Summary of Notes:

1 – 3 : main databases (wrangling)
4 –5 : merging and analyses
6 : graph
7: stats for graphs

Summary for python code:

1 : Kew_web_scrape_chromosome_size.py
2 : ccdb_edit.html
3 – 7: ccdb_kew_pw_analyses.html

1.) KEW:

Webscrape to get Kew chromosome size:
Data taken from: https://cvalues.science.kew.org/search
Script: Kew_web_scrape_chromosome_size.py
File: **Kew_1C_pg.csv**

2.) CCDB: Chromosome Count DataBase

Originally taken from 3 files:

Angiosperms_Jan2021.csv
Gymnosperms_Jan2021.csv
Pterids_Jan2021.csv

Cleaned and merge files:

I. made sure to remove any unwanted whitespace in names
II. created 3 columns:
    A. major_group: Angiosperms, Gymnosperms, Horsetails, Lycophytes, or Ferns
    B. spory: heterosporous or homosporous
    C. category: combine major_group and spory group names for when we make graphs
III. create a species column using the “resolved_name” column
    A. removed the 5 angiosperms with no resolved_name from the database
    B. nine where the species is a cross; I made the species nan
        1. (where ‘genus’ column startswith('×'))
    C. 124 where ‘Ã' is in the species name: change those to correct names
    D. finally, there are still 19 Solanum genera with no species name (these happen to also have contained the ‘Ã’ within the resolved name – followed by author names. NO species)
    E. Total of 28 rows of data where species is nan (9 cross and the 18 Solanum) in species column

Save this dataframe as:
ccdb_clean.csv
377,563 rows x 20 columns
see: ccdb_edit.html

Made a mini version of this with 10 columns:
ccdb_clean_mini.csv
10 cols are: [‘family’, ‘genus', ‘species', ‘resolved_name', ‘major_group', ‘spory', ‘category', ‘parsed_n', ‘list_parsed_n', ‘min_parsed_n’]

(NOTE: parsed_n, and subsequently the other n columns, have 10,035 rows of missing data)

3.) PW’s 10 sample database

Data from this table will be added to kew where there’s info in the 1C column, and to ccdb where there’s info in the n column.
“**Paul_chromosome_1C_additions.csv**”
(Original file that I didn’t use is: **Paul_orig_chromosome_1C_additions.csv**)

Citations:

Heterosporous ferns:
Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. Nat Plants 4, 460–472.

Sellaginella:
Little DP, Moran RC, Brenner ED, Stevenson DW. Nuclear genome size in Selaginella. Genome. 2007 Apr;50(4):351–6. doi: 10.1139/g06–138. PMID: 17546093.

genus	sp	category	n	DNA Amount1C (pg)
Azolla	filicoides	Ferns\nHeterosporous		0.76
Salvinia	cucullata	Ferns\nHeterosporous	9	0.26
Pilularia	americana	Ferns\nHeterosporous		0.84
Regnellidium	diphyllum	Ferns\nHeterosporous	19	1.8
Marsilea	spp	Ferns\nHeterosporous	20	1.37
Selaginella	apoda	Lycophytes\nHeterosporous	9	
Selaginella	involvens	Lycophytes\nHeterosporous	9	
Selaginella	moellendorffii	Lycophytes\nHeterosporous	9	
Selaginella	vogelii	Lycophytes\nHeterosporous	9	
Selaginella	willdenowii	Lycophytes\nHeterosporous	9	

4.) Add pw data to ccdb and kew databases

merge kew + pw = **Kew_pw_1C_pg_mini.csv**
- columns: ['genus', 'species', 'DNA Amount1C (pg)']
- size: (12278, 3)

merge ccdb (no horsetails, no missing n info (min_parse_n is nan)) + pw = **ccdb_pw_noHorsetails_noMissingN.csv**
- columns: ['family', 'genus', 'species',m'resolved_name', 'major_group', 'spory', 'category', 'parsed_n', 'list_parsed_n', 'min_parsed_n']
- size: (367405, 10)

