New microsatellite marker tools for genotype identification and analyses of genetic relationships in two ornamentals, the popular common lilac (*Syringa vulgaris*) and the invasive garden escapee Himalayan balsam (*Impatiens glandulifera*)

Helena Korpelainen*, Maria Pietiläinen

Department of Agricultural Sciences, Viikki Plant Science Centre, P.O. Box 27 (Latokartanonkaari 5), FI-00014 University of Helsinki, Finland

Keywords: Genetic fingerprinting, Marker development by sequencing, *Impatiens*, *Syringa*, SSR markers

* Corresponding author. E-mail address: helena.korpelainen@helsinki.fi

ABSTRACT

Nowadays, high-throughput sequencing technologies are widely available. Yet, it is practical to have an access to simpler and cheaper, yet effective low-throughput analyses as well. For that purpose, species-specific microsatellites, also called simple sequence repeats (SSR), are valuable, multi-purpose types of markers. In the present study, we introduce new sets of SSR markers for two ornamental plant species, the popular common lilac (*Syringa vulgaris* L.) (16 markers) and the invasive garden escapee Himalayan balsam (*Impatiens glandulifera* Royle) (259 markers). The markers were developed as a by-product of a genotyping-by-sequencing project producing a large amount of DNA sequence data. Both the frequency of SSRs and the success rate for marker development were considerably greater in *I. glandulifera* when compared to *S. vulgaris*. The new markers will contribute to the characterization of germplasm and to other types of genetic analyses on these two species.

1. Introduction

DNA markers are a powerful tool for assessing genetic diversity, differences and characteristics within a species, and for the identification and fingerprinting of genotypes. Practical applications range from marker-assisted breeding to environmental conservation, parentage analyses and forensic investigations. Despite a wealth of available DNA information, there are numerous species with lacking or a narrow range of adequate genetic markers. Nowadays, high-throughput sequencing can be applied to most everything, but it is practical to have an access to simpler and cheaper, yet effective low-throughput analyses as well. For that purpose, species-specific microsatellites, also called simple sequence repeats (SSR) and composed of short tandem repeats, are valuable, multi-purpose types of markers. In the present study, we will introduce new sets of SSR markers for two ornamental plant species, the popular common lilac (*Syringa vulgaris* L.) and the invasive garden escapee Himalayan balsam (*Impatiens glandulifera* Royle). At the present,

there are only 9 and 11 species-specific SSR markers available for *S. vulgaris* (JuntheikkiPalovaara et al., 2013) and for *I. glandulifera* (Provan et al., 2007; Walker et al., 2009), respectively.

S. vulgaris (Oleaceae) is a popular ornamental shrub or a small tree, native to the Balkans (Lack, 2001). It was first cultivated in Central Europe in the sixteenth century, but it quickly found its way to gardens in Western Europe and later to gardens in temperate regions all over the world (Lack, 2001). The precise identity of *S. vulgaris* accessions may be unknown and difficult to solve. Therefore, breeders, producers, retailers and consumers will benefit from the development of efficient molecular markers for a precise identification and characterization of cultivars. While *S. vulgaris* is a popular ornamental, *I. glandulifera* (Balsaminaceae) is a tall annual plant, originating from the Himalayas and presently considered an invasive plant that grows rapidly and spreads effectively. It was introduced to Europe in 1839 as a garden ornamental (Beerling and Perrins, 1993). Thereafter, *I. glandulifera* has spread widely throughout Europe (Beerling and Perrins, 1993), and it occurs also in North America and New Zealand as an invasive plant (Weber, 2003). Despite its beautiful flowers, the main interest is not to cultivate but to control and eradicate *I. glandulifera* populations. Its success may be due to its previous popularity as an ornamental garden plant, its rapid growth rate, good ability to survive heavy frost and high seed production (Perrins et al., 1993). Population genetic studies utilizing DNA markers may provide better understanding about the drivers of dynamic invasion processes and about the evolutionary opportunities that invasive plants have acquired. Such knowledge combined with life history studies can aid management practices.

