ICES Journal of Marine Science, 2025, Vol. 82, Issue 6, fsaf105
https://doi.org/10.1093/icesjms/fsaf105
Received: 24 March 2025; **revised:** 27 May 2025; **accepted:** 10 June 2025
Original Article

Comparative genomics and structural variation analysis reveal biotechnological potential in *Crassostrea gigas*

Hyungtaek Jung[1,*,†], Min-Seung Jeon[2,†], Hyeongwoo Choi[2], Chi-une Song[2], Huijeong Doh[2], Jung Hyun Kwak[3], Seong-il Eyun[2,*]

[1]National Centre for Indigenous Genomics, John Curtin School of Medical Research, The Australian National University, Acton ACT 2601, Australia
[2]Department of Life Science, Chung-Ang University, Seoul 06974, Korea
[3]Subtropical Fisheries Research Institute, National Institute of Fisheries Science, Jeju 63068, Korea

*Corresponding authors. Hyungtaek Jung, National Centre for Indigenous Genomics, John Curtin School of Medical Research, The Australian National University, Acton ACT 2601, Australia. E-mail: hyungtaek.jung@anu.edu.au; Seong-il Eyun, Department of Life Science, Chung-Ang University, Seoul 06974, Korea. E-mail: eyun@cau.ac.kr
†These authors contributed equally to this work.

Abstract

Next-generation sequencing has significantly advanced omics and post-omics technologies, facilitating detailed analysis of whole-genome profiles through comparative genomics and reshaping our understanding of evolutionary and ecological dynamics for more precise biotechnological applications. Specifically, analyzing genomic similarities and differences through synteny and structural variations (SVs) has clarified the relationship between genetic variations and phenotypic changes. This study represents the first instance of employing chromosome-level comparative genomics and long-read SV detection in the Pacific oyster *Crassostrea gigas*, a species crucial for both ecology and economy, aiming to pinpoint genes and genomic regions linked to growth and disease. Utilizing the latest Pacific oyster genomes and their PacBio long reads, we uncovered significant genetic variability that explains differences between individuals. By integrating results from five state-of-the-art read mappers (LRA, Minimap2, NGMLR, Vulcan, and WinnowMap) and two SV callers (cuteSV and Sniffles2), we identified between 193 355 and 219 501 SVs, accounting for 183 Mb (31.2%) and 228 Mb (35.2%) of the total genome in two distinct reference genomes, respectively. This approach, leveraging PacBio long reads, surpassed short-read technologies in identifying candidate genes associated with SVs, such as bone morphogenetic proteins and superoxide dismutase, thereby highlighting their potential roles in growth and disease resistance. Our findings offer a comprehensive view of comparative genomics and long-read SVs, revealing their significance in oyster genome evolution and providing valuable insights for future marine biotechnological research, including the development of genome editing and surrogate broodstock technologies.

Keywords: oyster; *Crassostrea gigas*; comparative genomics; long-read; structural variation

Introduction

Recent advancements in a broad spectrum of omics fields-including genomics, transcriptomics, proteomics, epigenomics, nutrigenomics, lipidomics, glycomics, and metabolomics-and post-omics technologies have significantly transformed biotechnological research. These integrated omics methodologies offer a holistic view by characterizing and quantifying biological molecules (Li et al. 2011, Amer and Baidoo 2021, Veenstra 2021, Jeon et al. 2024), thereby facilitating a deeper understanding of organismal structures and functions. This approach has become pivotal in life sciences for elucidating complex biological phenomena (Amer and Baidoo 2021, Veenstra 2021). Marine biotechnology, although still emerging, has notably benefited from these advancements, leveraging omics technologies to explore marine biodiversity with remarkable depth and breadth. Particularly, genomics and transcriptomics have undergone a transformative shift, fueled by the decreasing costs of next-generation sequencing, enhancements in bioinformatics, and greater availability of high-performance computing (HPC) (Jung et al. 2019, Jung et al. 2020a, Jung et al. 2020b, Jeon et al. 2023). The advent of long and ultra-long read sequencing technologies, such as those from Pacific Biosciences (PacBio) and Oxford Nanopore, has further propelled genomic studies, offering unparalleled accuracy in decoding complex genomes. These developments have not only advanced our comprehension of marine biotechnology but have also fostered progress in aquaculture and fisheries (Wang et al. 2015, Eyun et al. 2017, Peng et al. 2020, Danecek et al. 2021, Peñaloza et al. 2021, Polinski et al. 2021, Qi et al. 2021, Song et al. 2021). Yet, despite these technological strides, the vast biodiversity present in aquatic ecosystems presents a formidable challenge in fully harnessing the potential of massive genomic resources.

Genomic variations, including single nucleotide variants/polymorphisms (SNVs/SNPs), small insertions and deletions (InDels), copy number variations (CNVs), and structural variations (SVs) (Consortium 2015), are crucial for studies in molecular evolution, population genetics, comparative genomics, and genomics-assisted breeding (Benoit 2020, Houston et al. 2020, Varshney et al. 2021, Jones and Wilson 2022). These variations, identified through reference genomes and genome-wide diversity panels, offer insights into the connection between allelic variations and phenotypes (Varshney et al. 2021). Despite these advancements, chromosome-level

comparative genomics and long-read SV detection remain nascent, limiting our understanding of phenotypic variability and functional genomics, particularly in marine biotechnology. Comparative genomics, comparing genome sequences across species, has become a key analytical tool, enabling the discovery of genes and enzymes critical to biological systems and driving biotechnological innovations. In marine biology, recent comparative genomic studies have highlighted significant findings, such as the expansion of meiotic cell cycle genes and a unique histone variant in amphitriploid fish (Wang et al. 2022), mechanisms of virulence and antimicrobial resistance in *Flavobacterium columnare* isolates in carp and trout (Declercq et al. 2021), duplication of immune genes in *Crassostrea hongkongensis* (Peng et al. 2020), adaptations in longevity, neural functions, and immunity in the American lobster (Polinski et al. 2021), and insights into evolution, salinity adaptation, and sex determination in *Portunus trituberculatus* (Lv et al. 2022). These discoveries underscore the power of comparative genomics in unraveling the complexities of marine life and its potential applications in biotechnology.

