BrassiBase: Introduction to a Novel Knowledge Database on Brassicaceae Evolution

Markus Kiefer[1], Roswitha Schmickl[1,7], Dmitry A. German[1], Terezie Mandáková[2], Martin A. Lysak[2], Ihsan A. Al-Shehbaz[3], Andreas Franzke[4], Klaus Mummenhoff[5], Alexandros Stamatakis[6] and Marcus A. Koch[1,6,*]

[1]Department of Biodiversity and Plant Systematics, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, D-69120 Heidelberg, Germany
[2]CEITEC – Central European Institute of Technology, Masaryk University, CZ-625 00 Brno, Czech Republic
[3]Missouri Botanical Garden, St. Louis, MO 63166-0299, USA
[4]Heidelberg Botanic Garden, Centre for Organismal Studies (COS) Heidelberg, Heidelberg University, Im Neuenheimer Feld 340, D-69120 Heidelberg, Germany
[5]Botany Section, Department of Biology/Chemistry, University of Osnabrück, D-49076 Osnabrück, Germany
[6]Heidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany
[7]Present address: Institute of Botany, Academy of Sciences of the Czech Republic, CZ-25243 Průhonice, Czech Republic.
*Corresponding author: E-mail, marcus.koch@cos.uni-heidelberg.de; Fax, +49-6221-54-5508.
(Received August 20, 2013; Accepted October 23, 2013)

The Brassicaceae family (mustards or crucifers) includes *Arabidopsis thaliana* as one of the most important model species in plant biology and a number of important crop plants such as the various *Brassica* species (e.g. cabbage, canola and mustard). Moreover, the family comprises an increasing number of species that serve as study systems in many fields of plant science and evolutionary research. However, the systematics and taxonomy of the family are very complex and access to scientifically valuable and reliable information linked to species and genus names and its interpretation are often difficult. BrassiBase is a continuously developing and growing knowledge database (http://brassibase.cos.uni-heidelberg.de) that aims at providing direct access to many different types of information ranging from taxonomy and systematics to phylo- and cytogenetics. Providing critically revised key information, the database intends to optimize comparative evolutionary research in this family and supports the introduction of the Brassicaceae as the model family for evolutionary biology and plant sciences. Some features that should help to accomplish these goals within a comprehensive taxonomic framework have now been implemented in the new version 1.1.9. A 'Phylogenetic Placement Tool' should help to identify critical accessions and germplasm and provide a first visualization of phylogenetic relationships. The 'Cytogenetics Tool' provides in-depth information on genome sizes, chromosome numbers and polyploidy, and sets this information into a Brassicaceae-wide context.

Keywords: BrassiBase • Brassicaceae • Evolution • Knowledge database.

Abbreviations: ITS, internal transcribed spacer 1 and 2 of nuclear encoded rRNA.

Introduction

Evolutionary research is driven by comparative approaches in the widest sense. With the choice of *Arabidopsis thaliana* (L.) Heynh. as the first model organism of flowering plants, and having its genome sequenced more than a decade ago (Arabidopsis Genome Initiative 2000), a first fixed point, with which information from other plant organisms could be compared, has been set. Since then many other organisms (e.g. rice from the monocots) have been well established as model systems, allowing in-depth comparative analyses (Goff et al. 2002, Sakai et al. 2013). Prior to the advent of molecular techniques in evolutionary research, studies within the Brassicaceae family mainly focused on the characterization and breeding of agronomically important species such as cabbage or rapeseed. The same is true for crucifer systematics and taxonomy (Koch et al. 2003b, Koch and Al-Shehbaz 2009). As a consequence, past taxonomy and systematics accumulated more or less isolated information instead of providing comprehensive frameworks, from either a taxonomic or an evolutionary perspective. However, a broad systematic and taxonomic information pool is the primary 'access key' to the biological data indispensable for any kind of comparative research. This has already been outlined in great detail and precisely phrased by the title of the review by Paterson et al. (2010): 'Names are key to the big new biology'. Meanwhile there can be no doubt that taxonomic databases are vital for research

because they are the most readily available sources for adequate biological information (e.g. The Plant List' http://www.theplantlist.org/) and thus play the key role in one of the earliest and most important steps in any research project's workflow: choosing the right organisms to answer one's research questions. This is especially true for the Brassicaceae, a large plant family from which numerous model species (reviewed in Franzke et al. 2011, Koch and German 2013) do emerge, and which is almost notorious for complex phylogenetic relationships.