2. Materials and methods

In all, 85 *S. vulgaris* genotypes (additional 9 samples failed in the analyses) including international reference samples, historical samples from Finland, Sweden and France, and unidentified cultivars from Finland were exposed to genotyping-by-sequencing (GBS) to produce

DNA sequence data and to perform simultaneously analyses based on SNP and SilicoDArT markers to allow a genome-wide study on genetic diversity and differentiation (Korpelainen et al., in preparation). SNP markers are nucleotide polymorphisms present in the tag sequences, while SilicoDArT markers represent presence/absence variation (PAV) in the tag sequences. Comparably, 84 *I. glandulifera* samples (additional 10 samples failed in the analyses) originating from Finland, England, Canada, India and Pakistan were subjected to similar analyses as *S. vulgaris* (Korpelainen and Pietiläinen, in preparation). At the same time, the sequence data obtained from these two species allowed the discovery of microsatellite repeat regions and primer development to generate sets of SSR (microsatellite) markers for further use in simpler and cheaper low-throughput analyses on *Syringa* and *Impatiens*.

Prior to SNP and SilicoDArT analyses, genomic DNA was extracted from leaf tissue using the CTAB protocol of Doyle and Doyle (1990) or a commercial kit (E.Z.N.A.[™] Plant DNA Mini Kit Spin Protocol (Omega Bio-Tek). The quality and quantity of extracted DNA were quantified with a spectrophotometer and further confirmed on 0.8% agarose gels. DNA concentrations were adjusted at 50 ng μl^{-1} . DNA samples of *S. vulgaris* and *I. glandulifera* were sent to Diversity Arrays Technology Pty Ltd. (Canberra, Australia; <http://www.diversityarrays.com>) for high-throughput sequencing utilizing the DArTseq system and for marker discovery (Kilian et al., 2012).

The sequence data consisted of 69 bp long pieces of data. All available sequences were used to explore for SSR repeats using MSATFINDER version 2.0.9. (Thurston and Field, 2005) combined with visual observations with the following criteria: mononucleotide repeats with at least eight copies and other types of repeats with a minimum of five copies. All SSRs fulfilling the criteria were recorded. The Primer3 software (Rozen and Skaletsky, 2000) was used to design primer pairs for the potential amplification of the discovered SSRs when possible (SSR located at a relatively

middle position of the tag sequence containing a SSR, and the primers being 18-27 bp in length and with the GC content 40-60%

3. Results and discussion

The total length of the sequenced area obtained in SNP and DArTseq analyses equalled 1 035 552 bp and 740 508 bp, respectively, in *S. vulgaris*, and 1 447 827 bp and 2 044 125 bp, respectively, in *I. glandulifera*. The numbers of SSR regions found in *S. vulgaris* based on SNP and DArTseq analyses equalled 111 (repeat motifs: 63.1% mononucleotide, 31.5% dinucleotide and 5.4% compound) and 116 (repeat motifs: 82.8% mononucleotide, 13.8% dinucleotide and 3.4% compound), respectively (Table 1). In *I. glandulifera*, the numbers of SSRs found based on SNP and DArTseq analyses equalled 477 (repeat motifs: 76.9% mononucleotide, 11.9% dinucleotide, 8.4% trinucleotides, 0.2% tetranucleotides, 0.6% pentanucleoties and 1.9% compound) and 709 (repeat motifs: 73.9% mononucleotide, 11.8% dinucleotide, 11.6% trinucleotide, 0.8% tetranucleotide, 0.4% pentanucleotide, 0.1% hexanucleotide and 1.3% compound), respectively (Table 2). Thus, the great majority of all found SSRs were composed of mononucleotide repeats, followed by dinucleotide repeats, while repeat motifs longer than three nucleotides were very rare. The frequency of SSRs relative to the total sequence length was considerably higher in *I. glandulifera* than in *S. vulgaris*, about triple and double in SNP and DArTseq analyses, respectively. The relative frequency of SSRs was slightly higher in SNP-based analyses when compared to DArTseq-based analyses.

