CCDB Tutorial

author: Rosana Zenil-Ferguson date: July 1, 2015

WARNING: To use this package you will have a dataset from CCDB and clean it before hand.

Download your dataset at http://ccdb.tau.ac.il/

You can use package chromer from Matt Pennell to download datasets from https://github.com/ropensci/chromer

How do I clean my dataset?

1. From the resolved_binomial column on CCDB download you need to create two extra columns one named Genus and another one named Species. It is easy to do in excel by selecting text-to-columns function. You must have those two columns to use the package otherwise it will return error (check the capitalization in the Genus and Species column name)
2. Try to erase strings like (i), (ii) and blank spaces in strings like +(space)0-1. This is easy to do in a text editor using replace function.

You have to do this because there are so many particular cases that are being scan in the internal functions of CCDB. I try to add all of the cases, and enumerate them below. However, you should expect some errors still in CCDBcurator v1.

Good news: If you need any angiosperm genus you don't need to do any of the steps above. I am providing an already revised dataset in this version.

Load the library of CCDB curator

```
library("CCDBcurator")
```

Access the angiosperm database

```
data(AngiospermsCCDB)
angiospermdata[1,]
```

```
##   resolved_binomial gametophytic sporophytic      resolved_name Examples
## 1   Acanthus_mollis        24.00        <NA> Acanthus mollis L.       NA
##      Genus Species Var Subespeciesname
## 1 Acanthus  mollis
```

```
#This is the first row of the angiosperm dataset notice it has the Genus and Species columns
```

You can notice that `angiospermdata[1,]` has the following columns 1. resolved_binomial= it is the species name. Comes from chromer download. 2. gametophytic= string representing chromosome counts in gametophytes. Usually complicated strings that need cleaning (check examples below). Download from chromer 3. sporophytic= string representing chromosome counts in sporophytes. Usually complicated strings that need cleaning (check examples below). Download from chromer. 4. resolved_name= Species name from chromer resolved. 1-4 is what you obtain using chromer download.

5. Examples= You won't need this column, it is only for the examples I am showing in here. noneed to add
6. Genus, Species, Var, Subespiciesname columns= **You create this ahead in excel by using the text to columns function from resolved_name.**

All angiosperm genera are included here, so no need of editing if you need them.

1. Cleaning a gametophytic chromosome number

You can clean a gametophytic entry from CCDB using the function `gametophytic.translator()`. This function will create a table with 5 columns 1.Genus= Genus name, 2.Species= Species name (format usable to match with phylogenies) 3.CountTranslation (numeric)= A numeric vector with the counts ready to use for analyses 4.Type= Gametophytic 5. CountOriginal= The original record in CCDB. This is done to control for errors.

Example 1

```
# This has two counts in gametophytic  column 11 and 22, but they are listed as characters
angiospermdata[56725,]
```

```
##           resolved_binomial gametophytic sporophytic
## 56725 Lobularia_canariensis       11(22)      22(44)
##                                             resolved_name Examples
## 56725 Lobularia canariensis subsp. palmensis (Webb) L. Borgen       NA
##           Genus     Species    Var Subespeciesname
## 56725 Lobularia canariensis subsp.       palmensis
```

```
# Corrected record
gametophytic.translator(angiospermdata[56725,])
```

```
##       Genus               Species CountTranslation         Type
## 1 Lobularia Lobularia_canariensis               11 gametophytic
## 2 Lobularia Lobularia_canariensis               22 gametophytic
##   CountOriginal
## 1        11(22)
## 2        11(22)
```

Example 2

```
# This has three counts in gametophytic  and possible B chromosomes, that are not be accounted for. The

angiospermdata[67997,]
```

```
##       resolved_binomial      gametophytic sporophytic
## 67997  Trema_orientalis "10, 11, 20+0-1B"        <NA>
##                       resolved_name Examples Genus    Species Var
## 67997 Trema orientalis (L.) Blume        1 Trema orientalis
##       Subespeciesname
## 67997
```

```
# Corrected record
gametophytic.translator(angiospermdata[67997,])
```

```
##   Genus          Species CountTranslation         Type     CountOriginal
## 1 Trema Trema_orientalis               10 gametophytic "10, 11, 20+0-1B"
## 2 Trema Trema_orientalis               11 gametophytic "10, 11, 20+0-1B"
## 3 Trema Trema_orientalis               20 gametophytic "10, 11, 20+0-1B"
```