The mustard family belongs to the order Brassicales (core eudicots, Rosids) with three families grouped into the 'core Brassicales' (e.g. Beilstein et al. 2010): Brassicaceae, Cleomaceae and Capparaceae (4,440 species in total), and an additional 14 plant families with much lower total species numbers (245 species in total; Stevens 2001). Systematics, taxonomy and evolutionary history of the Brassicaceae have long been controversial. Boundaries between genera were often poorly delimited and attempts to group species diversity into higher order entities and combining genera into so-called tribes resulted in various artificial concepts. These difficulties resulted in a lack of agreement on the number and boundaries of phylogenetically defined tribes and genera and gave rise to several contradictory classification systems which have been proposed in the past (for a review, refer to Warwick et al. 2010). As an important consequence of this taxonomic history there is a lack of comprehensive and easily accessible information about generic and tribal placement of species. Meanwhile, in the most up to date overview, also used in BrassiBase, the mustard family comprises 3,660 species classified within 321 genera (Al-Shehbaz 2012), which, in turn, have been assigned to 49 monophyletic tribes (**Fig. 1**). Only 20 genera are not yet part of a tribe (Al-Shehbaz 2012), in most cases because plant material suitable for molecular systematic approaches is not readily available.

With BrassiBase, a consortium of scientists from various institutions aims to establish a knowledge database to provide direct access to reliable information about the Brassicaceae (Koch et al. 2012) and also to provide a complementary tool to other databases either focusing on Brassicaceae (Osaka et al. 2013, Shen et al. 2013) or aiming to contribute with data from many plant families (Cooper et al. 2013). BrassiBase is a continuously developing online-accessible knowledge database of cross-referenced information and resources on Brassicaceae taxonomy, systematics and evolution in the widest sense curated by a team of editors with different specialized expertise. The knowledge system is based on an accurate enumeration of all species, genera and tribes. The backbone of the database is a continuously updated taxonomic framework. This checklist will provide the basis for further groupings. The general concept and the underlying taxonomic framework have been presented in the first release version 1.1, in June 2012, and has been intensively discussed among taxonomists (Koch et al. 2012). With the second version 1.1.9. (released in August 2013) we present a

fully functional knowledge database. BrassiBase (http://brassibase.cos.uni-heidelberg.de) now includes cytogenetic data and a DNA sequence-based identification tool for genera and species. Additionally, it provides information about well-documented germplasm and other biological material collections at Heidelberg Botanical Gardens and Herbarium (HEID) representing the vast majority of tribes. Many Brassicaceae species are remarkable study objects or model systems, but living material or seeds to grow up plants are rarely available. However, biological material, either sampled in the wild or stored in collections, has often been taxonomically misidentified, and only very rarely has the material been sufficiently characterized and documented, e.g. as a herbarium voucher. Consequently, a DNA-based identification tool for genera and species based on ITS (internal transcribed spacer) DNA sequences is implemented, relying on the most recent, tribal and family-wide phylogenetic concepts (e.g. Beilstein et al. 2010, Couvreur et al. 2010, Franzke et al. 2011, Al-Shehbaz 2012). Finally, *BrassiBase* includes a comprehensive documentation of the taxonomy and systematics of the entire family by covering the relevant taxonomic, systematic and phylogenetic literature.

Database Implementation

The knowledge base system was developed using freely available software on standard server hardware. Its platform is a standard Linux installation containing the most recent stable versions available of MySQL 5, Apache 2 and PHP 5. For the Ajax- and Javascript-driven user interface and visualization of the data, the following frameworks and libraries were used: JQuery/JQuery-UI, DataTables, JIT InfoVis Toolkit and D3.js in the latest available versions. The functionality of the 'Phylogenetics Tool' was implemented using RAxML (Stamatakis 2006, Berger et al. 2011), MAFFT (Katoh and Standley 2013), the Newick Utilities (Junier et al. 2010), EMBOSS (Rice et al. 2000) and several Linux shell scripts. For the creation of reference alignments of ITS DNA sequences, MUSCLE (Edgar 2004) was used. To provide a high-resolution image viewer we installed the IIP Image server, an FCGI web server module and the Flash-based IIPZoom viewer. Images were converted to pyramidal TIFF format using the Image magick package. As an additional viewer for JPEG images we used the javascript tool prettyPhoto. More detailed information on the software used is provided online (see BrassiBase 'Technicalities' box on the main page). Graphical data visualization is done with dynamically created SVG files.