We succeeded to design adequate primer pairs for only a portion of the discovered SSRs: in *I. glandulifera*, for 23.5% (112 primer pairs) and 20.7% (147 primer pairs) of SNP- and DArTseqbased SSRs, respectively, while in *S. vulgaris*, the percentages were much lower, only 14.4% (16 primer pairs) and 0% of SSRs, respectively (Tables S1-S2). Thus, both the frequency of

SSRs and the success rate for marker development were considerably greater in *I. glandulifera* when compared to *S. vulgaris*.

Previous DNA-based studies on the genus *Syringa* have involved the use of RAPD markers (Chen et al., 1999; Kochneva et al., 2004; Melnikova et al., 2009), ISSR markers (Rzepka-Plevneš et al., 2006; Yang et al., 2013), AFLP markers (Ming and Gu, 2006), sequencing of nuclear and chloroplast regions (Smolik et al., 2010; Lendvay et al., 2016), and recently the sequencing of the whole chloroplast genome of *S. pinnatifolia* (Zhang et al., 2019). More effective and reliable SSR markers for cultivar identification in *S. vulgaris* have been developed by Juntheikki-Palovaara et al. (2013, nine microsatellites). In addition, Lendvay et al. (2013) have developed five microsatellites for *S. josikaea*, and De La Rosa et al. (2002) have reported that some markers developed for olive (*Olea europaea* L.) amplified also in the genus *Syringa*. Although these markers appeared valuable for detecting differentiation among cultivars, the availability of powerful markers has been limited to allow high-precision cultivar identification. The 16 SSRs introduced in the present study are a useful addition to the 9 species-specific markers previously available for *S. vulgaris* (Juntheikki-Palovaara et al., 2013). They will contribute to the characterization of germplasm and to other types of genetic analyses in *S. vulgaris* and possibly in other species of the genus *Syringa*.

Considerable research attention has been paid on *I. glandulifera*, which is a major invasive plant that spreads effectively. Previously, Provan et al. (2007) have developed eight SSR markers for *I. glandulifera*, and Walker et al. (2009) additional three markers. These markers have been used in population genetic studies by Love et al. (2013), Hagenblad et al (2015), Nagy and Korpelainen (2015), and Helsen et al. (2019). *I. glandulifera* has also been the target of RAPD analyses (Zybartaite et al., 2011; Kupcinskiene et al., 2015) and ISSR analyses (Kupcinskiene et al., 2015). The new SSRs (a total of 259 markers) are a huge increase in the selection of markers that could be used in different types of genetic studies on *I. glandulifera*. However, testing the true efficiency of each marker in genetic analyses is beyond the present short communication paper.

146

147 Acknowledgements

148 The project was supported by the Nikolai and Ljudmila Borisoff Foundation. We
149 would like to thank Juulia Kotkavuori and Leena Lindén for advice and help in sampling materials.

150

151 Supplementary data

152 Supplementary material related to this article: Table S1 and Table S2.

153

154 References

- 155 Beerling, D.J., Perrins, J.M., 1993. Biological flora of the British Isles. *Impatiens glandulifera*
156 Royle (*Impatiens roylei* Walp.). J. Ecol. 81, 367-382.
- 157 Chen, X., Chen, Z., Han, J. 1999. Using random amplified polymorphic DNA (RAPD) markers for
158 lilac genetic analysis and classification in lilac cultivars. Acta Botanica Boreali-occidentalia
159 Sinica 19, 169-176.
- 160 De La Rosa, R., James, C.M., Tobutt, K.R., 2002. Isolation and characterization of polymorphic
161 microsatellites in olive (*Olea europaea* L.) and their transferability to other genera in the
162 Oleaceae. Mol. Ecol. Res. 2, 265-267. <https://doi.org/10.1046/j.1471-8286.2002.00217.x>.
- 163 Doyle, J.J., Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. Focus 12, 13-15.
- 164 Hagenblad, J., Hülskötter, J., Acharya, K.P., Brunet, J., Chabrierie, O., Cousins, S.A.O., Dar, P.A.,
165 Diekmann, M., De Frenne, P., Hermy, M., Jamoneau, A., Kolb, A., Lemke, I., Plue, J., Reshi,
166 Z.A., Graae, B.J., 2015. Low genetic diversity despite multiple introductions of the invasive
167 plant species *Impatiens glandulifera* in Europe. BMC Genet. 16, 103.
168 <https://doi.org/10.1186/s12863-015-0242-8>
- 169 Helsen, K., Hagenblad, J., Acharya, K.P., Brunet, J., Cousins, S.A.O., Decocq, G., De Frenne, P.,

