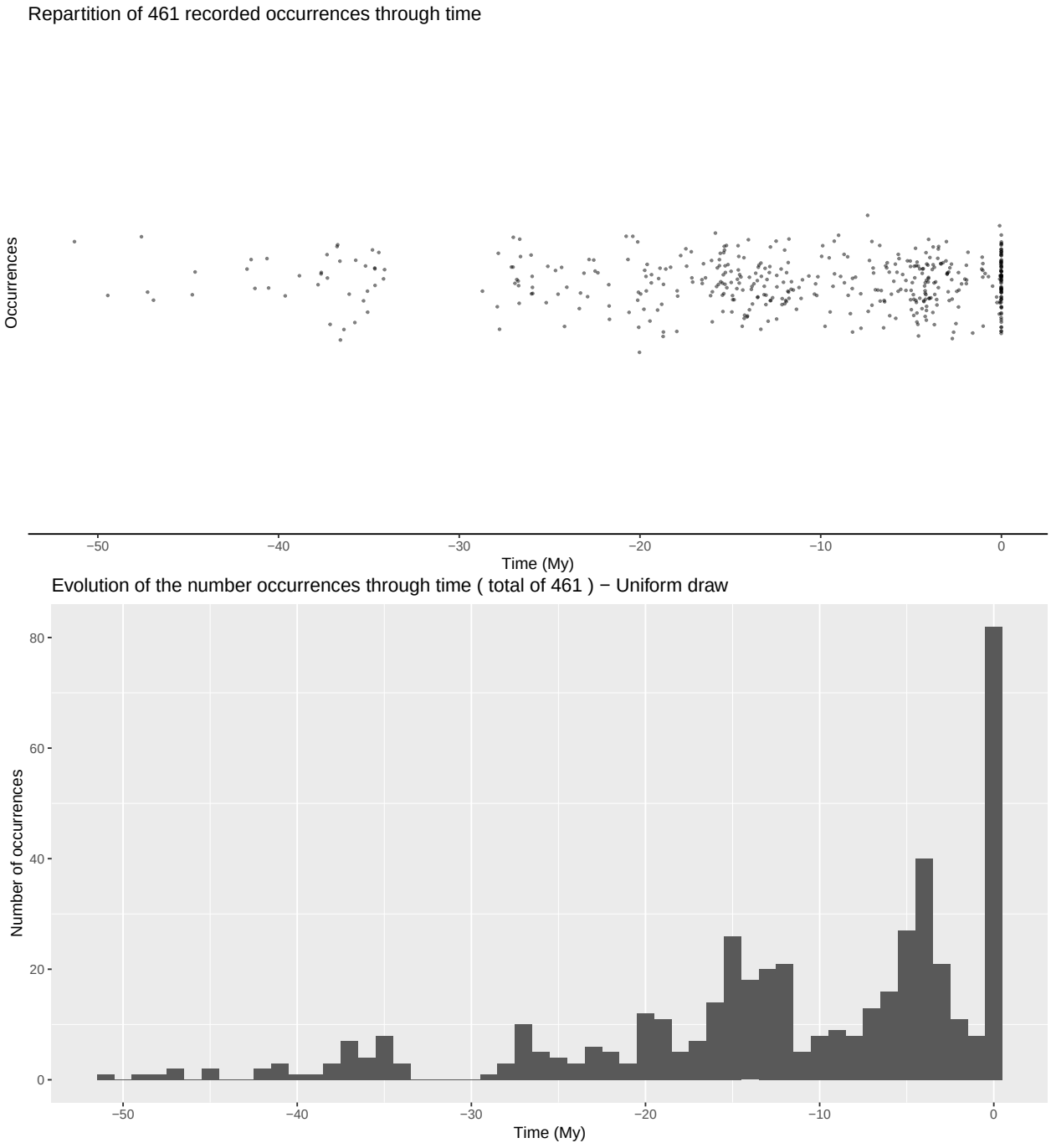

→ Numerous occurrences seem to have the same age interval so in order to avoid clusters let's draw them uniformly in their stratigraphic range rather than taking the mean.




→ The repartition seems much smoother now.

### Subsampling

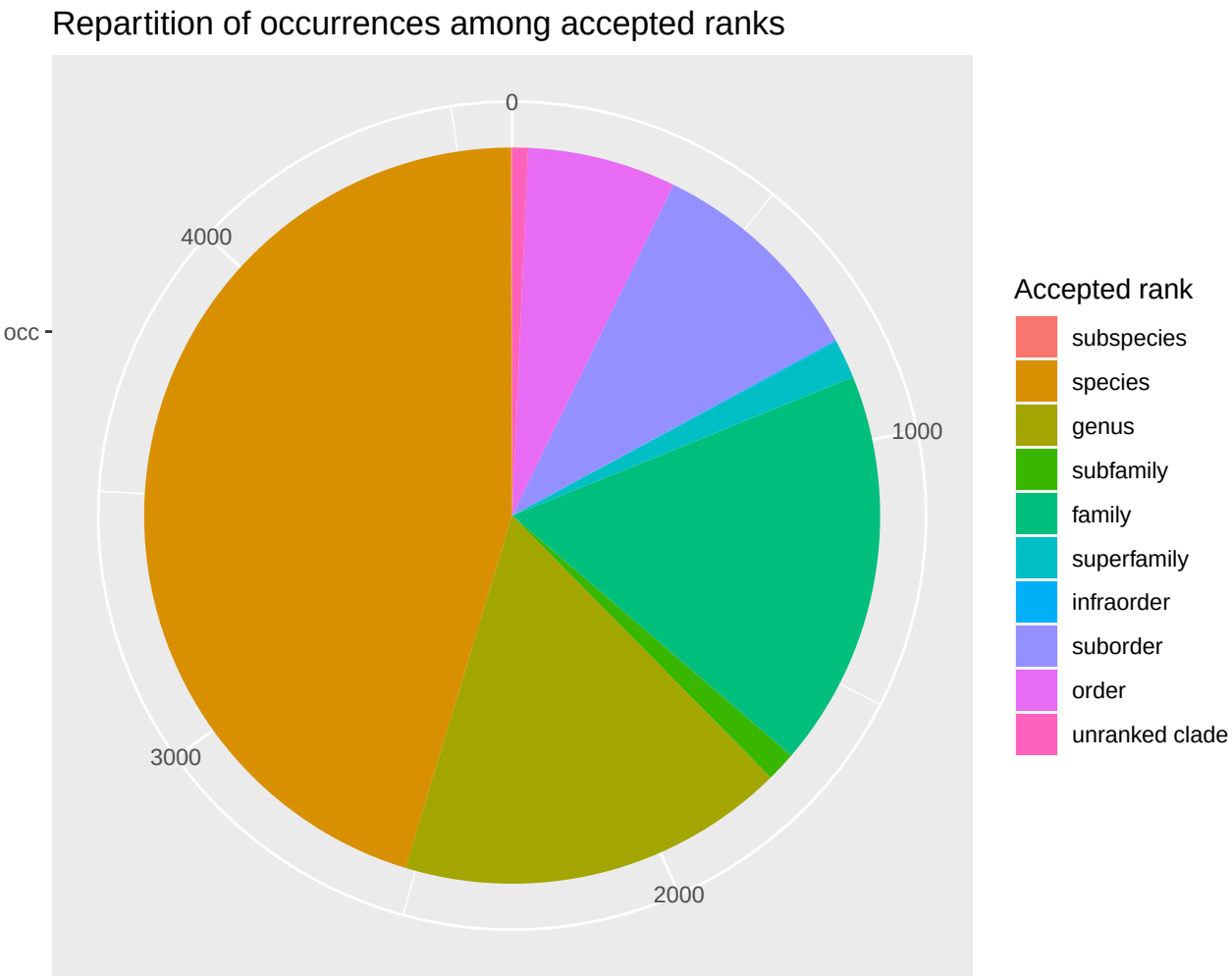
These occurrences are too numerous for our current implementation, let's subsample a fraction of them for now.




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

## Repartition among accepted ranks

### Pie chart

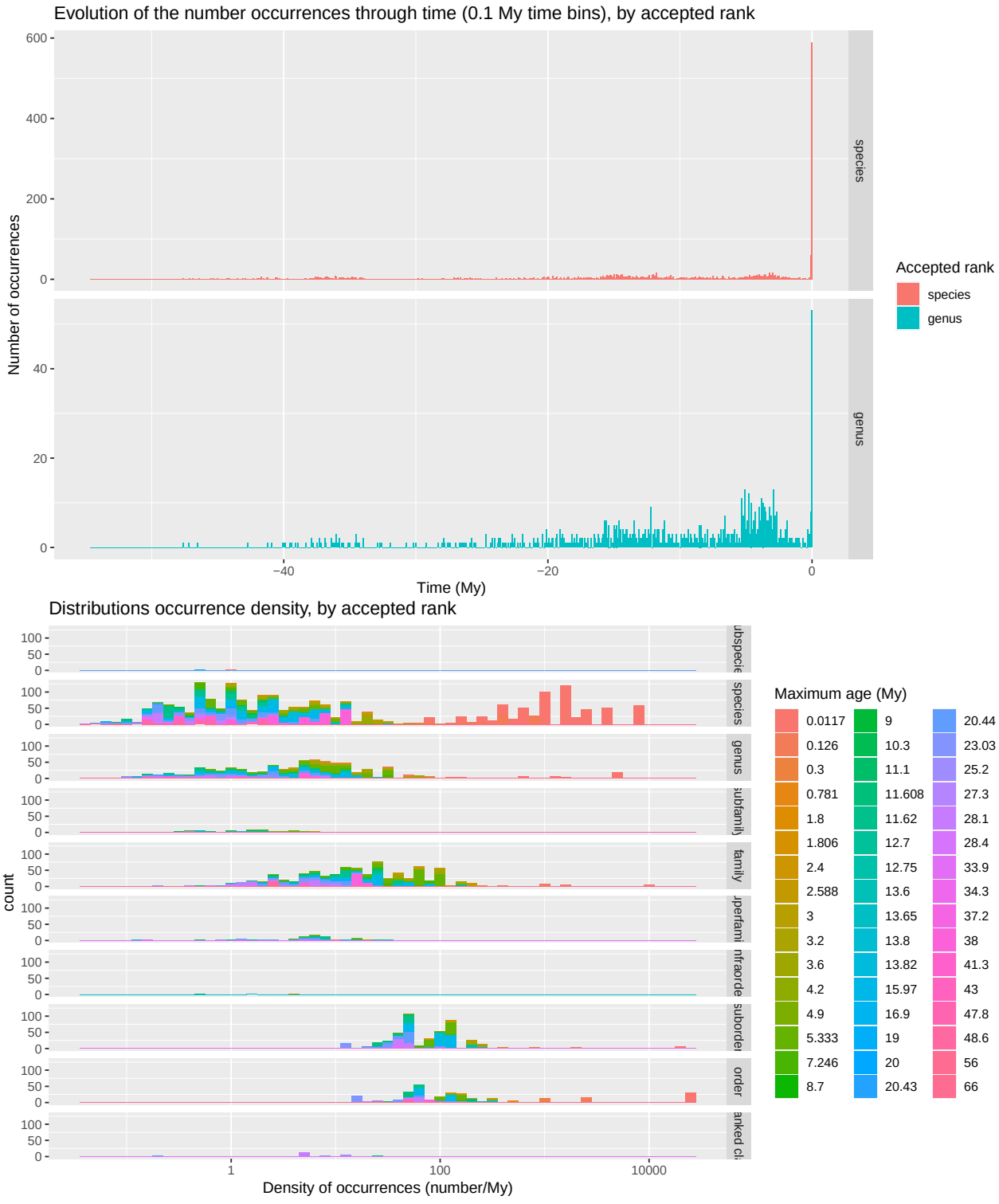


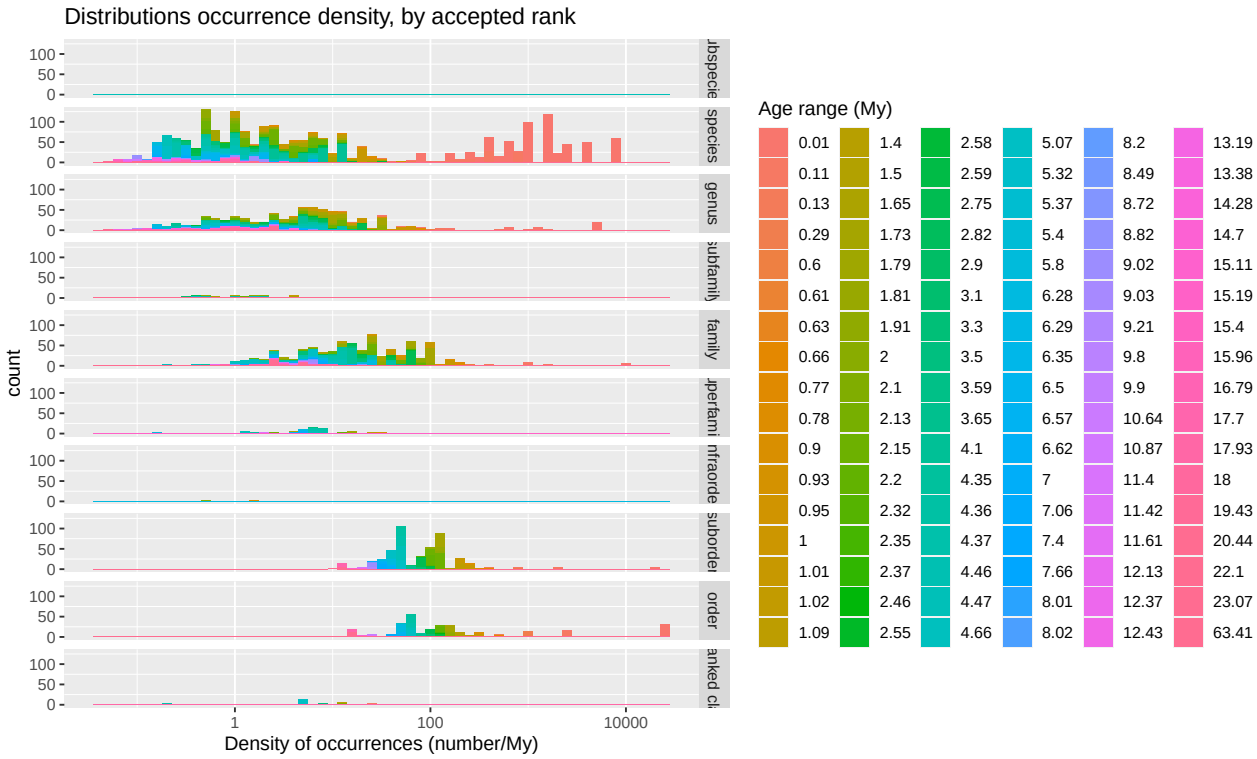

→ Half of the occurrences are identified at the level of the secies and 1/3 at the genus or family.

Some clades are unranked :

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

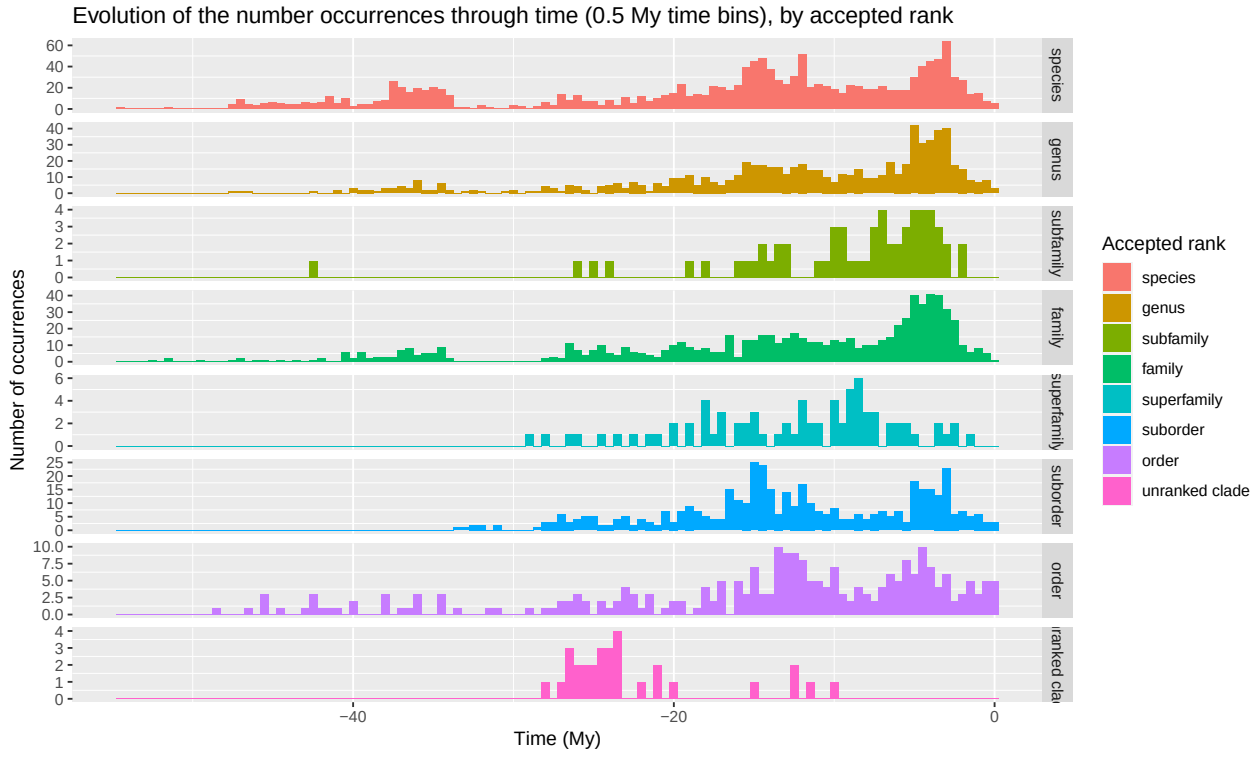
### Time repartition by rank





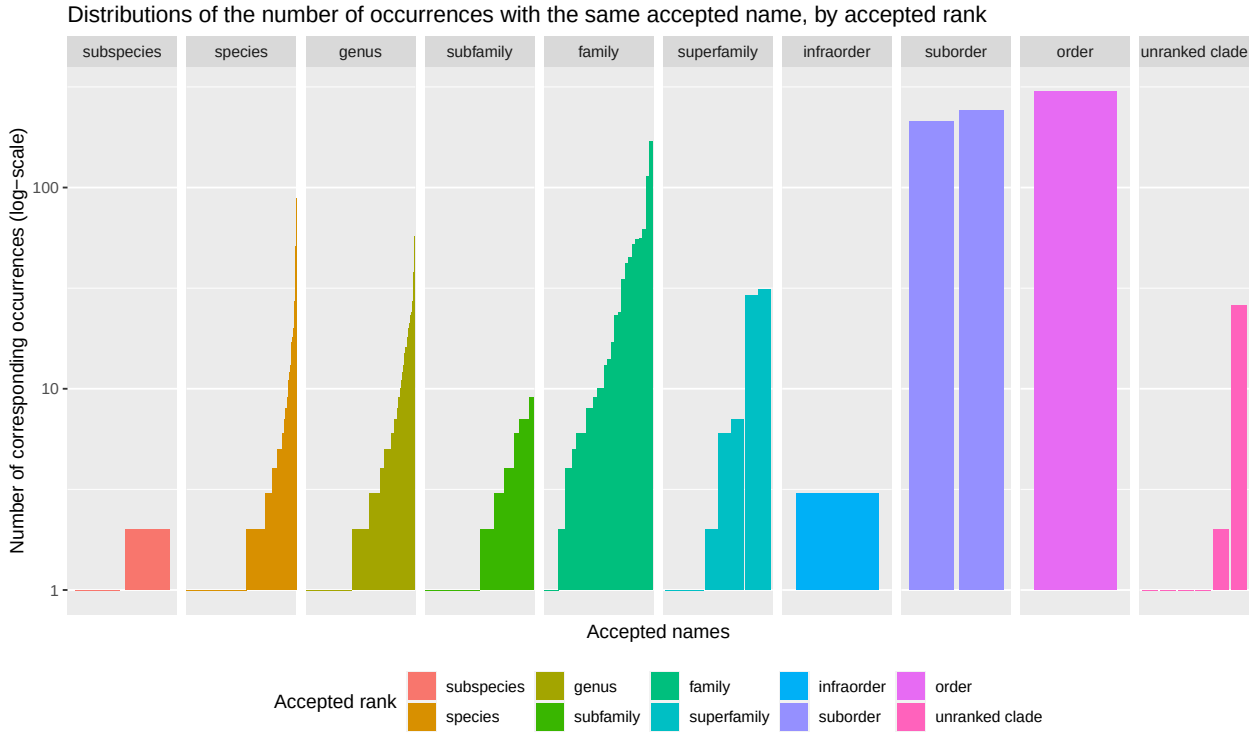

→ Apparent huge cluster of occurrences in recent times, with very precise dating = Artefact due to the "Pull of the Recent" effect ? → We decided to **remove all Late Pleistocene and Holocene occurrences** (thus setting the $\omega$-sampling to 0) in order to avoid this bias.

```
##                          cetacea_pbdb Cetacea_occ removed
## Number of occurrences            4609        3804     805
```



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

### Redundancy of occurrences with the same accepted name

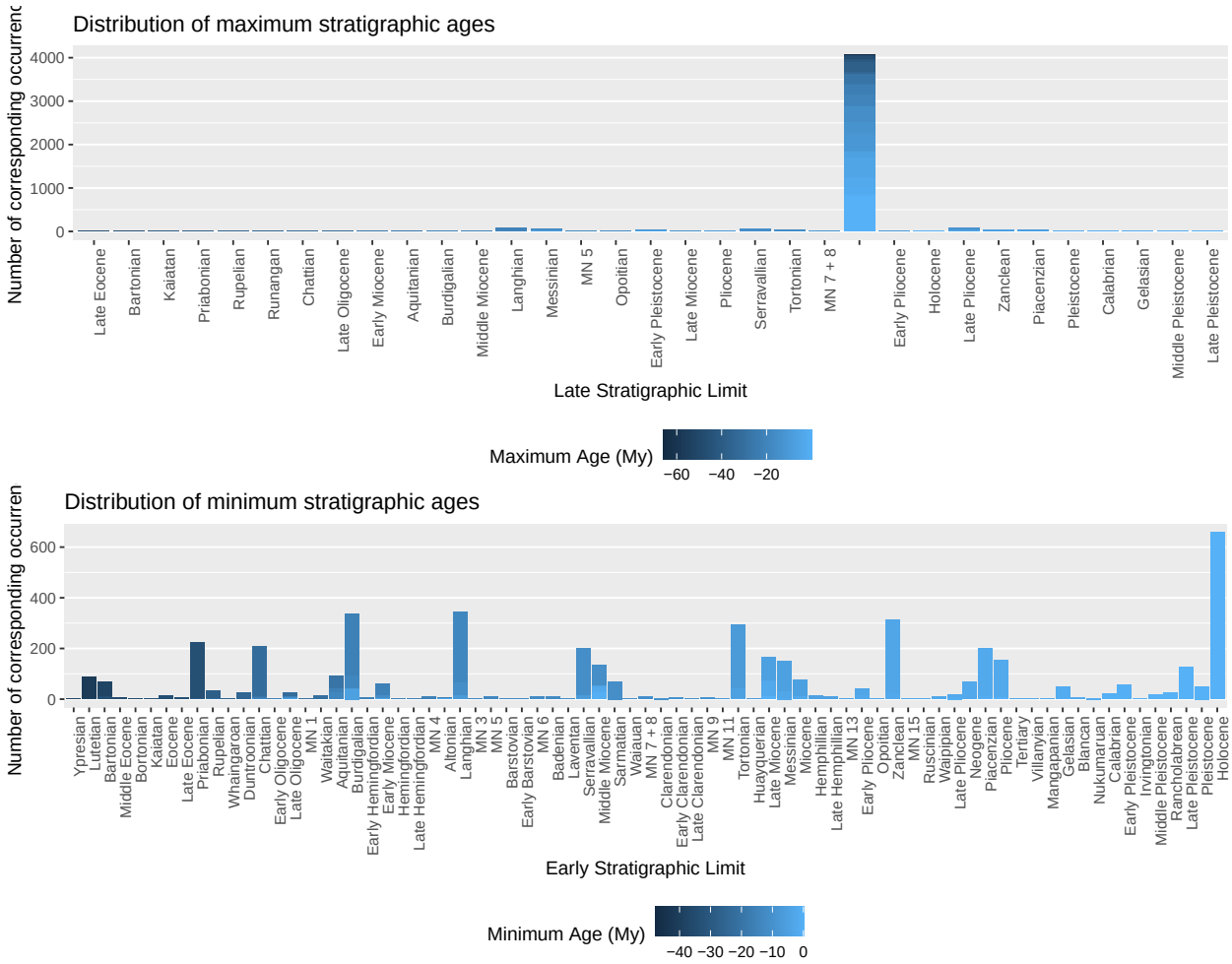



→ ~Half of species/genera/subfamilies have only one specimen by accepted name, but it could go up to ~50 within the same species and ~200 occurrences within the same suborder. **Those differences will have to be corrected because in our model all species are upposed to have the same abundance (identical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

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

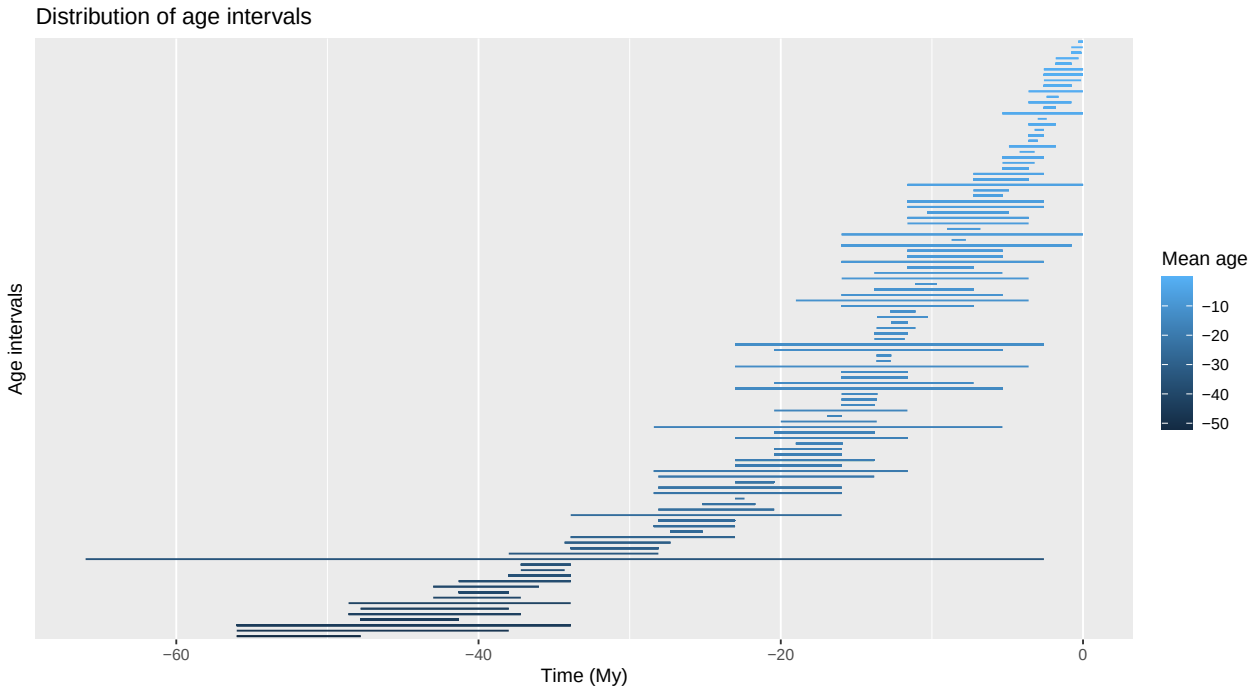
## Time intervals = stratigraphic age uncertainty

### Minimum and maximum stratigraphic limits



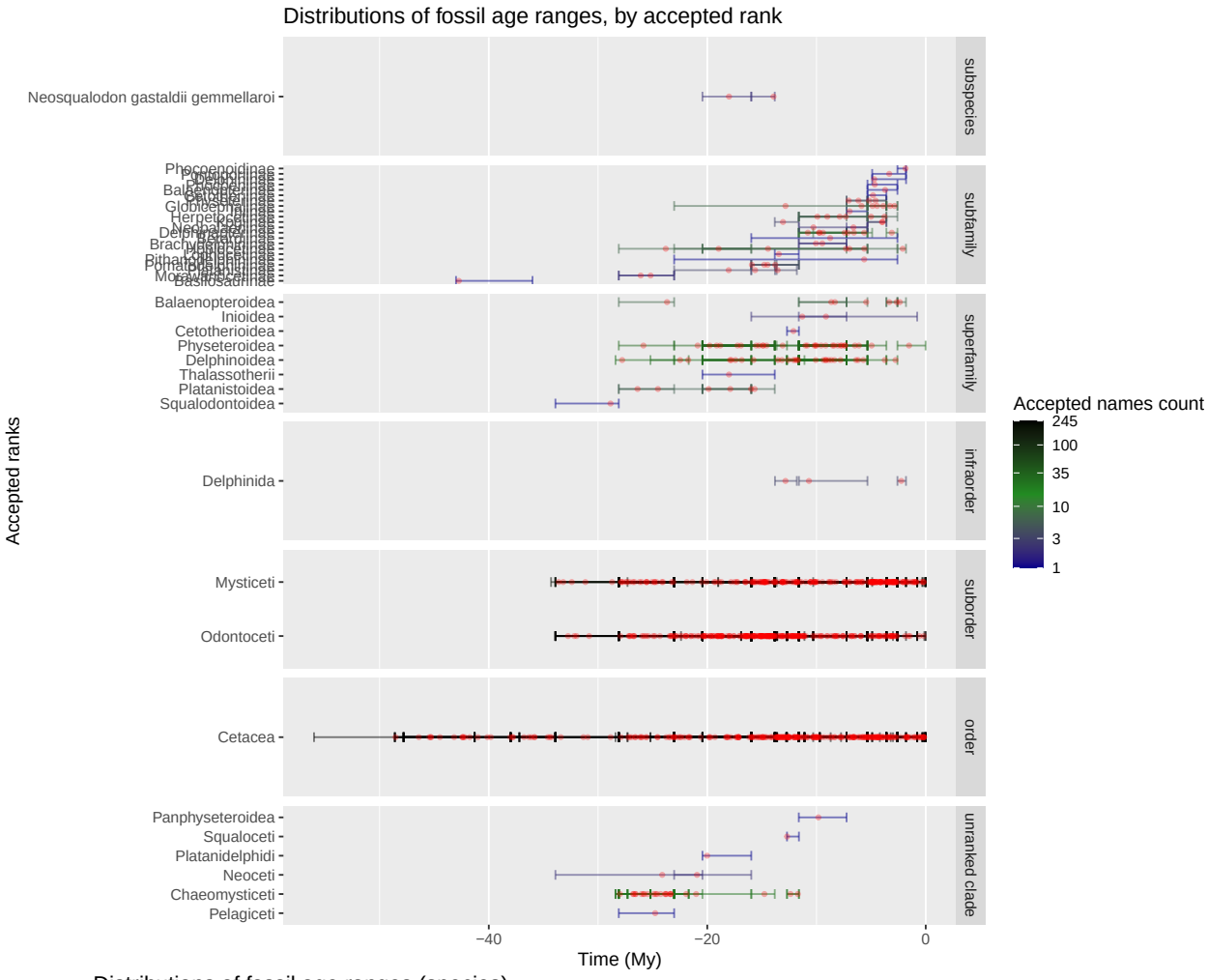

→ Most species have a early but not a late stratigraphic limit.