This new release now presents the taxonomic information in a more user-friendly 'Taxonomy Tool'. Additional tools providing substantial data have been incorporated, namely a 'Phylogenetic Placement Tool' and a 'Cytogenetics Tool'. Furthermore, the material and germplasm collection tool is now fully functional and accessible to the scientific community.
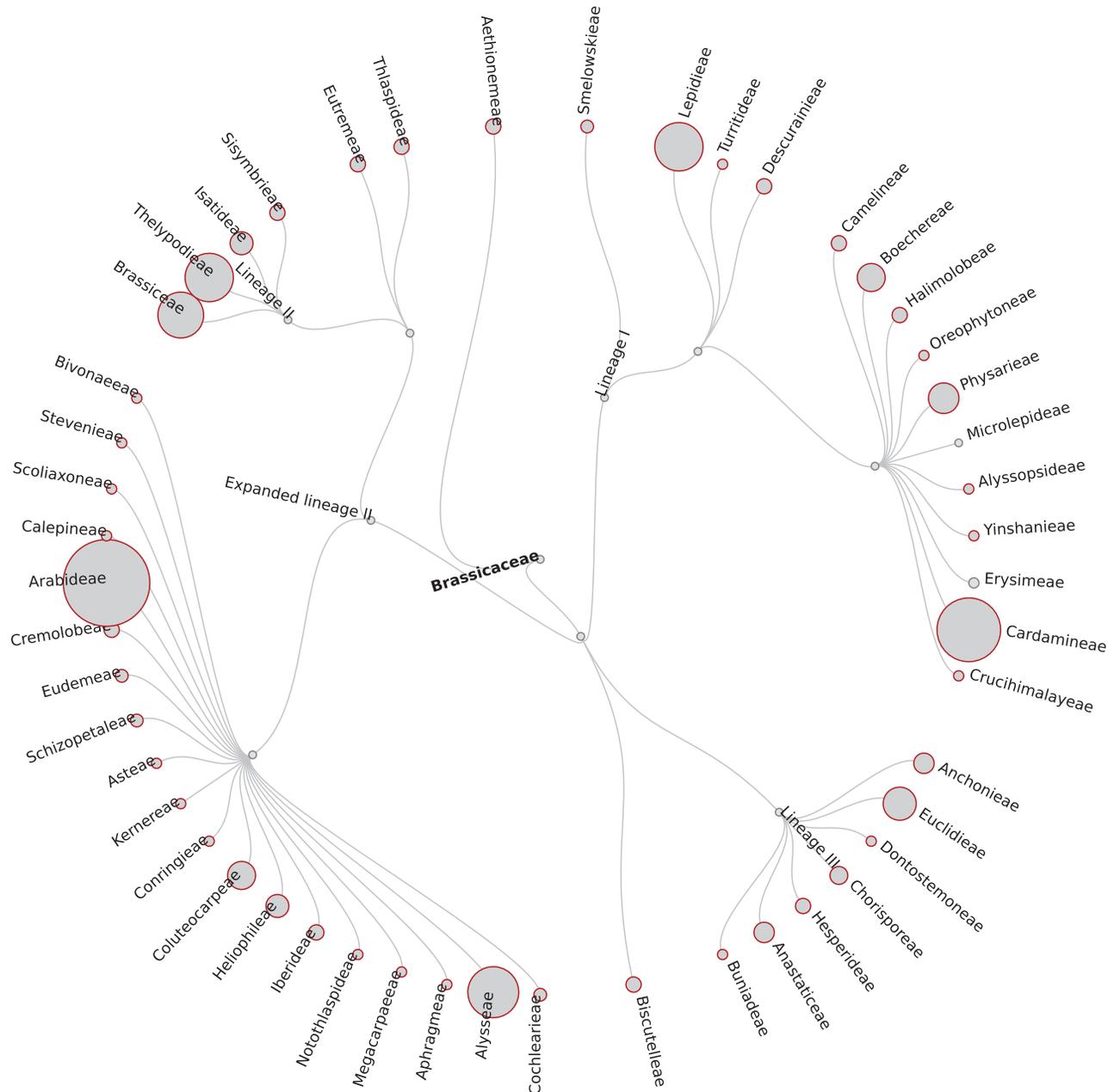
Fig. 1 A representation of the Brassicaceae tribe phylogeny as a circular diagram (Phylogenetics Tool information site). The size of the leaf nodes corresponds to the number of species attributed to a tribe.

