Haplotype-resolved genomes provide insights into the origin and function of genome diversity in bivalves

Shikai Liu[1,2,#,*], Chenyu Shi[1,#], Chenguang Chen[1], Ying Tan[1], Yuan Tian[1], Daniel J Macqueen[3], Qi Li[1,2,*]

[1] *Key Laboratory of Mariculture (Ocean University of China), Ministry of Education, and College of Fisheries, Ocean University of China, Qingdao 266003, China*

[2] *Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao Marine Science and Technology Center, Qingdao, Shandong 266237, China*

[3] *The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, United Kingdom*

[#] These authors contributed equally to this work.

* Correspondence author at: Key Laboratory of Mariculture (Ocean University of China), Ministry of Education, and College of Fisheries, Ocean University of China

E-mail address: liushk@ouc.edu.cn (S. Liu), qili66@ouc.edu.cn (Q. Li).

Telephone: +86-0532-82032595

Abstract

Most bivalve genomes exhibit extensive heterozygosity and diversity, yet the origin and function of these genomic features remain unclear. As an ancient bivalve group, oysters demonstrate high ecological adaptability with diverse genomes, which serve as a good model for studies in genome diversity and evolution. Here, we report the significant contraction but highly divergent genomic landscape of *Crassostrea* species and highlight the association of transposable elements (TEs) activity with this genomic feature. By constructing a haplotype-resolved genome of *C. sikamea*, we identified the widespread presence of high divergence sequences (HDS) between the haplotype genome. Combined with population resequencing data, we underscore the role of genome divergence driven by TEs in shaping and maintaining oyster genomic diversity. By comparing haplotype genomes across *C. sikamea*, *Pinctada fucata*, *Arcuatula senhousia*, and *Mimachlamys varia*, we find that while haplotype divergence is common, its mechanisms of occurrence and maintenance differ significantly among bivalve species. Furthermore, our results show that the widespread presence of HDS not only contributes to substantial genetic variation but also influences the regulation of gene expression in oysters. The lack of conservation in allele-specific expression among individuals in oysters suggests high plasticity in haplotype polymorphism, allowing significant variation in gene regulation to supporting high phenotype plasticity and environment adaption. Overall, these findings offer novel insights into the connection between the unique genomic features and their role in adaptive evolution.

Introduction

Genome diversity is the foundation of species evolution and shapes phenotypic diversity [1,2]. However, the underlying mechanisms driving genome diversity and its contribution to adaptive evolution remain insufficiently explored, particularly in species that maintain extensive genomic variation[3]. Presence-absence variation (PAV) is an extreme form of structural variants, in which certain genomic regions or genes are present in some individuals but absent in others. The extensive PAV differences observed in bivalves underscore the significant genome diversity of these species[4–6]. Within an individual genome, PAV can appear as high divergence sequences (HDS) between homologous chromosomes, depending on the combination of haplotypes inherited from both parents. In the context of Mendelian inheritance, HDS may become dispersed or intermixed at the population level due to recombination and subsequent independent assortment of genetic material during meiosis. Therefore, characterization of PAVs may effectively ascertain their adaptive function at the population level, while studying haplotype genome divergence can provide insights into the origin of genome diversity and its impact on genome components and regulation of gene expression.

The oyster genome exhibits extensive genetic variation, characterized by widespread diversity at both the intra- and inter-individual levels, including single nucleotide polymorphisms, copy number variations, and insertions/deletions[7,8]. High reproductive capacity and high genetic load may be key factors in the formation of highly heterozygous genomes in oysters[9]. Considering the origin of genome diversity, the relative contribution of mutation accumulation and gene flow to oyster genomes have yet to be clearly defined. Clarifying these mechanisms is essential for understanding how genetic diversity is maintained and shaped. Given the potential impact of genetic diversity on genotypic polymorphism and the resulting differences in gene

expression among individuals[10,11], the study of expression regulation in oyster is expected to provide novel insights into the functional role of genetic diversity.

Here, by assessing the genome features among species and individuals, we revealed the evolutionary characteristics of the genomes in *Crassostrea* species. We further uncovered the origin of HDS and their role in shaping population characteristics by constructing a haplotype-resolved genome of *C. sikamea*. By assembling and comparing haplotype genomes across four representative bivalve species, we demonstrate that haplotype divergence is common among genomes, but its composition varies among species. We further analyzed the impact of HDS on gene evolution and expression regulation in oysters. With a large-scale multi-tissue RNAseq datasets, we characterized the extensive differences in allele-specific expression among individuals in oysters, demonstrating a potential correlation between these differences and haplotype polymorphism. Our findings provide valuable insights into the genomic evolution and gene expression regulation in oysters as well as other bivalves.

RESULTS

Genome reduction and divergence in *Crassostrea* species

Understanding the changes in genome can provide insights into the evolutionary history of species[12]. Given the extensive genome diversity among bivalve species or even in individuals, differences in genome size are expected. We first conducted a comprehensive comparison of genome sizes across various orders of Bivalvia using information from literature and the GoAT database[13] (Fig. 1a). The results revealed a significant reduction in genome among species in the order Ostreida (Fig. 1b), particularly within the genus *Crassostrea*, compared to other bivalves. Typically, the genome sizes of Bivalvia species are around 1 Gb, whereas the genome size of the species in

genus *Crassostrea* is approximately 600 Mb (Table S1). Interestingly, despite this significant reduction in genome size, annotation of protein-coding genes of species in the oyster lineage (Fig. S1) revealed that the number of genes within genome remains similar across bivalve species, with comparable lengths of coding sequences (Fig. 1c). The high variation in gene expansion and contraction among *Crassostrea* species can be observed, which may reflect high levels of genome diversity among these species (Fig. 1c).

To further explore genetic variation within species, we performed pairwise genome comparisons among individuals from various bivalve species (Extended Data Fig. 1; Table S2). Despite potential inaccuracies due to the genetic background or genome assembly quality of the selected individuals, our results clearly show extensive genomic differences between individuals across these species (Fig. 1d). High proportions of un-aligned regions between individual genomes were observed, identified as un-aligned regions (NOTALs), highly diverged regions (HDRs), and Indels, which are collectively referred to as high divergence sequences (HDS) hereafter. The remaining sequences of the genome are divided into syntenic region alignments (SYNAL) and structural rearrangement alignments (SRAL) based on whether structural rearrangement occurs in the alignment. Notably, compared to the European flat oyster (*Ostrea edulis*), individuals within several species of the genus *Crassostrea* exhibit significant genome differences with a similar proportion of HDS sequences, suggesting that these extensive structural differences may be conserved within the genus. Further, we examined the coverage of HDS and transposable elements (TEs) within each 500 kb window across the whole genome. The distribution of HDS showed a strong positive correlation with TEs in *Crassostrea* species (Fig. 1e), indicating that TEs could be involved in the formation and maintenance of genome diversity in *Crassostrea* species.

Genome divergence shaped by TEs enhances genetic diversity in oysters

To reliably explore the unique genomic characteristics of *Crassostrea* species, we constructed the haplotype genomes of *C. sikamea*. We assembled a 1.08 Gb haplotype-resolved genome assembly using 29.53 Gb (~57X) of CCS reads and 50.49 Gb of Hi-C data. The assembly was organized into 20 pseudo-chromosomes, anchoring 1.04 Gb (96.3%) of the genome (Table 1; Table S3). Two haplotype genomes of 526.67 Mb and 516.10 Mb for haplotypes A and B (Hereafter HapA and HapB, respectively) were separated, both closely matching the estimated genome size by k-mer analysis (~ 517.20 Mb, Fig. S2a). The Hi-C contact matrix showed minimal interaction between chromosomes, indicating no obvious phasing errors in our genome assembly (Fig. S2b). The estimated Quality Value (QV) of the final assembly is 45.6 with k-mer completeness of 98.77% (Table S4), and each haplotype assembly also demonstrates high BUSCO completeness compared to other species (Fig. S2c). A total of 30,052 and 29,609 protein-coding genes were annotated from HapA and HapB, with 97.8% and 97.1% BUSCO completeness, respectively. Based on a comparison of collinear gene blocks, haplotype-resolved chromosomes showed a high level of co-linearity (Extended Data Fig. 2a).

The haplotype-resolved genome assembly allowed precise assessment of sequence divergence at the chromosomal level. Defining HapA as the reference, a total of 1,964 syntenic regions were detected which covered 373.60 Mb (70.94%) of the HapA genome and 384.05 Mb (74.41) of the HapB genome (Extended Data Fig. 2b). Additionally, a total of 3,620 structural variations (SVs) were identified, including 115 inversions, 2,617 translocations, and 888 duplications, covering 28.73 Mb of the HapA genome and 30.56 Mb of the HapB genome (Table S5). Notably, high proportions of high divergence sequence (HDS) were identified, accounting for 69.6% of the whole genome sequence. This finding aligns well with the abovementioned genomic variation among individual genomes.

To further investigate the potential origin of HDS, we annotated TEs of 217.45 Mb in HapA and 204.33 Mb in HapB, covering 41.29% and 39.59% of each haplotype genome, respectively (Table S6). The proportions of TEs between two haplotypes are generally consistent across each chromosome, but what is intriguing is, that the length of sequences contained in each TE family varies between the two haplotypes, especially the absolute length difference between the terminal inverted repeats (TIR) families (Fig. 2a; Table S6). In total, 179.99 MB and 171.6 MB of TEs were inserted into the HDS of the HapA and HapB genomes, representing a significant portion of the whole-genome TEs (Fig. 2b). Notably, 77.5% of intact TEs inserted entirely or partially within the HDS, suggesting that TE activation may be an important factor driving the formation of HDS (Fig. 2c).

To further investigate the conservation of HDS within the population, we analyzed the genetic characteristics of a wild population with 25 *C. sikamea* individuals (Table S7). Using the whole-genome resequencing, we observed that the alignment rate in HDS was significantly lower compared to other genomic regions (Fig. 2d). Moreover, nucleotide diversity of genes located in HDS regions was significantly higher than that in other genomic regions (Fig. 2e). Although the average length of haplotype blocks did not differ significantly across genomic regions (Fig. 2f), haplotype polymorphism was significantly higher within HDS regions (Fig. 2g, $p<0.01$). To illuminate the impact of introgression on HDS and genetic diversity, we included publicly available whole-genome resequencing data for 70 individuals from three closely related species, including 28 *C. ariakensis*, 21 *C. angulata*, and 21 *C. gigas* (Extended Data Fig. 3; Supplementary Note 1). The ABBA-BABA analysis indicated introgression between *C. gigas* and *C. sikamea* (Extended Data Fig. 4). However, the degree of introgression was minimal (Table S8), and there was no correlation between the potential introgression regions and HDS (Fig. 2h). Therefore, we propose that the genomic differentiation in *Crassostrea* species

primarily originates from TE activation, which significantly affects the genetic diversity of the population.

Comparative analysis of haplotype-resolved genomes across bivalve species

To assess whether the haplotype divergence observed in oysters reflects a broader evolutionary event in bivalves, we further constructed haplotype-resolved genomes for a mussel (*Arcuatula senhousia*) and a scallop (*Mimachlamys varia*) using the same assembly pipeline. These assemblies, along with the previously published haplotype-resolved genome of the pearl oyster (*Pinctada fucata*), were of high quality for comparative analysis (Table S9). Significant haplotype divergence was identified across all three species (Extended Data Fig. 5; Table S10). Notably, the length-density distribution of high divergence sequences (HDS) in oysters shows a concentration at shorter lengths, distinct from the distributions in other species (Fig. 3a). This suggests that the expansion of HDS in oysters may be constrained, potentially contributing to the maintenance of their smaller genome size. Additionally, a strong positive correlation was observed between haplotype divergence scores and transposable elements (TEs) in *C. sikamea* (R = 0.76), whereas this relationship was less pronounced in the haplotype assemblies of *P. fucata*, *A. senhousia*, and *M. varia* (Fig. 3b). The relationship between TEs and haplotype divergence is less consistent in bivalve species, which indicates that the mechanisms during genome evolution to shape and maintain genomic diversity may differ among species. Further quantification of the proportion of TEs within HDS across these four species revealed a clear correlation between the expansion of HDS and TE content in several representative species (Fig. 3c). In *C. sikamea*, *P. fucata*, and *A. senhousia*, TE content increases with HDS length, with an inflection point where the rate of TE accumulation becomes low. In *C. sikamea*, this inflection point occurs earlier, and the initial TE content is lower, suggesting a stronger inhibition mechanism for TEs in oysters. In

contrast, the TE content in *M. varia* does not change with increasing HDS length suggests that other factors beyond TE activity may play the role in shaping haplotype divergence in different bivalve species.

The divergence of haplotype genomes is expected to reveal even more pronounced differences across species. This is evidenced in the lower proportion of homologous genes within HDS compared to the whole genome in all four species, especially for homologous genes shared across all species (Fig. 3d). Although the proportion of homologous genes within HDS is low, the conserved homologous genes identified within HDS across *C. sikamea*, *P. fucata*, *A. senhousia*, and *M. varia* are significantly enriched in common functional pathways, such as DNA replication and phototransduction (Fig. 3e). This suggests that, while HDS may undergo rapid evolution, they also have essential biological functions, which potentially contribute to species divergence in response to environmental pressures.

