The Cerrado (Brazil) plant cytogenetics database

Fernando Roa[1], Mariana Pires de Campos Telles[1,2]

1 *Programa de pós-graduação em Genética e Melhoramento de Plantas, Escola de Agronomia, Universidade Federal de Goiás, 74001-970, Goiânia, GO, Brazil* **2** *Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás, 74605-010, Goiânia, GO, Brazil*

Corresponding author: *Fernando Roa* (fernando.roa@ufpe.br)

Abstract

Cerrado is a biodiversity hotspot that has lost ca. 50% of its original vegetation cover and hosts ca. 11,000 species belonging to 1,423 genera of phanerogams. For a fraction of those species some cytogenetic characteristics like chromosome numbers and C-value were available in databases, while other valuable information such as karyotype formula and banding patterns are missing. In order to integrate and share all cytogenetic information published for Cerrado species, including frequency of cytogenetic attributes and scientometrics aspects, Cerrado plant species were searched in bibliographic sources, including the 50 richest genera (with more than 45 taxa) and 273 genera with only one species in Cerrado. Determination of frequencies and the database website (http://cyto.shinyapps.io/cerrado) were developed in R. Studies were pooled by employed technique and decade, showing a rise in non-conventional cytogenetics since 2000. However, C-value estimation, heterochromatin staining and molecular cytogenetics are still not common for any family. For the richest and best sampled families, the following modal 2n counts were observed: Oxalidaceae 2n = 12, Lythraceae 2n = 30, Sapindaceae 2n = 24, Solanaceae 2n = 24, Cyperaceae 2n = 10, Poaceae 2n = 20, Asteraceae 2n = 18 and Fabaceae 2n = 26. Chromosome number information is available for only 16.1% of species, while there are genome size data for only 1.25%, being lower than the global percentages. In general, genome sizes were small, ranging from 2C = ca. 1.5 to ca. 3.5 pg. Intraspecific 2n number variation and higher 2n counts were mainly related to polyploidy, which relates to the prevalence of even haploid numbers above the mode of 2n in most major plant clades. Several orphan genera with almost no cytogenetic studies for Cerrado were identified. This effort represents a complete diagnosis for cytogenetic attributes of plants of Cerrado.

Keywords

Cerrado, cytogenetics, database, scientometrics, shiny-R

Introduction

Cerrado, a phytogeographic domain from Brazil, is the third largest biodiversity hotspot in the world considering species endemism and degree of threat (Myers et al. 2000; Forzza et al. 2012). In fact, it has suffered more anthropic impact than the Amazonian tropical forest (Spera et al. 2016). With originally 2m km^2 of wilderness area, of which ca. 50% are currently cultivated, this region hosts ca. 11,000 species belonging to 1,423 genera of 171 families of phanerogams (Sano et al. 2008; Ferreira et al. 2016). Classical cytogenetic studies, comprehending somatic or meiotic chromosome counts are known for a considerable portion of species occurring in Cerrado but in many cases they were not performed on plants from this region. Meanwhile, studies of genome size, differential staining and molecular cytogenetics are very rare.

Since the chromosome number is the most basic karyotype feature and it can be observed by conventional staining of meristem cells or pollen mother cells, most cytogenetic databases deal only with this attribute (Peruzzi and Bedini 2014; Rice et al. 2015). Estimates of genome size, which are currently addressed mostly by comparing the relative fluorescence of propidium iodide stained nuclei measured in a flow cytometer to that of known patterns, are compiled in the Kew C-value database (Doležel and Greilhuber 2010; Garcia et al. 2014a). Other valuable information such as: karyotype formula, dependent on the relative position of centromere along the chromosome and chromosome length (Guerra 1986), silver staining of nucleolar organizer regions (AgNOR) (Vieira et al. 1990), and heterochromatin staining after treatment with acids, bases or denaturation (Schwarzacher et al. 1980; Guerra 2000; Barros e Silva and Guerra 2010) are not included in databases. In contrast, fluorescence in situ hybridization (FISH), in which a DNA probe produced by molecular biology methods anneals with chromosome preparations and is detected by fluorescence has gained more attention (Roa and Guerra 2012; Garcia et al. 2014b; Roa and Guerra 2015).

