The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers

Introduction

For nearly a century, biologists, and botanists in particular, have been interested in the determination and documentation of chromosome numbers for extant taxa (reviewed in Goldblatt & Lowry, 2011) as well as extinct ones (Laane & Holand, 1986; Masterson, 1994). These data have been widely used to evaluate the evolutionary pattern of chromosome number change and to estimate the base chromosome number of clades of interest. Chromosome numbers have also been extensively utilized as an important phylogenetic character in the context of cytotaxonomy (Chatterjee & Kumar Sharma, 1969; Schlarbaum & Tsuchiya, 1984; Guerra, 2012). Perhaps the most influential use of chromosome number data has been in the inference of major genomic events such as whole genome duplications (polyploidy), as well as changes in single chromosome numbers (e.g. diploidy). Early researchers analyzed the distribution of chromosome numbers within a group of interest and employed various threshold techniques to estimate ploidy levels for the analyzed taxa (Stebbins, 1938; Grant, 1963; Goldblatt, 1980). More recently, phylogenetic information was incorporated into the analyses, allowing researchers to infer transitions in chromosome numbers along branches of the tree using either the maximum parsimony principle (Schultheis, 2001; Hansen et al., 2006; Ohi-Toma et al., 2006; Wood et al., 2009) or by using a probabilistic evolutionary model within the likelihood paradigm (Mayrose et al., 2010; Cusimano et al., 2012; Glick & Mayrose, 2014).

Due to their significance and the relative ease by which chromosome numbers can be obtained, it is not surprising that chromosome number is the most extensively and consistently recorded cytological property in most plant families and genera (Guerra, 2008). These data have been documented along the years in an array of journal manuscripts, printed books (Löve & Löve, 1948; Darlington & Wylie, 1955; Fedorov, 1969) and, more recently, in the form of online databases (Goldblatt & Johnson, 1979; Watanabe, 2002; Bennett & Leitch, 2011). To date, the most comprehensive data source is the Index to Plant Chromosome Numbers (IPCN; Goldblatt & Johnson, 1979), which provides reference point to original chromosome counts reported in the literature. IPCN was initially established at the University of California Berkeley in the 1950s and was later maintained by Canada Department of Agriculture, Missouri Botanical Garden, and currently by the International Association for Plant Taxonomy (IAPT). A large portion of the counts referenced during 1979–2006, the years that IPCN has been housed in the Missouri Botanical Garden, can be accessed and searched online. Counts reported in more recent years are currently published under IAPT/IOPB Chromosome Data series (Marhold, 2006) but are not stored within a central, easily searched, database. In addition to IPCN, several other online data sources are available, most of which are dedicated to either a specific geographical region (Slovakia – Marhold et al., 2007; Poland – Góralski et al., 2009 onwards) or to a certain taxonomic group (e.g. Hieracium – Schuhwerk, 1996; Asteraceae – Watanabe, 2002).

The amount of chromosome counts that exist to date is extensive, and searching the large number of resources that contain such information is a daunting task, particularly when a large number of taxa is examined. Consequently, many researchers search for chromosome number information only through the largest online database(s), while smaller but nonetheless valuable sources are ignored. This usually results in missing data for some of the species in question, which may lead to erroneous conclusions drawn from the analysis. Obviously, a large accessible database that unifies all currently known databases, including both printed and online sources, would be of great value to the botanical community and would make the task of data collection much easier. In addition, such a central resource would enable researchers to add new counts as soon as they are being reported, facilitating the task of data sharing.

Here, we present the Chromosome Counts Database (CCDB), as a community resource of plant chromosome numbers. The database incorporates data from dozens of sources, more than doubling the amount of data available within any single resource. The online database additionally enables researchers to add new counts or to comment on existing data entries, thereby facilitating data sharing. The extensive amount of data currently available in CCDB further allowed us to analyze the patterns of chromosome number distribution among major plant groups. We estimate the percentage of plant species exhibiting intraspecific variation in chromosome numbers as well as in their ploidy levels.