SVs, encompassing genomic rearrangements like deletions, insertions, inversions, duplications, and translocations of over 50 bp (Sudmant et al. 2015), are intimately linked with disease, evolutionary dynamics (e.g. gene losses, transposon activity), gene regulation (e.g. transcription factor rearrangement), and complex phenotypes (e.g. mating, reproduction isolation) (Jiang et al. 2020, Jiang et al. 2021, Kim et al. 2024). Long-read sequencing technologies, unlike their short-read counterparts, have profoundly enhanced our ability to characterize SVs (Beyter et al. 2021), thereby enriching our comprehension of mutation, evolutionary mechanisms, heritability gaps, and uncovering new biological insights (Logsdon et al. 2020). Despite these advances, optimizing sequencing for maximal yield and performance poses challenges, particularly due to the intricate SV patterns that emerge from the noise inherent in long reads. In the realm of aquatic research, long-read SV analysis has yielded significant discoveries, including the identification of a crucial 7.8 Mb inversion on chromosome 12 for Atlantic herring's ecological adaptation (Pettersson et al. 2019), the observation of widespread hemizygosity in mollusks (Calcino et al. 2021), and the documentation of extensive apoptosis inhibitor diversification in *Mercenaria mercenaria* (Song et al. 2021), highlighting the critical role of SVs in understanding marine life's genetic complexity.

The Pacific oyster (*C. gigas*), a key player in global seafood production (Troost 2010, Zhao et al. 2012), has been the focus of numerous selective breeding programs aimed at enhancing its economically valuable traits due to its significant contribution to aquaculture (Evans and Langdon 2006, Li et al. 2011, de Melo et al. 2016). SNPs, the most prevalent form of genetic variation (Lapègue et al. 2014, Wang et al. 2015, Gutierrez et al. 2017, Qi et al. 2021), have been widely used in various studies to understand and improve these traits (Hedgecock et al. 2015, She et al. 2015, Wang et al. 2016, Li et al. 2018, Shi et al. 2020). Despite some genomic variation maps (Qi et al. 2021) and SVs (Jiao et al. 2021) being identified with short-read sequencing technologies, the potential of chromosome-level comparative genomics and long-read SVs in marine biotechnology for the Pacific oyster has yet to be fully tapped. This study presents a pioneering bioinformatics analysis that constructs a comprehensive overview of chromosome-level genomics and long-read SV landscapes in the Pacific oyster. By utilizing two recent high-quality, chromosome-level genomes assembled with PacBio long reads and Hi-C scaffolding technology (Peñaloza et al. 2021, Qi et al. 2021), our work marks a significant advancement in marine biotechnology. This approach not only deepens our understanding of the oyster's biology and evolution but also opens new avenues for genetic improvement and the exploitation of marine resources through biotechnological innovations, setting the stage for groundbreaking developments in aquaculture breeding and marine bioproducts.

Materials and methods

Pacific oyster dataset and comparative genomics

Two genomes of the Pacific oyster, *C. gigas*, were obtained from the National Center for Biotechnology Information (NCBI) database, as detailed in Table 1. For improved clarity in our analysis, we have designated these genomes as *C. gigas*1 (Qi et al. 2021) and *C. gigas*2 (Peñaloza et al. 2021). The whole-genome alignment between the two reference genomes was conducted using MUM&Co (ver. 3.8) of MUMmer (ver. 4) with default parameters (O'Donnell and Fischer 2020). The alignment of *C. gigas*1 (as the reference genome) against *C. gigas*2 (as the query genome) is designated as Cg1RCg2Q. Conversely, the alignment of *C. gigas*2 (as the reference genome) against *C. gigas*1 (as the query genome) is referred to as Cg2RCg1Q.

Structural variations

Due to the limitations of short-read sequencing in accurately detecting long-range SVs (Jiao et al. 2021, Gao et al. 2022), our study exclusively utilized PacBio long reads for reciprocal alignments against two reference genomes of the Pacific oyster: raw reads from *C. gigas*1 aligned to the *C. gigas*2 reference genome (denoted as Cg2RCg1PB) and vice versa (Cg1RCg2PB). We employed five different aligners for the alignments: LRA (ver.1.3.4) (Ren and Chaisson 2021), Minimap2 (ver. 2.28) (Li 2018), NGMLR (ver. 0.2.7) (Sedlazeck et al. 2018), Vulcan (ver. 1.0.3) (Fu et al. 2021), and WinnowMap (ver. 2.03) (Jain et al. 2020), all set to their default parameters. SV detection, including insertions (INS), deletions (DEL), duplications (DUP), inversions (INV), and translocations/breakends (TRA/BND), was conducted using cuteSV (ver. 1.0.13) (Jiang et al. 2020) and Sniffles2 (ver. 2.6.0) (Smolka et al. 2024) with defaults for all five aligners. Low-quality SVs (quality score <5) were further removed using vcftools (ver. 0.1.16) (Danecek et al. 2021) to ensure high specificity, retaining only SVs marked as "PASS" in both variant-calling processes. Alignment coverage was calculated via the SAMtools (ver. 1.15.1) depth command (Danecek et al. 2021). The SURVIVOR software (ver. 1.0.7) (Jeffares et al. 2017) merge module was then used to consolidate SV findings, addressing methodological discrepancies. To increase confidence in merged SV calls and reduce methodological artifacts, merging required a minimum support of both two SV callers (cuteSV and Sniffles2) and a minimum SV size of 30 bp across all SV types. To determine an optimal merging threshold that maximizes specificity and reliability, we evaluated maximum merging distances ranging from 100 bp to 1000 bp in 100 bp increments. After assessing the stability of SV counts across genomic annotations, consistent overlap between callers, and minimal merging artifacts, we selected 400 bp as the opti-

Table 1. Summary of two *C. gigas* datasets used in this study.

Features	NCBI	Reference
Dataset 1 (*C. gigas*1)		
Genome assembly information	https://www.ncbi.nlm.nih.gov/assembly/GCA_011032805.1/	Qi et al. (2021)
Raw data accession	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA598006/ https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=598006	
Dataset 2 (*C. gigas*2)		
Genome assembly information	https://www.ncbi.nlm.nih.gov/assembly/GCF_902806645.1/	Peñaloza et al. (2021)
Raw data accession	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB35351/ https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=596972	

mal merging threshold. We summarized the SVs using a Venn diagram created with an online tool and visualized each SV type (https://bioinformatics.psb.ugent.be/webtools/Venn), except for translocations, using shinyCircos (Yu et al. 2018), excluding unanchored scaffolds from *C. gigas*2. For detailed SV inspection, Samplot (ver. 1.3.1) (Beyter et al. 2021) was employed after excluding unanchored scaffolds from *C. gigas*2. These analyses were conducted on HPC facilities at the Australian National University and Chung-Ang University, leveraging the PBSpro HPC environment. The scripts utilized in this study are available on GitHub, accompanied by comprehensive explanations (https://github.com/OZTaekOppa/lrsv_oysters).