Example 3

```
# This has 4 different counts with B chromosomes and fragments, that are not be accounted for.

angiospermdata[123476,]
```

```
##           resolved_binomial                gametophytic sporophytic
## 123476 Melampodium_cinereum "10, 10+2B, 20, 20+fragms."        <NA>
##                    resolved_name Examples       Genus  Species Var
## 123476 Melampodium cinereum DC.       NA Melampodium cinereum
##        Subespeciesname
## 123476
```

```
# Corrected record
gametophytic.translator(angiospermdata[123476,])
```

```
##         Genus              Species CountTranslation         Type
## 1 Melampodium Melampodium_cinereum               10 gametophytic
## 2 Melampodium Melampodium_cinereum               10 gametophytic
## 3 Melampodium Melampodium_cinereum               20 gametophytic
## 4 Melampodium Melampodium_cinereum               20 gametophytic
##                  CountOriginal
## 1 "10, 10+2B, 20, 20+fragms."
## 2 "10, 10+2B, 20, 20+fragms."
## 3 "10, 10+2B, 20, 20+fragms."
## 4 "10, 10+2B, 20, 20+fragms."
```

Example 4

```
# This has four records, interval-like  with B chromosomes. Intervals are a sequence so 10-13 are consi

angiospermdata[166557,]
```

```
##           resolved_binomial                 gametophytic sporophytic
## 166557 Pelargonium_pinnatum "10-13II+0-2B, 20-23II+0-2B"        c.22
##                             resolved_name Examples       Genus  Species
## 166557 Pelargonium pinnatum (L.) L'H\xe9r.        1 Pelargonium pinnatum
##        Var Subespeciesname
## 166557
```

```
# Corrected record
gametophytic.translator(angiospermdata[166557,])
```

```
##         Genus              Species CountTranslation         Type
## 1 Pelargonium Pelargonium_pinnatum               10 gametophytic
## 2 Pelargonium Pelargonium_pinnatum               11 gametophytic
## 3 Pelargonium Pelargonium_pinnatum               12 gametophytic
## 4 Pelargonium Pelargonium_pinnatum               13 gametophytic
## 5 Pelargonium Pelargonium_pinnatum               20 gametophytic
## 6 Pelargonium Pelargonium_pinnatum               21 gametophytic
## 7 Pelargonium Pelargonium_pinnatum               22 gametophytic
## 8 Pelargonium Pelargonium_pinnatum               23 gametophytic
##                   CountOriginal
## 1 "10-13II+0-2B, 20-23II+0-2B"
## 2 "10-13II+0-2B, 20-23II+0-2B"
## 3 "10-13II+0-2B, 20-23II+0-2B"
## 4 "10-13II+0-2B, 20-23II+0-2B"
## 5 "10-13II+0-2B, 20-23II+0-2B"
## 6 "10-13II+0-2B, 20-23II+0-2B"
## 7 "10-13II+0-2B, 20-23II+0-2B"
## 8 "10-13II+0-2B, 20-23II+0-2B"
```

Example 5

```
# This has three records and a reference of the type (1,1,1) that needs to be removed
angiospermdata[90026,]
```

```
##         resolved_binomial                gametophytic sporophytic
## 90026 Bahianthus_viscidus "10, ca.10, ca.12(1, 1, 1)"        <NA>
##                                       resolved_name Examples      Genus
## 90026 Bahianthus viscidus (Baker) R. M. King & H. Rob.        1 Bahianthus
##        Species Var Subespeciesname
## 90026 viscidus
```

```
# Corrected record
gametophytic.translator(angiospermdata[90026,])
```

```
##        Genus             Species CountTranslation         Type
## 1 Bahianthus Bahianthus_viscidus               10 gametophytic
## 2 Bahianthus Bahianthus_viscidus               10 gametophytic
## 3 Bahianthus Bahianthus_viscidus               12 gametophytic
##                CountOriginal
## 1 "10, ca.10, ca.12(1, 1, 1)"
## 2 "10, ca.10, ca.12(1, 1, 1)"
## 3 "10, ca.10, ca.12(1, 1, 1)"
```