Database Contents and Functions

The concept of BrassiBase is based on long-term research activities of a collaborative scientific community. Communication and interactivity do play a crucial role in incorporating the most recent knowledge into BrassiBase. This means that the content of BrassiBase as well as its means of data presentation will be updated continuously. Data sets are continuously complemented, because direct access to this information is crucial for effective research (Koch et al. 2012, Koch and German 2013). The information is cross-referenced, and a growing team of experts ('Scientific Editors') are contributing with their skills to guarantee high quality and further development of the database. The information will contribute greatly to implementing individual evolutionary hypotheses into a broader evolutionary context. Furthermore, the information provided by BrassiBase is intended to assist scientists in defining an evolutionary context prior to starting new research projects. The framework of the database is a reliable and accurate taxonomic information system, linked to detailed descriptions of genera (species descriptions and checklists are in preparation) and providing information about taxonomic synonyms.

Table 1 A summary of currently incorporated data and its coverage of respective taxonomic ranks within the Brassicaceae (total number and % of total number)

Taxonomic level	49 tribes	320 genera	3,660 species
Chromosome counts[a]	48 (98%)	237 (75%)	1,643 (45%)
Genome sizes[b]	41 (84%)	129 (41%)	318 (9%)
ITS sequence data[c]	49 (100%)	297 (93%)	1,689 (47%)
Reference trees	40 (82%)	285 (90%)	1,673 (46%)
Taxonomic descriptions	49 (100%)	72 (23%)	714 (20%)
Literature	9,098 citations in total		

[a]Chromosome counts are currently available from 9,035 data points.
[b]Genome size data actually represent 477 data points.
[c]A total of 1,715 DNA sequences of internal transcribed spacer 1 and 2 (ITS1 and 2) of nuclear-encoded rRNA (out of a total of 5,719 high quality sequences retained after evaluation) are currently incorporated in reference alignments.

The lack of readily available information on the latter was a major problem and source of errors for comparative evolutionary research in Brassicaceae in the past. (e.g. Arabidopsis: for a review and comments, refer to Koch et al. 2008). Using BrassiBase allows researchers easily to trace back the past nomenclatural history of genera.

BrassiBase currently contains information about all 49 tribes presently accepted as monophyletic, 331 accepted genera (including 20 genera not yet assigned to any tribe) plus 349 corresponding synonyms, 72 detailed genus descriptions which can be downloaded as PDF documents, 9,098 taxonomic literature citations, 580 images of living plants (6 megapixel), 1,266 high-resolution images of herbarium vouchers (70 megapixel), 49 ITS alignments consisting of a total of 1,715 DNA sequences, 40 corresponding maximum-likelihood trees, 9,512 cytogenetic data sets and 426 data sets of collected specimens (**Table 1**). Additional data sets are constantly added, and additional data categories (e.g. morphological traits) are currently under development.

Names as key to information

The 'Taxonomy Tool' allows a search to be made for scientific names of tribes, genera and species by simply using an input field with the capability of suggesting valid entries on the tribe and genus level. Since this field constitutes the central control element of the system, it is also available at all times in the right-hand column. If a tribe's name is given as a search term, information is collected about all member genera. The results are classified into the following five categories represented as 'tabs'.

The 'Names' tab provides information about taxa relevant to the query entered: taxonomic status and authorities, tribal assignment and, if applicable, geographic distribution, citation of the protolog and information about the type of the relevant taxon. Links to the reference ITS sequence alignment and a corresponding phylogenetic tree of the requested tribe(s) are also displayed here. In the right-hand column, direct search links for names of genera to the taxonomic databases IPNI and Tropicos are provided.

The 'Tree' tab shows a phylogenetic tree of the tribe requested in a family-wide context, and the tribal subtree consisting of representative species for the genera. Below, a dynamic representation of the phylogeny is also given, providing an overview of the relationships among tribes.

The 'Images' tab displays a list of available images relevant to the search terms. The images provided here are scaled to approximately $1,600 \times 2,400$ pixels, a medium quality resolution allowing identification of many relevant morphological characteristics. Photographs of living Brassicaceae specimens were mainly taken in the Heidelberg Botanic Garden while cultivating for seed increase from 2010 to 2013.

The 'Literature' tab provides a collection of taxonomic literature citations relevant to the search terms, compiled from reference collections of several contributors. The corresponding literature is neither stored on nor available from the project website (due to copyright and availability reasons) and, consequently, it is not full-text indexed. So no guarantee of the completeness in the results can be given, although the keyword lists used to link citations to relevant results are continuously improved. In any case, many of these citations will be difficult to find by other means.