Materials and Methods

Data collection

Chromosome counts were collected from a large number of electronic resources, older chromosome counts compendiums in the form of printed books, and an array of miscellaneous sources such as floras, monographs and other scientific manuscripts. The full list of resources is given in Table 1. Data from these sources were collected using the following procedures:
Online chromosome counts databases  Data from several online databases were retrieved directly from the database curator via personal communication in the form of comma-separated value (CSV) files. These include data from the Plant DNA C-values database (Bennett & Leitch, 2011; obtained from Ilia Leitch) and Chromosome number database of Polish plants (Góralski et al., 2009 onwards; obtained from Grzegorz Góralski). Other online chromosome counts databases were downloaded and processed using Perl/Python scripts. The following online sources were retrieved: IPCN (Goldblatt & Johnson, 1979–), Chilean plants cytogenetic database (Jara-Seguel & Urrutia, 2011), CHROBASE – Chromosome numbers for the Italian flora (Bedini et al., 2010 onwards), BSBI cytology database [accessed 20 June 2013] (http://rbg-web2.rbge.org.uk/BSBI/cytsearch.php), Index to chromosome numbers in Asteraceae (Watanabe, 2002), Published chromosome counts in Hieracium (Schuhwerk, 1996), ChromoPar – Paraguay chromosome counts database [accessed 12 June 2013] (http://www.ub.edu/botanica/cromopar/), Karyological database of the genus Cardamine (Kuccera et al., 2005) and Chromosome number survey of the ferns and flowering plants of Slovakia (Marhold et al., 2007).

Chromosome counts compendiums available as hard copy  In addition to online sources as already described, we have obtained well-known and widely used printed books containing chromosome counts indexes. The data in these books were retrieved in the following way: first, the books were scanned to generate image files. Then, using the optical character recognition (OCR) tool of Adobe Pro the files were converted to ‘textable’ PDF files. This OCR tool

