Short communication

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, PO 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

Running title: genetic markers for *Syringa* and *Impatiens*

Declarations of interest: none

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* (Juntheikki-Palovaara 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

74 unidentified cultivars from Finland were exposed to genotyping-by-sequencing (GBS) to produce
75 DNA sequence data and to perform simultaneously analyses based on SNP and SilicoDArT markers
76 to allow a genome-wide study on genetic diversity and differentiation (Korpelainen et al., in
77 preparation). SNP markers are nucleotide polymorphisms present in the tag sequences, while
78 SilicoDArT markers represent presence/absence variation (PAV) in the tag sequences. Comparably,
79 84 *I. glandulifera* samples (additional 10 samples failed in the analyses) originating from Finland,
80 England, Canada, India and Pakistan were subjected to similar analyses as *S. vulgaris* (Korpelainen
81 and Pietiläinen, in preparation). At the same time, the sequence data obtained from these two
82 species allowed the discovery of microsatellite repeat regions and primer development to generate
83 sets of SSR (microsatellite) markers for further use in simpler and cheaper low-throughput analyses
84 on *Syringa* and *Impatiens*.
85 Prior to SNP and SilicoDArT analyses, genomic DNA was extracted from leaf tissue using the
86 CTAB protocol of Doyle and Doyle (1990) or a commercial kit (E.Z.N.A.™ Plant DNA Mini Kit
87 Spin Protocol (Omega Bio-Tek). The quality and quantity of extracted DNA were quantified with a
88 spectrophotometer and further confirmed on 0.8% agarose gels. DNA concentrations were adjusted
89 at 50 ng μl^{-1}. DNA samples of *S. vulgaris* and *I. glandulifera* were sent to Diversity Arrays
90 Technology Pty Ltd. (Canberra, Australia; http://www.diversityarrays.com) for high-throughput
91 sequencing utilizing the DArTseq system and for marker discovery (Kilian et al., 2012).
92 The sequence data consisted of 69 bp long pieces of data. All available sequences were used to
93 explore for SSR repeats using MSATFINDER version 2.0.9. (Thurston and Field, 2005) combined
94 with visual observations with the following criteria: mononucleotide repeats with at least eight
95 copies and other types of repeats with a minimum of five copies. All SSRs fulfilling the criteria
96 were recorded. The Primer3 software (Rozen and Skaletsky, 2000) was used to design primer pairs
97 for the potential amplification of the discovered SSRs when possible (SSR located at a relatively

98 middle position of the tag sequence containing a SSR, and the primers being 18-27 bp in length and
99 with the GC content 40-60%
100
101 **3. Results and discussion**
102 The total length of the sequenced area obtained in SNP and DArTseq analyses equalled
103 1 035 552 bp and 740 508 bp, respectively, in S. *vulgaris*, and 1 447 827 bp and 2 044 125 bp,
104 respectively, in *I. glandulifera*. The numbers of SSR regions found in *S. vulgaris* based on SNP
105 and DArTseq analyses equaled 111 (repeat motifs: 63.1% mononucletide, 31.5% dinucleotide and
106 5.4% compound) and 116 (repeat motifs: 82.8% mononucletide, 13.8% dinucleotide and 3.4%
107 compound), respectively (Table 1). In *I. glandulifera*, the numbers of SSRs found based on SNP
108 and DArTseq analyses equaled 477 (repeat motifs: 76.9% mononucletide, 11.9% dinucleotide,
109 8.4% trinuleotides, 0.2% tetranucleotides, 0.6% pentanucleoties and 1.9% compound) and 709
110 (repeat motifs: 73.9% mononucletide, 11.8% dinucleotide, 11.6% trinucleotide, 0.8%
111 tetranucleotide, 0.4% pentanucleotide, 0.1% hexanucleotide and 1.3% compound), respectively
112 (Table 2). Thus, the great majority of all found SSRs were composed of mononucleotide repeats,
113 followed by dinucleotide repeats, while repeat motifs longer than three nucleotides were very rare.
114 The frequency of SSRs relative to the total sequence length was considerably higher in *I.*
115 *glandulifera* than in *S. vulgaris*, about triple and double in SNP and DArTseq analyses,
116 respectively. The relative frequency of SSRs was slightly higher in SNP-based analyses when
117 compared to DArTseq-based analyses.
118 We succeeded to design adequate primer pairs for only a portion of the discovered SSRs: in *I.*
119 *glandulifera*, for 23.5% (112 primer pairs) and 20.7% (147 primer pairs) of SNP- and DArTseq-
120 based SSRs , respectively, while in *S. vulgaris*, the percentages were much lower, only 14.4% (16
121 primer pairs) and 0% of SSRs, respectively (Tables S1-S2). Thus, both the frequency of SSRs and