The 'Descriptions' tab provides a list of downloadable PDF documents containing detailed generic descriptions either specifically compiled for a requested genus or at least mentioning it. Descriptions of missing taxa are continuously complemented.

Phylogenetic Placement

The 'Phylogenetics Tool' serves two purposes. On the one hand, it provides a browsable schematic overview (see display options on the lower part of the web page) of the pylogenetic relationships of Brassicaceae tribes as recently presented by Franzke et al. (2011) and in **fig. 1** of Koch et al. (2012). It is important to notice that ongoing research will provide a robust phylogenetic backbone for the whole family within the next few years, but at present there is no better information than presented herein with the above-mentioned schematic overview. As a consequence, we do not show any other detailed family-wide phylogenetic hypothesis.

On the other hand, the 'Phylogenetics Tool' allows the user to test their own ITS DNA sequences against underlying reference alignments. The ITS region of about 650 bp is easily accessible via simple PCR and using highly conserved primers. Even if the quantity and quality of the isolated DNA are critical (a very small amount of leaf material, herbarium material, DNA isolated from single seeds, etc.), this molecular marker works sufficiently in most cases. If needed, the region can also be amplified as separate fragments (ITS1 and ITS2) to be subjected to further DNA sequencing. In addition, the PCR product can be easily cloned prior to sequencing. Many of these advantages, also including its high phylogenetic discriminative power, led to

the refinement of the plant DNA barcodes, now also suggesting the inclusion of ITS in the standard plant DNA marker set (Hollingworth 2011). Because of the special type of molecular evolution of ITS (extensive concerted evolution is likely because of the very high copy number; e.g. Koch et al. 2003a), any phylogenetic hypothesis within a genus should be interpreted with caution, especially when reticulate evolution and allopolyploidization have played a major role in the evolution of the respective group of taxa.

All sequences used in generating the alignments for BrassiBase were taken from GenBank (Benson et al. 2013), and all sequences were carefully selected to represent all genera and isolated taxa within a given tribe. Sequences were selected only if a respective phylogenetic study has been published. These phylogenetic studies usually also included other molecular markers, morphology and biogeography, so that not only congruent phylogenetic placement but also misidentifications could be detected. Furthermore, we accepted only sequences with reliable documentations (source and origin, voucher) provided (either with the original GenBank annotation or with the respective phylogenetic publications). Alignments were initially created using MUSCLE (Edgar 2004) and extensively edited manually in subsequent steps. All alignments were checked for plausibility by phylogenetic analyses (maximum-likelihood methods) and compared with the relevant phylogenetic literature, which often takes other markers into account as well. A link to view the alignments can be found on the summary pages for each tribe (search for scientific names in the right-hand column or 'Taxonomy Tool'). The alignment provides the link to the GenBank accessions codes, and the whole alignment can be downloaded in different formats suitable for subsequent bioinformatics analyses (formats for download: FASTA, phylip, nexus, CLUSTAL and msf).

The button labeled 'Analyze own ITS sequence' gives access to a text field that can be used to paste an ITS sequence in FASTA format. This query sequence is first briefly checked for orientation and sufficient similarity to our ITS reference sequences using standard tools of the EMBOSS package (Rice et al. 2000). Subsequently the query sequence is evaluated against all available tribe alignments in turn, in each case starting with MAFFT (Katoh and Standley 2013) to add the new sequence to the respective pre-existing reference alignment, and then using the 'Evolutionary Placement Algorithm' mode of RAxML (compiled with the pthreads option to allow multithreading) to calculate a classification value for placement of the new sequence in the also pre-existing reference tree (Stamatakis 2006, Berger et al. 2011). The evolutionary model used is GTR + G. RAxML was adapted to the needs of the BrassiBase project by Alexandros Stamatakis. The changes in the new version of RAxML enable the use of a single, longer query sequence instead of short reads. The changes have been included in the most recent standard distribution of RAxML. Classification values for all tribes are then normalized to a maximum of 100 and compared to determine the best tribal placement. Results are displayed in the form of a bar chart and a phylogenetic tree. The