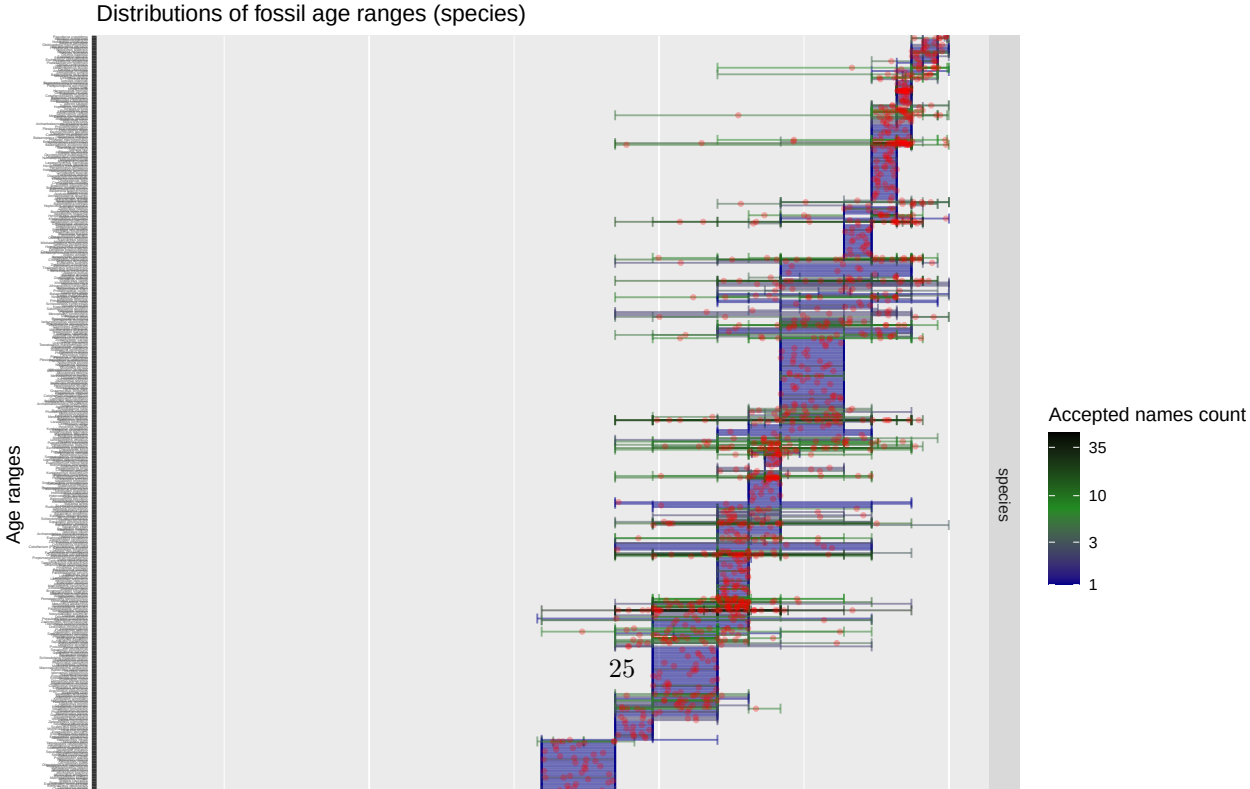
### Minimum and maximum ages

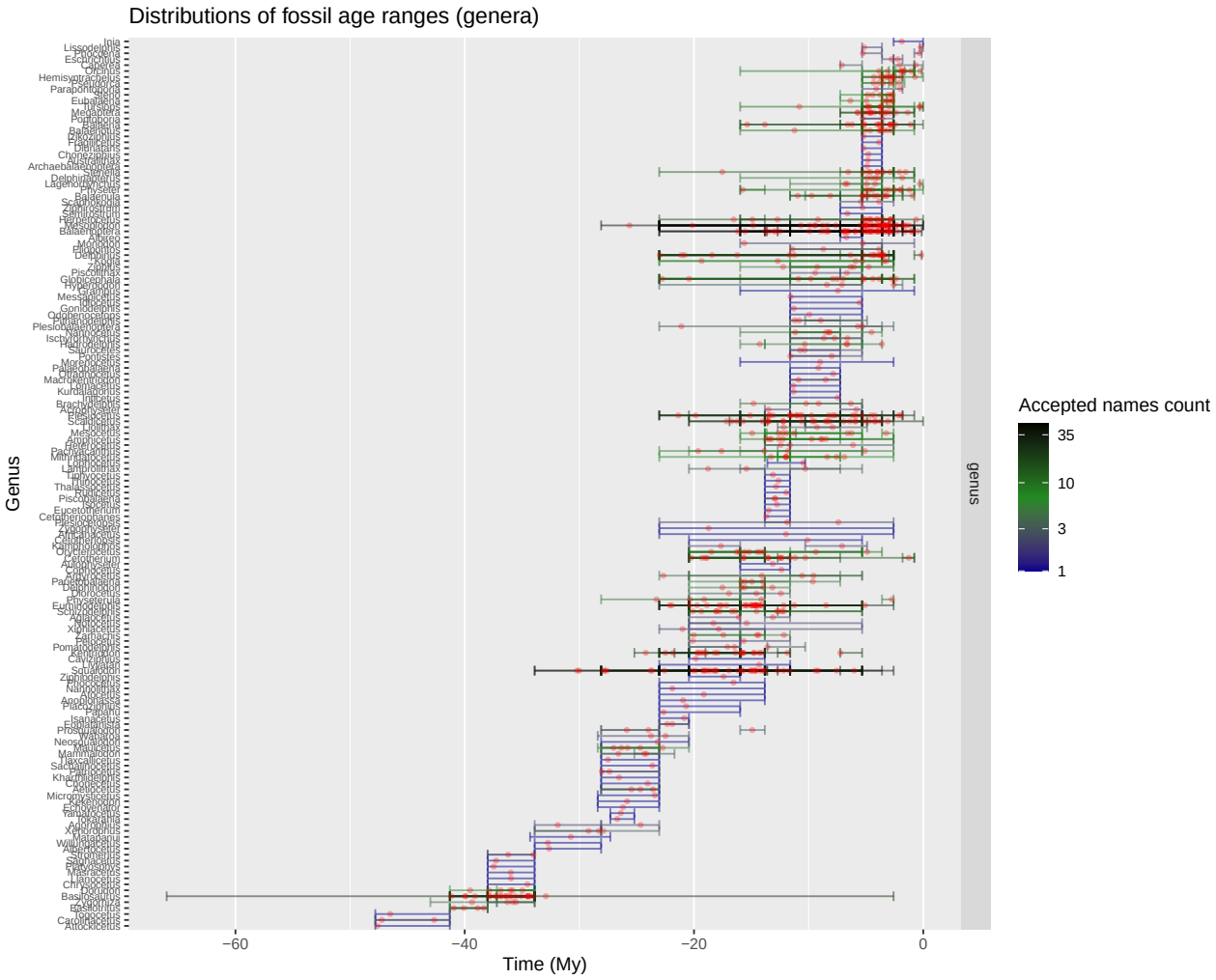


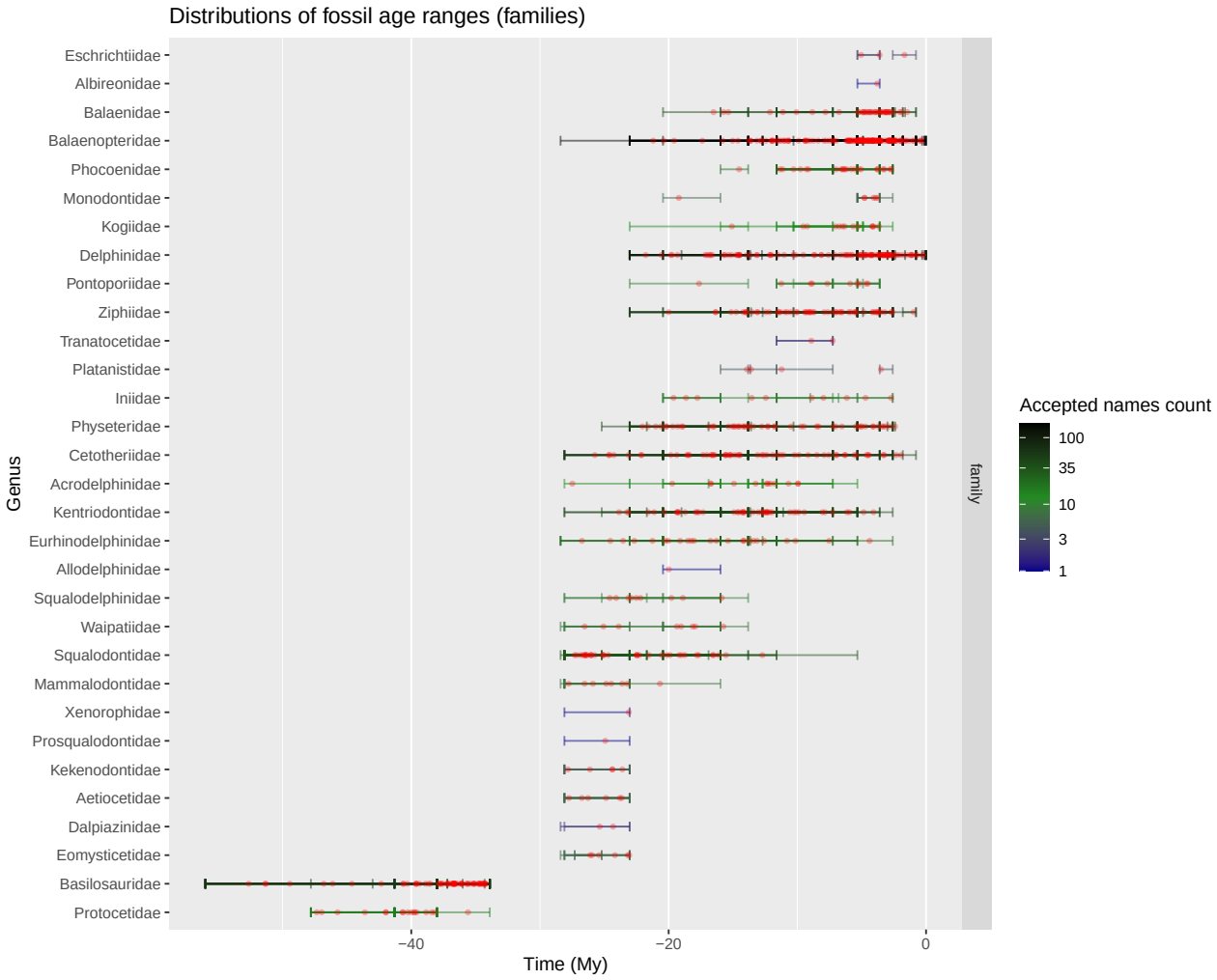
# Time ranges = duration of the time intervals

## Count occurrences by accepted name



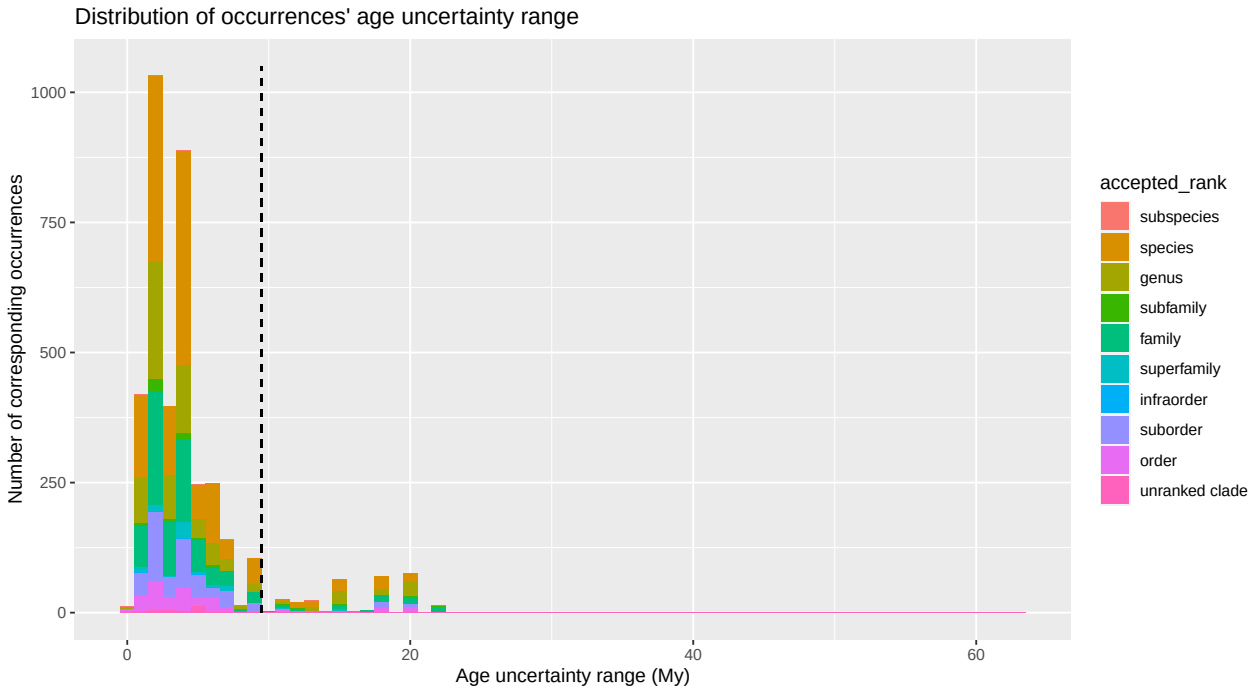

Distributions of fossil age ranges (genera)
Genus
Time (My)
−60
−40
−20
0
genus
Accepted names count
35
10
3
1




→ Some occurrences have too much age uncertainty, they risk to artificially increase species durations.

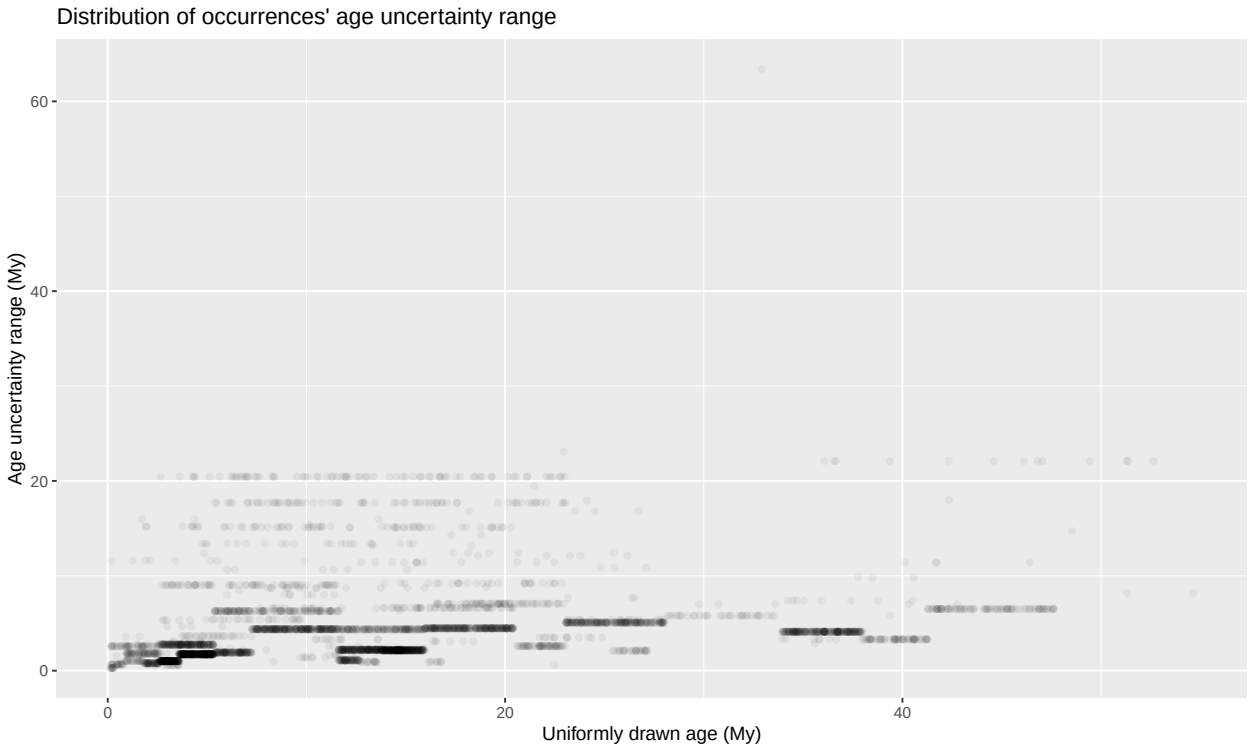
### Remove occurrences with highly uncertain dating (range > 10My)

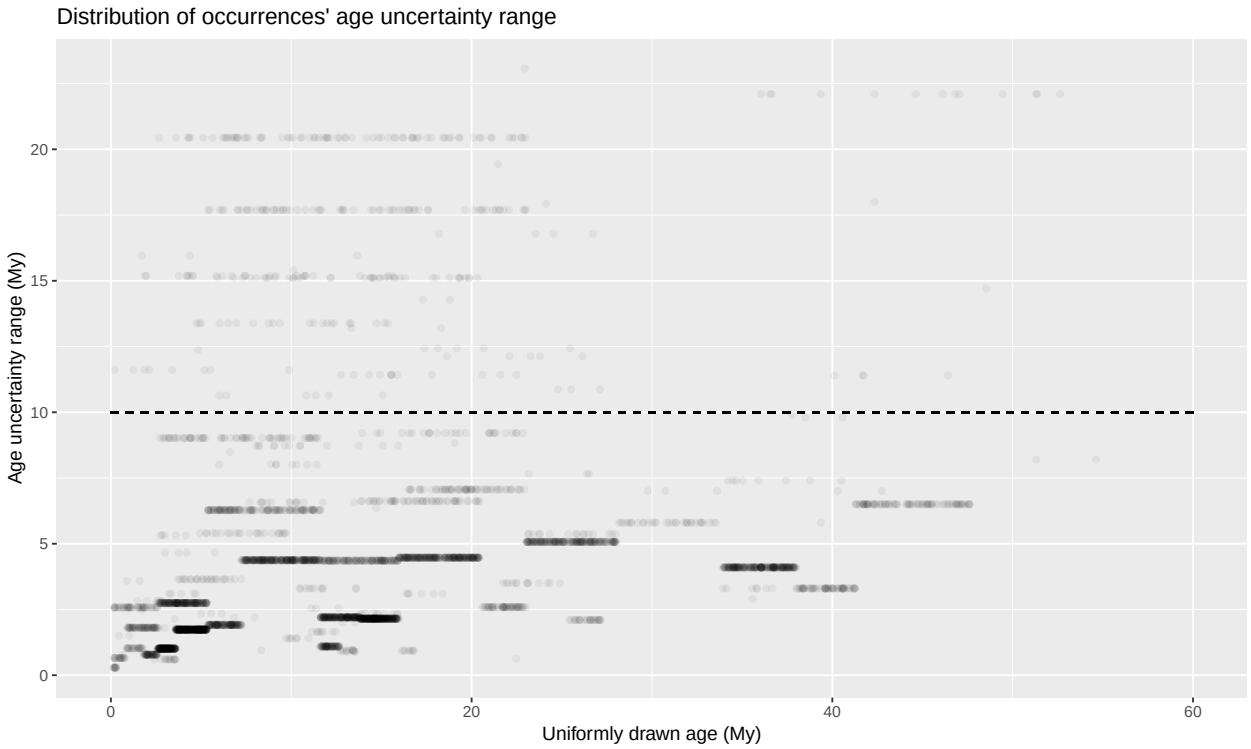
```
## [1] 3502  117
```

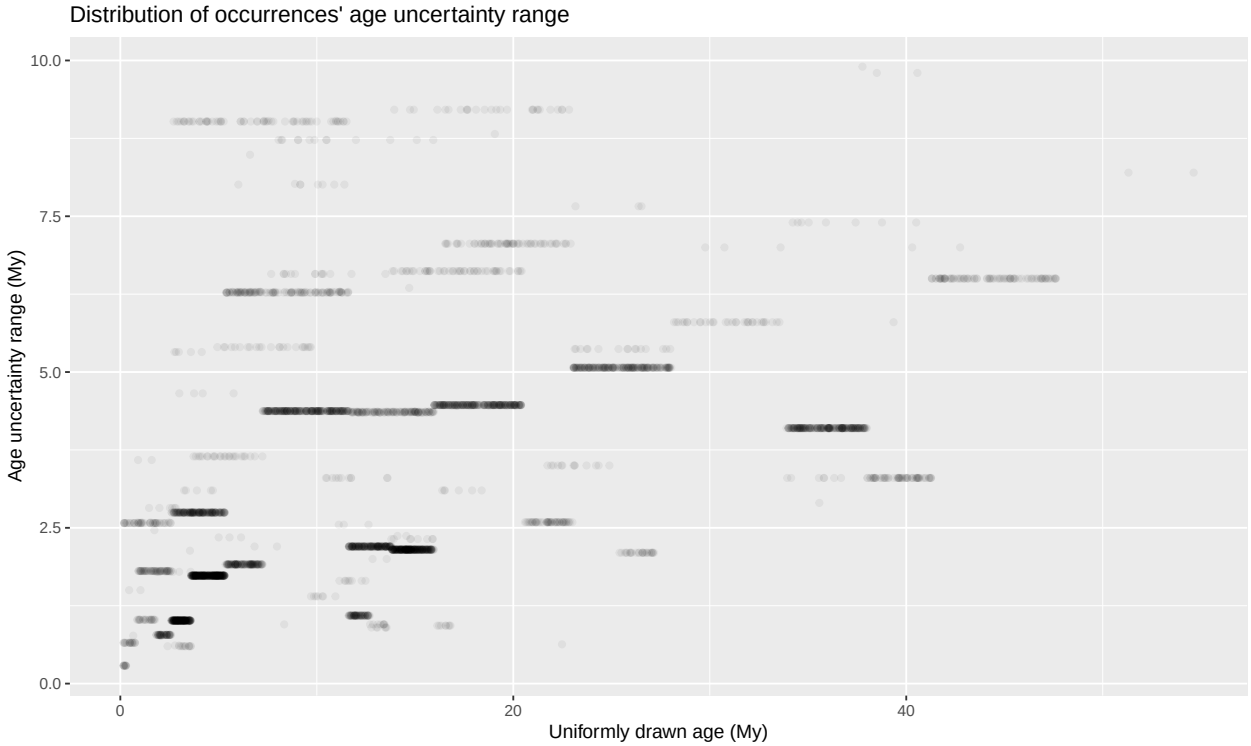


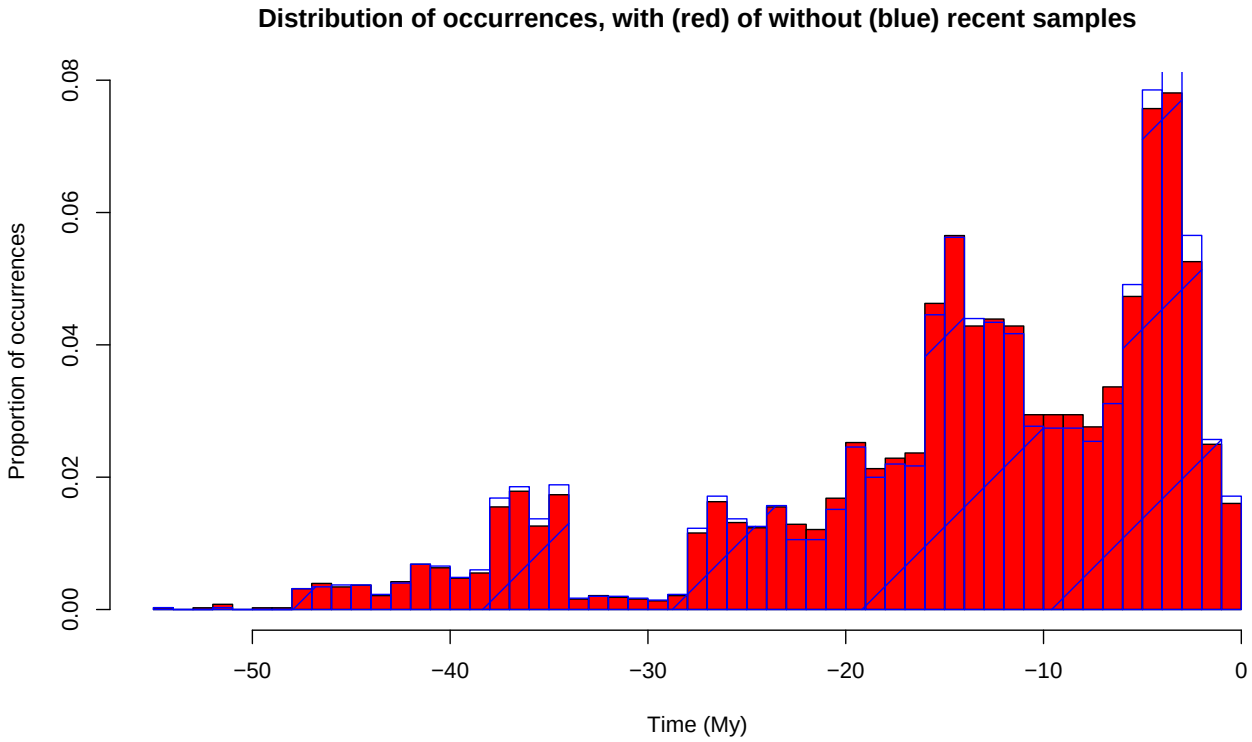

Most of occurrences show less than 10 My age uncertainty, let's try to keep only these ones.

```
##                       all_ranges smaller_10My removed
## Number of occurrences       3804         3502     302
```

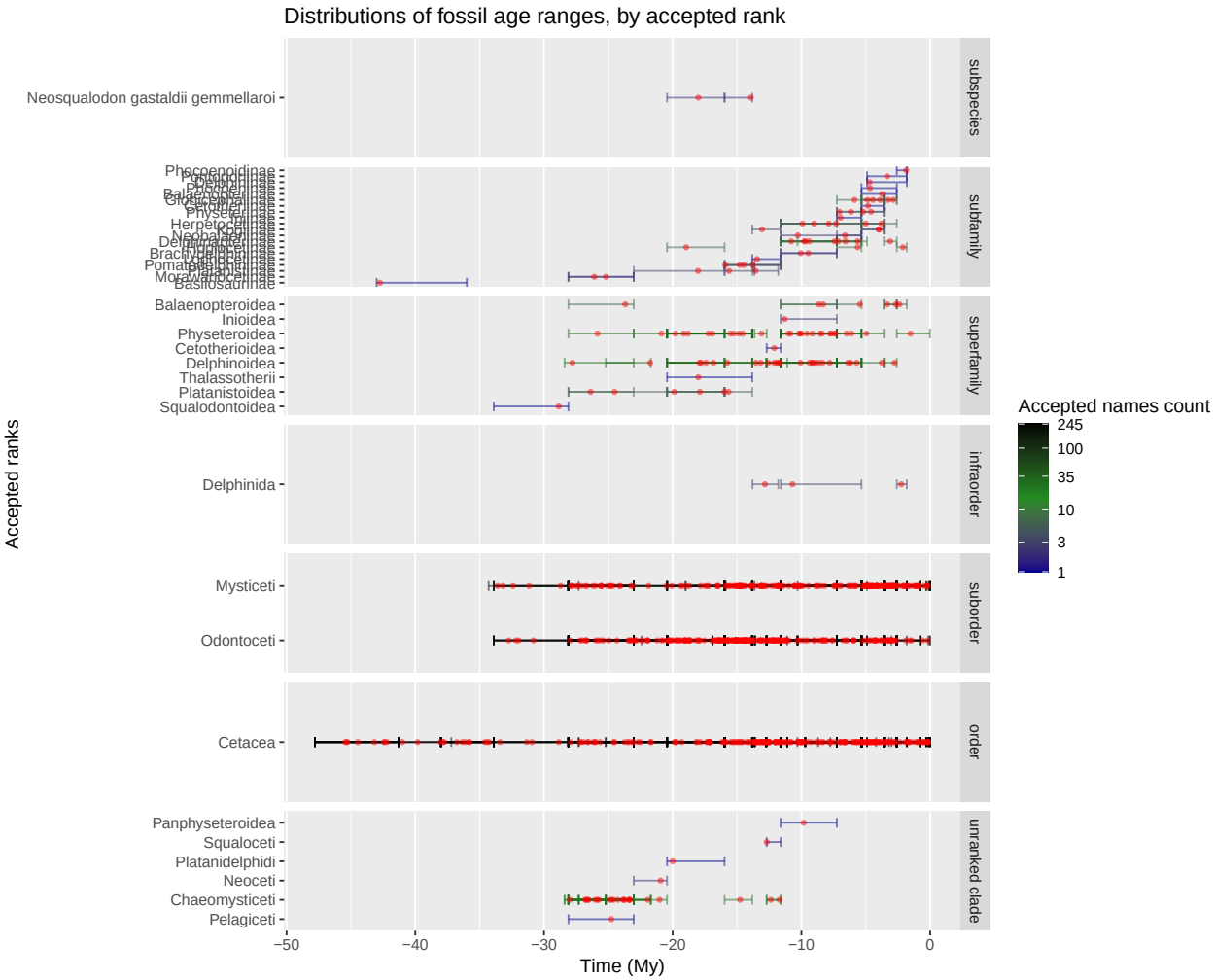


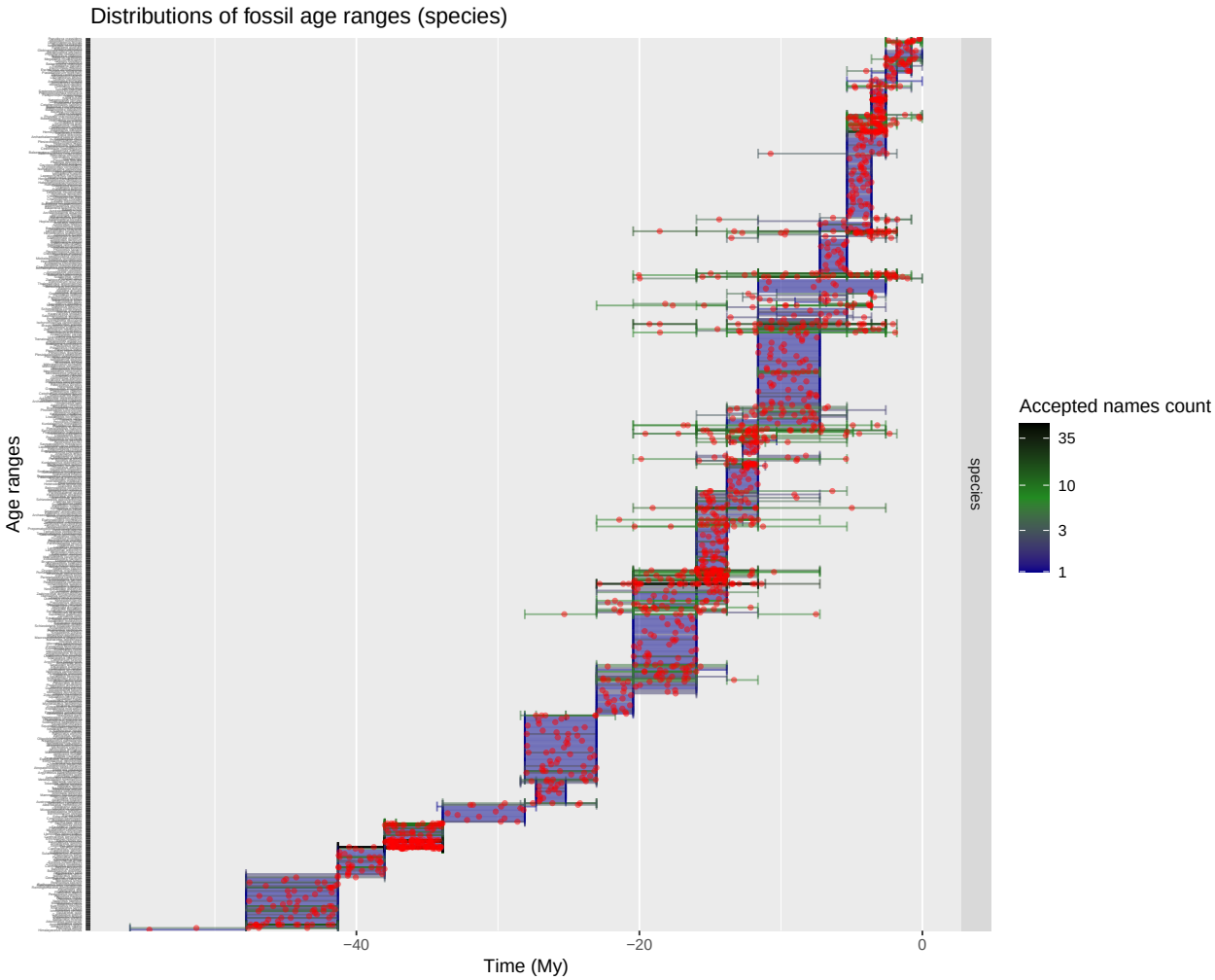

Distribution of occurrences' age uncertainty range
Age uncertainty range (My)
0
5
10
15
20
0
20
40
60
Uniformly drawn age (My)

