**Database tool**

**Boechera microsatellite website: an online portal for species identification and determination of hybrid parentage**

Fay-Wei Li¹,²,⁎,†, Catherine A. Rushworth¹,², James B. Beck³,⁴ and Michael D. Windham¹,⁎,†

¹Department of Biology, Duke University, Durham, NC 27708, USA, ²University and Jepson Herbaria and Department of Integrative Biology, University of California, Berkeley, CA 94720, USA, ³Department of Biological Sciences, Wichita State University, Wichita, KS 67260, USA and ⁴Botanical Research Institute of Texas, 1700 University Drive, Fort Worth, TX 76107, USA

⁎Corresponding author: Email: fayweili@gmail.com, mdw26@duke.edu

Present address: Fay-Wei Li, Boyce Thompson Institute, Ithaca, NY 14850, USA

†These authors contributed equally to this work

Citation details: Li,F.W., Rushworth,C.A., Beck,J.B., et al. Boechera microsatellite website: an online portal for species identification and determination of hybrid parentage. Database (2016) Vol. 2016: article ID baw169; doi:10.1093/database/baw169

Received 3 September 2016; Revised 10 December 2016; Accepted 13 December 2016

**Abstract**

**Boechera** (Brassicaceae) has many features to recommend it as a model genus for ecological and evolutionary research, including species richness, ecological diversity, experimental tractability and close phylogenetic proximity to **Arabidopsis**. However, efforts to realize the full potential of this model system have been thwarted by the frequent inability of researchers to identify their samples and place them in a broader evolutionary context. Here we present the **Boechera** Microsatellite Website (BMW), a portal that archives over 55,000 microsatellite allele calls from 4471 specimens (including 133 nomenclatural types). The portal includes analytical tools that utilize data from 15 microsatellite loci as a highly effective DNA barcoding system. The BMW facilitates the accurate identification of **Boechera** samples and the investigation of reticulate evolution among the ±83 sexual diploid taxa in the genus, thereby greatly enhancing **Boechera**'s potential as a model system.

**Database URL:** http://sites.biology.duke.edu/windhamlab/

**Introduction**

Among the close relatives of **Arabidopsis** Heynh., the genus **Boechera** Á. Löve & D. Löve (Brassicaceae) is notable for its species richness, ecological diversity and experimental tractability, features that have contributed to its emergence as a model genetic system for studies of...
evolutionary biology (1). However, efforts to fully leverage Boechera as a model system have been hampered by the taxonomic complexity of the genus and resultant misidentifications of materials used in various research studies. The taxonomic uncertainty in Boechera arises from a “perfect storm” of hybridization, polyploidy, and apomixis, all of which are rampant in the genus (2–7). Through these processes, the already substantial diversity of sexual diploid Boechera taxa (±83) has given rise to hundreds of additional taxa involving diverse combinations of at least 64 of these known sexual genomes (see 8–14).

In groups prone to such hybridization, determining which taxa represent the products of divergent evolution (i.e. sexual diploids) and which are the result of reticulate evolution is a prerequisite to developing a useable taxonomy (15–18). However, sexual diploids can be very difficult to detect morphologically in the apomictic hybrid milieu (14), and traditional molecular tools for species identification and hybrid diagnosis have proven of little use in Boechera due to widespread allele-sharing across species (19–21) and possible plastid capture (21). To address this problem, we have developed a 15-locus nuclear microsatellite ‘barcode’ that distinguishes all known sexual diploids and simultaneously provides a framework for determining the genomic makeup of hybrid individuals. These short genetic markers offer another advantage, in that they can be reliably amplified from herbarium specimens of Boechera collected over the past 130 years. Here we present a database of 56 353 microsatellite allele calls from 4471 specimens including 133 nomenclatural types. We also present novel tools for genotype-based species identification and determination of the genomic constitution of hybrid individuals.

Materials and methods

Microsatellite data

The database consists of microsatellite genotypes and basic locality data for 4471 Boechera specimens representing ca. 95% of all previously named taxa. Each population of every taxon included in the database is represented by voucher specimens deposited in one or more of the 36 herbaria supporting this effort. DNA extraction and multiplex microsatellite genotyping were performed using protocols outlined in Beck et al. (12). In addition to the 13-locus microsatellite set employed in that study, we incorporated two additional loci (BF19 and a3; Table 1). These loci are useful for sample identification but are excluded from heterozygosity calculations because these loci occasionally exhibit more than the expected number of alleles in plants of known ploidy. Extra alleles likely indicate amplification of multiple regions of the genome, and these loci are considered unreliable for heterozygosity estimates. The need for consistency in calling alleles is critical, and we have developed a set of locus-by-locus guidelines accessible through the Boechera Microsatellite Website (BMW) entry page by clicking on ‘About & Contact’.

Inferring reproductive modes

The two primary reproductive pathways in Boechera are sexuality (largely self-compatible and self-pollinating) and apomixis, the latter defined here as asexual seed production through either diplospory or apospory (25). The reproductive pathway of each accession in the database has been inferred based on a series of genetic and morphological assessments summarized in Figure 1. Foremost among these are analyses of meiotic chromosome pairing and segregation (14). Plants exhibiting normal pairing behavior resulting in the formation of spore tetrads are considered sexual; those with little or no pairing that produce malformed spores or functional diplospores in dyads are inferred to be apomictic. Including both published (summarized in 26) and unpublished chromosome studies, we have cytogenetic data for ~5% of all accessions in the database. Inferring the reproductive mode of the remaining accessions requires some additional observations.