122 the success rate for marker development were considerably greater in *I. glandulifera* when
123 compared to *S. vulgaris*.
124 Previous DNA-based studies on the genus *Syringa* have involved the use of RAPD markers
125 (Chen et al., 1999; Kochneva et al., 2004; Melnikova et al., 2009), ISSR markers (Rzepka-Plevneš
126 et al., 2006; Yang et al., 2013), AFLP markers (Ming and Gu, 2006), sequencing of nuclear and
127 chloroplast regions (Smolik et al., 2010; Lendvay et al., 2016), and recently the sequencing of the
128 whole chloroplast genome of *S. pinnatifolia* (Zhang et al., 2019). More effective and reliable SSR
129 markers for cultivar identification in *S. vulgaris* have been developed by Juntheikki-Palovaara et al.
130 (2013, nine microsatellites). In addition, Lendvay et al. (2013) have developed five microsatellites
131 for *S. josikaea*, and De La Rosa et al. (2002) have reported that some markers developed for olive
132 (*Olea europaea* L.) amplified also in the genus *Syringa*. Although these markers appeared valuable
133 for detecting differentiation among cultivars, the availability of powerful markers has been limited
134 to allow high-precision cultivar identification. The 16 SSRs introduced in the present study are a
135 useful addition to the 9 species-specific markers previously available for *S. vulgaris* (Juntheikki-
136 Palovaara et al., 2013). They will contribute to the characterization of germplasm and to other
137 types of genetic analyses in *S. vulgaris* and possibly in other species of the genus *Syringa*.
138 Considerable research attention has been paid on *I. glandulifera*, which is a major invasive plant
139 that spreads effectively. Previously, Provan et al. (2007) have developed eight SSR markers for *I.*
140 *glandulifera*, and Walker et al. (2009) additional three markers. These markers have been used in
141 population genetic studies by Love et al. (2013), Hagenblad et al (2015), Nagy and Korpelainen
142 (2015), and Helsen et al. (2019). *I. glandulifera* has also been the target of RAPD analyses
143 (Zybartaite et al., 2011; Kupcinskiene et al., 2015) and ISSR analyses (Kupcinskiene et al., 2015).
144 The new SSRs (a total of 259 markers) are a huge increase in the selection of markers that could be
145 used in different types of genetic studies on *I. glandulifera*. However, testing the true efficiency of
146 each marker in genetic analyses is beyond the present short communication paper.

Acknowledgements

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

Appendix A. Supplementary data

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

References

Beerling, D.J., Perrins, J.M., 1993. Biological flora of the British Isles. *Impatiens glandulifera* Royle (Impatiens roylei Walp.). J. Ecol. 81, 367-382.

Chen, X., Chen, Z., Han, J. 1999. Using random amplified polymorphic DNA (RAPD) markers for liliac genetic analysis and classification in lilac cultivars. Acta Botanica Boreali-occidentalia Sinica 19, 169-176.

De La Rosa, R., James, C.M., Tobutt, K.R., 2002. Isolation and characterization of polymorphic microsatellites in olive (*Olea europaea* L.) and their transferability to other genera in the Oleaceae. Mol. Ecol. Res. 2, 265-267. https://doi.org/10.1046/j.1471-8286.2002.00217.x.

Doyle, J.J., Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. Focus 12, 13-15.

Hagenblad, J., Hülskötter, J., Acharya, K.P., Brunet, J., Chabrerie, O., Cousins, S.A.O., Dar, P.A., Diekmann, M., De Frenne, P., Hermy, M., Jamoneau, A., Kolb, A., Lemke, I., Plue, J., Reshi, Z.A., Graae, B.J., 2015. Low genetic diversity despite multiple introductions of the invasive plant species *Impatiens glandulifera* in Europe. BMC Genet. 16, 103. https://doi.org/10.1186/s12863-015-0242-8

Helsen, K., Hagenblad, J., Acharya, K.P., Brunet, J., Cousins, S.A.O.,Decocq, G., De Frenne, P.,

Kimberley, A., Kolb, A., Michaelis, J., Plue, J., Verheyen, K., Speed, J.D.M., Graae, B.J., 2019. No genetic erosion after five generations for *Impatiens glandulifera* populations across the invaded range in Europe. BMC Genet. 20, 20. https://doi.org/10.1186/s12863-019-0721-4

Juntheikki-Palovaara, I., Antonius, K., Lindén, L., Korpelainen, H., 2013. Microsatellite markers for common lilac (*Syringa vulgaris* L.). Plant Gen. Res. 11, 279-282. https://doi.org/10.1017/S1479262113000166

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., Caig, V., Heller-Uszynska, K., Jaccoud, D., Hopper, C., Aschenbrenner-Kilian, M., Margaret Evers, M., Peng, K., Cayla, C., Hok, P., Uszynski, G., 2012. Diversity arrays technology: a generic genome profiling technology on open platforms. Methods Mol. Biol. 888, 67–89. https://doi.org/10.1007/978-1-61779-870-2_5.

