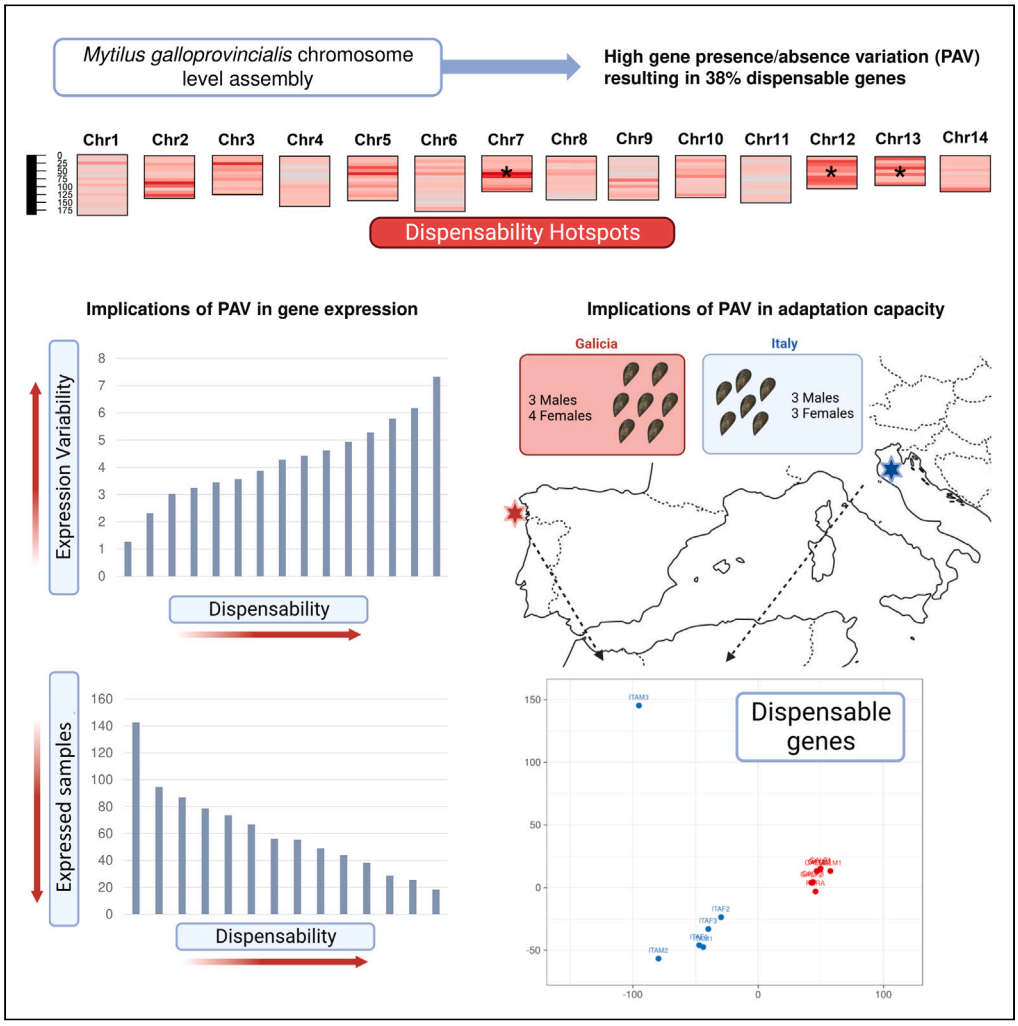
iScience

CellPress
OPEN ACCESS

Article

Gene presence/absence variation in *Mytilus galloprovincialis* and its implications in gene expression and adaptation



Amaro Saco, Magalí Rey-Campos, Cristian Gallardo-Escárate, Marco Gerdol, Beatriz Novoa, Antonio Figueras

antoniofigueras@iim.csic.es

Highlights

PAV genes are widespread in mussel chromosomes, but with dispensability hotspots

Dispensable genes show expression variation due to the dispersion among individuals

Dispensable and core myticins are the highest expressed genes of the whole genome

Dispensable genes are involved in local adaptation

Saco et al., iScience *26*, 107827 October 20, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.107827

iScience

CellPress
OPEN ACCESS

Article

Gene presence/absence variation in *Mytilus galloprovincialis* and its implications in gene expression and adaptation

Amaro Saco,[1] Magalí Rey-Campos,[1] Cristian Gallardo-Escárate,[2] Marco Gerdol,[3] Beatriz Novoa,[1] and Antonio Figueras[1,4,*]

SUMMARY

Presence/absence variation (PAV) is a well-known phenomenon in prokaryotes that was described for the first time in bivalves in 2020 in *Mytilus galloprovincialis*. The objective of the present study was to further our understanding of the PAV phenomenon in mussel biology. The distribution of PAV was studied in a mussel chromosome-level genome assembly, revealing a widespread distribution but with hotspots of dispensability. Special attention was given to the effect of PAV in gene expression, since dispensable genes were found to be inherently subject to distortions due to their sparse distribution among individuals. Furthermore, the high expression and strong tissue specificity of some dispensable genes, such as myticins, strongly supported their biological relevance. The significant differences in the repertoire of dispensable genes associated with two geographically distinct populations suggest that PAV is involved in local adaptation. Overall, the PAV phenomenon would provide a key selective advantage at the population level.