Results and discussion

The alignment of reads to the two Pacific oyster reference genomes (Peñaloza et al. 2021, Qi et al. 2021) resulted in mapping rates varying between 77.59% and 99.45% for Cg1RCg2PB, and 85.38% to 99.66% for Cg2RCg1PB (Table 2). These rates were obtained under uniform computational settings, with the NGMLR aligner showing the lowest mapping rates but the longest runtimes for both datasets. Conversely, Vulcan displayed the highest mapping rates, while Minimap2 was the most time-efficient aligner for both sets of data (Table 2). INS and DEL emerged as the predominant types of SVs, regardless of the alignment method chosen. Notably, the NGMLR aligner identified the least variety of SV types, whereas Minimap2 and the Vulcan aligner detected the most, excluding inversions (INV) and duplications (DUP), across both datasets (Table 3).

We evaluated merging distances from 100 to 1000 bp and observed that the total number of consensus SVs stabilized at approximately 400 bp across all tested alignment tools (LRA, minimap2, NGMLR, Vulcan, and WinnowMap) (Supplementary Fig. 1). Increasing the merging distance beyond 400 bp did not significantly increase SV numbers, indicating minimal gain and potential risk of merging biologically distinct events. Additionally, the proportion of SVs detected in non-repeat gene regions (~26%–27%) versus repeat-rich regions (~58%–60%) remained consistent across merging distances, confirming that SV detection in genic regions is robust and not adversely impacted by the chosen merging parameter (Supplementary Table 1). Therefore, we concluded that a 400 bp merging distance provides the optimal balance between sensitivity and specificity, ensuring accurate and biologically interpretable SV detection.

Upon consolidating the variant call format (VCF) files using SURVIVOR, a comprehensive count of 193 355 putative SVs, spanning 183 Mb of the *C. gigas*1 genome, and 219 501 SVs, covering 228 Mb of the *C. gigas*2 genome, were catalogued (Table 3). The size of SVs identified ranged from 38 bp (INS) to 99 658 bp (INV) in Cg1RCg2PB, averaging 953 bp, whereas for Cg2RCg1PB, sizes ranged from 38 bp (INS) to 99 489 bp (DEL), with an average SV size of 1050 bp. Using MUM&Co, the whole-genome alignment resulted in a total of 11 263 alignments for Cg1RCg2Q and 11 520 alignments for Cg2RCg1Q (Table 4).

Figure 1 illustrates the distribution of SVs across the genome, revealing that the majority of SVs were under 500 bp in length. Interestingly, 47% of variants in the Cg1RCg2PB dataset and 41% in the Cg2RCg1PB dataset were identified (overlapped) by at least one alignment tool, as shown in Fig. 2. Notably, LRA, followed by Minimap2 or NGMLR, detected the highest number of unique SV types compared to other aligners.

A comparative analysis of the two *C. gigas* genome assemblies revealed a diverse array of SV types across the chromosomes, detailed in Table 3. The genomic landscapes of SVs in the Pacific oysters are depicted in a Circos plot shown in Fig. 3. However, no translocations (TRAs) were detected, likely due to the challenge of identifying numerous small regions. Across both oyster genomes, a consistent pattern emerged for all SV types, with INS and DEL constituting the majority, accounting for 98.16% and 97.82% in Cg1RCg2PB and Cg2RCg1PB datasets, respectively. Further analysis revealed specific bam file profiles and SV types for putative INSs (ranging from 495 to 2677 bp) and DELs (ranging from 655 to 19 406 bp) in two immune-related genes using Cg2RCg1PB, as illustrated in Fig. 4. Notably, a significant deletion of 19 406 bp was observed at the end of the bone morphogenetic protein gene on chromosome 1 (Fig. 4a), and a 655 bp deletion at the start of the superoxide dismutase/extracellular superoxide gene on chromosome 9 (Fig. 4b). Additionally, the application of long-read sequencing technology was pivotal in uncovering a super-sized (~100 kb) DEL (Fig. 4c), INV (Fig. 4d), and DUP (Fig. 4e), impacting several genes.

Genomic SVs are key components of genetic diversity within organisms, significantly impacting biological functions and traits, including disease susceptibility. Although SNPs and small insertions or deletions (indels <50 bp) have been widely studied, the exploration of larger genome SVs (over 50 bp)—crucial for understanding disease, evolutionary processes, gene regulation, chromosome rearrangements, and

Table 2. Alignment statistics obtained through various methods in this study.

	Cg1RCg2PB [a]		Cg2RCg1PB [a]	
	Mapping rate (%)	Computing time (hrs) [b]	Mapping rate (%)	Computing time (hrs) [b]
LRA	85.93	11.0	91.60	21.8
Minimap2	96.37	4.6	97.70	9.3
NGMLR	77.59	35.8	85.38	85.4
Vulcan	99.45	16.9	99.66	30.4
WinnowMap	91.31	7.7	95.62	14.4

[a]Cg1RCg2PB (*C. gigas*1 reference genome with *C. gigas*2 PacBio reads) and Cg2RCg1PB (*C. gigas*2 reference genome with *C. gigas*1 PacBio reads).
[b]Computing time indicates the total wall time on the HPC system at the Australian National University, utilizing 100 GB RAM and 8 CPU cores).

Table 3. Summary of statistics concerning SVs identified using various methods.

	Methods[a]	INS	DEL	INV	DUP	TRA[b]	Total
Cg1RCg2PB							
LRA	cuteSV	100 840	92 410	827	5	56	194 138
	Sniffles2	101 496	90 268	801	1	5749	198 315
Minimap2	cuteSV	111 048	105 340	721	386	6604	224 099
	Sniffles2	94 602	98 286	871	841	6003	200 603
NGMLR	cuteSV	86 485	85 505	521	2 080	2947	177 538
	Sniffles2	88 021	83 863	1 153	1 758	6638	181 433
Vulcan	cuteSV	102 772	95 907	680	643	8148	208 150
	Sniffles2	86 973	89 114	830	893	6044	183 854
WinnowMap	cuteSV	92 704	89 755	733	523	5864	189 579
	Sniffles2	79 308	84 178	993	1115	7969	173 563
SURVIVOR[c]	cuteSV & Sniffles2[d]	92 493	97 302	840	1000	1720	193 355
Cg2RCg1PB							
LRA	cuteSV	118 436	105 160	1010	22	250	224 878
	Sniffles2	115 814	96 531	950	3	8837	222 135
Minimap2	cuteSV	128 072	120 409	1080	449	11 909	261 919
	Sniffles2	104 484	102 773	925	702	6368	215 252
NGMLR	cuteSV	105 578	100 558	786	2 708	5112	214 742
	Sniffles2	105 025	89 697	1127	1 584	10 441	207 874
Vulcan	cuteSV	124 122	113 983	1105	765	15 340	255 315
	Sniffles2	100 989	96 403	925	784	6768	205 869
WinnowMap	cuteSV	112 578	106 758	1497	876	13 349	235 058
	Sniffles2	93 141	90 550	1263	1092	9710	195 756
SURVIVOR[c]	cuteSV & Sniffles2[d]	104 746	109 975	1162	1196	2422	219 501

[a]Methods represent the combination of alignment tools (LRA, Minimap2, NGMLR, Vulcan, and WinnowMap) and variant calling processes (cuteSV and Sniffles2)
[b]TRA (a.k.a. BND) was not included in the overall length calculation.
[c]SURVIVOR was used as the tool to merge the SV VCF outcomes.
[d]cuteSV and Sniffles2 were used to cross-validate the confidence of the SVs.