170 Kimberley, A., Kolb, A., Michaelis, J., Plue, J., Verheyen, K., Speed, J.D.M., Graae, B.J., 2019.
 171 No genetic erosion after five generations for *Impatiens glandulifera* populations across the
 172 invaded range in Europe. BMC Genet. 20, 20. <https://doi.org/10.1186/s12863-019-0721-4>
 173 Juntheikki-Palovaara, I., Antonius, K., Lindén, L., Korpelainen, H., 2013. Microsatellite markers
 174 for common lilac (*Syringa vulgaris* L.). Plant Gen. Res. 11, 279-282.
 175 <https://doi.org/10.1017/S1479262113000166>
 176 Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., Caig, V., Heller-Uszynska, K.,
 177 Jaccoud, D., Hopper, C., Aschenbrenner-Kilian, M., Margaret Evers, M., Peng, K., Cayla, C.,
 178 Hok, P., Uszynski, G., 2012. Diversity arrays technology: a generic genome profiling
 179 technology on open platforms. Methods Mol. Biol. 888, 67–89. <https://doi.org/10.1007/978-1->
 180 [61779-8702_5](https://doi.org/10.1007/978-1-61779-8702_5).
 181 Kochneva, E.Z., Ryzhova, N.N., Molkanova, O.I., Kudryavtsev, A.M., Upelniek, V.P., Okuneva,
 182 I.B., 2004. The genus *Syringa*: molecular markers of species and cultivars. Russian J. Genet. 40,
 183 30–32.
 184 Kupcinskiene, E., Zybartaite, L., Paulauskas, A., 2015. Comparison of genetic diversity of three
 185 *Impatiens* species from Central Europe and Baltic region. Žemdirbystė-Agriculture 102, 87–94.
 186 <https://doi.org/10.13080/z-a.2015.102.011>.
 187 Lack, H.W., 2001. Lilac and horse-chestnut: discovery and rediscovery. Bocconea 13, 613-616.
 188 Lendvay, B., Kadereit, J.W., Westberg, E., Cornejo, C., Pedryc, A., Höhn, M., 2016.
 189 Phylogeography of *Syringa josikaea* (Oleaceae): early Pleistocene divergence from East Asian
 190 relatives and survival in small populations in the Carpathians, Biol. J. Linn. Soc. 119, 689–703.
 191 <https://doi.org/10.1111/bij.12499>.
 192 Lendvay, B., Pedryc, A., Höhn, M., 2013. Characterization of nuclear microsatellite markers for the
 193 narrow endemic *Syringa josikaea* Jacq. fil. ex Rchb. Notulae Botanicae Horti Agrobotanici Cluj-
 194 Napoca 41: 301-305. <https://doi.org/10.15835/nbha4119023>.

195 Love, H.M., Maggs, C.A., Murray, T.E., Provan, J., 2013. Genetic evidence for predominantly
196 hydrochoric gene flow in the invasive riparian plant *Impatiens glandulifera* (Himalayan balsam),
197 Ann. Bot. 112, 1743–1750. <https://doi.org/10.1093/aob/mct227>.

198 Melnikova, N.V., Borhert, E.V., Martynov, S.P., Okuneva, I.B., Molkanova, O.I., Up.lniek, V.P.,
199 Kudryavtsev, A.M. 2009. Molecular genetic marker-based approaches to the verification of lilac
200 *Syringa vulgaris* L. in vitro germplasm collections. Russian J. Genet. 45, 85–90.

201 Ming, J., Gu, W.C., 2006. Genetic diversity in natural populations of *Syringa oblata* detected by
202 AFLP markers. Acta Horticulturae Sinica 33, 1269–1274.