INTRODUCTION

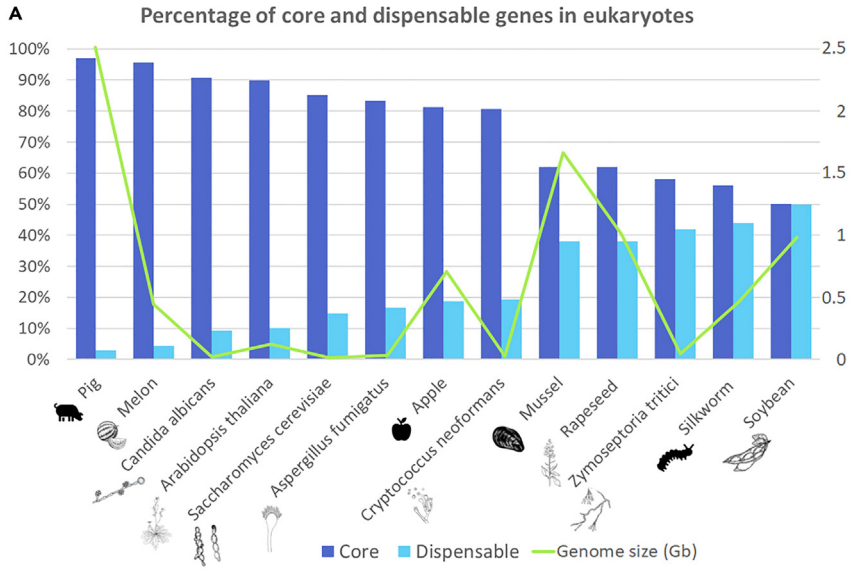
Genomic structural variation (SV) consists of large-scale differences of different types (translocations, inversions, duplications, insertions, deletions, and copy number variation) within the genome of a species and detectable through comparison among different haplogroups.[1] SVs are arbitrarily defined as such whenever they involve regions larger than 1 kb to differentiate them from smaller-scale modifications (e.g., SNPs, short tandem repeats, etc.) that are widespread in all genomes.[2] Large-scale SV can affect both intergenic and gene-encoding regions, frequently resulting in intraspecific variation in gene content, defined as gene presence-absence variation (PAV), with biological implications that are potentially much more relevant than small-scale variants.

Changes in gene content among different individuals or populations belonging to the same species are the foundations of a pangenome. Originally described in microbial and viral genomes, pangenomes are defined by a set of core genes shared by all individuals and dispensable genes that are absent in some individuals.[3,4] The presence of dispensable genes in some populations may have advantageous effects due to accessory functions provided by the encoded proteins, allowing adaptation to different ecological niches.[5] In addition to prokaryotes, open pangenomes (i.e., pangenomes with a high dispensable:core gene ratio) have been described in plants, microalgae, and fungi, where the dispensable genes cf. adaptive advantages mainly associated with biotic and abiotic stress resistance.[6–9] Until recently, only close pangenomes, with a low ratio of dispensable genes, had been reported in the animal kingdom. This was the case for humans[10,11] and pigs,[12] with SV mostly linked with intergenic regions and variations on the order of 1–10% in gene content. Similar variations have been reported in other chordates, such as the barn swallow bird *Hirundo rustica*,[13] the ancient eel *Anguilla japonica*,[14] and the Atlantic cod *Gadus morhua*.[15,16] Although widespread structural polymorphisms have been described in invertebrates, such as urochordates[17] and bivalve mollusks,[18] none of these studies investigated the impact of these phenomena on gene content. Moreover, the few studies reporting extreme cases of interindividual genetic diversity, such as those carried out in nematodes,[19,20] have been mostly based on SNP analyses, disregarding the potentials of SV and the associated gene PAV. The first open pangenome of the animal kingdom with widespread PAV, affecting 38% of all annotated genes, was reported in 2020 in the mussel *Mytilus galloprovincialis* through a comparative genomic analysis that involved the resequencing of several individuals.[21] After this study, which represents a paradigm shift, more pangenomes with large dispensable fractions were reported in other invertebrates, such as the silkworm *Bombyx mori*.[22] These reports demonstrated that the genomes of invertebrate animals can reach gene PAV levels similar to those observed in several plants and microbes (Figure 1).