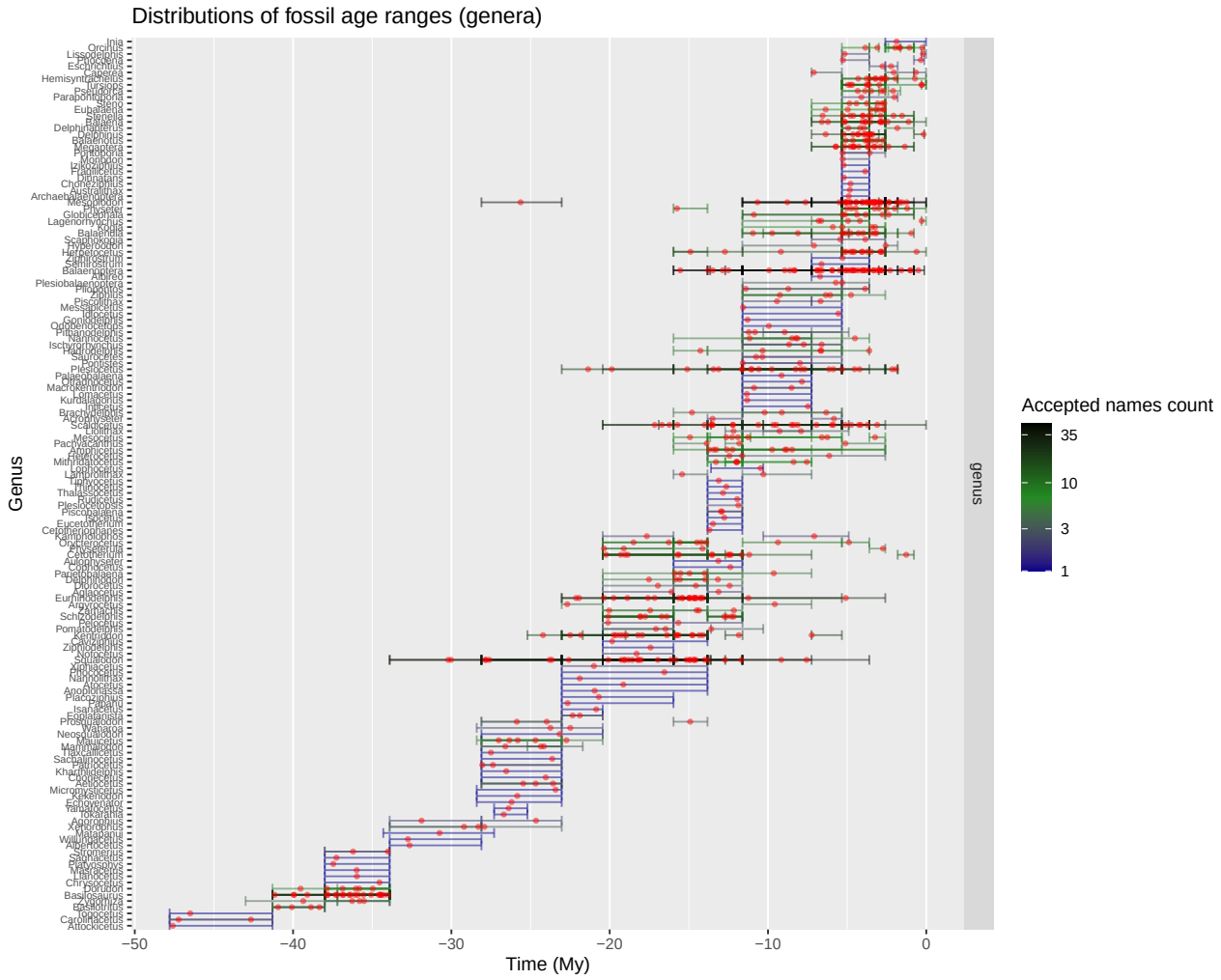

Distribution of occurrences' age uncertainty range
Age uncertainty range (My)
0.0
2.5
5.0
7.5
10.0
0
20
40
Uniformly drawn age (My)

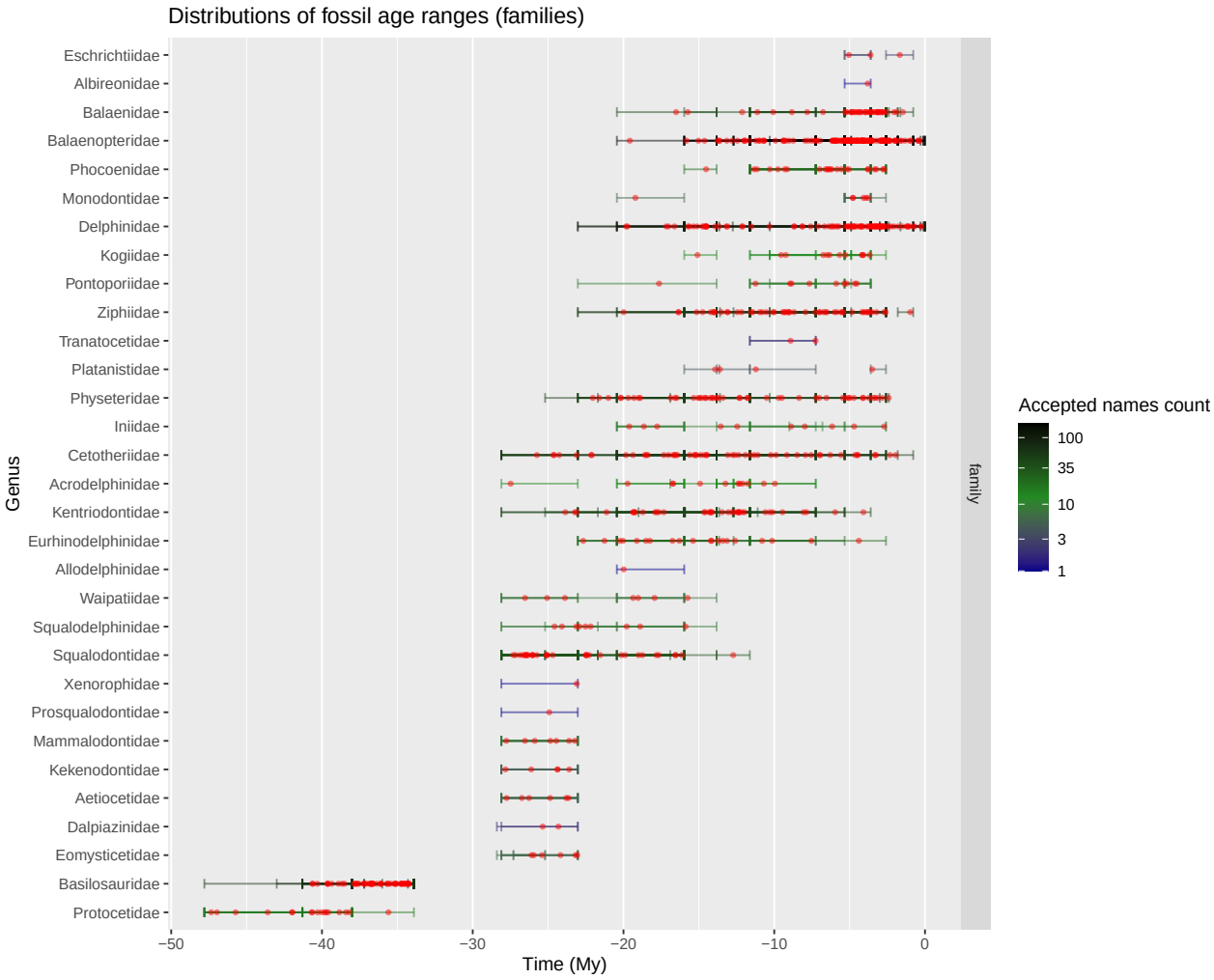



→ The removal of highly uncertain occurrences seems to be only a little biased, even if uncertainty globally increases with age.


Distributions of fossil age ranges, by accepted rank
Accepted ranks
Neosqualodon gastaldii gemmellaroi
Balaenopteroidea
Inioidea
Physeteroidea
Cetotherioidea
Delphinoidea
Thalassotherii
Platanistoidea
Squalodontoidea
Delphinida
Mysticeti
Odontoceti
Cetacea
Panphyseteroidea
Squaloceti
Platanidelphidi
Neoceti
Chaeomysticeti
Pelagiceti
subspecies
subfamily
superfamily
infraorder
suborder
order
unranked clade
−50
−40
−30
−20
−10
0
Time (My)
Accepted names count
245
100
35
10
3
1

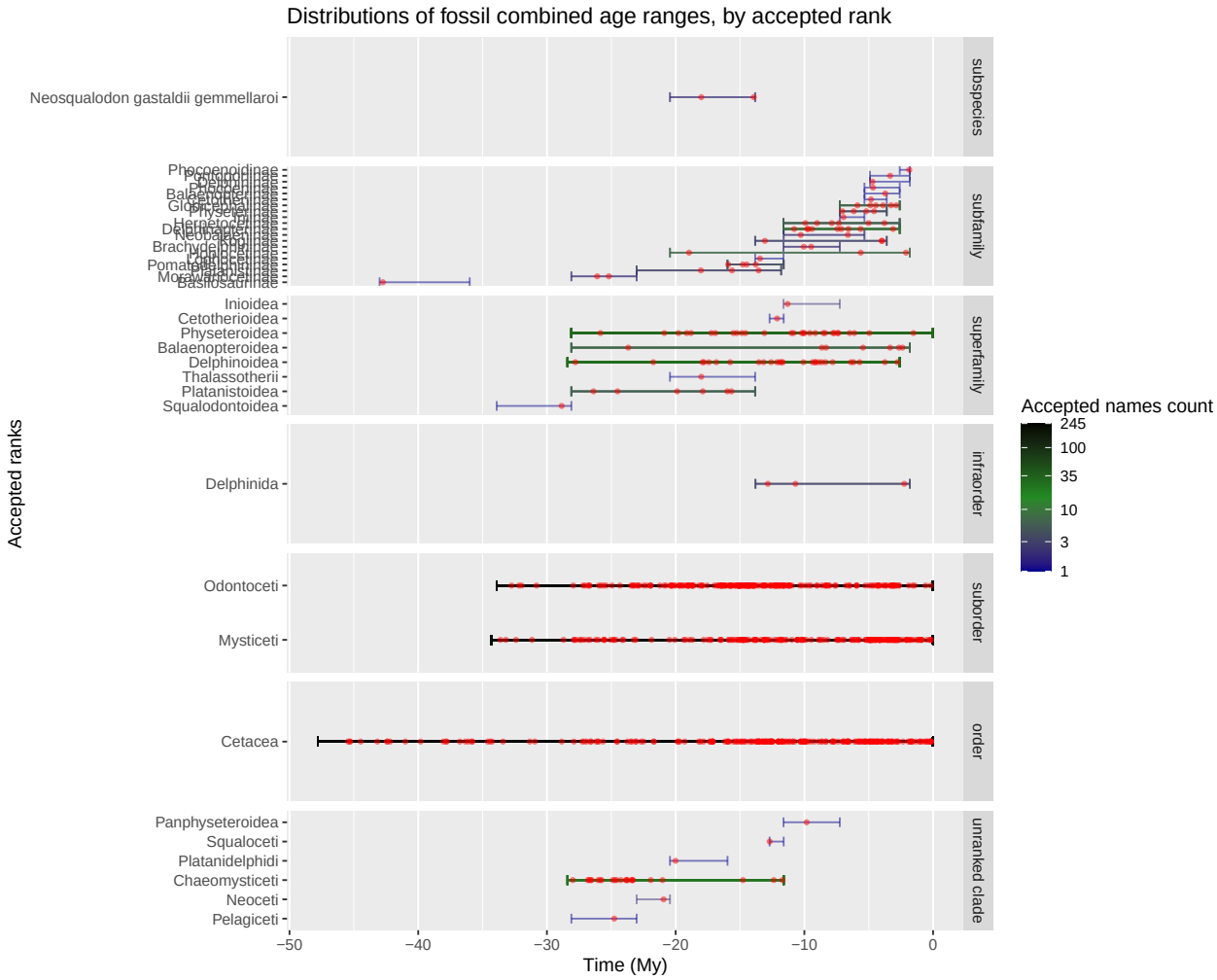

Distributions of fossil age ranges (species)
Age ranges
species
−40
−20
0
Time (My)
Accepted names count
35
10
3
1

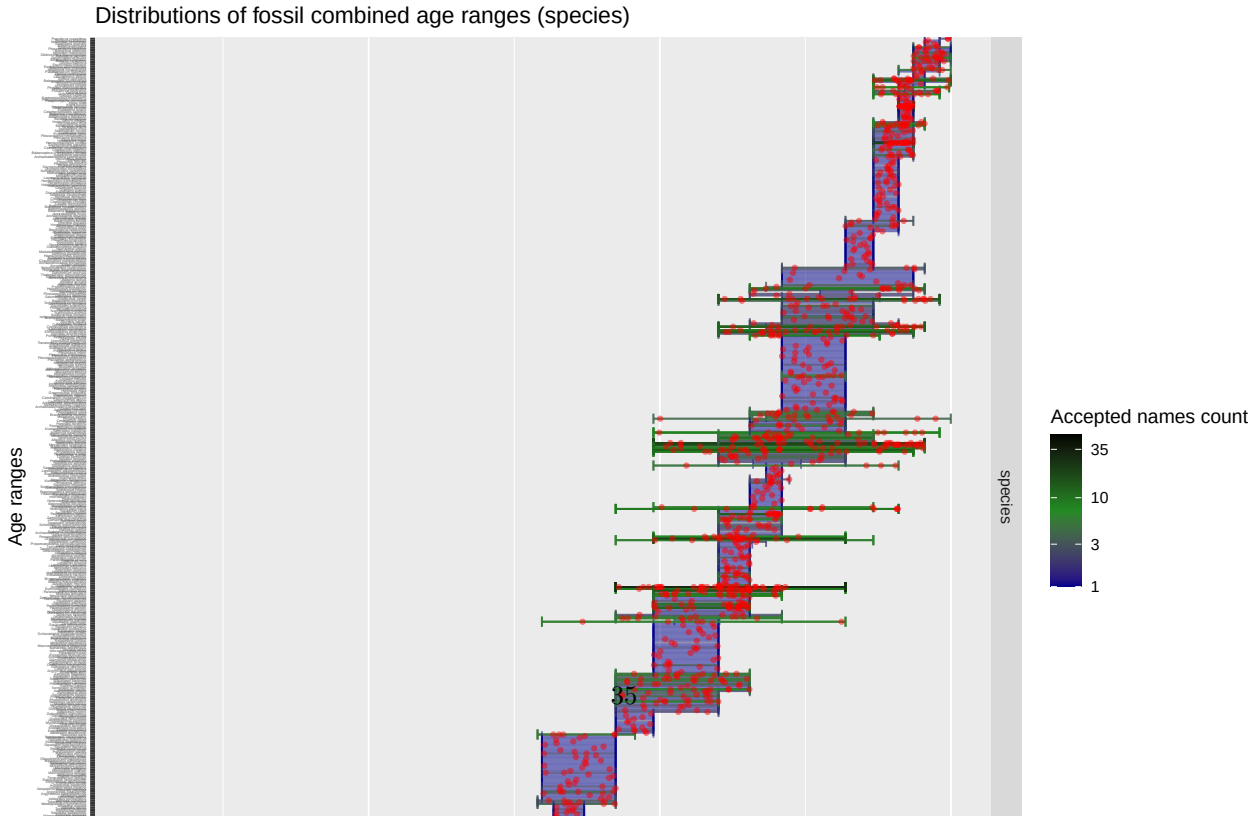

Distributions of fossil age ranges (genera)
Genus
Time (My)
−50
−40
−30
−20
−10
0
genus
Accepted names count
35
10
3
1

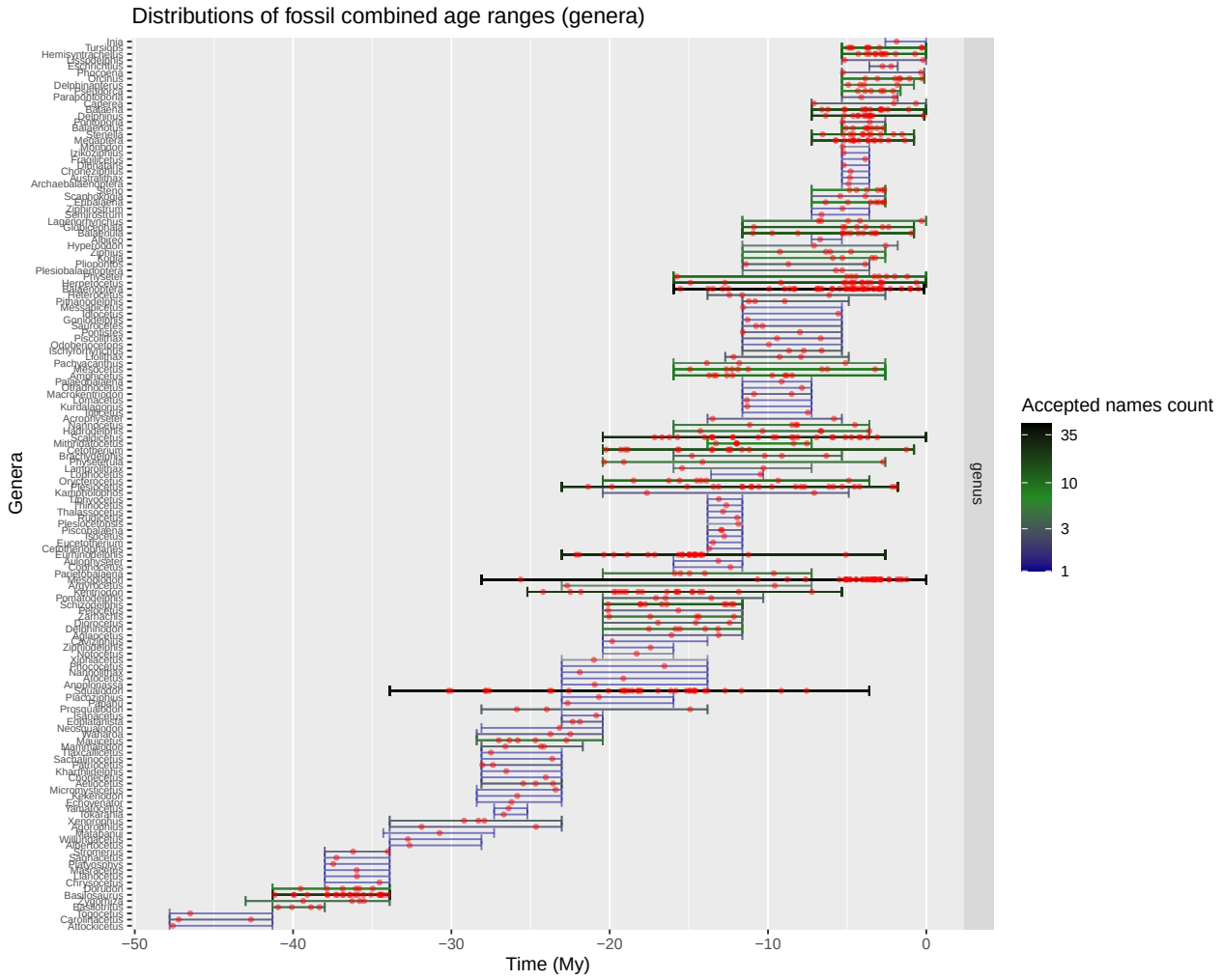


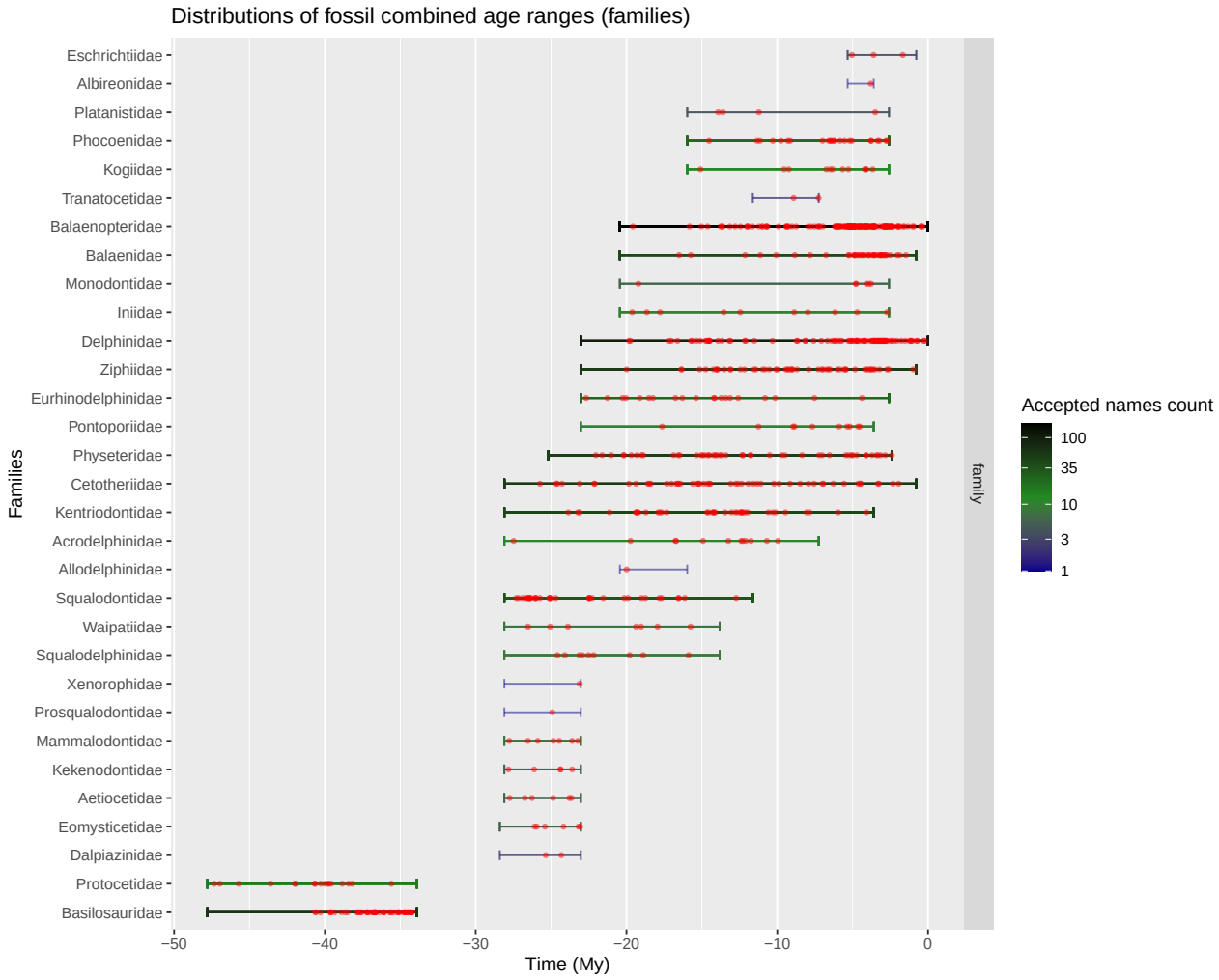

→ Some species (or other ranks) have several occurrences with several time ranges, **let's combine them into a unique range covering all the others**.

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



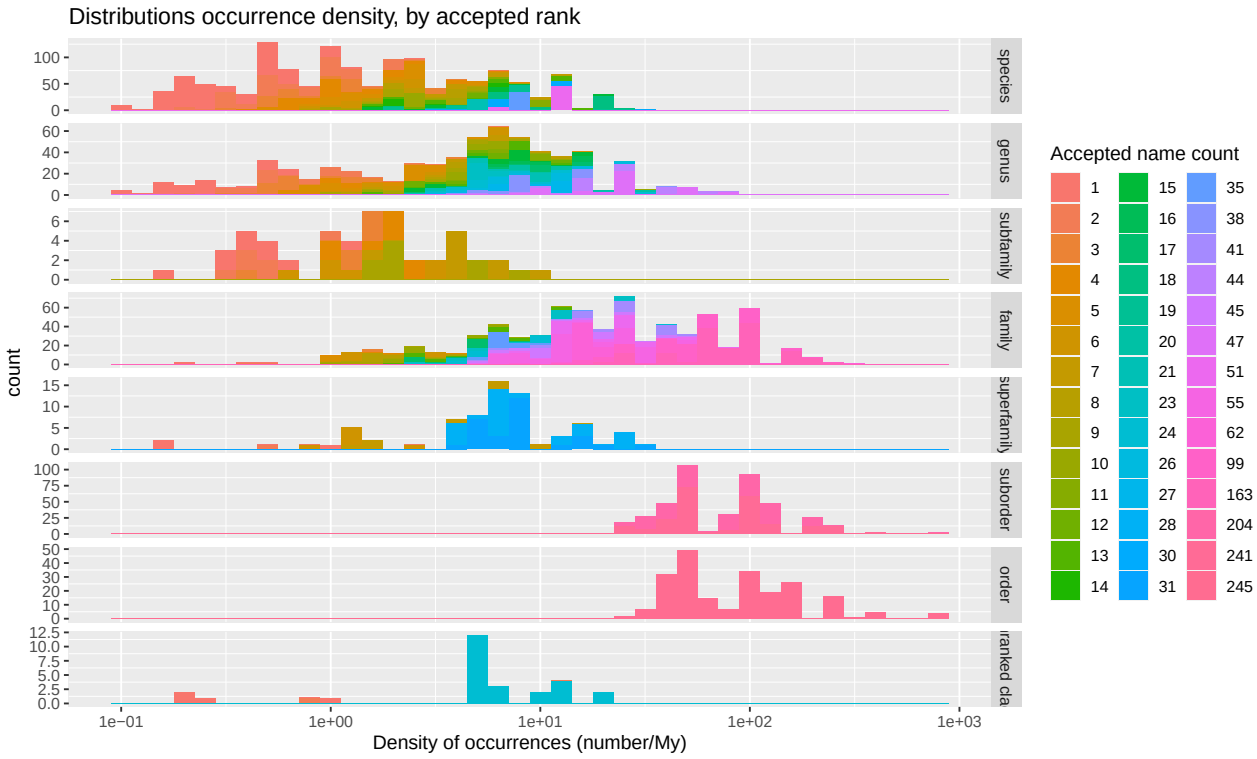




Distributions of fossil combined age ranges (genera)
Genera
Time (My)
−50
−40
−30
−20
−10
0
genus
Accepted names count
35
10
3
1


Distributions of fossil combined age ranges (families)
Eschrichtiidae
Albireonidae
Platanistidae
Phocoenidae
Kogiidae
Tranatocetidae
Balaenopteridae
Balaenidae
Monodontidae
Iniidae
Delphinidae
Ziphiidae
Eurhinodelphinidae
Pontoporiidae
Physeteridae
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

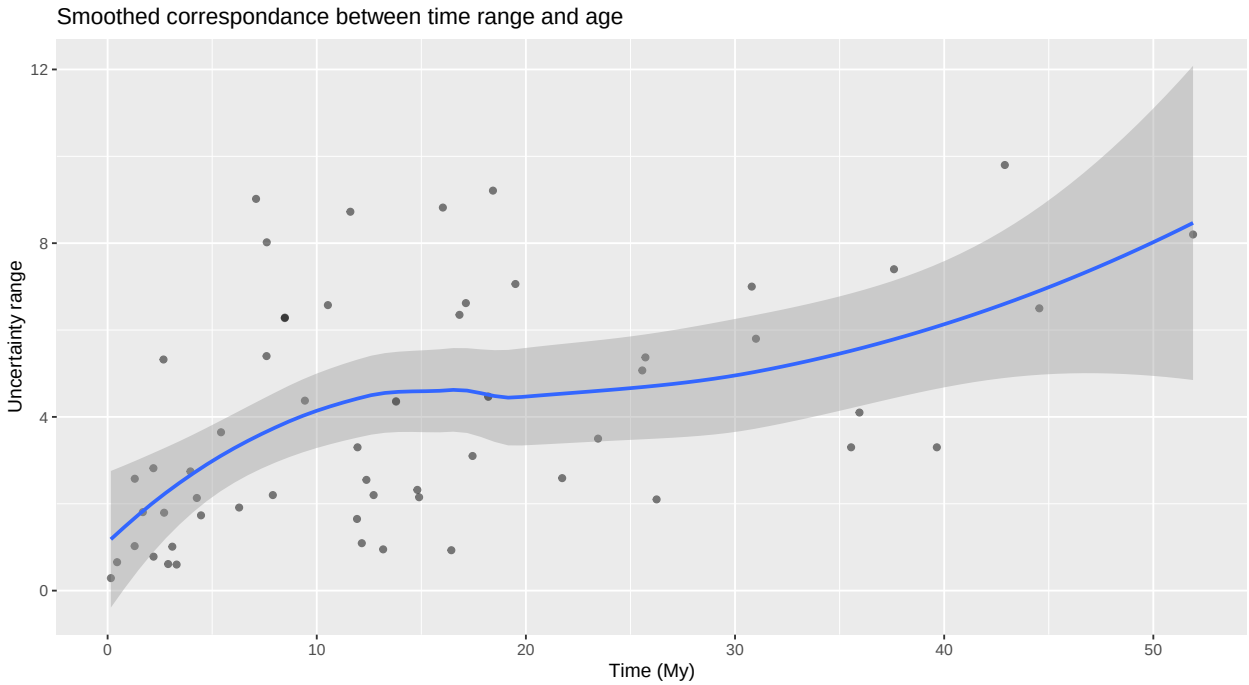
### Density distributions

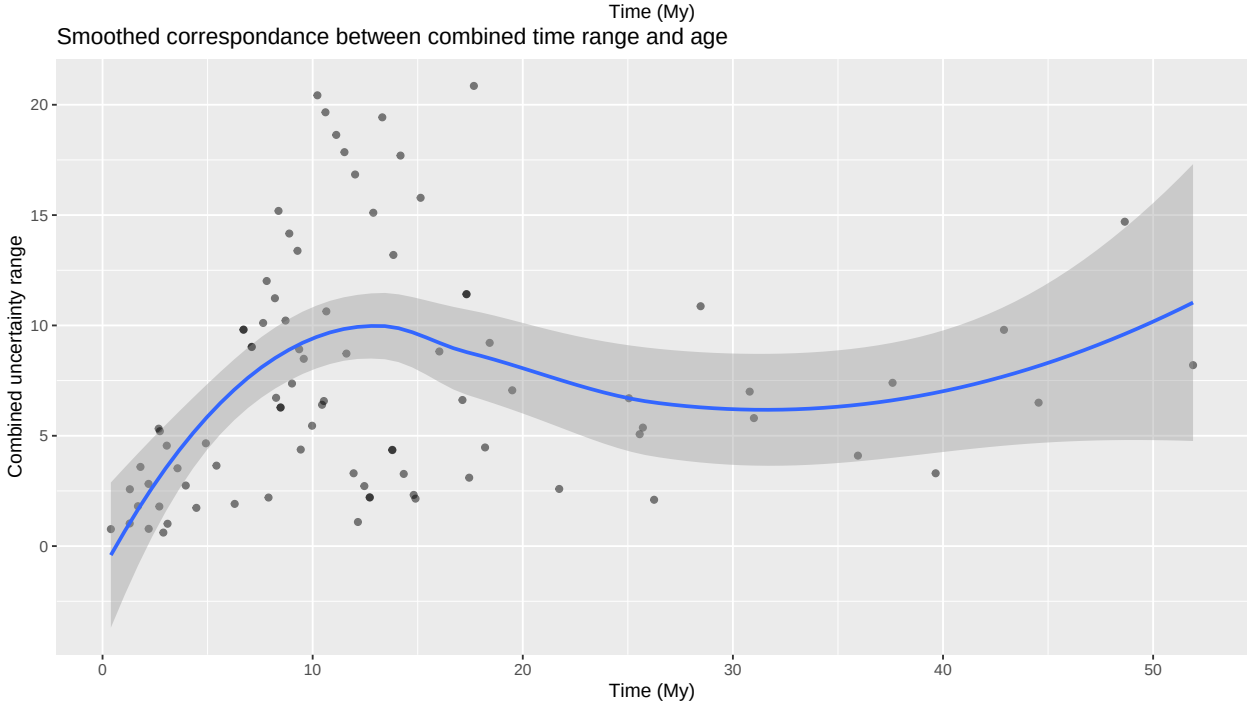



→ Density logically increases as taxa ranks increase.

## Correlation between time range and age

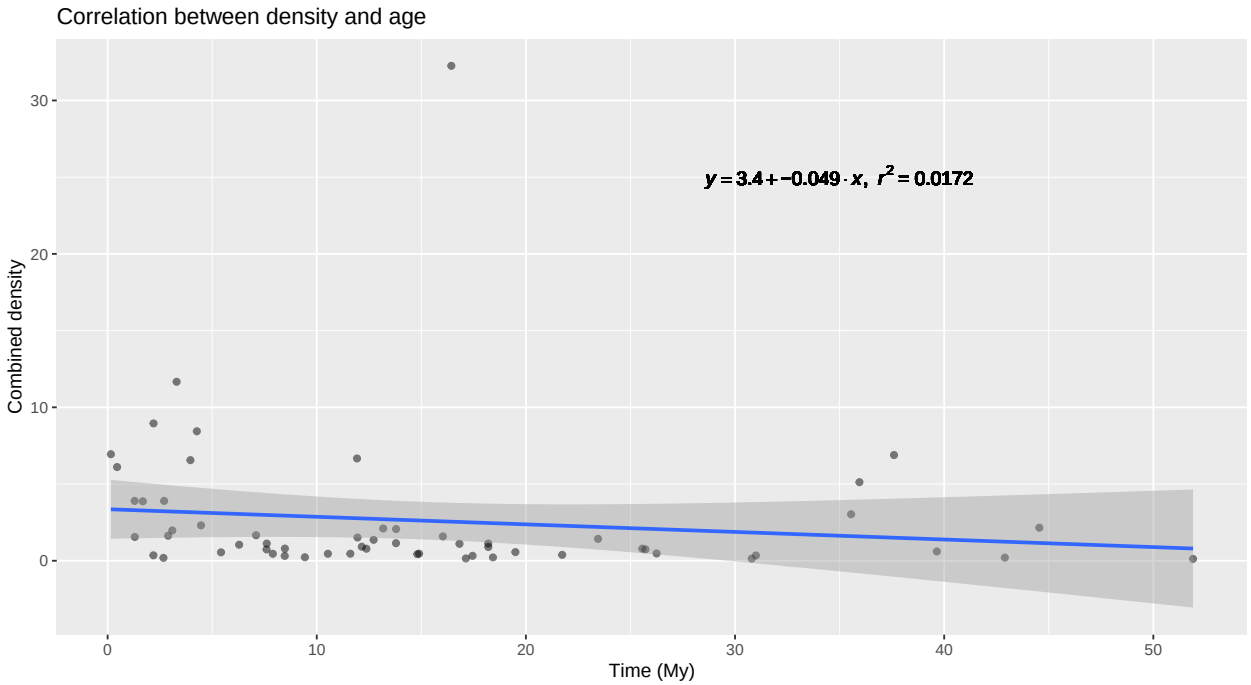
If we want to correct species abundance differences based on the number of occurrences in the time range ("density"), those factors should not depend on time in order to avoid penalizing periods with higher densities.

