

# SCIENTIFIC REPORTS

OPEN

## Efficiency of ddRAD target enriched sequencing across spiny rock lobster species (Palinuridae: *Jasus*)

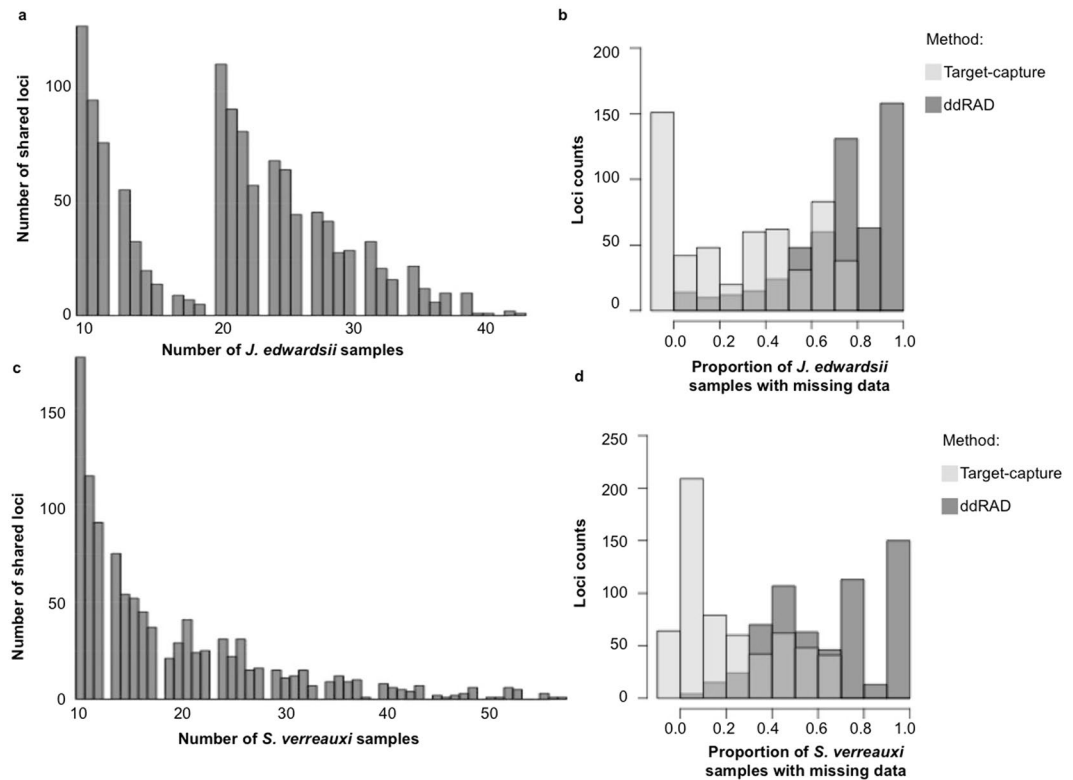
Carla A. Souza<sup>1</sup>, Nicholas Murphy<sup>1</sup>, Cecilia Villacorta-Rath<sup>2</sup>, Laura N. Woodings<sup>1</sup>, Irina Ilyushkina<sup>3</sup>, Cristian E. Hernandez<sup>4</sup>, Bridget S. Green<sup>2</sup>, James J. Bell<sup>3</sup> & Jan M. Strugnell<sup>5,1</sup>

Double digest restriction site-associated DNA sequencing (ddRADseq) and target capture sequencing methods are used to explore population and phylogenetic questions in non-model organisms. ddRADseq offers a simple and reliable protocol for population genomic studies, however it can result in a large amount of missing data due to allelic dropout. Target capture sequencing offers an opportunity to increase sequencing coverage with little missing data and consistent orthologous loci across samples, although this approach has generally been applied to conserved markers for deeper evolutionary questions. Here, we combine both methods to generate high quality sequencing data for population genomic studies of all marine lobster species from the genus *Jasus*. We designed probes based on ddRADseq libraries of two lobster species (*Jasus edwardsii* and *Sagmariasus verreauxi*) and evaluated the captured sequencing data in five other *Jasus* species. We validated 4,465 polymorphic loci amongst these species using a cost effective sequencing protocol, of which 1,730 were recovered from all species, and 4,026 were present in at least three species. The method was also successfully applied to DNA samples obtained from museum specimens. This data will be further used to assess spatial-temporal genetic variation in *Jasus* species found in the Southern Hemisphere.

Target enriched, or target-capture next-generation sequencing, has been successfully applied to assess genome-scale data in non-model species lacking a reference genome. The method uses 60- to 120-mer probes for in-solution hybridization to capture specific genome targets for sequencing. This approach has proven to be useful in addressing phylogenetic questions (reviewed by McCormack<sup>1</sup>) and has been employed in studies incorporating museum specimens<sup>2–6</sup>. However, the design of probes for target capture methods requires high-quality genomic or transcriptomic resources, which limits its application to a wide range of taxa<sup>7,8</sup>. Loci obtained using most capture methods to date are either subject to selection (e.g. in the case of exon derived transcriptomic markers) or are highly conserved within species (e.g. for probes that use ultra-conserved elements). Therefore, capture-based approaches have not been widely used for studying population-level processes such as genetic drift and gene flow, and have had limited application to population genetic studies.

The majority of population genomics studies have been undertaken using restriction digest derived methods (e.g. RADSeq, GBS, ddRAD). These methods comprise a range of related protocols, employing restriction digestion and library size selection in order to reduce genome complexity, enabling the study of genome-wide genetic variation without any prior genomic knowledge<sup>9–11</sup>. However, issues such as PCR duplication bias, large amounts of missing data due to allelic dropout, low reproducibility and the requirement of high molecular weight DNA have been reported from these studies<sup>12–15</sup>. The main shortcomings of this approach are the difficulties of cross species comparison<sup>16</sup> and repeatability of sequencing across multiple libraries, even of replicate samples. Additionally, poor quality DNA means that the application of restriction digest-based approaches is mostly unsuitable for preserved material (e.g. museum collections).

<sup>1</sup>Department of Ecology, Environment and Evolution, School of Life Sciences, La Trobe University, Melbourne, VIC 3086, Australia. <sup>2</sup>Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, TAS 7001, Australia. <sup>3</sup>School of Biological Sciences, Victoria University of Wellington, Wellington, 6140, New Zealand. <sup>4</sup>Departamento de Zoología, Facultad de Ciencias Naturales y Oceanográficas, Universidad de Concepción, Concepción, Chile. <sup>5</sup>Centre for Sustainable Tropical Fisheries and Aquaculture and College of Science and Engineering, James Cook University, Townsville, QLD 7001, Australia. Correspondence and requests for materials should be addressed to C.A.S. (email: carladosanjos@gmail.com)



**Figure 1.** Histograms displaying the number of shared loci among samples in ddRAD libraries of *J. edwardsii* (a) and *S. verreauxii* (c) samples. Overlapping histograms of target-capture and ddRAD representing the percentage of missing data among validated loci in *J. edwardsii* (b) and *S. verreauxii* (d).

Recently, methods combining the benefits of RADseq (e.g. no requirement for genomic resources) and target-capture (e.g. repeatability across samples) approaches have been developed to improve sequencing coverage in non-model organisms. Examples of these protocols include HyRAD<sup>5</sup>, RADcap<sup>17</sup> and Rapture<sup>18</sup>, and although similar, each technique offers distinct benefits and limitations. HyRAD supports the use of highly degraded DNA samples and offers cost-effective benefits such as the use of ‘home-made’ probes (biotinylated ddRAD libraries); the Rapture protocol produces the highest coverage from minimal reads per individual; RADcap has proven to be cost-effective and uses adapters with degenerated bases in ddRAD libraries enabling identification of PCR duplicates. These methods all offer promise for combining the efficiency and repeatability of capture-based approaches with the utility for population genomics studies of non-model species. Here we report on a similar, yet alternate approach for restriction-based capture libraries and assess the utility of this approach for multiple species population genomics studies, and, in particular for incorporating museum samples into these studies.

In this study we used existing ddRADseq sequence databases of two closely related marine lobster species, *Jasus edwardsii*<sup>15</sup> (N = 42) and *Sagmariasus verreauxii* (N = 55), to design probes for target enrichment re-sequencing. We then evaluated the efficiency and reproducibility of the target enrichment approach, and determined the practical utility of this method for population genomic studies within all marine spiny lobster species of the genus *Jasus* in modern (N = 40) and museum collection (N = 39) specimens. We implemented a modified target-capture protocol to: (1) generate cost effective, high quality population genomic data across *Jasus* species using a single target panel; (2) reduce the amount of missing data reported from previous ddRAD experiments; and (3) recover consistent sequencing data from highly degraded museum collection samples by using high confidence probabilistic base-calling.

