**Generating genomic data sets**

**De novo assembly**

After demultiplexing we trimmed reads to the same 85 base pairs length using trimmomatic (Bolger et al., 2014). To identify the optimal parameters for assembling loci de novo in STACKS, we followed the parameter-testing pipeline established by Rochette and Catchen (2017) and Paris et al. (2017). We ran the STACKS multiple times varying the parameters on each step. We ranged the ‘*m’* parameter from 3 to 10 (m3 – m10), the ‘*M’* parameter from 3 to 5 (M3–M5), the ‘*n*’parameter from 2 to 7 (n2–n7) and ‘*r*’ was kept constant at 0.80, with the rest of the parameters on default setting. ‘*r*’ = 0.80 signifies that a locus must be present in a minimum of 80% of individuals.

To examine the output from each de novo assembly run, we used vcftools to create statistics for each run (Danecek et al., 2011). These statistics were generated and examined to select the optimal parameters for our data set. The statistics used were variant quality, variant mean depth, variant missingness, minor allele frequency and were then observed in R. The results generated very similar results and we chose the parameter combination -*m* 5 -*M* 3 -*n* 3 for the Aciurina dataset and -*m* 7 -*M* 3 -*n* 3 for Ericameria. After selecting one of the parameters, we used SNPfiltR to further streamline and automate the process of choosing appropriate filtering parameters for our datasets (DeRaad, 2022; vcfR Knaus and Grunwald, 2017). Briefly, we filtered for genotype depth and quality, allele balance, maximum depth, and missing data per sample and per SNP, resulting in a filtered SNP data set containing 96 original samples, filtered from 26,023 to 22,057 SNPs with 12.81% overall missing data for Aciurina. For Ericameria, from the initial 198 samples, only 101 passed filtering thresholds. The 96 samples contained 534,532 sites filtered down to 2,462 with 10,57% missing data. A second de novo run was doe for Ericameria including only the 96 samples that passed filtering and additional samples from Utah, Arizona, and New Mexico from Faske et al. (2021). For this second run including extra samples, there were XXX samples of which XXX passed filtering and contained XXXXXX sites, after filtering resulted with XXXX with XXXX% missing data. Lists of samples and full parameter optimization and filtering details are available in supplementary methods for each data set (organize sample list and link to R code and visualizations detailing filtering pipeline).

Separate analyses were run for subsets of both the Aciurina and Ericameria data sets. The full pipeline the de novo alignment, PCA and trees were run for M+C+F and S for Aciurina; and C+F+S and a small subset of C+F+S+M+G (5 samples of each) for Ericameria. The parameters were not optimized for all these runs and if they were the same parameters as the full data set were chosen since results were similar.

**Reference based alignment**

The Ericameria dataset presented was also processed using a reference-based alignment. We aligned sequences to *Erigeron canadensis* genome (GCA\_010389155.1; Peng Y, et al. 2014) using the Burrows–Wheeler Aligner (BWA) version 0.7.17 short-read aligner with default parameters and the MEM alignment algorithm (Li and Durbin 2010). This was the closest complete genome available for the group at the moment of the analysis. Statistics were generated and filtering using SNPfiltR was done similarly to the de novo pipeline. For the reference-based run with our samples only there were XXX samples of which XXX passed filtering and contained XXXXXX sites, after filtering resulted with XXXX with XXXX% missing data.

For the reference-based run including the Faske et al. (2021) there were 346 samples of which 96 were ours and 250 were Faske et al. (2021). Of the 346, 259 passed filtering and contained 6,728 sites, after filtering resulted with 2,026 with 13,84% missing data.

Characterizing genomic structure and differentiation

We used different approaches to both **identify and describe** genomic structure, including principal component analysis (PCA), sNMF and phylogenetic trees. We used PCA for an initial exploratory screening of data in R (cite adegenet package). Since PCA analysis makes no assumptions about sampled and ancestral populations, we ran sNMF in R for estimates of ancestry coefficients (cite LEA package). PCA and sNMF scripts are available in supplementary methods (link to R code for pipeline).

For the phylogenies, we constructed a maximum likelihood (ML) tree with IQ-TREE 2.2.0 (Minh et al. 2020) and used ModelFinder (Kalyaanamoorthy et al. 2017) to select a model for Akaike Information Criterion (AIC). Branch support was evaluated with 1000 ultrafast bootstrap (UFBoot) (Hoang et al. 2018) and 1000 Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT) (Guindon et al. 2010) replications. We report the selected model and branch support values in the results section.

References

GENOME: De novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. Peng Y, et al. Plant Physiol 2014 Nov