Population genetic structure and identification of loci under selection in the invasive tunicate, *Botryllus schlosseri*, using newly developed EST-SSRs

Yaping Lin \textsuperscript{a, b}, Aibin Zhan \textsuperscript{a, b, *}

\textsuperscript{a} Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing, 100085, China

\textsuperscript{b} University of Chinese Academy of Sciences, 19A Yuquan Road, Shijingshan District, Beijing, 100049, China

\textbf{Article info}

Article history:
Received 1 March 2016
Received in revised form 13 May 2016
Accepted 14 May 2016
Available online 26 May 2016

Keywords:
Biological invasion
*Botryllus schlosseri*
Genetic diversity
Golden star tunicate
Microsatellite
Invasive species
Tunicate

\textbf{Abstract}

*Botryllus schlosseri*, also known as golden star tunicate, is generally considered of European origin and has successfully invaded coastal waters of all continents except Antarctica. Owing to its significantly negative ecological impacts, it is crucial to understand its dispersal dynamics and mechanisms of invasion success. Here, we identified 1020 microsatellite-containing sequences from 98,626 expressed sequence tags (ESTs), and developed and characterized 17 polymorphic microsatellites (i.e. EST-SSRs) based on populations both from native (French coast) and invasive ranges (Canadian coast). The number of alleles, observed heterozygosity, and expected heterozygosity ranged from 2 to 12, 0.200 to 0.783, and 0.523 to 0.888 for the French population, and from 2 to 10, 0 to 0.524 and 0.043 to 0.827 for the Canadian population, respectively. We found significant population genetic differentiation between the native and invasive populations (pairwise $F_{ST} = 0.1712$). Moreover, principal coordinates analysis and Bayesian clustering test suggest long-distance dispersal between distant populations. When all loci were subjected for selection analyses, two loci (BS3244 and BS5339) were under selection based on the LOSITAN test. The results obtained in this study can help understand how ecological and evolutionary processes shape population genetic structure, and further how these processes contribute to invasion success.

© 2016 Elsevier Ltd. All rights reserved.

\section{Introduction}

Biological invasions have become a global ecological and conservation crisis in many ecosystems such as marine and coastal waters (Willis and Birks, 2006). As many estuaries, bays and near-shore waters are increasingly affected by human activities, coastal waters represent one of the most invaded habitats on the Earth (Rilov and Crooks, 2009). Among many notorious invaders in marine and coastal ecosystems, the golden star tunicate *Botryllus schlosseri* represents one of the most widespread marine invasive species, resulting in huge ecological damages and economic loss in world’s oceans (Zhan et al., 2015).

* Corresponding author. Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 18 Shuangqing Road, Haidian District, Beijing, 100085, China.

\textit{E-mail addresses:} zhanaibin@hotmail.com, azhan@rcees.ac.cn (A. Zhan).
Botryllus schlosseri, which is generally considered of European origin, was likely introduced via shipping to the east coast of North America in the early 1900s (Bock et al., 2012; Zhan et al., 2015). Since then, this species has successfully colonized coastal zones of all non-polar continents, for example, New Zealand in 1922, Australia in 1928, San Francisco in the early 1940’s, South California in 1960’s, Gulf of Maine in 1978, and Prince Edward Island of Canada in 2001 (Carver et al., 2006; Zhan et al., 2015). Established populations in new invaded areas are being reported frequently (Ben-Shlomo et al., 2010). Multiple vectors are responsible for its introduction and wide spread, such as hull fouling on a large number of boats, especially small pleasure crafts. Also, aquaculture transfers can provide new breeding stocks to colonize uninfested areas (Zhan et al., 2015). Natural dispersal is usually considered to be limited, only via rafting on eelgrass, algae or other forms of floating debris (Carver et al., 2006).