The aim of this study was to assess the current cytogenetic knowledge of Cerrado plants, aiming to provide a consistent database that didn't exclude any attribute and diagnose recent progress in the field.

Methods

Data compilation

Plant names in the Cerrado plant list (Sano et al. 2008) and the cytogenetic literature were checked according to the Brazilian Flora 2020 site (REFLORA 2016) through the www.plantminer.com app (Carvalho et al. 2010). In order to get to the original cytological sources, websites in Table 1 were used. The information in the primary source was organized in the following fields: family (APG 2016), genus, original name as reported in the publication (field = name_on_source), accepted name following the online Brazilian Flora 2020 (field = accepted_name), place of sampling (field = provenance), dis-

tribution of C-, Chromomycin A$_3$, AgNOR-bands (fields = C-bands, CMA_bands, AgNOR), availability of images (field = images), karyotype formula (field = karyotype), FISH sites position (field = FISH), genome size (field = C-value), meiotic or gametophyte chromosome number (field = meiosis-n), sporophytic chromosome number-2n (field = 2n), level of ploidy (field = ploidy), average chromosome size (field = ACS), Total chromosome length (field = TCL), Total chromosome area (field = TCA), reference, and authors' affiliation (field = affiliation). Metadata were gathered partially by scripts from search engines as Scopus (Elsevier) and Crossref (PILA).

Data statistics

All statistics and the online database were performed in R (R Core Team 2016). Packages used were shiny (Chang et al. 2016), shinyjs (Attali 2016), shinydashboard (Chang 2015), ggplot2 (Wickham 2009), plyr (Wickham 2015a), dplyr (Wickham and Francois 2016), stringr (Wickham 2015b), data.table (Dowle et al. 2015), DT (Xie 2015), mongolite (Ooms 2016), gtools (Warnes et al. 2015), robustHD (Alfons 2016) and scales (Wickham 2016). Several frequencies were determined, such as: employed techniques or country of author affiliation per decade; chromosome number (2n), genome size, TCL and TCA per taxa. Where not available, TCL and TCA were measured from photos by ImageJ (Schneider et al. 2012). Due to the scarcity of genome size data, a correlation analysis between TCA/TCL and known genome sizes was performed in order to have a rough idea of genome size in pg based on TCL or TCA. In accessions without 2n number, they were calculated from meiosis or pollen observations (n).

Even n numbers are more common than odd ones because of the prevalence of even 2n that eventually undergo duplication (Otto and Whitton 2000; Rice et al. 2015). As expected, this trend is more pronounced for high 2n which are usually the result of whole genome duplication events, but may vary by dysploidy. Accessions were subdivided in two groups according to n under or over the mode (n) (low 2n, high 2n); and the even/odd ratio for each group was calculated in order to assess indirectly the degree of polyploidy/dysploidy in each major clade.

Results

Coverage

A subset of 38.9% (4,590 taxa) of the Cerrado Plant list was searched, including the 50 richest genera and 273 genera with only one species. Information of 1,431 accessions, from 366 available primary sources, were included in the database. 702 (16.8%) species of our searched sample (4182) had any cytogenetic data and about 500 of them had at least one accession collected in Brazil. Complete lack of information happens

for 70% of the single-species genera addressed (Suppl. material 1, 2, 3). The genera *Hyptis* Jacq., 1787 (Lamiaceae), *Stachytarpheta* Vahl, 1804 (Verbenaceae), *Microlicia* D.Don, 1823 (Melastomataceae), *Leiothrix* Ruhland, 1903 (Eriocaulaceae) and *Ditassa* R.Br., 1809 (Apocynaceae) may be considered orphan because they are among the 50 most diverse in Cerrado, but almost no chromosome numbers are known for them (Suppl. material 1). At the family level Eriocaulaceae (Monocots – Commelinids), Lamiaceae, Apocynaceae (Core eudicots-Superasterids-Asterids-Lamiids), Polygalaceae (Core eudicots-Superrosids-Fabids) and Lauraceae (Magnoliids) are also poorly studied (Suppl. material 2), while among the major clades (previously in parentheses), studies for Magnoliids are scarce. Every figure and table is interactive in the database website (http://cyto.shinyapps.io/cerrado).