## Results and Discussion

**ddRAD loci catalogue and probe design.** The *J. edwardsii* ddRAD libraries published by Villacorta-Rath *et al.* (2016)<sup>15</sup>, and *S. verreauxii* libraries generated in this work were first assembled in PyRAD<sup>19</sup> in order to obtain a ddRAD loci catalogue. This was used as a template to design the MYbait<sup>®</sup> probes. Three PyRAD assemblies (*J. edwardsii*, *S. verreauxii* and a combined species assembly) resulted in a catalogue of 4,629 loci (Supplementary Methods S1). *De novo* assembly, using liberal similarity thresholds (75%), revealed putative paralogous loci, which were discarded. Redundant loci (present in multiple assemblies) were synonymized into single loci as described in the Methods section. Across all three assemblies only 123 loci were shared between the two species, and 2,267 were species-specific (1,219 were from *J. edwardsii* and 1,048 were from *S. verreauxii*). The percentage of missing loci over all samples was  $52.4 \pm 19.2$  for *J. edwardsii* and  $66.3 \pm 18.3$  for *S. verreauxii* (Fig. 1a and c). After discarding putative paralogous loci and simple repeat regions, 2,366 ddRAD loci were selected as templates for manufacturing MYbaits<sup>®</sup> probes.

Species	Country of Origin	Location	Date	N
<i>J. edwardsii</i>	Australia	Victoria, West, North, South of Tasmania and Tasman Sea	1991	8
	Australia	Victoria, Southwest Tasmania and East Tasmania	2013–2014	8
<i>J. caveorum</i>	British Overseas Territory	South East Pacific, Pitcairn Island, Foundation Seamounts	1995	7
<i>J. frontalis</i>	Chile	Juan Fernandez Archipelago, Islas Desventuradas	1967	8
	Chile	Juan Fernandez Archipelago, Islas Desventuradas	2010	8
<i>J. lalandii</i>	Namibia/South Africa	West/Southern Africa	1991	8
	South Africa	Cape Town	1967	8
	South Africa	East coast and islands	2015	8
<i>J. paulensis</i>	French Overseas Territory	Saint Paul and Amsterdam Islands	1967	8
	French Overseas Territory	Saint Paul and Amsterdam Islands	2014	8
<i>J. tristani</i>	South Africa	Tristan da Cunha Seamount	2015	8
<i>S. verreauxi</i>	Australia	Tasmania, New South Wales	2013	8
	New Zealand	Cape Maria Van Diemen to North Cape	2013	8
Total				103

**Table 1.** Sampling localities, species designations and number of individuals sampled (N).

Species	<i>J. edwardsii</i> (N = 11)	<i>S. verreauxi</i> (N = 16)
Processed reads	740,745	544,149
On-target reads	558,595	416,393
Reads mapped to original probe set	296,247	239,490
PCR duplicates	117,566	91,787
ddRAD loci BLAST hits	1,516	1,250
ddRAD loci mapped (Bowtie2)	922	784
Average coverage	29.21X	19.10X
% missing ddRAD loci*	24.04 ± 19.53	31.20 ± 27.30

**Table 2.** Target-capture enriched sequencing and genotyping efficiency between *J. edwardsii* and *S. verreauxi* sampling. \*Missing data was estimated as the average missing percentage across validated loci.

**Target enriched sequencing efficiency.** MYbait® probes were used to capture 2,366 ddRAD loci from 87 individuals, representing the six *Jasus* species (47 obtained from museum samples) and 16 individuals of *S. verreauxi* (Table 1). The enriched loci were sequenced on a single 250 X 2 Illumina MiSeq sequencing run, generating 4.6 Gbp (giga base pairs) of sequence data (1.29% sequencing error rate and 84.49% of data above Q30 score). Subsequent trimming and removal of low quality reads and external contaminants resulted in 800 and 174 Mbp high-quality data for modern and museum specimens, respectively. The lower sequencing output for museum samples is likely due to low yield DNA extracts, which were highly fragmented and possessed a low  $A_{260}/A_{280}$  ratio (Supplementary Fig. S2). For example, none of the 50-year-old *J. frontalis* museum samples (N = 8) generated sequencing data sufficient to enable variant calling and were excluded from downstream analysis.

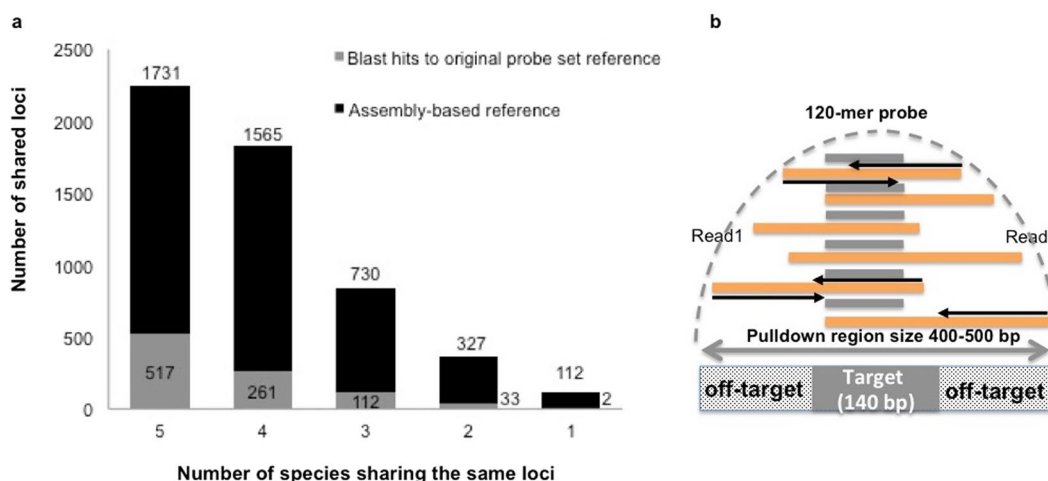
**Probe efficiency for *J. edwardsii* and *S. verreauxi*.** As the probes were developed with *J. edwardsii* and *S. verreauxi* ddRAD libraries, we initially examined the probe capture efficiency in these two taxa.

**ddRAD loci recovered.** The number of reads recovered that were directly BLAST<sup>20</sup> matched to the designed probes varied from 72.69% in *J. edwardsii* and 76.52% in *S. verreauxi*, covering 2,045 ddRAD loci by at least one sequence read (Table 2). We found 721 loci shared between species and 1,324 were species-specific (795 in *J. edwardsii* and 529 in *S. verreauxi*). A lower percentage of reads were correctly mapped using Bowtie 2<sup>21</sup> aligner (55.04% in *J. edwardsii* and 57.52% in *S. verreauxi*) and fewer loci (1,250 loci) were mapped to the probe set with mapping quality greater than five ( $maQ \geq 5$ ), of which 456 were shared between species (Table 2). Compared to the initial ddRAD sequencing (123 shared loci) the target-capture method resulted in a four-fold increase in shared markers between these genera, including a number of loci identified as putatively under selection by Villacorta-Rath *et al.*<sup>15</sup>. In addition, the target-capture approach resulted in less missing data within each species than the ddRAD sequencing approach (Fig. 1b and d). This is most likely due to allele dropout in ddRAD libraries usually caused by polymorphisms in the restriction sites, which has the potential to bias population genetics analyses<sup>22, 23</sup> or render them completely unusable.

We were able to re-sequence and validate 54 loci reported in Villacorta-Rath *et al.*<sup>15</sup> in eight *J. edwardsii* samples replicates. Besides, the average level of missing data among these loci in our target-capture experiment was 15.96% within *J. edwardsii* species. It means that the ddRAD loci outliers discovered in a previous study<sup>15</sup> were

Species	N	Processed reads	Mapped reads to original probe set (%)	PCR duplicates (% over mapped reads)	Blast hits to original probe set**	Mapped reads to new assembly (%)
<i>J. edwardsii</i>	8	605,396	242,538 (40.06)	97,952(40.39)	910	331,505(42.91)
<i>J. frontalis</i>	8	286,288	28,360(9.91)	5,403 (19.05)	565	68,374(15.85)
<i>J. lalandii</i>	8	285,545	106,425(37.27)	22,283(20.94)	854	139,969(39.15)
<i>J. paulensis</i>	8	478,098	190,903(39.93)	58,824 (30.81)	875	264,249(52.33)
<i>J. tristani</i>	8	246,252	94,839(38.51)	24,486(25.82)	829	132,244(31.10)
Total	40	1,901,579	663,065(34.86)	208,948 (31.51)	925*	936,241(49.23)

**Table 3.** Number of reads mapped to original probe set (ddRAD loci) and assembly-based reference with corresponding levels of similarity threshold (\*). \*\*Overall loci counts across species. \*( $maQ \geq 5$ ).



**Figure 2.** Number of variable loci shared amongst each of six *Jasus* species. (a) The left panel displays the number of loci shared among five (5), four (4), three (3) and two (2) species, or found only in a single species (1) in the assembly-based reference (black and grey stacked column) and the corresponding significant BLAST hits to the original probe set (grey). (b) The right panel is a graphic representation of target enrichment hybridization reaction showing library size, captured fragments size and off-target reads. The grey lines represent the 120-mer probes, the orange lines represent the library fragments and the black arrows represent overlapping/non-overlapping paired-end reads (Read 1 and Read 2).