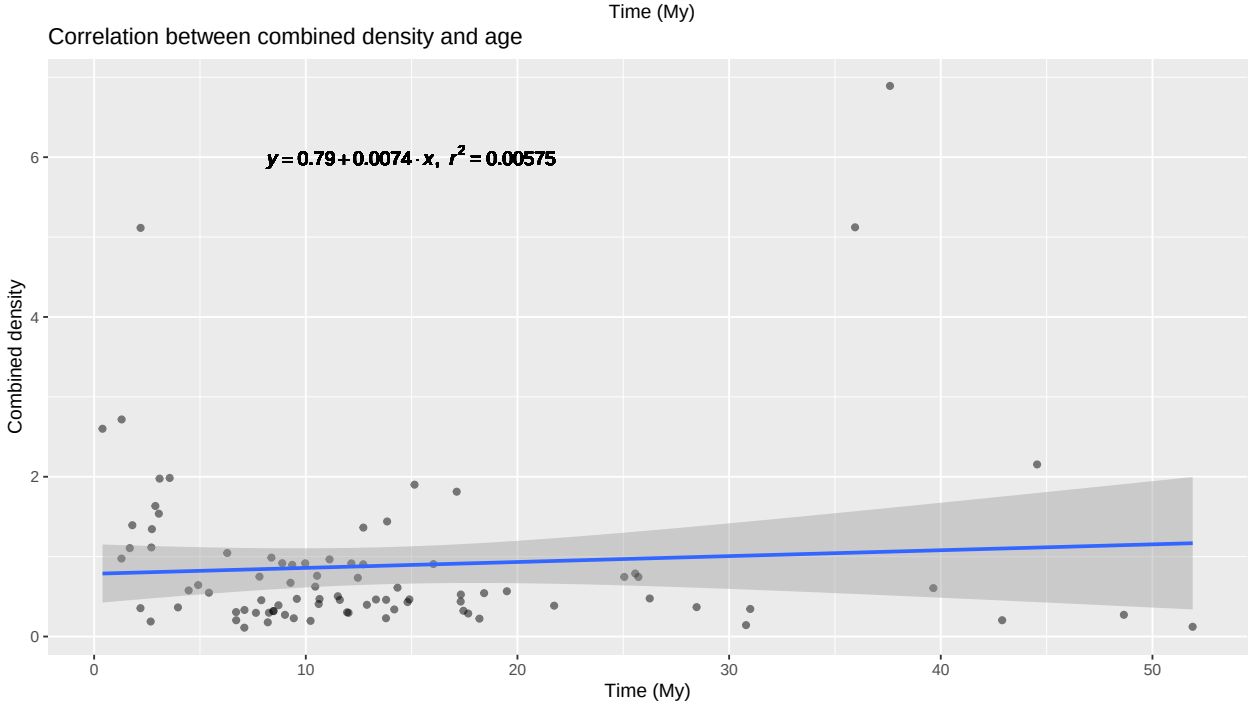



→ It seems that age range varies importantly with time, but when taking the full combined range into account the correlation seems quite weak after the first million years.

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




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

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

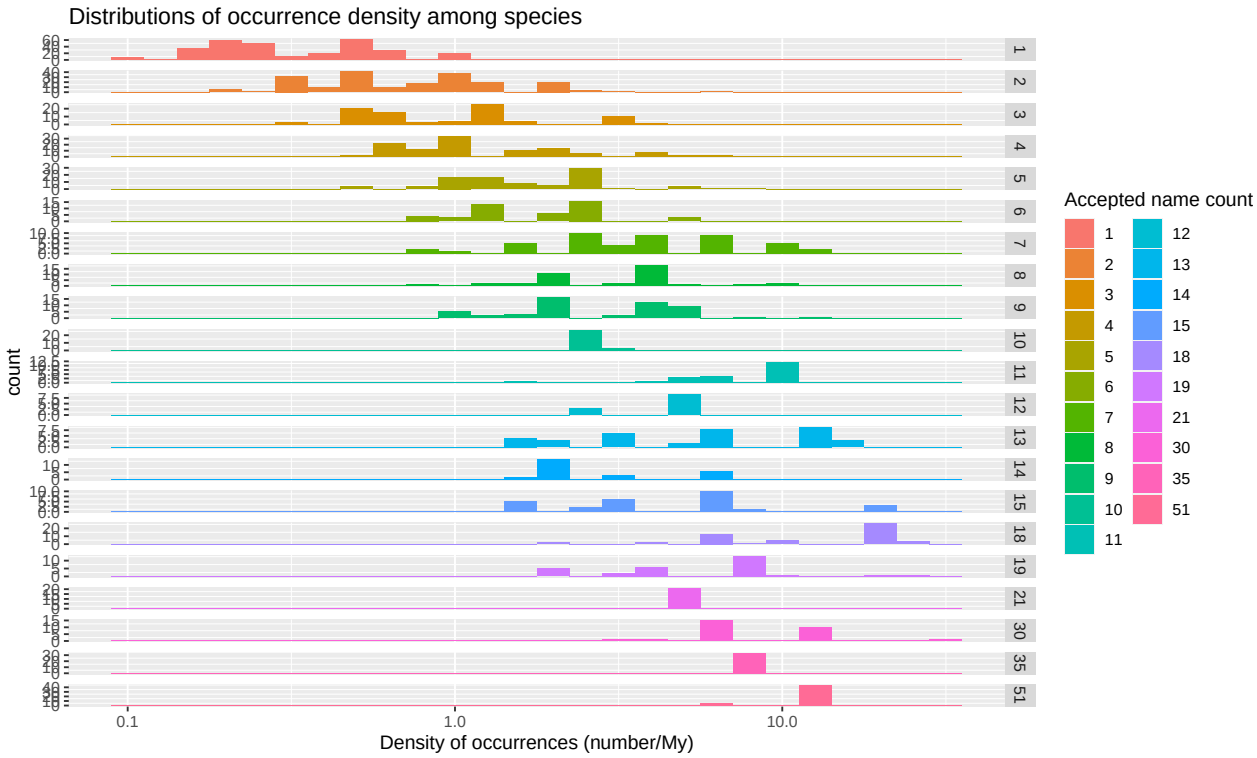
### Compare densities for single vs. combined ranges.

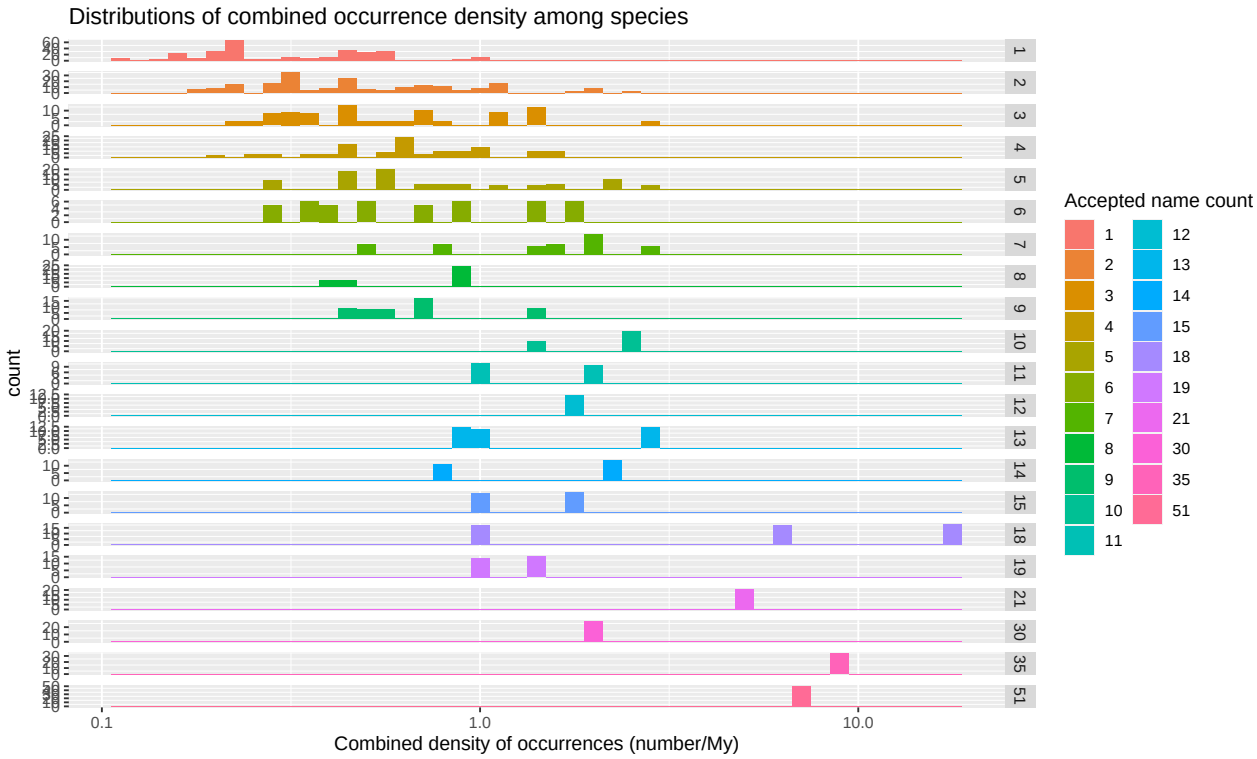

→ Densities are smaller and more concentrated with the combined ranges (larger time span + less ranges in

total because of the collepses into unique ranges).

### Compare densities by accepted name count (species only)

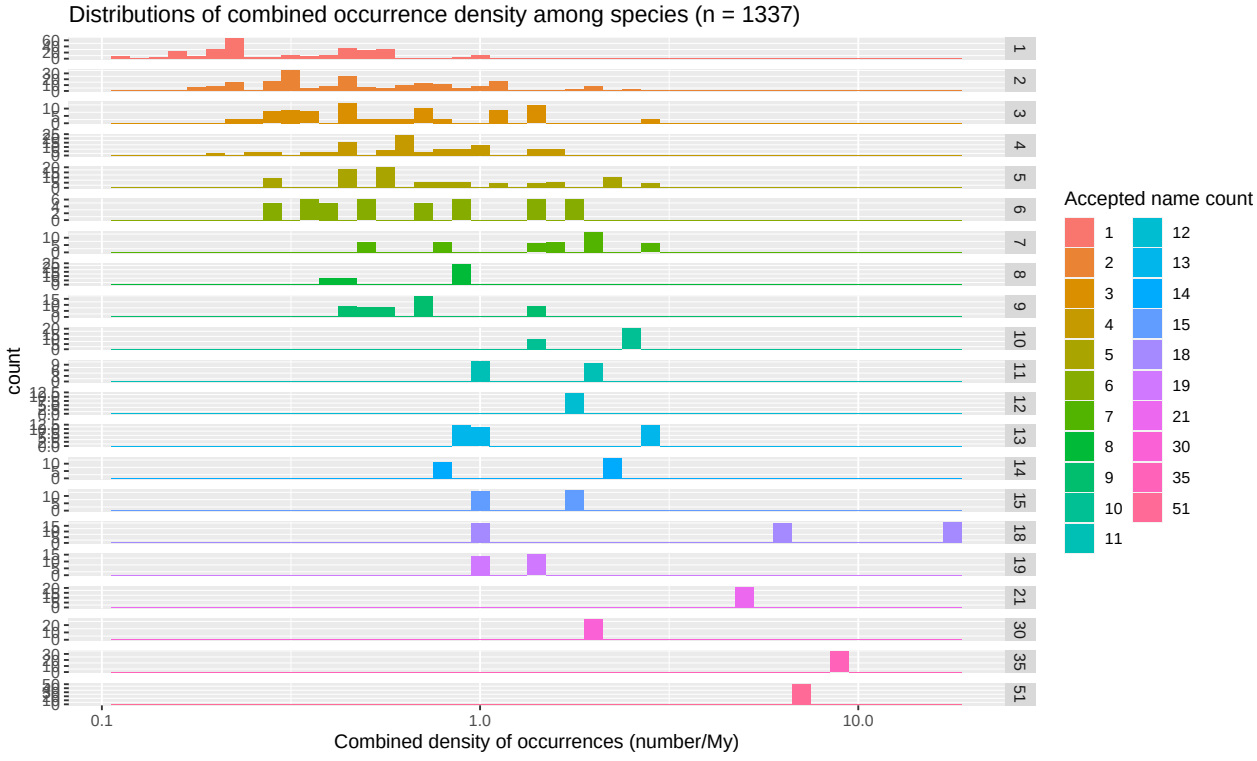
Let's focus now on the occurrences accepted at the species level because they are the one for which we can correct the abundance bias by subsampling the most concentrated combined intervals.



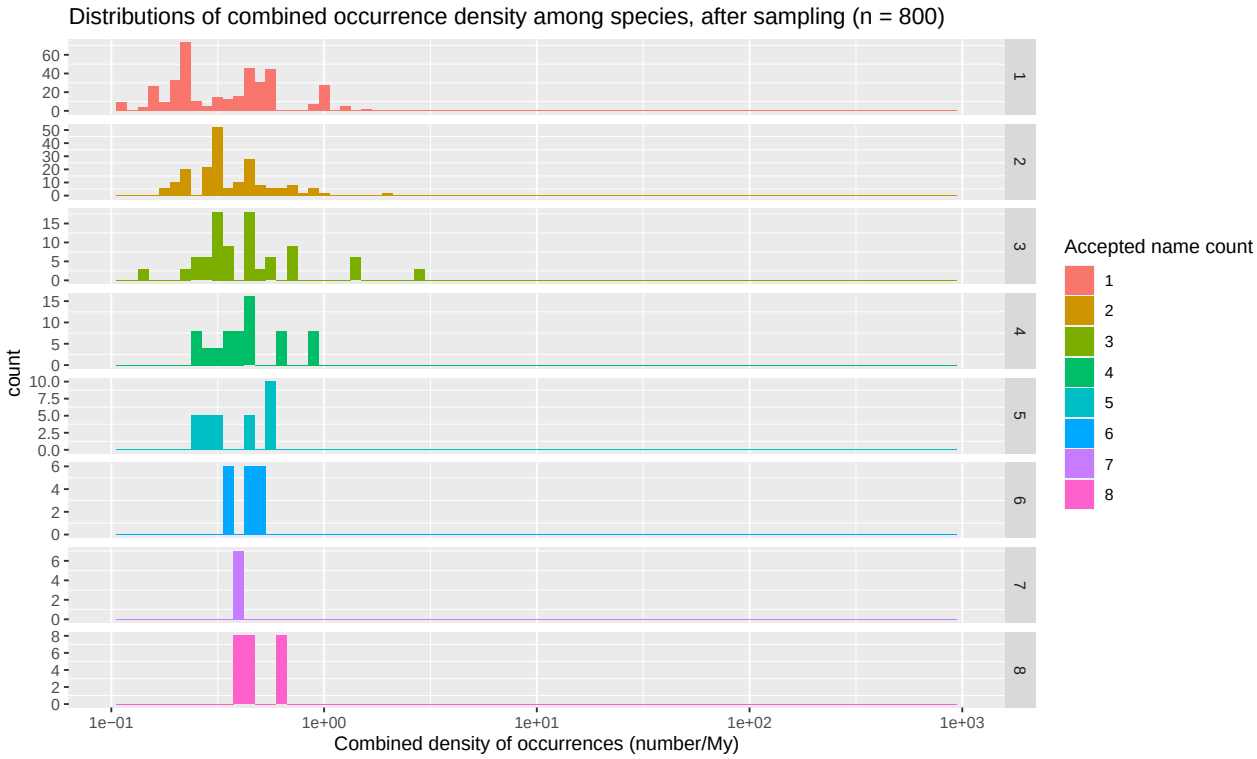



→ Their is a huge span of densities driven by the number of occurrences for the same species that we can reduce by subsampling the most concentrated intervals.

### Impact of correcting subsampling on density distributions (species only)



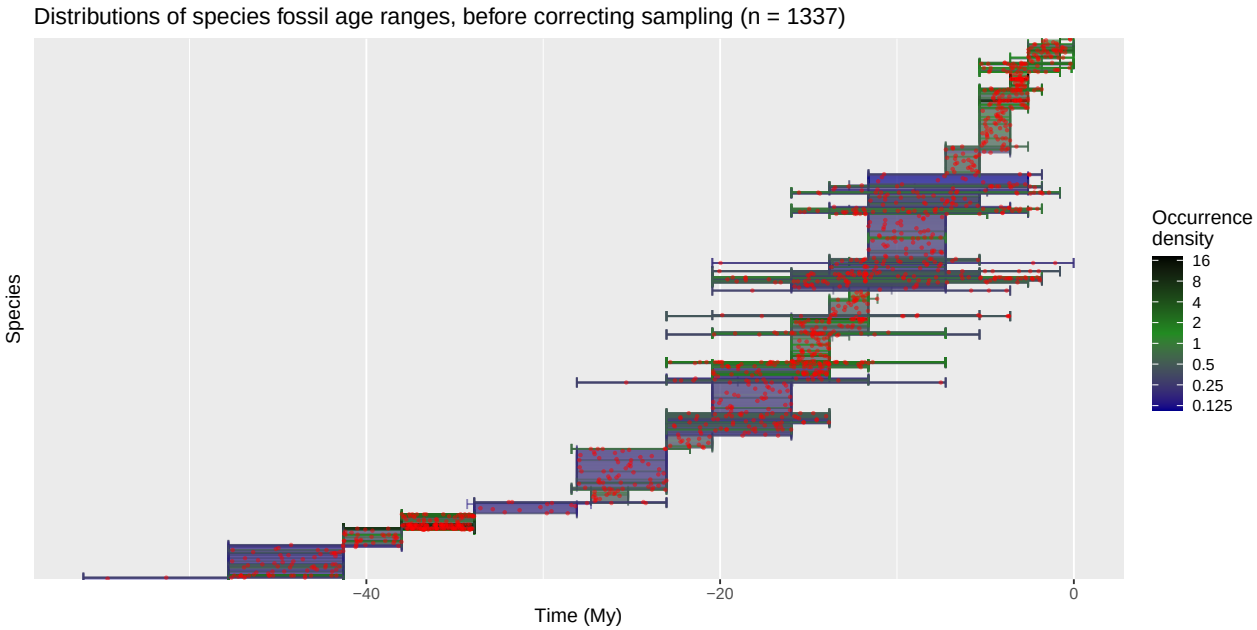
```
## Warning: Removed 16 rows containing missing values (geom_bar).
```

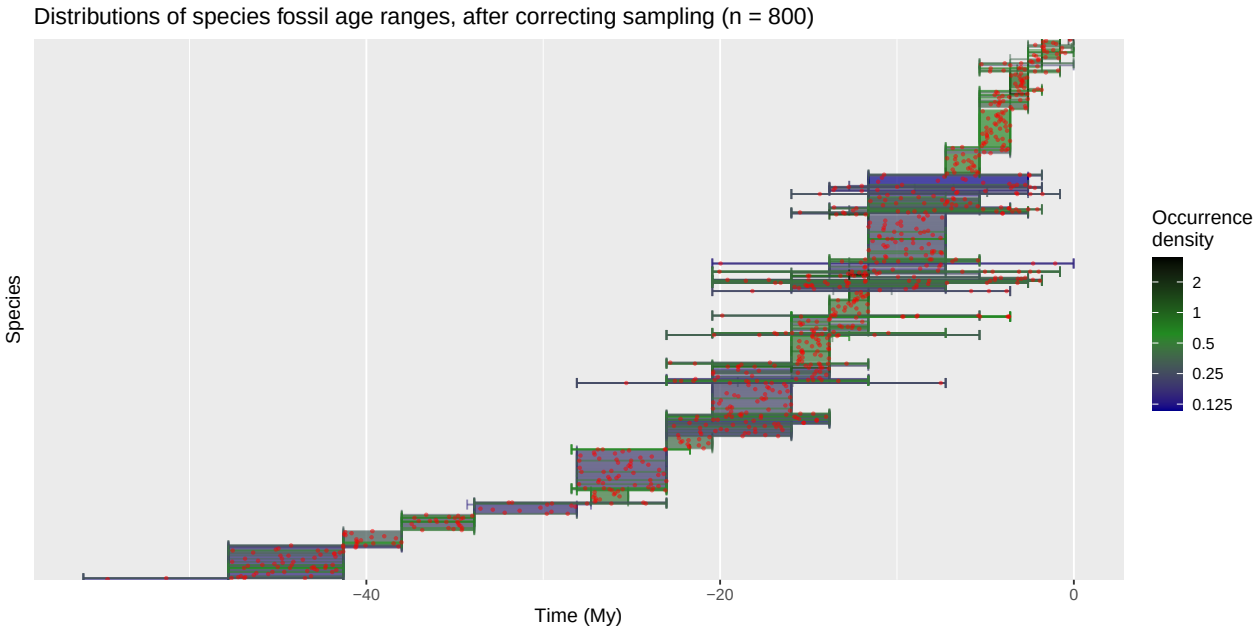



→ Subsampling successfully reduces the density span from 2 to 1 order of magnitude.

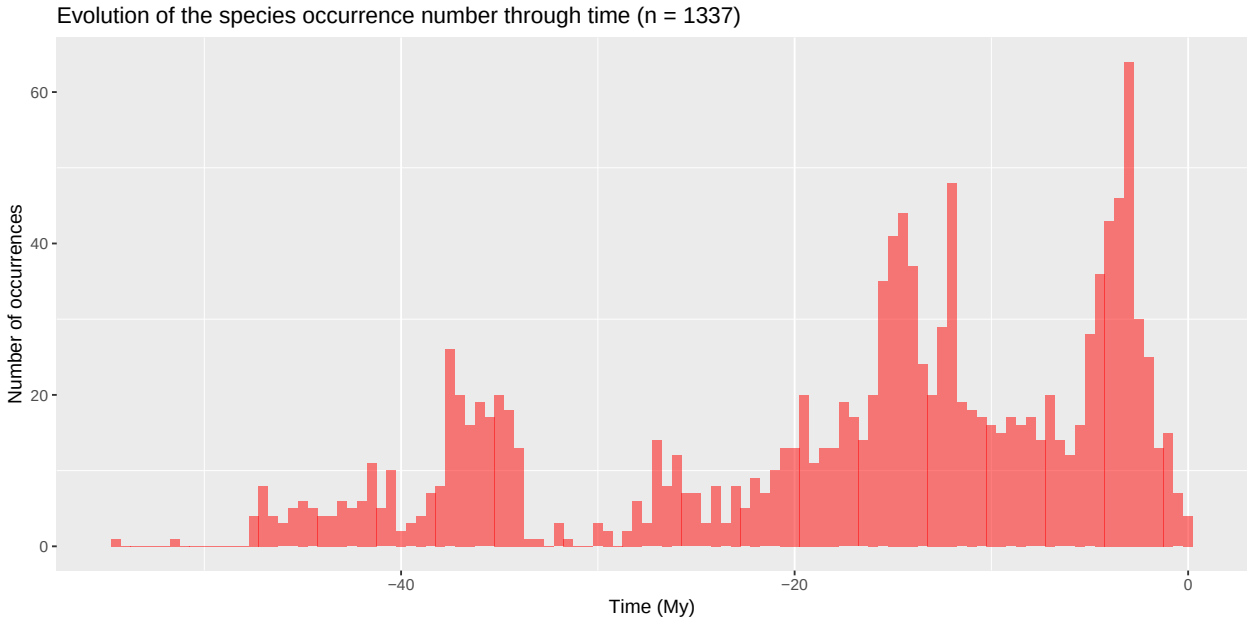
### Impact of subsampling on occurrences repartition (species only)

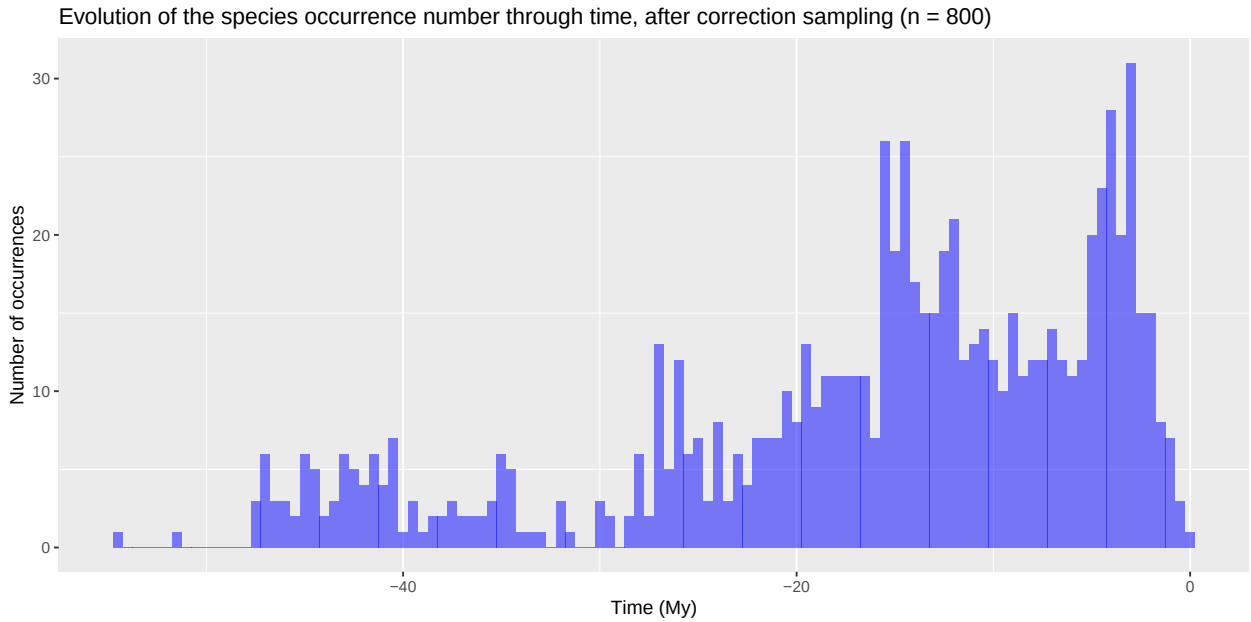
See what our distributions loook like after subsampling :



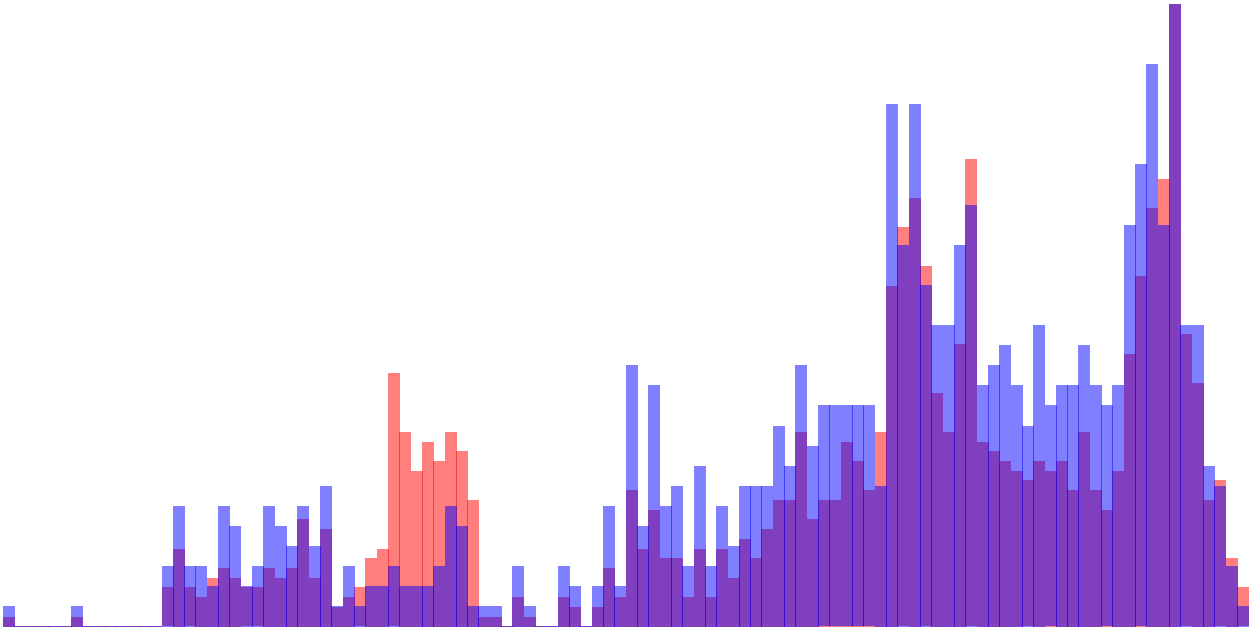



→ Some highly dense cluster became much more similar to the others.






If we superpose these 2 plots :



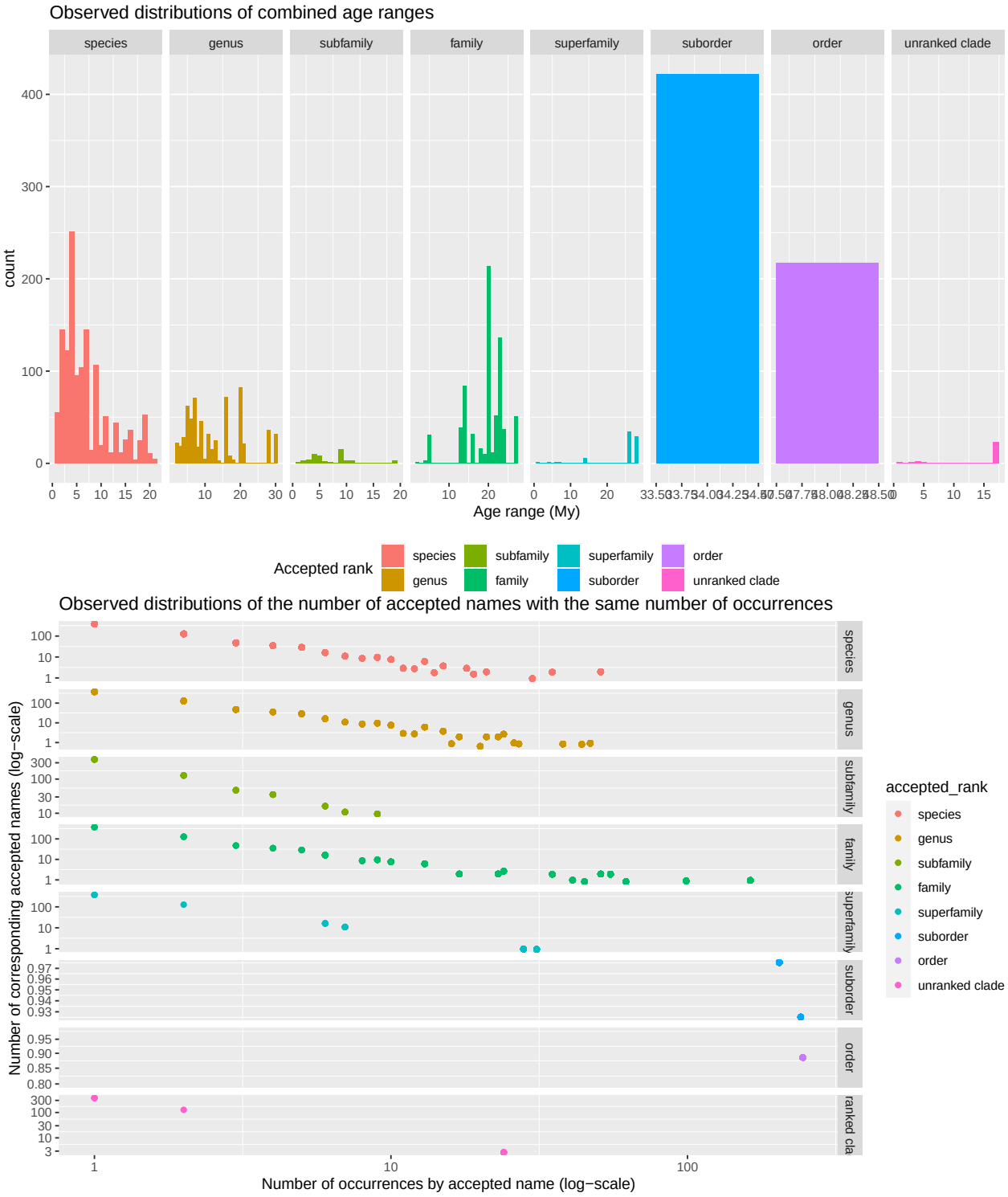
→ We get the new species occurrence repartition after subsampling correction, that could be used for doing inference with the occurrence birth-death model.

