1   Leimbach – Maus et al. – *Gypsophila paniculata* microsatellites

2

3   **Microsatellite primer development for invasive perennial herb, *Gypsophila paniculata***

4   **(Caryophyllaceae)[1]**

5

6   Hailee B. Leimbach-Maus[2,3], Syndell R. Parks[2], and Charlyn G. Partridge[2]

7

8   [2]Annis Water Resources Institute, Grand Valley State University (AWRI-GVSU), 740 W.

9   Shoreline Dr., Muskegon, MI 49441 USA

10

11   Email addresses: HBLM: hailee.leimbachmaus@gmail.com

12   SRP: syndell.parks@gmail.com

13   CGP: partridc@gvsu.edu

14   Number of words: 1,191.

15

16   [1]Manuscript received _____; revision accepted_____.

17   [3]Author for correspondence: hailee.leimbachmaus@gmail.com

22

23

ABSTRACT

- *Premise of the study: Gypsophila paniculata* L. (baby's breath) is an herbaceous perennial that has invaded much of northern and western United States and Canada, outcompeting and crowding out native and endemic species. Microsatellite primers were developed to analyze the genetic structure of invasive populations.

- *Methods and Results:* We have identified 16 polymorphic nuclear microsatellite loci for *G. paniculata* out of 73 loci that successfully amplified from a primer library created using Illumina sequencing technology. The developed primers amplified microsatellite loci in 3 invasive populations in Michigan. Primers amplified di-, tri-, and tetra-nucleotide repeats.

- *Conclusions:* These markers will be useful in characterizing the genetic structure of invasive populations throughout North America to aid targeted management efforts, and in native Eurasian populations to better understand invasion history. Five of these developed primers also amplified in *G. elegans*.

Key words: Genetic diversity; *Gypsophila paniculata*; invasive species; microsatellites.

47   **INTRODUCTION**

48   The herbaceous perennial forb, *Gypsophila paniculata* L., was introduced to North

49   America in the late 1800's (Darwent and Coupland, 1966). Invasive populations have since been

50   documented throughout the northern and western United States and Canada, specifically in

51   agriculture fields, rangeland, roadsides, and sandy coastlines along the Great Lakes (Darwent,

52   1975; Emery and Doran, 2013). Despite its wide invasive range, little information exists on how

53   populations throughout North America are related or spreading. Due to its aggressive invasion,

54   negative impacts on native biota (Emery and Doran, 2013), and a lack of data regarding its

55   spread, it is important to develop molecular markers that can characterize the genetic structure of

56   invasive populations of *G. paniculata*. These markers will be directly used to investigate

57   invasions within the Lake Michigan coastal dune system where an 1,800-acre infestation occurs

58   (TNC, 2013). These markers and optimized protocols can be used to characterize populations of

59   *G. paniculata* throughout its invasive and native ranges to further assess its invasion history and

60   spread.

61   Calistri et al. (2014) examined the genetic relationship of five *Gypsophila spp*. (including

62   *G. paniculata*) within their native range and 13 commercial hybrid strains using a combination of

63   amplified fragment length polymorphisms (AFLPs), inter simple sequence repeats (ISSRs),

64   target region amplification polymorphism (TRAP), and universal chloroplast simple sequence

65   repeats (cpSSRs). However, the majority of these markers are dominant and thus do not fully

66   distinguish between homozygotes and heterozygotes, a characteristic that would allow for fine-

67   scale population genetic analyses (Freeland et al., 2011). Thus, the development of microsatellite

68   markers for *G. paniculata* is necessary to adequately characterize invasive populations

69   throughout North America.

70 **METHODS AND RESULTS**

71 *Microsatellite Library Development, Assembly and Identification—*

72 Adventitious buds growing from the caudex of five *G. paniculata* plants were collected

73 from Sleeping Bear Dunes National Lakeshore (hereafter Sleeping Bear Dunes or SBDNL) in

74 2015 to develop the microsatellite library. Tissue was stored in indicator silica until DNA

75 extraction. Genomic DNA was extracted using DNeasy plant mini kits (QIAGEN, Hilden,

76 Germany), with modifications including extra wash steps with AW2 buffer. Extracted DNA was

77 run through Zymo OneStep PCR Inhibitor Removal Columns twice (Irvine, California, USA),

78 and checked using a Thermo Fisher Scientific Nanodrop 2000 (Waltham, Massachusetts, USA).