The differences in chromosome pairing cited above lead to easily observed disparities in pollen morphology between sexuals and apomicts. Following Windham and Al-Rehab (8) and Beck et al. (12), accessions with a high percentage of well-formed, narrowly ellipsoid, tricolpate and symmetrical pollen (and no diplosporous pollen) are inferred to be sexual; those with a preponderance of malformed (highly variable) and/or ovoid-spheroid, multicolpate and asymmetric pollen are classified as apomictic. Approximately 60% of Boechera herbarium specimens have flowers with sufficient pollen to ascertain reproductive mode. Assessment of reproductive mode in remaining specimens involves looking for evidence of stabilized hybridity in the available microsatellite data. Specimens exhibiting allelic additivity (involving either distinct taxa or inbred lines) at most loci are inferred to be apomicts, with the exception of a small number of sexual allotetraploids in the dataset. Accessions in which nearly all alleles are associated with a single taxon or inbred line are considered sexual.

Inferring ploidy levels

Many previous studies have reported variation in ploidy between and within species of Boechera (see 23 and references therein). Voucher specimens for most published
Table 1. Information for the 15 microsatellite loci presented in the database

| Marker   | Primer sequences                      | Chr. no. | nt repeat | Number of alleles, N<sub>a</sub> | Original publication |
|----------|---------------------------------------|----------|-----------|----------------------------------|----------------------|
| ICE3     | GACTAATCATACCGAGTCGAGCAGAGG       | 5        | CT        | 62                               | (22)                 |
| ICE14    | TCAGGTTGCTCTCTGAGGTAGT              | 2        | GAT       | 17                               | (22)                 |
| a1       | AGATTGTTGACACCCGACC               | 2        | GAT       | 15                               | (23)                 |
| a3       | AGCCTTTTGCAATGGGAG               | 2        | AG, AT    | 45                               | (23)                 |
| b6       | GCCTATTCACAGCTGAGGAG             | 1        | CT, GT    | 62                               | (23)                 |
| c8       | TTCCGGGTATACCTCACCTAG            | 5        | CTT       | 58                               | (23)                 |
| c9       | GCTGATCTCGAATCACCCTTTG            | 4        | CT       | 56                               | (23)                 |
| BF3      | TTTTTAGACATGTGGCTGTTG             | 4        | GA       | 58                               | (24)                 |
| BF9      | AAACACATTCCCGGTCGGCTC             | 3        | GA       | 55                               | (24)                 |
| BF11     | TCCTCCATTGAGACGACAGAC           | 2        | GA       | 27                               | (24)                 |
| BF15     | CAGATCTTCCTTGCTGTG               | 5        | GA       | 48                               | (24)                 |
| BF18     | AACCTCCAAGATTTGCGCTTC           | 1        | CT       | 30                               | (24)                 |
| BF19     | ACCGGCTTTGCTGTTGTC              | -        | GA       | 63                               | (24)                 |
| BF20     | TTCCGGGAAAAATATGAGGAG           | 2        | CT       | 47                               | (24)                 |
| Bdru266  | TTTATTTGTGGCTGTTG                 | -        | AT       | 54                               | (24)                 |

Loci a3 and BF19 occasionally exhibit more than the expected number of alleles in plants of known ploidy and were therefore excluded from heterozygosity calculations. Null alleles were treated as missing data and are not included in the total number of alleles.

Figure 1. Summary of the steps involved in inferring reproductive mode in accessions of *Boechera*.
chromosome counts in the genus have been included in the database, as have the vouchers for several hundred unpublished counts (Windham et al., in prep.). These data reveal a very strong association between the chromosome number documented for each individual and the maximum number of alleles observed at 13 of the 15 microsatellite loci analyzed (excluding loci a3 and BF19; Table 1). Cytogenetically known diploids in Boechera range from completely homozygous to having two alleles at 12 of these 13 putatively single-copy loci, whereas documented apomictic triploid and apomictic tetraploid individuals exhibit at least two loci with three or four alleles, respectively. These observations provide a pathway for inferring the ploidy of accessions that have not been studied cytogenetically but have appropriate microsatellite data. These inferences are based on the maximum number of alleles at two or more loci, a conservative threshold adopted to prevent the inflation of inferred ploidy based on a single aberrant locus or allele call. For example, a specimen would be called triploid if two or more of the loci used to assess heterozygosity (Table 1) exhibited three alleles. Plants that are homozygous at all loci are, by default, diploid. Although this approach works well for distinguishing diploids from triploids and apomictic tetraploids, sexual tetraploids are less tractable. The database includes 56 sexual tetraploid accessions representing two taxa: holmgrenii × lemmonii and laevigata × stricta. These hybrid species exhibit disomic inheritance (fixed heterozygosity) and rarely show more than two alleles per locus, as is common in sexual allotetraploids derived from relatively divergent species (26). Thus they are largely indistinguishable from highly heterozygous (apomictic) diploids based on the maximum number of alleles per locus. However, sexual tetraploids are rarely encountered in Boechera (27, 28), and they are readily separated from apomictic diploids by other criteria, including pollen morphology and chromosome number (Fig. 1).