most likely position of the query sequence within the tribe phylogeny is highlighted. For registered users, result reports are stored and can be displayed at a later time. In our simulation analyses with known query sequences, successful placements were always phylogenetically correct, while unsuccessful tests quite obviously showed too little overall difference between all classification values. The inability of the analysis tool to assign a sequence reliably to the phylogenetically correct position in the Brassicaceae tree can be due to a lack of congruence with our ITS sequences so that MAFFT's (Katoh and Standley 2013) profile alignment algorithm cannot establish a new alignment of adequate quality including the query sequence. Another reason could be inconsistencies in our reference phylogenies or the inclusion of GenBank sequences originating from incorrectly determined plant specimens, perhaps not even belonging to the Brassicaceae at all. Ultimately the query sequence might be most similar to one of the few up to now unassigned genera (not linked to any tribe yet). Though we exercised the utmost care in the selection of our 1,715 ITS reference sequences as well as in the design of reference alignments and corresponding trees, errors cannot be completely ruled out. The approach of splitting up the Brassicaceae into separate tribe alignments was not chosen only for performance and manageability reasons. The most important reason was that a reliable and consistent family-wide alignment of ITS sequences in the Brassicaceae is not possible as outlined above. BrassiBase trees cannot replace a detailed phylogenetic analysis. The 'ITS Phylogenetics Tool' is intended only to provide the first hints for design of subsequent sophisticated phylogenetic analyses (e.g. ingroup–outgroup comparisons).

Cytogenetic Information

Cytogenetic information about 1,650 Brassicaceae species can be obtained using the 'Cytogenetics Tool'. Results include chromosome numbers, genome sizes and information about ploidy levels. The data are compiled from thorough literature research as well as experimental work carried out in Heidelberg. In 'default mode' a search can be conducted by using scientific names of tribes, genera or species as seach terms (**Fig. 2**); 'in advanced mode' it is possible to use cytogenetic data as search criteria, e.g. genome size (in pg) above a specified quantity or the presence of neopolyploidy (recent auto- and allopolyploids) or combinations thereof. Is has to be noted here that cytotype variations within a species do affect the resulting percentages of diploids vs. polyploids, and differing cytotypes (diploidy vs. polyploidy) have to be treated in this case as different taxonomic entities. Base chromosome numbers for the various tribes vary from four to 19 (e.g. Lysak et al. 2006, Lysak et al. 2009).

Results will be displayed in a freely sortable table layout (**Fig. 2**). The amount of data sets displayed can be changed and the results can be filtered for keywords to narrow the search further. Several of the data provided by this tool can also be viewed as family-wide summaries plotted onto the

Fig. 2 An example view of the BrassiBase user interface. Shown is the results page of a Cytogenetics search (query was tribe Arabideae) with a tabular output.

phylogeny of tribes (e.g. average genome size or percentage polyploid taxa) (**Fig. 3**).