203 Nagy, A.-M., Korpelainen, H. 2015. Population genetics of Himalayan balsam (*Impatiens*
204 *glandulifera*): comparison of native and introduced populations, Plant Ecol. Div. 8, 317-321.
205 <https://doi.org/10.1080/17550874.2013.863407>.

206 Perrins, J., Fitter, A., Williamson, M., 1993. Population biology and rates of invasion of three
207 introduced *Impatiens* species in the British Isles. J. Biogeogr. 20, 33-44.

208 Provan, J., Love, H.M., Maggs, C.A., 2007. Development of microsatellites for the invasive
209 riparian plant *Impatiens glandulifera* (Himalayan balsam) using intersimple sequence repeat
210 cloning. Mol. Ecol. Notes 7, 451-453. <https://doi.org/10.1111/j.1471-8286.2006.01614.x>.

211 Rozen, S., Skaletsky, H.J., 2000. Primer3 on the WWW for general users and for biologist
212 programmers, in: Misener, S., Krawetz, S.A. (Eds.), Methods in Molecular Biology, vol. 132:
213 Bioinformatics methods and protocols. Humana Press, Totowa, New Jersey, USA, pp. 365–386.

214 Rzepka-Plevneš D., Smolik M., Tanska K., 2006. Genetic similarity of chosen *Syringa* species
215 determined by the ISSR-PCR technique. Dendrobiology 56, 61-67.

216 Smolik, M., Andrys, D., Franas, A., Krupa-Mańkiewicz, M., Malinowska, K., 2010. Polymorphism
217 in *Syringa* rDNA regions assessed by PCR technique. Dendrobiology 64, 55-64.

218 Thurston, M. I., Field, D., 2005. Msatfinder: detection and characterisation of microsatellites.
219 Distributed by the authors at <http://www.genomics.ceh.ac.uk/msatfinder/>. CEH Oxford,

220 Mansfield Road, Oxford OX1 3SR.

221 Walker, N.F., Hulme, P.E., Hoelzel, A.R., 2009. Population genetics of an invasive riparian species

222 *Impatiens glandulifera*. Plant Ecol. 203, 243-252. <https://doi.org/10.1007/s11258-008-9540-9>.

223 Weber, E., 2003. Invasive Plant Species of the World: a Reference Guide to Environmental Weeds.

224 CABI, Wallingford, UK.

225 Yang, Y.-J., Li J.-N., Gong, L., Zhou, J., 2013. Clonal structure and genetic diversity of natural

226 *Syringa pinnatifolia* var. *alanshanica* assessed by ISSR. Plant Sci. J. 31, 85-92.

227 <https://doi.org/10.3724/SP.J.1142.2013.10085>.

228 Zhang, J., Jiang, Z., Su, H., Zhao, H., Cai, J., 2019. The complete chloroplast genome sequence of

229 the endangered species *Syringa pinnatifolia* (Oleaceae). Nordic J. Bot., e02201.

230 <https://doi.org/10.1111/njb.02201>

231 Zybartaitė, L., Zukauskienė, J., Jodinskienė, M., Janssens, S.B., Paulauskas, A., Kupcinskiene, E.,

232 2011. RAPD analysis of genetic diversity among Lithuanian populations of *Impatiens*

233 *glandulifera*. Žemdirbystė-Agriculture 98, 391-398.

234

235

Table 1
Numbers and percentages of different microsatellite repeats found in 1 035 552 bp and 740 508 bp of DNA sequences obtained from SNP and SilicoDArT marker analyses, respectively, in *Syringa vulgaris*.