Taxonomic coverage

In the database as well as this article, known or inferred sexual diploid samples are designated by single epithets (e.g. ‘gracilipes’), most validly published in Boechera but with a few still pending. Inferred apomicts (diploid, triploid or tetraploid) and the few sexual tetraploids are identified by the appropriate number and combination of epithets, each separated by the hybrid symbol ‘×’ (e.g. ‘gracilipes × perennans × texana’). After removing 43 accessions that lack data at eight or more loci, the remaining 4428 accessions included 2594 presumed or known sexual individuals and 1834 likely apomicts (Table 2). These represent 485 unique entities (sexual diploid taxa and their hybrid derivatives), a phenomenal amount of diversification given the relatively short evolutionary history of the
group (29). A full list of sexual diploid taxa and the current sampling of each can be accessed by clicking on ‘By Taxon’ on the entry page of the website. As of this publication, all known sexual diploid taxa except the critically endangered *B. perstellata* are represented in the database, with an average of 31 samples each. Several previously recognized rare endemics (e.g. *Boechera hoffmannii*, *Boechera serotina*, *Boechera yorkii*) remain poorly sampled, as do many of the newly discovered sexual diploids (e.g. ‘roguensis’, ‘tahoensis’, ‘wallowaensis’). In terms of geography, the region with the greatest taxonomic complexity (i.e. western North America) is generally well sampled. However, the various taxa confined to eastern North America and eastern Asia are currently undersampled. Our ongoing research efforts are focused on filling these gaps, and the database will be continuously updated as new results become available.

### Database structure

In order to facilitate the broadest usage of these data, the database has been made openly accessible through a web portal (http://sites.biology.duke.edu/windhamlab/), constructed using custom Python and HTML scripts, and connected by the Common Gateway Interface. The core is a Python library for interacting with the database (scripts available at https://bitbucket.org/fayweili/boecheradata base.git). Users can search and retrieve basic metadata using the Extrac# (a unique ID number annotated on every voucher specimen), taxon name, or locality (Fig. 2). Searches based on taxon name can be limited to sexual diploids (i.e. just the epithet entered) or expanded to include all hybrids containing that genome. The output of a taxon query can be used to produce working state and county level distribution lists free of the misidentifications common in other on-line resources. The search by locality function allows investigators to quickly assemble checklists of confirmed *Boechera* identifications for any state, province, or county. This can improve the efficiency of morphology-based identification by reducing the diversity of taxa and the number of dichotomous key couplets that need to be negotiated. To further facilitate this process, exemplar specimens of each sexual diploid taxon as well as all documented hybrid combinations are being imaged and will be available via the portal. In addition to the basic search and display functions, we also include two novel algorithms: ‘Taxon Enquiry based on Similarity of Loci and Alleles’ (TESLA) for sample identification, and ‘Parental Relationship Identification Using Subtraction’ (PRIUS) for inferring the genomic constitution of hybrid individuals (see below). Although many microsatellite databases have been developed for plants, the majority of them focus on crop plants with the main purpose to assist breeding and genotyping marker selection (e.g. 30, 31). To our knowledge, the only comparable database for biodiversity research and species identification is the Olive Genetic Diversity Database (32).

### Taxon enquiry based on similarity of loci and alleles (TESLA)

Taxonomic identification of *Boechera* specimens is notoriously challenging, even for specialists (11). With about ±83 sexual diploid taxa that hybridize whenever they come into contact to produce true-breeding apomictic diploids, triploids and tetraploids, it is impossible to identify many accessions based on gross morphology alone (14). To address this problem, we have developed TESLA, a microsatellite-based identification method. Once a user enters allelic information for their sample, TESLA compares this multi-locus genotype (MLG) to every accession in the database, calculating an Allele Similarity (AS) score between two accessions as follows:

\[
\text{AS} = \frac{\text{No. shared alleles} \times (\text{No. loci} - \text{No. loci without data})}{\text{Total allele count in query} \times \text{No. loci}}
\]

The AS score represents the proportion of alleles that are shared relative to the query, scaled by the amount of missing data. The output of a TESLA query consists of the specimens with the 100 highest AS scores listed in descending order. In addition to researching samples already included in the database, user-generated MLGs can be entered individually (using the ‘by alleles’ function) or a query spreadsheet can be uploaded to the database in batch mode (Fig. 2).

Both the rank value and magnitude of the AS scores derived from TESLA are useful in identifying unknown *Boechera* accessions. For example, in queries involving 100 random sexual diploid samples of known identity, the highest AS scores coincided with the ‘correct’ identification 96% of the time. In an additional 3%, the highest scoring matches belonged to the same major clade as the expected taxon (clade assignments of all sexual diploid taxa can be
accessed through ‘Search Database By Taxon’). These phylogenetically localized mismatches were attributable to limited sampling of the taxon involved, poor differentiation between the target taxon and its close relatives, and/or missing data. In the single instance in which the individual with the highest AS score belonged to an unrelated taxon, the analysis was impeded by a lack of data at the majority of loci. In terms of magnitude, the highest AS scores from TESLA searches resulting in correct identifications ranged from 1.000 to 0.375. Misidentifications that attribute a sample to the wrong clade became common below 0.500 and we therefore recommend this as the threshold for considering a TESLA result a hypothesis worthy of serious consideration.

**Parental relationship identification using subtraction (PRIUS)**

*Boechera* hybrids are ubiquitous, often comprising >50% of plants randomly sampled from natural populations (M. D. Windham, personal observation). They are also hyper-diverse, drawing nearly every known sexual genome into the apomictic milieu (12, 13). Few extant analytical programs can determine the genomic constitution of hybrid individuals based on microsatellite data (33–35), and none of the available implementations are capable of handling the tri-genomic hybrids common in *Boechera*. To address this need, we have developed a program called ‘PRIUS’, available via the online database portal.

The basic workflow of PRIUS is as follows: (i) identify the most similar sexual diploid specimens in the database based on AS score, (ii) subtract from the hybrid’s MLG the corresponding alleles from each top-scoring specimen and (iii) repeat the previous steps using the subtracted genotype for one or two more iterations (see below). The end result is a list of diploid specimen combinations that best explain the hybrid’s MLG. The corresponding species of these diploid specimens then form a working hypothesis of the putative parental species.