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Chromosome counts resources incorporated in Chromosome Counts Database (CCDB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource name</td>
<td>Data entries</td>
</tr>
<tr>
<td>(a) Online resources</td>
<td></td>
</tr>
<tr>
<td>IPCN <a href="http://www.tropicos.org/Project/IPCN">www.tropicos.org/Project/IPCN</a></td>
<td>111 224</td>
</tr>
<tr>
<td>Plant DNA C-values database <a href="http://data.kew.org/cvalues">http://data.kew.org/cvalues</a></td>
<td>5889</td>
</tr>
<tr>
<td>BSBI cytology (British Isles) <a href="http://rbg-web2.rbge.org.uk/BSBI/cytsearch.php">http://rbg-web2.rbge.org.uk/BSBI/cytsearch.php</a></td>
<td>2766</td>
</tr>
<tr>
<td>Chilean Plants Cytogenetic Database <a href="http://www.chileanpcd.com">http://www.chileanpcd.com</a></td>
<td>224</td>
</tr>
<tr>
<td>ChromoPar (Paraguay) <a href="http://www.ub.edu/botanica/cromopar">http://www.ub.edu/botanica/cromopar</a></td>
<td>1278</td>
</tr>
<tr>
<td>CHROBASE (Italy) <a href="http://www.biologia.unipi.it/chrobase">http://www.biologia.unipi.it/chrobase</a></td>
<td>6517</td>
</tr>
<tr>
<td>Slovakia’s Karyological database <a href="http://www.chromosomes.sav.sk">http://www.chromosomes.sav.sk</a></td>
<td>7734</td>
</tr>
<tr>
<td>Chromosome number database of Polish plants <a href="http://chromosomes.binoz.uuj.edu.pl">http://chromosomes.binoz.uuj.edu.pl</a></td>
<td>2183</td>
</tr>
<tr>
<td>Index to chromosome numbers in Asteraceae</td>
<td>23 800</td>
</tr>
<tr>
<td><a href="http://www.lib.kobe-u.ac.jp/infolib/meta_pub/G0000003asteraceae_e">http://www.lib.kobe-u.ac.jp/infolib/meta_pub/G0000003asteraceae_e</a></td>
<td>2966</td>
</tr>
<tr>
<td>Published chromosome-counts in Hieracium</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.botanischestaatsammlung.de/projects/chrzlit.html">http://www.botanischestaatsammlung.de/projects/chrzlit.html</a></td>
<td></td>
</tr>
<tr>
<td>(b) Hard-copy resources</td>
<td></td>
</tr>
<tr>
<td>Cytotaxonomical atlas of the Pteridophyta (Löve et al., 1977)</td>
<td>4216</td>
</tr>
<tr>
<td>Chromosome atlas of flowering plants (Darlington &amp; Wylie, 1955)</td>
<td>11 741</td>
</tr>
<tr>
<td>Chromosome Numbers of Flowering Plants (Fedorov, 1969)</td>
<td>30 544</td>
</tr>
<tr>
<td>Chromosome atlas of flowering plants of the Indian subcontinent Vol1 (Kumar &amp; Subramaniam, 1987a)</td>
<td>7409</td>
</tr>
<tr>
<td>Chromosome Atlas of Flowering Plants of the Indian Subcontinent Vol2 (Kumar &amp; Subramaniam, 1987b)</td>
<td>4497</td>
</tr>
<tr>
<td>Index to plant chromosome numbers, 1965 (Ornduff, 1967)</td>
<td>3900</td>
</tr>
<tr>
<td>Index to plant chromosome numbers, 1966 (Ornduff, 1968)</td>
<td>3836</td>
</tr>
<tr>
<td>Index to plant chromosome numbers, 1967–1971 (Moore, 1973)</td>
<td>21 996</td>
</tr>
<tr>
<td>Index to plant chromosome numbers, 1972 (Moore, 1974)</td>
<td>3622</td>
</tr>
<tr>
<td>Index to plant chromosome numbers, 1973–1974 (Moore, 1977)</td>
<td>10 243</td>
</tr>
<tr>
<td>Index to plant chromosome numbers, 1975–1978 (Goldblatt, 1981)</td>
<td>12 525</td>
</tr>
<tr>
<td>Chromosome numbers of northern plant species (Löve &amp; Löve, 1948)</td>
<td>1352</td>
</tr>
<tr>
<td>Flora Europaea – checklist and chromosome index (Moore, 1982)</td>
<td>4490</td>
</tr>
<tr>
<td>(c) Miscellaneous resources</td>
<td></td>
</tr>
<tr>
<td>IAPT/IOPB chromosome data (Marhold, 2006)</td>
<td>4123</td>
</tr>
<tr>
<td>Efflora (<a href="http://www.effloras.org/">http://www.effloras.org/</a>)</td>
<td>11 405</td>
</tr>
<tr>
<td>Flora ibérica</td>
<td>9603</td>
</tr>
<tr>
<td>Interactive flora of NW Europe (<a href="http://wbd.etibioinformatics.nl/bis/flora.php">http://wbd.etibioinformatics.nl/bis/flora.php</a>?)</td>
<td>2764</td>
</tr>
<tr>
<td>Araceae chromosome numbers (Cusimano et al., 2012)</td>
<td>1026</td>
</tr>
<tr>
<td>Brassicaceae chromosome numbers (Warwick &amp; Al-Shehbaz, 2006)</td>
<td>8685</td>
</tr>
<tr>
<td>Cyperaceae chromosome numbers (Roalson, 2008)</td>
<td>2818</td>
</tr>
<tr>
<td>Veronicaceae chromosome numbers (Albach et al., 2008)</td>
<td>2491</td>
</tr>
<tr>
<td>Chromosome atlas of the New Zealand flora (M. Dawson, pers. comm.)</td>
<td>2255</td>
</tr>
<tr>
<td>Solanaceae chromosome numbers (E. Goldberg, pers. comm.)</td>
<td>2001</td>
</tr>
</tbody>
</table>

Only major resources are given. A full list of resources is available at http://ccdb.tau.ac.il/about/.