79 For microsatellite library development, each sample was diluted to 50 ng/µL and submitted to

80 Cornell University, Department of Ecology and Evolutionary Biology. Libraries were then

81 submitted to the Sequencing and Genotyping Facility at the Cornell Life Science Core

82 Laboratory Center for sequencing using a 2x250 paired-end format on an Illumina MiSeq

83 (Appendix S1). Raw sequence files for the microsatellite library have been deposited to NCBI's

84 Short Read Archive (Bioproject No: PRJNA431197). A total of 58,907 contigs containing

85 microsatellite loci were obtained. Msatcommander (v 1.0.3) (Faircloth, 2008) identified 3,892

86 potentially unique primers that yielded products of 150-450 bp, had a GC content between 30-

87 70%, and that had a T_m between 58-62ºC, with an optimum of 60ºC (Appendix S2).

88

89 *Primer Optimization—*

90 Prior to PCR optimization, contigs containing potential primers were aligned using

91 ClustalOmega (Sievers et al., 2011) to ensure they were targeting unique microsatellite regions.

92 We tested 107 primer pairs that consisted of either tetrameric, trimeric, or dimeric motifs, and

93   yielded products between 150-300 bp. Of these, 73 successfully amplified, and 16 were

94   determined to be polymorphic and easily scorable (Appendix S3). DNA from leaf tissue

95   collected in 2016 from three populations (Zetterberg Preserve, SBDNL, Petoskey State Park)

96   along eastern Lake Michigan was used for primer optimization (see Table 2 for population

97   geographic coordinates). A minimum of 30 tissue samples were collected from each population.

98   Tissue storage and DNA extraction methods are the same as previously stated.

99   PCR reactions consisted of 1x KCl buffer (Thermo Fisher, Waltham, Massachusetts,

100   USA), 2.0-2.5 mM $MgCl_2$ (Table 1[b]) (Thermo Fisher, Waltham, Massachusetts, USA), 300 µM

101   dNTP (New England BioLabs, Ipswich, Massachusetts, USA), 0.08 mg/mL BSA (Thermo

102   Fisher, Waltham, Massachusetts, USA), 0.4 µM forward primer fluorescently labeled with either

103   FAM, VIC, NED, or PET (Applied Biosystems, Foster City, California, USA), 0.4 µM reverse

104   primer (Integrated DNA Technologies, Coralville, Iowa, USA), 0.25 units of *Taq* polymerase

105   (Thermo Fisher, Waltham, Massachusetts, USA), and a minimum of 50 ng DNA template, all in

106   a 10.0 µL reaction volume. The thermal cycle profile consisted of 94˚C for 5 minutes, 35 cycles

107   of 94˚C for 1 minute, primer-specific annealing temperature (Table 1) for 1 min, 72˚C for 1 min,

108   and a final elongation step of 72˚C for 10 minutes. Successful amplification was determined by

109   visualizing the amplicons on a 2% agarose gel stained with ethidium bromide. Fragment analysis

110   of the amplicons was performed on an ABI3130xl Genetic Analyzer (Applied Biosystems,

111   Foster City, California, USA).

112

113   ***Microsatellite marker data analysis—***

114   Alleles were scored using Genemapper v5 (Applied Biosystems, Foster City, California,

115   USA), and Micro-Checker v2.2.3 (Van Oosterhout et al., 2004; Van Oosterhout et al., 2006) was

116   used to identify null alleles and potential scoring errors from stuttering or large allele dropout.

117   There was no significant evidence of null alleles ($p > 0.05$) in the Zetterberg Preserve and

118   SBDNL populations. However, null alleles were suggested for loci BB_3968, BB_5021, and

119   BB_8681 in the Petoskey State Park population (Table 2). Homozygote excess for Petoskey

120   State Park is not surprising, given this population's reduced number of alleles at each locus and

121   small comparative population size. We characterized genetic diversity by examining the number

122   of alleles, and expected and observed heterozygosity for each locus averaged over each

123   population (Table 2) using the package STRATAG in the R statistical program (Archer et al.,

124   2016). The number of observed alleles ranged from $1 - 10$. Some loci were monomorphic for

125   one population, but polymorphic when analysis included all populations (e.g., BB_4258).