Table 4. Summary of assembly-based whole genome alignment for SV detection in this study.

Features	Cg1RCg2Q	Cg2RCg1Q
Deletions	4714	4831
Insertions	5747	5798
Duplications	179	169
Contractions	97	73
Inversions	155	218
Translocations	371	431
Total SVs	11 263	11 520

complex phenotypes (Jiang et al. 2020, Jiang et al. 2021)—remains limited due to detection challenges (Bertolotti et al. 2020, Qi et al. 2021). In this study, we present the first detailed SV landscape of the Pacific oyster, created by aligning chromosome-level genome assemblies from two distinct research groups and analyzing comprehensive whole-genome long-read data. We also pinpointed syntenic regions within the SV landscape, highlighting genes linked to growth and immune functions. This work sheds light on the Pacific oyster's genome evolution and domestication, influenced by selection pressures, providing valuable insights for future genetic studies and breeding programs.

SVs are pivotal genetic alterations that contribute significantly to traits crucial for domestication and artificial breeding (Chakraborty et al. 2019, Liu et al. 2019), exemplified by grapevine berry color (Zhou et al. 2019) and dietary shifts in gray wolves and dholes (Wang et al. 2022). However, the exploration of SVs in aquatic species, particularly in fast-growing Pacific oyster strains through whole-genome re-sequencing with Illumina short-read technology, is sparse (Jiao et al. 2021). While short-read analyses (<1 kb) have provided valuable insights into the genetic basis of SVs in Pacific oyster breeding (Jiao et al. 2021), the inherent limitations of this technology, primarily due to its inability to span larger SVs directly, necessitate indirect detection methods (split reads, read pairs, read depths, and local *de novo* assembly) (Gao et al. 2022). In contrast, long-read sequenc-

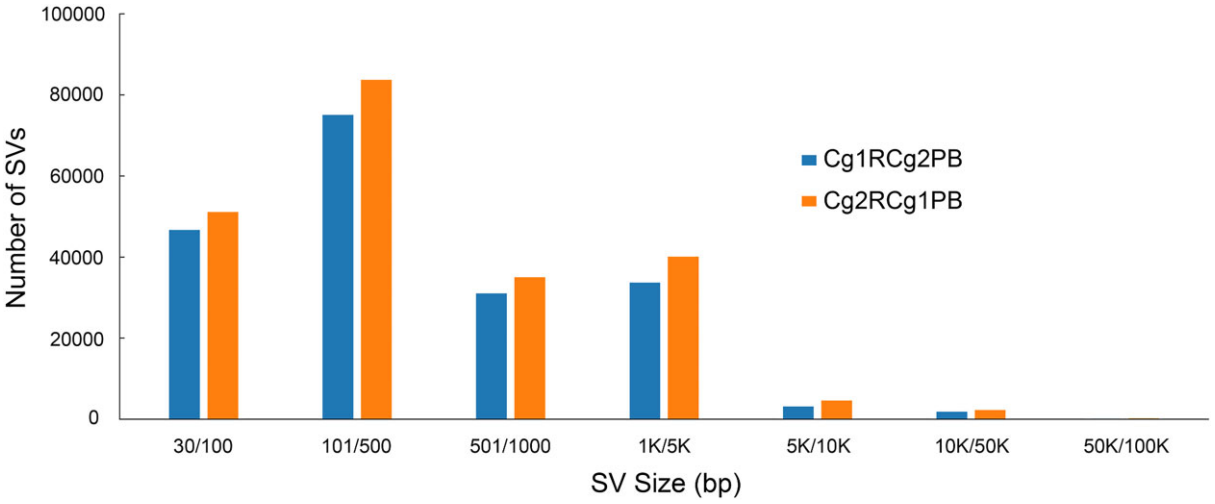

Figure 1. Size distribution for SVs inferred from the Pacific oyster reference genomes Cg1RCg2PB and Cg2RCg1PB.

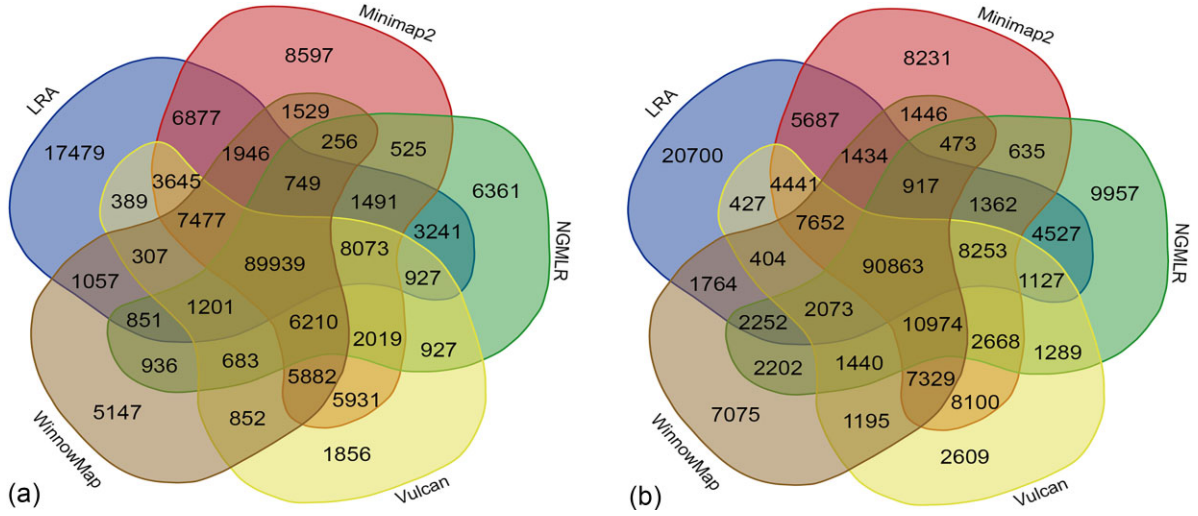

Figure 2. Comparison of the five SV datasets between the Pacific oyster reference genomes Cg1RCg2PB (a) and Cg2RCg1PB (b). The numbers in each section of the diagram represent the total number of all SV types identified in each comparison, without categorization by SV types.

ing technologies like PacBio and Oxford Nanopore, capable of producing reads exceeding 100 kb, offer significant advantages in identifying large SVs and complex genomic regions (Jung et al. 2019, Jung et al. 2020b, Beyter et al. 2021), thereby enhancing our understanding of missing heritability and unveiling new biological insights (Logsdon et al. 2020). Despite challenges in optimizing yield, our study undertook a comparative analysis using five long-read aligners across two PacBio datasets and chromosome-level genome assemblies (Peñaloza et al. 2021, Qi et al. 2021) to establish a benchmark for SV analysis in aquatic organisms. This approach highlighted the variability in read alignment efficacy and SV detection across different sequencing datasets and genome assemblies, underscoring the potential of long-read sequencing in revealing the genetic architecture of significant traits in aquatic species.