For the richest and best sampled families, the following modal 2n counts were observed: Oxalidaceae 2n = 12, Lythraceae 2n = 30, Sapindaceae 2n = 24, Solanaceae 2n = 24, Cyperaceae 2n = 10, Poaceae 2n = 20, Asteraceae 2n = 18 and Fabaceae 2n = 26 (Suppl. material 2). At the genus level, *Paspalum* L., 1759 (Poaceae), *Cuphea* P.Browne, 1756 (Lythraceae) and *Rhynchospora* Vahl, 1805 (Cyperaceae) stand among the best studied with ca. 50% of coverage. For ca. 20% of species in most major clades, the chromosome number has been studied (Suppl. material 3).

Techniques chronology

Frequencies of use of karyological techniques show that the first cytogenetic analysis were mainly based on meiosis from 1928 on, and subsequently by 1990 the analysis of mitosis (sporophyte) gained more prevalence. Measurements of karyotype formula (morphology) achieved significance after the 1990's. A trend to include images in publications, only with a fall during the 2000's decade was observed. Sophisticated techniques that show a substantial rise from the 2000's decade are the estimation of the C-value, CMA banding (including C-CMA banding) and FISH. However, they have been applied to a limited number of species, ranging from 0.5 to 2.3% of the 1,431 entries in the database (Fig. 1).

Author affiliations

In order to track affiliation of authors over time, they were clustered by decade in a time-line. For accessions collected in Brazil [774 (54%)] the country of affiliation of the author showed an increase in the prevalence of Brazilian based research after the 1990's. Other significant contributions for Brazilian samples have been made by Argentinian based authors (Fig. 2). Most publishing authors for the 1935 – 1970 period were affiliated in the U.S.A, and afterwards, in Brazil. Among the authors with more entries, Coleman, Irwin and Turner studied in the 1935 – 1970 period mainly Asteraceae and Fabaceae. In the 1970 decade, Coleman studied Fabaceae and Nassar, Eu-
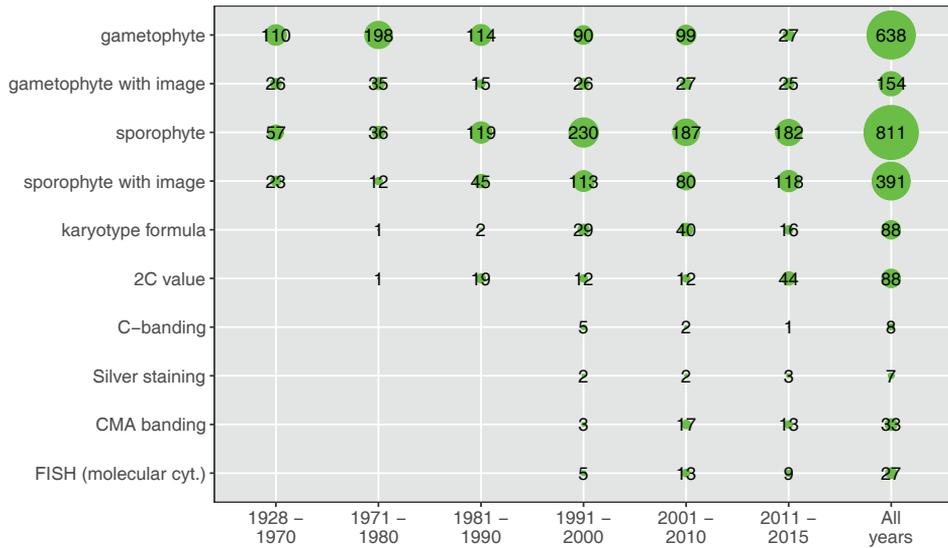
Figure 1. Chronology of cytogenetic techniques or parameters published for Cerrado plant species. Numbers of accessions in the database.
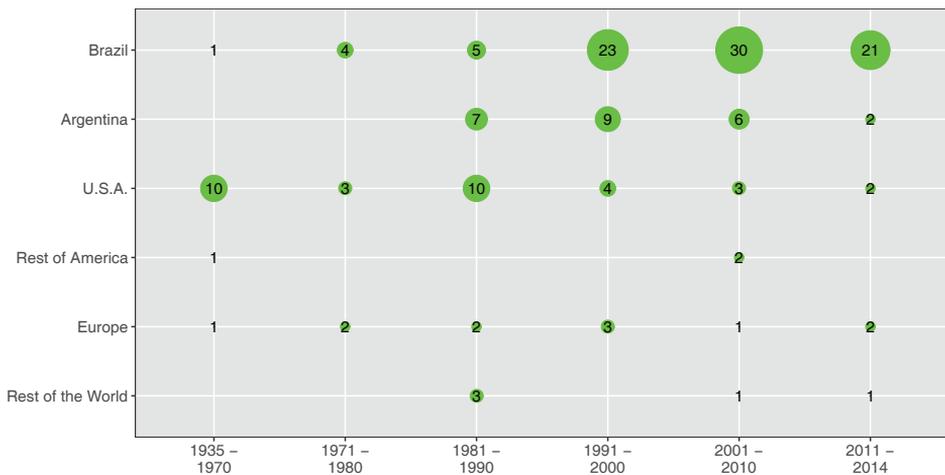