Haplotype divergence drives gene evolution and allelic variation in gene expression

To understand the function of high divergence sequences (HDS), we further annotated genes and determined their expression profiles within HDS in *C. sikamea*. Based on the similarity and alignment coverage of protein sequences (identity>90%; coverage>80%), 4,326 and 3,879 redundant genes were identified in the HapA and HapB genomes, respectively. We then quantified the distribution of these redundant genes across different regions of both haplotype genomes. Notably, 3,076 and 2,669 redundant genes in the HapA and HapB genomes, respectively, were located within HDS and showed significant enrichment in these regions (Fig. 4a; p.adj < 0.01). Additionally, based on the gene family clustering from OrthoFinder, we separately counted the number of lineage-specific genes as well as their distribution across various regions of the genome (Fig. 4b). Haplotype-specific genes represent the genetic differences

between individuals and are indicative of rapidly evolving presence/absence variation (PAV) genes. We detected 895 and 929 haplotype-specific genes in the hapA and hapB genome, respectively. Of these, 76.4% and 71.8% were significantly enriched within HDS (p.adj < 0.01). The haplotype-specific genes in HDS regions between haplotypes showed similar enrichment of KEGG terms, such as DNA replication proteins, Mitochondrial biogenesis, Pattern recognition receptors, and NOD-like receptor signaling pathway (Table S11). Analysis of RNA-seq data from the labial palps, gill, and digestive gland of the genome-sequenced individual revealed that genes located within the region of structural rearrangement alignments (SRAL) and HDS exhibited lower median expression levels (Fig. 4c). Previous studies have indicated that low expression and non-essentiality are prominent features of evolutionarily young genes[15]. Gene gaining and duplication occur more frequently in HDS regions, which may promote the evolution of new genes.

The haplotype-resolved genome assembly enabled us to define alleles across the genome and explore the impact of haplotype divergence on allelic variation. Based on synteny and annotation of genes, 19,148 pairs (64.19% of all annotated genes) of those were considered reliable allelic genes (Supplementary Data 1; Extended Data Fig 6a, b). The majority of these identified alleles exhibit a high degree of coding sequence similarity, with an average similarity of 97.67%, and an average Ka/Ks ratio of 0.23 (Extended Data Fig 6c, d). Interestingly, the allele displays distinctive distribution differences across the genome (Extended Data Fig 7). In the HDS regions, the proportion of identified alleles is lower than the whole genome (Fig. 4d) and exhibits significantly high sequence divergence compared to other regions (Fig. 4e). With the identified alleles between the haplotype genome, the RNA-seq data from the genome-sequenced individual were used for allelic expression analysis. Direct quantification of allelic expression using the diploid genome provides an effective strategy to control for the effect of mapping bias on allelic

expression analysis (Fig. S3). Allele-specific expression (ASE) is potentially influenced by naturally selected variations between alleles. As shown in Fig S4, the alleles, modified by nonsynonymous mutations or structural variations, exhibit varying degrees of ASE among tissues. Considering the variation in allele expression patterns across different tissues (Fig. 4f) and the higher proportion of ASE observed in HDS (Fig. 4g), it's speculated that HDS may affect allele expression, thereby influencing biological phenotypes.

Highly plastic allele-specific expression is associated with the genome diversity

To understand the relationship between the genomic feature of high diversity and adaptive evolution in oysters in perspective of gene expression regulation, we developed a strategy to differentiate alleles with SNPs using RNAseq data generated from multiple tissues of individual oysters (see Methods; Extended Data Fig 8). Our results demonstrate the reliability of SNP detection from RNA-seq data (Supplementary Note 2; Table S12), which allows for reliable allele identification for individual genotyping. We then further investigated allele expression patterns across different individuals and tissues to explore their relationship with genome diversity. The allele-specific expression genes (ASEGs) were identified in each tissue sample (Table S13). Among the different tissues from the same individual, highly prevalent tissue-specific ASEGs are common, suggesting the widespread presence of tissue-specific *cis*-regulation. We observed similar expression trends for these tissue-specific ASE genes across individuals, which usually exhibited pronounced tissue-preferred expression, especially in the digestive gland and mantle (Extended Data Fig 9).

To obtain further insights into the potential functions of ASEGs, we compared four ASE gene sets detected from same tissue of different individuals (Extended Data Fig 10, Table S13,14). Across each dataset, a substantial amount of ASEGs were identified from specific individuals, while only a limited

number of ASEGs were identified from all tested individuals (Fig. 5a). Further clustering of the ASEGs revealed that the similarity of ASEGs identified from different tissues of same individual is significantly higher than the similarity of ASEGs identified same tissues of the different individual (Fig. 5b). This suggests that the presence of ASEGs may be influenced more by genetic background related to haplotype divergence than by selective pressures associated with tissues during evolution in the oysters.

To unravel the potential evolutionary scenario of the occurrence of ASEGs, we investigated the sequence differences across genes exhibiting diverse patterns of allelic expression. The heterozygosity of ASE genes is significantly higher than that of genes without allele-specific expression (non-ASEGs) in exon regions. However, the heterozygosity in promoter regions does not appear to have a direct correlation with the occurrence of allelic expression (Fig. 5c). To verify the connection between the extensive differences in ASEGs among individuals and haplotype polymorphism, we further explore the relationship between genetic divergence among individuals and the occurrence of ASE. The ASEGs exhibit higher nucleotide polymorphism in the exon and promoter regions compared to non-ASEGs and the consistent-ASEGs show significantly higher nucleotide polymorphism (Fig. 5d). Higher polymorphism may reflect the result of neutral evolution or balancing selection, however, due to the small effect size, it is difficult to distinguish whether this selection is strong or widespread. Notably, we observe lower evolutionary constraints in ASEGs compared to non-ASEGs across species (Fig. 5e), suggesting that the occurrence of ASEGs could be mainly derived from regulatory diversity, rather than being constrained by selective pressures.

DISCUSSION

Understanding the origin and function of genome diversity is essential to deciphering the intricate relationship between genetic innovation and adaptive

evolution. In this study, we propose an evolutionary strategy in bivalves, where species enhance their adaptability by maintaining extensive genomic divergence rather than by expanding their genetic reservoir. A prime example can be found in *Crassostrea* oyster species, which maintain a relatively contracted yet diverse genome but exhibit extreme environmental adaptation. Through assembly of a haplotype-resolved genome of *C. sikamea* and analysis of haplotype divergence, we show that haplotype genome divergence shaped by TEs enhances genetic diversity in oysters. The similar evolutionary mechanisms are present in selfing species, such as *C. elegans*[20] and *A. thaliana*[21], which implies that this may be a common strategy for adaptive evolution across various species.

Comparative analysis of haplotype-resolved genomes across multiple bivalve species revealed distinct mechanisms driving and maintaining haplotype divergence among species, with notable differences in the contribution of TEs activity to the formation of genome diversity. For bivalves, TE activity is not the sole contributing factor to genomic diversity. Other mechanisms, such as recombination and duplication, may play crucial roles in maintaining diversity and function. This underscores the need for broader analyses across more species. Notably, we identified highly conserved homologous genes located in HDS across these bivalve species, suggesting conservation in the origin or function of genome diversity across the species. Given the strong association between genomic diversity and TEs as observed in *Crassostrea* species, the oysters may serve as a unique model for future investigation into the broader mechanisms of genomic variation, providing valuable insights into the impact of transposable elements on genomic structure.

The potential function of haplotype divergence can be reflected in its impact on gene birth-and-death and expression regulation[22,23]. Our results demonstrated that haplotype divergence enhances gene birth in the oyster genome and plays a significant role in regulation of allele expression. However,

it is important to note that the novelty in genetic composition and expression regulation does not necessarily indicate functional necessity. In the context of genetic diversity, the different combination of divergent haplotypes can play a crucial role in shaping the complex genomic characteristics of individuals[20]. Due to the lack of genomic data from multiple individuals, we are unable to further assess the fine details of these differences. However, by utilizing gene expression datasets from various tissues, we evaluated the impact of genomic polymorphisms on expression regulation. The high plasticity of allele-specific expression observed among tissues and individuals, combined with the low evolutionary constraints in these genes, suggests that ASEGs have non-conserved functions. Given that oysters are typically profligate broadcast spawners and inhabit quite variable environments, we speculate that the formation of these stochastic ASEGs might be one of its adaptive evolutionary strategies toward shaping the high phenotypic plasticity among individuals.

Overall, this study highlights the significance of genome divergence in promoting adaptation and evolutionary success, offering valuable insights into the evolutionary strategies employed by bivalves. These findings pave the way for future research to explore the broader implications of genomic variation and its roles in gene expression regulation.

Method

Sample collection and genome sequencing

The Kumamoto oysters (*Crassostrea sikamea*) used for genome sequencing in this study were collected from a wild population in Ningde (Fujian, China). Detailed sampling information is provided in metadata (Table S1). For population genomics analysis, a total of 25 oysters were used for whole-genome resequencing. The adductor muscle from each sample was flash-frozen in liquid nitrogen. The DNA was extracted using the phenol-chloroform method and dissolved in double-distilled water. The concentration and quality

of the DNA were assessed using a Qubit fluorescence quantitative analyzer and 1% agarose gel electrophoresis. The sequencing libraries were constructed using the NEBNext Ultra DNA Library Prep Kit and sequenced on the Illumina Hiseq Xten platform (Illumina) for 150 bp pair-end reads.

For whole genome sequencing, high molecular weight DNA was extracted from the adductor muscle of an individual for long-read sequencing using single-molecule real-time sequencing technology, and for long-range high-throughput chromatin capture (Hi-C) sequencing. The long-read sequencing library was prepared by using the SMRTbell prep kit 3.0 and sequenced on the PacBio Sequel II System. The Hi-C library was prepared using the Dovetail™ Hi-C Library Preparation Kit and sequenced on the Illumina Hiseq Xten platform.

For gene annotation, transcriptome sequencing was conducted using various tissues from the same individual used for whole genome sequencing. Tissues including the gill, digestive gland, and labial palps were separately processed for RNA-seq. The RNA-seq libraries were constructed using the NEBNext® Ultra™ RNA Library Prep Kit and sequenced on the NovaSeq 6000 platform for 150 bp paired-end reads. Additionally, total RNA was extracted from various tissues of multiple individuals, including the mantle, gill, heart, digestive gland, labial palps, and adductor muscle. Equal amount of RNA from these tissue samples were pooled and subjected to full-length transcriptome sequencing. The Iso-Seq library was constructed using the SMRTbell prep kit 3.0 and sequenced on a PacBio Sequel II platform.

For gene expression analysis, five tissues, including adductor muscles, mantle, digestive gland, gill and labial palps, were collected from three wild individuals of *C. sikamea*, respectively. The procedures for tissue collection, storage, RNA extraction, and sequencing were as described above. Additionally, RNA-seq data sequenced from *C. sikamea* at different growth stages within the same population were utilized. Samples were collected at three different time points in March (n=4), June (n=3), and October (n=5). Three

tissues, including adductor muscle, digestive gland and labial palps, were collected from each of the individuals, respectively. RNA-seq was performed as described above.

Haplotype genome assembly and comparative analysis

Before assembly, Jellyfish was employed to calculate k-mer frequencies (k=19). Then, genome size, repetitiveness, and rate of heterozygosity were predicted with genomeScope2.0[24]. We assembled the *C. sikamea* genome using Hifiasm incorporating Illumina short-read sequences and PacBio single-molecule real-time long-read sequences as well as sequences from high-throughput chromatin conformation capture (Hi-C) technologies. The Juicer (v1.6)[25] combined with 3D-DNA (v180419)[26] was used for scaffolding. The same assembly procedure was applied to assembly of the haplotype genome of *Mimachlamys varia*[27] and *Arcuatula senhousia* by using the publicly available data from the Darwin Tree of Life Project (DToL, https: //www.darwintreeoflife.org/).

To identify structural variations in the genomes, minimap2[28] was used to perform whole-genome alignment of two haplotype genomes using haplotype A as the reference genome. The SyRI[29] was used to identify syntenic regions and structural rearrangements based on the direction, position, and data volume relationships between the sets of alignments. Based on the positional relationship with the annotation blocks (syntenic or structurally rearranged, such as inversions, translocations, and duplications), the un-aligned regions were classified as Indels, highly divergent regions (HDRs), and un-aligned regions (NOTAL). Here, we collectively term Indels, HDRs and NOTAL as high divergence sequences (HDS), representing the specific sequences or structural variations between two haplotype genomes.

Genome annotation of genes and repetitive elements

Genome annotation was performed using a comprehensive strategy combing the methods of both evidence-based and *ab initio* gene predication. *De novo* identification and classification of repeats was performed using RepeatModeler[30]. RepeatMasker[31] was used with the *de novo* repeat library to soft-mask the genome. BRAKER3[32] was used to annotate genes on the soft-masked genome. The RNA-seq and Iso-seq data was supplied to the BRAKER pipeline along with OrthoDB v10[33]. Functional annotation of protein-coding genes was carried out using the EggNOG (v 5.0)[34] and InterProScan (v 5.52–86.0)[35]. To achieve precise and comprehensive annotation of the species-specific repeat libraries, we further used a combination of structural and homology-based methods to annotate transposable elements with EDTA[36]. Based on the same analysis procedure, the genomes of several other representative species (i.e, *P. fucata*, *A. senhousia*, and *M. varia*) were also annotated, followed by comparative analysis.

Population genomics analyses

Sequence alignments and variant calling

The whole-genome resequencing data of *C. sikamea* population (n=25) was employed for variant calling. The Fastp[37], with default parameters, was utilized to eliminate residual adapter sequences and low-quality reads from the raw data. The qualified reads were aligned to the HapA genome of *C. sikamea* using BWA[38]. The alignment results were sorted and converted to BAM format using Samtools[39]. Duplicate marking and variant detection were performed using GATK4 following the GATK Best Practices[40]. Finally, Vcftools[41] software was employed to filter out SNPs with multiple alleles (>2) and SNPs with a missing rate exceeding 20% and minor allele frequencies were less than 5%. PIXY[42] is used to calculate nucleotide diversity for each region. The LDBlockShow[43] were used to identify the haplotype block.