# New developments

## 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 (as a proxy for species duration).



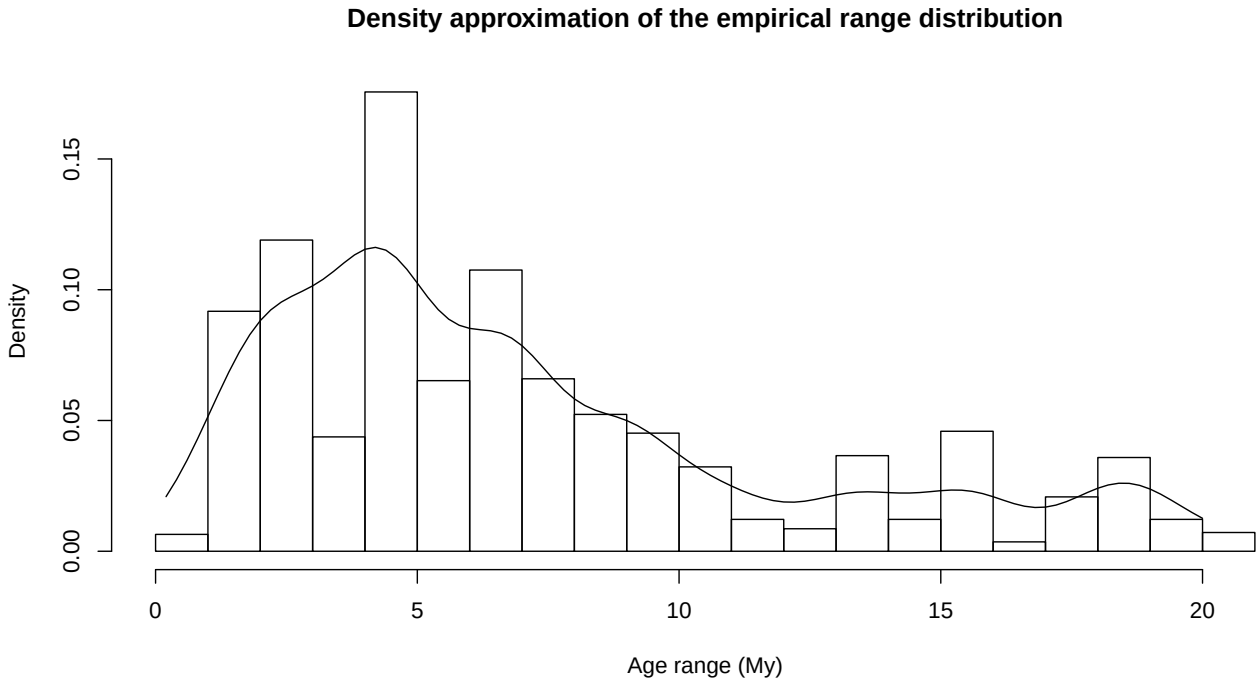

In a Poisson process with occurrence sampling rate $\omega$ and for a given time interval of length $t$, the probability of observing $N_t = k$ occurrences is given by the Poisson distribution of mean of parameter $\omega \times t$ :

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

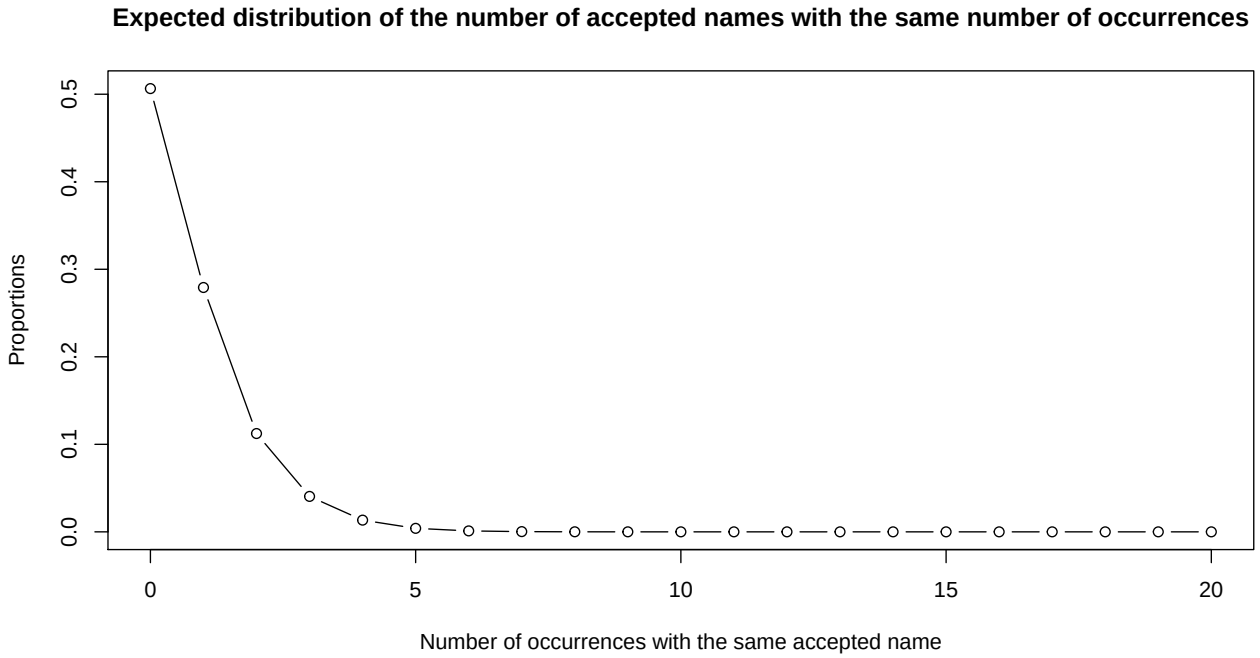
So in order to have the absolute probability of observing $N_0 = n$ occurrences we have to integrate over the full distribution of age ranges $t$, called $f(t)$ :

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

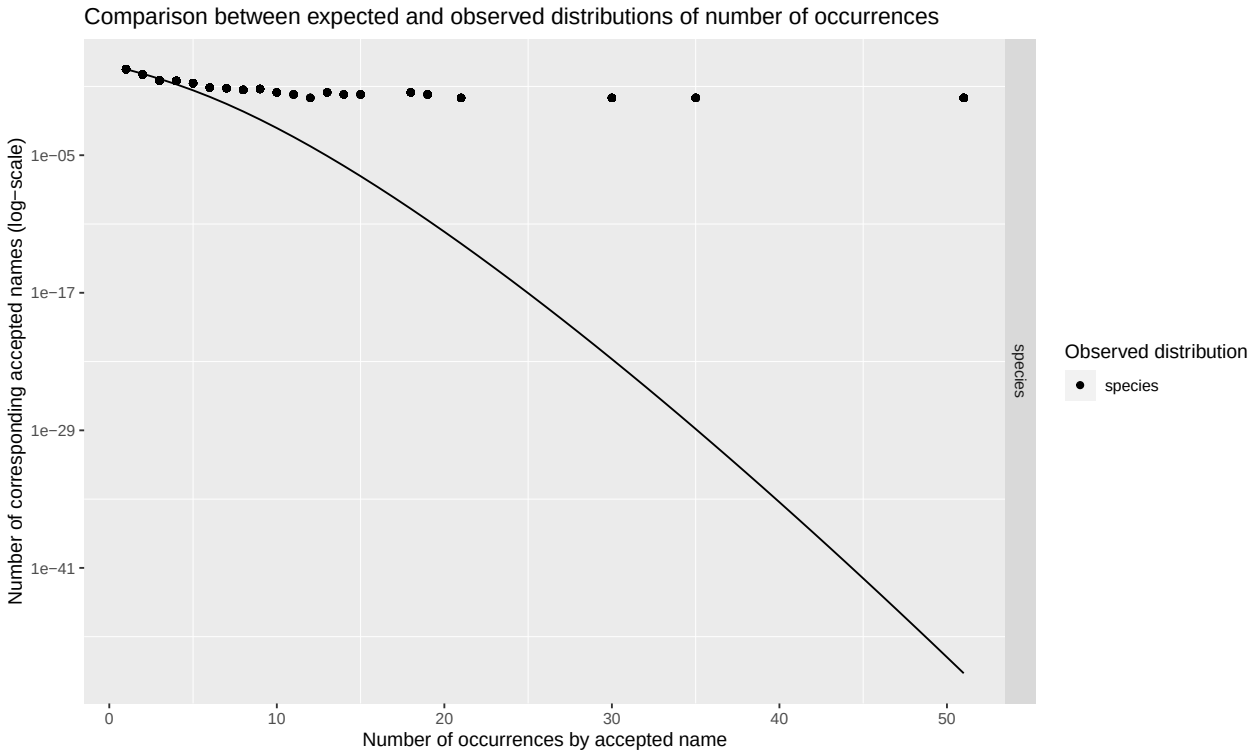
First, approximate this distribution :

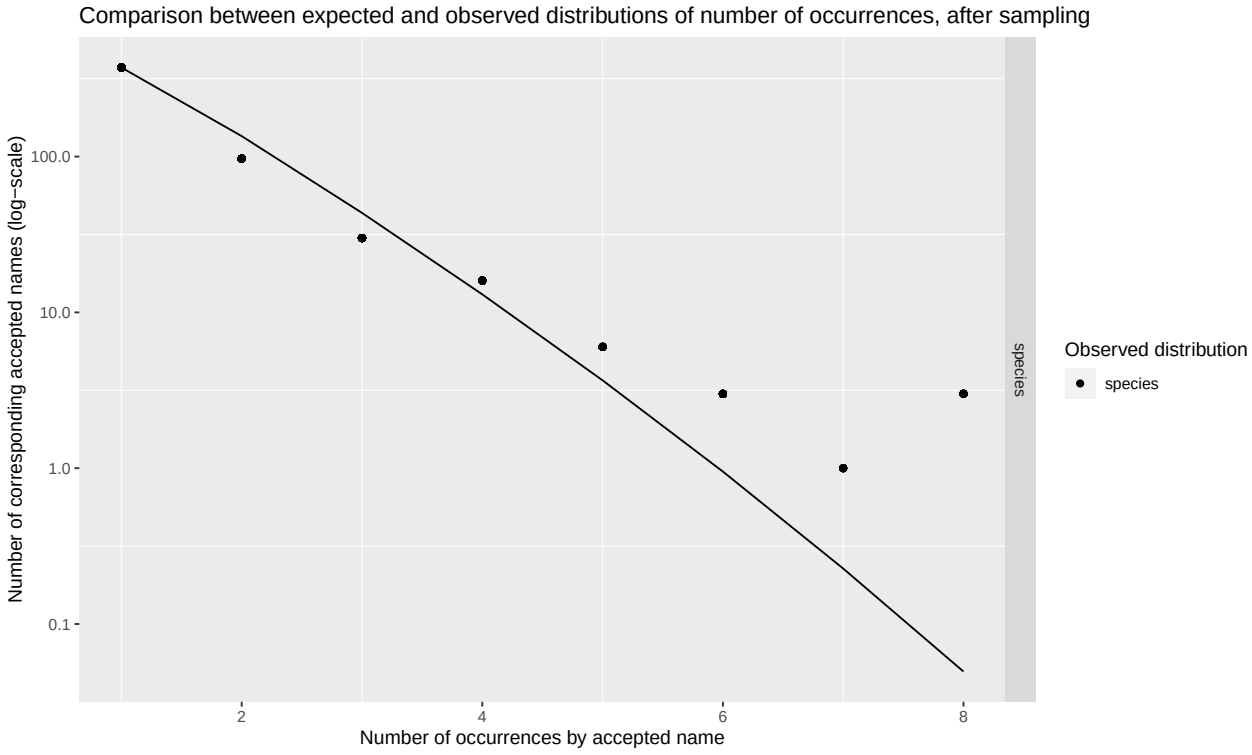



Then integrate and plot the expected distribution for a given omega :




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

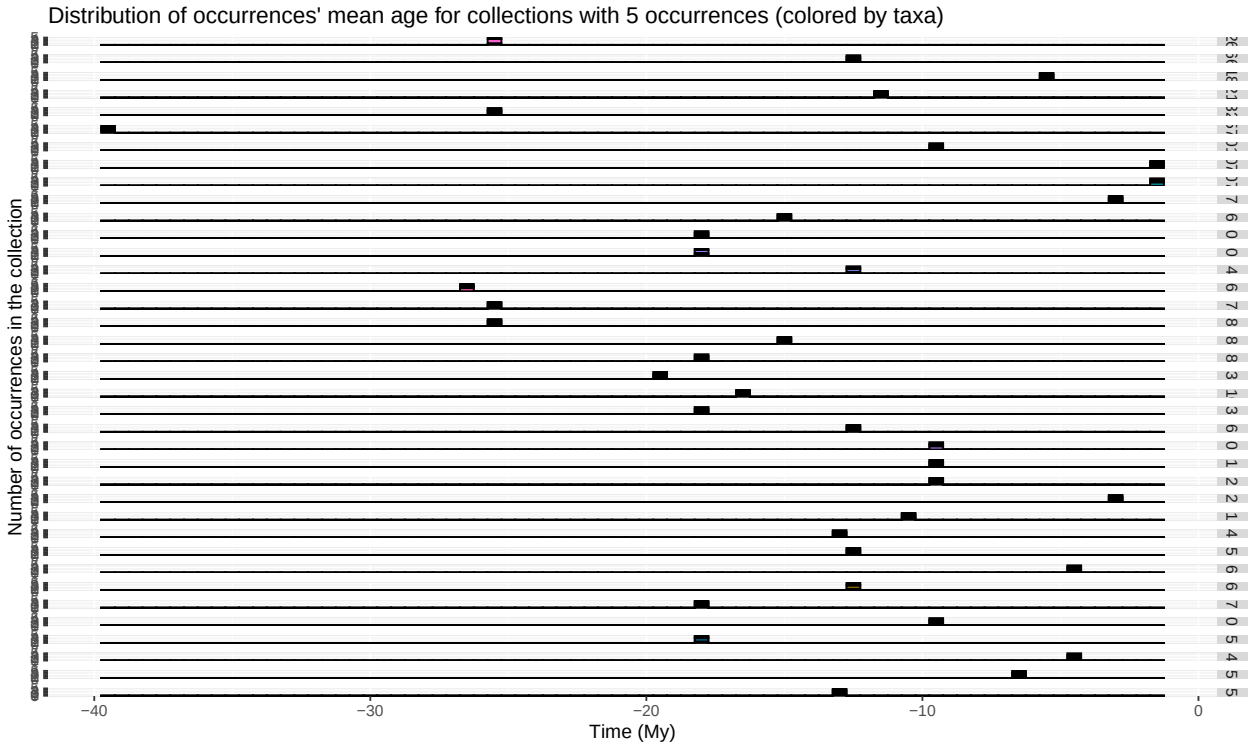


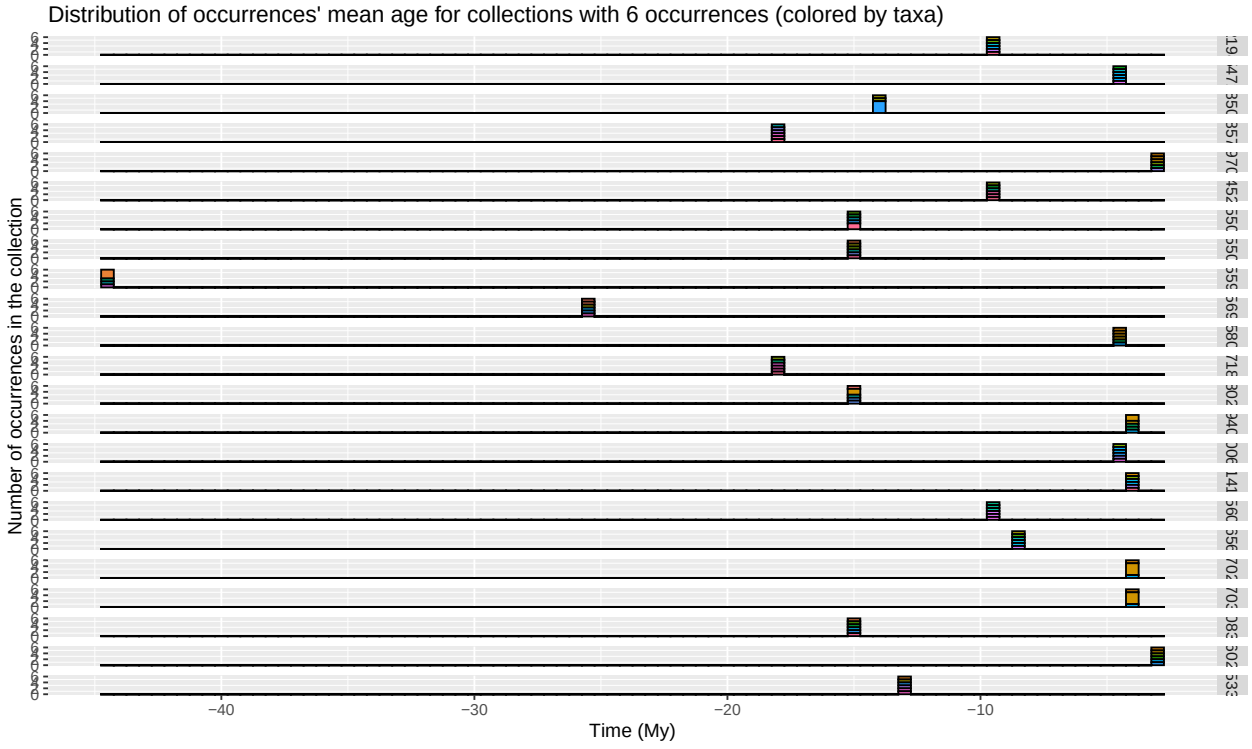



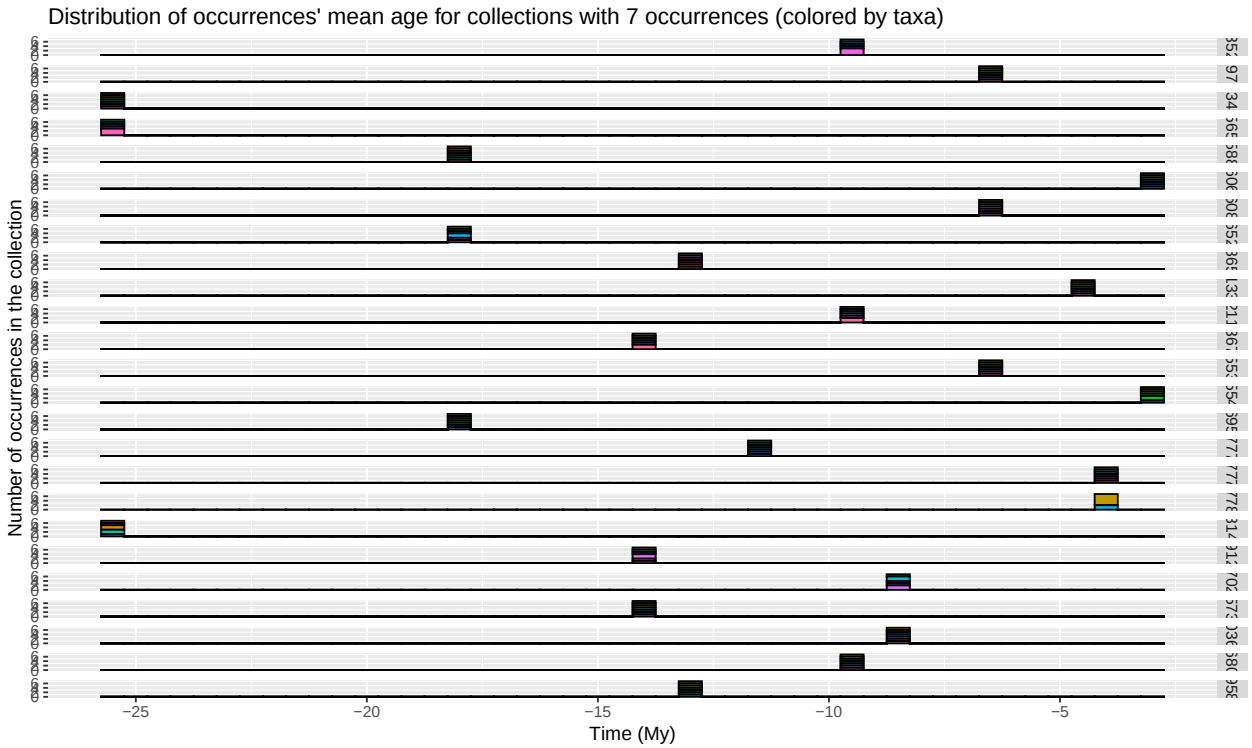

$\Longrightarrow$ Initial observations really do not fit the expectations, **species with more than 5 occurrences must remain very rare** ! But our subsampling seems to correct most of this bias.

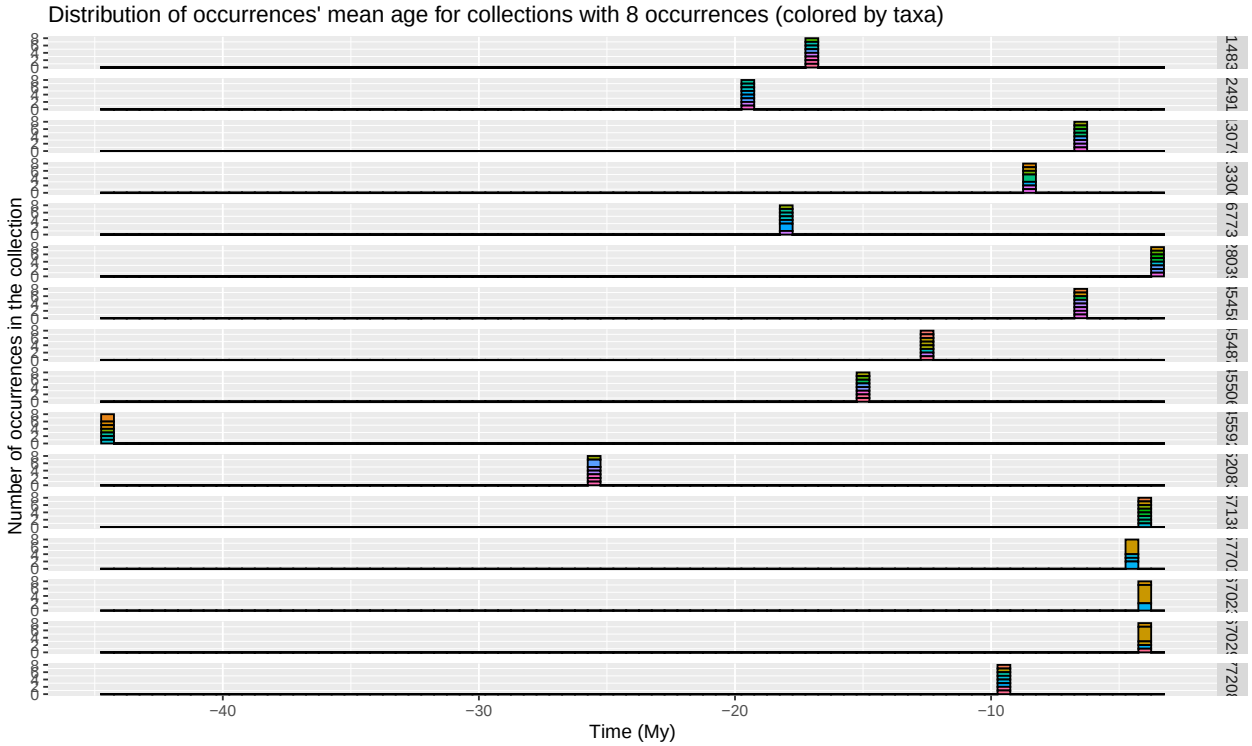
However, this method requires to make several arbitrary choices that may introduce new biases so we will instead subsample at other levels (palaeontological collection, geological formation). In each case only one occurrence will be sampled for the similarly identified, a process we will refer to as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

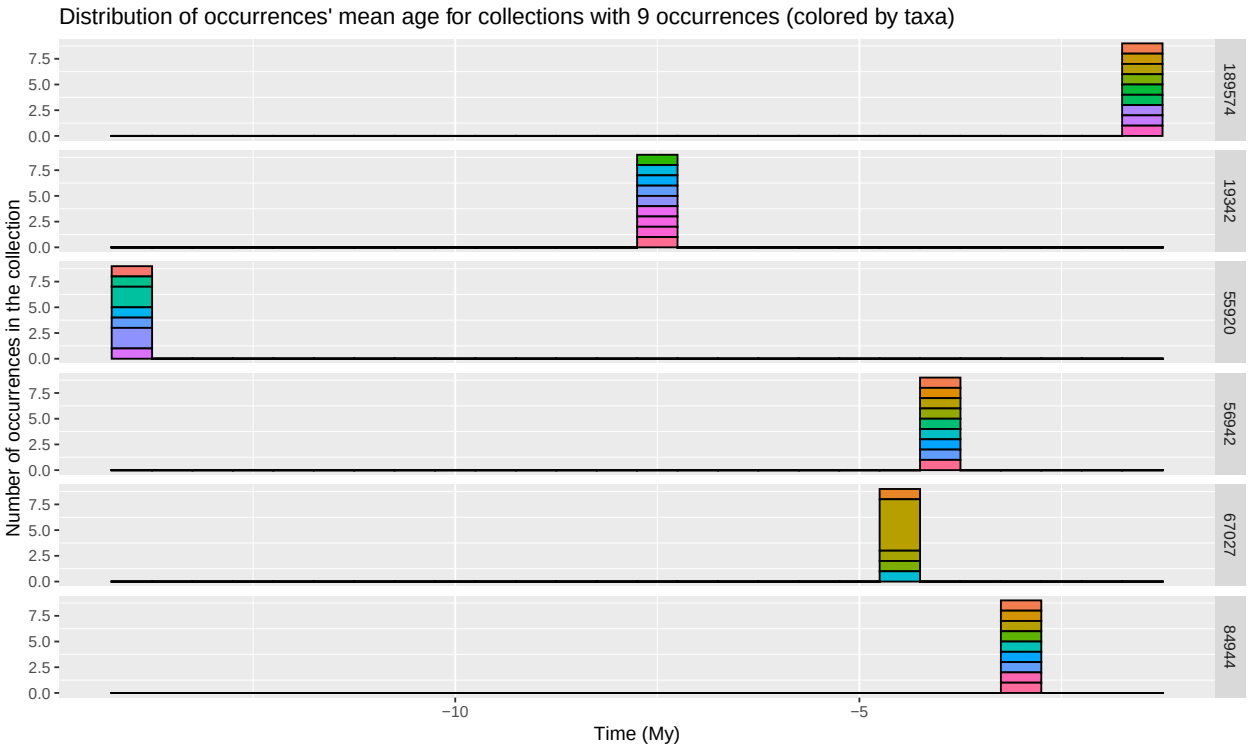
## Aggregate similarly identified occurrences in each collection

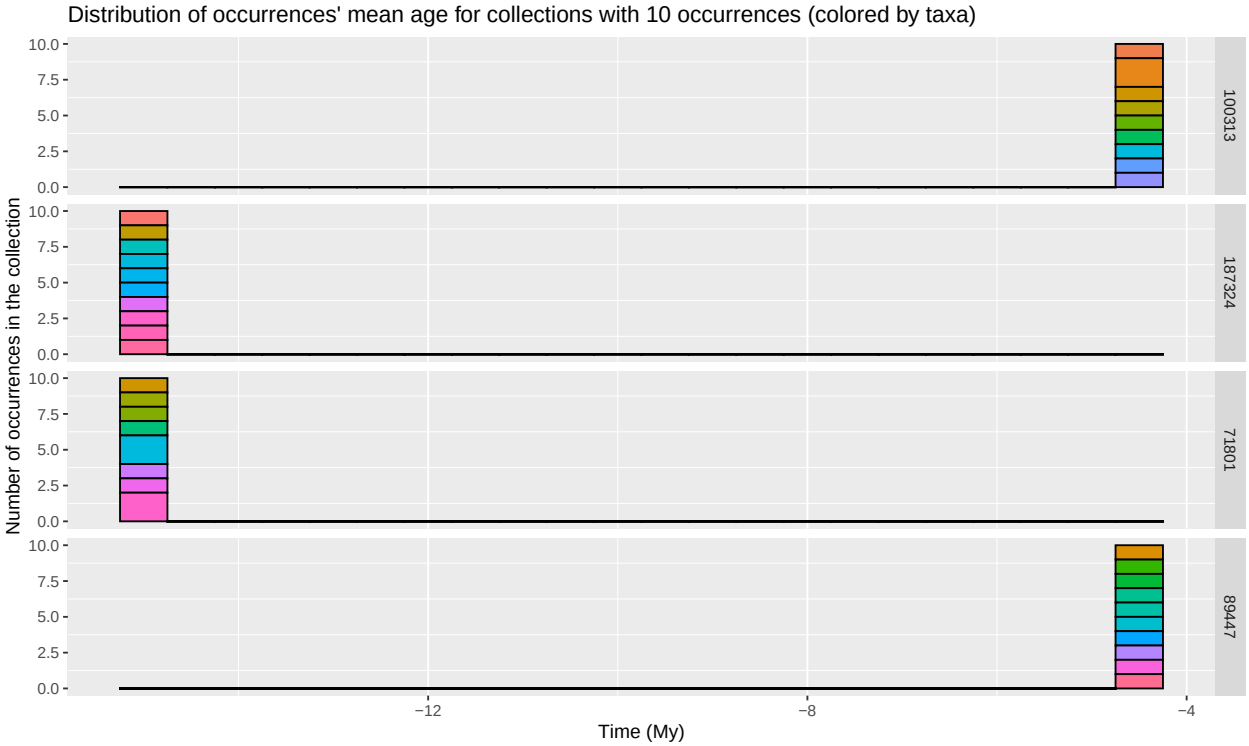


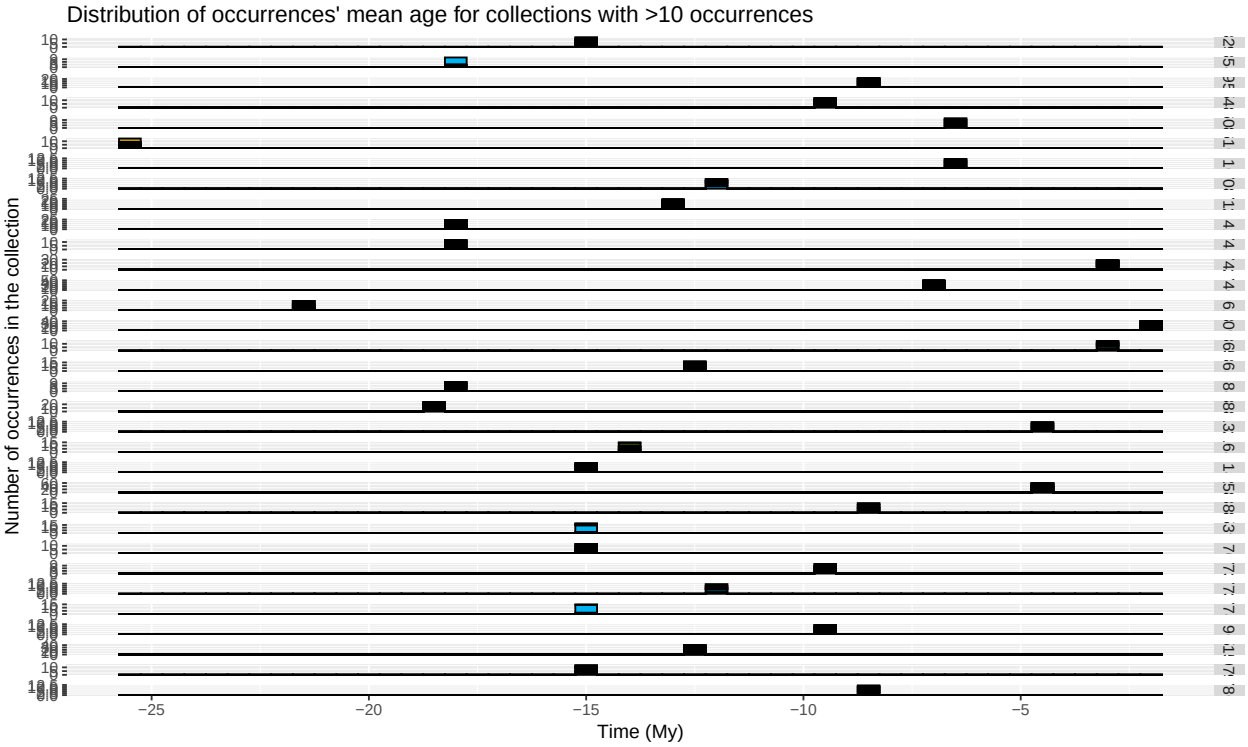




Distribution of occurrences' mean age for collections with 7 occurrences (colored by taxa)
Number of occurrences in the collection
−25
−20
−15
−10
−5
Time (My)


Distribution of occurrences' mean age for collections with 8 occurrences (colored by taxa)
Number of occurrences in the collection
−40
−30
−20
−10
Time (My)


Distribution of occurrences' mean age for collections with 9 occurrences (colored by taxa)
Number of occurrences in the collection
7.5
5.0
2.5
0.0
189574
19342
55920
56942
67027
84944
−10
−5
Time (My)


Distribution of occurrences' mean age for collections with 10 occurrences (colored by taxa)
Number of occurrences in the collection
10.0
7.5
5.0
2.5
0.0
100313
187324
71801
89447
−12
−8
−4
Time (My)




→ Each collection corresponds to a unique time interval (inferred from the unique age mean).