[1]Institute of Marine Research, Spanish National Research Council, Vigo, Spain
[2]Center for Aquaculture Research, University of Concepción, Concepción, Chile
[3]Department of Life Sciences, University of Trieste, Trieste, Italy
[4]Lead contact
*Correspondence: antoniofigueras@iim.csic.es
https://doi.org/10.1016/j.isci.2023.107827

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

A



B

Species	RefSeq Representative Genome size (Gb)	Number of analyzed individuals	Dispensable	Reference
Sus scrofa	2.5	12	3.00%	Tian et al., 2020
Cucumis melo	0.438	5	4.40%	González et al., 2013
Candida albicans	0.014	34	9.37%	McCarthy et al., 2019
Arabidopsis thaliana	0.119	80	10.00%	Tan et al., 2012
Saccharomyces cerevisiae	0.012	100	14.92%	McCarthy et al., 2019
Aspergillus fumigatus	0.029	12	16.71%	McCarthy et al., 2019
Malus domestica	0.703	36	18.70%	Sun et al., 2020
Cryptococcus neoformans	0.019	25	19.44%	McCarthy et al., 2019
Mytilus galloprovincialis	1.658	16	38.00%	Gerdol et al., 2020
Brassica napus	1.001	53	38.00%	Hurgobin et al., 2018
Zymoseptoria tritici	0.039	5	41.90%	Plissonneau et al., 2018
Bombyx mori	0.46	100	44.00%	Tong et al., 2022
Glycine max	0.978	26	49.88%	Liu et al., 2020

Figure 1. Gene presence/absence variation (PAV) phenomenon

(A) Proportion of core/dispensable genes in the mussel genome (mg10) and the genomes of other eukaryotic species where a pangenomic architecture has been described (left scale bar). The green line indicates the genome size (Gb, right scale bar) considering the RefSeq Representative Genome.

(B) Table showing data on genome size, number of resequenced individuals, percentage of dispensable genes, and reference information for each pangenome.

In genome assemblies of diploid organisms, such as mussels, PAV occurs between homologous chromosomes within an individual, resulting in large hemizygous genomic regions, which are present in only one of the two homologous chromosomes. Since such regions and the genes they contain are subjected to Mendelian inheritance, at the population level, each individual may have one (hemizygous), two (homozygous), or zero copies (nullizygous) of each dispensable DNA block (and associated genes), depending on the combination between parental haplotypes.[21,23] Hemizygous regions are widespread in mollusk genomes and particularly important in bivalves, suggesting that pangenomes could be extensively present in this mollusk class.[23] In fact, recent reports have revealed additional bivalve pangenomes, with SVs affecting gene content in the clam *Mercenaria mercenaria*[24] and in the oysters *Pinctada fucata*[25] and *Crassostrea virginica*.[26]

Marine bivalve mollusks are sessile filter feeders with a global distribution, which inhabit different environments, and are usually found in high-density populations.[27] These species are subjected to threats due to the high concentration and diversity of bacterial and viral pathogens present in the sea[28,29] and the harmful environmental factors associated with their habitats.[30] The genetic variation maintained in the mussel pangenome has been suggested to strengthen the adaptation capacity and resilience to a broad range of stressors. Such considerations derive from the strong enrichment of the dispensable fraction in gene families subjected to lineage-specific expansions in bivalves, which provide functions related to stress and immune response.[21] Several immune gene families are widely expanded and subjected to PAV in mussels, such as IL-17,[31] C1q proteins,[32] and antimicrobial peptides, including myticins[33] and defensins.[34] Nevertheless, gene PAV is likely not an exclusive feature of mussel genomes, as similar phenomena are the main drivers of interindividual sequence diversity in oyster

defensins,[35] and preliminary studies have hinted that expanded immune gene families are commonly associated with hemizygous genomic regions in different bivalve species.[23]

The objective of this study was to deepen our understanding of the PAV phenomenon in *Mytilus galloprovincialis* and its implications in mussel biology. The chromosomal distribution of PAV was studied to detect hotspots of hemizygosity and dispensability. Moreover, we investigated the differences in the repertoire of dispensable genes for two mussel populations, revealing a clear pangenome imprint on adaptation capacity, which was further supported by the important effect of gene PAV on expression profiles in all major mussel tissues.

RESULTS

Mussel chromosome-level assembly and presence/absence variation

The mg3 chromosome level assembly placed 99.42% of the mg3 scaffolds into 14 chromosomes. The resulting genome consisted of 1.8Gb and encoded for 77,769 protein coding genes, higher than the 1.2 Gb and 60,302 coding genes found in the reference mg10 scaffold level assembly. This difference was caused by a filtering step of the hemizygous fraction performed during the construction of the mg10 assembly, while this fraction was kept complete in mg3 in order to include all the gene content. Data concerning the chromosome level assembly and its comparison with previous *Mytilus galloprovincialis* genomic assemblies is presented in Tables S1–S3.