While employing five distinct alignment tools, we observed variations in performance, resource consumption, and outcomes, as detailed in Tables 2 and 3. After utilizing SURVIVOR for data merging, we identified 193 355 and 219 501 SVs in the *C. gigas*1 and *C. gigas*2 genomes, respectively, covering 183 Mb (31.2%) and 228 Mb (35.2%) of their total genomic content. This contrasts with a previous study using short-read sequencing that identified 511 170 short SVs and 979 486 CNVs (including INSs and DELs), with 63 100 short SVs and 58 182 CNVs deemed common across at least 20 samples (Jiao et al. 2021). Our analysis revealed a threefold increase in SV detection (merged with SURVIVOR) compared to common variations identified in the short-read study, demonstrating concordance, particularly in the high frequency of INSs and DELs observed with both sequencing approaches. Notably, Minimap2 outperformed other aligners in SV detection, excluding INV and DUP. When evaluating the efficiency and user-friendliness of long-read alignment tools, the combination of Minimap2, Vulcan, and WinnowMap with cuteSV stood out as particularly accessible for beginners due to lower computational requirements and ease of use (Table 2). Conversely, NGMLR, though effective, may not be the best option for balancing speed and accuracy in SV detection. This discrepancy underscores the necessity of cautious interpretation (Jiang et al. 2020, Jiang et al. 2021) when applying these

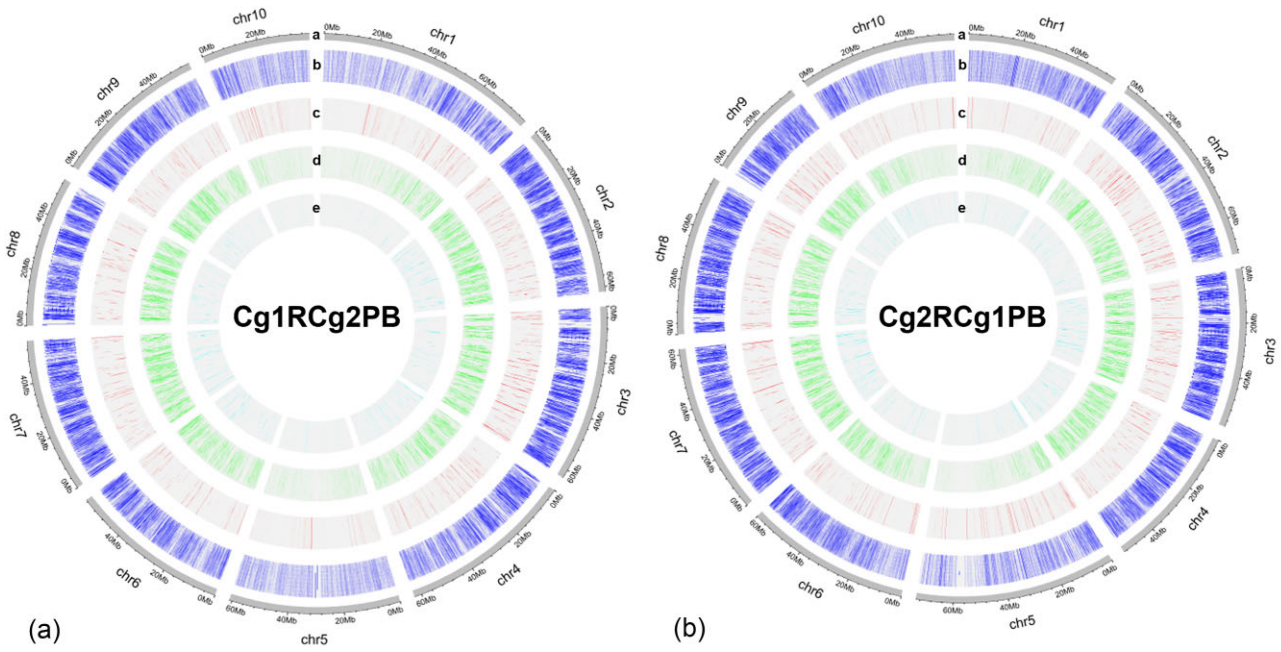

Figure 3. Genomic SV landscape of the Pacific oysters for the reference genomes Cg1RCg2PB (a) and Cg2RCg1PB (b). Track information is listed as follows (outer to inner circles) (a) chromosome number and size, (b) deletion, (c) duplication, (d) insertion, and (e) inversion.