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                     3804               3556     248
## Number of occurrences (species only)      1437               1364      73
```

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

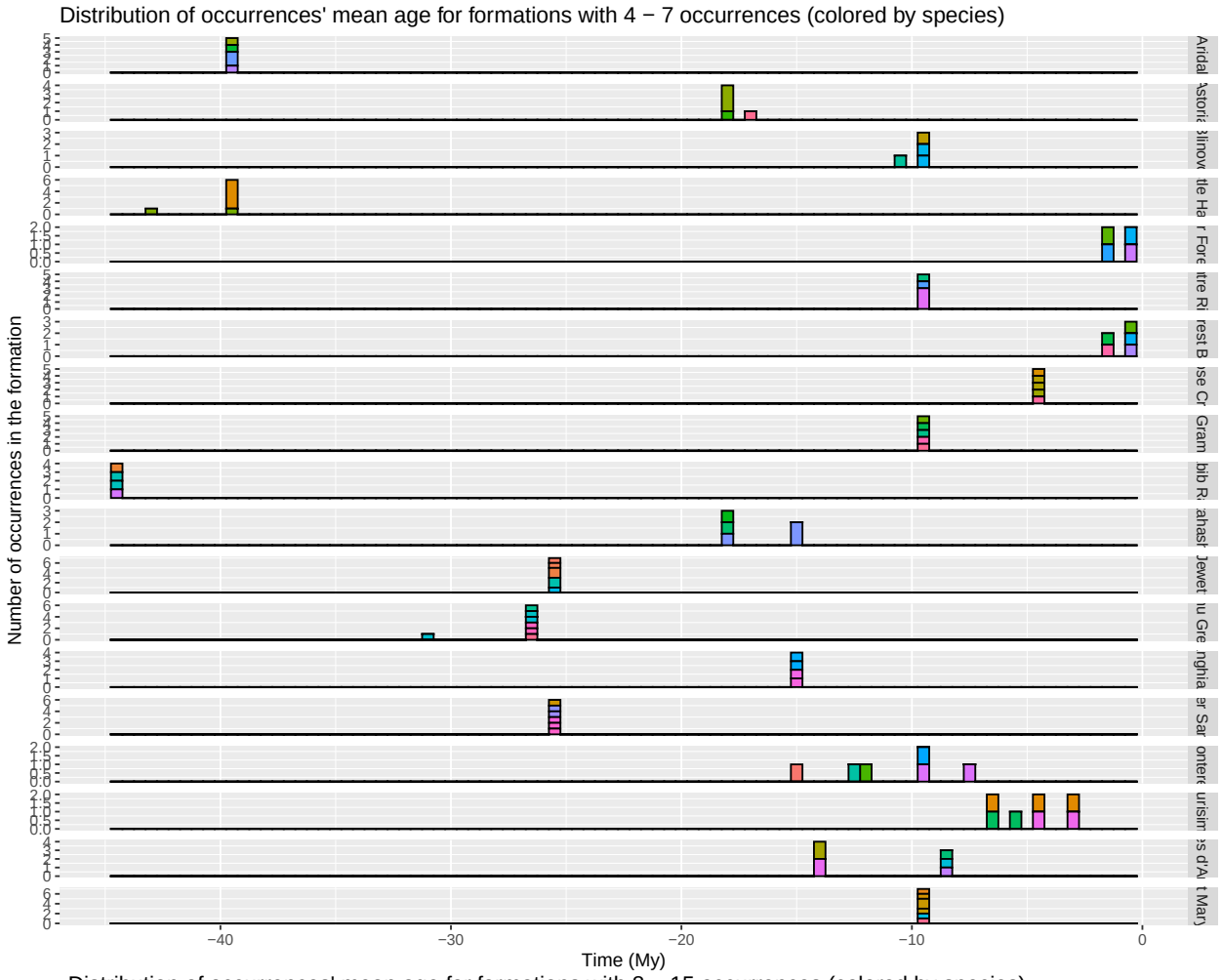
```
##
##         Aprixokogia kelloggi                     Balaena
##                            1                           2
##   Balaenoptera acutorostrata             Balaenopteridae
##                            2                           2
##                    Balaenula      Bohaskaia monodontoides
##                            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
```

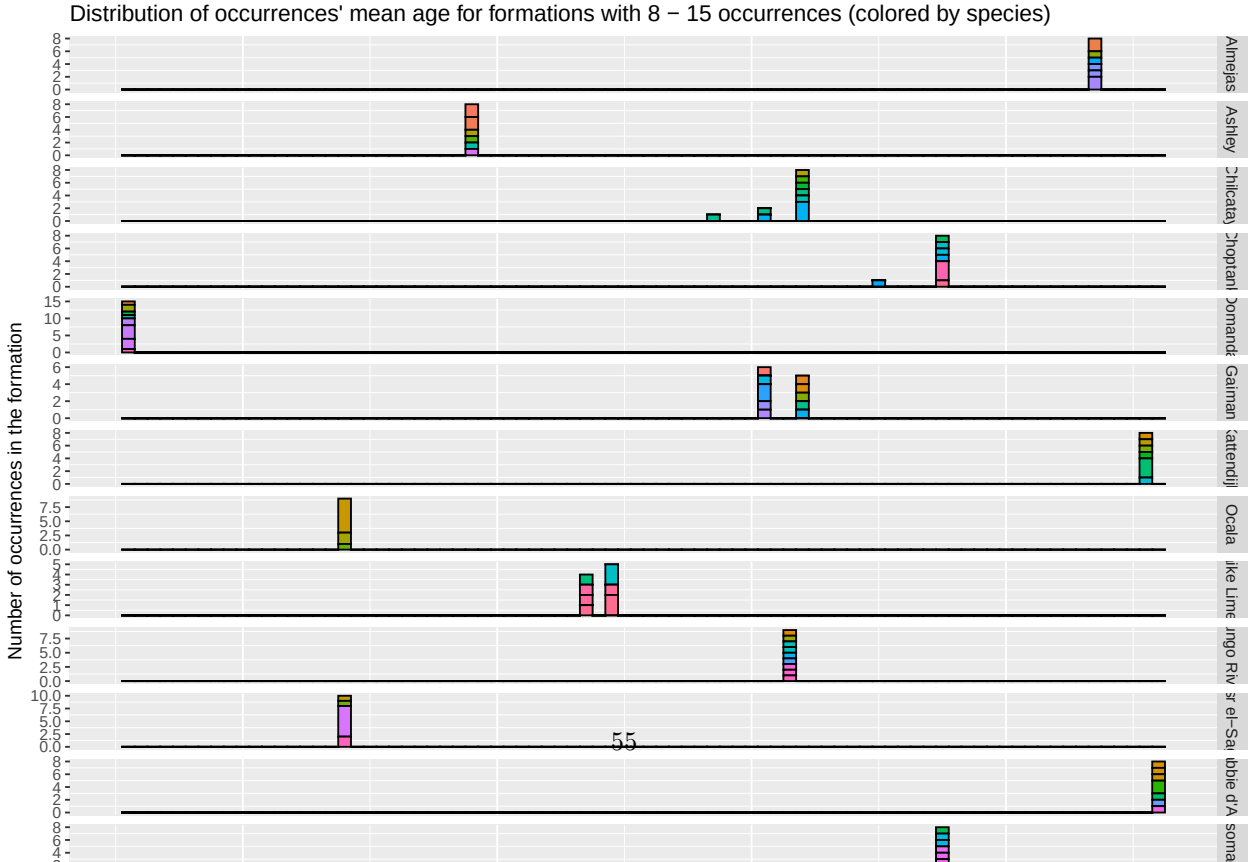
```
##     Lagenorhynchus harmatuki                   Megaptera
##                            1                           2
##      Mesoplodon longirostris                    Monodon
##                            2                           1
##     Ninoziphius platyrostris               Orycterocetus
##                            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
```

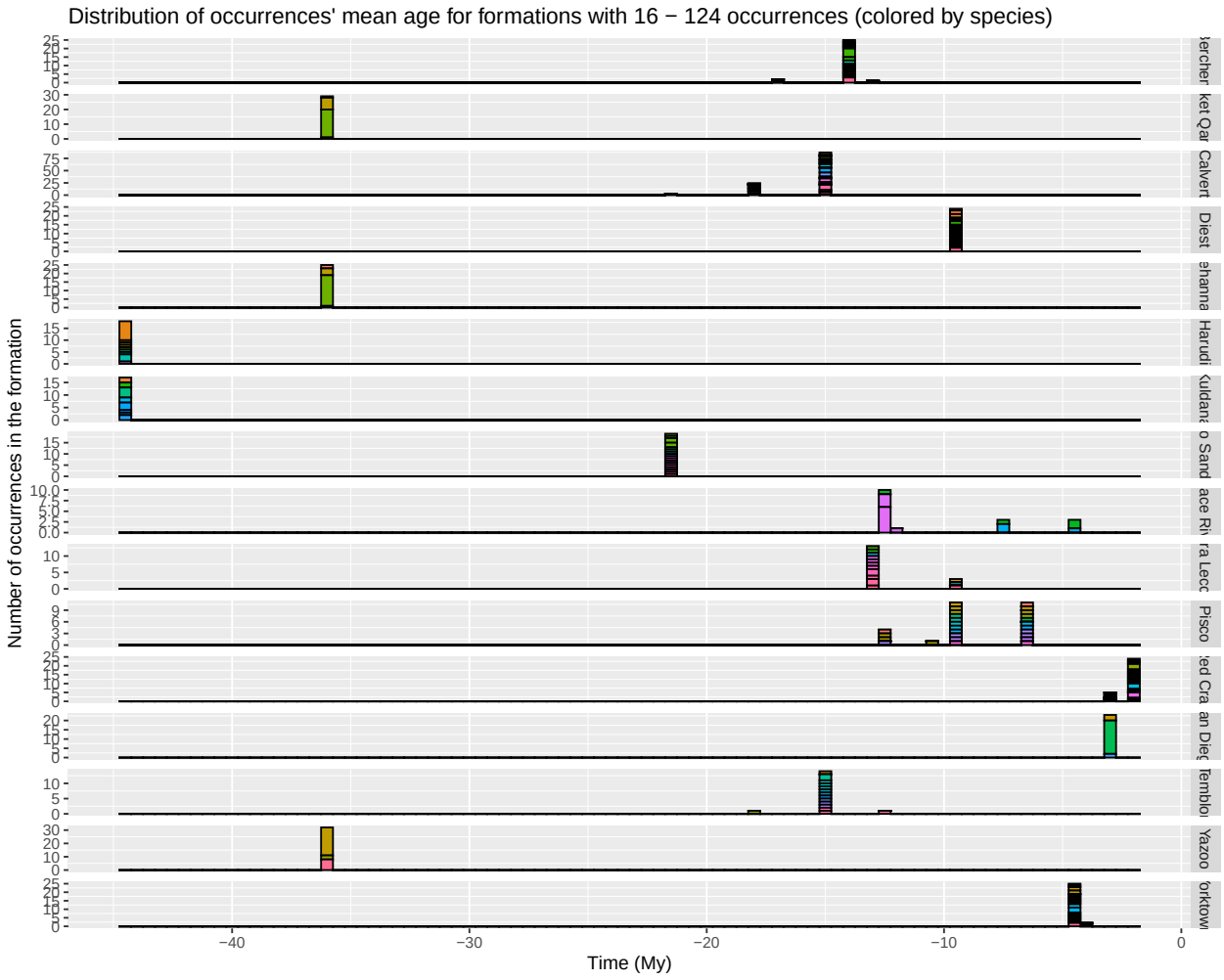
→ There are very few redundancies among the accepted names in collections so aggregating those won't reduce the abundance bias.

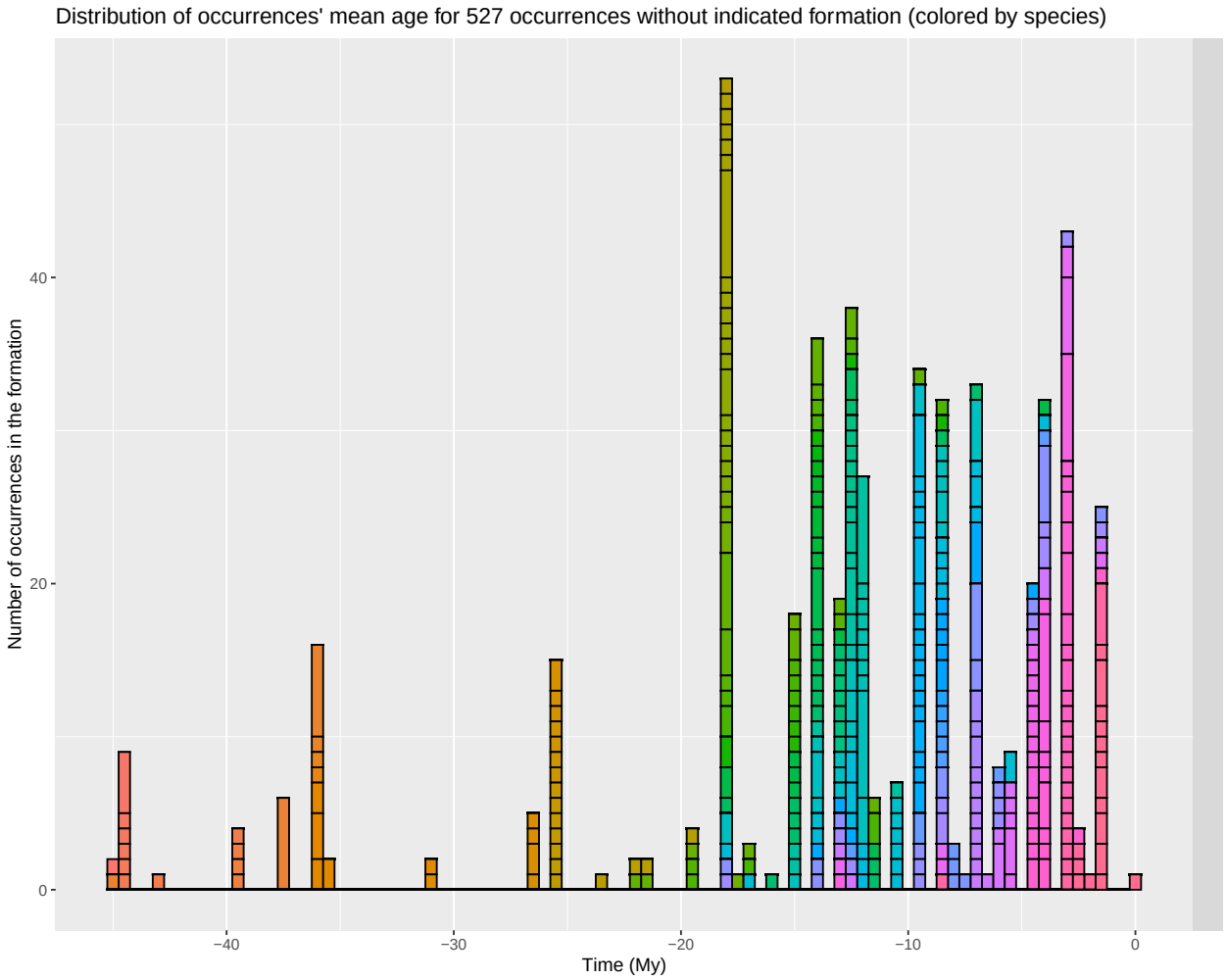
Instead, we may try to **aggregate occurrences with the same accepted name at the level of the geological formation** (ie subsample only one for each).

## Aggregate similarly identified occurrences in each formation



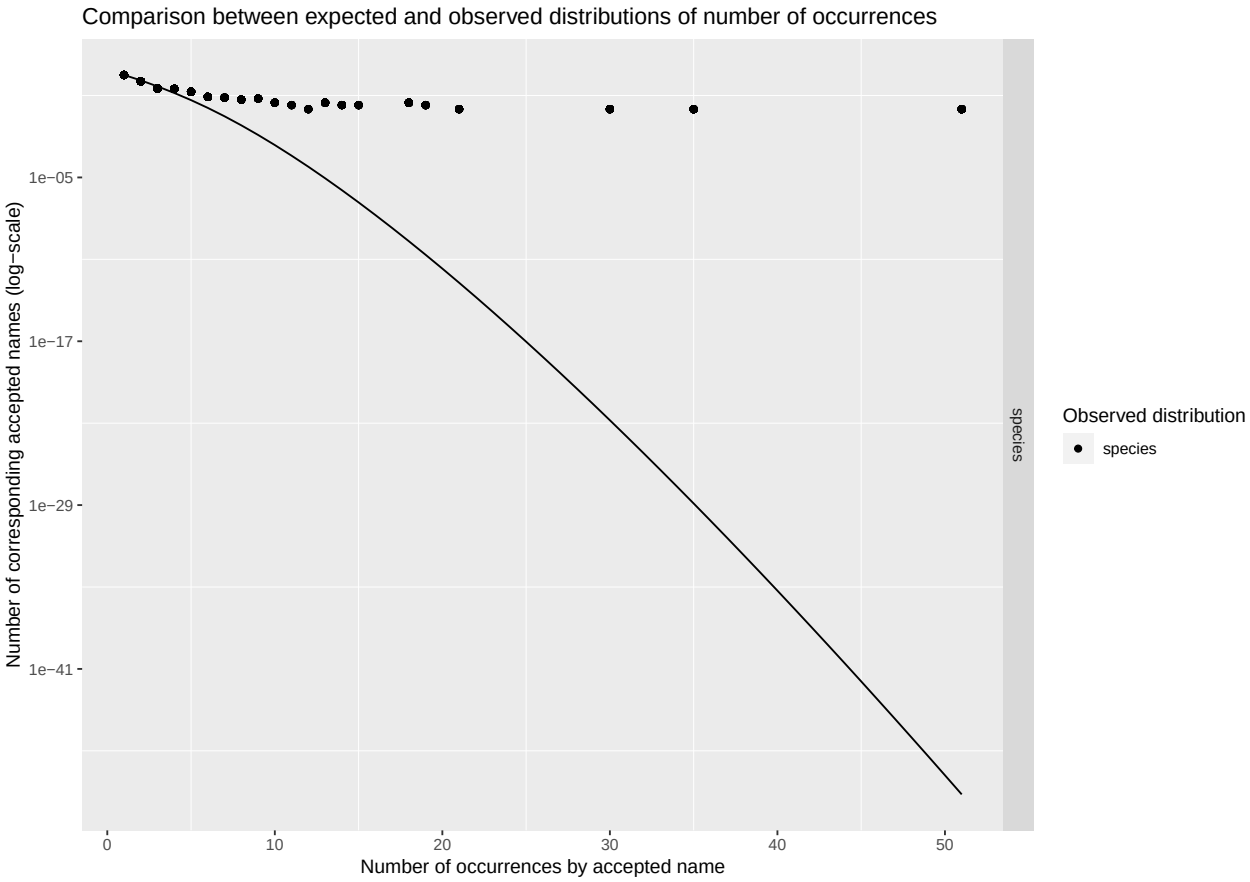

Distribution of occurrences' mean age for formations with 16 − 124 occurrences (colored by species)
Number of occurrences in the formation
Time (My)
−40
−30
−20
−10
0
Calvert
Harudi
Pisco
Yazoo

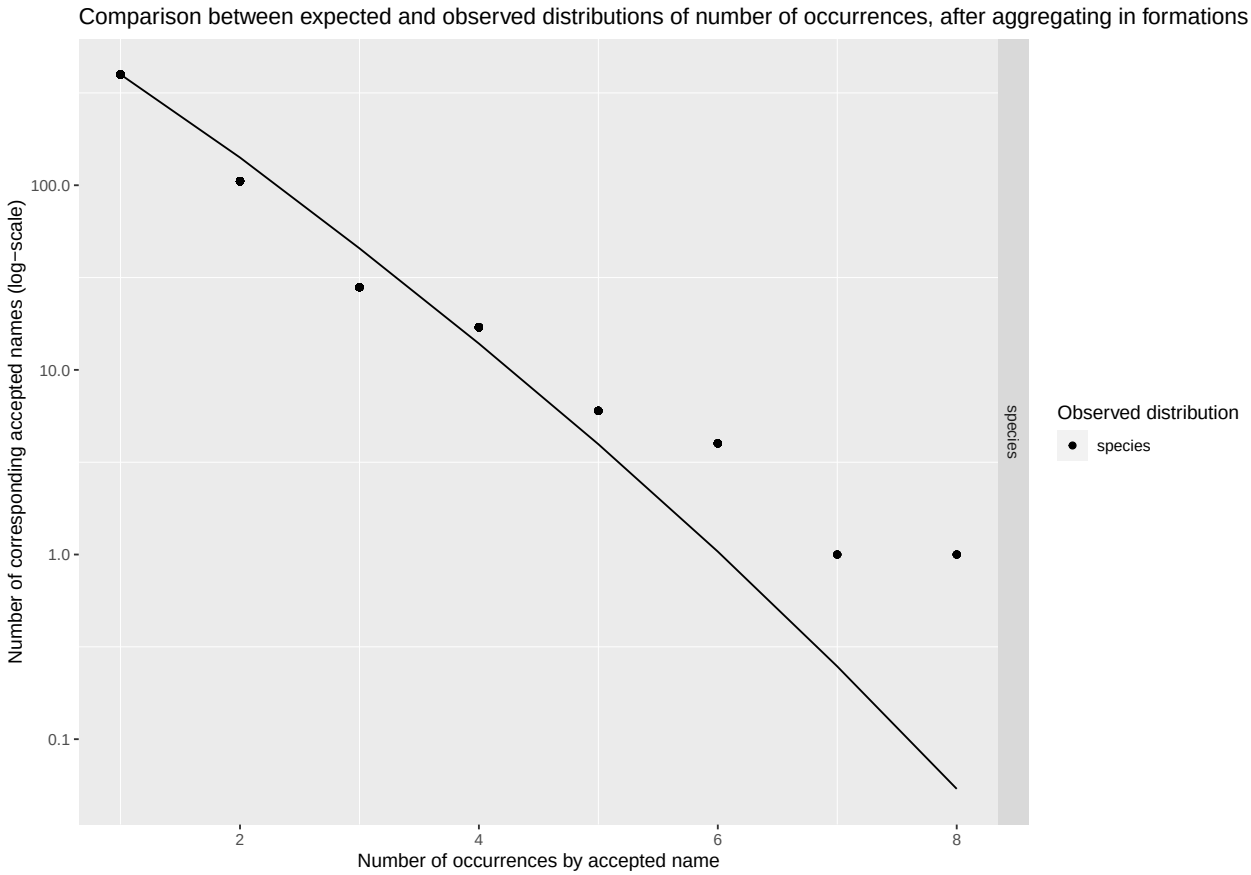



→ For most formations each species seems to be restricted to only one age, so as expected we are not loosing too much information when aggregating them to a single occurrence.

```
##                                    Cetacea_occ Cetacea_occ_aggreg removed
## Number of occurrences                     3804               1983    1821
## Number of occurrences (species only)      1437                829     608
```

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


Comparison between expected and observed distributions of number of occurrences
Number of corresponding accepted names (log−scale)
1e−05
1e−17
1e−29
1e−41
0
10
20
30
40
50
Number of occurrences by accepted name
species
Observed distribution
species

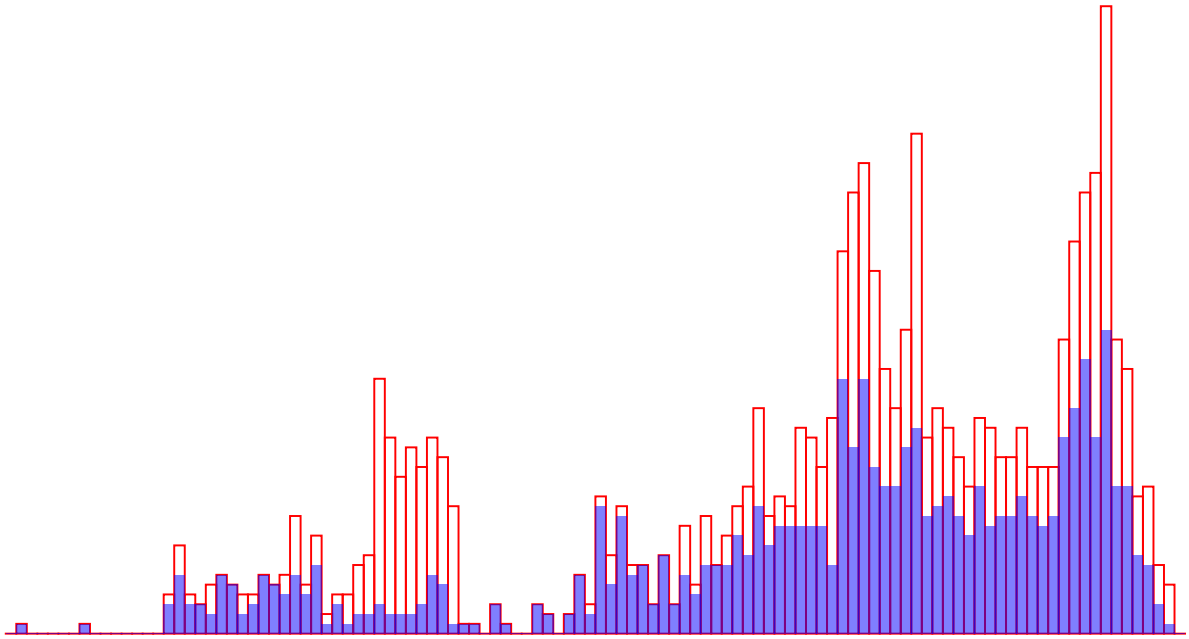



→ The bias can be corridered as completely corrected.

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

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

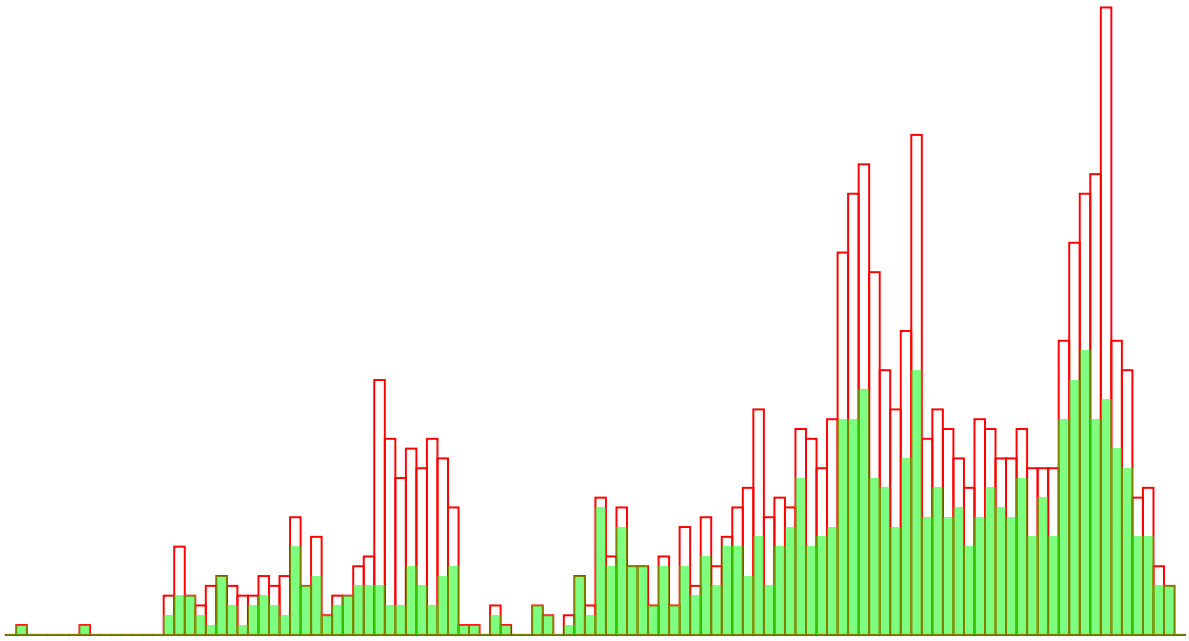
Initial occurrences distribution (red) and comparison after sub−sampling (blue)



```
## 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 aggregating in formations (green)

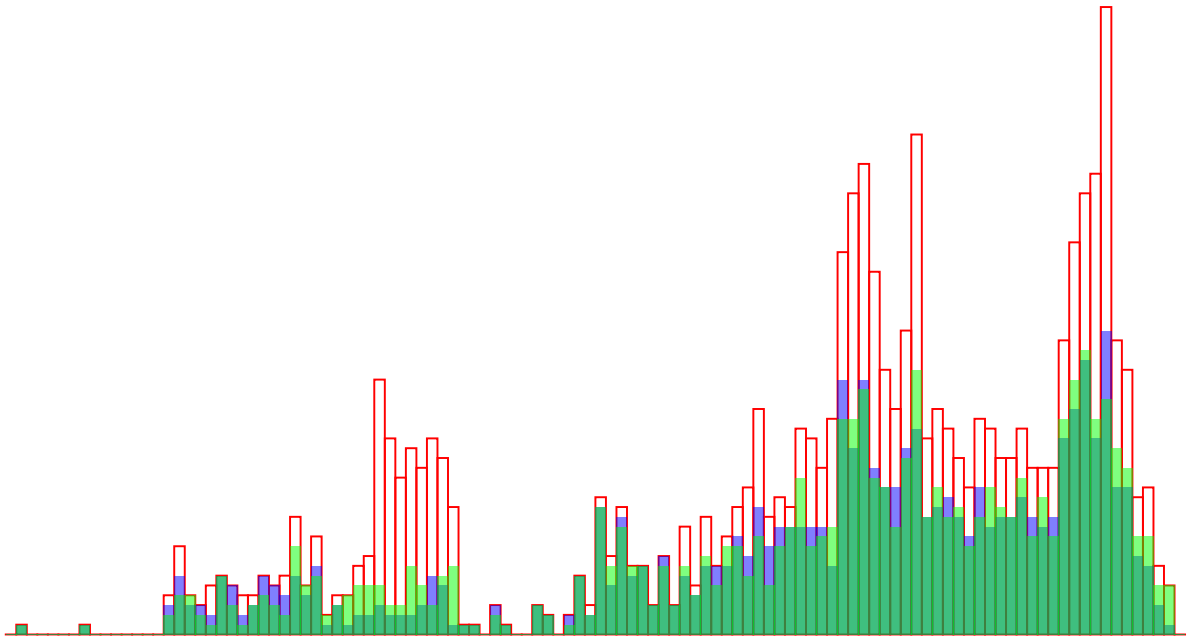


```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## 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)



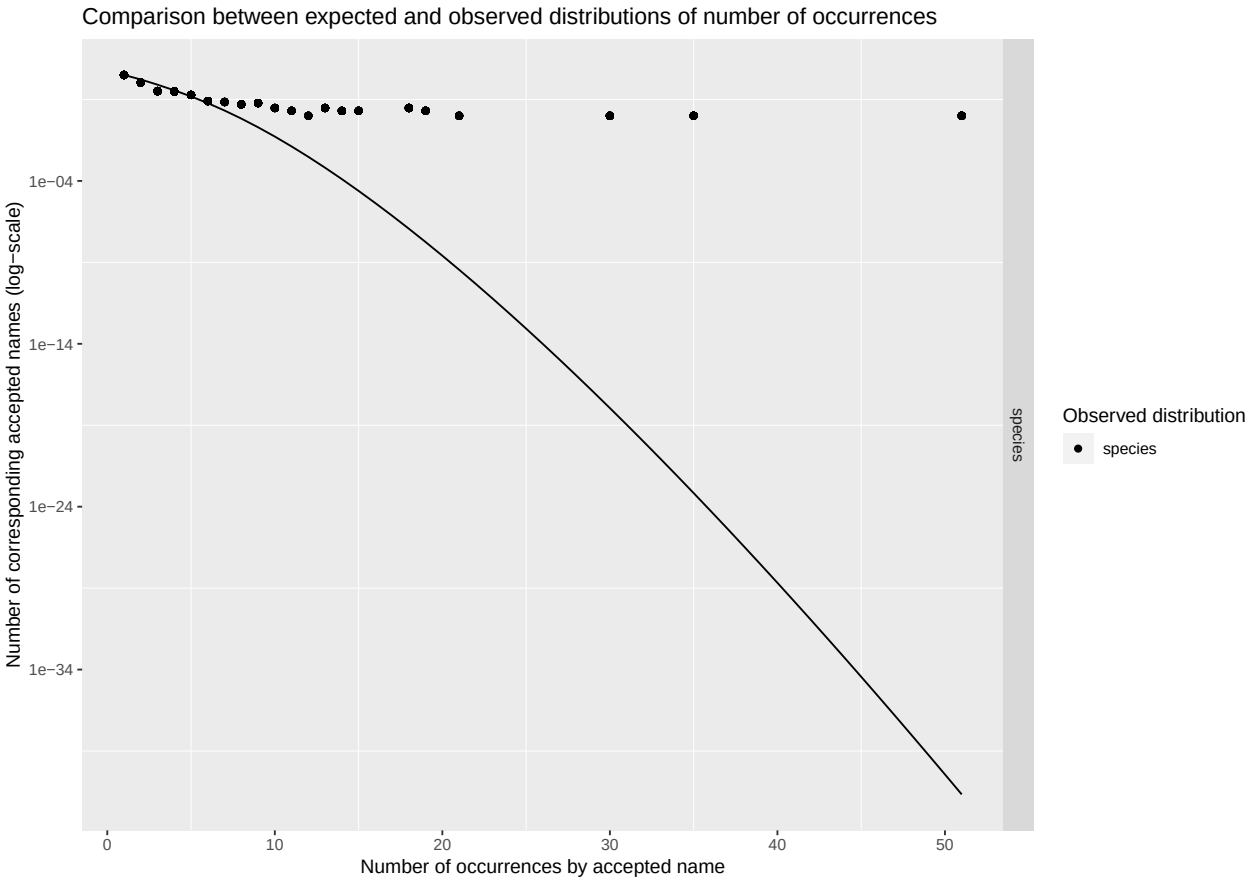
→ Comparing with the initial occurrences distribution and with the distribution after our first sub-sampling it appears that both methods lead to very similar distributions. This conforts us about the robustness of those approaches.

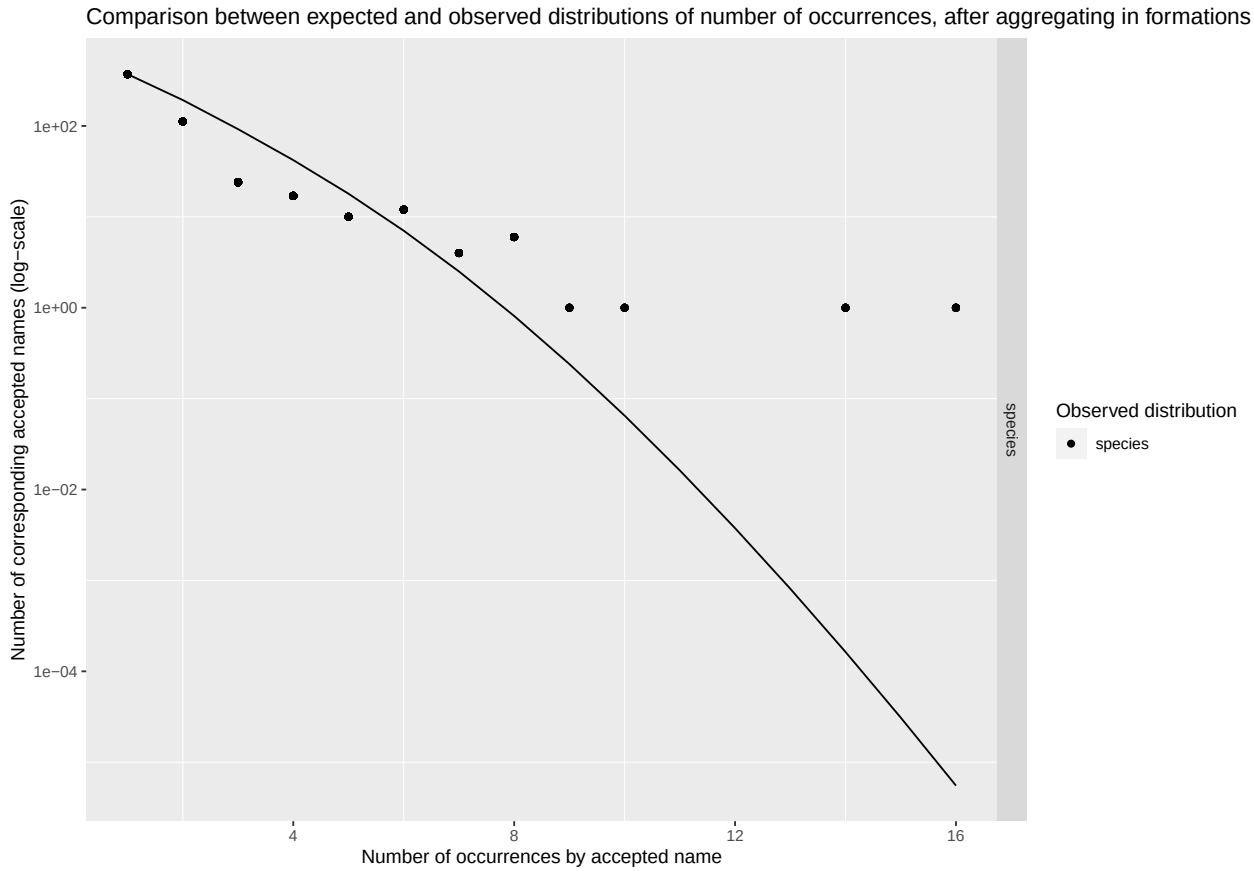
However, we have to take into account the occurrences that do not have any indicated geological formation to subsample them separately. To approximate geological formation we chose to proceed to the aggregation based on the combination of the country and the early stratigraphic interval.

## Aggregate occurrences without formation by country + early interval