Haplotype diversity analysis

The analysis of haplotype diversity started with the use of WhatsHap[44] to generate phase sets for each individual. Subsequently, the obtained population haplotype phase sets were imputed using Beagle[45], with imputation applied only to sites with a missing rate exceeding 20%. After imputation, the haplotype block diversity across different regions was calculated and normalized by the length of each region.

Population structure analysis

The qualified SNPs were pruned using the Plink[46] software with the following parameters: '-indep-pairwise 50 5 0.2', aiming to eliminate the influence of linkage disequilibrium (LD). Based on the pruned dataset, phylogenetic tree was constructed using neighbor-joining (NJ) method implemented. Structure analysis was performed in Admixture[47]. *K* values were set from $K = 2$ to $K = 9$. The minimum CV error value appeared at $K = 4$. PCA analysis was performed using Plink. For LD decay analysis, the dataset of qualified SNPs with LD pruning was utilized, and LD decay was analyzed by the PopLDdecay software[48]. Fixation statistics (Fst) and nucleotide diversity (π) were calculated with the python script (https://github.com/simonhmartin/genomics_genera). To assess the extent of shared DNA segments among individuals across the four species, identity-by-descent (IBD) block analysis was conducted using Beagle[45] with the following parameters: window = 100,000; overlap = 10,000; ibdtrim = 100; ibdlod = 5.

Analyses of introgression

We conducted the *D*-statistic (also known as the ABBA-BABA test) to investigate introgression among three oyster species. Following the detection of introgression among individuals at the genome level, we used a modified *f*-statistic (*fd*) to estimate the proportion of introgression sites at the population level. Using the *C. ariakensis* oyster as an outgroup, we expect equal counts of the two site patterns (ABBA and BABA) when incomplete lineage sorting results

in inconsistent patterns at the sites. If the inconsistency is due to introgression, one of the site patterns is expected to be more prevalent than the other. Leveraging the allele frequencies detected by GATK for each SNP, we used the R script calculate_abba_baba.r (https://github.com/palc/tutorials-1) to estimate the *D*-statistic, *fd* and *P*-value.

Demographic history

We implemented pairwise sequentially Markovian coalescent (PSMC)[49] to estimate the dynamics of effective population size (*Ne*) over the past several million years for four oyster species. Three *C. sikamea*, three *C. gigas*, two *C. angulata*, and two *C. ariakensis* were included in the analysis. The PSMC parameters were set as follows: "-N25 -t15 -r5 -p "4+25*2+4+6"". The generation time (*g*) was set to 1, and the mutation rate (μ) was set to 0.2×10^{-8} for all four species.

Analyses of gene family, redundant genes, and expression profiling

The cluster analysis of gene family was performed with annotated protein datasets of 21 representative molluscan species. To identify orthogroups and orthologs from the datasets, the OrthoFinder (v2.5.4)[50] pipeline was used by invoking DIAMOND and OrthoMCL to call orthogroups based on sequence identity. The detailed procedure and subsequent phylogenetic analyses were described in Supplementary Note 3. The orthogroups that are present only within a specific lineage were identified based on OrthoFinder outputs. These groups contain genes from the target species exclusively, without including genes from other compared species. The conserved one-to-one orthologous genes among different oyster family species were identified by filtering based on the best BLAST hits and gene synteny results. To identify redundant genes, we used CD-HIT[51] with a sequence identity threshold of 90% and a minimum alignment coverage of 80%. For gene expression analysis, RNA-seq datasets

were mapped to the reference genome using STAR[52], and expression levels were quantified with StringTie[53].

Identification and comparison of alleles between haplotype genomes

From haplotype-resolved genome assembly, the allele was defined using a synteny-based strategy with comparison of sequence similarity. Following the similar allele identification protocols as previously reported[54,55], alleles were identified by assessing gene synteny, sequence similarity, and functional annotation results. Specifically, comparison of gene synteny between the *C. sikamea* haplotype-resolved genome assembly was performed with WGDI[56]. Then, the paired genes within each synteny block with the best blast score were considered as putative alleles. We further filter the putative alleles based on the results of gene functional annotation and protein sequence length. We used the EMBOSS needle program[57] for pairwise comparison of allelic genes with default parameters. The similarity score was calculated as the number of unsubstituted bases divided by the length of the alignment block. The Ka and Ks values for the protein-coding sequences of alleles were calculated by WGDI.

Allele-specific expression (ASE) analysis

SNPs between haplotype genomes were detected from RNA-seq data to differentiate expressed alleles between haplotype genes using the GATK pipeline. Briefly, RNA-seq from various tissues of the same individual were combined to gain power in SNP detection, which were subject to stringent trimming with fastp program. STAR was used with default parameters for the read mapping on the Haplotype A genome of *C. sikamea*, after the multi-sample 2-pass mapping procedure. Uniquely mapped reads were then post-processed following the GATK best practices for RNA-seq data to detect variants, which were further filtered using “VariantFiltration” with two of the three suggested filters, “QD < 2” and “FS > 30”. Finally, we selected the SNPs with genotypes

associated with each individual that (10.reads.DP) Genotype Call Rate (CR) ≥ 20% and CR of all samples ≥ 50%[58].

The SNPs obtained in the previous step were used to generate the pseudo-genome per individual by the "FastaAlternateReferenceMaker" tools from GATK. Reads from each tissue sample were aligned to the individual genome and then used to detect allele expression among each sample by phASER[59]. To assess ASE in each sample, we screened for read number imbalance between the alleles using a binomial test (binom.test R function). P-values were corrected using the Benjamini-Hochberg method with a false discovery rate of 0.05. The alleles with padj > 0.1 are considered to have no occurrence of ASE.

We also used the DNA-seq data from each individual for variant detection and comparative analysis. The pipeline of variant calling was same as described above. The SNP variant information from different regions in the whole genome was extracted using bcftools[60]. HiBLUP[61] was used to calculate heterozygosity across each region. PIXY[42] was used to calculate nucleotide diversity for each region.

Detection of evolutionary constraint

The Cactus[62] was used for inter-species alignments across nine species in ostreidae lineage, including *C. sikamea*, *C. angulata*, *C. gigas*, *C. nippona*, *C. hongkongensis*, *C. ariakensis*, *C. virginica*, *O. denselamellosa*, and *O. edulis*. Using the annotated coding genes of *C. sikamea*, we first extracted 4-fold degenerate sites using hal4dExtract and used phyloFit[63] to estimate the initial neutral model. Once the neutral model was calculated, the whole alignment from Cactus was used to estimate rho (the expected substitution rate of conserved elements relative to neutrality) and calculate PhastCons score for each site across the genome by using phastCons[64].

Data availability

Sequencing data are available from NCBI SRA under BioProject accessions PRJNA1090262, PRJNA1090167, PRJNA1089862; The genome assembly and annotation data of the accessions is available at Figshare 10.6084/m9.figshare.26953708.

Code availability

Custom code used for the analysis is available at https://github.com/Scy-bio/paper-anlysis-pipeline.

References

1. Stange, M., Barrett, R. D. H. & Hendry, A. P. The importance of genomic variation for biodiversity, ecosystems and people. *Nat Rev Genet* **22**, 89–105 (2021).

2. Nevo, E. Evolution of genome–phenome diversity under environmental stress. *Proceedings of the National Academy of Sciences* **98**, 6233–6240 (2001).

3. Calcino, A. D., Kenny, N. J. & Gerdol, M. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. *Philos Trans R Soc Lond B Biol Sci* **376**, 20200153 (2021).

4. Farhat, S. *et al.* Comparative analysis of the Mercenaria mercenaria genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks. *BMC Genomics* **23**, 192 (2022).

5. Takeuchi, T. *et al.* A high-quality, haplotype-phased genome reconstruction reveals unexpected haplotype diversity in a pearl oyster. *DNA Research* **29**, dsac035 (2022).

6. Saco, A. *et al.* Gene presence/absence variation in *Mytilus galloprovincialis* and its implications in gene expression and adaptation. *iScience* **26**, 107827 (2023).

7. Jiao, Z., Tian, Y., Hu, B., Li, Q. & Liu, S. Genome Structural Variation Landscape and Its Selection Signatures in the Fast-growing Strains of the Pacific Oyster, *Crassostrea gigas*. *Mar Biotechnol* **23**, 736–748 (2021).

8. Qi, H., Li, L. & Zhang, G. Construction of a chromosome-level genome and variation map for the Pacific oyster *Crassostrea gigas*. *Molecular Ecology Resources* **21**, 1670–1685 (2021).

9. Yin, X. & Hedgecock, D. Overt and concealed genetic loads revealed by QTL mapping of genotype-dependent viability in the Pacific oyster *Crassostrea gigas*. *Genetics* **219**, iyab165 (2021).

10. Huang, W.-C., Bennett, K. & Gregg, C. Epigenetic and Cellular Diversity in the Brain through Allele-Specific Effects. *Trends in Neurosciences* **41**, 925–937 (2018).

11. Pham, G. M. *et al.* Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *The Plant Journal* **92**, 624–637 (2017).

12. Blommaert, J. Genome size evolution: towards new model systems for old questions. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20201441 (2020).

13. Challis, R., Kumar, S., Sotero-Caio, C., Brown, M. & Blaxter, M. Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Res* **8**, 24 (2023).

14. Martelossi, J. *et al.* Multiple and diversified transposon lineages contribute to early and recent bivalve genome evolution. *BMC Biology* **21**, 145 (2023).

15. Doughty, T. W. *et al.* Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat Commun* **11**, 2144 (2020).

16. Meyer, A. *et al.* Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **590**, 284–289 (2021).

17. Nowoshilow, S. *et al.* The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).

18. Liu, C. *et al.* Giant African snail genomes provide insights into molluscan whole-genome duplication and aquatic–terrestrial transition. *Molecular Ecology Resources* **21**, 478–494 (2021).

19. Shapiro, J. A. Engines of innovation: biological origins of genome evolution. *Biological Journal of the Linnean Society* **139**, 441–456 (2023).

20. Lee, D. *et al.* Balancing selection maintains hyper-divergent haplotypes in *Caenorhabditis elegans*. *Nat Ecol Evol* **5**, 794–807 (2021).

21. Baduel, P. *et al.* Genetic and environmental modulation of transposition shapes the evolutionary potential of *Arabidopsis thaliana*. *Genome Biol* **22**, 138 (2021).

22. Bao, Y. *et al.* Unraveling cis and trans regulatory evolution during cotton domestication. *Nat Commun* **10**, 5399 (2019).

23. Todesco, M. *et al.* Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).

24. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432 (2020).

25. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).

26. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

27. Fletcher, C. *et al.* The genome sequence of the variegated scallop, *Mimachlamys varia* (Linnaeus, 1758). *Wellcome Open Res* **8**, 307 (2023).

28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

29. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).

30. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).

31. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **25**, 4.10.1-4.10.14 (2009).

32. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv* 2023.06.10.544449 (2024) doi:10.1101/2023.06.10.544449.

33. Kriventseva, E. V. *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47**, D807–D811 (2019).

34. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).

35. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

36. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**, 275 (2019).

37. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

38. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).

39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

40. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* **43**, 11.10.1-11.10.33 (2013).

41. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

42. Korunes, K. L. & Samuk, K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources* **21**, 1359–1368 (2021).

43. Dong, S.-S. *et al.* LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Briefings in Bioinformatics* **22**, bbaa227 (2021).

44. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology* **22**, 498–509 (2015).

45. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am J Hum Genet* **81**, 1084–1097 (2007).

46. 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).

47. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).

48. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).

49. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).

51. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

52. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

53. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**, 1650–1667 (2016).

54. Zhang, X. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat Genet* **53**, 1250–1259 (2021).

55. Zhou, Q. *et al.* Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet* **52**, 1018–1023 (2020).

56. Sun, P. *et al.* WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Molecular Plant* **15**, 1841–1851 (2022).

57. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276–277 (2000).

58. Jehl, F. *et al.* RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species. *Front. Genet.* **12**, (2021).

59. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun* **7**, 12817 (2016).

60. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

61. Yin, L. *et al.* HIBLUP: an integration of statistical models on the BLUP framework for efficient genetic evaluation using big genomic data. *Nucleic Acids Research* **51**, 3501–3512 (2023).

62. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

63. Siepel, A. & Haussler, D. Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. *Molecular Biology and Evolution* **21**, 468–488 (2004).

64. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* **12**, 41–51 (2011).

65. Kumar, S. *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution* **39**, msac174 (2022).

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (32341060 and 42276112), the Key Research and Development Program of Shandong Province (2021ZLGX03), the National Key Research and Development Program of China (2022YFD2400300), and the earmarked fund for the Agriculture Research System of China (CARS-49). We acknowledge the support of the High-Performance Biological Supercomputing Center at the Ocean University of China.

Author contributions

QL and SL initiated this project and obtained funding for the studies, SL conceived the project, coordinated and supervised the entire project. YT1 (Ying Tan) contributed to the sample collection and genome and transcriptome sequencing. CS and YT2 (Yuan Tian) assembled the genome of *C. sikamea* CS, SL, DM, and CC analyzed the data, CS and SL wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary Information

Supplementary Figs. 1–10, Tables 1–12, Notes 1–3 and references.

Supplementary Data 1

List of identified alleles between the haplotype genomes with information of sequence comparison and allele expression.