Example 6

```
# This has two records and a pattern (2-8 f that needs to be removed
angiospermdata[237124,]
```

```
##           resolved_binomial    gametophytic sporophytic
## 237124 Plantago_crassifolia "10, 10+(2-8 f"        <NA>
##                       resolved_name Examples    Genus     Species Var
## 237124 Plantago crassifolia Forssk.        1 Plantago crassifolia
##        Subespeciesname
## 237124
```

```
# Corrected record
gametophytic.translator(angiospermdata[237124,])
```

```
##      Genus              Species CountTranslation         Type
## 1 Plantago Plantago_crassifolia               10 gametophytic
## 2 Plantago Plantago_crassifolia               10 gametophytic
##     CountOriginal
## 1 "10, 10+(2-8 f"
## 2 "10, 10+(2-8 f"
```

As you can see I have done my best to remove most of the patterns that should not be counts. However in a dataset this big many things could go wrong. To detect any mistakes I have added the column CountOriginal so you can compare the original record with the translation. Please let me know if you find other patterns undetected at rzenil@ufl.edu

2.Cleaning a sporophytic chromosome number

Simmilarly to the gametophytic clean, you can do sporophytic cleanings through the function `sporophytic.translator()`. This function will create a table with 5 columns 1.Genus= Genus name, 2.Species= Species name (format usable to match with phylogenies) 3.CountTranslation (numeric)= A numeric vector with the counts ready to use for analyses 4.Type= sporophytic 5. CountOriginal= The original record in CCDB. This is done to control for errors.

Example 1

```
# This has one record and B chromosomes
angiospermdata[1831,]
```

```
##        resolved_binomial gametophytic sporophytic          resolved_name
## 1831 Sambucus_canadensis         <NA>   38+(0-2B) Sambucus canadensis L.
##      Examples    Genus    Species Var Subespeciesname
## 1831       NA Sambucus canadensis
```

```
# Corrected record
sporophytic.translator(angiospermdata[1831,])
```

```
##      Genus             Species CountTranslation        Type CountOriginal
## 1 Sambucus Sambucus_canadensis               38 sporophytic     38+(0-2B)
```

3. Creating clean records for a genus or all angiosperms.

If you want to clean a whole genus or subset you want to use the function `CCDBcurator()` . The input for this function needs either one row or multiple rows from your dataset. **Example1**

```
# Clean all the chromosome counts for genus Acanthus
index<-which(angiospermdata$Genus=="Acanthus") #which data in angiospermdata are Acanthus?

acanthus.sample<-angiospermdata[index,] # this are all the Acanthus in angiospermdata

Acanthus.clean<-CCDBcurator(acanthus.sample) # Prints a warning when there are no sporophytic or gameto
```

```
## [1] "No Sporophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Sporophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Sporophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Sporophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
## [1] "No Gametophytic Count"
```