SNP analysis			SilicoDArT analysis		
Repeat motif	Number	% repeats	Repeat motif	Numbers	% repeats
A ₈ /T ₈	54	48.6	A ₈ /T ₈	52	44.8
A ₉ /T ₉	5	4.5	A ₉ /T ₉	22	19.0
A ₁₀ /T ₁₀	4	3.6	A ₁₀ /T ₁₀	14	12.1
A ₁₁ /T ₁₁	2	1.8	A ₁₁ /T ₁₁	5	4.3
A ₁₃	1	0.9	A ₁₂ /T ₁₂	2	1.7
A ₁₅	1	0.9	G ₉	1	0.9
T ₁₆	1	0.9	[A/G] ₂₂	1	0.9
T ₁₇	1	0.9	[AC] ₆	2	1.7
T ₁₈	1	0.9	[AC] ₁₀	1	0.9
[AG] ₅	1	0.9	[AG] ₆	1	0.9
[AT] ₅	1	0.9	[CA] ₁₀	1	0.9
[AT] ₆	1	0.9	[CG] ₅	1	0.9
[AC] ₅	2	1.8	[CT] ₈	1	0.9
[AG] ₁₀	1	0.9	[GA] ₈	1	0.9
[AG] ₁₅	1	0.9	[GA] ₁₃	1	0.9
[CA] ₅ /[TG] ₅	7	6.3	[GT] ₅	1	0.9
[CA] ₆	2	1.8	[GT] ₇	1	0.9
[GA] ₅	4	3.6	[TC] ₅	3	2.6
[GA] ₆	3	2.7	[TC] ₆	1	0.9
[GA] ₈	1	0.9	[TG] ₅	1	0.9
[GA] ₁₂	1	0.9	[GT] ₅ [GA] ₄	1	0.9
[GA] ₁₄	1	0.9	T ₄ CT ₈	1	0.9
[GA] ₃₀	1	0.9	G ₉ AG ₅	1	0.9
[TC] ₅	7	6.3			
[TC] ₁₁	1	0.9			
T ₈ CT ₄	1	0.9			
A ₈ TA ₇	1	0.9			
A ₅ N ₂ A ₄ NA ₈	1	0.9			
A ₂ TA ₈ TA ₃	1	0.9			
A ₆ TA ₈	1	0.9			
T ₆ N ₆ T ₈	1	0.9			
Total	111		Total	116	

Table 2
Numbers and percentages of different microsatellite repeats found in 1 447 827 bp and 2 044 125 bp of DNA sequences obtained from SNP and SilicoDArT analyses, respectively, in *Impatiens glandulifera*.

SNP analysis			SilicoDArT analysis		
Repeat motif	Numbers	% repeats	Repeat motif	Numbers	% repeats
A ₈ /T ₈	182	38.2	A ₈ /T ₈	208	29.3
A ₉ /T ₉	91	19.1	A ₉ /T ₉	148	20.8
A ₁₀ /T ₁₀	44	9.2	A ₁₀ /T ₁₀	92	13.0
A ₁₁ /T ₁₁	22	4.6	A ₁₁ /T ₁₁	38	5.4
A ₁₂ /T ₁₂	15	3.1	A ₁₂ /T ₁₂	15	2.1
A ₁₃ /T ₁₃	4	0.8	A ₁₃ /T ₁₃	9	1.3
A ₁₄	3	0.6	A ₁₄	3	0.4
A ₁₅	2	0.4	A ₁₅	1	0.1
A ₁₇	1	0.2	A ₁₆	2	0.3
T ₁₆	1	0.2	A ₁₇	2	0.3
T ₂₈	1	0.2	A ₁₈	3	0.4
G ₁₀	1	0.2	C ₈ /G ₈	2	0.3
[CT] ₅	4	0.8	G ₁₀	1	0.1
[CT] ₈	4	0.8	[AT] ₆	1	0.1
[CT] ₉	4	0.8	[AT] ₇	1	0.1
[CA] ₇	1	0.2	[CA] ₁₁	1	0.1
[CA] ₈	1	0.2	[CA] ₁₅	1	0.1
[GA] ₅	5	1.0	[CT] ₅	4	0.6
[GA] ₆	2	0.4	[CT] ₆	2	0.3
[GA] ₇	1	0.2	[CT] ₇	1	0.1
[GA] ₉	2	0.4	[CT] ₈	3	0.4
[GA] ₁₀	1	0.2	[CT] ₉	2	0.3
[GA] ₁₄	1	0.2	[CT] ₁₀	1	0.1
[GT] ₆	2	0.4	[GA] ₅ /[TC] ₅	19	2.7
[GT] ₈	1	0.2	[GA] ₆ /[TC] ₆	11	1.5
[TA] ₅	17	3.6	[GA] ₇	4	0.6
[TA] ₆	4	0.8	[GA] ₈	2	0.3
[TA] ₇	3	0.6	[GA] ₉	4	0.6
[TC] ₅	1	0.2	[GA] ₁₀	1	0.1
[TC] ₆	2	0.4	[GA] ₁₄	1	0.1
[TG] ₈	1	0.2	[GT] ₅	1	0.1
[ATT] ₅	1	0.2	[GT] ₆	2	0.3
[CAA] ₅	2	0.4	[TA] ₅	10	1.4
[CAA] ₇	1	0.2	[TA] ₆	5	0.7
[CAAAA] ₅	1	0.2	[TA] ₇	2	0.3
[CAAT] ₅	1	0.2	[TA] ₈	3	0.4
[CAT] ₅	3	0.6	[TG] ₆	1	0.1
[CAT] ₆	2	0.4	[TG] ₇	1	0.1