```
##                                   Cetacea_occ Cetacea_occ_aggreg removed
## Number of occurrences                    3804               2644    1160
## Number of occurrences (species only)     1437                982     455
```

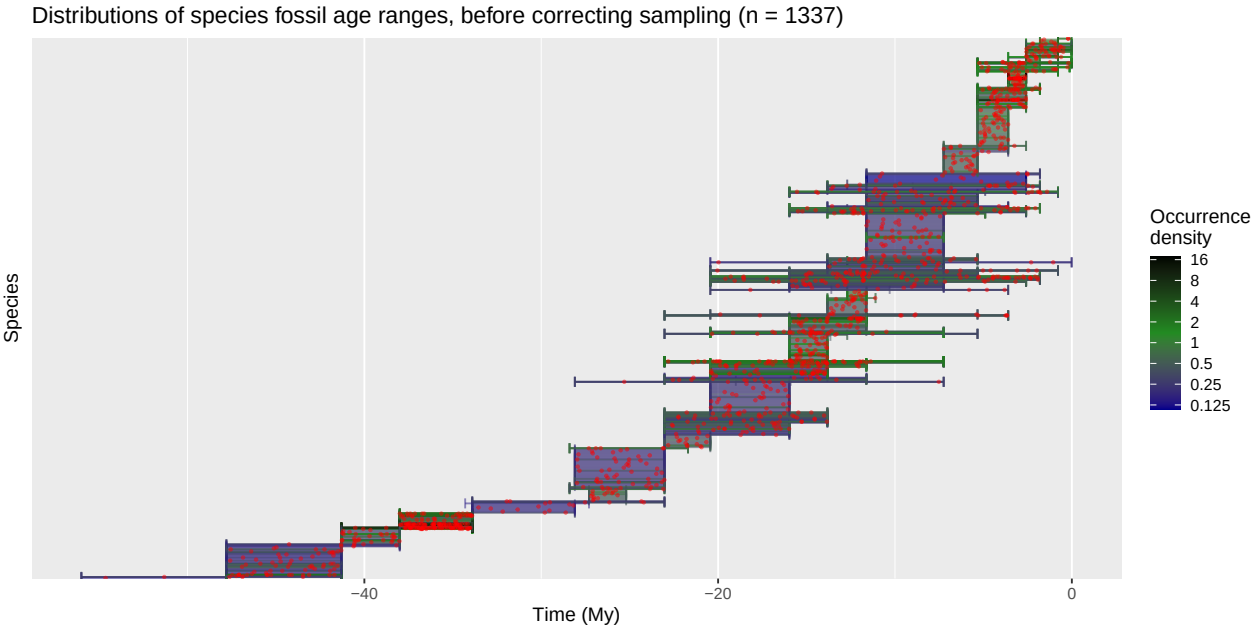
More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution :


Comparison between expected and observed distributions of number of occurrences
Number of corresponding accepted names (log−scale)
1e−04
1e−14
1e−24
1e−34
0
10
20
30
40
50
Number of occurrences by accepted name
species
Observed distribution
species

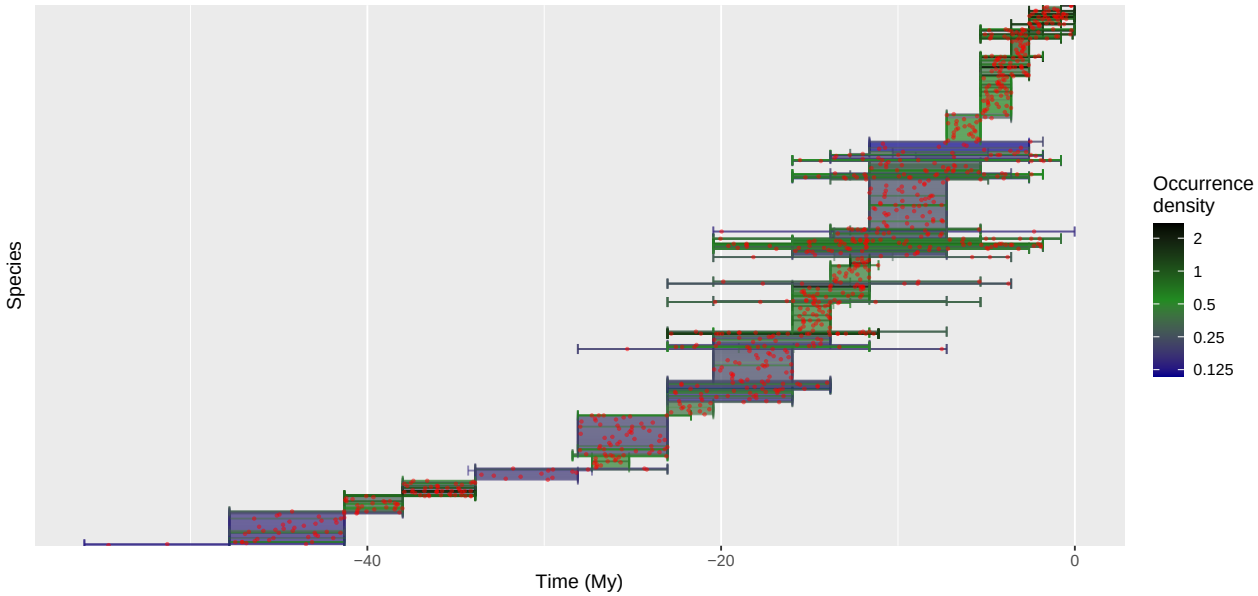



→ The correspondance is still good, except for two taxa :

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



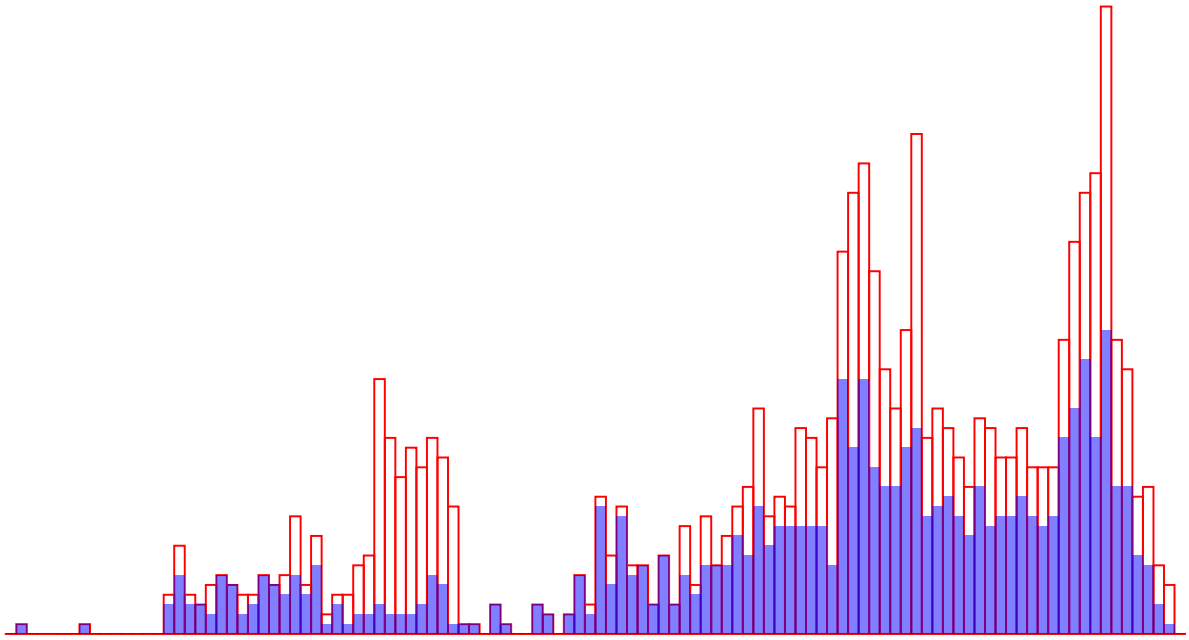
Distributions of species fossil age ranges, after correcting sampling (n = 800)




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

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

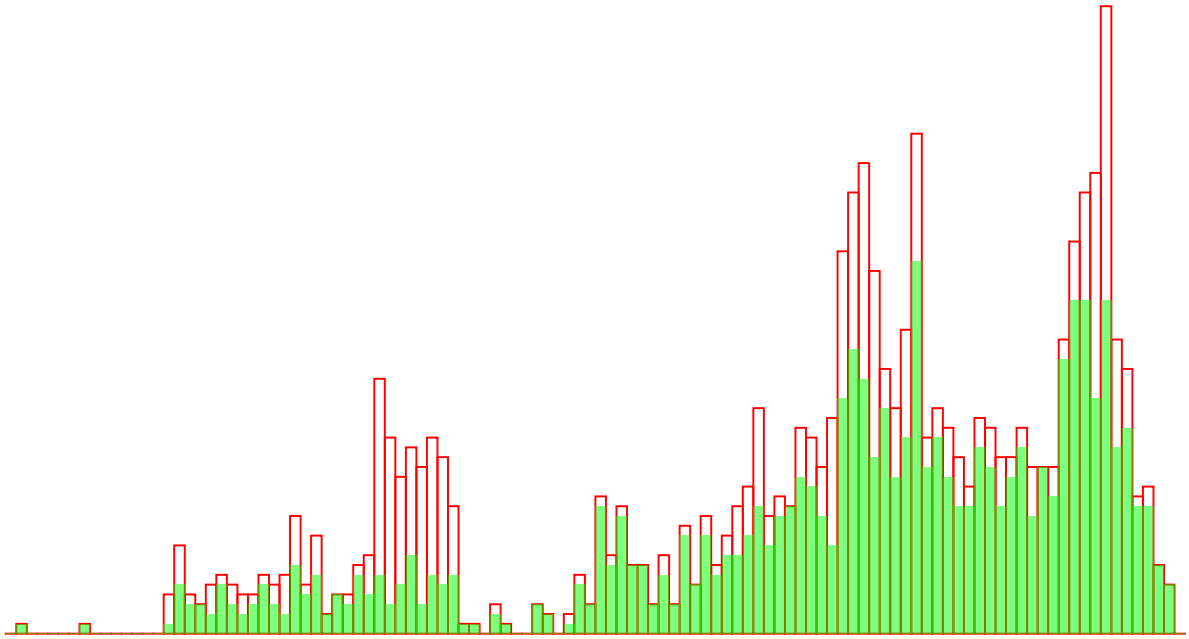
Initial occurrences distribution (red) and comparison after sub−sampling (blue)



```
## 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 aggregating in formations (green)

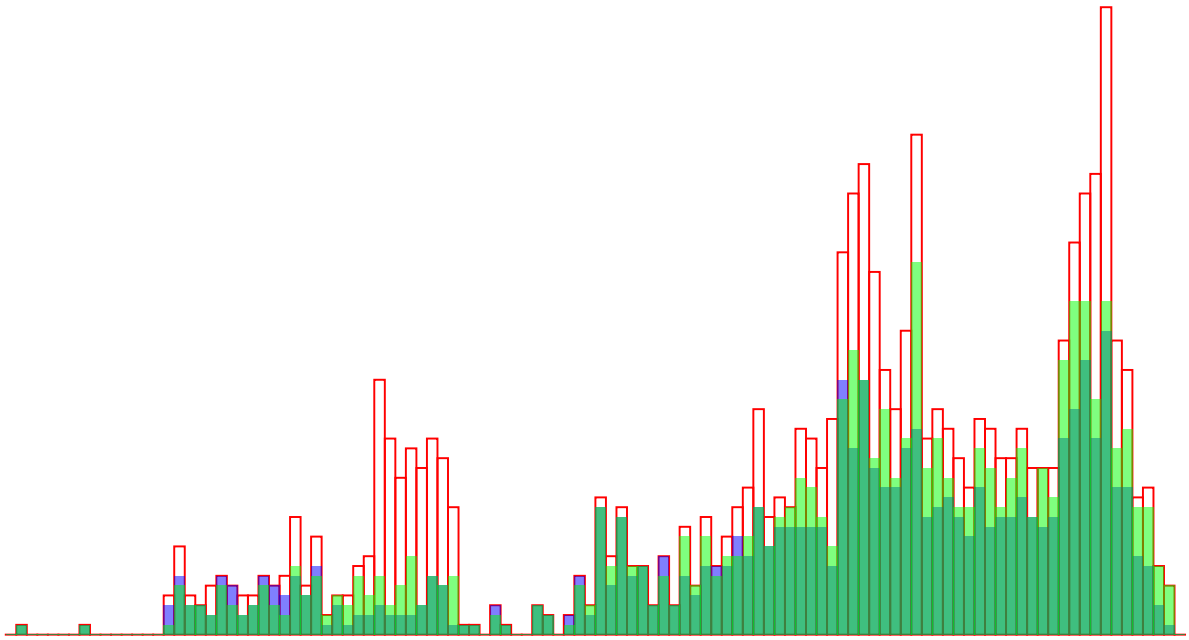


```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## 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)



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

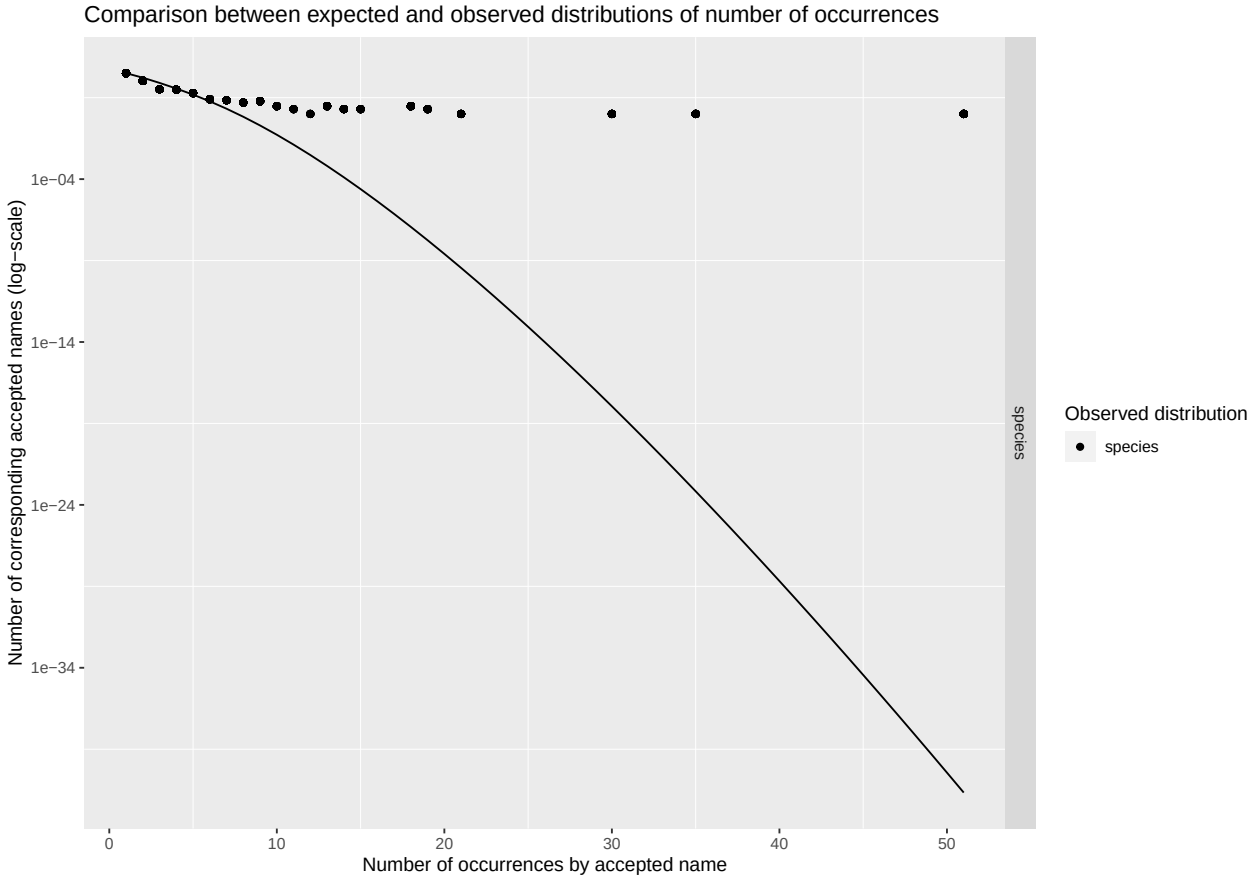
→ Comparing with the initial occurrences distribution and with the distribution after our first sub-sampling it appears that both methods lead to very similar distributions. This conforts us about the robustness of those approaches.

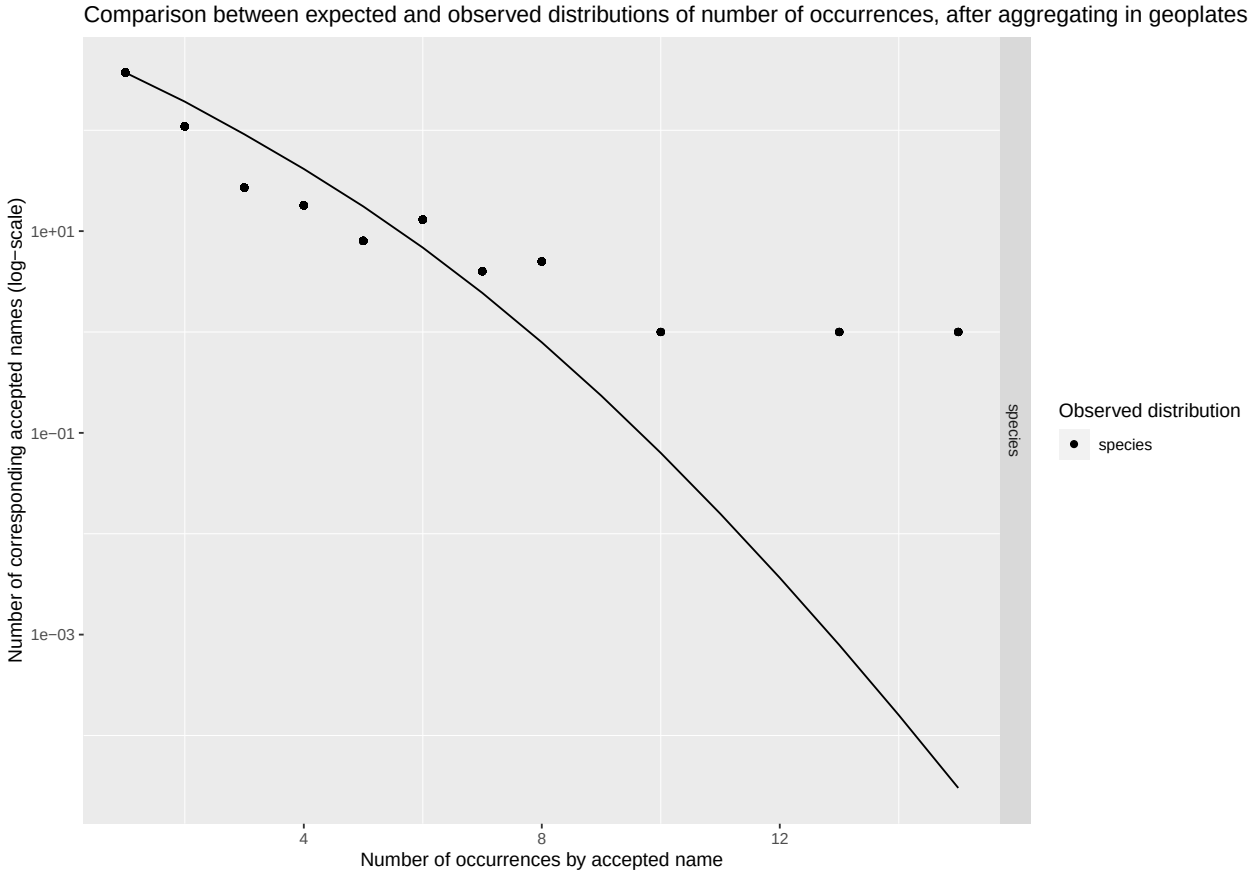
Replacing coutries by gelogical plates seems to make more sens from a palaeontological perspective, so let's try it.

## Aggregate occurrences without formation by geoplate + early interval

```
##                                    Cetacea_occ Cetacea_occ_aggreg removed
## Number of occurrences                     3804               2608    1196
## Number of occurrences (species only)      1437                968     469
```

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





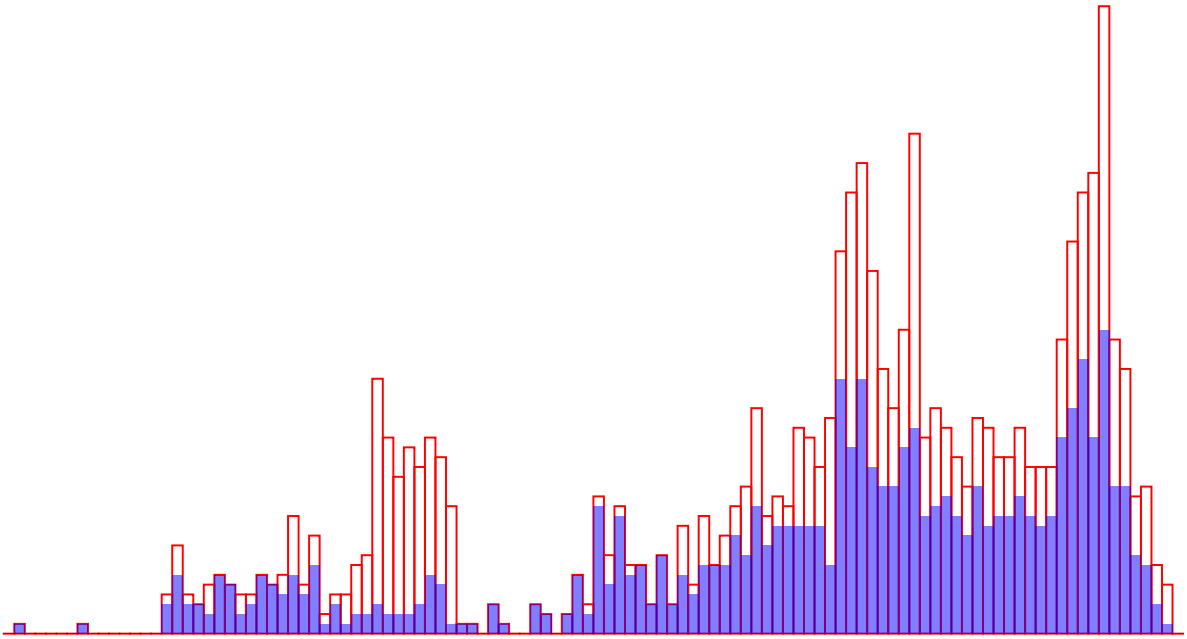

→ The correspondance is still good, except for two taxa :

```
##
##    Scaldicetus grandis Schizodelphis sulcatus
##                      13                     15
```

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

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

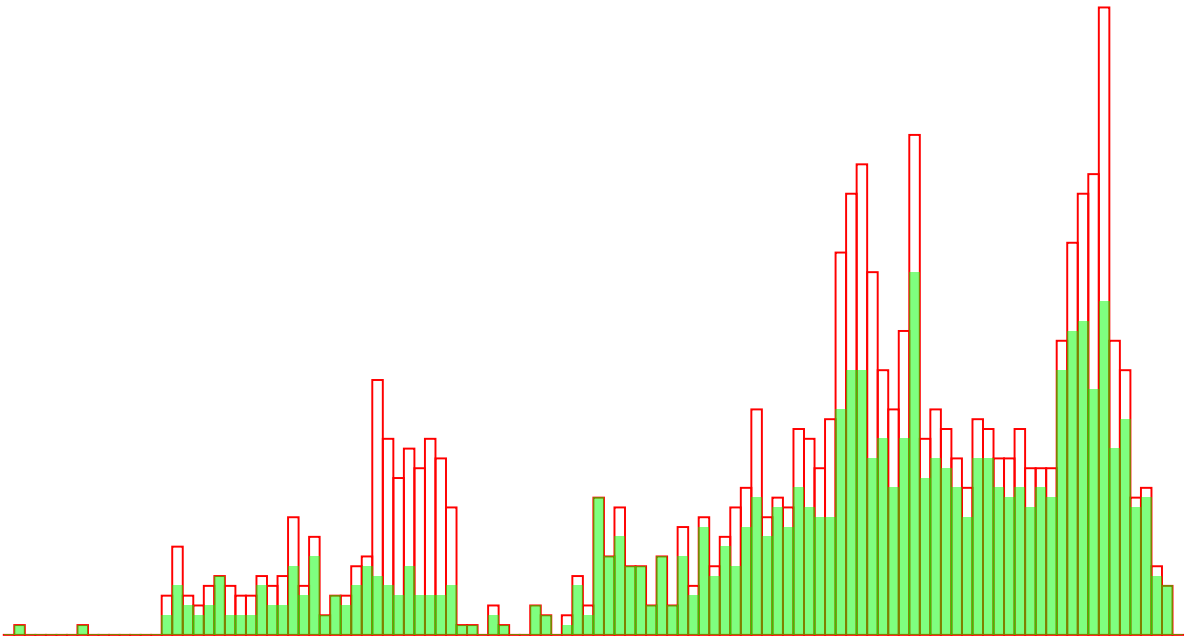
Initial occurrences distribution (red) and comparison after sub−sampling (blue)



```
## 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 aggregating in geoplates (green)