5.) Databases for the Graphs

NOTE: This gets used to create the upper two graphs
Using the ccdb_pw_noHorsetails_noMissingN.csv, get only ONE minimum per genus:
(final_ccdb_mins in html file)
ccdb_noHorsetails_1min_per_genus.csv
size: (8338, 10)
columns: ['genus', 'min_parsed_n', 'family', 'species', 'resolved_name', 'major_group', 'spory', 'category', 'parsed_n', 'list_parsed_n']

NOTE: Did NOT end up using this one for making the graph.
Using the ccdb_pw_noHorsetails_noMissingN.csv, get ALL possible minimums per genus:
ccdb_noHorsetails_ALLmin_per_genus.csvsize:
size: (69098, 4)
columns: ['genus', 'species', 'category', 'min_parsed_n']

NOTE: This does NOT used for making the lower 2 graphs
Want all diploids, even if aneuploid….up to a certain point:
Want to groupby each genus. Within each genus, make a list of all chromosome counts. Within the list of counts, divide each number in list by the lowest value in the list. If that divided value is <= 1.2, then we will keep those individuals within that given genus.
This contains all unique species for each genus!
ccdb_cleaned_EVERY_min_1pt2_ratio.csv
shape: (98373, 11)

columns: ['family', 'genus', 'species', 'resolved_name', 'major_group', ‘spory', ‘category', 'parsed_n', 'list_parsed_n', ‘min_parsed_n', 'diploid_mins']

NOTE: This gets used for the lower graphs
Use ccdb_cleaned_EVERY_min_1pt2_ratio.csv.
Merge with ccdb_pw_noHorsetails_noMissingN.csv (kew_pw2 from html).
Select just needed columns: ['genus', 'species', 'category', 'min_parsed_n','DNA Amount1C (pg)']
Drop the duplicate lines.
Result:
lower_graphs_db.csv
(3656, 5)

6.) Graphs:

upper graphs use: **ccdb_noHorsetails_1min_per_genus.csv**
lower graphs use: **lower_graphs_db.csv**
output graph:
hetero_chromo_num_paper_withPWdata_ALLmins_ratio_1pt2.png

7.) Stats:

UPPER GRAPHS: chromosome counts

CATEGORY
Angiosperms\nHeterosporous 7900
Ferns\nHomosporous 343
Gymnosperms\nHeterosporous 76
Lycophytes\nHomosporous 7
Lycophytes\nHeterosporous 6
Ferns\nHeterosporous 6
Name: category, dtype: int64

HOMOSPOROUS ONLY
count 350.000000
mean 40.471429
std 23.980365
min 1.000000
25% 30.000000

```
50%       36.000000
75%       41.000000
max      164.000000
Name: min_parsed_n, dtype: float64

HETEROSPOROUS ONLY
count    7988.000000
mean       12.878443
std         8.842195
min         1.000000
25%         9.000000
50%        11.000000
75%        15.000000
max       275.000000
Name: min_parsed_n, dtype: float64

BOTH (HETEROSPOROUS & HOMOSPOROUS)
count    8338.000000
mean       14.036699
std        11.384060
min         1.000000
25%         9.000000
50%        11.000000
75%        16.000000
max       275.000000
Name: min_parsed_n, dtype: float64
```

LOWER GRAPHS: genome size

```
CATEGORY
Angiosperms\nHeterosporous    3387
Gymnosperms\nHeterosporous     178
Ferns\nHomosporous              72
Lycophytes\nHeterosporous       14
Ferns\nHeterosporous             4
Lycophytes\nHomosporous          1
Name: category, dtype: int64

HOMOSPOROUS ONLY
count    73.000000
mean     18.356027
```

std 21.938207
min 2.430000
25% 6.970000
50% 10.460000
75% 14.890000
max 74.840000
Name: DNA Amount1C (pg), dtype: float64

HETEROSPOROUS ONLY
count 3583.000000
mean 5.730851
std 9.370177
min 0.080000
25% 0.750000
50% 1.750000
75% 6.165000
max 102.900000
Name: DNA Amount1C (pg), dtype: float64

BOTH (HETEROSPOROUS & HOMOSPOROUS)
count 3656.000000
mean 5.982940
std 9.932141
min 0.080000
25% 0.770000
50% 1.815000
75% 6.700000
max 102.900000
Name: DNA Amount1C (pg), dtype: float64