© 2014 The Authors
New Phytologist © 2014 New Phytologist Trust

www.newphytologist.com
was chosen because it exhibited the most accurate performance compared to five other OCR tools in an initial screen of several books. In the next step we used ‘Some PDF to Text Converter’ (available through www.somepdf.com), which converted the PDF files into plain text files that could be parsed automatically using Python scripts. Because this whole automated process suffers from some inaccuracies – particularly due to errors related to the OCR conversion (e.g. occasional confusion between ‘l’, ‘L’, and ‘!’) – thousands of counts were manually verified. In addition, our general approach in processing such sources was to maximize retrieval accuracy rather than data completeness. Consequently, not all data available through the target source were retrieved. It should be emphasized that occasional errors may still remain (this is particularly so for the compendium published by Fedorov, 1969, for which OCR errors are more abundant due to the Cyrillic font and tables/columns included within the text) and CCDB allows users to report such cases. The following sources were retrieved this way: Chromosome numbers of northern plant species (Löve & Löve, 1948), Chromosome atlas of flowering plants (Darlington & Wylie, 1955), Cytotaxonomical atlas of the Pteridophyta (Löve et al., 1977), Chromosome numbers of flowering plants (Fedorov, 1969), Flora Europaea – checklist and chromosome index (Moore, 1982), Chromosome atlas of flowering plants of the Indian subcontinent; volumes 1 and 2 (Kumar & Subramaniam, 1987a) and Index to plant chromosome numbers for the years 1965–1974 (Ornduff, 1967, 1968; Moore, 1970, 1971, 1973, 1974, 1977). The IPCN volume for the years 1975–1978 (Goldblatt, 1981) was also parsed but counts were inserted into the database only in case the online IPCN database did not already contain them.

Floras, journal manuscripts and other resources In addition to dedicated chromosome counts databases and hard copy books, a large number of other sources exist that contain information regarding the chromosome number for a given taxon. These resources include floras, monographs and an array of scientific manuscripts. However, automatic retrieval of chromosome number data from such resources is not a trivial task because the data are organized in a source-specific manner (e.g. the botanical description of a given species as appears in its relevant flora obtained through http://www.efloras.org). Hence, the downloading and processing of each data source were performed using dedicated Perl/Python scripts written specifically for each data source, followed by a manual verification of hundreds of records. As mentioned above, we preferred to maximize data accuracy over data completeness and therefore some fraction of the data available in these sources was not used.

Thousands of chromosome counts were acquired from online floras – efloras [accessed 20 October 2013] (http://www.efloras.org), Flora Iberica [accessed 20 June 2013] (http://www.floraiberica.es), and from the Interactive flora of NW Europe [accessed 20 June 2013] (http://wbd.etibioinformatics.nl/bis/flora.php). In addition to floras, chromosome counts that appear within several Systematic Botany Monographs were retrieved (Saunders, 2000; Bohs, 2001; Freire-Fierro, 2002; Aldasoro et al., 2004; Zuloaga et al., 2004; Thompson, 2005; Wagner et al., 2005; Meudt, 2006; Miller & Chambers, 2006).

Scientific manuscripts that contain large amounts of chromosome counts were parsed in a source-specific manner and incorporated into the database. IAPT/IOPB Chromosome Data reports 1–16 (Marhold, 2006) were obtained from the International Organization of Plant Biosystematists website (http://www.iopb.org/) as PDF files, converted to text files and parsed using Perl scripts. In addition, a large number of journal manuscripts that contain counts for a given taxonomic group or geographic region were obtained and parsed in a source-specific procedure. These include data reported in a large number of Mediterranean chromosome number reports (Kamari et al., 1991), as well as large collections available for Araeaceae (Cusimano et al., 2012), Brassicaceae (Warwick & Al-Shehbaz, 2006), Colchicaceae (Chacón et al., 2014), Cyperaceae (Roalson, 2008), Pinguicula (Casper & Stipmer, 2009), and Veronicaceae (Albach et al., 2008). The full list of scientific manuscripts that were incorporated into CCDB is available through the database help pages (http://ccdb.tau.ac.il/about/).

Finally, chromosome counts datasets that were compiled by individual researchers were obtained via personal communication. These include chromosome numbers of indigenous New Zealand plants obtained from Murray Dawson and chromosome numbers for a large number of Solanaceae species obtained from Emma Goldberg.