The proper functioning of PRIUS depends on the ability to distinguish between sexual diploids (the pool from which potential parents are drawn) and apomicts (generally the products of hybridization; see 12, 36). Fortunately, this information resides within the microsatellite data themselves. Based on a broad taxonomic sampling of 13 microsatellite loci in 1393 *Boechera* specimens, Beck et al. (12) demonstrated clear bimodality in the number of heterozygous loci, a metric that was strongly associated with inferred breeding system. In this study, the mean heterozygosity of diploid plants with predominantly sexual pollen was 0.232 (3.02/13), whereas the mean for diploid plants with predominantly apomictic pollen was 0.723 (9.4/13).

Only one sexual diploid taxon (*parishii*) included multiple samples in which heterozygosity exceeded 0.5, and only 6 of 82 apomictic diploid taxa exhibited heterozygosity <0.5 (12). Based on this empirical evidence, we set the PRIUS heterozygosity threshold at 0.5, thus creating essentially non-overlapping pools of potential parents (sexual diploids) and apomictic (or sexual tetraploid) hybrids subject to a PRIUS query.

To begin the PRIUS process, the program infers the ploidy of the query accession based on the maximum number of alleles at two or more of the 13 loci used to calculate heterozygosity (see Table 1). If the accession is determined to be triploid, PRIUS identifies the five sexual diploid specimens in the full dataset with the highest AS scores relative to the queried MLG. For each tri-allelic locus in the query accession, PRIUS independently subtracts from the MLG any matching alleles observed in each of the five sexual diploids, generating five new query MLGs. These roughly approximate apomictic diploid progenitors that could have hybridized with the subtracted genome to form the original triploid MLG. Each of the five subtracted MLGs is then queried in a second iteration, where PRIUS identifies the five sexual diploid specimens with the highest AS scores relative to each of the subtracted MLGs. For each tri-allelic and tri-allelic locus in the five queried MLGs, the program subtracts any matching alleles from each of the five most similar sexual diploids. This step generates 25 (5²) new query MLGs that very roughly approximate a third sexual diploid genome that might have combined with the two subtracted genomes to form the original triploid MLG. A third iteration then identifies the five sexual diploid specimens with the highest AS scores relative to each of the 25 new subtracted MLGs. After three iterations, there will be 125 (5³) different specimen combinations, and PRIUS ranks them by cumulative AS scores. If a query accession is determined to be diploid (i.e. exhibits a maximum of two alleles at 12 of the 13 loci used to assess heterozygosity), PRIUS begins with the second iteration. As with TESLA, a user generated MLG can be entered individually (using the ‘by alleles’ function) or a query spreadsheet can be uploaded to the database in batch mode (Fig. 2).

PRIUS is not designed to directly address the parentage of the few tetraploid apomicts encountered (12/4428 specimens or 0.3% of the dataset) because a fourth round of subtraction produces a high percentage of spurious matches. PRIUS searches are also negatively affected by the presence of undetected null alleles in a queried accession. This is because the AS scores of appropriate sexual diploid accessions with null alleles are lowered, reducing their likelihood of being identified as possible parents. The effect can be significant, resulting in a bias against certain
sexual diploid taxa (e.g. *retrorfracta*) that are extensively involved in the formation of apomictic hybrids (11). This problem is not easily addressed, but we have instituted an approach that improves the success rate in both PRIUS and TESLA for many triploid accessions.

In sexual taxa, null alleles are hypothesized to exist in the homozygous state if both of the following conditions were met: (i) the remaining multiplexed loci from that individual amplified, and (ii) individuals in the same genotyping run representing other taxa that lack null alleles amplified. Because most *Boechera* triploids (even those arising within named species) combine divergent genomes with different alleles, we expect a given triploid sample to be tri-allelic or, at least, di-allelic at the majority of these highly variable loci (Table 1). A broad survey of the database suggests that the occurrence of apparently homozygous loci in such accessions is more often the result of undetected null alleles than the improbable incorporation of three identical alleles from three different parental genomes. Both PRIUS and TESLA therefore add a null allele (as ‘0’) to apparently homozygous loci in triploid queries. This protocol has no effect on the AS scores of accessions supporting the ‘three identical alleles’ hypothesis but increases the probability that sexual diploids with null alleles will be considered as possible parents.

It should be stressed that although PRIUS is a powerful tool for data exploration and generating hypotheses regarding the genomic constitution of hybrid *Boechera* individuals, the hypothesized hybrid origin should be evaluated with the widest array of data that can be brought to bear, including morphology, cytology, ecology, geography, genomics and reproductive biology (13).

### Results

The following examples demonstrate the functionality of TESLA (Table 3) and PRIUS (Fig. 2) for identifying hybrids. The sample queried (Extrac# PJA154a) was subject to a variety of analyses by Alexander *et al.* (13) and thus provides a good basis for comparing the outputs of these algorithms to other approaches used to identify taxa and propose hybrid parentages. Table 3 shows the result of a TESLA search on PJA154a, demonstrating that all of the most similar accessions in the database (with AS scores ranging from 1.0000 to 0.9143) have been identified as triploid hybrids incorporating genomes derived from *gracilipes, perennans* and *texana*. To facilitate presentation of this TESLA query, the full 15-locus dataset has been reduced to the eight loci that show allelic variability within this hybrid cluster. In addition, the standard output of TESLA (i.e. the 100 most similar accessions in the database) has been condensed by grouping hybrids with identical genotypes and assigning members of individual sexual diploid taxa with identical AS scores to a single row (the number of condensed accessions indicated parenthetically in the Extrac# column).