Table 1. Statistics of genome assembly and annotation of *C. sikamea.*

Feature	*C. sikamea*	
Estimated genome size (bp)	517,203,934	
Heterozygosity (%)	3.31	
Duplication (%)	0.57	
De novo assembly		
Assembly size (Gb)	1.08 (1,076,379,459)	
Contig number	1,243	
Largest length of contigs	10,147,741	
Average length of contigs	865,952.90	
Contig N50 (Mb)	2,159,660	
BUSCO completeness of assembly (%)	98.9%	
Haplotype-resolved chromosomal-level assembly and annotation		
	Haplotype A	Haplotype B
Length of chromosomes (Mb)	526.67	516.10
BUSCO completeness of assembly (%)	98.2%	98.0%
BUSCO completeness of annotation (%)	97.8%	97.1%
Number of annotated genes	30,052	29,609
Total number of anchored genes	59,661	

Figure Legends

Fig. 1 Comparative analysis of the genome size, gene composition, and individual genomic variation across different bivalve genomes. a. Phylogenetic tree of bivalve species with available genomic size data. Branch lengths were derived from the divergence times provided by TimeTree5[65]. **b.** Comparison of species genome sizes across families of Pteriomorphia, including Mytilida, Ostreida, Petinida, Venerida, and Cardiida. P-values were calculated using a two-sided unpaired Wilcoxon test; **c,** Bivalve-wide analysis of gene gain and loss, indicating the number of genes gained (in red) and lost (in blue) for each species; Counting the number of genes and total length of CDS for each species. **d,** Proportions of syntenic alignment (SYNAL), structural rearrangement alignments (SRAL) and high divergence sequences (HDR, NOTAL and Indels) between individual genome in several bivalvia species. **e,** The distribution density comparison of high divergence sequences (HDS) with the TEs along the genome across different Bivalvia species. The distribution density of each genomic component was calculated based on the proportion of sequences per 500 kb window.

Fig. 2 Transposable elements and genetic diversity within high divergence sequences between haplotype genomes. a. The sequence length occupied by different classes of transposable elements (TEs) on each chromosome is shown on the left, with the white bars representing the chromosome length. The line graph on the right statistics shows the absolute sequence length difference of each TE family within different TE classes between two haplotypes; **b.** Proportion of TEs in different region in HPA and HPB genome; **c.** Proportion of intact TEs which inserted in HDS; **d.** Mapping coverage rate from population genome resequencing data in different genome regions; **e.** Nucleotide diversity within different genome regions; **f.** The length of haplotype blocks within different genome regions; **g.** Normalized haplotype diversity within different genome regions; **h.** The density distribution of HDS and population genetic introgression regions at whole-genome level.

Fig. 3 Comparative analysis of high divergence sequences among the haplotype-resolved genomes of bivalves. a. Density distribution of high divergence sequences (HDS) lengths in each haplotype genome assembly; **b.** The distribution density comparison of HDS with the TEs along the genome across each haplotype-resovled genomes. The distribution density of each genomic component was calculated based on the proportion of sequences per 500 kb window; **c.** Comparison of the average proportion of TEs in HDS of different lengths across each haplotype-resolved genome; **d.** The distribution of homologous genes across the haplotype genomes and HDS of *C. sikamea*, *P. fucata*, *A. senhousia*, and *M. varia*. “N:N:N” indicates orthologous genes present in the all species and “patchy” indicates the existence of other orthologs that are presented in at least one genome; **e.** KEGG enrichment analysis of homologous genes within the HDS between haplotypes of *C. sikamea*, *P. fucata*, *A. senhousia*, and *M. varia*.

Fig. 4 Impact of haplotype divergence on gene evolution and allelic variation in gene expression. a. The proportion of redundant genes in different genome regions of *C. sikamea*; **b.** The proportion of lineage-specific genes at different phylogenetic nodes in different genome regions of *C. sikamea*; **c.** Comparing gene expression levels between different genome regions in the tissues of the digestive gland, gill, and labial palps. Significance testing was conducted using the Wilcox test; **d.** The proportion of identified alleles in different genome regions of *C. sikamea*; **e.** The proportion of allelic sequence similarity in different genome regions of *C. sikamea*; **f.** The expression pattern of alleles of alleles among the tissues of the digestive gland (TD), gill (TG), and labial palps (TL); **g.** The precent of ASEGs in different genome regions.

Fig. 5 Expression profile, genetic variation and evolutionary constraint of the allele-specific expression genes (ASEGs). a. The number of ASEGs identified from specific individuals (individual-ASEGs, noted by solid line) and all individuals (shared-ASEGs, noted by dotted line) from various tissues (including TA, adductor muscles; TM, mantle; TD, digestive gland;TG, gill; and TL, labial palps) using each of the four RNASeq datasets; **b.** The heatmap represents the similarity among the ASE gene sets and tissue-specific ASE gene sets from each tissue of individuals in *RNAseq data set 1*. Gene sets were clustered based on the similarities and the clustering is represented by a cladogram on the top and left of the heatmap; **c.** Heterozygosity of ASEGs and non-ASEGs within exon and promoter regions. **d.** Nucleotide diversity of ASEGs and non-ASEGs among individuals within exon and promoter regions. Consistent-ASEGs indicate the ASEGs that are shared among individuals and exhibit a consistent pattern of ASE across various tissues; Specific-ASEGs indicate the ASEGs that are not shared among individuals. Significance testing was conducted using the Wilcox test. **e.** Evolutionary constraint of ASEGs and non-ASEGs. The PhastCons scores of genes were determined across genomes of nine oyster species of Ostreidae. The ASEGs and non-ASEGs were identified for each individual using four RNAseq datasets.

Fig. 1

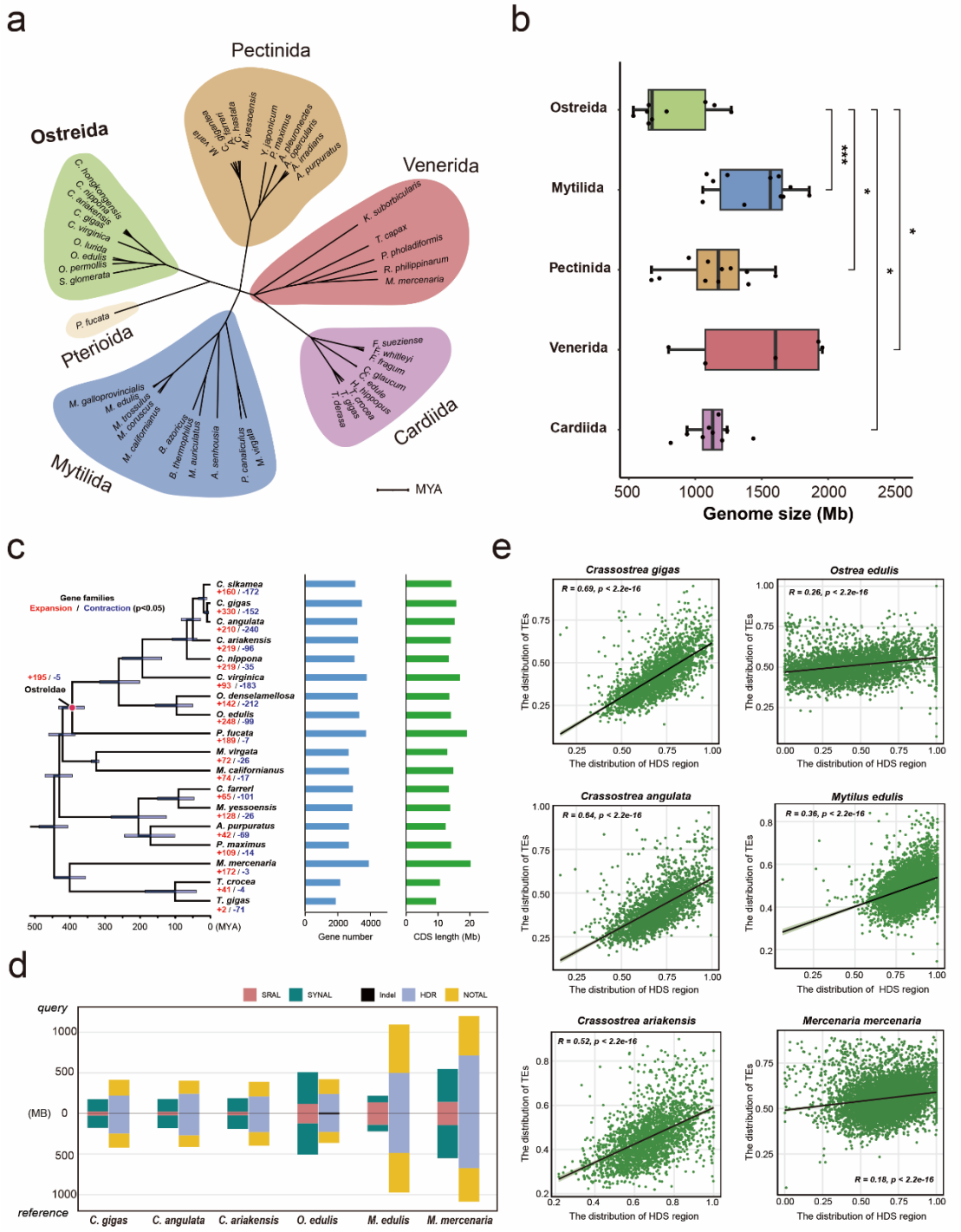

a
Pectinida
Ostreida
Venerida
Pterioida
Cardiida
Mytilida
MYA
b
Ostreida
Mytilida
Pectinida
Venerida
Cardiida
500
1000
1500
2000
2500
Genome size (Mb)
c
Gene families
Expansion / Contraction (p<0.05)
Ostreidae
500
400
300
200
100
0 (MYA)
Gene number
CDS length (Mb)
d
SRAL
SYNAL
Indel
HDR
NOTAL
query
(MB)
reference
C. gigas
C. angulata
C. ariakensis
O. edulis
M. edulis
M. mercenaria
e
Crassostrea gigas
R = 0.69, p < 2.2e-16
Ostrea edulis
R = 0.26, p < 2.2e-16
Crassostrea angulata
R = 0.64, p < 2.2e-16
Mytilus edulis
R = 0.36, p < 2.2e-16
Crassostrea ariakensis
R = 0.52, p < 2.2e-16
Mercenaria mercenaria
R = 0.18, p < 2.2e-16
The distribution of TEs
The distribution of HDS region

Fig. 2

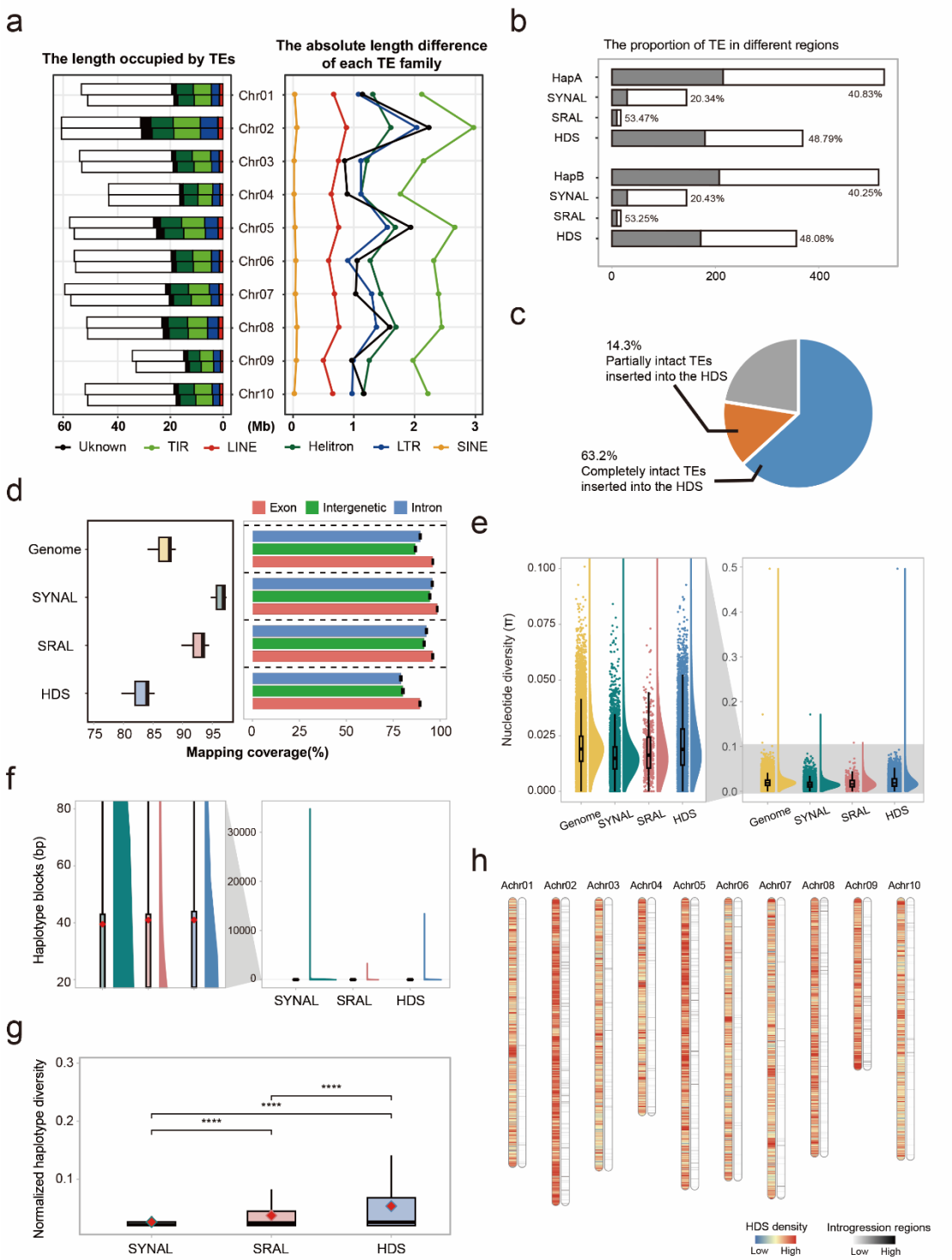

a
The length occupied by TEs
The absolute length difference of each TE family
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
60 40 20 0 (Mb) 0 1 2 3
Uknown TIR LINE Helitron LTR SINE
b
The proportion of TE in different regions
HapA
SYNAL
SRAL
HDS
HapB
SYNAL
SRAL
HDS
20.34%
40.83%
53.47%
48.79%
20.43%
40.25%
53.25%
48.08%
0 200 400
c
14.3%
Partially intact TEs inserted into the HDS
63.2%
Completely intact TEs inserted into the HDS
d
Exon Intergenetic Intron
Genome
SYNAL
SRAL
HDS
75 80 85 90 95
0 25 50 75 100
Mapping coverage(%)
e
Nucleotide diversity (π)
0.100
0.075
0.050
0.025
0.000
0.5
0.4
0.3
0.2
0.1
0.0
Genome SYNAL SRAL HDS
f
Haplotype blocks (bp)
80
60
40
20
30000
20000
10000
0
SYNAL SRAL HDS
g
Normalized haplotype diversity
0.3
0.2
0.1
****
****
****
SYNAL SRAL HDS
h
Achr01 Achr02 Achr03 Achr04 Achr05 Achr06 Achr07 Achr08 Achr09 Achr10
HDS density
Low High
Introgression regions
Low High