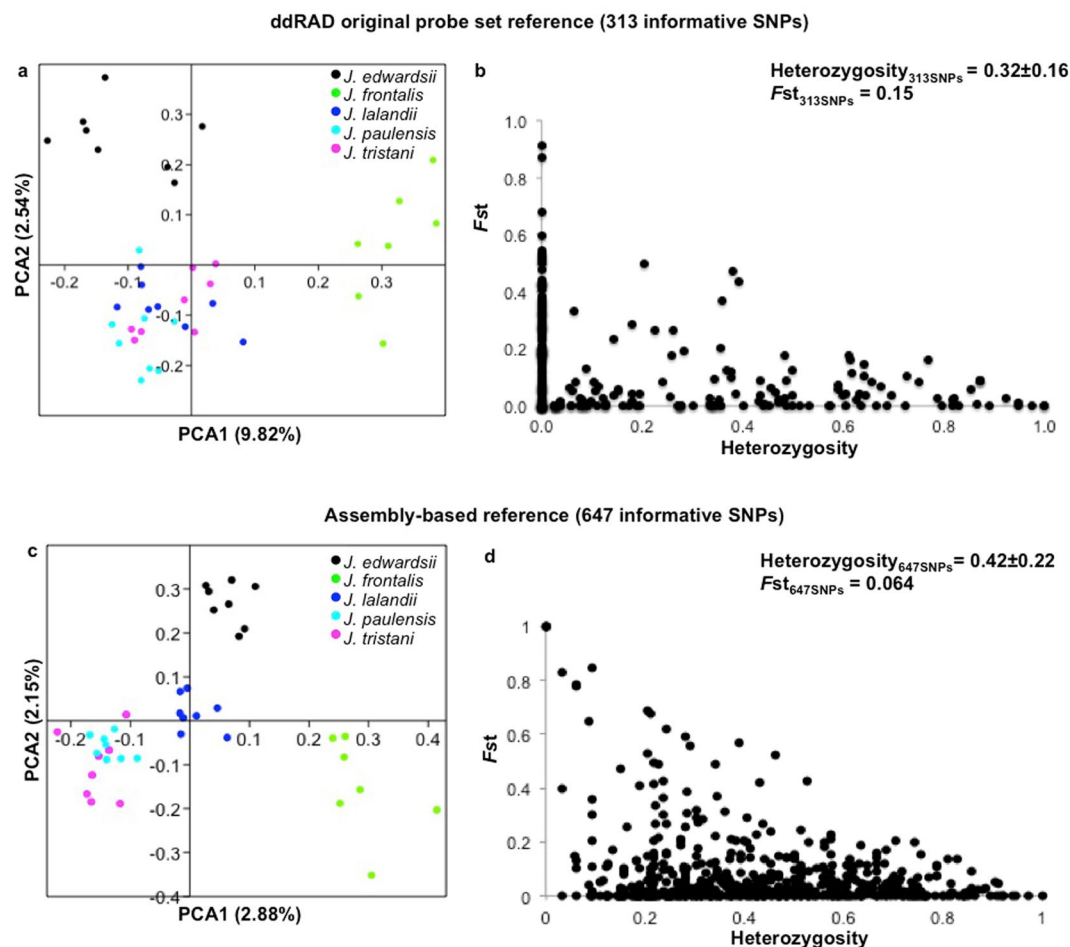
successfully re-sequenced and could be enriched and sequenced in range of populations to investigate genome signatures of selection.

**Probe efficiency in non-target *Jasus* species (N = 40).** To assess the probe efficiency in non-target species (i.e. those not used for the initial probe design), we assessed the recovery of loci across the modern *Jasus* samples including *J. edwardsii*, *J. frontalis*, *J. lalandii*, *J. paulensis* and *J. tristani*.

**ddRAD loci recovered.** Reads mapped to the original probe set using Bowtie 2 covered 925 target loci, varying from 565 in *J. frontalis* to 910 in *J. edwardsii* (Table 3). A total of 517 loci were shared among the five *Jasus* species and a further 112 were shared across at least three species. Only two loci were detected from a single species (Fig. 2a).

**SNP diversity.** The variant calling followed rigorous post processing filters to reduce the number of false positive calls, resulting in 5,486 SNPs detected within and among species (Supplementary Table S3). After discarding SNPs in Linkage Disequilibrium, those with greater than 10% missing data or a minor allele frequency (MAF) lower than 0.05, 313 informative SNPs remained across all modern *Jasus* samples. These SNPs had a mean heterozygosity of  $0.32 \pm 0.16$  and a global  $F_{st}$  of 0.15. PCA analyses of this data revealed three main clusters (Fig. 3a): (1) *J. edwardsii*, (2) *J. frontalis* and (3) a third cluster comprised by *J. lalandii*, *J. paulensis* and *J. tristani*. Several SNPs (133) were fixed within species (Fig. 3a and b), which is unsurprising given that these SNPs are testing species level differences, however, the remaining SNPs allow population-level assessment within multiple taxa.

**Assembly-based reference for *Jasus* genus based on target enriched sequencing data.** In order to enable more effective mapping results across the 40 *Jasus* specimens described above, a new reference sequence was built based on a *de novo* assembly, to account for the entire sequence diversity across non-target *Jasus* species, including 'off-target' loci (i.e. flanking regions adjacent to the targeted ddRAD loci). All *Jasus* reads were clustered



**Figure 3.** Principal Component Analysis and scatter plot of locus-specific  $F_{st}$  vs observed heterozygosity of *Jasus* species in modern samples. (a) and (b) 313 SNPs from the ddRAD original probe set; (c) and (d) 647 SNPs from the assembly-based reference. Maximum missing data per site was set to 0.10, MAF < 0.05, and overall  $F_{st}$  was estimated according to Weir & Cockerham (1984).

under a 90% similarity threshold, and the clusters were screened for putative paralogous loci, over-splitting loci, mitochondrial DNA copies and assembly artefacts, such as external contaminants and chimeras, each of which were discarded (detailed in Supplementary Methods S4). The new assembly-based reference comprised 5,940 loci, of which 1,773 loci were assigned as off-target reads, rather than artefacts (Fig. 3b), and therefore, an alternative source of informative sequencing data. BLAST analysis of off-target loci showed 925 significant hits to a *J. edwardsii* transcriptome draft (SRA Bioproject accession number: PRJNA386609), indicating that almost half off-target loci are in fact expressed regions of *J. edwardsii* genome (Supplementary Table S5).

**Mapping efficiency of assembly-based reference in modern samples.** The assembly-based reference increased mapping success from 34.86% to 49.23% across the 40 modern *Jasus* specimens (Table 3). Mapping results showed that 1,475 (among 5,940) loci were monomorphic among specimens. Among the variable loci, we found 1,731 homologous loci were consistently shared across the five species, 4,026 were shared across at least three species and 112 loci were found in only a single species (Fig. 2a). The number of loci in each species ranged from 4,106 in *J. edwardsii* to 2,968 in *J. frontalis* (Table 4). Employing an assembly-based reference maximised the use of available reads because it permitted the mapping of loci that were too divergent - in the non-target species - to be efficiently mapped to the original probe set.

**Genome divergence of non-target *Jasus* species.** Examination of target enriched sequencing efficiency revealed evidence of genome divergence among *Jasus* species. Aspects such as sequencing yield (given in number of processed reads), mean coverage, GC content (%GC) and SNP diversity suggest important differences across the captured genomes (Table 4). For example, the heterogeneous mean coverage ( $0.631$ ;  $P < 0.01$ ) and GC content ( $0.346$ ;  $P < 0.01$ ) across species, for example, were both positively correlated with sequencing yield (Supplementary Table S6). However, mean coverage and GC content were not reciprocally correlated. The mean coverage was significantly reduced in *J. tristani* and *J. frontalis* ( $3.90X \pm 0.42$  and  $3.21X \pm 0.14$  respectively); while the former presented the lowest sequencing yield (Table 3), the latter presented the greatest GC content. This



Species	Date	N	Mapping results				Gene diversity							
			Overall coverage	Mean coverage among samples	Read heterozygous rate	% GC	Loci counts	Variable loci	Deamination	SNPs	SNP/ locus	tsi/ tsv	MMAF	HHet.
Modern samples														
<i>J. edwardsii</i>	2013	8	48.89 ± 1.46	8.39 ± 0.76 <sup>a</sup>	0.0066 ± 0.0006	44.96 ± 0.88 <sup>a</sup>	4,106	3,363	0.861	30,334	9.02	1.28	0.28	0.27
<i>J. frontalis</i>	2010	8	11.25 ± 0.36	3.21 ± 0.14 <sup>b</sup>	0.0271 ± 0.0035	47.32 ± 0.31 <sup>b</sup>	2,698	1,375	0.897	6,632	4.82	1.24	0.32	0.21
<i>J. lalandii</i>	2015	8	20.69 ± 0.49	4.02 ± 0.26 <sup>a</sup>	0.0151 ± 0.0015	43.41 ± 0.10 <sup>a</sup>	3,297	2,338	0.883	17,223	7.37	1.31	0.28	0.2
<i>J. paulensis</i>	2015	8	40.56 ± 1.16	7.00 ± 0.33 <sup>a</sup>	0.0082 ± 0.0005	45.00 ± 0.14 <sup>a</sup>	4,035	3,261	0.86	28,673	8.79	1.28	0.28	0.24
<i>J. tristani</i>	2015	8	18.50 ± 0.53	3.90 ± 0.42 <sup>c</sup>	0.0152 ± 0.0036	45.28 ± 0.45 <sup>a</sup>	3,735	2,604	0.888	16,121	6.19	1.28	0.3	0.21
Museum samples														
<i>J. caveorum</i>	1995	7	30.28X ± 0.94	7.01 ± 1.74	0.0090 ± 0.0020	44.29 ± 1.00	3,828	2,745	0.864	18,140	6.61	1.27	0.33	0.37
<i>J. edwardsii</i>	1991	8	17.63X ± 0.48	3.55 ± 1.00	0.0170 ± 0.0070	43.68 ± 1.10	3,377	2,103	0.887	12,782	6.08	1.3	0.41	0.35
<i>J. lalandii</i>	1967	8	14.15X ± 0.74	9.38 ± 1.97	0.0136 ± 0.0045	43.00 ± 1.48	2,248	982	0.936	4,408	4.49	1.22	0.34	0.28
<i>J. lalandii</i>	1991	8	35.19X ± 1.93	4.02 ± 1.46	0.0103 ± 0.0025	37.05 ± 0.64	1,043	315	0.835	941	2.99	1.1	0.4	0.55
<i>J. paulensis</i>	1967	8	5.96X ± 0.22	2.38 ± 0.17	0.0170 ± 0.0031	36.66 ± 0.37	1,380	483	1.00609	1,978	4.1	1.19	0.36	0.19