Figure 2. Chronology of author affiliation for Cerrado plant species collected in Brazil. Numbers of references in the database.

phorbiaceae. In the 80's and 2000's Graham studied Lythraceae, and in the 90's Guerra published mainly data on Cyperaceae, Orchidaceae and Velloziaceae. In the current decade, Félix published mainly about Euphorbiaceae and Fabaceae (Suppl. material 4). Accordingly, those authors correspond to the most important nodes in the co-author network (database website).

Chromosome numbers and genome size

699 (16.1%) species of the searched list have known chromosome numbers with a general median of $2n$ = 28. Fig. 3 shows two ploidy related prevalent chromosome numbers; neopolyploids in Commelinids, which corresponds to *Paspalum* counts (2n = 20 and 40), and several paleopolyploid genera as *Habenaria* with 2n = 42 (Monocots) and *Mikania* with 2n = 36 (Eudicots-Asterids-Campanulids). An unusual conspicuous high frequency of 2n = 36 for the whole sample, corresponds mainly to genera *Manihot* Mill., 1754 (Euphorbiaceae), Tibouchina Aubl., 1775 (Melastomataceae) and *Mikania* Willd., 1803 (Asteraceae). Intraspecific chromosome number variation occurs in 14.3% of species, being polyploidy the cause in 88% of them. At the genus level, most high chromosome numbers are multiples of one of the lowest modal numbers of the genus, i.e. base numbers (database website). The even/odd ratio of n is higher for the groups with "*n* greater than the mode" in almost every clade, as expected, except for Asterids-Campanulids (ratio 1.5), while Asterids-Lamiids shows the greatest bias towards even numbers (ratio 6.5) in karyotypes with n greater than the mode (Table 2).

Only 88 C-values were found for the taxa of Cerrado, most of them ranging from 2C = ca. 1.5 pg to ca. 3.5 pg with a median value of 2.38 pg (Fig. 4). That corresponds to 1.25% of the addressed species list. For a sample of known TCL (or TCA) and C-value, a correlation analysis, although significant, resulted in low r^2 values (weak to moderate). Applying a linear model to predict 2C-value based on TCL, resulted in prediction intervals of ca. 2pg. TCL and TCA values together with some predicted 2C- values are shown in Suppl. material 5, 6 and the database website.
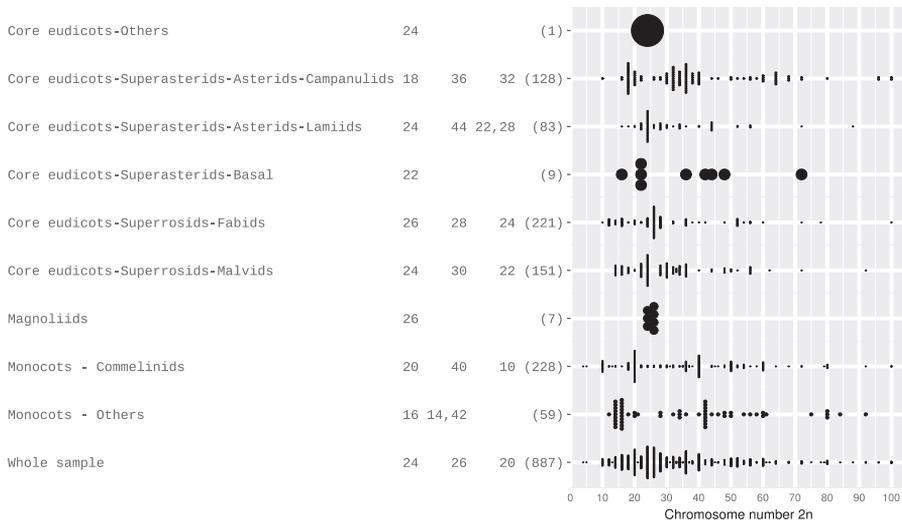