Name resolution

Combining data from multiple sources required a method for standardization of the information, especially regarding the taxonomy of the records. Many plant species have been given different names by different authors. Some of these names are considered synonyms, others are recognized as accepted names, while another fraction is still unresolved. Another common problem is differences in spelling conventions between sources, or simply spelling mistakes, resulting from either manual typing errors in the original source, or incorrect processing of our automatic pipelines. To overcome these difficulties, we used Taxonome (Kluyver & Osborne, 2013), a taxonomic name resolution software that provides the ability to match synonymous taxon names to accepted names while accounting for differences in naming conventions and likely misspellings. As the underlying database for names, we used a local repository of synonymous and accepted names that was created based on The Plant List (TPL) v1.1 (http://www.theplantlist.org/) with some modifications (i.e. for Solanaceae we used Solanaceae Source (http://solanaceae.source.org/) as the primary taxonomic source supplemented with The Plant List for missing taxon names). In case a taxon name could not be matched to a recognized plant name (e.g. due to erroneous OCR processing), the corresponding data entry was excluded from the database.

Data access

CCDB is available through http://ccdb.tau.ac.il/. Users can access the data by browsing through the taxonomic hierarchy or by searching for a specific genus or species. At each level, all counts can
be retrieved as a CSV file. Additionally, users can access the data through the dedicated application programming interface (API), available through http://ccdb.tau.ac.il/services/. Researchers are invited to contribute to the completeness and correctness of the resource. This can be achieved by submitting new data, originating from resources not yet incorporated into the database as well as reporting errors found in the database. We note that unlike in IPCN, new data entries will not be thoroughly reviewed. Thus, data contributors are strongly encouraged to include supporting information such as voucher specimen or an image file of the cells analyzed.

Results

CCDB encompasses a wide array of resources, the majority of which were unavailable before in a digitized format. At present, CCDB contains 334 963 data entries, encompassing chromosome counts for 171 338 unique taxon names, including species names and infraspecific names. Following a taxonomic name resolution process that collapsed synonymous names to their accepted names, the number of unique names in CCDB is 77 958 (of these 68 146 are accepted names and 9812 are unresolved according to TPL V1.1). This represents a substantial increase in data coverage compared to IPCN – the largest online resource to date – that has information for a total of 60 167 plant names (48 829 following name resolution). Table 1 specifies the number of counts extracted from each source, as well as the number of unique names before and after name resolution.

CCDB includes a total of 8750 genera from 539 families. The coverage of CCDB varies widely across the major plant groups. The current coverage for angiosperms is 19% (58 980 out of 304 419 accepted species as reported in TPL V1.1 – not including data available for infraspecific names). The exact coverage, however, vary between 12% and 23% depending on the assumed number of angiosperm species, with estimates ranging from 12% to 500 000 if yet undiscovered species are considered (as discussed in Galbraith et al., 2011). The estimated coverage for pteridophytes (here and in the online database referred to as the monilophytes and lycophytes clades), bryophytes and gymnosperms is 22% (2350/10 620), 4% (1436/34 556) and 38% (427/1104), respectively. Within the 20 largest angiosperm families (Supporting Information Table S1), the best covered one is Apiaceae, with counts available for 42% of the taxa (1474 out of 3509), while the coverage for the largest plant family, the Compositae, is 32% (11 776 out of 36 700). Of the 20 largest families, the least covered one is Bromeliaceae with 7%. Our compilation also highlights some additional families where chromosome count data are particularly lacking and where additional efforts should be particularly beneficial. Some of the least represented families in CCDB include the Daltoniaceae (having only one count out of 328 accepted names), Vochysiaceae (1/225) and Calophyllaceae (1/131).