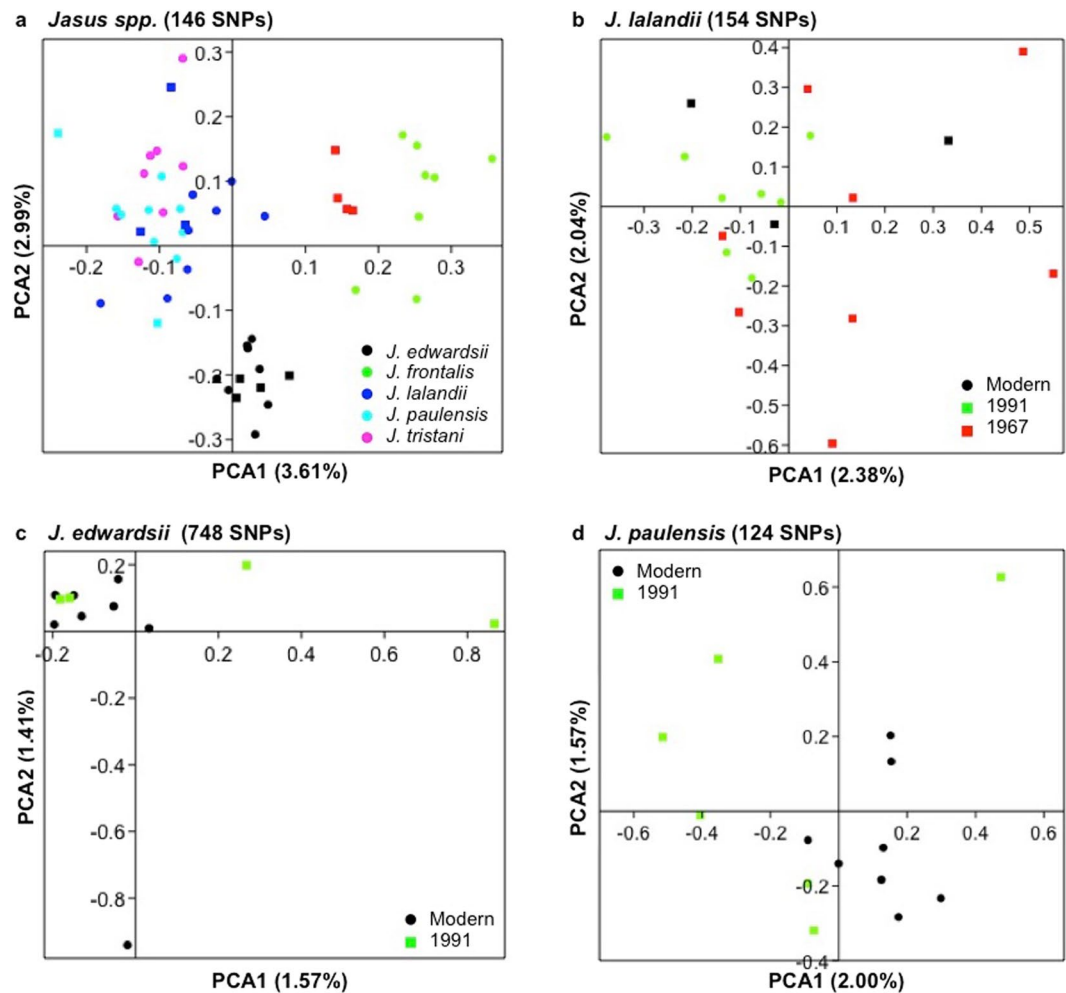
**Table 4.** Descriptive statistics of mapped data based on the new assembly for modern and museum samples. Mean coverage and GC content (GC%) per sample was compared across species in modern samples. Data expressed as mean (SEM). <sup>a,b,c</sup>Different superscripts within a column denote significant differences ( $P < 0.01$ ). Adjustment for multiple comparisons: Bonferroni.

implies that, unlike *J. tristani*, *J. frontalis* might have deeper genome divergence from the other *Jasus* species (as suggested by the PCA analyses in figure 3). This may have also affected the mapping efficiency in this species.

**Sequencing coverage.** The overall mapping coverage (mean depth of targets) of the assembly-based reference varied from 11X in *J. frontalis* to 48X in *J. edwardsii* (Table 4). In terms of sequence similarity, this means that the efficiency of probes in capturing the genome targets was reduced in *J. frontalis* and increased in *J. edwardsii*. *J. edwardsii* showed the best results in terms of captured yield and coverage (Tables 3 and 4) as sequences of this species were used to design probes (along with *S. verreauxi*). On the contrary, the *J. frontalis* genome was the most divergent as evidenced by higher sequence dissimilarities to the assembly-based reference. In line with this observation, differences in average sequence coverage have been reported in target-capture experiments for species with more than 5% sequence divergence<sup>25–28</sup>. However, *J. frontalis* did not have the lowest sequencing depth or captured yield as expected, but rather the lowest number of mapped reads (Table 3) and also the lowest mean target coverage (Table 4). Although this could suggest differentiation bias as a result of low sequence coverage, such bias tends to underestimate rather than overestimate differentiation and rare variants<sup>29</sup>.

**SNP diversity.** Overall 23,555 cross-species variants were identified within the 1,731 shared loci (Fig. 2a) from the assembly-based reference. The diversity of this reference is in accordance with the patterns in the original mapped loci (Table 4), with the lowest number of variants within species found in *J. frontalis* 2,698 and the greatest in *J. edwardsii* (4,106). After removing SNPs with missing data greater than 10%, MAF lower than 0.05 and linked SNPs, a total of 647 informative SNPs remained (2X increase from the original mapped loci). This SNP subset displayed a high proportion of polymorphic loci ( $0.42 \pm 0.22$  mean heterozygosity) and low differentiation level (global  $F_{st} = 0.064$ ) among species (Fig. 3d). This indicates that they are potentially more informative within species rather than among species comparisons, when compared with those from the mapped probe set. For example, in Fig. 3d, several loci (232 SNPs) present  $F_{st}$  among species equal to 0. The PCA analyses based on 647 SNPs revealed four main clusters: (1) *J. frontalis*, (2) *J. edwardsii*, (3) *J. lalandii* and (4) a cluster comprised by *J. paulensis* and *J. tristani* (supporting Groenenveld *et al.* (2012)<sup>30</sup> findings, based on mtDNA data evidence, that *J. paulensis* and *J. tristani* should be synonymized). The assembly-based data set provided greater separation than the mapped probes, placing *J. lalandii* in a different cluster from the *J. paulensis*/*J. tristani* species. This result indicates that differentiation levels among non-target species can be more accurately achieved by building an assembly-based reference of target-enriched sequencing data; because it takes into account the full diversity of the sequencing dataset.

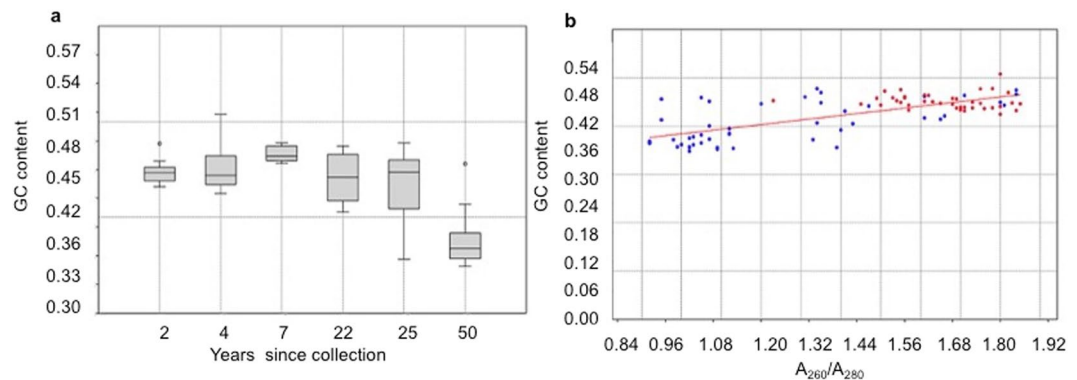
**Efficiency of target enriched sequencing method applied to museum samples.** In order to evaluate the efficiency of target enriched sequencing in museum samples, reads from the 39 museum specimens of five *Jasus* species (Table 4) were mapped to the assembly-based reference. The alignments were then compared with modern samples in terms of mean coverage, GC content, read heterozygous rate and SNP diversity. Prior to this comparison, DNA damage patterns were established for each museum sample by tracking and quantifying cumulative substitutions frequencies of C to T at the 5' end and G to A at the 3' end in mapped reads using the mapDamage 2.0 software<sup>31</sup> (data not shown). No evidence of base mis-incorporation bias due to DNA damage was found in any of the museum samples.