As indicated earlier, the top hits for a TESLA query on PJA154a are all specimens identified as *gracilipes × perennans × texana*. These comprise 13 different MLGs representing 30 individual samples (Table 3). In fact, all but two of the samples in the dataset with this hypothesized parentage are clustered at the top of this figure with AS scores ≥ 0.9143. One exception is Extrac# FW765 (Table 3; row 19), which is identical to PJA222e (row 12) except that it lacks data at two loci (c8 and b6) and has incomplete data at two others (ICE14 and Bdu266). The remaining exception to the tight clustering of samples identified as *gracilipes × perennans × texana* is Extrac# JB1453 (Table 3; last row), which has an AS score of 0.6111 reflecting divergent allele profiles at seven of the eight illustrated loci. We suspect that Extrac# JB1453 represents an independent hybrid origin of this triploid apomict, a hypothesis discussed in more detail below.

Random sampling of apomictic hybrids in the current dataset indicate that ca. 60% of TESLA searches correctly specify their genomic make-up, placing plants with identical parentage at the top of the results list. In the remaining 40% of cases in which highly similar hybrid genotypes have not yet been sampled, the probable parents typically are discernable among the sexual diploid taxa with high AS scores. Our query involving PJA154a (Table 3) can be used to illustrate this approach as well. Ignoring for the moment the hybrid origin suggested by rows 3–15, the next highest AS score (0.8667) is shared by individuals of *perennans* and *gracilipes*. A third sexual diploid taxon (*texana*) appears at an AS score of 0.7333. Repeated occurrences of these three species [as well as *gracilipes × perennans*, a possible apomictic diploid bridge discussed by Alexander *et al.* (13)] suggest the same hybrid parentage reflected in rows 3–15.

PRIUS provides an alternative, taxon-centric approach to assessing hybrid parentage and is especially informative in cases where TESLA fails to provide a clear result. Figure 2 presents a screen shot of the PRIUS query result for Extrac# PJA154a, the same individual used in the preceding TESLA search. The cumulative scores on the left serve the same function as the AS scores in TESLA. In the case of PJA154a, the highest cumulative score is 2.1333, which is represented by 92 rows (different combinations of individual accessions), with 10 presented in Figure 2, for clarity. In every case, putative parent 1 is an accession of *perennans* while putative parent 2 is *texana*. In 82/92 combinations (89.1%), an accession of *gracilipes* appears as
Table 3. Output from a TESLA query on Extrac# PJA154a

| AS_score | Extrace# | Taxon ID | State  | County | ICE3  | ICE3  | c8    | c8    | ICE14 | ICE14 | BF3   | BF3   | BF3   | b6    | b6    | BF19  | BF19  | BF19  | BF19  | BF15  | BF15  | BF15  | Bdmu266 | Bdmu266 | Bdmu266 |
|----------|----------|----------|--------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|
| 1        | PJA154a  | NM Eddy  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 146    | |        |
| 0.9722   | PJA163h  | NM Doña  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 154    | |        |
| 0.9722   | PJA382z  | NM Torrance | 71 | 75 | 93 | 244 | 250 | 211 | 223 | 107 | 111 | 129 | 299 | 309 | 311 | 145 | 0 | 93 | 104 | 115 | 133 | 150    | |        |
| 0.9722   | PJA380a  | NM Otero  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 152    | |        |
| 0.9459   | PJA380z  | NM Torrance | 71 | 75 | 93 | 244 | 250 | 211 | 223 | 107 | 111 | 129 | 299 | 309 | 311 | 145 | 93 | 104 | 115 | 133 | 142    | |        |
| 0.9444   | PJA230a  | NM Doña  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 150    | |        |
| 0.9444   | PJA230c  | NM Doña  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 152    | |        |
| 0.9211   | JB1346   | NM San Miguel | 71 | 75 | 93 | 244 | 250 | 211 | 223 | 107 | 111 | 129 | 299 | 309 | 311 | 145 | 93 | 104 | 115 | 133 | 150    | |        |
| 0.9189   | PJA222c  | NM Doña  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 150    | |        |
| 0.9189   | PJA385x  | NM Otero  | 71     | 75     | 93    | 244   | 248   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 131   | 148    | |        |
| 0.9167   | JB1345   | NM Socorro | 71 | 75 | 93 | 244 | 250 | 211 | 223 | 111 | 127 | 299 | 309 | 311 | 145 | 93 | 104 | 115 | 136 | 152    | |        |
| 0.9143   | PJA395c  | NM Lincoln | 71 | 75 | 77 | 244 | 250 | 211 | 223 | 107 | 111 | 129 | 299 | 309 | 145 | 93 | 104 | 115 | 131 | 150    | |        |
| 0.8667   | FW812    | AZ Maricopa | 75 | 244 | 258 | 223 | 11 | 155 | 307 | 145 | 153 | 93 | 141 | 150 | |        |
| 0.8667   | FW483    | AZ Gila     | 71 | 250 | 252 | 223 | 1 | 105 | 107 | 171 | 21 | 147 | 104 | 125 | 133 | |        |
| 0.8      | FW799    | AZ Gila     | 75 | 244 | 0   | 11 | 135 | 305 | 309 | 157 | 163 | 93 | 131 | 154 | |        |
| 0.7407   | FW765    | NM Doña  | 71     | 75     | 93    | 244   | 250   | 211   | 223   | 107   | 111   | 129   | 299   | 309   | 311   | 145   | 0     | 93    | 104   | 115   | 133   | 150    | |        |
| 0.7333   | PJA151h  | NM Doña  | 75     | 248   | 254 | 223 | 99 | 129 | 305 | 319 | 149 | 93 | 125 | 146 | |        |
| 0.7333   | JB1454   | TX Jeff Davis | 73 | 232 | 211 | 97 | 123 | 313 | 0 | 101 | 115 | 127 | 146 | |        |
| 0.7143   | JB595    | AZ Cochise | 71 | 248 | 232 | 223 | 105 | 129 | 311 | 313 | 145 | 151 | 155 | 93 | 104 | 133 | 141 | |        |
| 0.6667   | FW811    | AZ Maricopa | 75 | 264 | 223 | 119 | 305 | 0 | 93 | 133 | | | |        |
| 0.6667   | JB905    | AZ Coconino | 71 | 238 | 223 | 99 | 321 | 145 | 151 | 104 | 0 | |        |
| 0.6667   | JB183    | CO Fremont | 75 | 242 | 223 | 101 | 311 | 143 | 99 | 133 | | | |        |
| 0.6667   | JB1447   | TX Browster | 73 | 232 | 240 | 211 | 107 | 115 | 313 | 319 | 0 | 101 | 137 | 144 | |        |
| 0.6538   | JB648    | AZ Graham | 73 | 252 | 262 | 223 | 111 | 119 | 309 | 315 | 145 | 93 | 104 | 156 | |        |
| 0.6111   | JB1453   | TX Caliberson | 73 | 232 | 242 | 248 | 211 | 223 | 107 | 113 | 125 | 299 | 309 | 315 | 147 | 151 | 155 | 93 | 104 | 137 | 146 | 148 | |        |
putative parent 3. In another six cases (6.5%), perennans is repeated as the third parent and, in four others (4.4%), fendleri is identified as putative parent 3. All four of these sexual diploid taxa belong to the same major clade of Boechera (13, 21).