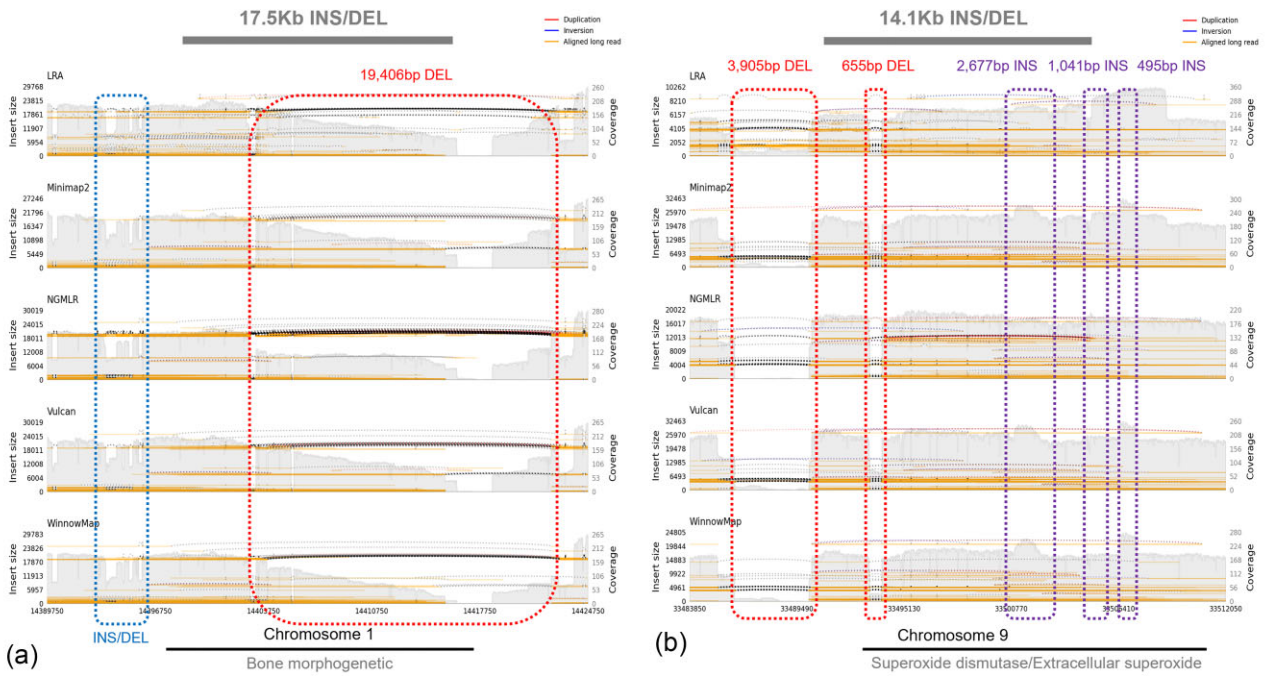

Figure 4. Images of SVs identified using the Pacific oyster reference genome Cg2RCg1PB. Putative insertion and deletion calls are shown in two immune-related genes: bone morphogenetic protein on Chromosome 1 (a) and superoxide dismutase and extracellular superoxide on Chromosome 9 (b). Examples of putative large deletions (DEL; c), inversions (INV; d), and duplications (DUP; e) across different chromosomes.

tools to non-model and marine species genomes (Jung et al. 2020b), as the success observed in model organisms might not directly translate due to unique challenges such as repetitive sequences, sequencing errors, and chimeric reads (e.g. split reads and break ends) (Jiang et al. 2020). Our study, therefore, lays down the essential groundwork for guiding future genomic research in marine species, indicating that while certain tools may be more suitable for initial explorations, comprehensive analyses may require a nuanced approach considering the specificities of each genome.

Identifying candidate genes linked with SVs underscores the complex relationship between genetic variations and biological traits, suggesting a significant impact on genome evolution and domestication processes in Pacific oysters. Notably, genes encoding bone morphogenetic proteins, which belong to the transforming growth factor β superfamily, are implicated in

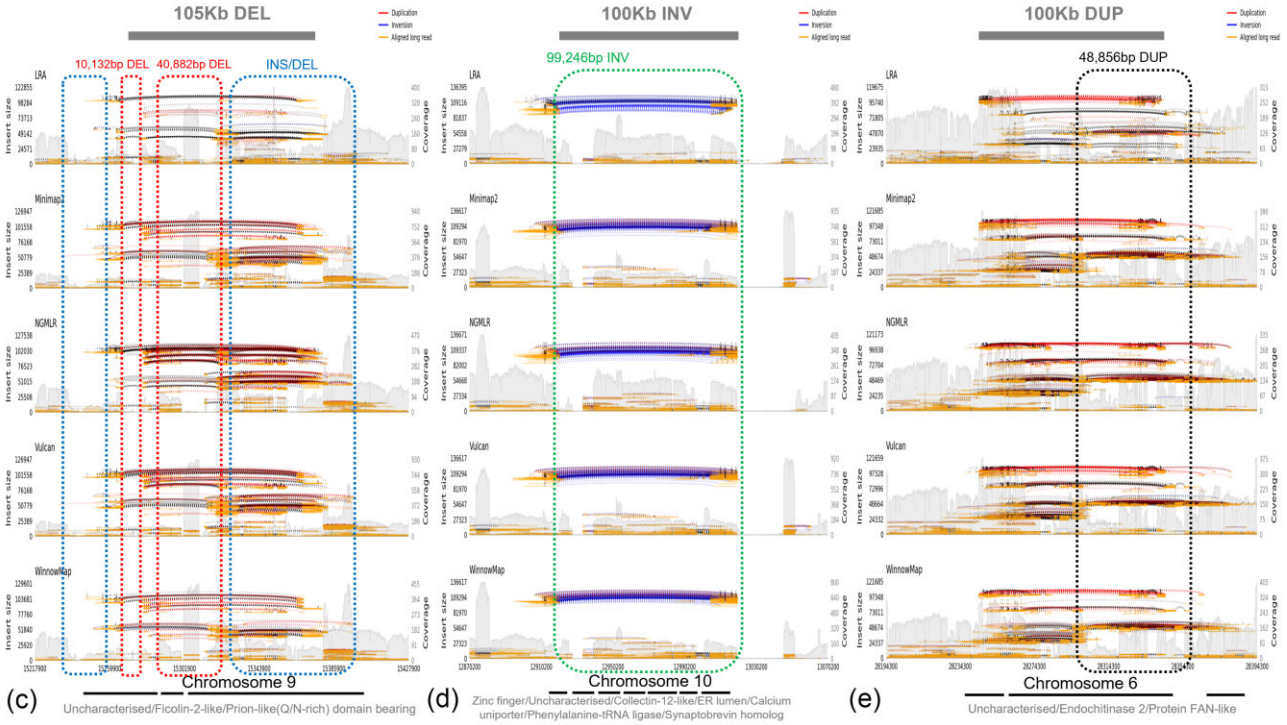

Figure 4. *Continued.*

vital biological functions, including embryonic development and immune responses (Tirapé et al. 2007). We identified several INSs (~195 bp at the start) and DELs (19 406 bp) within divergent SV regions on Chromosome 1, highlighting the genetic variability in regions encoding these critical proteins. Additionally, the gene for extracellular superoxide dismutase, essential for the oyster immune system and known for its role in lipopolysaccharide (LPS) binding and expression in hemocytes (Gonzalez et al. 2005, Tirapé et al. 2007), exhibited distinct INSs and DELs within significantly divergent SV loci on Chromosome 9. Remarkably, large SVs including DEL, INV, and DUP ranging 40 kb to 100 kb, spanned multiple genes, revealing genetic complexities not previously captured by short-read sequencing. These findings suggest a deeper layer of genetic influence on growth and immunity traits in Pacific oysters, calling for extensive research across diverse cultural and population settings to fully understand the implications of these candidate genes and SVs.

The widespread use of long-read sequencing technologies and the improvement in genome assembly quality have enabled a thorough investigation of the full spectrum of SVs. This advancement allows for the accurate identification of SVs and syntenic regions, revealing new biological insights. Recent research indicates that achieving 20× coverage with PacBio or Oxford Nanopore technologies is sufficient to comprehensively detect SVs. This level of coverage effectively balances computational efficiency with the cost-effectiveness of sequencing (Jiang et al. 2020, Jiang et al. 2021, Reis et al. 2023).