Although phylogenetics and population genetics studies have been performed on this highly invasive species (Bock et al., 2012), many questions remain poorly investigated, such as ecological and/or evolutionary changes of population genetic structure and genetic loci under selection during biological invasions. In order to deeply investigate these issues, we developed and characterized 17 polymorphic gene-associated microsatellite markers for B. schlosseri. We further assessed the population genetic structure of representative populations collected from both native and invasive distribution ranges. Finally, we employed multiple methods to identify loci under selection during biological invasions of B. schlosseri.

2. Materials and methods

2.1. Data mining and microsatellite identification

We obtained the expressed sequence tags (ESTs) of B. schlosseri from the NCBI database (http://www.ncbi.nlm.nih.gov/nucest/?term=Botryllus%20schlosseri). All downloaded ESTs were assembled into non-redundant sequences using web-based CD-HIT-EST with default parameters (Huang et al., 2010). Non-redundant EST sequences were used for microsatellite mining with SCIROKO version 3.4 (Kofler et al., 2007) based on the parameters described by Zhan et al. (2005) and Lin et al. (2016).

2.2. Microsatellite validation and polymorphism assessment

Amplification success and polymorphism were assessed using two populations collected from the coast of Canet-en-Roussillon, France (native habitat, N = 24) and Sydney, Nova Scotia, Canada (invasive habitat, N = 24). Genomic DNA was extracted from zooids using the standard phenol/chloroform method. DNA concentration and quality was assessed using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The microsatellite genotyping was performed using the methods of Zhan et al. (2010). Briefly, the PCR amplification was carried out in 12.5 µL reaction volume containing approximately 40 ng of genomic DNA, 1 × PCR buffer, 0.2 mM of each dNTP, 1.5 mM of Mg2+, 0.5 pmol M13-tailed (5′-CAC-GACGTTGTAAAACGAC-3′) forward primer, 1 pmol reverse primer and 1 pmol fluorescently (6-FAM, HEX, TMR, ROX) labelled M13 primer, and 0.25 unit of Taq DNA polymerase (Takara Bio Inc.). The PCR profile consisted of an initial denaturation at 95 °C for 5 min, followed by 35 cycles of 95 °C for 30 s, 50 °C for 30 s, and 72 °C for 30 s, with a final extension at 72 °C for 5 min. Amplified fragments were analyzed on an ABI 3730xl automated sequencer (Applied Biosystems, Foster City, CA, USA) with GeneScan™-500 LIZ™ as the internal size standard (Applied Biosystems). Alleles were scored using GeneMapper™ version 4.0 (Applied Biosystems).

2.3. Genetic diversity

Genetic diversity, including the number of alleles (A), observed heterozygosity (H_o), and expected heterozygosity (H_e) were calculated using GENEPOP online version (http://www.genepop.curtin.edu.au/). The polymorphic information content (PIC) value of each locus was estimated using CERVUS version 3.0.3 (Marshall et al., 1998). The presence of null alleles at all loci was evaluated using the software MICRO-CHECKER version 2.2.0 (van Oosterhout et al., 2004). The probability of significant deviation from Hardy–Weinberg equilibrium (HWE) was assessed using the Markov chain-based method implemented in GENEPOP. Significance criteria were adjusted for the number of simultaneous tests using the standard Bonferroni correction.

2.4. Loci under selection

To identify loci under selection between two populations, we adopted two theoretical approaches based on the FST-outlier test for modeling neutral loci. The first method that we used here was developed by Beaumont and Nichols (1996). This method detects the loci under selection based on the distributions of heterozygosity and FST under neutral expectations (Antao et al., 2008). As recommended by Antao et al. (2008), we used the program LOSITAN to perform 100,000 coalescent simulations to generate FST values under the infinite alleles model (IAM) with 99% confidence intervals and false discovery rate (FDR) of 0.05 for both 'neutral mean FST' and 'force mean FST' options. The second method that we used here was suggested by Foll and Gaggiotti (2008). This method decomposes locus population FST values into locus-specific components (z) and population-specific components (β). We performed this test using BAYESCAN program. Following 10 pilot runs of 5000 iterations with a 50,000 burn-in and a thinning interval of 20, we used the prior odds of 10 in favour of a neutral mode.
2.5. Population genetic structure analyses