Figure 3. Dot-plot of observed and calculated 2n numbers for Cerrado plant species considering APG major clades. The three most common numbers (frequency ties separated by comma) and sample size in parentheses are indicated. Dot size varies depending on the maximum frequency of each group.
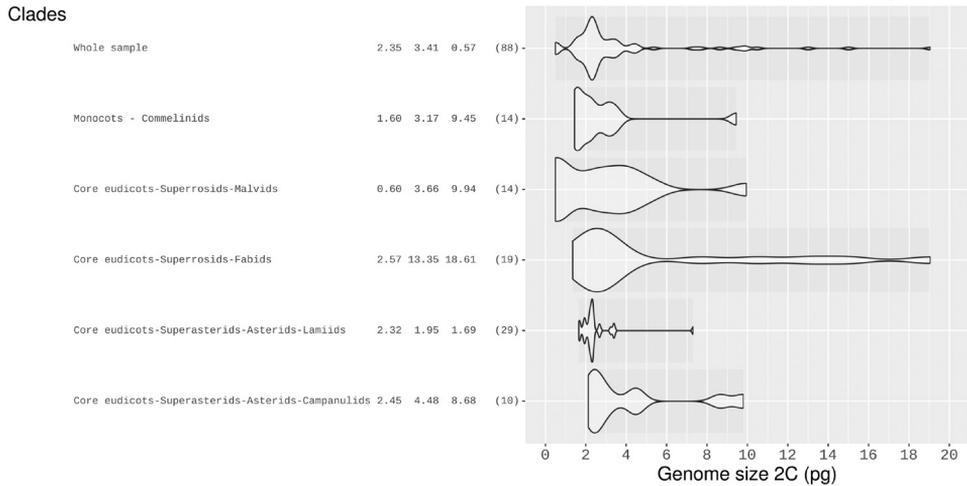
Figure 4. Violin-plots for genome sizes in pg for Cerrado angiosperms. Y-axis contains clades, three highest peaks of the Gaussian kernel density and sample size. Bandwidth of the density estimator customizable in http://cyto.shinyapps.io/cerrado

Table 1. Websites used for the search of literature.

Name	URL
Chromosome Counts Database*	ccdb.tau.ac.il
Plant DNA C-values Database*	data.kew.org/cvalues
International Organization of Plant Biosystematists	www.iopb.org
Scholar Google	scholar.google.com
Scopus	www.scopus.com
Biodiversity Heritage Library	biodiversitylibrary.org
JSTOR	www.jstor.org
Real Jardín Botánico CSIC	bibdigital.rjb.csic.es/spa/index.php
Naturalis Biodiversity Center	www.repository.naturalis.nl
Botanicus Digital Library	www.botanicus.org
Smithsonian Contributions to Botany	repository.si.edu/handle/10088/6943
Crossref	www.crossref.org

*citation in text, data taken from available primary sources.

Table 2. Even/odd ratio for haploid (n) numbers considering mode of n.