In order to estimate the completeness of the data obtained through CCDB compared to the maximal availability of chromosome count information (i.e. all counts ever reported in the literature), we compared the coverage of CCDB relative to that obtained in five previous studies. Each of these studies assembled chromosome-number information in a detailed manner for a specific plant clade, and we thus regard those as approximately representing all available data for these groups (Pinguicula – Casper & Stimper, 2009; Araceae – Cusimano et al., 2012; Solanaceae – E. Goldberg, pers. comm.; Colchicaceae – Chacón et al., 2014; Danthonioideae – Linder & Barker, 2014). In these comparisons, we calculated the fraction of species in the reference dataset for which information exists in CCDB while considering data entries obtained from other resources only (because the data obtained from the above five studies were already incorporated in CCDB). As demonstrated in Table 2, for several clades, such as Araceae and Colchicaceae, data completeness of CCDB is very high, nearly reaching that obtained by meticulous manual searches. However, for other clades (i.e. Pinguicula) our data retrieval was not as complete, missing roughly half of the data that have been previously reported. Notably, even for the least covered group, data availability in CCDB constitutes a major improvement compared to what is currently available through IPCN (Table 2). These results emphasize the need for a community effort aimed towards improving accessibility to the vast amount of chromosome number information that has been determined over the years, but appears sporadically within scientific manuscripts and thus is regularly missed.

Using the chromosome counts data assembled in CCDB, we next examined the distribution of the haploid chromosome numbers within each of the major plant groups. In case more than one count was available for a certain taxon, the median was taken as the representative count. As has been previously observed in ferns (Otto & Whitton, 2000), there are more even haploid numbers than odd ones (across the whole database the median chromosome number for 42 161 taxa is even and for 33 317 it is odd; Table 3), resulting in a ‘saw-toothed’ pattern (Fig. S1a). As noted by Otto & Whitton (2000), this pattern can be explained by frequent polyploidization events, because a genome duplication will always result in an even number while other changes in chromosome numbers (e.g. via dysploidy) can lead to both even and odd

<p>| Table 2. Chromosome Counts Database (CCDB) coverage in five manually curated plant clades |
|---------------------------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Clade</th>
<th>Number of taxa with data according to original study</th>
<th>Number of taxa in CCDB</th>
<th>Number of taxa in IPCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Araceae</td>
<td>740</td>
<td>649 (88%)</td>
<td>524 (71%)</td>
</tr>
<tr>
<td>Colchicaceae</td>
<td>144</td>
<td>137 (95%)</td>
<td>118 (82%)</td>
</tr>
<tr>
<td>Pinguicula</td>
<td>68</td>
<td>36 (53%)</td>
<td>22 (32%)</td>
</tr>
<tr>
<td>Solanaceae</td>
<td>846</td>
<td>617 (73%)</td>
<td>474 (56%)</td>
</tr>
<tr>
<td>Poacea subfamily</td>
<td>161</td>
<td>127 (79%)</td>
<td>112 (70%)</td>
</tr>
</tbody>
</table>

*aFor comparison, the number of taxa in the original study is reported following the name resolution process described in the Materials and Methods section (i.e. omitting those that were not successfully resolved). 

*bThe number of species reported in CCDB excluding counts obtained from the original studies. IPCN, Index to Plant Chromosome Numbers.
numbers. Interestingly, the chromosome number distribution varies markedly between the major plant groups (Fig. 1). In monilophytes (Fig. 1a), a clade known to possess particularly high chromosome numbers (reviewed in Barker, 2013), the most common haploid number is 41, followed by 36 with two additional peaks at 82 and 72 that are exact duplications of the two most common numbers. Additionally, while 63% (1887 out of 2986) of the species possess an even chromosome number, the even-to-odd ratio increases substantially considering counts larger than the modal number (for counts above 41, 79% of the species have an even haploid number), suggesting that chromosome number increases are mainly the result of polyploidy transitions. In lycophytes, three distinct peaks are observed (Fig. 1b): the lowest peak c. 9–11 comprises mostly chromosome counts originating from Isoetales and Selaginellales, a second peak c. 22–23 that includes counts from Isoetales and Lycopodiales, and a third peak c. 34 of Lycopodiales species. In angiosperms (Fig. S1b), as is also reflected in the distribution obtained for eudicots (Fig. 1c), the modal number is more diffused and is centered c. 7–12 and the saw-toothed pattern is noticeable for chromosome numbers larger than 12. While 56% of angiosperms have an even haploid number, the even-to-odd ratio changes substantially above the major mode – the ratio between even and odd numbers below 12 is 0.95 (i.e. slightly more odds than evens), whereas for 13 and over it is 1.7. As far as chromosome numbers are concerned, it seems that plants possessing low chromosome numbers have undergone a polyploidy event so long ago that its signal has been eroded by subsequent dysploidy events. When considering the two main angiosperm clades, monocots were shown to have undergone more frequent polyploidy events compared to eudicots (Otto & Whitton, 2000). Indeed, the saw-toothed pattern for monocots (Fig. 1d) is