The identified neutral microsatellite loci were used for population genetic analyses. Population differentiation was determined by $F_{ST}$ using ARLEQUIN version 3.5.2.1 (Excoffier and Lischer, 2010). Exact tests for population differentiation (10,000 dememorization steps, 500 batches and 5000 iterations per batch) were implemented at the 1% significance level. Principal coordinates analysis (PCoA) was conducted using GenAlEx version 6.501 (Peakall and Smouse, 2012). Population genetic structure was analyzed using a Bayesian clustering approach implemented in STRUCTURE version 2.3.4 (Pritchard et al., 2000). The admixture model was applied with five replicate chains of 1,000,000 Markov Chain Monte Carlo (MCMC) iterations and 100,000 burn-in repetitions. The most probable number of clusters was determined by the Ln $P(X|K)$.

3. Results and discussion

3.1. Distribution of microsatellites

Among 98,626 ESTs deposited at the NCBI database, we identified 32,400 non-redundant ESTs after removing redundant sequences using CD-HIT-EST. A total of 1020 (3.15%) microsatellite-containing EST sequences were identified based on the mining criteria (Table 1). The distribution density was 6.9 microsatellites per Mbp (Table 1). The most abundant repeat motif was penta-nucleotide (31.9%), followed by tri-nucleotide (27.0%) and hexa-nucleotide (21.5%). Di-nucleotide and tetra-nucleotide accounted for only 14.9% and 4.7%. The relatively low frequencies of these motifs mainly result from mutation pressure and possibly positive selection for specific single amino-acid stretches (Morgante et al., 2002).

Among all repeat motifs, AAGTC was the most abundant, accounting for 14.9% and 46.8% of all identified microsatellites and penta-nucleotides, respectively (Supplementary Fig. S1). Also, the dinucleotide AT showed a high abundance, accounting for 5.8% and 38.8% of all repeat motifs and di-nucleotides, respectively (Supplementary Fig. S1). Among tri-nucleotide repeats, the most frequent motifs were AAC (20.7%), ATC (18.2%) and AAT (15.6%; Supplementary Fig. S1).

We also examined the distribution pattern and quantity of identified microsatellites with respect to the repeat number of motifs (Supplementary Fig. S2). In general, the occurrence frequency of microsatellites decreased as the motif repeat number increased across all types of motifs (Supplementary Fig. S2), and a total of 993 microsatellites (97.4%) had fewer than 11 repeats (Supplementary Fig. S2). We randomly chose 37 microsatellite loci, including 13 di-, 20 tri-, 1 tetra-, 1 penta- and 2 hexa-nucleotides, for primer design and further polymorphism assessment (Table 2).

3.2. Genetic diversity

Among all 37 loci tested, 17 loci were polymorphic in at least one of the two tested populations. The number of alleles ranged from 2 to 12 and 2 to 10 for the French and Canadian populations, respectively (Table 2). The observed heterozygosity ($H_o$) varied from 0.200 to 0.783 and 0 to 0.524, while the expected heterozygosity ($H_e$) ranged from 0.523 to 0.888 and 0.043 to 0.827 for the French and Canadian populations, respectively (Table 2). The polymorphism information content (PIC) values for each locus ranged from 0.042 to 0.827 for the French and Canadian populations, respectively (Table 2). The loci are usually considered to be highly informative, reasonably informative or slightly informative if PIC $> 0.5$, PIC $> 0.25$, and PIC $< 0.25$ (Botstein et al., 1980). The highly informative microsatellite markers developed here will be useful for answering questions associated with invasion success and dispersal dynamics of B. schlosseri.