Clade or group	n even/odd ratio	
	Greater	Lesser
	than n mode	
Monocots - Commelinids	3.47	0.40
Other monocots	4.00	0.12
Core eudicots-Superasterids-Asterids-Lamiids	6.50	1.00
Core eudicots-Superasterids-Asterids-Campanulids	1.50	1.57
Core eudicots-Superrosids-Malvids	3.22	0.50
Core eudicots- Superrosids-Fabids	3.40	0.56

Discussion

There is strong interest for genomics and genetics of Cerrado species (Souza et al. 2016). Knowledge of basic data as chromosome number, ploidy level and genome size is critical for the selection of model species (Kelly et al. 2012). Though recent cytogenetic studies tend to incorporate phylogenetic frameworks and correlate cytogenetic with ecological variables, this kind of studies are scarce for Cerrado (Glick and Mayrose 2014; Suda et al. 2015; Silveira et al. 2016). As noted by Guerra (1990), previous to the 1990 decade, cytogenetic studies of authors from U.S.A. prevailed. After that period, mitotic analysis increased compared to meiotic studies, which relate to the establishment of cytogenetic laboratories allowing the treatment with antimitotics of root meristems. Still, the amount of chromosome number data for species occurring in Cerrado is lower than the world total, 16.1 vs. 20% (Rice et al. 2015) and several "orphan" genera were identified, such as *Stachytarpheta* (Verbenaceae), *Microlicia* (Melastomataceae) and *Leiothrix* (Eriocaulaceae). Unlike other techniques, C-banding and silver staining studies did not increase in recent decades, which might be related to low repetitiveness of methods (Guerra 1990). Despite recent efforts, the percentage of known Cerrado C-values is still lower than the 2.1% for all angiosperms (Garcia et al. 2014a). Data of TCA, and TCL for a larger number of species were intended to be used as proxy for genome size after an analysis that resulted in weak to moderate significant correlation with 2C-value in pg. Though they can give a rough estimate of the magnitude of the genome, the prediction interval is high considering most genomes fall within a 2 pg range.

Discontinuities in chromosome numbers seen as multimodal distributions indicate that higher numbers are polyploids generated by genome duplication. A polyploid origin for most high n numbers is also evidenced by the high (> 1) even/odd ratios of *n* (Rice et al. 2015). In contrast, low even/odd value (for *n* greater than the mode) is rare and suggests increased dysploidy and/or low frequency of polyploidy, compared with groups that follow the opposite general trend.

Conclusion

The Database of Cytogenetics of Cerrado plants and its website is presented, making an important step in facilitating access to most known cytogenetic attributes (http://cyto.shinyapps.io/cerrado/). The amount of chromosome number data for Cerrado is lower than the world total. The chromosome number, is still lacking in several rich genera and families, and therefore, they can be considered "orphan". This complete lack of information also happens for 70% of the single-species genera addressed. Analysis of chromosome numbers at several taxonomical levels revealed a straightforward relationship between polyploidy and high chromosome numbers. Regarding other techniques, like heterochromatin staining, molecular cytogenetics and C-value estimation, they have been applied to a very small percentage of species, however, those studies have

been steadily increasing since 2000. Though cytogenetic data are the basis for some evolutionary and ecological studies, there is a lack of those kind of interdisciplinary studies for Cerrado.

Acknowledgements

References

Alfons A (2016) robustHD: Robust Methods for High-Dimensional Data. https://cran.r-project.org/package=robustHD [accessed 1. June 2016]

APG (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181(1): 1–20. https://doi.org/10.1111/boj.12385

Attali D (2016) shinyjs: Perform Common JavaScript Operations in Shiny Apps using Plain R Code. https://cran.r-project.org/package=shinyjs [accessed 1. June 2016]

Barros e Silva AE, Guerra M (2010) The meaning of DAPI bands observed after C-banding and FISH procedures. Biotechnic & Histochemistry 85(2): 115–125. https://doi.org/10.3109/10520290903149596

Carvalho GH, Cianciaruso MV, Batalha MA (2010) Plantminer: A web tool for checking and gathering plant species taxonomic information. Environmental Modelling & Software 25(6): 815–816. https://doi.org/10.1016/j.envsoft.2009.11.014

Chang W (2015) shinydashboard: Create Dashboards with "Shiny." https://cran.r-project.org/package=shinydashboard [accessed 1. June 2016]