Fig. 3

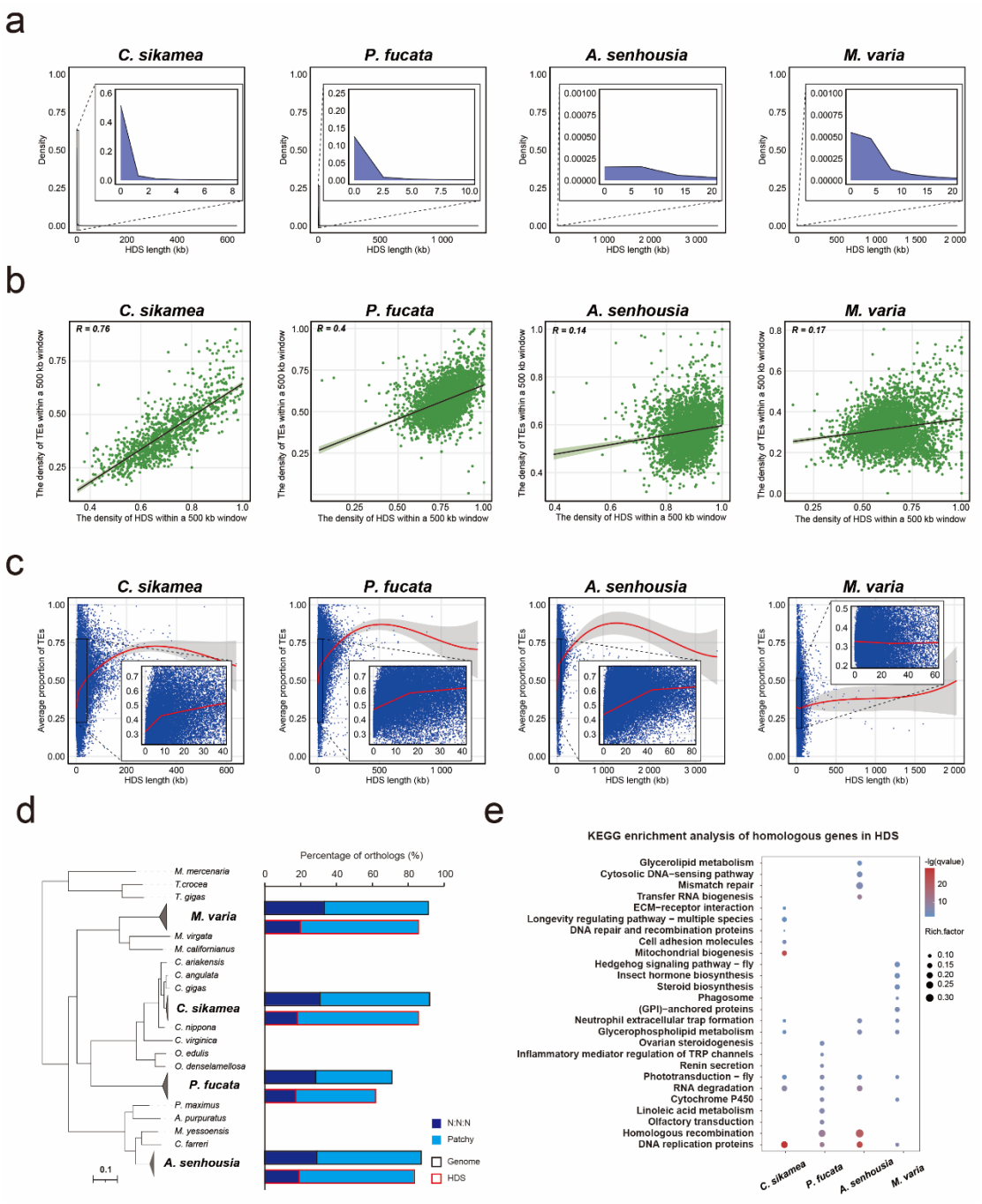

a
C. sikamea
P. fucata
A. senhousia
M. varia
Density
HDS length (kb)
b
R = 0.76
R = 0.4
R = 0.14
R = 0.17
The density of TEs within a 500 kb window
The density of HDS within a 500 kb window
c
Average proportion of TEs
HDS length (kb)
d
Percentage of orthologs (%)
M. mercenaria
T.crocea
T. gigas
M. varia
M. virgata
M. californianus
C. ariakensis
C. angulata
C. gigas
C. sikamea
C. nippona
C. virginica
O. edulis
O. denselamellosa
P. fucata
P. maximus
A. purpuratus
M. yessoensis
C. farreri
A. senhousia
0.1
N:N:N
Patchy
Genome
HDS
e
KEGG enrichment analysis of homologous genes in HDS
Glycerolipid metabolism
Cytosolic DNA−sensing pathway
Mismatch repair
Transfer RNA biogenesis
ECM−receptor interaction
Longevity regulating pathway − multiple species
DNA repair and recombination proteins
Cell adhesion molecules
Mitochondrial biogenesis
Hedgehog signaling pathway − fly
Insect hormone biosynthesis
Steroid biosynthesis
Phagosome
(GPI)−anchored proteins
Neutrophil extracellular trap formation
Glycerophospholipid metabolism
Ovarian steroidogenesis
Inflammatory mediator regulation of TRP channels
Renin secretion
Phototransduction − fly
RNA degradation
Cytochrome P450
Linoleic acid metabolism
Olfactory transduction
Homologous recombination
DNA replication proteins
C. sikamea
P. fucata
A. senhousia
M. varia
-lg(qvalue)
Rich.factor

Fig. 4

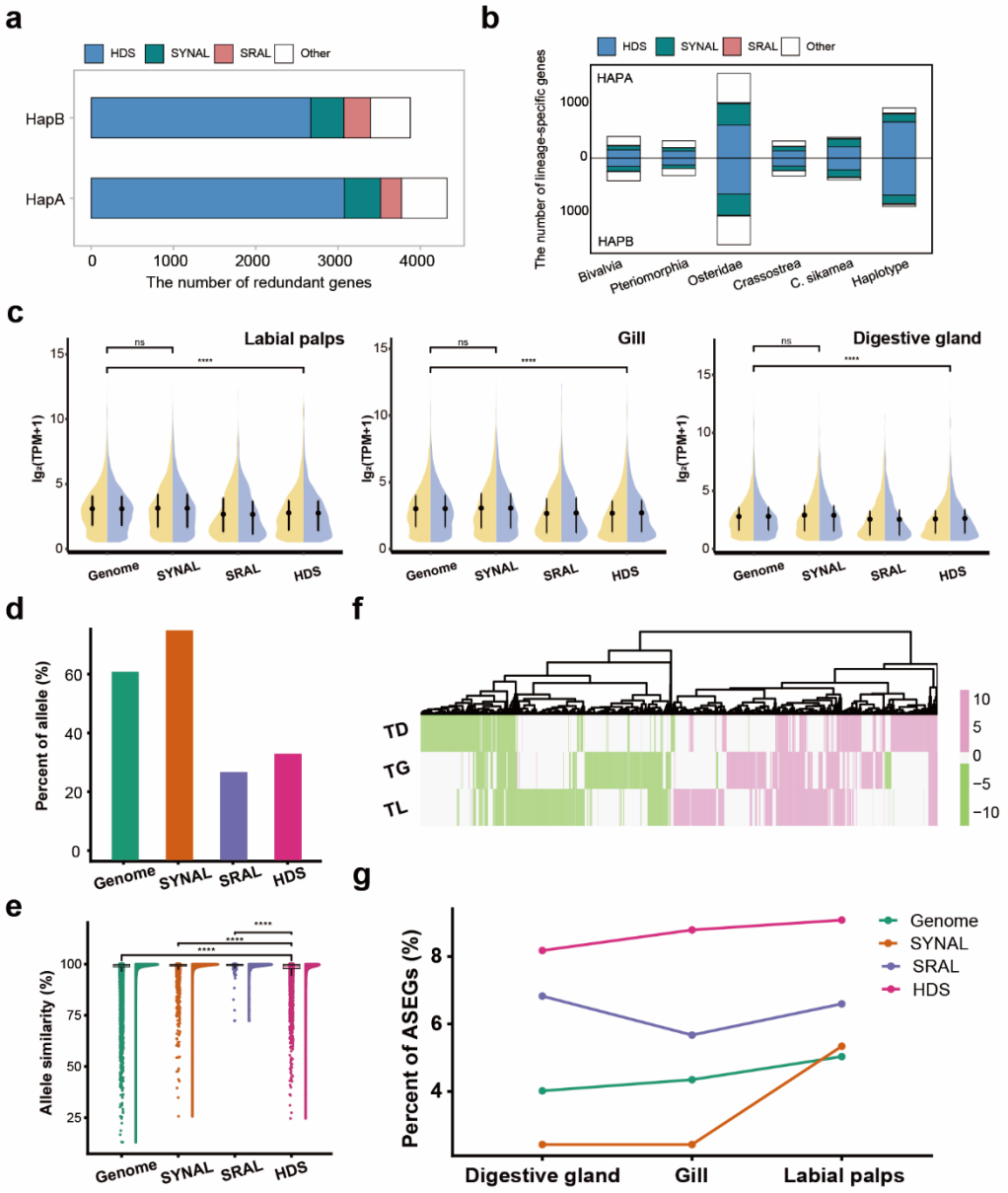
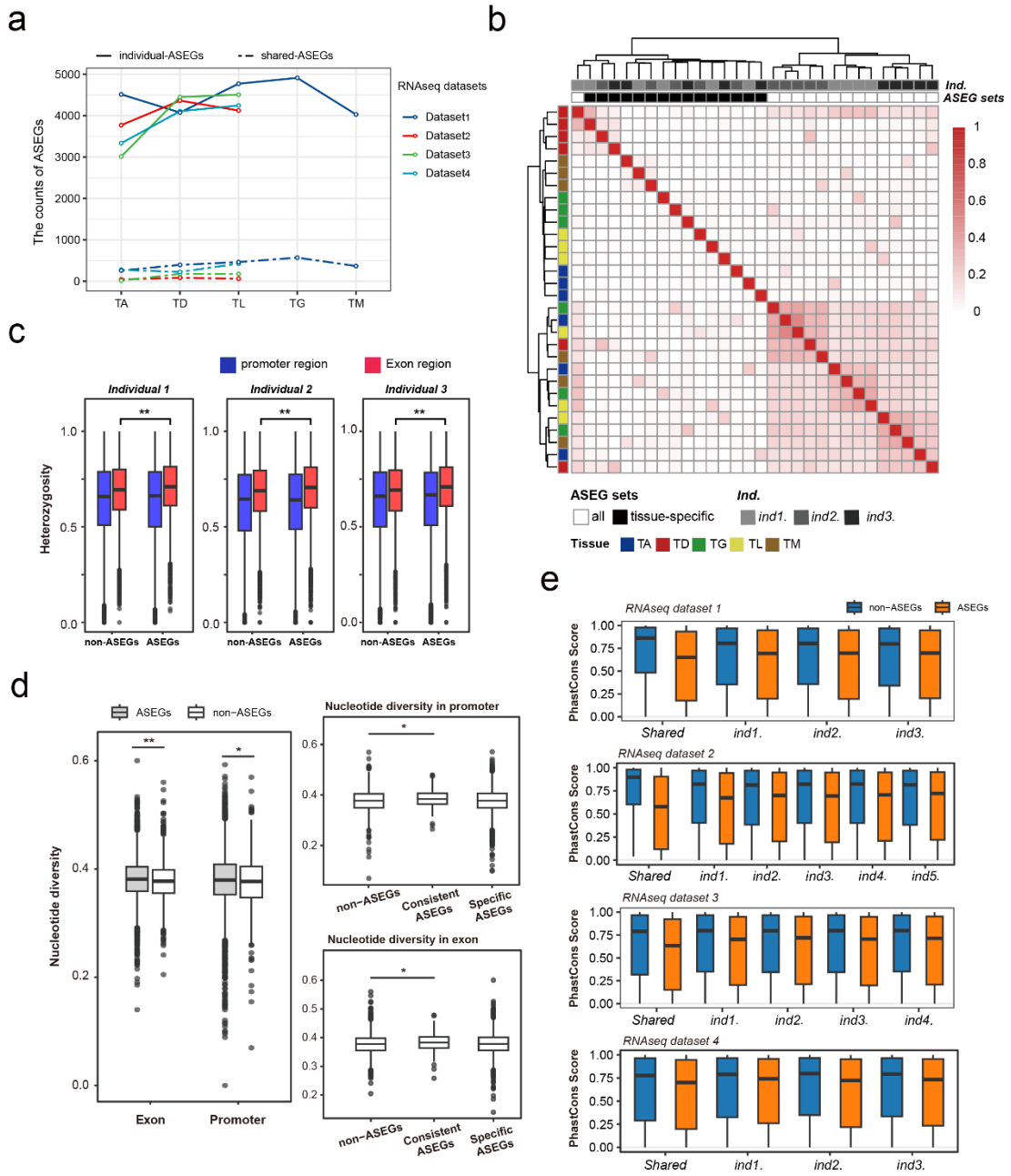
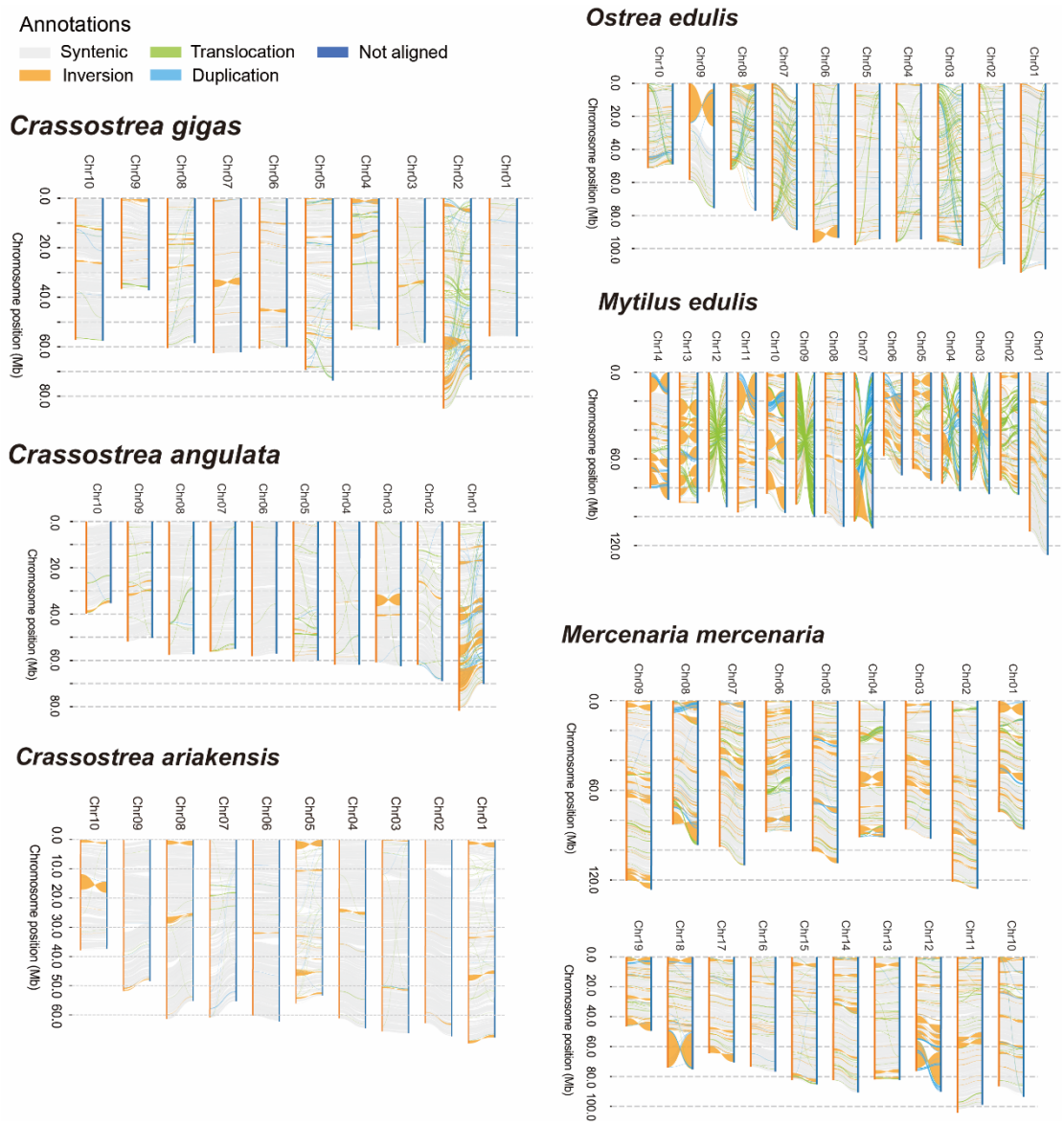