Material Collection

The Brassicaceae material collection managed by the BrassiBase project aims to provide the scientific community with well-documented biological material of Brassicaceae specimens. This material includes seeds, flower buds and root tips stored in ethanol, silica-dried leaf tissue and herbarium vouchers. Most of the material originates from Brassicaceae seed increase cultivation at the Heidelberg Botanic Garden. Therefore, it is not possible to provide the material completely free of charge and a small fee has to be collected to cover the institution's expenses.
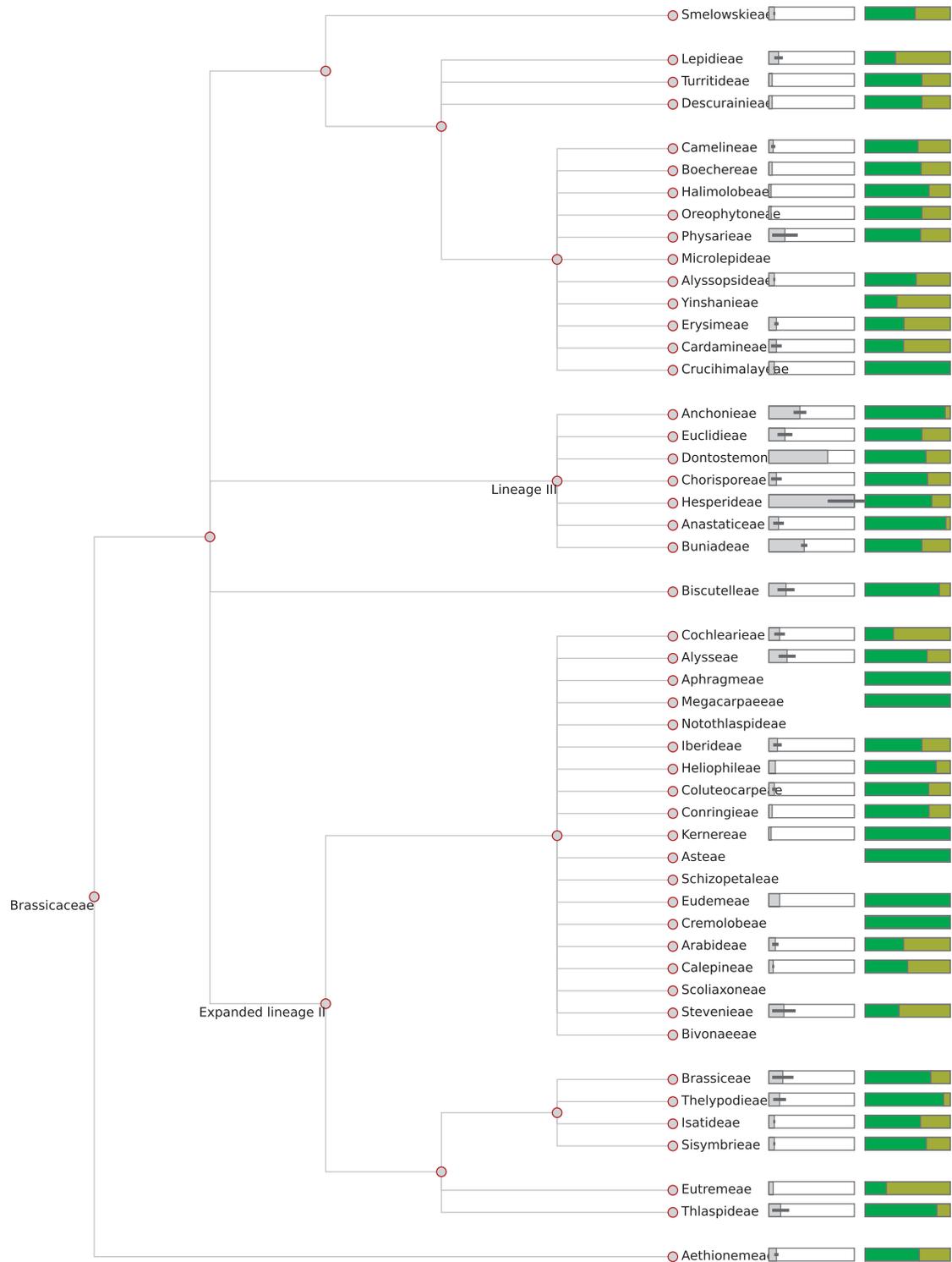
Fig. 3 An overview of cytogenetic data on the tribal level (Cytogenetics Tool information site): gray boxes represent the average genome size in a tribe as a fraction of the family's maximum (found in tribe Hesperideae); colored boxes show the fraction of polyploids in a tribe (in yellow, in contrast to diploids in green). The black bar indicates the respective standard deviation from the given mean value.

Within the BrassiBase user interface it is easy to search for available material by scientific name. Registered users can filter for availability of a certain type of material. The data are displayed in a convenient table layout, with sortable columns and keyword filtering. For most of the data sets, additional documentation such as geographic data can be displayed in the right-hand column. If available, preview images will provide links to high-resolution images (approximately 10,000 × 7,000

pixels) of corresponding herbarium vouchers. Vouchers can be viewed conveniently as tiled, pyramidal TIFF images, which will even be accessible over a low bandwidth connection with reasonable performance.

Registered Users

We offer a few advantages to registered users. First, registered users can store their results reports, and these can be re-displayed at a later time. For this service we do need an overview of the maximum number of accounts to guarantee the functionality and stability of the system. Secondly, registered users do have more possibilities to filter research material. Here we also require a full registration because all material is only provided in full agreement with CBD (convention biological diversity), and it is expected that there will be a number of orders of biological material that will leave the IPEN (International Plant Exchange Network). However, not only do most European Botanical Gardens and collections agree with the rules of CBD, but the DFG (German Research Foundation), who provide substantial financial resources for BrassiBase, also requires acceptance of these strict regulations. More detailed information can be found in the database (terms of use). Thirdly, registered users will be informed from time to time about updates and important achievements.

Conclusion and Future Directions

As mentioned earlier, BrassiBase is a continually progressing project, as new data are constantly being added. This refers not only to an expansion of the data we already have in our hands, but also to the addition of new data classes resulting in the development of new features. At the moment we are working on a large morphological trait data set and corresponding search tools that will, once implemented, be of tremendous help in determining Brassicaceae specimens. It is also planned to constrain ITS-based phylogenetic hypothesis with phylogenetic tree topologies based on whole-genome data from the plastids, mitochondria and nuclear genomes/transcriptomes. That will close the gap between a 'Phylogenetic Placement Tool' and high quality phylogenetic analysis.