\[
\begin{array}{|c|c|c|}
\hline
\text{Clade} & \text{Number of total counts} & \text{Percentage of even counts} \\
\hline
\text{Angiosperm} & 70338 & 56\% \\
\text{Monocots} & 15528 & 58\% \\
\text{Eudicots} & 53492 & 55\% \\
\text{Monilophytes} & 2986 & 63\% \\
\text{Lycophytes} & 220 & 53\% \\
\text{Gymnosperms} & 488 & 59\% \\
\text{Bryophytes} & 1446 & 48\% \\
\hline
\end{array}
\]

\(^{4}\text{In these analyses ambiguous counts were not considered \textit{(e.g. odd 2n counts, those appearing within parentheses, or those containing Roman numerals due to the higher OCR error rate).}\)

![Fig. 1 The distribution of haploid chromosome numbers across the major plant groups. The distribution is calculated across (a) monilophytes, (b) lycophytes, (c) eudicots, (d) monocots, (e) gymnosperms and (f) bryophytes.}](image-url)
particularly apparent with an even-to-odd ratio of 1.7 above the modal count of 7. In gymnosperms (Fig. 1e) – a group in which polyploidy is considered rare (Husband et al., 2013) – there is a high percentage of even counts (59%). However, this is due to the modal count of 12 (47% of all species) and the saw-toothed pattern is not apparent. In bryophytes, no apparent saw-toothed pattern was observed (Fig. 1f), with a relatively diffused mode between 6 and 13.

Next, we examined the extent by which chromosome number varies within resolved named species and infraspecific taxa (i.e. considering subspecies and varieties distinct from the corresponding species). Our analysis revealed that cytotype polymorphism is frequent within named species and infraspecific taxa, existing in 22.7% of taxa in our database; 15% of taxa were reported with two distinct counts and 7.7% with three or more cytotypes. Moreover, repeating this analysis at the species level (i.e. by collapsing all infraspecific names to their corresponding binomials), revealed that intraspecific variation in chromosome numbers exists in 23.5% (16 379 out of 69 639) of species in our database (15.2% of species were reported with two distinct counts and 8.3% with three or more). With the exception of gymnosperms, the frequency of species with multiple counts is relatively similar across the major lineages (23.6%, 26.5%, 22.1%, 20.1% and 12.1% for angiosperms, monilophytes, lycophytes, bryophytes and gymnosperms, respectively). These frequencies are obviously an underestimation due to the incompleteness of the database (i.e. not all reported cytotypes are included in CCDB) and since the karyotypes of many distinct cytotypes were not determined.