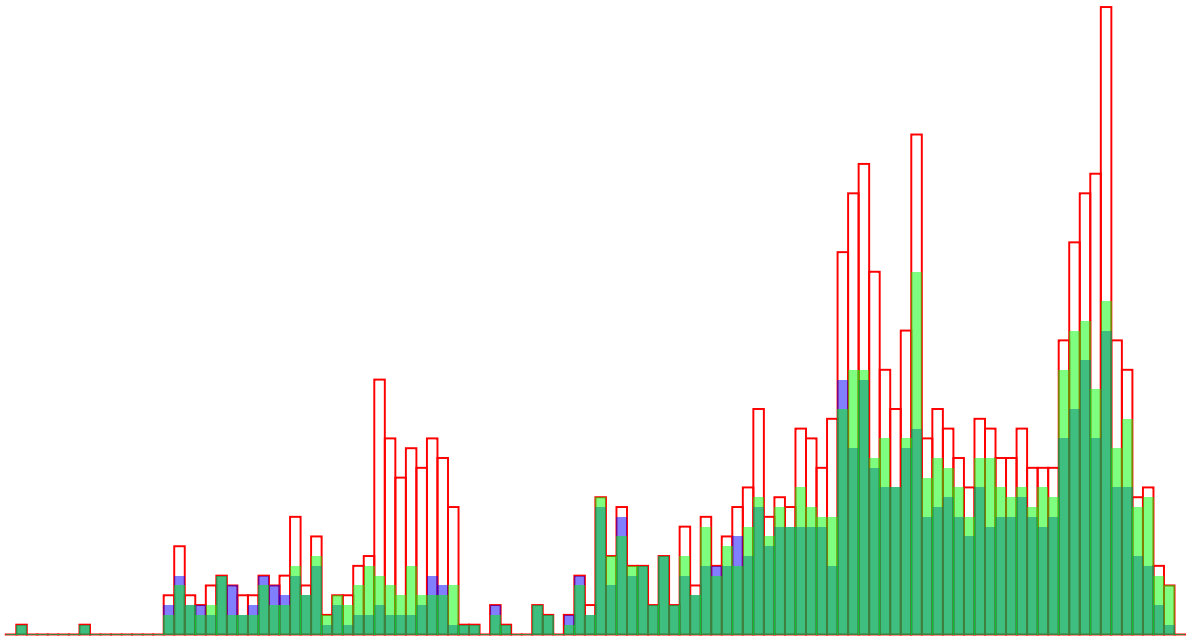


```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## 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)

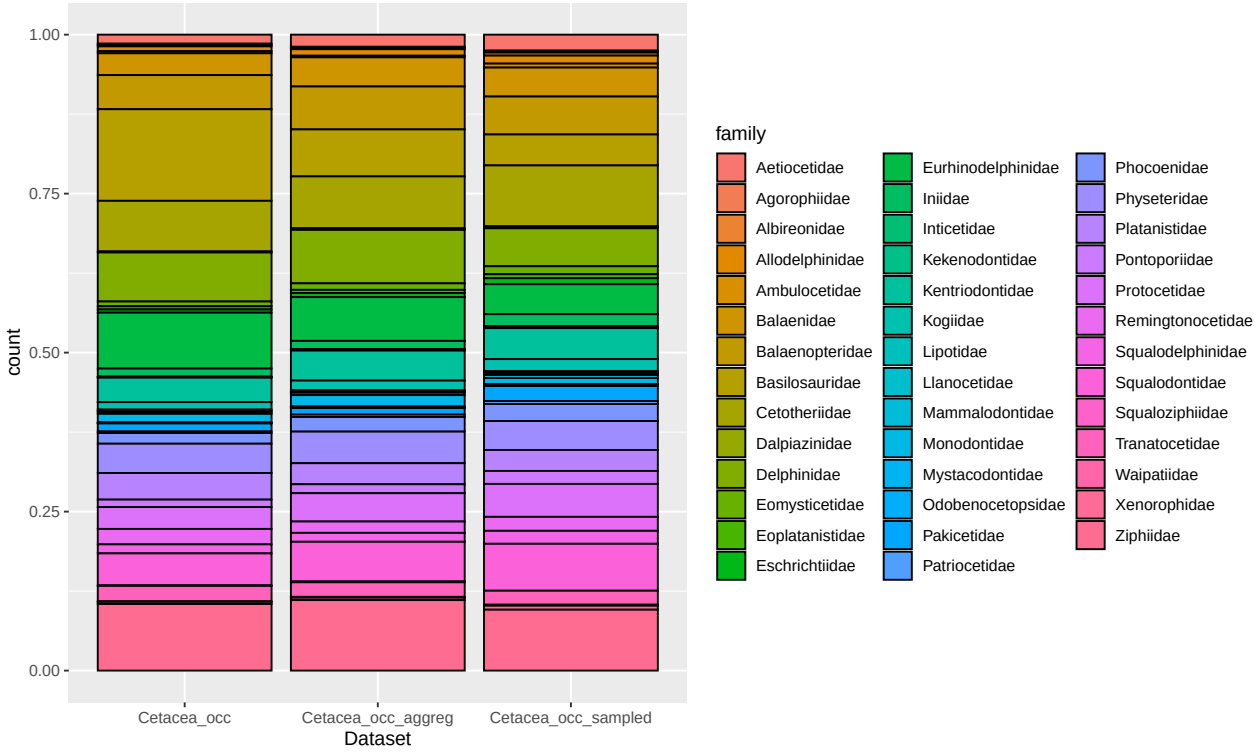


→ Delimiting by geological plates (+ age) instead of countries (+ age) leads to similar distributions, so we will keep it.

## 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 :

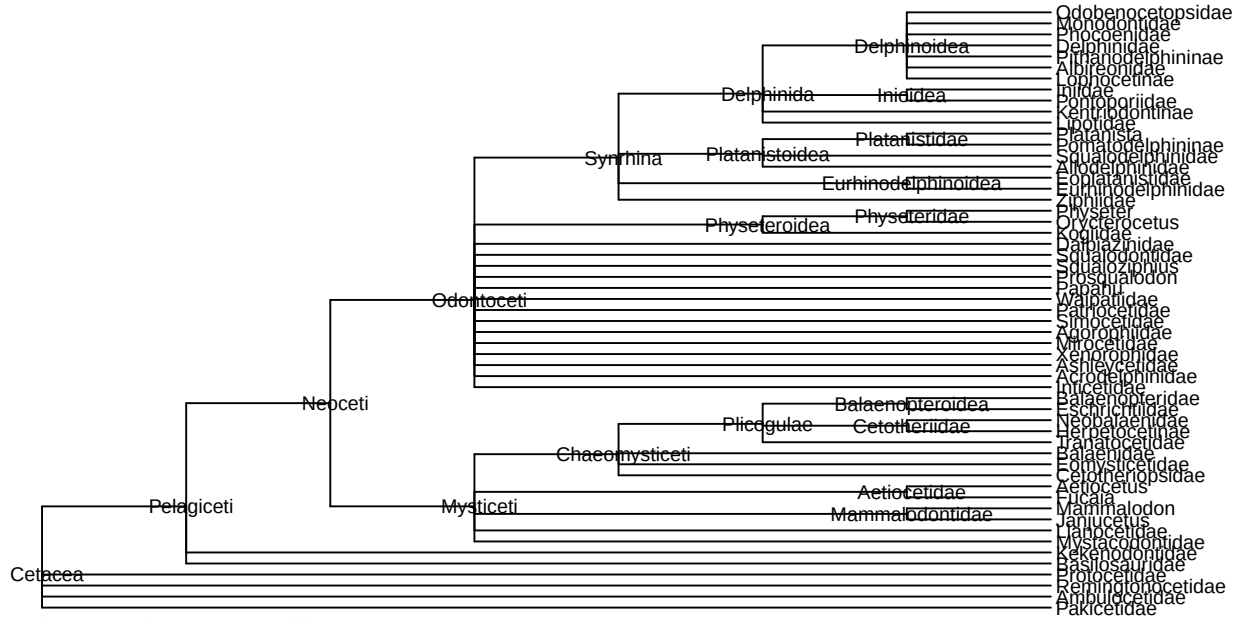



→ The family repartitions vary a bit after subsampling, but because of the limited number of species by family the fact that we corrected the oversampling of some species could have a disproportionate effect.

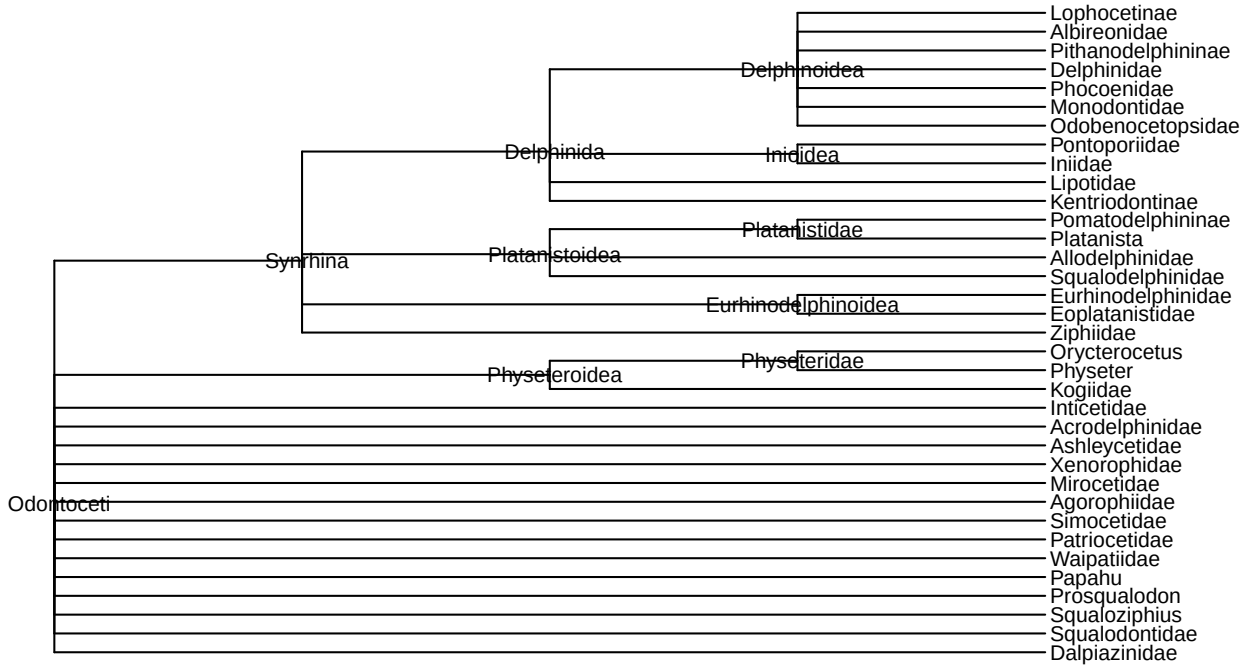
Let's look rather at a higher taxonomic rank, by importing the topology of cetacean families (from Marx et al. 2016) :

```
## 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.
```

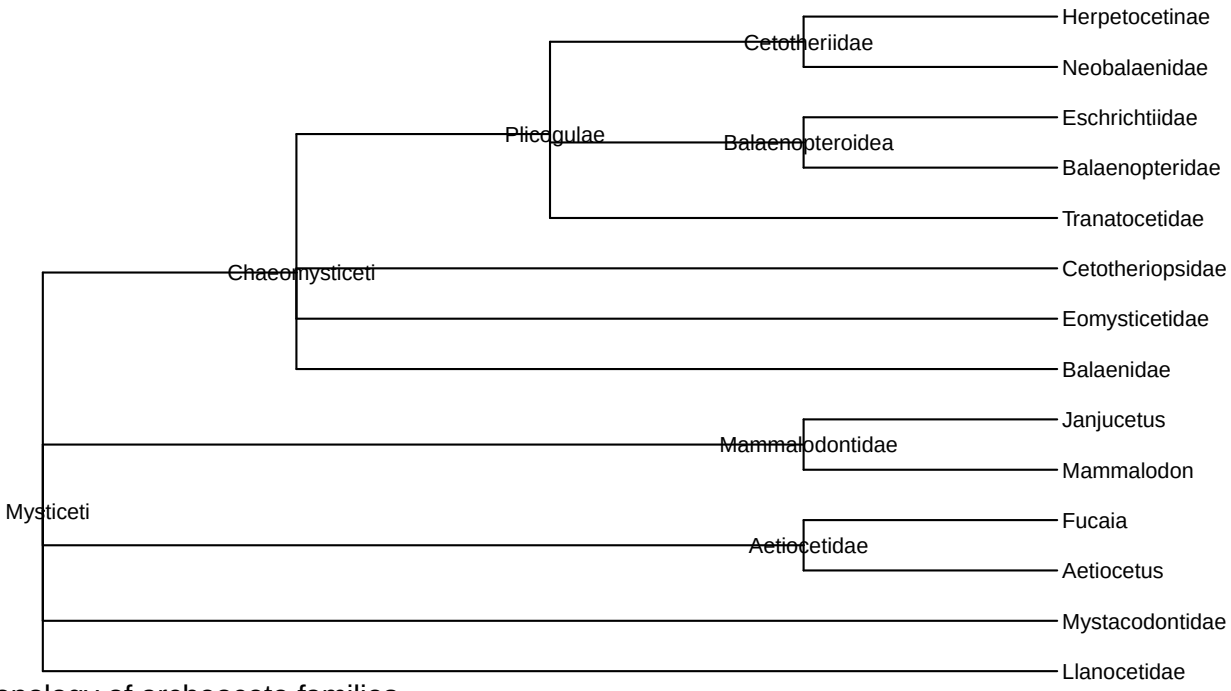
Topology of cetacean families



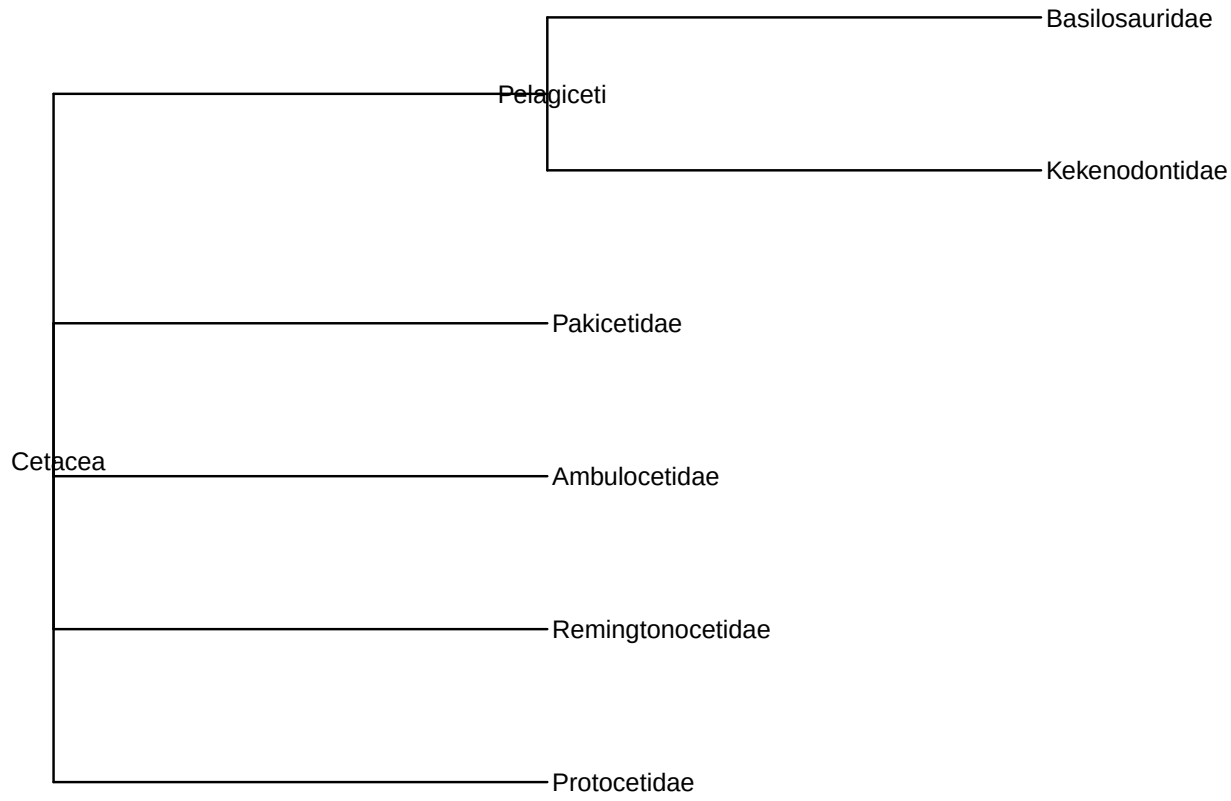

Topology of odontocete families

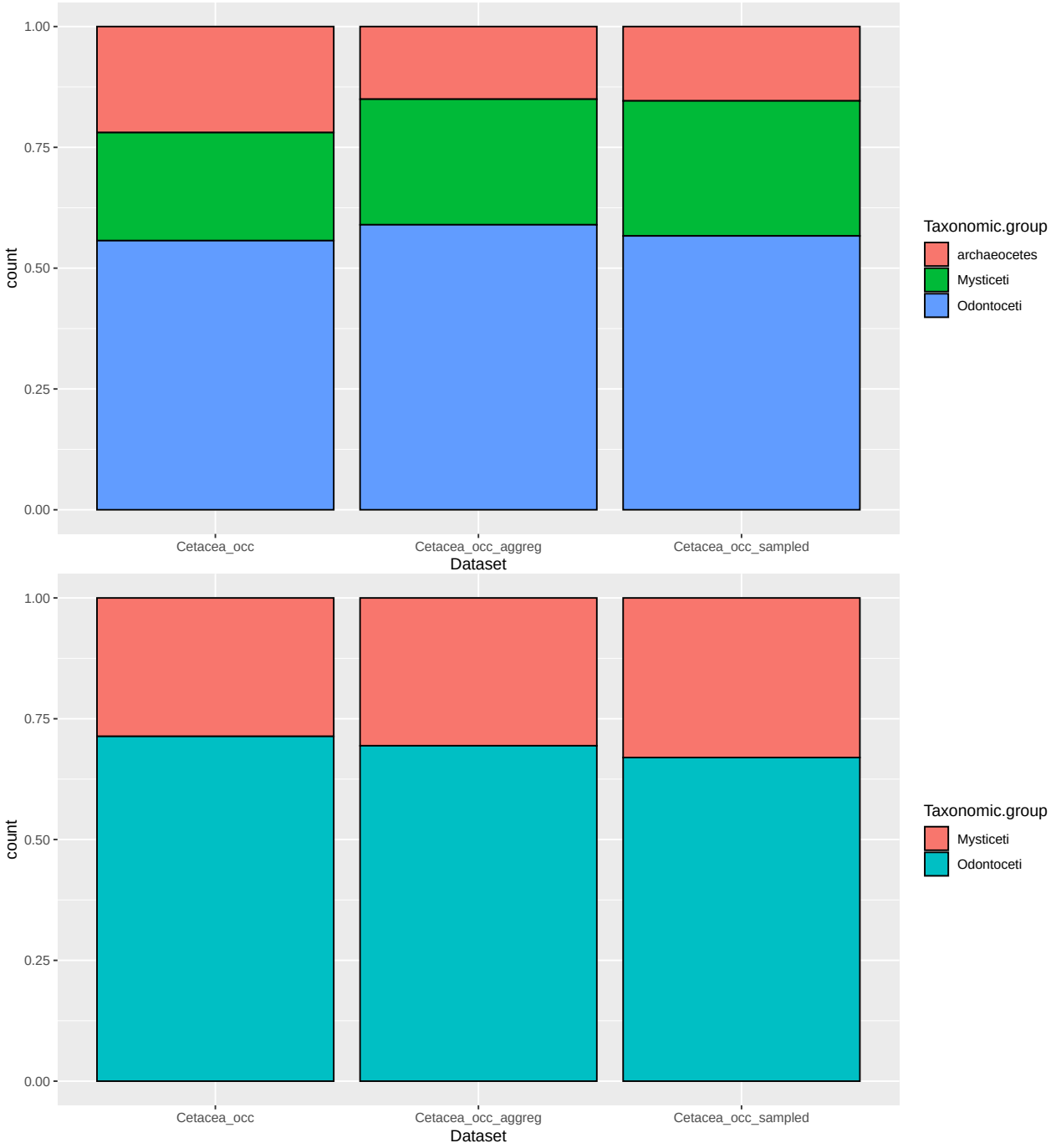


Topology of mysticete families

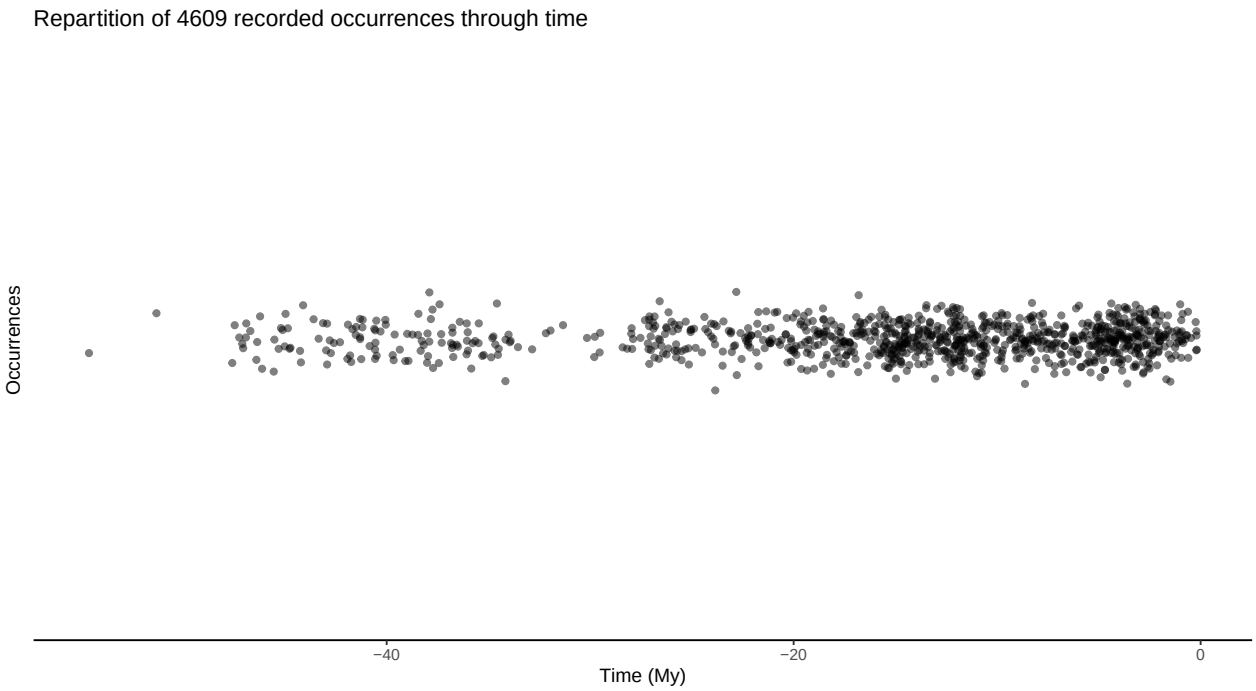



Topology of archeocete families





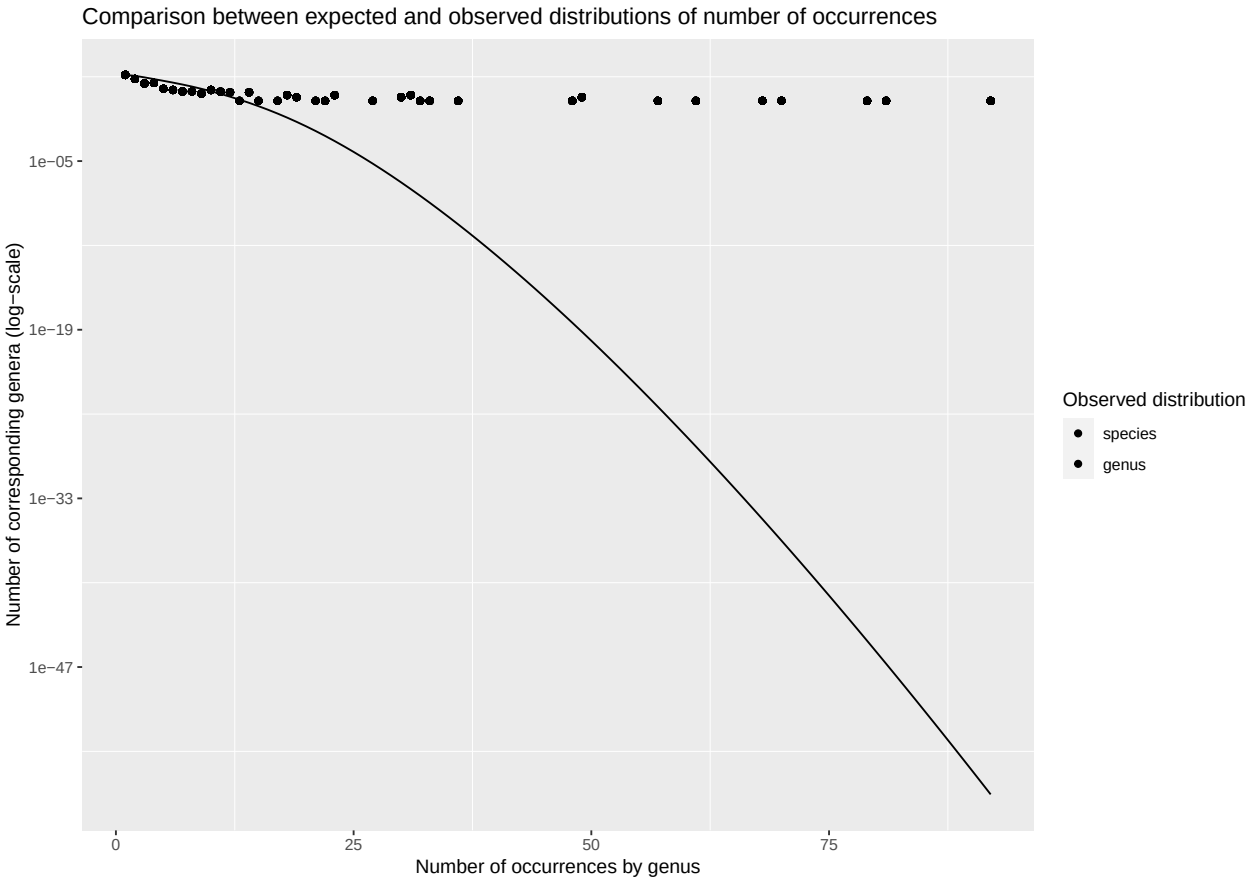

→ There is a smaller proportion of archeocete occurrences after either sampling, because a huge cluster is subsampled around 35 My ago, but this effect is expected. However, for the Mysticeti vs. Odotoceti there is no huge apparent bias, especially with the aggregating method.

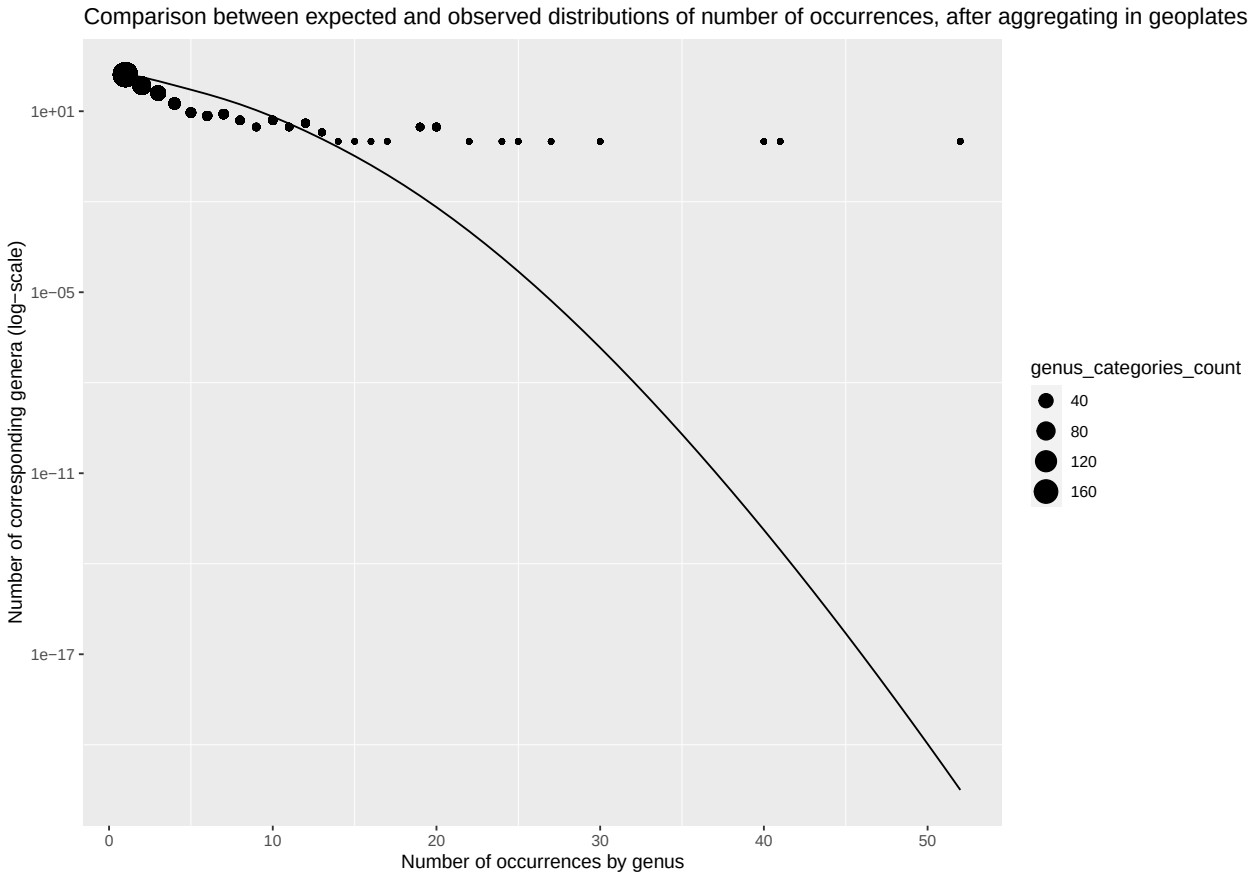



## Faster genus-level analysis

```
##                                                Cetacea_occ
## Number of occurrences                                 3804
## Number of occurrences (species and genera only)       2163
##                                                Cetacea_occ_aggreg_gen removed
## Number of occurrences                                            1840    1964
## Number of occurrences (species and genera only)                  1303     860
```

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


Comparison between expected and observed distributions of number of occurrences
Number of corresponding genera (log−scale)
1e−05
1e−19
1e−33
1e−47
0
25
50
75
Number of occurrences by genus
Observed distribution
species
genus



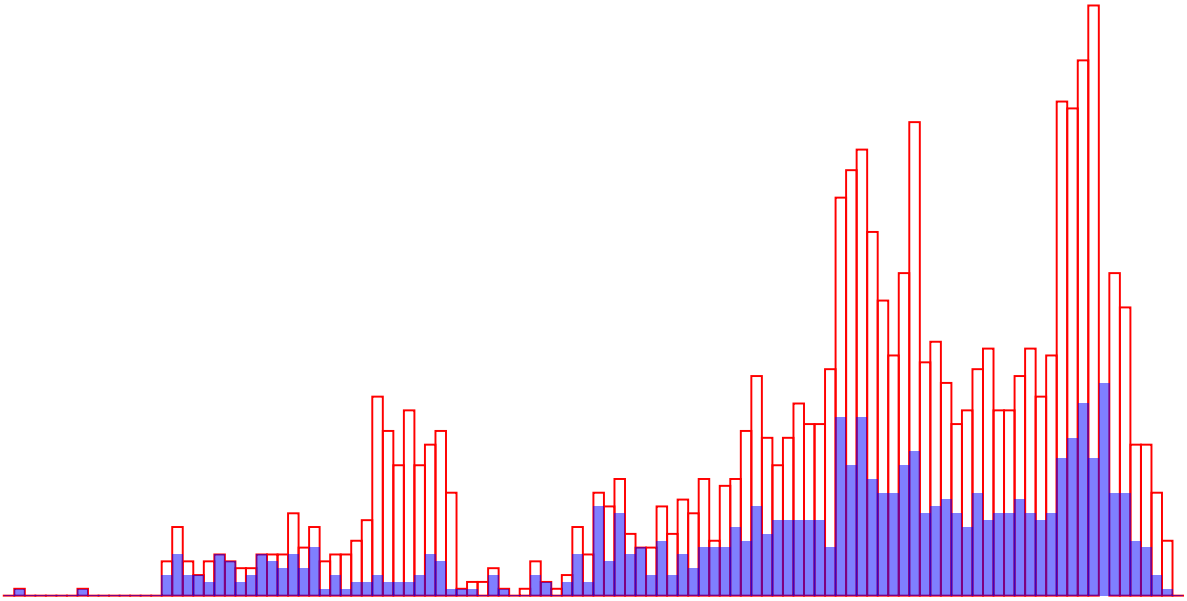

→ The correspondance is much less improved than with species aggregation because some genera have a lot of occurrences due to their high number of species species :

```
##
## Balaenoptera   Kentriodon   Mesoplodon  Scaldicetus    Squalodon
##           52           27           30           41           40
```

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

```
## Warning: Removed 3 rows containing missing values (geom_bar).
```

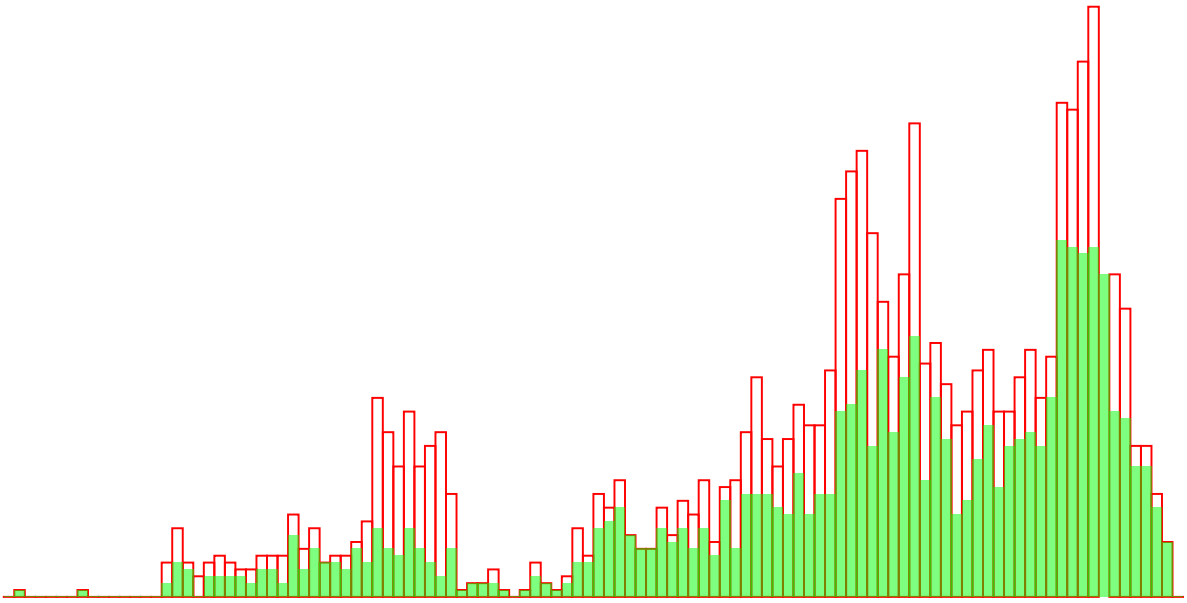
Initial occurrences distribution (red) and comparison after sub−sampling (blue)



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

```
## Warning: Removed 3 rows containing missing values (geom_bar).
```

Initial occurrences distribution (red) and comparison after aggregating in geoplates (green)

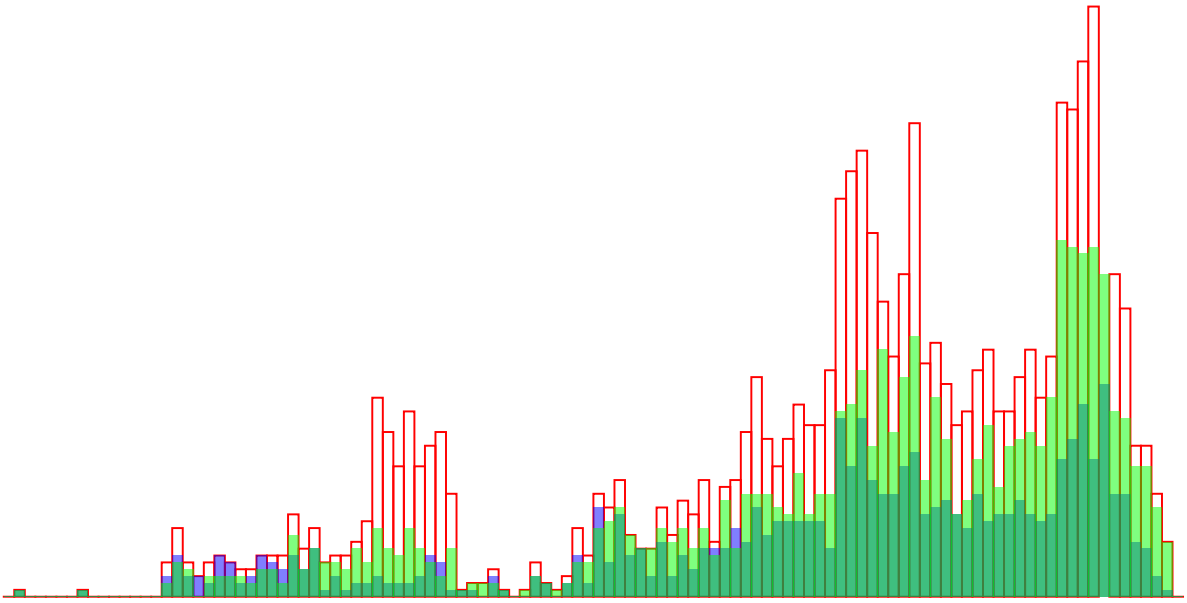


```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

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

```
## Warning: Removed 3 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, so we will keep it.

## 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) ?**