In Alexander et al. (13) Extrac# PJA154a is labeled PJA195a (the original collector’s number) and is identified as Boechera porphyrea (the correct binomial name for hybrids of this proposed parentage). In a phylogenetic analysis incorporating data from the nuclear pistillata locus, different clone sequences from this sample associated with gracilipes, perennans and texana, respectively (13). STRUCTURE analysis of the microsatellite data available at the time for this taxon revealed admixture proportions averaging 25% gracilipes, 37% perennans and 23% texana, with a lesser contribution attributed to fendleri. Thus, the results reported by Alexander et al. (13) are nearly identical to those obtained from our TESLA and PRIUS queries on PJA154a. Our new algorithms have two clear advantages, however. First, using TESLA and PRIUS requires no specialized taxonomic knowledge of the group. To conduct their analyses, both Alexander et al. (13) and Beck et al. (12) had to decide a priori which Boechera species were likely involved in the hybridization events they were trying to reconstruct. Second, TESLA and PRIUS results are obtained in seconds, as opposed to the time-consuming STRUCTURE algorithm.

Discussion

In addition to providing microsatellite-based tools to facilitate the identification of Boechera specimens, the database also can be used to explore large-scale evolutionary patterns within the genus. Insights regarding the relationships among ploidy, heterozygosity and reproductive mode, and processes of hybrid formation and evolution are discussed below.

Relationship between reproductive mode and ploidy

Boechera ploidy is non-randomly associated with reproductive mode (28). The clearest such association involves plants with three full sets of chromosomes (2n = 21) that have no chance of pairing normally in meiosis. The current dataset includes 959 triploid accessions (Table 2), all of which are inferred to be apomictic based on the stepwise analyses outlined in Figure 1. Whereas triploids represent 21.7% of the current dataset, tetraploids are rare (1.5%) and higher ploidy levels have not been observed. Based on cytogenetic and pollen data, the majority of tetraploid individuals are sexual; though 12 apomictic tetraploids (representing seven different genomic combinations) are present as well (Table 3). The two types of tetraploids are readily separated based on the maximum number of alleles per locus (two or rarely three in sexual tetraploids vs. four in apomicts). The remaining accessions (77.8%) are considered diploid based on the allele number criterion discussed earlier. Although ca. 75% of these appear to be sexual, our sampling was biased toward sexual individuals, and apomictic diploids are likely to be more prevalent in nature. This is a critical feature of the Boechera model system given the rarity of apomictic diploids among flowering plants (25).

Relationship between reproductive mode and heterozygosity

As suggested by previous studies, apomicts exhibit higher levels of heterozygosity than sexually reproducing Boechera accessions (Fig. 3). This is not surprising given that the observed heterozygosity is largely the result of hybridization between different inbred genotypes, which may be a trigger for the development of apomixis (12, 37). Results from a Wilcoxon rank sum test of the current dataset indicate that apomict heterozygosity is significantly higher than that of sexuals (apomict mean 0.78 ± 0.14 SD, median 0.77 vs. sexual mean 0.11 ± 0.18 SD, median 0; Wilcoxon W = 4709860, P < 2.2e-16; Fig. 3). Although mean heterozygosity of triploid (0.85 ± 0.10 SD) and tetraploid apomicts (0.90 ± 0.07 SD) is higher than that of diploid apomicts (0.70 ± 0.13 SD), this result does not change qualitatively when comparing diploid only sexuals and apomicts.
The data presented in Figure 3 provide strong empirical support for using the 0.5 heterozygosity threshold implemented in PRIUS to separate potential parents (sexual diploids) from the non-segregating hybrids subject to analysis. Just 147 accessions (3.3% of the dataset) exhibit heterozygosity values higher or lower than expectations based on reproductive mode inferred via the pathway outlined in Figure 3. Unexpectedly high values are observed in 56 accessions of the two sexual tetraploid taxa (bolmgenrii × lemmonii and laevigata × stricta) discussed earlier and 28 individuals of the sexual diploid taxon parishii. This species exhibits unusually high allelic diversity and possesses a suite of traits that suggest or enforce outcrossing. Unexpectedly low values are observed in a series of apomictic diploids that include: (i) a variety of hybrids between closely related sexual diploids that are insufficiently divergent to meet the 0.5 threshold when crossed, and (ii) a series of apomictic diploid hybrids that include retrofracta or other species with null alleles in their parentage.