126   The Zetterberg Preserve population displayed slightly higher heterozygosity values than

127   Sleeping Bear Dunes, but the Petoskey State Park population had much lower heterozygosity

128   values in comparison. A probability test for Hardy-Weinberg Equilibrium (HWE), calculation of

129   the fixation index (F_{IS}), and linkage disequilibrium were performed in GENEPOP 4.2 (Raymond

130   and Rousset, 1995; Rousset, 2008). The default parameters for Markov Chain Monte Carlo

131   (MCMC) iterations were used to calculate HWE. All loci were in HWE except locus BB_3968

132   for SBDNL, and locus BB_21680 for Petoskey State Park (Table 2). The F_{IS} estimates were

133   calculated using the probability model following Robertson and Hill (1984). Statistical tests for

134   genetic linkage disequilibrium were performed using the log likelihood ratio statistic (G-test) and

135   MCMC algorithm by Raymond and Rousset (1995). Two pairs of loci were significantly out of

136   linkage disequilibrium ($p < 0.05$) for both Zetterberg Preserve and Sleeping Bear Dunes:

137   BB_5021 and BB_2888, and BB_3913 and BB_1355. Out of 16 loci, five successfully amplified

138   in a related species *G. elegans* (BB_4443, BB_4258, BB_7213, BB_5151, BB_1355) (Table 3).

139

CONCLUSIONS

141   The 16 microsatellite primers developed for *G. paniculata* provide a tool for estimating

142   genetic diversity and structure of invasive populations, which will aid in understanding its

143   invasion history, identifying source populations, and examining dispersal patterns. Though we

144   developed these markers to study the Lake Michigan dune system invasion, it is invasive

145   throughout North America. With these markers, we can begin to understand the invasion of *G.*

146   *paniculata* in North America in order to improve management efforts and prevent the further

147   spread of this species.

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162 **LITERATURE CITED**

163 Archer, F.I., P.E. Adams, and B.B. Schneiders. 2016. STRATAG: An R package for

164 manipulating, summarizing and analysing population genetic data. *Molecular Ecology*

165 *Resources.* 17: 5–11.

166 Calistri, E., M. Buiatti, and P. Bogani. 2014. Characterization of Gypsophila species and

167 commercial hybrids with nuclear whole-genome and cytoplasmic molecular markers.

168 *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology*.

169 150: 1–11.

170 Darwent, A.L. 1975. The biology of Canadian weeds. 14. *Gypsophila paniculata* L. *Canadian*

171 *Journal of Plant Science.* 55: 1049–1058.

172 Darwent, A.L., and R.T. Coupland. 1966. Life history of *Gypsophila paniculata*. *Weeds.* 14:

173 313–318.

174 Emery, S.M., and P.J. Doran. 2013. Presence and management of the invasive plant *Gypsophila*

175 *paniculata* (baby's breath) on sand dunes alters arthropod abundance and community

176 structure. *Biological Conservation.* 161: 174–181.

177 Faircloth, B.C. 2008. Msatcommander: Detection of microsatellite repeat arrays and

178 automated, locus-specific primer design. *Molecular Ecology Resources*. 8: 92–94.

179 Freeland, J.R., S.D. Petersen, and H. Kirk. 2011. Molecular Ecology, 2nd ed. Wiley-Blackwell,

180 Hoboken, New Jersey, USA.

181 Raymond, M., and F. Rousset. 1995. GENEPOP (version 4.2): Population genetics software for

182 exact tests and ecumenicism. *Journal of Heredity*. 86: 248–249.

183 Robertson, A., and W.G. Hill. 1984. Deviations from Hardy-Weinberg proportions: sampling

184 variances and use in estimation of inbreeding coefficients. *Genetics.* 107: 703–718.

185   Rousset, F. 2008. Genepop'007: a complete reimplementation of the Genepop software for

186   Windows and Linux. *Molecular Ecology Resources.* 8: 103–106.

187   Sievers, F., A. Wilm, D.G. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez et al. 2011. Fast,

188   scalable generation of high-quality protein multiple sequence alignments using Clustal

189   Omega. *Molecular Systems Biology.* 7: 539.

190   The Nature Conservancy, and partners. 2013. Lake Michigan Coastal Dunes Restoration Report.

191   The Nature Conservancy. Website

192   https://www.nature.org/ourinitiatives/regions/northamerica/unitedstates/michigan/lake-

193   michigan-dune-report-2013.pdf [accessed 01 November 2017].