Overall, the genetic diversity estimates for B. schlosseri in the invasive population were lower than those in native population, which was supported by the number of alleles (7.0 vs. 4.3), expected heterozygosity (0.737 vs. 0.468), and polymorphism information content (0.683 vs. 0.417). The lower population diversity is likely due to genetic bottlenecks associated with the introduction of this species to North America. When a limited number of individuals was introduced into new habitats to source local populations, genetic bottlenecks derived from the founder effect can lead to a decrease of genetic diversity in newly established populations (Zhan et al., 2012), and genetic bottlenecks were detected in many ascidian species in their invaded ranges (see the review by Zhan et al., 2015).

Among the 17 loci, 10 loci and 14 loci conformed to Hardy–Weinberg equilibrium in French and Canadian populations after Bonferroni corrections (Table 2). A further test based on Micro-Checker showed the presence of null alleles at these loci, suggesting that null alleles be responsible for the deviation from Hardy–Weinberg equilibrium.

Table 1
Statistics for microsatellites (SSRs) in expressed sequence tags (ESTs) of the highly invasive golden star tunicate, Botryllus schlosseri.

| Motif       | No. of SSRs | Percentage (%) | Average repeat number | No. of SSRs per Mbp |
|-------------|-------------|----------------|-----------------------|---------------------|
| Di-nucleotide | 152         | 14.9           | 9.0                   | 6.9                 |
| Tri-nucleotide | 275         | 27.0           | 6.3                   | 12.5                |
| Tetra-nucleotide | 48          | 4.7            | 4.5                   | 2.2                 |
| Penta-nucleotide | 325        | 31.9           | 3.5                   | 14.8                |
| Hexa-nucleotide | 220         | 21.5           | 3.6                   | 10.0                |
| Total       | 1020        | 100            | –                     | 9.3                 |
3.3. Loci under selection

We used the LOSITAN and BAYESCAN programs to identify loci under selection for all 17 polymorphic microsatellite markers. Since EST-SSRs are associated with genes, some loci will inevitably be under selection and appear non-neutral (Ellis and Burke, 2007). Therefore, it is essential to implement a neutrality test before using EST-SSRs for population genetic structure analysis (Li et al., 2004). For the LOSITAN analysis, two outlier loci were identified: BS3244 (candidate balancing selection) and BS3339 (candidate positive selection, Fig. 1a). When these two microsatellite-containing ESTs were subjected to BLAST searches against NCBI databases, no significant hits were detected. When using BAYESCAN program, no outlier locus was detected (Fig. 1b). The disparity between the numbers of loci under selection identified by the two different methods is mainly derived from the fact that outlier test based on F-statistics can generate a certain rate of false positives owing to both biological and statistical reasons (Biere et al., 2013). Compared to the LOSITAN method, BAYESCAN has approved to be less prone to have false positives (Narum and Hess, 2011).

3.4. Population genetic structure

Based on the results of loci under selection, we selected 15 neutral polymorphic markers for population genetic structure analyses between native and invasive populations. We found a high level of genetic differentiation between native population and invasive populations. The pairwise $F_{ST}$ value based on the 15 neutral loci was 0.1712, which was significantly different ($P < 0.01$). The observed significant population genetic structure was consistent with that derived from genomic microsatellite markers (Bock et al., 2012). The results of principal coordinates analysis (PCoA) showed that two populations were clustered into two distinct groups excluding one individual from each population, CER17 from the French population and DOS22 from the Canadian population (Fig. 2a). Bayesian probability assignment based on STRUCTURE revealed two genetically divergent clusters ($K = 2$; Fig. 2b), largely consistent with the results based on PCoA (Fig. 2a). Most individuals from native population were assigned into one cluster, whereas most individuals from the invasive population were assigned to the other cluster with the exception of CER17 and DOS22 (Fig. 2b). One individual from the native French population (CER17) fall
within the invasive cluster, while one individual from invasive Canadian population (DOS22) assigned to the native cluster with high probabilities based on both PCoA and STRUCTURE (Fig. 2), suggesting that these individuals may experience long-distance dispersal likely via shipping. Despite the possible existence of gene flow among distant populations, the observed contrasting patterns of genetic structure between the populations from native and invasive ranges suggest local adaptation and/or genetic drift during range expansions of B. schlosseri (Bock et al., 2012).