As a long-term perspective, we aim to provide even more comprehensive taxonomic information by including more species-level data. Explicitly, we also invite researchers to incorporate their data into BrassiBase.

Funding

M.A.K. and K.M. and providing resources to develop the BrassiBase database system].

Disclosures

The authors have no conflicts of interest to declare.

References

Al-Shehbaz, I.A. (2012) A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon* 61: 931–954.

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* 107: 18724–18728.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. et al. (2013) GenBank. *Nucleic Acids Res.* 41: D36–D42.

Berger, S.A., Krompass, D. and Stamatakis, A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* 60: 291–302.

Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B. et al. (2013) The Plant Ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* 54: e1.

Couvreur, T.L.P., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A. and Mummenhoff, K. (2010) Molecular phylogenetics, temporal diversification and principles of evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* 27: 55–71.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32: 1792–1797.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A. and Mummenhoff, K. (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci.* 16: 108–116.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92–100.

Hollingworth, P.M. (2011) Refining the DNA barcode for land plants. *Proc. Natl Acad. Sci. USA* 108: 19451–19452.

Junier, T. and Zdobnov, E.M. (2010) The Newick Utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26: 1669–1670.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30: 772–780.

Koch, M.A. and Al-Shehbaz, I.A. (2009) Molecluar systematics and evolution of 'wild' crucifers (Brassicaceae or Cruciferae). *In* Biology and Breeding of Crucifers. Edited by Gupta, S. pp. 1–18. Taylor and Francis Group, London.

Koch, M.A., Dobes, C. and Mitchell-Olds, T. (2003a) Multiple hybrid formation in natural populations: concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in North American *Arabis divaricarpa* (Brassicaeae). *Mol. Biol. Evol.* 20: 63–73.

Koch, M.A. and German, D. (2013) Taxonomy and systematics are key to biological information: *Arabidopsis*, *Eutrema* (*Thellungiella*), *Noccaea* and *Schrenkiella* (Brassicaceae) as examples. *Front. Plant Sci.* 4: 267.

Koch, M.A., Kiefer, M., German, D., Al-Shehbaz, I.A., Franzke, A., Mummenhoff, K. et al. (2012) BrassiBase: tools and biological resources to study characters and traits in the Brassicaceae—version 1.1. *Taxon* 61: 1001–1009.

Koch, M.A., Mummenhoff, K. and Al-Shehbaz, I.A. (2003b) Molecular systematics, evolution, and population biology in the mustard family (Brassicaceae): a review of a decade of studies. *Ann. Miss. Bot. Gard.* 90: 151–171.

Koch, M.A., Wernisch, M. and Schmickl, R. (2008) *Arabidopsis thaliana*'s wild relatives: an updated overview on systematics, taxonomy and evolution. *Taxon* 57: 933–943.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. and Schubert, I. (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc. Natl Acad. Sci. USA* 103: 5224–5229.

Lysak, M.A., Koch, M.A., Leitch, I.J., Beaulieau, J.M. and Meister, A. (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol. Evol.* 26: 85–98.

Osaka, M., Matsudo, T., Sakazono, S., Masuko-Suzuki, H., Maeda, S., Sewaki, M. et al. (2013) Cell type-specific transcriptome of Brassicaceae stigmatic papilla cells from a combination of laser microdissection and RNA sequencing. *Plant Cell Physiol.* 54: 1894–1906.

Paterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L. and Remsen, D.P. (2010) Names are key to the big new biology. *Trends Evol. Ecol.* 25: 696–689.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16: 276–277.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y. et al. (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54: e6.

Shen, D., Sun, H., Huang, M., Zheng, Y., Li, X. and Fei, Z. (2013) RadishBase: a database for genomics and genetics of radish. *Plant Cell Physiol.* 54: e3.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.

Stevens, P.F. (2001) onwards. Angiosperm Phylogeny Website. Version 12, July 2012 [and more or less continuously updated since]. http://www.mobot.org/MOBOT/research/APweb/.

Warwick, S.I., Mummenhoff, K., Sauder, C., Koch, M.A. and Al-Shehbaz, I.A. (2010) Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. *Plant Syst. Evol.* 285: 209–232.