194   Van Oosterhout, C., W.F. Hutchinson, D.P.M. Wills, and P. Shipley. 2004. MICRO-CHECKER:

195   software for identifying and correcting genotyping errors in microsatellite data.

196   *Molecular Ecology Resources.* 4: 535–538.

197   Van Oosterhout, C., D. Weetman, and W.F. Hutchinson. 2006. Estimation and adjustment of

198   microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Resources.*

199   6: 255–256.

200

201

202

203

204

205

206

207

208 **DATA ACCESSIBILITY**

209   A summary of the microsatellite library development and sequence analysis protocols

210   (unpublished data) provided to us by Cornell University, Department of Ecology and

211   Evolutionary Biology are in Appendix S1. Fasta sequences for the 16 microsatellite primers

212   developed here are in Appendix S5. The fasta file listing all identified contigs containing

213   microsatellite regions are in Appendix S4. Potential primer pairs for the identified microsatellite-

214   containing contigs are in Appendix S2. The 107 *G. paniculata* – specific primer pairs tested

215   during primer optimization are in Appendix S3. Raw sequence files for the microsatellite library

216   have been deposited to NCBI's Short Read Archive (Bioproject No: PRJNA431197) and

217   microsatellite sequences have been deposited to GenBank (Table 1). Voucher specimen for each

218   population have been deposited into the Grand Valley State University Herbarium (GVSC),

219   Grand Valley State University Department of Biology, Allendale, MI, USA. (Appendix S6).

220

221

222

223

224

225

226

227

228

229

230

231 **TABLES**

232

233

Table 1. Characteristics of 16 nuclear microsatellite loci developed for *Gypsophila paniculata.*

Locus[a]	Primer sequences (5' - 3')	Repeat motif	Allele size range (bp)	Annealing temperature (℃)	Fluorescent label	GenBank Accession No.
BB_3335[b]	F: TCCACCAAACTCTTAAACTGCC R: CACAGACACAAAGGATCCAACC	$(AGG)_5$	215 - 244	62	NED	MH704701
BB_3913	F: GGCTGTCGGGTAATAAACACAG R: TCCCAACTCAAGTCATAGCCTAG	$(ACAG)_5$	159 - 171	62	PET	MH704702
BB_5567	F: GGCTAGGGAAAGTAGGAAGACC R: CGTGTCCTGTTTCTCCATGATC	$(AAT)_5$	198 - 222	62	VIC	MH704703
BB_4443	F: TAGGGTGGGTGCTTGTACTAAC R: AAAGTGGTGCTGCAGAAGAATC	$(AAG)_{16}$	171 - 211	62	NED	MH704704
BB_21680	F: ACTACACACAGACTCGATCCTC R: CTTTGATTGTTTGGTGTAAGTTGC	$(AAAG)_5$	199 - 218	62	PET	MH704705
BB_3968	F: CATGGAGGACAATGAGAAGACG R: ACGGTGGTAATGAAGTTTGGTG	$(AGG)_6$	207 - 219	62	FAM	MH704706
BB_1355	F: GCTGATCTTTGTCGTCAGGAAG R: ACTCTAGGTGTTAGGAAGGCAC	$(AAAC)_5$	220 - 224	62	NED	MH704711
BB_5151	F: TCCACCTTATAACTCACCACCC R: TGAGGAAGGATAACAGCTCTCG	$(ACC)_5$	205 - 210	62	PET	MH704712
BB_14751[b]	F: CCTCAAACCCTAACAATGCTCC R: TCAGCCGATCCTCTAACACG	$(AAG)_{12}$	195 - 248	62	FAM	MH704713
BB_4258[b]	F: TCACAAGAGGCCCAATTTCTTC R: ACTTGAACCCGAACCTATACCC	$(AAT)_5$	178 - 195	62	VIC	MH704714
BB_6627	F: CAAACTCAACCAACCAGACACC R: CACCTCAGCAACAACAGAGTG	$(AAAC)_5$	151 - 155	62	FAM	MH704715
BB_5021	F: ATTGTCGGTGGTCATTGGTTTC R: CTTAGTCCGCAGTGTAAACAAAG	$(AC)_8$	162 - 207	62	VIC	MH704707
BB_7213[b]	F: TTGCATTCCCACCATTTCATCC R: AGCCAACCTCGTATTAATTGCC	$(AC)_7$	161 - 248	62	PET	MH704708
BB_2888[b]	F: CTTCATTCATGTACAAGAGCGC R: AGAACTGGCTATGGATCGAAATG	$(AC)_{16}$	219 - 232	63	FAM	MH704709
BB_8681[b]	F: ATCTCCAGTTTCCGTGATTTGC R: TACGTCACAAGAGCTTTCAACC	$(ACC)_8$	204 - 222	62	NED	MH704710
BB_31555	F: TGTATAACTGAGATAACCCAGACG R: TTGTTACCTTGTTCCGGCAAAG	$(AC)_7$	150 - 156	62	VIC	MH704716