**Figure 4.** PCA across and within species based on SNP genotypes, where no deamination filter was applied. For each dataset, the sample passed filters and SNP pruning was adjusted to allow 0.10 maximum missing data per site and  $MAF < 0.05$ : (a) PCA across six *Jasus* species, 146 SNPs and 53 samples passed filters; (b) PCA of *J. lalandii*, 154 SNPs and 19 samples passed filter; (c) PCA of *J. edwardsii*, 748 variants and 12 samples passed filters; (d) PCA of *J. paulensis*, 124 variants and 14 samples passed filters.

**Sequencing coverage and GC content.** In contrast to the modern samples, a significant correlation between GC content and mean coverage ( $0.24179$ ;  $P < 0.032$ ) was found, suggesting that museum samples were highly impacted by sequencing coverage (Supplementary Table S7 and Fig. S8). The overall GC content among museum samples were quite heterogeneous (Fig. 5a) and significantly correlated to the year of collection ( $0.483$ ;  $P < 0.01$ ) and  $A_{260}/A_{280}$  ratio ( $0.482$ ;  $P < 0.01$ ) (Table S7 and Fig. S8). We suggest that sequencing coverage was affected by DNA fragmentation in museum samples leading to non-uniform representation of targets and heterogeneous GC content in sequencing libraries. However, patterns are difficult to interpret. For example, the highest overall coverage was observed in 50-year-old *J. lalandii* samples ( $35.19X \pm 1.93$ ) and the lowest in 50-year-old *J. paulensis* ( $5.96X \pm 0.22$ ) (Table 3). In both sample groups, the mean GC content significantly deviated from the mean of corresponding modern samples ( $P < 0.01$  in both cases; Supplementary Table S9), suggesting significant bias among these samples. However *J. paulensis* museum samples were the only samples that demonstrated an increased read heterozygous rate ( $P < 0.0066$ ; in Supplementary Table S9). The impact of sample age in DNA fragmentation was evident, but not homogeneous, among museum samples due to variability in DNA purity as indicated by the  $A_{260}/A_{280}$  ratio as shown in Fig. 5b. It may also be related to the museum preservation methods applied to the specimens.

**Utility of museum samples.** In an attempt to find an approach to screen the museum samples for library preparation in future studies where there is likely to be DNA damage and fragmentation, we found that samples with  $A_{260}/A_{280}$  ratios as low as 1.01 and moderate DNA concentration above  $30\text{--}40\text{ ng}/\mu\text{L}$  (as determined from the  $A_{260}$  values) exhibited acceptable average target coverage with no significant GC% deviations (Fig. 5b). This allows false-positive variants to be filtered, improving SNP calling accuracy<sup>32</sup>, and, a 'rule of thumb' for selecting suitable historic samples for library preparation; samples with an  $A_{260}/A_{280}$  ratio greater than 1.0 and spectrophotometric quantification greater than  $30\text{ ng}/\mu\text{L}$  are more likely to generate acceptable sequencing coverage. The  $A_{260}/A_{280}$



**Figure 5.** Sample quality of modern and museum *Jasus* specimens. **(a)** Box-plot of mean GC content and year since sample collection. **(b)** Linear regression between  $A_{260}/A_{280}$  ratio and GC content; blue dots represent museum samples and red dots represent modern samples.

ratio is routinely used as a DNA purity indicator for protein, phenol or other contaminants that strongly absorb light at or near 280 nm wavelengths, however, the actual ratio is also subject to nucleotide composition and pH variation<sup>33</sup>. Therefore, when using historic samples for library preparation, close attention should be paid to both the  $A_{260}/A_{280}$  ratio and DNA spectrophotometric quantification cutoff, both prior to sequencing, and when assessing the results of sequencing.

**Mapping efficiency of assembly-based reference in museum samples.** Because of the limitations of the museum samples used in this study, the amount of missing data was much higher in these samples. This is likely because the fragmentation bias of DNA templates in library preparation led to uneven enrichment and sequencing coverage of targeted loci. The number of loci in each species ranged from 3,828 in 25-year-old *J. edwardsii* to 1,043 in 50-year-old *J. lalandii* (Table 4). Each of the four species for which museum samples were used, has at least one fold fewer loci (nearly five fold in *J. paulensis*) when compared with the modern samples. The loci also seem to be less variable, as determined by the number of SNP per locus in all species. Whilst cumulative substitution bias was not evident in museum samples, the overall deamination level in museum samples of *J. lalandii* and *J. paulensis* from the year 1967 was higher than modern samples (Table 4). Thus, in order to examine the usefulness of the target capture approach for incorporating museum samples into population genomic studies, we modified the variant-filtering settings to avoid false positive variants as a precaution. Strand biased variants, heterozygous SNPs at the end of reads and SNPs with deamination pattern found towards the 5' and the 3' reads' ends were filtered out resulting in a slightly different SNP set from that used for the modern samples.

**Genetic diversity of *Jasus* including museum samples.** Amongst the *Jasus* species, only museum samples were available for *J. caveorum*. This species formed a distinct group in the PCA (Fig. 4a). However, the necessity of using the reduced set of SNPs meant that again, *J. lalandii* appeared indistinguishable from *J. paulensis*/*J. tristani*. Filtering SNPs for deamination had a negligible impact on the overall results, as evident from global *Fst* estimates (from 0.0589 to 0.0653 after filtering) and PCA distributions (Supplementary Fig. S10). At the intraspecific level, deamination filtering also had little impact on *Fst* values, suggesting that base modifications due to DNA damage are not severe enough to influence overall differentiation or are not present among the variants called. Thus, provided historic samples with extreme low coverage (<2.0X) are removed, variant call accuracy can be adjusted to diminish base mis-incorporation bias to a negligible level.

**Museum samples versus modern samples.** PCA plots of modern versus museum samples demonstrate that there is a gradient of variation related to the first principal component (PC1) for both *J. lalandii* (Fig. 4b) and *J. paulensis* (Fig. 4d), but not for *J. edwardsii* (Fig. 4c). Given that the aims of a museum/modern sample comparison are to assess temporal changes in population structure, these results are clearly interesting and should be investigated further using population-sequencing data. Our results demonstrate that target capture of ddRAD loci of lobster specimens from museum collections will likely provide usable sequencing data for population genomic approaches, but that an increase in sequencing effort of historical samples might be necessary to enable confident variant calling. Ideally, the coverage in historical samples should be similar to modern specimens, however DNA fragmentation or low DNA quantity will negatively impact the sequencing depth; in some cases the increase of sequencing effort might not affect the coverage of targets in these samples. This caveat requires further investigation in future studies by comparing population genetic parameters such as *Fst*, demography inference, selection<sup>34</sup> and the scaled population mutation rate  $\theta$ <sup>35</sup>.

## Conclusions

**Keep it simple and effective.** Here, ddRAD libraries from two closely related lobster species, *J. edwardsii* and *S. verreauxi* were used as genome resources to design probes to capture and enrich genomic libraries of other five closely related species. The target-enriched sequencing generated thousands of informative markers for population genomics application with a small sequencing effort of 4.6 Gbp only. The enriched sequencing



of previously discovered ddRAD loci enabled the recovery of 1,250 out of 2,366 ddRAD loci in the species from which the probes were designed, including sample replicates between both methods. Thus, it could potentially be used to enrich and re-sequence ddRAD loci outliers discovered in previous studies<sup>15</sup> in a range of populations to investigate genome signatures of selection. Also, this method circumvents one of the main disadvantages of ddRADseq, that being the high level of missing data due to allele dropout.

Our target-enriched loci were consistently recovered across all non-target *Jasus* species, whilst the use of an assembly-based reference including off-target sequences substantially increased mapping efficiency and average target coverage of informative loci. This enabled the detection of 1,731 loci shared across five *Jasus* species, which has the potential to enable direct comparison of locus-specific population genetic variation (including putatively selected loci) over multiple species, something that is not possible using restriction enzyme-based approaches. Thus, we provided a robust method approach to interrogate comparative studies of dispersal, self-recruitment and adaptation in all *Jasus* populations. Given the limited number of genetic markers available for wide comparisons of *Jasus* species distributions, the ddRAD loci enriched data represents a reliable genome resource, highly repeatable among individuals and replicates. This represents a permanent resource that may be further used for comparison among different studies in *Jasus* species, and perhaps to other closely related species, such as *S. verreauxi* or *Projasus* spp.