**Conflicts of interest**

The authors declare that they have no conflict of interest.
References

Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A., Luikart, G., 2008. LOSITAN: a workbench to detect molecular adaptation based on a $F_{ST}$ outlier method. BMC Bioinform. 9, 323.

Beaumont, M.A., Nichols, R.A., 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. Roy. Soc. B Biol. Sci. 263, 1619–1626.

Ben-Shlomo, R., Reem, E., Douek, J., Rinkevich, B., 2010. Population genetics of the invasive ascidian Botryllus schlosseri from South American coasts. Mar. Ecol. Prog. Ser 412, 85–92.

Bierne, N., Roze, D., Welzic, I.J., 2013. Pervasive selection or is it… why are $F_{ST}$ outliers sometimes so frequent? Mol. Ecol. 22, 2061–2064.

Bock, D.G., MacIsaac, H.J., Cristescu, M.E., 2012. Multilocus genetic analyses differentiate between widespread and spatially restricted cryptic species in a model ascidian. P. Roy. Soc. B Biol. Sci. 279, 2377–2385.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32, 314–331.

Carver, C.E., Mallet, A.L., Vercaemer, B., 2006. Biological synopsis of the colonial tunicates, Botryllus schlosseri and Botrylloides violaceus. Can. Man. Rep. Fish. Aquat. Sci. 2747, 42.

Ellis, J., Burke, J., 2007. EST-SSRs as a resource for population genetic analyses. Heredity 99, 125–132.

Excoffier, L., Lischer, H.E.L., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10, 564–567.

Foll, M., Gaggiotti, O., 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180, 977–993.

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.

Koller, F., Schlötterer, C., Lelley, T., 2007. SciRoko: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23, 1683–1685.

Li, Y.C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.

Lin, Y., Chen, Y., Xiong, W., Zhan, A., 2016. Genomewide gene-associated microsatellite markers for the model invasive ascidian, Ciona intestinalis species complex. Mol. Ecol. Res. 16, 784–793.

Marshall, T.C., Slate, J., Kruuk, L., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639–655.

Morgante, M., Hanafey, M., Powell, W., 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30, 194–200.

Narum, S.R., Hess, J.E., 2011. Comparison of $F_{ST}$ outlier tests for SNP loci under selection. Mol. Ecol. Resour. 11184–11194.

Peakall, R., Smouse, P.E., 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537–2539.

Richiard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rilov, G., Crooks, J.A., 2009. Marine bioinvasions: conservation hazards and vehicles for ecological understanding. In: Rilov, G., Crooks, J.A., Jeffrey, A. (Eds.), Biological Invasions in Marine Ecosystems. Springer, Berlin, pp. 1–9.

van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M., Shipley, P., 2004. MicroChecker: software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4, 53–58.

Willis, K.J., Birks, H.J.B., 2006. What is natural? the need for a long-term perspective in biodiversity conservation. Science 314, 1261–1265.

Zhan, A., Bao, Z., Wang, X., Hu, J., 2005. Microsatellite markers derived from bay scallop Argopecten irradians expressed sequence tags. Fish. Sci. 71, 1341–1346.

Zhan, A., Briski, E., Bock, D.G., Ghabooli, S., MacIsaac, H.J., 2015. Ascidians as models for studying invasion success. Mar. Biol. 162, 2497–2470.

Zhan, A., Darling, J.A., Bock, D.G., Lacoursière-Roussel, A., MacIsaac, H.J., Cristescu, M.E., 2012. Complex genetic patterns in closely related colonizing invasive species. Ecol. Evol. 2, 1331–1346.

Zhan, A., MacIsaac, H.J., Cristescu, M.E., 2010. Invasion genetics of the Ciona intestinalis species complex: from regional endemism to global homogeneity. Mol. Ecol. 19, 4678–4694.