Kochneva, E.Z., Ryzhova, N.N., Molkanova, O.I., Kudryavtsev, A.M., Upelniek, V.P., Okuneva, I.B., 2004. The genus *Syringa*: molecular markers of species and cultivars. Russian J. Genet. 40, 30–32.

Kupcinskiene, E., Zybartaite, L., Paulauskas, A., 2015. Comparison of genetic diversity of three *Impatiens* species from Central Europe and Baltic region. Žemdirbystė-Agriculture 102, 87−94. https://doi.org/10.13080/z-a.2015.102.011.

Lack, H.W., 2001. Lilac and horse-chestnut: discovery and rediscovery. Bocconea 13, 613-616.

Lendvay, B., Kadereit, J.W., Westberg, E., Cornejo, C., Pedryc, A., Höhn, M., 2016. Phylogeography of *Syringa josikaea* (Oleaceae): early Pleistocene divergence from East Asian relatives and survival in small populations in the Carpathians, Biol. J. Linn. Soc. 119, 689–703. https://doi.org/10.1111/bij.12499.

Lendvay, B., Pedryc, A., Höhn, M., 2013. Characterization of nuclear microsatellite markers for the narrow endemic *Syringa josikaea* Jacq. fil. ex Rchb. Notulae Botanicae Horti Agrobotanici Cluj-Napoca 41: 301-305. https://doi.org/10.15835/nbha4119023.

196 Love, H.M., Maggs, C.A., Murray, T.E., Provan, J., 2013. Genetic evidence for predominantly
197 hydrochoric gene flow in the invasive riparian plant *Impatiens glandulifera* (Himalayan balsam),
198 Ann. Bot. 112, 1743–1750. https://doi.org/10.1093/aob/mct227.
199 Melnikova, N.V., Borhert, E.V., Martynov, S.P., Okuneva, I.B., Molkanova, O.I., Up.lniek, V.P.,
200 Kudryavtsev, A.M. 2009. Molecular genetic marker-based approaches to the verification of lilac
201 *Syringa vulgaris* L. in vitro germplasm collections. Russian J. Genet. 45, 85–90.
202 Ming, J., Gu, W.C., 2006. Genetic diversity in natural populations of *Syringa oblata* detected by
203 AFLP markers. Acta Horticulturae Sinica 33, 1269–1274.
204 Nagy, A.-M., Korpelainen, H. 2015. Population genetics of Himalayan balsam (*Impatiens*
205 *glandulifera*): comparison of native and introduced populations, Plant Ecol. Div. 8, 317-321.
206 https://doi.org/10.1080/17550874.2013.863407.
207 Perrins, J., Fitter, A., Williamson, M., 1993. Population biology and rates of invasion of three
208 introduced *Impatiens* species in the British Isles. J. Biogeogr. 20, 33-44.
209 Provan, J., Love, H.M., Maggs, C.A., 2007. Development of microsatellites for the invasive
210 riparian plant *Impatiens glandulifera* (Himalayan balsam) using intersimple sequence repeat
211 cloning. Mol. Ecol. Notes 7, 451-453. https://doi.org/10.1111/j.1471-8286.2006.01614.x.
212 Rozen, S., Skaletsky, H.J., 2000. Primer3 on the WWW for general users and for biologist
213 programmers, in: Misener, S., Krawetz, S.A. (Eds.), Methods in Molecular Biology, vol. 132:
214 Bioinformatics methods and protocols. Humana Press, Totowa, New Jersey, USA, pp. 365–386.
215 Rzepka-Plevneš D., Smolik M., Tanska K., 2006. Genetic similarity of chosen *Syringa* species
216 determined by the ISSR-PCR technique. Dendrobiology 56, 61-67.
217 Smolik, M., Andrys, D., Franas, A., Krupa-Małkiewicz, M., Malinowska, K., 2010. Polymorphism
218 in *Syringa* rDNA regions assessed by PCR technique. Dendrobiology 64, 55-64.
219 Thurston, M. I., Field, D., 2005. Msatfinder: detection and characterisation of microsatellites.