Hybrid formation and evolution

Earlier articles hint at the important role hybridization has played in the evolutionary diversification of Boechera (12, 13), but we clearly establish that hybridization is ubiquitous and largely unconstrained by phylogeny. Of the 485 taxa included in the current database, 400 are confirmed or inferred apomorphic hybrids. Eighteen of these have arisen through crosses between divergent inbred lines within particular sexual diploid taxa; the remaining 382 are the products of hybridization between distinct taxa. Of the 83 sexual diploid taxa, 64 have been involved in at least one hybridization event and the 19 sexual diploid taxa not currently known to hybridize are those that rarely co-occur with other taxa and/or are currently under-sampled.

Hybridization has also occurred repeatedly among the same sexual diploid taxa, producing hybrid lineages with similar taxonomic constitutions but different parental alleles. The results for a TESLA query for Extrac# PJA154a, provide a good example (Table 3). The queried specimen, an apomictic triploid hybrid incorporating genomes derived from gracilipes × perennans × texana (13), belongs to a cluster of 13 very similar MLGs with AS scores ≥ 0.9143 and one outlier (Extrac# FW765, AS 0.7407) identical to one of these except for missing data. There is, however, a 14th MLG represented by a single sample (Table 3; Extrac# JB1453, row 28) with a much lower AS score of 0.6111. The lower score reflects the fact that there are allelic differences between JB1453 and the other 13 MLGs at nearly every locus analyzed. Given the lack of chromosome pairing and allelic segregation in these apomictic triploids, it is very likely that the apomictic triploid gracilipes × perennans × texana had at least two independent origins. A cursory survey of the data included in the BMW indicates that multiple origins of Boechera hybrid apomicts are the rule rather than the exception.

Conclusion

The taxonomic complexity of Boechera has long stymied those investigating the diversity, biogeography, and evolutionary dynamics of the group (8, 11). Here we present a database containing information for 4428 specimens collected throughout the native range in North America, Greenland and Russia and encompassing >95% of all named taxa. A variety of search algorithms facilitate assigning the correct taxon name, ploidy, reproductive mode, and hybrid parentage to any Boechera specimen with appropriate microsatellite data. Two novel identification algorithms, TESLA and PRIUS, facilitate the identification of unknown Boechera specimens using simple, standardized methods. This database, publicly available and regularly updated, will provide valuable new tools for researchers interested in Boechera as a model genus.

Acknowledgements

We thank the curators of the following herbaria (acronyms from 38) for permission to remove pollen and small leaf fragments: ALA, ARIZ, ASC, ASU, BHSU, BRY, CAN, CAS, COLO, CS, DAO, DS, DUKE, GH, JEPS, MO, MONTU, NCU, NDG, NY, ORE, OSC, POM, RM, RSA, SJNM, UBC, UC, UNLV, UNM, USFS, UT, UTC, WILLU, WS, WTU and also Ryan Hardy at Duke Office of Information Technology for helping to setup the website server. Kathleen Pryer, Tom Mitchell-olds, Carl Rothfels, and three anonymous reviewers also provided insightful suggestions.

Funding

This work was funded by National Science Foundation DEB-0816560 (to M.D.W.), Doctoral Dissertation Improvement Grant DEB-1407158 (to F.-W.L.) and DEB-1311269 (to C.A.R.). F.-W.L and C.A.R. were each supported by a National Science Foundation Graduate Research Fellowship.

Conflict of interest. None declared.