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2016) shiny: Web Application Framework for R. https://cran.r-project.org/package=shiny [accessed 1. June 2016]

Doležel J, Greilhuber J (2010) Nuclear genome size: Are we getting closer? Cytometry Part A 77(7): 635–642. https://doi.org/10.1002/cyto.a.20915

Dowle M, Srinivasan A, Short T [with contributions from R Saporta SL, Antonyan E] (2015) data.table: Extension of Data.frame. https://cran.r-project.org/package=data.table [accessed 1. June 2016]

Ferreira ME, dos Anjos AF, Ferreira LG, Bustamante M, Fernandes GW, Machado RB (2016) Cerrado O fim da história ou uma nova história? Ciênciahoje 56(334): 24–29.

Forzza RC, Baumgratz JFA, Bicudo CEM, Canhos DAL, Carvalho AA, Coelho MAN, Costa AF, Costa DP, Hopkins MG, Leitman PM, Lohmann LG, Lughadha EN, Maia LC, Martinelli G, Menezes M, Morim MP, Peixoto AL, Pirani JR, Prado J, Queiroz LP, Souza S, Souza VC, Stehmann JR, Sylvestre LS, Walter BMT, Zappi DC (2012) New Brazilian Floristic List Highlights Conservation Challenges. BioScience 62(1): 39–45. https://doi.org/10.1525/bio.2012.62.1.8

Garcia S, Gálvez F, Gras A, Kovařík A, Garnatje T (2014b) Plant rDNA database: update and new features. Database: the journal of biological databases and curation 2014 bau063. https://doi.org/10.1093/database/bau063

Garcia S, Leitch IJ, Anadon-Rosell A, Canela MÁ, Gálvez F, Garnatje T, Gras A, Hidalgo O, Johnston E, Mas de Xaxars G, Pellicer J, Siljak-Yakovlev S, Vallès J, Vitales D, Bennett MD (2014a) Recent updates and developments to plant genome size databases. Nucleic Acids Research 42(Database issue): 1159–1166. https://doi.org/10.1093/nar/gkt1195

Glick L, Mayrose I (2014) ChromEvol: Assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. Molecular Biology and Evolution 31(7): 1914–1922. https://doi.org/10.1093/molbev/msu122

Guerra M (1986) Reviewing the chromosome nomenclature of Levan et al. Brazilian Journal of Genetics 9(4): 741–743.

Guerra M (1990) A situação da citotaxonomia de angiospermas nos trópicos e, em particular, no Brasil. Acta Botanica Brasilica 4(2): 75–86. https://doi.org/10.1590/S0102-33061990000300009

Guerra M (2000) Patterns of heterochromatin distribution in plant chromosomes. Genetics and Molecular Biology 23(4): 1029–1041. https://doi.org/10.1590/S1415-47572000000400049

Kelly LJ, Leitch AR, Fay MF, Renny-Byfield S, Pellicer J, Macas JJ, Leitch IJ (2012) Why size really matters when sequencing plant genomes. Plant Ecology & Diversity 5(4): 415–425. https://doi.org/10.1080/17550874.2012.716868

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. Nature 403(6772): 853–858. https://doi.org/10.1038/35002501

Ooms J (2016) mongolite: Fast and Simple MongoDB Client for R. https://cran.r-project.org/package=mongolite

Otto SP, Whitton J (2000) Polyploid Incidence and Evolution. Annual Review of Genetics 34(1): 401–437. https://doi.org/10.1146/annurev.genet.34.1.401

Peruzzi L, Bedini G (2014) Online resources for chromosome number databases. Caryologia 67(4): 292–295. https://doi.org/10.1080/0144235X.2014.974358

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

REFLORA (2016) Flora do Brasil 2020. http://floradobrasil.jbrj.gov.br/ [accessed 22. April 2016]

Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I (2015) The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. New Phytologist 206(1): 19–26. https://doi.org/10.1111/nph.13191

Roa F, Guerra M (2012) Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. BMC Evolutionary Biology 12 225. https://doi.org/10.1186/1471-2148-12-225

Roa F, Guerra M (2015) Non-Random Distribution of 5S rDNA Sites and Its Association with 45S rDNA in Plant Chromosomes. Cytogenetic and Genome Research 146(3): 243–249. https://doi.org/10.1159/000440930

Sano SM, Pedrosa de Almeida S, Ribeiro J (2008) Cerrado: ecologia e flora. Embrapa Informação Tecnológica: Planaltina, DF: Embrapa Cerrados.

Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nature Methods 9(7): 671–675. https://doi.org/10.1038/nmeth.2089

Schwarzacher T, Ambros P, Schweizer D (1980) Application of Giemsa banding to orchid karyotype analysis. Plant Systematics and Evolution 134(3–4): 293–297. https://doi.org/10.1007/BF00986805

Silveira RM, Machado RM, Forni-Martins ER, Verola CF, Costa IR (2016) Environmental variations drive polyploid evolution in neotropical Eugenia species (Myrtaceae). Genet Mol Res. https://doi.org/10.4238/gmr15048842

Souza UJB de, Telles MP de C, Diniz-Filho JAF (2016) Tendências da literatura científica sobre genética de populações de plantas do Cerrado. Hoehnea 43(3): 461–477. https://doi.org/10.1590/2236-8906-06/2016

Spera SA, Galford GL, Coe MT, Macedo MN, Mustard JF (2016) Land-use change affects water recycling in Brazil's last agricultural frontier. Global Change Biology 22(10): 3405–3413. https://doi.org/10.1111/gcb.13298

Suda J, Meyerson LA, Leitch IJ, Pyšek P (2015) The hidden side of plant invasions: The role of genome size. New Phytologist 205(3): 994–1007. https://doi.org/10.1111/nph.13107

Vieira R, Queiroz Á, Morais L, Barão A, Mello-Sampayo T, Viegas W (1990) Genetic control of 1R nucleolus organizer region expression in the presence of wheat genomes. Genome 33(5): 713–718. https://doi.org/10.1139/g90-107

Warnes GR, Bolker B, Lumley T (2015) gtools: Various R Programming Tools. https://cran.r-project.org/package=gtools [accessed 1. June 2016]

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

Wickham H (2015a) plyr: Tools for Splitting, Applying and Combining Data. https://cran.r-project.org/package=plyr [accessed 1. June 2016]

Wickham H (2015b) stringr: Simple, Consistent Wrappers for Common String Operations. https://cran.r-project.org/package=stringr [accessed 1. June 2016]

Wickham H (2016) scales: Scale Functions for Visualization. https://cran.r-project.org/package=scales [accessed 1. June 2016]

Wickham H, Francois R (2016) dplyr: A Grammar of Data Manipulation. https://cran.r-project.org/package=dplyr [accessed 1. June 2016]

Xie Y (2015) DT: A Wrapper of the JavaScript Library "DataTables." https://cran.r-project.org/package=DT [accessed 1. June 2016]

Supplementary material 1

Summary statistics for genera
Authors: Fernando Roa, Mariana Pires de Campos Telles
Data type: cytogenetic attributes
Explanation note: more details in http://cyto.shinyapps.io/cerrado

Supplementary material 2

Summary statistics for families
Authors: Fernando Roa, Mariana Pires de Campos Telles
Data type: cytogenetic attributes
Explanation note: more details in http://cyto.shinyapps.io/cerrado

Supplementary material 3

Summary statistics for major clades
Authors: Fernando Roa, Mariana Pires de Campos Telles
Data type: cytogenetic attributes
Explanation note: more details in http://cyto.shinyapps.io/cerrado

Supplementary material 4

Authors with more publications of cytogenetic data of *Cerrado* plants per decade
Authors: Fernando Roa, Mariana Pires de Campos Telles
Data type: affiliation and taxa – categorical data table
Explanation note: more details in http://cyto.shinyapps.io/cerrado
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Supplementary material 5

Figure S1. Total chromosome length for *Cerrado angiosperms*
Authors: Fernando Roa, Mariana Pires de Campos Telles
Data type: Histogram
Explanation note: See text for details.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Supplementary material 6

Figure S2. Total chromosome area for *Cerrado angiosperms*
Authors: Fernando Roa, Mariana Pires de Campos Telles
Data type: Histogram
Explanation note: See text for details.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.