[CAT] ₈	1	0.2	[ATT] ₅	2	0.3
[CCG] ₆	1	0.2	[ATT] ₆	2	0.3
[CTT] ₅	5	1.0	[ATTTTT] ₅	1	0.1
[CTT] ₆	2	0.4	[ATCT] ₅	1	0.1
[CTTCT] ₅	1	0.2	[CAA] ₅	1	0.1
[CTTTT] ₅	1	0.2	[CAA] ₆	1	0.1
[GAA] ₅	5	1.0	[CAAAA] ₅	2	0.3
[GAG] ₅	1	0.2	[CATA] ₅	1	0.1
[GAT] ₅	4	0.8	[CAT] ₅	5	0.7
[GAT] ₇	1	0.2	[CAT] ₆	5	0.7
[GAT] ₁₁	1	0.2	[CAT] ₇	2	0.3
[GCC] ₇	1	0.2	[CAT] ₈	1	0.1
[GCG] ₅	1	0.2	[CCG] ₆	1	0.1
[GCT] ₅	1	0.2	[CTT] ₅	7	1.0
[TAA] ₆	1	0.2	[CTT] ₇	2	0.3
[TATC] ₅	1	0.2	[CTT] ₈	1	0.1
[TGC] ₆	1	0.2	[CTT] ₁₀	1	0.1
[TTA] ₅	2	0.4	[CGG] ₅	1	0.1
[TTA] ₆	1	0.2	[CGG] ₆	1	0.1
[TTA] ₇	2	0.4	[GAA] ₅	15	2.1
[A/G] ₁₅	1	0.2	[GAA] ₆ /[TTC] ₆	8	1.1
[A/G] ₁₆	1	0.2	[GAA] ₇	1	0.1
[A/G] ₁₇	1	0.2	[GAA] ₈	2	0.3
[A/G] ₂₇	1	0.2	[GAAA] ₅	2	0.3
[A/G] ₃₁	1	0.2	[GAG] ₅	4	0.6
[A/C] ₁₈	1	0.2	[GAT] ₅	1	0.1
[A/T] ₁₇	1	0.2	[GAT] ₆	2	0.3
[G/T] ₁₉	1	0.2	[GAT] ₈	1	0.1
[A/T] ₁₈ GT ₆	1	0.2	[GAT] ₉	1	0.1
			[GAT] ₁₁	1	0.1
			[GCA] ₇	1	0.1
			[GCC] ₅	2	0.3
			[GCT] ₅	1	0.1
			[GGC] ₆	1	0.1
			[GGTTT] ₅	1	0.1
			[GTTT] ₅	1	0.1
			[TAA] ₇	1	0.1
			[TAT] ₇	1	0.1
			[TGG] ₆	1	0.1
			[TTC] ₅	2	0.3
			[TTA] ₅	2	0.3
			[TTA] ₉	1	0.1
			[TTTA] ₅	1	0.1
			[A/G] ₁₉	1	0.1
			[A/G] ₂₅	1	0.1
			[A/G] ₂₇	1	0.1
			[A/C] ₁₆	1	0.1
			[C/T] ₂₉	1	0.1