220 Distributed by the authors at http://www.genomics.ceh.ac.uk/msatfinder/. CEH Oxford,
221 Mansfield Road, Oxford OX1 3SR.
222 Walker, N.F., Hulme, P.E., Hoelzel, A.R., 2009. Population genetics of an invasive riparian species
223 *Impatiens glandulifera*. Plant Ecol. 203, 243-252. https://doi.org/10.1007/s11258-008-9540-9.
224 Weber, E., 2003. Invasive Plant Species of the World: a Reference Guide to Environmental
225 Weeds. CABI, Wallingford, UK.
226 Yang, Y.-J., Li J.-N., Gong, L., Zhou, J., 2013. Clonal structure and genetic diversity of natural
227 *Syringa pinnatifolia* var.*alanshanica* assessed by ISSR. Plant Sci. J. 31, 85-92.
228 https://doi.org/10.3724/SP.J.1142.2013.10085.
229 Zhang, J., Jiang, Z., Su, H., Zhao, H., Cai, J., 2019. The complete chloroplast genome sequence of
230 the endangered species *Syringa pinnatifolia* (Oleaceae). Nordic J. Bot., e02201.
231 https://doi.org/10.1111/njb.02 2 01
232 Zybartaite, L., Zukauskiene, J., Jodinskiene, M., Janssens, S.B., Paulauskas, A., Kupcinskiene, E.,
233 2011. RAPD analysis of genetic diversity among Lithuanian populations of *Impatiens*
234 *glandulifera*. Žemdirbystė-Agriculture 98, 391-398.
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 s	% 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_4CT_8	1	0.9
$[GA]_{30}$	1	0.9	G_9AG_5	1	0.9
$[TC]_5$	7	6.3			
$[TC]_{11}$	1	0.9			
T_8CT_4	1	0.9			
A_8TA_7	1	0.9			
$A_5N_2A_4NA_8$	1	0.9			
$A_2TA_8TA_3$	1	0.9			
A_6TA_8	1	0.9			
$T_6N_6T_8$	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
[CCG]$_6$	1	0.2
[CTT]$_5$	5	1.0
[CTT]$_6$	2	0.4
[CTTCT]$_5$	1	0.2
[CTTTT]$_5$	1	0.2
[GAA]$_5$	5	1.0
[GAG]$_5$	1	0.2
[GAT]$_5$	4	0.8
[GAT]$_7$	1	0.2
[GAT]$_{11}$	1	0.2
[GCC]$_7$	1	0.2
[GCG]$_5$	1	0.2
[GCT]$_5$	1	0.2
[TAA]$_6$	1	0.2
[TATC]$_5$	1	0.2
[TGC]$_6$	1	0.2
[TTA]$_5$	2	0.4
[TTA]$_6$	1	0.2
[TTA]$_7$	2	0.4
[A/G]$_{15}$	1	0.2
[A/G]$_{16}$	1	0.2
[A/G]$_{17}$	1	0.2
[A/G]$_{27}$	1	0.2
[A/G}$_{31}$	1	0.2
[A/C]$_{18}$	1	0.2
[A/T]$_{17}$	1	0.2
[G/T]$_{19}$	1	0.2
[A/T]$_{18}$GT$_6$	1	0.2

[ATT]$_5$	2	0.3
[ATT]$_6$	2	0.3
[ATTTTT]$_5$	1	0.1
[ATCT]$_5$	1	0.1
[CAA]$_5$	1	0.1
[CAA]$_6$	1	0.1
[CAAAA]$_5$	2	0.3
[CATA]$_5$	1	0.1
[CAT]$_5$	5	0.7
[CAT]$_6$	5	0.7
[CAT]$_7$	2	0.3
[CAT]$_8$	1	0.1
[CCG]$_6$	1	0.1
[CTT]$_5$	7	1.0
[CTT]$_7$	2	0.3
[CTT]$_8$	1	0.1
[CTT]$_{10}$	1	0.1
[CGG]$_5$	1	0.1
[CGG]$_6$	1	0.1
[GAA]$_5$	15	2.1
[GAA]$_6$/[TTC]$_6$	8	1.1
[GAA]$_7$	1	0.1
[GAA]$_8$	2	0.3
[GAAA]$_5$	2	0.3
[GAG]$_5$	4	0.6
[GAT]$_5$	1	0.1
[GAT]$_6$	2	0.3
[GAT]$_8$	1	0.1
[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

		$[A/T]_{20}$	1	0.1
		$[A/T]_{22}$	1	0.1
		$[A/T]_{23}$	1	0.1
		$[A/T]_{31}$	1	0.1
Total	477	Total	709	

254
255