

General information

‡ Title

SNP genotypes of two west-Pacific pen shells, *Atrina japonica* and *Atrina lischkeana*, and their hybrids

‡ Principal investigator

Masashi Sekino

Fisheries Resources Institute, Japan Fisheries Research and Education Agency, 2-12-4 Fuku-ura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

‡ Co-investigators

1. Kazumasa Hashimoto

Fisheries Technology Institute, Japan Fisheries Research and Education Agency. 1551-8 Taira, Nagasaki 851-2213, Japan

2. Reiichiro Nakamichi

Fisheries Resources Institute, Japan Fisheries Research and Education Agency, 2-12-4 Fuku-ura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

3. Masayuki Yamamoto

Fisheries Division, Kagawa Prefectural Government, 4-1-10 Bancho, Takamatsu, Kagawa 760-8570, Japan

4. Yuichiro Fujinami

Goto Field Station, Fisheries Technology Institute, Japan Fisheries Research and Education Agency, 122-7 Nunoura, Tamanoura, Goto,

Nagasaki 853-0508, Japan

5. Takenori Sasaki

The University Museum, the University of Tokyo, 7-3-1 Hongo, Bunkyo,
Tokyo 113-0033, Japan

‡ Year of sampling

2009–2018

‡ Geographic locations of specimens

Geographic locations where the specimens were derived are distinguished by the prefix of specimen names (see below). The names of the geographical population samples are represented by the prefixes. For example, a specimen ID “HKDT-15” indicates that the specimen is one of 45 specimens collected from off Hakodate in Hokkaido Prefecture, Japan.

Prefix “HKDT-”: off Hakodate in Hokkaido Prefecture, Japan ($N = 45$).

Prefix “HMKJ-”: off Himakajima Island in Mikawa Bay (Aichi Pref.), Japan ($N = 35$).

Prefix “KGWA-”: off Kagawa Prefecture in the Seto Inland Sea, Japan ($N = 183$).

Prefix “SAGA-”: off Saga Prefecture in the Ariake Sea, Japan ($N = 40$).

Prefix “GOTO-”: off Fukuejima Island of the Goto Islands (Nagasaki Pref.), Japan ($N = 33$).

‡ Data and file overview

The file “final_SNP_genotypes.vcf” stores SNP genotype data of two

west-Pacific pen shells, *Atrina japonica* and *Atrina lischkeana*, and their hybrids (1,469 SNPs for 336 specimens in vcf format; produced on June 6, 2022). We inferred that the KGWA and SAGA population samples contained hybrids (including introgressants) between the two species. See Sekino *et al.* (in press) listed in the references section for details.

‡ Sharing and access information

No license/restriction is applied to the data. A manuscript based on the data has been accepted for publication in *Molecular Ecology* (Sekino *et al.* in press).

‡ Methodological information

The SNP data was obtained by RAD sequencing (restriction enzyme: SbfI). We followed Sekino *et al.* (2016) for RAD library construction. The libraries were subjected to short-read sequencing with a NextSeq 500 sequencer in combination with NextSeq 500 High Output Kit (Illumina; 75 cycles of single-end sequencing).

We truncated the resulting short reads to 64 bases including partial SbfI-recognition sequence (six bases) using the subprogram *process_radtags* of Stacks v1.35 or higher (Catchen *et al.* 2011). The subsequent variant calling was based on the remaining 58 bases. With FASTX-Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit), low quality reads, which had a base-quality score of less than 20 at 5% or more of the bases, were discarded. Reads retained after this filtering were mapped onto a reference genome of *Atrina japonica* (Sekino *et al.* in press; DDBJ/EMBL/GenBank accession numbers, BROG01000001–BROG01003391; Sequence Read Archive, DRR380487 and DRR380488) based on the MEM algorithm available in BWA v0.7.12 (Li & Durbin 2010)

with default parameters. After extracting mapped reads (the subprogram *view* in SAMtools v0.1.19; Li *et al.* 2009), we set aside reads with alternative hits (“XA” tag) and chimeric reads (“SA” tag). The resulting data (bam format) were converted to the mpileup format (the subprogram *mpileup* of SAMtools). Based on the mpileup data, we performed variant calling with the subprogram *mpileup2snp* of VarScan 2 v2.4.4 (Koboldt *et al.* 2012). In this computation, the following parameters were set for calling a variant at a position: minimum base-quality score, 30; minimum number of reads that supported a variant, 5; threshold significance value to call a variant (Fisher’s exact test), 0.05. After variant calling and initial filtering, we performed more rigorous SNP filtering as follows (see also Sekino *et al.* in press):

1. Biallelic sites were selected (VCFtools v0.1.16; Danecek *et al.* 2011).
2. For each specimen, sites with the coverage depth of < 30 were rejected (VCFtools).
3. Sites with the coverage depth of $\leq D + 3\sqrt{D}$, where D is the average depth over all SNP sites and specimens, were allowed (Li 2014).
4. Specimens with more than 10% missing genotypes were omitted.
5. Sites with minor allele frequency of > 0.05 across the population samples were selected (VCFtools).
6. Sites that were available in 90% or more of the specimens across the population samples as well as in each population sample were retained.
7. Sites were thinned so that neighboring sites were at least 1 Kb apart in a contig of the reference genome.
8. Sites that were out of Hardy-Weinberg equilibrium (HWE) were removed (an exact test available in VCFtools; critical P of 0.05 without correction of significance level for multiple simultaneous comparisons). The HWE filtering was applied to two population samples of putatively pure species

(*A. japonica*, HKDT; *A. lischkeana*, GOTO). Sites that failed to meet HWE in either sample or both were omitted.

9. If a pair of sites gave r^2 (an indicator of linkage disequilibrium) of > 0.1 , one of the two sites was removed (BCFtools v1.8; Li 2011). As with the HWE filtering, this filtering was applied to the HKDT and GOTO samples.

10. If a SNP site was in a contig that had no SbfI-recognition sequence, the site was excluded.

‡References

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. **G3**, 1, 171–182.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. **Bioinformatics**, 27, 2156–2158.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. **Genome Research**, 22, 568–576.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. **Bioinformatics**, 25, 2078–2079.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. **Bioinformatics**, 26, 589–595.

Li H (2011) A statistical framework for SNP calling, mutation discovery,

association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, 27, 2987–2993.

Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. **Bioinformatics**, 30, 2843–2851.

Sekino M, Nakamichi R, Iwasaki Y, Tanabe AS, Fujiwara A, Yasuike M, Shiraishi M, Saitoh K (2016) A new resource of single nucleotide polymorphisms in the Japanese eel *Anguilla japonica* derived from restriction site-associated DNA. **Ichthyological Research**, 63, 496–504.

Sekino M, Hashimoto K, Nakamichi R, Yamamoto M, Fujinami Y, Sasaki T (in press) Introgressive hybridization in the west Pacific pen shells (genus *Atrina*): restricted interspecies gene flow within the genome. **Molecular Ecology**