Author contributions

H.J., S.E.: Conceptualisation; H.J., S.E.: Funding acquisition; M.J., H.J., C.S., H.C., H.D., S.E.: Formal analysis; M.J., H.J., C.S., H.C., H.D., J.H.K., S.E.: Investigation; H.J., S.E., Project administration; H.J., S.E.: Supervision; M.J., H.J., S.E.: Writing – original draft; H.J., S.E.: Writing – review & editing

Supplementary data

Supplementary data is available at *ICES Journal of Marine Science* online.

Conflict of interest: None declared.

Funding

This work was supported by the Korea Institute of Marine Science & Technology Promotion (RS-2025–02215227) funded by the Ministry of Oceans and Fisheries and National Institute of Fisheries Science (R2025013).

Data availability

Not applicable. The scripts used in this study are accessible on GitHub (https://github.com/OZTaekOppa/lrsv_oysters).

References

Amer B, Baidoo EEK. Omics-driven biotechnology for industrial applications. *Front Bioeng Biotechnol* 2021;**9**:613307. https://doi.org/10.3389/fbioe.2021.613307

Benoit M. On the importance of variation: a high-resolution map of copy number variants in arabidopsis. *Plant Cell* 2020;**32**:1771–2. https://doi.org/10.1105/tpc.20.00257

Bertolotti AC, Layer RM, Gundappa MK *et al.* The structural variation landscape in 492 Atlantic salmon genomes. *Nat Commun* 2020;**11**:5176. https://doi.org/10.1038/s41467-020-18972-x

Beyter D, Ingimundardottir H, Oddsson A *et al.* Long-read sequencing of 3622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* 2021;**53**:779–86. https://doi.org/10.1038/s41588-021-00865-4

Calcino AD, Kenny NJ, Gerdol M. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. *Philos Trans R Soc B: Biol Sci* 2021;**376**:20200153. https://doi.org/10.1098/rstb.2020.0153

Chakraborty M, Emerson JJ, Macdonald SJ *et al.* Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* 2019;**10**:4872. https://doi.org/10.1038/s41467-019-12884-1

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;**526**:68–74. https://doi.org/10.1038/nature15393

Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008

Declercq AM, Tilleman L, Gansemans Y *et al.* Comparative genomics of *Flavobacterium columnare* unveils novel insights in virulence and antimicrobial resistance mechanisms. *Vet Res* 2021;**52**:18. https://doi.org/10.1186/s13567-021-00899-w

de Melo CMR, Durland E, Langdon C. Improvements in desirable traits of the Pacific oyster, *Crassostrea gigas*, as a result of five generations of selection on the West Coast. *Aquaculture* 2016;**460**:105–15. https://doi.org/10.1016/j.aquaculture.2016.04.017

Evans S, Langdon C. Direct and indirect responses to selection on individual body weight in the Pacific oyster (*Crassostrea gigas*). *Aquaculture*. 2006;**261**:546–55. https://doi.org/10.1016/j.aquaculture.2006.07.037

Eyun S, Soh HY, Posavi M *et al.* Evolutionary history of chemosensory-related gene families across the Arthropoda. *Mol Biol Evol* 2017;**34**:1838–62. https://doi.org/10.1093/molbev/msx147

Fu YL, Mahmoud M, Muraliraman VV *et al.* Vulcan: improved long-read mapping and structural variant calling via dual-mode alignment. *Gigascience* 2021;**10**:giab063. https://doi.org/10.1093/gigascience/giab063

Gao YH, Ma L, Liu GE. Initial analysis of structural variation detections in cattle using long-read sequencing methods. *Genes* 2022;**13**:828. https://doi.org/10.3390/genes13050828

Gonzalez M, Romestand B, Fievet J *et al.* Evidence in oyster of a plasma extracellular superoxide dismutase which binds LPS. *Biochem Biophys Res Commun* 2005;**338**:1089–97. https://doi.org/10.1016/j.bbrc.2005.10.075

Gutierrez AP, Turner F, Gharbi K *et al.* Development of a medium density combined-species SNP array for Pacific and European oysters (*Crassostrea gigas* and *Ostrea edulis*). *G3* 2017;**7**:2209–18. https://doi.org/10.1534/g3.117.041780

Hedgecock D, Shin G, Gracey AY *et al.* Second-generation linkage maps for the Pacific oyster reveal errors in assembly of genome scaffolds. *G3* 2015;**5**:2007–19. https://doi.org/10.1534/g3.115.019570

Houston RD, Bean TP, Macqueen DJ *et al.* Harnessing genomics to fast-track genetic improvement in aquaculture. *Nat Rev Genet* 2020;**21**:389–409. https://doi.org/10.1038/s41576-020-0227-y

Jain C, Rhie A, Zhang HW *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics*. 2020;**36**:i111–8. https://doi.org/10.1093/bioinformatics/btaa435

Jeffares DC, Jolly C, Hoti M *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* 2017;**8**:14061. https://doi.org/10.1038/ncomms14061

Jeon D, Song C, Jeong HG *et al.* An integrated phylogenomic approach for potential host-associated evolution of monstrilloid copepods. *Oceanogr Mar Biol, Annu Rev* 2024;**62**:350–75.

Jeon MS, Jeong D, Doh H *et al.* A practical comparison of the next-generation sequencing platform and assemblers using yeast genome. *Life Sci Allian* 2023;**6**:e202201744. https://doi.org/10.26508/lsa.202201744

Jiang T, Liu SQ, Cao SQ *et al.* Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. *BMC Bioinf* 2021;**22**:552. https://doi.org/10.1186/s12859-021-04422-y

Jiang T, Liu Y, Jiang Y *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 2020;**21**:189. https://doi.org/10.1186/s13059-020-02107-y

Jiao ZX, Tian Y, Hu BY *et al.* Genome structural variation landscape and its selection signatures in the fast-growing strains of the Pacific oyster, *Crassostrea gigas*. *Mar Biotechnol* 2021;**23**:736–48. https://doi.org/10.1007/s10126-021-10060-5

Jones HE, Wilson PB. Progress and opportunities through use of genomics in animal production. *Trends Genet*. 2022;**38**:1228–52. https://doi.org/10.1016/j.tig.2022.06.014

Jung H, Jeon MS, Hodgett M *et al.* Comparative evaluation of genome assemblers from long-read sequencing for plants and crops. *J Agric Food Chem* 2020a;**68**:7670–7. https://doi.org/10.1021/acs.jafc.0c01647

Jung H, Ventura T, Chung JS *et al.* Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Comput Biol* 2020b;**16**:e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Jung H, Winefield C, Bombarely A *et al.* Tools and strategies for long-read sequencing and *de novo* assembly of plant genomes. *Trends Plant Sci* 2019;**24**:700–24. https://doi.org/10.1016/j.tplants.2019.05.003