References

1. Rushworth,C.A., Song,B.H., Lee,C.R.R. et al. (2011) Boechera, a model system for ecological genomics. Mol. Ecol., 20, 4843–4857.
2. Rollins,R.C. (1941) A monographic study of Arabi in western North America. Rhodora, 43, 289–325.
3. Böcher,T.W. (1951) Cytological and embryological studies in the amphi-apomictic Arabis holboellii complex. Biol. Skr. Dan. Vid. Selsk., 6, 1–59.
4. Böcher,T.W. (1964) Further studies in the Arabis holboellii complex. Bot. Tidsskr., 64, 141–161.
5. Bocher, T.W. (1954) Experimental taxonomic studies in the Arabis holboellii complex. Svensk. Bot. Tidskr., 48, 31–44.
6. Rollins, R.C. The Cruciferae of Continental North America; Stanford University Press, Palo Alto, 1993.
7. Roy, B.A., and Rieseberg, L.H. (1989) Evidence for apomixis in Arabis. J. Hered., 80, 506–508.
8. Windham, M.D., and Al-Shehbaz, I.A. (2006) New and noteworthy species of Boechera (Brassicaceae) I: sexual diploids. Harvard Pap. Bot., 11, 61–88.
9. Windham, M.D., and Al-Shehbaz, I.A. (2007) New and noteworthy species of Boechera (Brassicaceae) II: apomorphic hybrids. Harvard Pap. Bot., 11, 257–274.
10. Windham, M.D., and Al-Shehbaz, I.A. (2007) New and noteworthy species of Boechera (Brassicaceae) III: additional sexual diploids and apomictic hybrids. Harvard Pap. Botany, 12, 235–257.
11. Al-Shehbaz, I.A. and Windham, M.D. (2010) Boechera. In: Flora of North America Editorial Committee (ed). Flora of North America North of Mexico. Oxford University Press, Oxford, pp. 348–412.
12. Beck, J.B., Alexander, P.J., Allphin, L. et al. (2012) Does hybridization drive the transition to asexuality in diploid Boechera? Evolution, 66, 985–995.
13. Alexander, P.J., Windham, M.D., Beck, J.B. et al. (2015) Weaving a tangled web: Divergent and reticulate speciation in Boechera fendleri sensu lato (Brassicaceae: Boechereae). Syst. Bot., 40, 572–596.
14. Windham, M.D., Beck, J.B., Li, F.W. et al. (2015) Searching for diamonds in the apomorphic rough: a case study involving Boechera lignifera (Brassicaceae). Syst. Bot., 40, 1031–1044.
15. Rieseberg, L.H., Carter, R. and Zona, S. (1990) Molecular tests of the hypothesized hybrid origin of two diploid Helianthus species (Asteraceae). Evolution, 44, 1498–1511.
16. Wendel, J.F., Schnabel, A. and Seelanan, T. (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). Proc. Natl. Acad. Sci. USA., 92, 280–284.
17. Doyle, J.J., Doyle, J.L. and Brown, A.H. (1999) Origins, colonization, and lineage recombination in a widespread perennial soybean polyploid complex. Proc. Natl. Acad. Sci. USA., 96, 10741–10745.
18. Beck, J.B., Windham, M.D., Yatskievych, G. et al. (2010) A diploids-first approach to species delimitation and interpreting polyploid evolution in the fern genus Astroideps (Pteridaceae). Syst. Bot., 35, 223–234.
19. Kiefer, C., Dobeš, C. and Koch, M.A. (2009) Boechera or not? Phylogeny and phylogeography of eastern North American Boechera species (Brassicaceae). Taxon, 58, 1109–1121.
20. Kiefer, C., and Koch, M.A. (2012) A continental-wide perspective: the genepool of nuclear encoded ribosomal DNA and single-copy gene sequences in North American Boechera (Brassicaceae). PlaSOne, 7, e36491.
21. Alexander, P.J., Windham, M.D., Beck, J.B. et al. (2013) Molecular phylogenetics and taxonomy of the genus Boechera and related genera (Brassicaceae: Boechereae). Syst. Bot., 38, 192–209.
22. Claus, M.J., Cobban, H., and Mitchell-Olds, T. (2002) Cross-species microsatellite markers for elucidating population genetic structure in Arabidopsis and Arabis (Brassicaceae). Mol. Ecol., 11, 591–601.
23. Dobeš, C., Mitchell-Olds, T., and Koch, M.A. (2004) Intraspecific diversification in North American Boechera stricta (= Arabis drummondii), Boechera xdivaricarpa, and Boechera holboellii (Brassicaceae) inferred from nuclear and chloroplast molecular markers—an integrative approach. Am. J. Bot., 91, 2087–2101.
24. Song, B.H., Clauss, M.J., Pepper, A. et al. (2006) Geographic patterns of microsatellite variation in Boechera stricta, a close relative of Arabidopsis. Mol. Ecol., 15, 357–369.
25. Asker, S. and Jerling, L. (1992) Apomixis in Plants; CRC Press, Boca Raton, Florida.
26. Dobeš, C., Koch, M. and Sharbel, T.F. (2006) Embryology, karyology, and modes of reproduction in the North American genus Boechera (Brassicaceae): a compilation of seven decades of research. Ann. Missouri Bot. Gard., 93, 517–534.
27. Doyle, J.J. and Egan, A.N. (2010) Dating the origins of polyploidy events. New Phytol., 186, 73–85.
28. Schranz, M.E., Dobeš, C., Koch, M.A. et al. (2005) Sexual reproduction, hybridization, apomixis, and polyploidization in the genus Boechera (Brassicaceae). Am. J. Bot., 92, 1797–1810.
29. Song, B.H., Lee, C.R.R., Pais, A. et al. Rapid adaptive speciation and divergence population genetics in Boechera, a model system in ecology and evolution. in review.
30. Doddamani, D., Katta, M.A.V.S.K., Khan, A.W. et al. (2014) CicArMSatDB: the chickpea microsatellite database. BMC Bioinformatics, 15, 212.
31. Krishnan, R.R., Sumarth, R., Bindroo, B.B. et al. (2014) MulSatDB: a first online database for mulberry microsatellites. Trees, 28, 1793–1799.
32. Ben Ayed, R., Ben Hassen, H., Ennouri, K. et al. (2016) OGDD ( Olive Genetic Diversity Database): a microsatellite markers’ genotypes database of worldwide olive trees for cultivar identification and virgin olive oil traceability. Database, 2016, bav090–bav099.
33. Anderson, E.C. and Thompson, E.A. (2002) A model-based method for identifying species hybrids using multilocus genetic data. Genetics, 160, 1217–1229.
34. Spielmann, A., Harris, S.A., Boshier, D.H. et al. (2015) Orchard: paternity program for autotetraploid species. Mol Ecol Resour, 15, 915–920.
35. Walling, C.A., Pemberton, J.M., Hadfield, J.D. et al. (2010) Comparing parentage inference software: reanalysis of a red deer pedigree. Mol. Ecol., 19, 1914–1928.
36. Kantama, L., Sharbel, T.F., Schranz, M.E. et al. (2007) Diploid apomicts of the Boechera holboellii complex display large-scale chromosome substitutions and aberrant chromosomes. Proc. Natl. Acad. Sci. USA., 104, 14026–14031.
37. Carman, J.G. (1997) Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispory, tetraspory, and polyploidy. Biol. J. Linn. Soc., 61, 51–94.
38. Thiers, B. (2016) Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden’s Virtual Herbarium. http://sweetgum.nybg.org/science/ih/ (August 2016, last accessed).