a
HDS
SYNAL
SRAL
Other
HapB
HapA
0
1000
2000
3000
4000
The number of redundant genes
b
HDS
SYNAL
SRAL
Other
HAPA
HAPB
The number of lineage-specific genes
1000
0
1000
Bivalvia
Pteriomorphia
Osteridae
Crassostrea
C. sikamea
Haplotype
c
Labial palps
Gill
Digestive gland
ns
****
log2(TPM+1)
Genome
SYNAL
SRAL
HDS
d
Percent of allele (%)
Genome
SYNAL
SRAL
HDS
f
TD
TG
TL
10
5
0
−5
−10
e
Allele similarity (%)
Genome
SYNAL
SRAL
HDS
g
Percent of ASEGs (%)
Genome
SYNAL
SRAL
HDS
Digestive gland
Gill
Labial palps

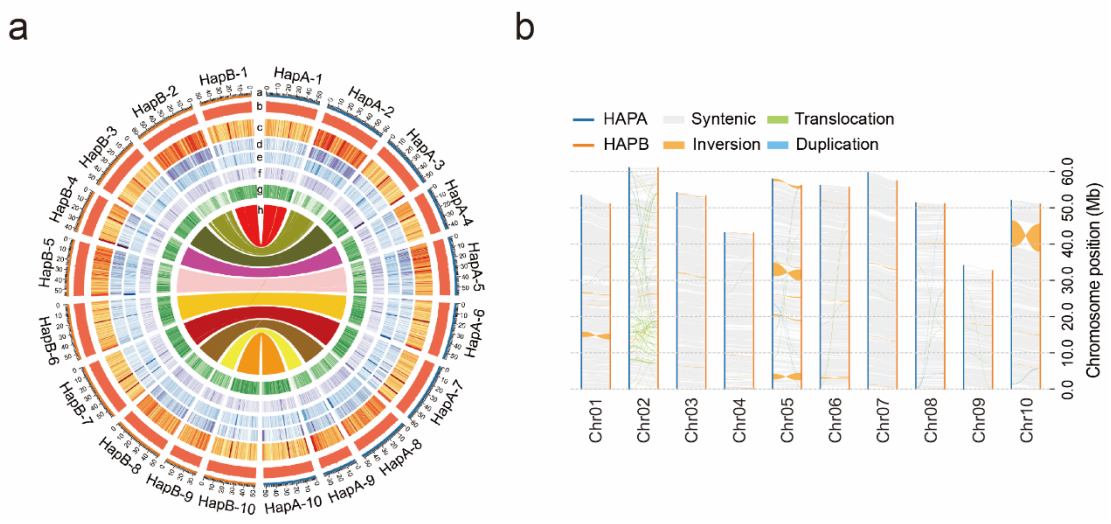
Fig. 5


a
individual-ASEGs
shared-ASEGs
The counts of ASEGs
RNAseq datasets
Dataset1
Dataset2
Dataset3
Dataset4
TA
TD
TL
TG
TM
b
Ind.
ASEG sets
ASEG sets
all
tissue-specific
Ind.
ind1.
ind2.
ind3.
Tissue
TA
TD
TG
TL
TM
c
promoter region
Exon region
Individual 1
Individual 2
Individual 3
Heterozygosity
non-ASEGs
ASEGs
d
ASEGs
non-ASEGs
Nucleotide diversity
Exon
Promoter
Nucleotide diversity in promoter
Nucleotide diversity in exon
non-ASEGs
Consistent ASEGs
Specific ASEGs
e
RNAseq dataset 1
RNAseq dataset 2
RNAseq dataset 3
RNAseq dataset 4
PhastCons Score
Shared
ind1.
ind2.
ind3.
ind4.
ind5.

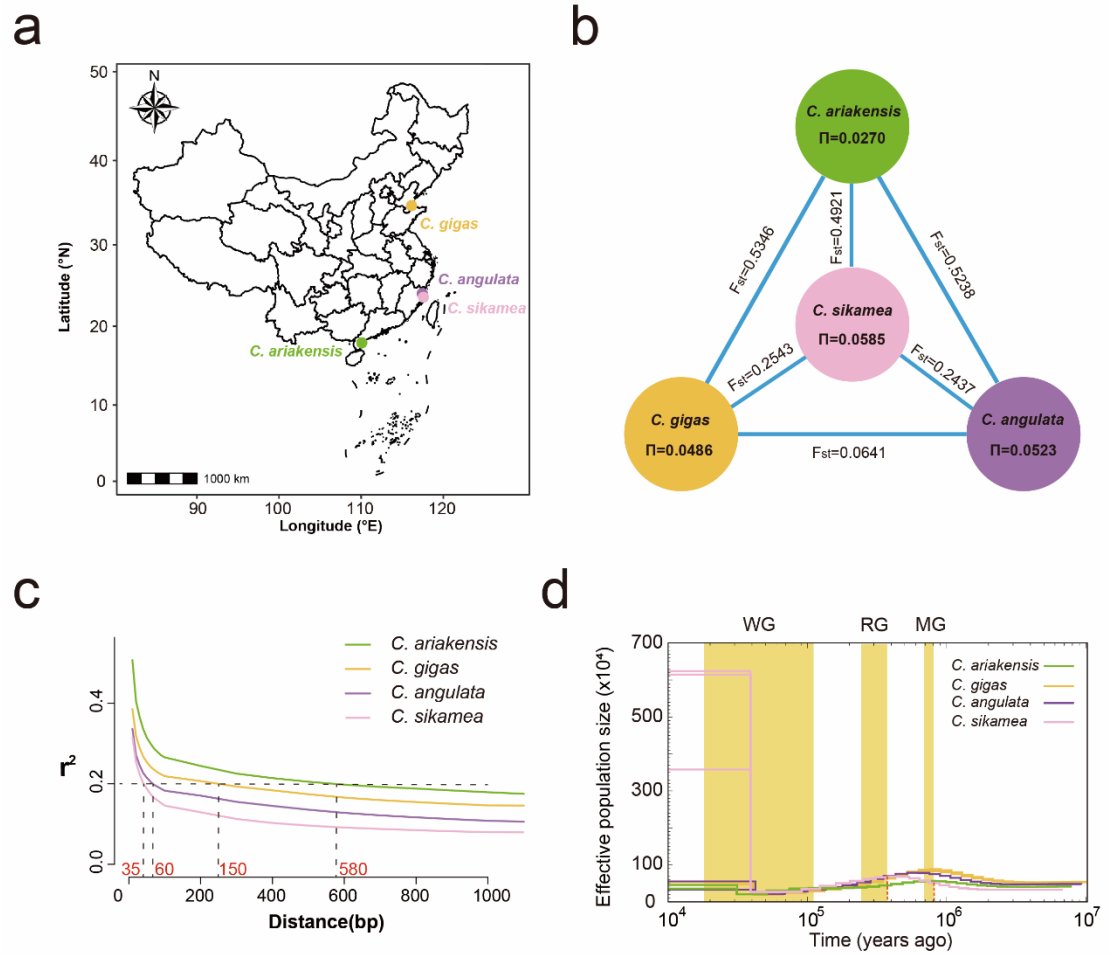
Extended Data:



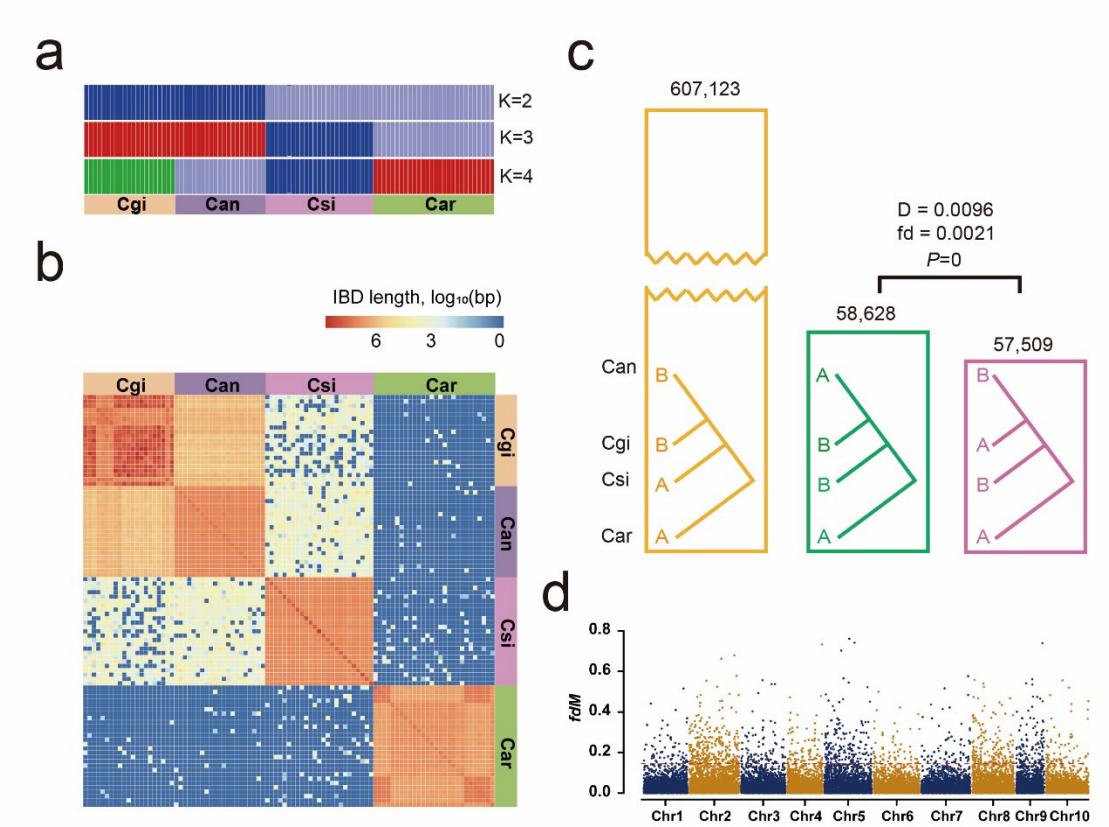
Extended Data Fig. 1 The comparing genomes between individuals across different bivalvia species. Structural rearrangements and synteny are represented by colored lines connecting regions. The chromosomes are numbered according to the sequence of the reference genome, such as for the *C. gigas* (GCA_902806645.1), *C. angulata* (GCA_025765675.3), *C. ariakensis* (GCA_020567875.1), *O. edulis* (GCA_947568865.1), *M. edulis* (GCA_963676685.2), and *M. mercenaria* (GCA_014805675.2); In addition, the proportion of structural rearrangements and collinearity proportions in haplotype or individual genomes of different bivalve species were also analyzed. The detailed information for genome assemblies and comparation results refers to Supplementary Table 2;