So the clean records and Haploid number of chromosomes and the possibility of aneuploidy (sporophyte record non-divisible by 2) are stored in the object

```
Acanthus.clean
```

```
##       Genus              Species CountTranslation         Type
## 1  Acanthus      Acanthus_mollis               24 gametophytic
## 2  Acanthus      Acanthus_mollis               56  sporophytic
## 3  Acanthus      Acanthus_mollis               80  sporophytic
## 4  Acanthus    Acanthus_spinosus              112  sporophytic
## 5  Acanthus      Acanthus_mollis               56  sporophytic
## 6  Acanthus Acanthus_ilicifolius               44  sporophytic
## 7  Acanthus Acanthus_ilicifolius               44  sporophytic
## 8  Acanthus Acanthus_ilicifolius               44  sporophytic
## 9  Acanthus   Acanthus_pubescens               56  sporophytic
## 10 Acanthus      Acanthus_mollis               80  sporophytic
## 11 Acanthus      Acanthus_mollis               28 gametophytic
## 12 Acanthus    Acanthus_spinosus               56  sporophytic
## 13 Acanthus    Acanthus_spinosus               56 gametophytic
## 14 Acanthus Acanthus_ebracteatus               44  sporophytic
## 15 Acanthus   Acanthus_volubilis               44  sporophytic
## 16 Acanthus Acanthus_ilicifolius               24 gametophytic
## 17 Acanthus Acanthus_ilicifolius               44  sporophytic
## 18 Acanthus Acanthus_ilicifolius               48  sporophytic
## 19 Acanthus Acanthus_ilicifolius               48  sporophytic
```

```
## 20 Acanthus Acanthus_ebracteatus              44  sporophytic
## 21 Acanthus Acanthus_ilicifolius              44  sporophytic
## 22 Acanthus Acanthus_ilicifolius              48  sporophytic
## 23 Acanthus Acanthus_ilicifolius              44  sporophytic
## 24 Acanthus Acanthus_ilicifolius              48  sporophytic
## 25 Acanthus      Acanthus_mollis              56  sporophytic
## 26 Acanthus      Acanthus_mollis              56  sporophytic
## 27 Acanthus    Acanthus_spinosus              80  sporophytic
## 28 Acanthus    Acanthus_spinosus             112  sporophytic
## 29 Acanthus      Acanthus_mollis              56  sporophytic
## 30 Acanthus    Acanthus_spinosus              56  sporophytic
## 31 Acanthus    Acanthus_spinosus              80  sporophytic
## 32 Acanthus    Acanthus_spinosus             112  sporophytic
##    CountOriginal HaploidNumber Aneuplody
## 1          24.00            24         0
## 2       "56, 80"            28         0
## 3       "56, 80"            40         0
## 4            112            56         0
## 5             56            28         0
## 6             44            22         0
## 7             44            22         0
## 8             44            22         0
## 9             56            28         0
## 10            80            40         0
## 11         28.00            28         0
## 12            56            28         0
## 13         56.00            56         0
## 14            44            22         0
## 15            44            22         0
## 16         24.00            24         0
## 17      "44, 48"            22         0
## 18      "44, 48"            24         0
## 19            48            24         0
## 20            44            22         0
## 21       "44,48"            22         0
## 22       "44,48"            24         0
## 23      "44, 48"            22         0
## 24      "44, 48"            24         0
## 25           ~56            28         0
## 26            56            28         0
## 27            80            40         0
## 28           112            56         0
## 29            56            28         0
## 30            56            28         0
## 31            80            40         0
## 32           112            56         0
```

With this output you can do multiple things for example check by genus how many haploid numbers you have and how many times they appear

```
genus.table<- table(Acanthus.clean$Genus,Acanthus.clean$HaploidNumber)
genus.table
```

```
##
```

```
##            22 24 28 40 56
##   Acanthus  9  6  9  4  4
```

Or check how many records by species

```
species.table<-table(Acanthus.clean$Species, Acanthus.clean$HaploidNumber)
species.table
```

```
##
##                        22 24 28 40 56
##   Acanthus_ebracteatus  2  0  0  0  0
##   Acanthus_ilicifolius  6  5  0  0  0
##   Acanthus_mollis       0  1  6  2  0
##   Acanthus_pubescens    0  0  1  0  0
##   Acanthus_spinosus     0  0  2  2  4
##   Acanthus_volubilis    1  0  0  0  0
```

You can plot to detect the most likely x-number and if the gametophytic and sporophytic counts match

```
small<-as.data.frame(table(Acanthus.clean$HaploidNumber,Acanthus.clean$Type)) # Acanthus.clean is used
names(small)<-c("HaploidNumber","Type","Freq") # The rest remains the same for any other example

small_ordered = small[with(small, order(HaploidNumber,Type)),]
data_ordered=matrix(small_ordered$Freq, ncol=2,byrow=TRUE)
colnames(data_ordered)=levels(small_ordered$Type)
rownames(data_ordered)=levels(small_ordered$HaploidNumber)
colorsplot<-c("lightblue")
barplot(t(data_ordered), col=c("blue","gray"), main="Acanthus",xlab="Haploid Number",ylab="Frequency")
legend("topright",fill=c("blue","gray"),legend=colnames(data_ordered))
```