We have also developed a suitable protocol for the use of museum samples in population genomics studies. These samples provide representation of temporal shifts in populations, and critically are sometimes the only available material for a species. Based on the present results, informative SNPs from museum samples could be obtained provided a minimum target sequencing coverage exists and reads are not subject to base mis-incorporations bias. Here, we compiled a series of protocols for ddRAD loci enriched sequencing that enables quality control of sequencing data by computationally removing contaminants, PCR duplicates and spurious variants due to base mis-incorporation. This is particularly relevant when dealing with museum samples that are often contaminated, present low DNA yield and severe DNA damage. Thus, the methods described here can be applied to further investigate temporal changes in population structure, a critical issue in intensively fished species<sup>36</sup>.

## Methods

**Target species.** The genus *Jasus* encompasses six lobster species (*J. caveorum*, *J. edwardsii*, *J. frontalis*, *J. lalandii*, *J. paulensis* and *J. tristani*) that are distributed throughout the Southern Hemisphere<sup>37</sup>. *J. edwardsii* is the most widespread species within the genus, whilst the other species maintain limited geographical distributions, with some species known from only a single seamount (e.g. *J. caveorum*). These species all support valuable fisheries, and have been exploited for more than one hundred years<sup>38</sup>.

Due to their markedly long pelagic larval durations (PLD; in the order of 18–24 months<sup>39</sup>), panmixia (i.e., random mating between all individuals) is still widely accepted for some spiny lobster species, especially in the context of fishery management resources<sup>33</sup>. However, the paradigm that marine species with pelagic larval stages are genetically homogeneous across large geographic scales<sup>40</sup> is changing and cannot be systematically extrapolated across species. For instance, despite the high dispersal potential of *J. edwardsii* due to its long PLD, the occurrence of panmixia between Australian and New Zealand populations has been rejected<sup>15, 41, 42</sup>. Given the importance of understanding gene flow for fisheries management, these species represent an ideal case for the utilisation of genomic resources to better understand population structure. Currently, for species other than *J. edwardsii*, the genetic markers available are limited to mtDNA<sup>30</sup>, which is inadequate for accurately assessing population structure.

**Sample collection.** For the target-capture experiment, we collected 87 samples of modern and museum specimens of *Jasus* comprising six species *J. edwardsii*, *J. caveorum*, *J. frontalis*, *J. lalandii*, *J. paulensis* and *J. tristani*. Specimens of *S. verreauxi* (N = 16) were included to evaluate the custom probe set and the efficiency between ddRAD and target-capture methods for the validated loci in both methods (Supplementary Table S11).

**Modern samples.** A total of 40 lobsters samples were collected between 2010 and 2015. The French Southern and Antarctic Lands (Terres Australes et Antarctiques Françaises-TAAF) provided *J. paulensis* pleopod tissue collected in 2014. *J. lalandii* and *J. tristani* pereopod samples, collected during in 2015 in South Africa and Tristan da Cunha islands, respectively, were donated by the South African Department of Agriculture, Fisheries and Forestry. *J. frontalis* were sampled in Juan Fernandez archipelago in 2010<sup>43</sup>, these samples were donated by the Universidad de Concepcion (Chile).

**Museum samples.** Museum samples (N = 47) were donated by the National Institute of Water and Atmospheric Research (NIWA) and Te Papa Museum (New Zealand). The samples were collected in 1967 (*J. lalandii* and *J. paulensis*), 1991 (*J. edwardsii* and *J. lalandii*) and 1995 (*J. caveorum*). Samples were preserved in ethanol/isopropanol (mostly evaporated). In some cases preservation methods were not indicated.

**Loci selection from ddRAD libraries for probe design.** We used the ddRAD libraries published by Villacorta-Rath *et al.*<sup>15</sup> for *J. edwardsii*, and produced a subsequent library for the closely related *S. verreauxi* for our probe design. These species have estimated divergence time of approximately 40<sup>44</sup> to 108<sup>45</sup> million years. A modified Peterson *et al.*<sup>46</sup> ddRAD protocol was utilized for library preparation and sequencing of both *J. edwardsii* and *S. verreauxi* DNA samples, as described in Villacorta-Rath *et al.*<sup>15</sup>. Full information of samples from which ddRAD libraries were used for probe set design are described in Supplementary Table S12. The ddRADseq indexed libraries of 42 *J. edwardsii* and 55 *S. verreauxi* individuals were sequenced on the Illumina MiSeq platform. Raw ddRAD reads were trimmed using Trimmomatic 0.32<sup>47</sup>, in order to remove poor quality sections or

removed the entire read when the average Phred score was lower than 33. Contaminant reads were identified using Kraken 0.10.4 beta<sup>48</sup> and subsequently removed. Sequences were paired and trimmed to a minimum size of 140 base pairs (bp) using Pear 0.9.4<sup>49</sup>. Paired reads were then demultiplexed and assigned to corresponding samples following the dual indexed adapter sequences by using a locally developed pipeline (<https://github.com/molecularbiodiversity/rad-pipeline>).

**PyRAD assembly and ddRAD loci catalogue.** Remaining ddRAD reads were assembled in PyRAD 3.0.4<sup>19</sup> and used to build ddRAD loci catalogues within and between species. Samples were assembled in three datasets as follows: (1) *J. edwardsii* only; (2) *S. verreauxi* only; and (3) all samples from both species. For assemblies 1 and 2 we adopted a 95% similarity threshold within species and for assembly 3 we established an 85% similarity threshold. A maximum of three mismatches and a maximum of 0.5 site heterozygosity were allowed per cluster. To avoid clusters of paralogous loci we discarded all clusters with excessive shared heterozygous SNPs in more than four individuals and the paralogous filter was set to three. The ddRAD loci were considered for probe design (candidate ddRAD loci) only if they were shared across 10 individuals in assemblies 1 and 2 or 20 individuals in assembly 3. PyRAD assemblies' settings and outputs are detailed in the Supplementary Methods S1.

**Paralogous loci removal and repeat masking.** The candidate ddRAD loci were assembled *de novo* adopting a 75% similarity threshold using Geneious R7<sup>50</sup>. The default settings of RepeatExplorer<sup>51</sup> were implemented to distinguish paralogous families<sup>52</sup>, and contigs with more than 90% similarity across at least 55% of the fragment length were discarded. Highly similar clusters with spurious alignments resulting from low complexity DNA sequences (comprised by mononucleotide repeats) were also discarded. As capture probes are known to support up to about 12% sequence divergence<sup>13,53</sup>, 75% similarity among ddRAD loci was used as a limit. Remaining repetitive regions among ddRAD loci were identified and removed using RepeatMasker Web server<sup>54</sup>, with *Homarus americanus* as model organism, whilst the ddRAD loci derived from mitochondrial genome were identified using the *S. verreauxi* mitochondrial DNA complete genome (AB859775 accession number)<sup>55</sup>.

**MYbaits® probes panel.** A total of 2,366 loci were finally selected as templates for bait manufacturing: 2,358 nuclear loci and eight mitochondrial loci. The MYbaits® set were estimated to cover 322 kb from 5.3 Gb of *J. edwardsii* genome, including 80 loci identified as outlier SNPs putatively under selection in *J. edwardsii*<sup>15</sup>. In total, 4,732 120-mer MYbaits® probes (MYcroarray) were manufactured with 2X tiling density and an overlap of 60 bp between probes. Probes sequences were deposited at the Dryad data repository (<http://dx.doi.org/10.5061/dryad.3dk40>).

**Target enriched sequencing.** **DNA extraction.** DNA was extracted from adult lobsters (pleopod clip or pereopod muscle) and phyllosoma larvae (leg). All DNA extractions were performed with the DNeasy Qiagen kit using spin columns EconoSpin (Epoch Life Sciences). As the museum specimens resulted in low DNA yield and purity levels, the isolation protocol was optimized accordingly. For these samples, DNA recovery was improved by overnight incubation, final DNA elution in 30 µl of AE buffer with three consecutive washes followed by column centrifugation. DNA extraction and library preparation using museum specimens was performed using consumables dedicated to the museum specimens only with different batches for each species.

**Library preparation.** Genomic DNA samples from modern specimens were randomly fragmented by sonication using a BioRuptor NGS (Diagenode). Most museum DNA samples were naturally fragmented (i.e. due to degradation over time) therefore a fragmentation step was not necessary. We followed an established dual-indexed library preparation protocol detailed in Rohland and Reich<sup>56</sup> for hybridization capture reaction, incorporating few modifications to reduce costs and hands-on time. We omitted the size-selection, as it is assumed that sonication shearing generates a sufficiently narrow fragment-size distribution<sup>57</sup> and re-dimensioned final volumes of reactions and the proportion of magnetic microspheres (hereafter referred to as magnetic beads) within the purification steps (detailed in Supplementary in Table S13). Libraries were purified with magnetic beads, eluted and multiplexed by pooling eight of 16 libraries, prior to hybridization capture reaction<sup>58</sup> at equimolar ratios to a final yield of 200–500 ng (Supplementary Table S14 and S15). Two negative controls were included in every step.