Kim E, Jeon D, Park YJ *et al.* Dietary exposure of the water flea to microcystin-LR. *Anim Cells Syst* 2024;**28**:25–36. https://doi.org/10.1080/19768354.2024.2302529

Lapègue S, Harrang E, Heurtebise S *et al.* Development of SNP-genotyping arrays in two shellfish species. *Mol Ecol Resour* 2014;**14**:820–30. https://doi.org/10.1111/1755-0998.12230

Li CY, Wang JP, Song K *et al.* Construction of a high-density genetic map and fine QTL mapping for growth and nutritional traits of *Crassostrea gigas*. *BMC Genom* 2018;**19**:626. https://doi.org/10.1186/s12864-018-4996-z

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191

Li Q, Wang QZ, Liu SK *et al.* Selection response and realized heritability for growth in three stocks of the Pacific oyster *Crassostrea gigas*. *Fish Sci* 2011;**77**:643–8. https://doi.org/10.1007/s12562-011-0369-0

Liu M, Zhou Y, Rosen BD *et al.* Diversity of copy number variation in the worldwide goat population. *Heredity* 2019;**122**:636–46. https://doi.org/10.1038/s41437-018-0150-6

Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;**21**:597–614. https://doi.org/10.1038/s41576-020-0236-x

Lv J, Li R, Su Z *et al.* A chromosome-level genome of provides insights into its evolution, salinity adaptation and sex determination. *Mol Ecol Resour* 2022;**22**:1606–25. https://doi.org/10.1111/1755-0998.13564

O'Donnell S, Fischer G. MUM&Co: accurate detection of all SV types through whole-genome alignment. *Bioinform* 2020;**36**:3242–3.

Peñaloza C, Gutierrez AP, Eöry L *et al.* A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. *Gigascience* 2021;**10**:giab020. https://doi.org/10.1093/gigascience/giab020

Peng J, Li Q, Xu L *et al.* Chromosome-level analysis of the genome reveals extensive duplication of immune-related genes in bivalves. *Mol Ecol Resour* 2020;**20**:980–94. https://doi.org/10.1111/1755-0998.13157

Pettersson ME, Rochus CM, Han F *et al.* A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Res* 2019;**29**:1919–28. https://doi.org/10.1101/gr.253435.119

Polinski JM, Zimin AV, Clark KF *et al.* The American lobster genome reveals insights on longevity, neural, and immune adaptations. *Sci Adv* 2021;**7**:eabe8290. https://doi.org/10.1126/sciadv.abe8290

Qi HG, Li L, Zhang GF. Construction of a chromosome-level genome and variation map for the Pacific oyster. *Mol Ecol Resour* 2021;**21**:1670–85. https://doi.org/10.1111/1755-0998.13368

Reis ALM, Rapadas M, Hammond JM *et al.* The landscape of genomic structural variation in Indigenous Australians. *Nature* 2023;**624**:602–10. https://doi.org/10.1038/s41586-023-06842-7

Ren JW, Chaisson MJP. lra: a long read aligner for sequences and contigs. *PLoS Comput Biol* 2021;**17**:e1009078. https://doi.org/10.1371/journal.pcbi.1009078

Sedlazeck FJ, Rescheneder P, Smolka M *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;**15**:461–8. https://doi.org/10.1038/s41592-018-0001-7

She ZC, Li L, Qi HG *et al.* Candidate gene polymorphisms and their association with glycogen content in the Pacific oyster. *PLoS One* 2015;**10**:e0124401. https://doi.org/10.1371/journal.pone.0124401

Shi RH, Li CY, Qi HG *et al.* Construction of a high -resolution genetic map of *Crassostrea gigas*: QTL mapping and GWAS applications revealed candidate genes controlling nutritional traits. *Aquaculture* 2020;**527**:735427. https://doi.org/10.1016/j.aquaculture.2020.735427

Smolka M, Paulin LF, Grochowski CM *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* 2024;**42**:1571–80. https://doi.org/10.1038/s41587-023-02024-y

Song H, Guo X, Sun L *et al.* The hard clam genome reveals massive expansion and diversification of inhibitors of apoptosis in Bivalvia. *BMC Biol* 2021;**19**:15. https://doi.org/10.1186/s12915-020-00943-9

Sudmant PH, Rausch T, Gardner EJ *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81. https://doi.org/10.1038/nature15394

Tirapé A, Bacque C, Brizard R *et al.* Expression of immune-related genes in the oyster during ontogenesis. *Dev Comp Immunol* 2007;**31**:859–73. https://doi.org/10.1016/j.dci.2007.01.005

Troost K. Causes and effects of a highly successful marine invasion: case-study of the introduced Pacific oyster in continental NW European estuaries. *J Sea Res* 2010;**64**:145–65. https://doi.org/10.1016/j.seares.2010.02.004

Varshney RK, Bohra A, Yu JM *et al.* Feature review: designing future crops: genomics-assisted breeding comes of age. *Trends Plant Sci* 2021;**26**:631–49. https://doi.org/10.1016/j.tplants.2021.03.010

Veenstra TD. Omics in systems biology: current progress and future outlook. *Proteomics* 2021;**21**:e2000235. https://doi.org/10.1002/pmic.202000235

Wang JF, Qi HG, Li L *et al.* Discovery and validation of genic single nucleotide polymorphisms in the Pacific oyster. *Mol Ecol Resour* 2015;**15**:123–35. https://doi.org/10.1111/1755-0998.12278

Wang JP, Li L, Zhang GF. A high-density SNP genetic linkage map and QTL analysis of growth-related traits in a hybrid family of Oysters (*Crassostrea gigas* × *Crassostrea angulata*) using genotyping-by-sequencing. *G3* 2016;**6**:1417–26. https://doi.org/10.1534/g3.116.026971

Wang Y, Li X-Y, Xu W-J *et al.* Comparative genome anatomy reveals evolutionary insights into a unique amphitriploid fish. *Nat Ecol Evol* 2022;**6**:1354–66. https://doi.org/10.1038/s41559-022-01813-z

Yu YM, Ouyang YD, Yao W. shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics* 2018;**34**:1229–31. https://doi.org/10.1093/bioinformatics/btx763

Zhao XL, Yu H, Kong LF *et al.* Transcriptomic responses to salinity stress in the Pacific oyster *Crassostrea gigas*. *PLoS One* 2012;**7**:e46244. https://doi.org/10.1371/journal.pone.0046244

Zhou Y, Minio A, Massonnet M *et al.* The population genetics of structural variants in grapevine domestication. *Nat Plants* 2019;**5**:965–79. https://doi.org/10.1038/s41477-019-0507-8

Handling editor: John Horne