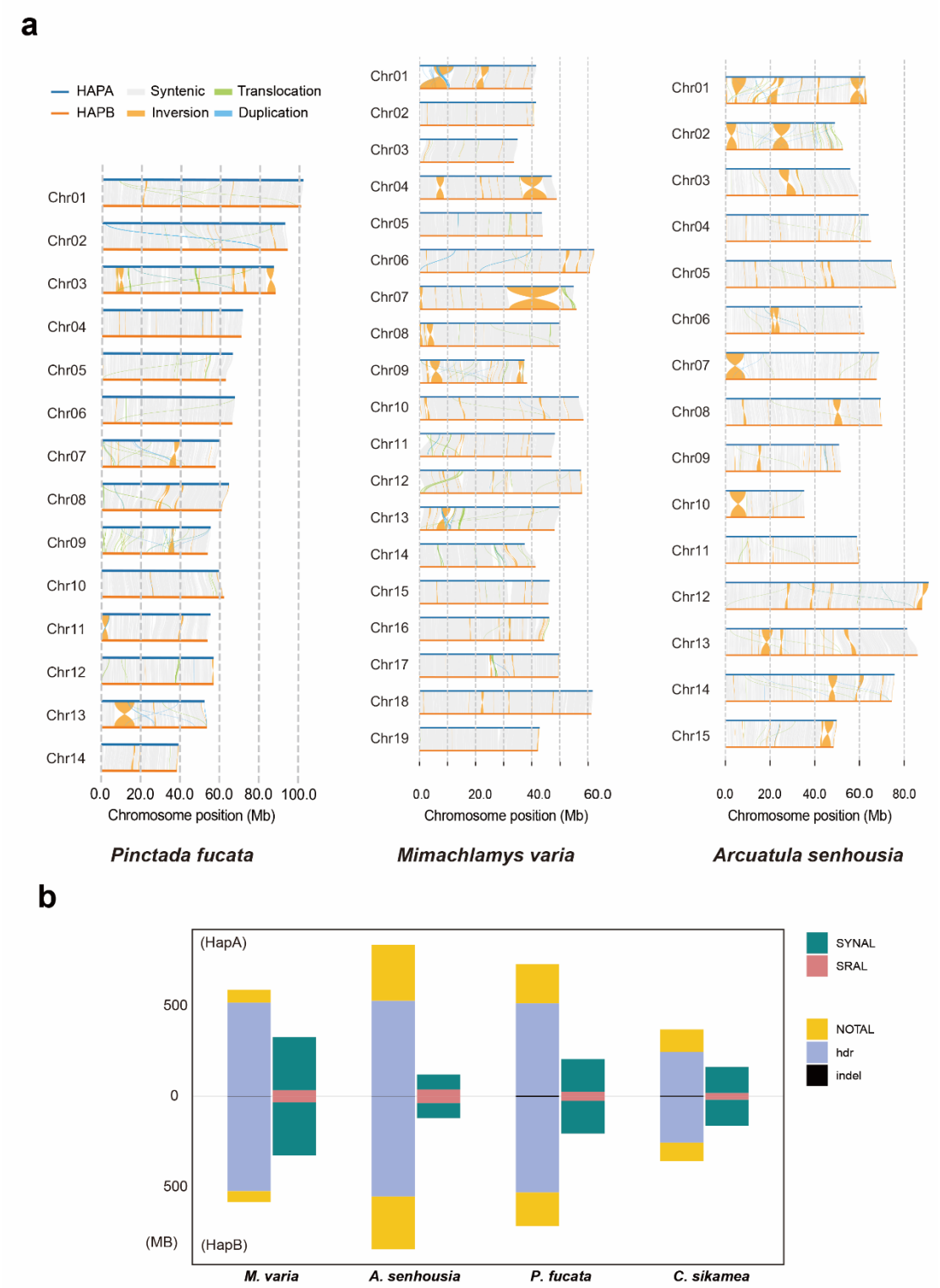
Extended Data Fig. 2 The assembly and comparasion of *C.sikamea* haplotyp-resloved genome. a, Circos plots showing 20 chromosomes (a), the distribution of GC content (b), transposable elements (c), LTR (d), TIR (e), Helitron (f), Gene (g) and the synteny of Gene; **b,** Structural rearrangements and synteny are represented by colored lines connecting regions of the haplotype-resolved chromosome-level genome of *C. sikamea*.



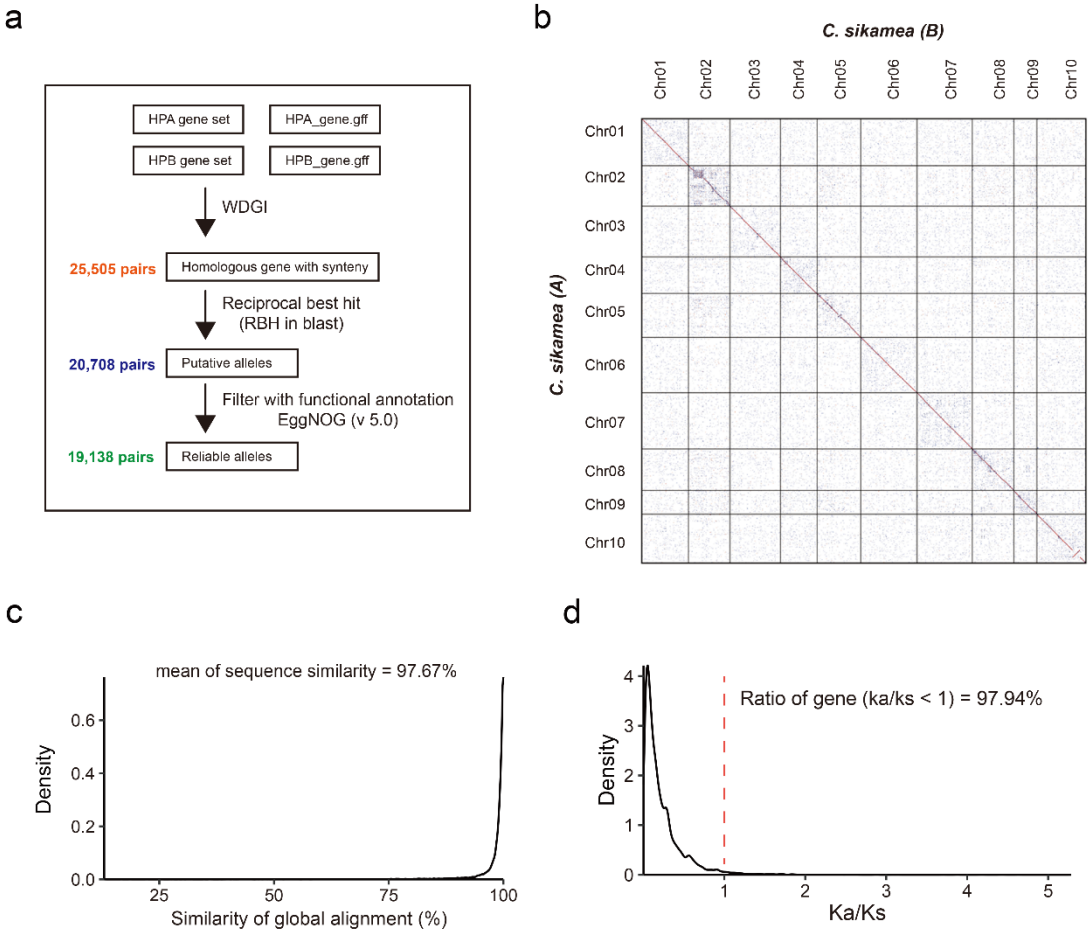
Extended Data Fig. 3 The population structure and demographic of four *Crassostrea* species. a, Sampling sites of *C. gigas*, *C. angulata*, *C. sikamea*, and *C. ariakensis*. **b,** Nucleotide diversity (Π) and genetic divergence (*F*st, between sampled individuals) among four oyster species; **c.** Decay of linkage disequilibrium in four oyster species; **d.** The demographic history of the four oyster species was inferred using the pairwise sequentially Markovian coalescent (PSMC) model. Periods of the Mindel glaciation (MG, 0.68-0.80 million years ago), Riss glaciation (MG, 0.24-0.37 million years ago), and Würm glaciation (WG, 10,000-120,000 years ago) were indicated by yellow shading.



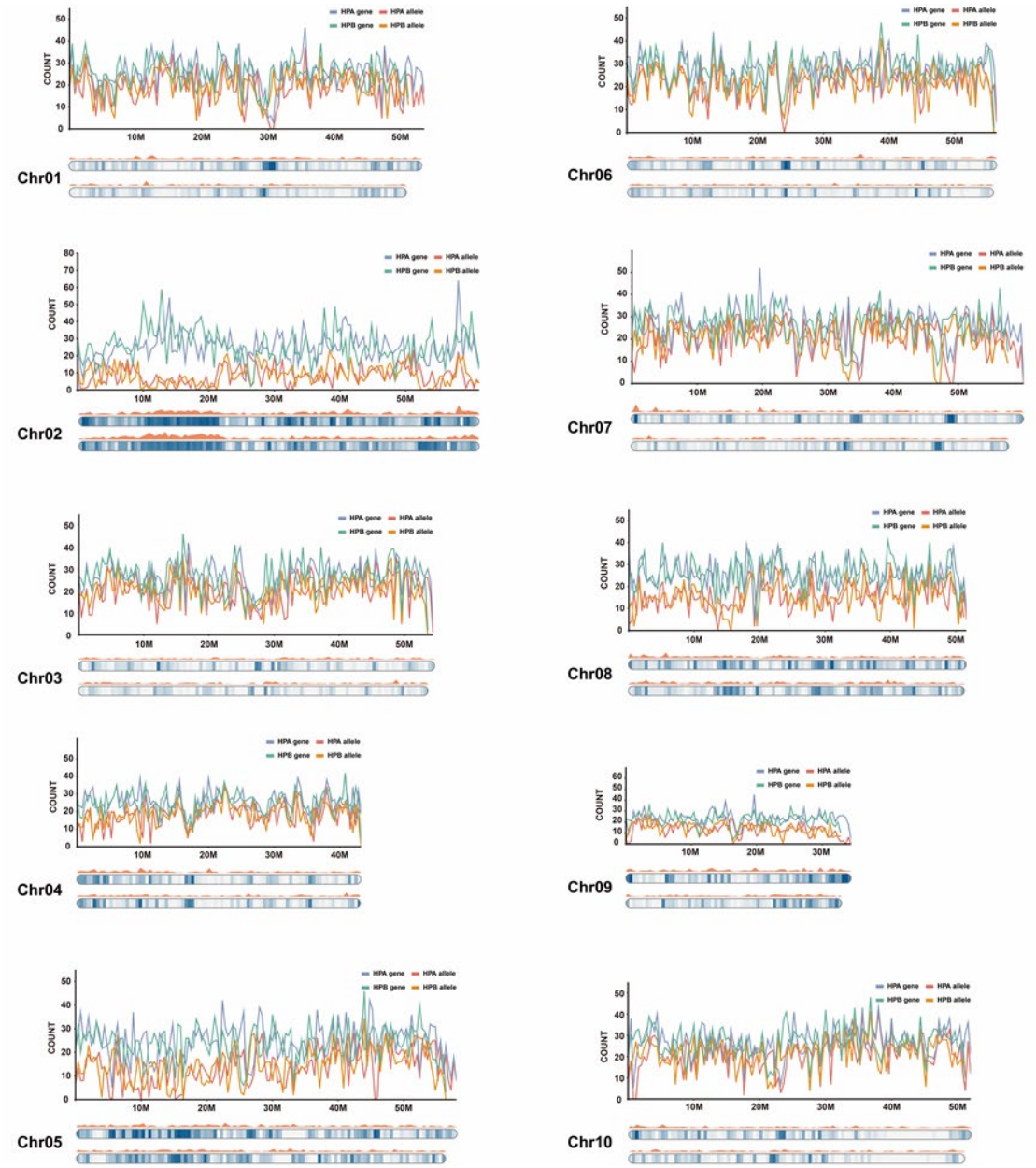
Extended Data Fig. 4 The introgression of four Crassostrea species. a. Admixture analysis was performed with K=4, where the CV error value was the lowest. The Admixture plot shows the inferred genetic components for each individual. The colors in the Structure plot represent the different genetic components present in each individual sample; *C. ariakensis* (Car), *C. sikamea* (Csi), *C. angulata* (Can) and *C. gigas* (Cgi)**. b.** Estimated haplotype sharing between individuals of the four oyster species. The colors in the heatmap represent the length of identifcal by decent (IBD) blocks shared between individuals; **c.** ABBA-BABA analysis of the four oyster species. The number of sites supporting different tree structures was shown in a bar plot. To test the null hypothesis of symmetry in genetic relationships, *D* and *fd* statistics were provided, along with the corresponding block jackknife corrected *P* value; **d,** Genomic characteristics of the putative introgressed regions between *C.sikamea* and *C. gigas*. Manhattan plot showing the f_{dM} value between *C.sikamea* and *C. gigas*. The red dashed line shows the cutoff value, calculated by highest 1% of f_{dM} values.