[a]Optimal annealing temperature was 62℃ for all loci except for BB_2888, where it was 63 ℃. [b]Used 2.5 mM $MgCl_2$ per sample to amplify locus.

234

235

236

237

238

239

240

Table 2. Genetic properties of the 16 developed nuclear microsatellite loci for 3 populations of *Gypsophila paniculata*.

Locus	Zetterberg Preserve ($n = 30$)					Sleeping Bear Dunes ($n = 33$)					Petoskey State Park ($n = 30$)				
	A	H_O[a]	H_E	F_{IS}	r	A	H_O[a]	H_E	F_{IS}	r	A	H_O[a]	H_E	F_{IS}	r
BB_3335	8	0.70	0.82	0.09	0.07	7	0.80	0.80	-0.01	0.00	3	0.50	0.56	0.02	0.05
BB_3913	4	0.74	0.68	-0.04	-0.06	4	0.59	0.61	0.05	0.00	2	0.13	0.13	-0.06	-0.06
BB_5567	4	0.63	0.60	-0.04	-0.04	4	0.63	0.63	0.01	-0.01	3	0.53	0.62	0.14	0.05
BB_4443	10	0.60	0.64	0.03	0.03	8	0.76	0.77	-0.02	0.00	5	0.80	0.71	-0.06	-0.09
BB_21680	4	0.53	0.56	0.01	0.02	3	0.52	0.56	0.02	0.05	4	0.67*	0.58	0.23	-0.10
BB_3968	3	0.50	0.42	-0.14	-0.11	4	0.35*	0.49	0.15	0.13	2	0.04	0.04	–	0.28
BB_5021	5	0.77	0.73	0.00	-0.05	5	0.54	0.60	0.15	0.02	1	–	–	–	0.30
BB_7213	5	0.63	0.61	-0.05	-0.02	3	0.59	0.58	-0.13	0.04	2	0.37	0.38	0.04	0.01
BB_2888	6	0.90	0.80	-0.11	-0.07	6	0.71	0.82	0.20	0.06	2	0.57	0.50	-0.13	-0.08
BB_8681	4	0.47	0.51	0.23	0.01	4	0.38	0.36	-0.07	-0.05	3	0.41	0.53	0.28	0.17
BB_1355	2	0.17	0.26	0.37	0.13	2	0.14	0.13	-0.07	-0.08	2	0.37	0.51	0.29	0.13
BB_5151	2	0.20	0.18	-0.10	-0.11	2	0.49	0.51	0.11	0.00	2	0.03	0.03	–	-0.02
BB_14751	6	0.79	0.71	-0.06	-0.07	6	0.76	0.76	-0.02	-0.01	3	0.20	0.29	0.21	0.12
BB_4258	3	0.40	0.33	-0.11	-0.22	1	0.00	0.00	–	0.00	1	–	–	–	0.00
BB_6627	2	0.30	0.26	-0.16	-0.16	2	0.46	0.49	-0.03	0.04	1	–	–	–	0.00
BB_31555	4	0.60	0.67	0.07	0.05	4	0.46	0.59	0.11	0.12	2	0.23	0.26	0.10	0.04

Note: n = number of individuals sampled; A = number of alleles; H_O = observed heterozygosity; H_E = expected heterozygosity; F_{IS} = fixation index; r = null allele frequency; – = data not available because only 1 allele present at locus or because $A > 1$ is due to rare allele.

Data based on 3 populations with the following geographic coordinates: Zetterberg Preserve = 44.68231, -86.25322; Sleeping Bear Dunes =44.87372, -86.06170; Petoskey State Park = 44.40859, -84.91238. All 3 populations are located in the northwest region of Michigan's lower peninsula.

[a]Significant deviation from Hardy–Weinberg equilibrium: *P < 0.05.

241

bioRxiv preprint doi: https://doi.org/10.1101/401398; this version posted August 27, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.