**Target-capture.** The MYbaits® manufacturer's protocol was followed, replacing the blocker #3 of the kit with the custom blocking oligonucleotides described in Rohland and Reich<sup>54</sup>. In-solution hybridization reactions of libraries with biotinylated probes were incubated for 24 hours at 65 °C. Hybridized fragments were captured with Streptavidin-coated magnetic beads and washed four times with Wash Buffer (MYbaits® kit component) for five minutes to remove unspecific material. Captured DNA fragments were denatured for five minutes at 95 °C and eluted in TE buffer. The enriched multiplexed libraries were PCR-amplified for 18 cycles (Supplementary Table S15). Prior to sequencing, enriched libraries were Qubit-quantified, normalized and quantified in real-time PCR (Supplementary Table S14). Cluster generation was performed in a single Illumina MiSeq. 1500 lane, with 2 × 250 cycles of base incorporation.

**Sequencing data analysis.** **Sequencing data processing.** Sequencing data was processed using the locally developed pipeline CarlaSeq (<https://github.com/molecularbiodiversity/carlaSeq>) described in detail in Supplementary Methods S4. Briefly, it involved the first four steps of the pipeline that consisted of adapter trimming, removal of contaminants (human and microorganisms), merging of paired-end reads and de-multiplexing. The paired-end filtered reads were merged and trimmed to fragments up to 220 bp long. Reads with average Phred score lower than 33 were discarded.

**Read mapping.** Processed reads were mapped to the reference ddRAD loci, using Bowtie 2<sup>21</sup> with ‘-very-fast-local’ settings. SAMtools (<http://github.com/samtools/samtools>) was used to sort alignments (maQ ≥ 5). Picard 2.6.0 (<http://broadinstitute.github.io/picard>) was used to mark PCR duplicates from alignments by identifying fragments that are identical in insert length and related sequence composition. The capture reaction ensures that single strand fragments at a given locus are unlikely to be of equal length unless they are duplicates.

**Variant-calling.** Because most population genetic models are vulnerable to NGS errors (sequencing errors) and the amplification of PCR duplicates, initial variant-calling was only performed in modern samples to estimate overall genetic variation across species obtained with our capture probe set. Museum samples were initially omitted in order to reduce the chance of detecting false positive variants resulting from incorporated errors and DNA damage causing base modification. Thus, modern sample alignments were merged and post processing filters were used to perform probabilistic variant-calling using GATK 3.6<sup>59</sup>. GATK best practice guidelines were adapted as follows: first, indel intervals were locally realigned and the variants detected were indexed and converted into a Variant Call Format (VCF); second, variant-filtering was set to stringent thresholds for strand bias, coverage, mapping quality and variant position. Specifically, SNP sites with quality by depth < 4, root mean square mapping quality over all the reads at the site < 18, Phred quality < 40, Fisher Strand > 60, haplotype score > 13 and less than 5X coverage across at least eight individuals were discarded. Heterozygous SNPs at read ends (reads rank position < -12bp) were also discarded. Only biallelic sites that passed all filters were retained so that data were directly comparable among all *Jasus* species.

Deamination and transition/transversion ratio were calculated over filtered variants using VCFtools 0.1.14<sup>60</sup>. SNP pruning, diversity indices (*F*<sub>st</sub>, heterozygosity and MAF) and PCA were estimated for all *Jasus* species using Plink 1.9<sup>61</sup>.

**Method efficiency.** The efficiency of the target capture experiment was firstly assessed for the two species (*J. edwardsii* and *S. verreauxi*) from which the original probes were designed. This data set included 11 *J. edwardsii* specimens and 16 *S. verreauxi*, of which eight and 16 samples, respectively, were replicates of the original ddRAD libraries sequenced. The number of on-target reads between species was compared using BLAST. Then, mapping success (Bowtie 2) and the amount of missing data (loci counts) per number of samples were compared between ddRAD and target-capture methods for both species.

The efficiency of the target capture experiment non-target *Jasus* species was assessed using only modern samples (*N* = 40). The overall target coverage was given by the mean depth of targets for each species using BEDtools 2.26<sup>62</sup>. Mean coverage (sample coverage depth), heterozygous read rate and GC content of samples were estimated from sample alignments using BBmap (<https://github.com/BioInfoTools/BBMap>).

In order to compare the target-genome divergence among species and its effects on target enriched sequencing, a nonparametric correlation was performed among sequencing yield (given in number of processed reads), mapped reads, GC content, mean coverage and mapping quality for each species using SPSS v22.0 (IBM Corp., NY, USA). Univariate analysis of variance was applied to test whether GC content and mean coverage were significantly different among *Jasus* species. The sequencing yield effect was included as a covariate, since it was previously explored as a random factor within a one-way ANOVA analysis and found to be significant. Data were checked with Levene’s test and logarithm transformed to ensure normality and homogeneity of variance. Principal Component Analysis (PCA) including all variables was performed using PAST 3.12 (<http://folk.uio.no/ohammer/past/>).

**Assembly-based reference building.** Clustered assembly using Vsearch 1.1.3 (<https://github.com/torognes/vsearch>) was performed only on modern samples in order to build a reference loci catalogue avoiding low quality reads and potential artefacts from museum samples. The similarity threshold was set to 90% to account for sequence diversity within and among *Jasus* species. Resulting clusters with less than 40 reads were discarded. To ensure that clusters correspond to different genomic regions a *de novo* assembly was performed using Geneious R7<sup>48</sup> to check overlapping (redundancy) between clusters. The Geneious assembler was set to custom sensitivity parameters of 120 bp minimum overlap among clusters, which corresponds to the capture probe length, and 85% similarity threshold. Clusters assembled into the same contig were catalogued as a single locus. A reciprocal BLAST<sup>20</sup> using the original probe set database was applied to identify and filter only on-target clusters. These steps are detailed in CarlaSeq (<https://github.com/molecularbiodiversity/carlaSeq>) pipeline in Supplementary Methods S4 (see steps five to eight).

**Assessing diversity using museum samples.** Evidence of DNA damage in museum samples was established by tracking and quantifying cumulative substitutions frequencies of C to T at the 5’ end and G to A at the 3’ end in mapped reads using mapDamage 2.0<sup>31</sup> (<https://github.com/ginolhac/mapDamage>). We also estimated the deviations of GC content and target genome heterozygous rate with BBmap (<https://github.com/BioInfoTools/BBMap>) between modern and museum samples as a result of low coverage. A Student’s *t*-test (SPSS v22.0, IBM Corp., NY, USA) was used to test whether GC content and heterozygous rate between modern and museum samples for each species were significantly different. Data were checked with Levene’s test, and degrees of freedom were adjusted using the Welch-Satterthwaite method for data groups with unequal variances. Results are expressed as means ‘±’ standard errors (SEM), with statistically significant differences stated at *P*-value < 0.01. In order to evaluate the effect of DNA purity on target enriched sequencing among museum samples, nonparametric correlation analyses were performed among sequencing yield, mapped reads, mean coverage, mapping quality, GC content, year of collection and A<sub>260</sub>/A<sub>280</sub> ratio including modern samples (*N* = 79). PCA including all variables was performed using PAST 3.12 (<http://folk.uio.no/ohammer/past/>).

According to Parks *et al.*<sup>32</sup> variant-calling is not significantly affected by low damage (base modifications) in DNA, but is highly impacted by low coverage depth at any level of divergence. Thus, to avoid introducing severe bias by extreme low coverage depth, all samples with less than 2.0X target coverage (mean depth among targets) were removed from population genetics analysis as follows: three *J. caveorum* individuals, three *J. edwardsii* individuals, five *J. lalandii* from 1991 and two *J. paulensis* individuals from 1967. Note that Bi *et al.*<sup>2</sup> previously adopted a similar threshold for museum samples of non-model species, when utilizing a probabilistic method to call variants. *J. paulensis*, *J. lalandii* and *J. edwardsii* remaining samples collected in different decades were used to make comparisons between modern and museum samples based on unbiased *Fst*<sup>24</sup> values and PCA. As a proof-of-concept of the negligible DNA damage of museum samples used in this study, deamination filters were applied to perform a before/after comparison as recommended in Bi *et al.*<sup>2</sup>. Data was cleaned up for G to A and C to T substitution sites detected in historic samples. This is due to the fact that deamination substitutions are directional, they substitute G to A and C to T, but not A to G or T to C. Major and minor alleles were detected for all loci and directional substitutions G to A and C to T intervals were removed from all samples (modern and museum).