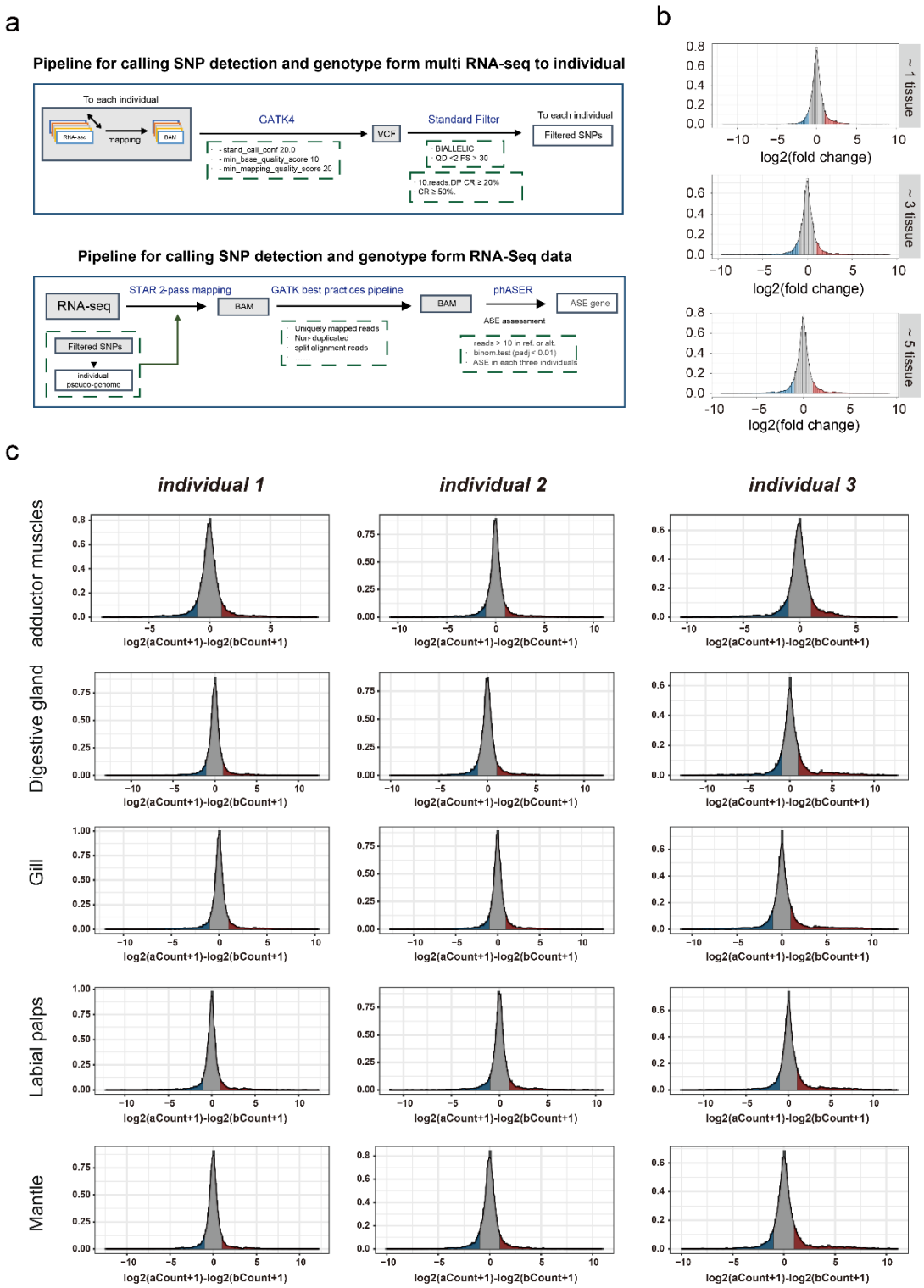
Extended Data Fig. 5. The comparing haplotype genomes of *A. senhousia*, *M. varia*, and *P. fucata*. a. Structural rearrangements and synteny are represented by colored lines connecting regions of the haplotype-resolved chromosome-level genome of *A. senhousia, M. varia. a, and P. fucata.*; **b.** Proportions of syntenic alignment (SYNAL), structural rearrangement alignments (SRAL) and high divergence sequences (HDR, NOTAL and Indels) between haplotype genomes.



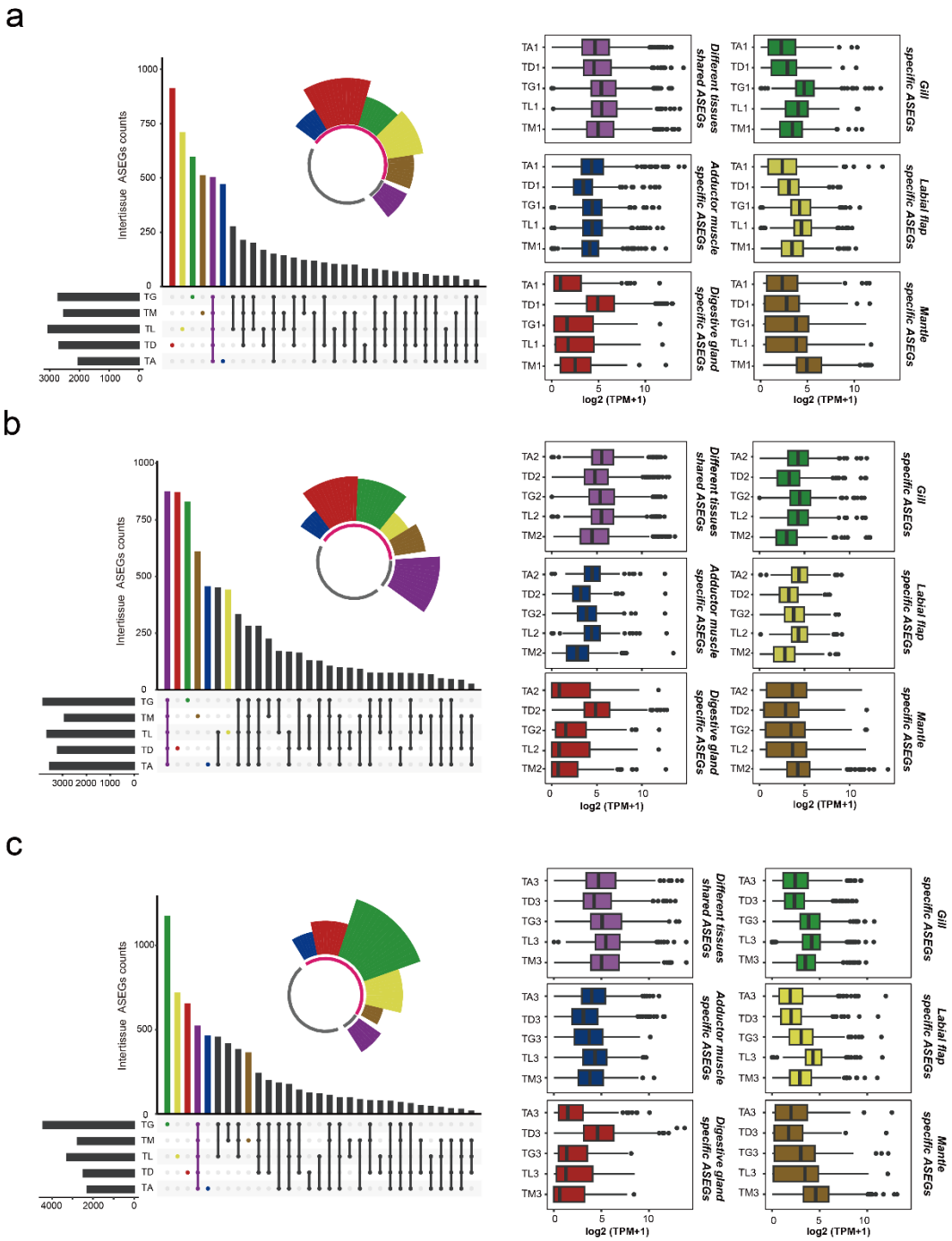
Extended Data Fig 6. Identified alleles pipeline details and comparative analysis of gene homology between haplotype genomes and sequence comparison of allelic genes. a, Diagram of the approach taken to identify alleles between haplotype genomes. **b,** The homologous gene dotplot between haplotype genomes; **c,** The sequences similarity of identified alleles; **d,** The Ka/Ks ratio of identified alleles.



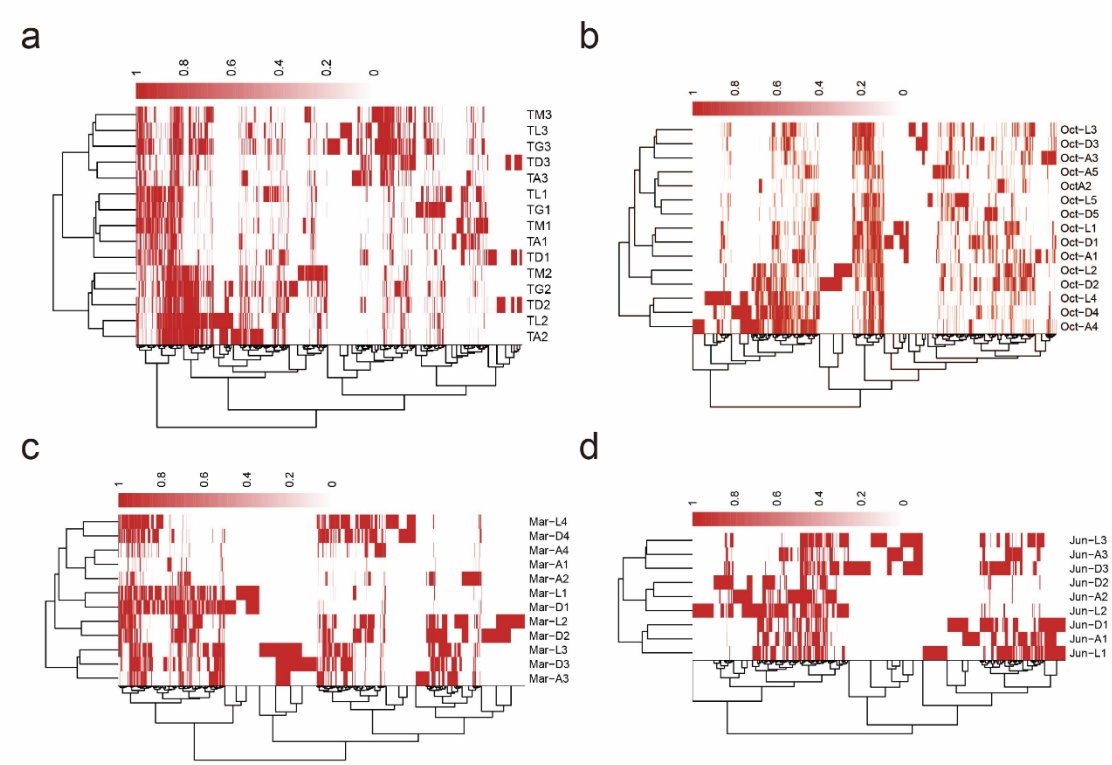
Extended Data Fig 7. The density distribution of gene and identified alleles in genome-wide with the compared with the distribution of not-aligned region (chromosome diagram, blue) and redundant genes (chromosome diagram, orange).



Extended Data Fig 8. Transcriptome-based SNP detection and ASE analysis process and feasibility analysis. a, Workflow used to detect SNPs and analysis Allele-Specific Expression (ASE) from different RNA-seq data for each individual; **b,** With the tissue addition for calling SNP detection, the histograms of a global reference allelic ratio at every gene locus in the adductor muscle. **c,** Analysis of the effectiveness of using multi RNA-seq data for individual allele expression (AE) in mapping bias control.



Extended Data Fig 9. The comparative analysis of allele-specific expression (ASE) gene sets from five different tissues of the individual 1 (a), individual 2 (b) and individual (b). The upper-right coxcomb chart shows the proportion of tissue-specific ASEGs and tissue-shared AEGs in the individual 1 (a), individual 2 (b) and individual (c). The right figure is comparing the overall expression levels of the ASEGs shared in all tissues or tissues- specific ASEGs among different tissues in the individual 1 (a) , individual 2 (b) and individual (c).; The colors represent the ASEGs shared in all tissues (purple) or tissues- specific ASEGs in different tissues (blue, adductor muscles; red, digestive gland; green, gill; yellow, labial palps; and brown, mantle);



Extended Data Fig 10. Comparing ASE gene sets from different organizations and tissue-specific ASE gene sets. a. *RNA-seq dataset1* is from five different tissues of three individuals (TA, adductor muscles; TD, digestive gland; TG, gill; TL, labial palps; and TM, mantle); **b.** *RNA-seq dataset2* is from three different tissues of five individuals. **c.** *RNA-seq dataset3* is from four different tissues of four individuals. d. *RNA-seq dataset4* is from three different tissues of three individuals.