The multiple cytotypes that exist within nearly one quarter of named plant species encompass cases that affect only the karyotype but not the genomic content (e.g. chromosome fusion) and those that affect both as a result of major genomic processes such as polyploidy. As suggested by Soltis et al. (2007), a significant fraction of such intraspecific ploidal variants arose through autoploidy events. In many cases, these autoploids should be treated as distinct species under most commonly used species concepts. Thus, we examined the extent to which intraspecific variation in chromosome numbers can be attributed to ploidal variants using a simple nonphylogenetic approach. To this end, for each polymorphic species the ploidy index for all its cytotypes was defined as the multiplication factor relative to the lowest chromosome number found in that species (e.g. if the reported gametophytic counts for a certain species were 10, 15 and 20 the respective multiplication factors were 1.5 and 2). As shown in Fig. 2, a very large fraction of the observed intraspecific variation is due to polyploidy. Clearly, the most common factor is 2, which corresponds to a single whole genome duplication; next are the factors 3, 4, 5 and 6, each corresponding to chromosome number changes due to polyploidy. In addition, the frequency c. 1 is relatively high and could be explained by dysploidy events (such as chromosome fusion and fusion) while another peak is observed c. 1.5, corresponding to the occurrence of triploid taxa. In order to evaluate the relative contribution of polyploidy to intraspecific chromosome number variation compared to other processes of chromosome-number change, a threshold of 1.4 was used. Assuming that this threshold can be used to distinguish polyploidy events (including transition to triploids) from dysploidy transitions, 69% of the observed intraspecific variation is due to polyploidy, whereas 31% are due to other types of chromosome number transition. In total, our analysis revealed that 16.2% of plant species harbor intraspecific variation in their ploidy levels – higher than the estimate provided by Wood et al. (2009) who reported that 12–13% of angiosperms and 17% of fern species (in that study including the lycophytes) harbor multiple ploidy levels (compared to 16.2% in angiosperms and 19.7% in pteridophytes: 20.1% in monilophytes and 12.9% in lycophytes observed in our analysis). The difference in estimates stems mainly from the different cutoff used by Wood et al. who disregarded triploids in their estimate (by using a threshold of 1.75) but also due to the additional data incorporated in CCDB.

**Discussion**

Here, we presented the Chromosome Counts Database, as a community resource for plant researchers. While CCDB represents a step towards enhanced data coverage and accessibility, for certain clades data completeness is still lacking. CCDB may thus guide other global initiatives, such as those concerning the collection of C-values (Galbraith et al., 2011) by pointing out taxonomic groups where collection efforts could be particularly rewarding. The current coverage for angiosperms in CCDB is c. 20% while Bennett
(1998) estimated this number to be c. 25%. While the difference in these estimates may also stem from the number of angiosperm species assumed, there are obviously additional data that CCDB does not contain. For example, in bryophytes merely 4% of the species have chromosome-number information in CCDB, while Husband et al. (2013) estimated the coverage for bryophytes to be three times higher. Importantly, the estimation reported by Husband et al. (2013) was based on two printed sources (Fritsch, 1991; Przywara & Kuta, 1995), which in the current compilation of CCDB were not available. Such gaps in coverage can be readily filled by the community either by uploading directly through the CCDB website or by providing data in the form of a printed/scanned copy, which can be automatically processed using the developed procedures. Our goal in the construction of CCDB was to provide an extensive, yet flexible framework within which additional data can be added by the community, thus facilitating data sharing for a wide array of in-depth studies concerning the pattern of chromosome number change.

Acknowledgements

We thank Michael S. Barker, Emma Goldberg and Murray Dawson for providing us with extensive chromosome number collections; Ilia Leitch for providing a CSV file of the Plant DNA C-value database and Grzegorz Góralski for providing a CSV file of the chromosome number database of Polish plants; Sarah P. Otto for providing a scanned copy of the book Cytotaxonomical atlas of the Pteridophyta, and Aretuza Sousa, Susanne S. Renner, and an anonymous reviewer for constructive comments. This study was supported by a fellowship from the Manna Center Program in Food Safety & Security to L.G. and by the Israel Science Foundation grant number 1265/12.

Anna Rice†, Lior Glick†, Shiran Abadi†, Moshe Einhorn, Naama M. Kopelman, Ayelet Salman-Minkov, Jonathan Mayzel, Ofer Chay and Itay Mayrose*†

Department of Molecular Biology and Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

(*Author for correspondence: tel +972 3 640 7212; email itaymay@post.tau.ac.il)
†These authors contributed equally to this work.

References

Fedorov AA. 1969. Chromosome numbers of flowering plants. Leningrad, USSR: Academy of Natural Sciences of the USSR.


Supporting Information

Additional supporting information may be found in the online version of this article.

**Fig. S1** The distribution of haploid chromosome numbers in CCDB across all taxa and angiosperms.

**Table S1** Data coverage in CCDB for the 20 largest plant families

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

**Key words:** chromosome counts, Chromosome Counts Database (CCDB), chromosome numbers, database, dysploidy, intraspecific variation, polyploidy.