## References

- McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C. & Brumfield, R. T. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. and Evol.* **66**, 526–38 (2013).
- Bi, K. *et al.* Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* **22**, 6018–6032 (2013).
- Bailey, S. E. *et al.* The use of museum samples for large-scale sequence capture: A study of congeneric horseshoe bats (family Rhinolophidae). *Biol. J. Linn. Soc.* **117**, 58–70 (2016).
- Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R. & Moussalli, A. An Exon-Capture System for the Entire Class Ophiuroidea. *Mol. Biol. Evol.* **33**, 281–294 (2016).
- Suchan, T. *et al.* Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLoS One* **11**, e0151651 (2016).
- McCormack, J. E., Tsai, W. L. E. & Faircloth, B. C. Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* doi:10.1111/1755-0998.12466 (2015).
- Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C. & Brumfield, R. T. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst. Biol.* **63**, 83–95 (2014).
- Lemmon, A. R., Emme, S. A. & Lemmon, E. M. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* **61**, 727–744 (2012).
- Andrews, K. R., Paul, A., Miller, M. R. & Luikart, G. Trade-offs and utility of alternative RADseq methods: reply to Puritz *et al.* 2014. *Mol. Ecol.* 5943–5946 (2014).
- Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, 1–7 (2008).
- Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248 (2007).
- Puritz, J. B. *et al.* Comment: Demystifying the RAD fad. 1–18, doi:10.1111/mec.12965 (2014).
- Andrews, K. R. & Luikart, G. Recent novel approaches for population genomics data analysis. *Mol. Ecol.* **23**, 1661–1667 (2014).
- DaCosta, J. M. & Sorenson, M. D. Amplification Biases and Consistent Recovery of Loci in a Double-digest RAD-seq Protocol. **9** (2014).
- Villacorta-Rath, C. *et al.* Outlier SNPs enable food traceability of the southern rock lobster, *Jasus edwardsii*. *Mar. Biol.* **163**, 223 (2016).
- Rubin, B. E. R., Ree, R. H. & Moreau, C. S. Inferring phylogenies from RAD sequence data. *PLoS One* **7** (2012).
- Hoffberg, S. *et al.* Adapterama IV: Sequence Capture of Dual-digest RADseq Libraries with Identifiable Duplicates (RADcap). bioRxiv 1–22, doi:10.1101/jnc.13494 (2016).
- Ali, O. A. *et al.* RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. *Genetics* **202**, 389–400 (2016).
- Eaton, D. A. R. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849 (2014).
- Camacho, C. *et al.* BLAST plus: architecture and applications. *BMC Bioinformatics* **10**, 1 (2009).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).
- Arnold, B., Corbett-Detig, R. B., Hartl, D. & Bomblies, K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22**, 3179–90 (2013).
- Gautier, M. *et al.* The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* **22**, 3165–3178 (2013).
- Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution (N. Y.)* **38**, 1358–1370 (1984).
- Cosart, T. *et al.* Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**, 347 (2011).
- Vallender, E. J. Expanding whole exome resequencing into non-human primates. *Genome Biol.* **12**, R87 (2011).
- Bi, K. *et al.* Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* **13**, 403 (2012).
- Bragg, J. G., Potter, S., Bi, K. & Moritz, C. Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour.* **16**, 1059–1068 (2016).
- Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36** (2008).
- Groeneveld, J. C., Von der Heyden, S. & Matthee, C. A. High connectivity and lack of mtDNA differentiation among two previously recognized spiny lobster species in the southern Atlantic and Indian Oceans. *Mar. Biol. Res.* **8**, 764–770 (2012).
- Ginolhac, A. *et al.* Improving the performance of true single molecule sequencing for ancient DNA. *BMC Genomics* **13**, 177 (2012).
- Parks, M. & Lambert, D. Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics* **16**, 19 (2015).
- Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. W. H. Freeman and Company 53, (Sara Tenny, 2008).
- Crawford, J. E. & Lazzaro, B. P. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front. Genet.* **3**, 1–13 (2012).
- Liu, X., Fu, Y., Maxwell, T. J. & Boerwinkle, E. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. *Genome Res.* **20**, 101–109 (2010).
- Allendorf, F. W. & Hard, J. J. Human-induced evolution caused by unnatural selection through harvest of wild animals. *Proc. Natl. Acad. Sci. USA* **106**(Suppl), 9987–9994 (2009).



37. Phillips, B. F. in *Lobsters: Biology, Management, Aquaculture & Fisheries: Second Edition* (ed. Phillips, B. F.) 1–474 (John Wiley & Sons, Ltd, 2013), doi:10.1002/9781118517444.
38. FAO. FAO Global Aquaculture Production statistics database updated to 2013: Summary information. Food Agric. Organization United Nations 2013, (2015).
39. Booth, J. D. & Ovenden, J. R. Distribution of *Jasus* spp. (Decapoda: Palinuridae) phyllosomas in southern waters: Implications for larval recruitment. *Mar. Ecol. Prog. Ser.* **200**, 241–255 (2000).
40. Deagle, B. E. *et al.* Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water. *Mol. Ecol.* **24**, 4943–4959 (2015).
41. Thomas, L. & Bell, J. J. Testing the consistency of connectivity patterns for a widely dispersing marine species. *Heredity (Edinb.)*. **111**, 345–354 (2013).
42. Morgan, E. M. J., Green, B. S., Murphy, N. P. & Strugnell, J. M. Investigation of Genetic Structure between Deep and Shallow Populations of the Southern Rock Lobster, *Jasus edwardsii* in Tasmania, Australia. *PLoS One* **8**, e77978 (2013).
43. Porobić, J., Canales-Aguirre, C. B., Ernst, B., Galleguillos, R. & Hernández, C. E. Biogeography and historical demography of the Juan Fernández rock lobster, *Jasus frontalis* (Milne Edwards, 1837). *J. Hered.* **104**, 223–233 (2013).
44. Palero, F., Abelló, P., Macpherson, E., Gristina, M. & Pascual, M. Phylogeography of the European spiny lobster (*Palinurus elephas*): Influence of current oceanographical features and historical processes. *Mol. Phylogenet. Evol.* **48**, 708–717 (2008).
45. Bracken-Grissom, H. D. *et al.* The emergence of lobsters: Phylogenetic relationships, morphological evolution and divergence time comparisons of an ancient group (Decapoda: Achelata, astacidea, glypheidea, polychelida). *Syst. Biol.* **63**, 457–479 (2014).
46. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**, e37135 (2012).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
48. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
49. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
50. Kears, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
51. Novak, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
52. Dodsworth, S. *et al.* Genomic Repeat Abundances Contain Phylogenetic Signal. *Syst. Biol.* **64**, 112–126 (2015).
53. Hancock-Hanser, B. L. *et al.* Targeted multiplex next-generation sequencing: Advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol. Ecol. Resour.* **13**, 254–268 (2013).
54. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015. RepeatMasker Open-4.0 at <http://www.repeatmasker.org>.
55. Doyle, S. R., Griffith, I. S., Murphy, N. P. & Strugnell, J. M. Low-coverage MiSeq next generation sequencing reveals the mitochondrial genome of the Eastern Rock Lobster, *Sagmariasus verreauxi*. *Mitochondrial DNA* **1736**, 1–2 (2014).
56. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
57. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–8 (2010).
58. Shearer, A. E. *et al.* Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. *BMC Genomics* **13**, 618 (2012).
59. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
60. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
61. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
62. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

## Acknowledgements

Funding for this research was provided by an Australian Research Council Discovery Project grant (Project No. DP150101491) awarded to J.M.S., N.P.M., B.S.G. and J.J.B., an Australian Research Council Linkage Project grant (Project No. LP120200164) from B.S.G., a Fisheries Research and Development Corporation grant 2015–025 as well as the Tasmanian Rock Lobster Fisherman's Association, the Department of Primary Industries, Park Water and Environment (Tasmania, Australia), Seafood Innovations Limited (Wellington, New Zealand) and the New Zealand Rock Lobster Industry Council. The first author acknowledges the Postdoctoral grant from the Brazilian Program Science without Borders (CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico). L.N.W. thanks La Trobe University Securing Food, Water and Environment Grant, Holsworth Wildlife Research Endowment, the New South Wales Department of Primary Industries and IMAS. We would like to thank Gary Carlos (University of Tasmania), Colin Fry (University of Tasmania), Daniel Ierodiaconou (Deakin University), Kent Way Andrew Kent, Geoff Liggins, Marcus Miller, Giles Ballinger, Darrel Sykes, Andrew Cockcroft (DAFF), Rick Webber (Te Papa Museum) and Sadie Mills (NIWA Invertebrate Collection samples collected from various sources including the scientific observers from the Ministry for Primary Industry and FRST-funded NIWA projects from the Ministry of Business, Innovation and Employment of New Zealand) for field assistance and sample collection; T.A.A.F. (Terres Australes et Antarctiques Françaises), their French fisheries observer service “COPEC” and the fisheries observers Guy Duhamel, Nicolas Gasco, Alexis Martin, Patrice Pruvost and Charlotte Chazeau on board the F.V. “AUSTRAL” to have made available biological scientific samples of *Jasus paulensis* from catches in the Saint-Paul/Amsterdam\* French EEZ; the help of crew has also been appreciated; Steve Doyle and Michael Amor for ddRADseq laboratory assistance and; Cristina Valdiosera her helpful comments on methods using ancient DNA. We would like to thank the editor Toshifumi Minamoto and the anonymous reviewers for their constructive suggestions.



## Author Contributions

C.A.S. wrote the main text, generated the data and performed the analyses. C.A.S., N.M. and J.M.S. designed the study. C.A.S., N.M., J.M.S., J.J.B. and B.G. conceived the idea and reviewed the manuscript. C.V.R., L.N.W., I.I. and C.E.H. collected the samples and contributed to data interpretation. C.A.S. and L.N.W. performed the laboratory work. C.A.S., N.M. and J.M.S. led the writing of the manuscript with contributions from all authors.

## Additional Information

**Supplementary information** accompanies this paper at doi:[10.1038/s41598-017-06582-5](https://doi.org/10.1038/s41598-017-06582-5)

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017