After running the presence/absence pipeline,[36] 48,313 out of the 77,769 genes annotated in the mg3 chromosomal assembly were classified as core (62%) and 29,456 (38%) as dispensable, based on their identification in all or absence in at least one of the resequenced individuals, respectively. Figure 2A shows an example of the coverage graphs obtained from the analysis of read mapping profiles from each individual. Typically, such graphs are characterized by the presence of a homozygous peak (2n) of coverage, which corresponds to genes present with two alleles in the diploid genome. This peak invariably matches the expected coverage used for sequencing. A second hemizygous peak (n), placed at exactly half the coverage of the 2n peak, includes genes associated with hemizygous genomic regions, only present with a single allele in the diploid genome and potentially affected by PAV at the population level.[23] A third peak would appear at zero coverage only when mapping genomic reads from a different resequenced individual against the reference assembly. This third peak includes dispensable genes that are present in the reference genome, but do not match any read from the resequenced individual, where they are absent. As expected, only homozygous and hemizygous peaks could be observed when genomic reads from the reference individual were mapped to the reference assembly (Figure 2B). The comparison between the coverage curves obtained for all annotated genes (blue curves) and for the complete BUSCO genes (orange curves) revealed a marked difference between the two gene sets, with the latter (enriched in genes with housekeeping functions) displaying a higher proportion of genes located in the homozygous peak. When mapping reads from a resequenced individual, the hemizygous peak is reduced since several dispensable genes are absent in that particular individual, therefore appearing in a peak at zero coverage (Figure 2C). The merging of all the genes missing in at least one individual allowed the generation of a complete list of dispensable genes. Most dispensable genes were present in nearly all individuals (Figure 2D), thereby falling within a category of dispensable genes that some authors define as "soft core" or "persistent".[37,38] Fewer dispensable genes were missing in a higher fraction of individuals, displaying intermediate or high dispensability, consistent with "shell" and "cloud" dispensable gene definitions, respectively,[39] to the point that approximately 1,000 genes were exclusively found in the reference individual and absent in all the resequenced mussels.

The spatial distribution of dispensable genes was analyzed in the chromosome-level assembly of *Mytilus galloprovincialis* using the *M. chilensis* assembly as a reference for scaffolding. PAV weight was analyzed using the average dispensability at megabase level. A 5-megabase window resolution map of PAV-enriched chromosomal regions is represented in Figure 3A. The obtained levels of dispensability were dependent on the proportion of dispensable genes with respect to the total number of coding genes found in each megabase of genomic DNA (Figure 3B), as well as on the average dispensability of those genes (i.e., the number of analyzed resequenced genomes in which they were absent) (Figure 3C). The chromosomal distribution of C1q genes, a gene family previously reported to be strongly subjected to PAV,[21] is represented in Figures 3D and 3E, displaying the dispensability rate of dispensable genes and the distribution of core genes, respectively. The strongest dispensability signals for C1q genes corresponded to chromosomal regions with high PAV weight, as observed in chromosomes 3, 7, and 13 among others.

Based on the megabase-scale dispensability levels, chromosomes 7, 12, and 13 were significantly enriched in dispensable genes with respect to the overall PAV weight of the whole genome, according to an ANOVA Welch's F-test. On the other hand, chromosome 11 presented a significantly lower weight of PAV, being enriched in the core fraction (Figure 3F). Additional tests allowed the identification of 5-megabase windows of genomic sequence showing significantly different PAV weight than chromosomal average in chromosomes 2 and 9 (Figure 3A), revealing the presence of PAV hotspots outside the three aforementioned PAV-enriched chromosomes.

Gene expression associated with core and dispensable genes

The expressions of the core and dispensable gene sets in the mussel genome were studied by analyzing all high-quality transcriptomic datasets available for *M. galloprovincialis*. Figures 4A and 4B represent the fraction of samples (out of the 235 that passed all quality filters) where any given or dispensable genes were expressed, respectively. This analysis revealed clear differences between the two gene sets. While most of the core genes were expressed in almost all (90–100%) transcriptomic samples (Figure 4A), the opposite was observed for the dispensable set (Figure 4B). In fact, many dispensable genes were expressed in a low fraction of samples (0–10%), and only a very few of them showed broad expressions. The impact of gene PAV on expression was further investigated by separately analyzing groups of dispensable genes characterized by different levels of dispensability. Figure 4C shows the relationship between dispensability and the detection/nondetection of expression, highlighting the marked differences between "soft core" genes, which were generally expressed in a relatively broad panel of

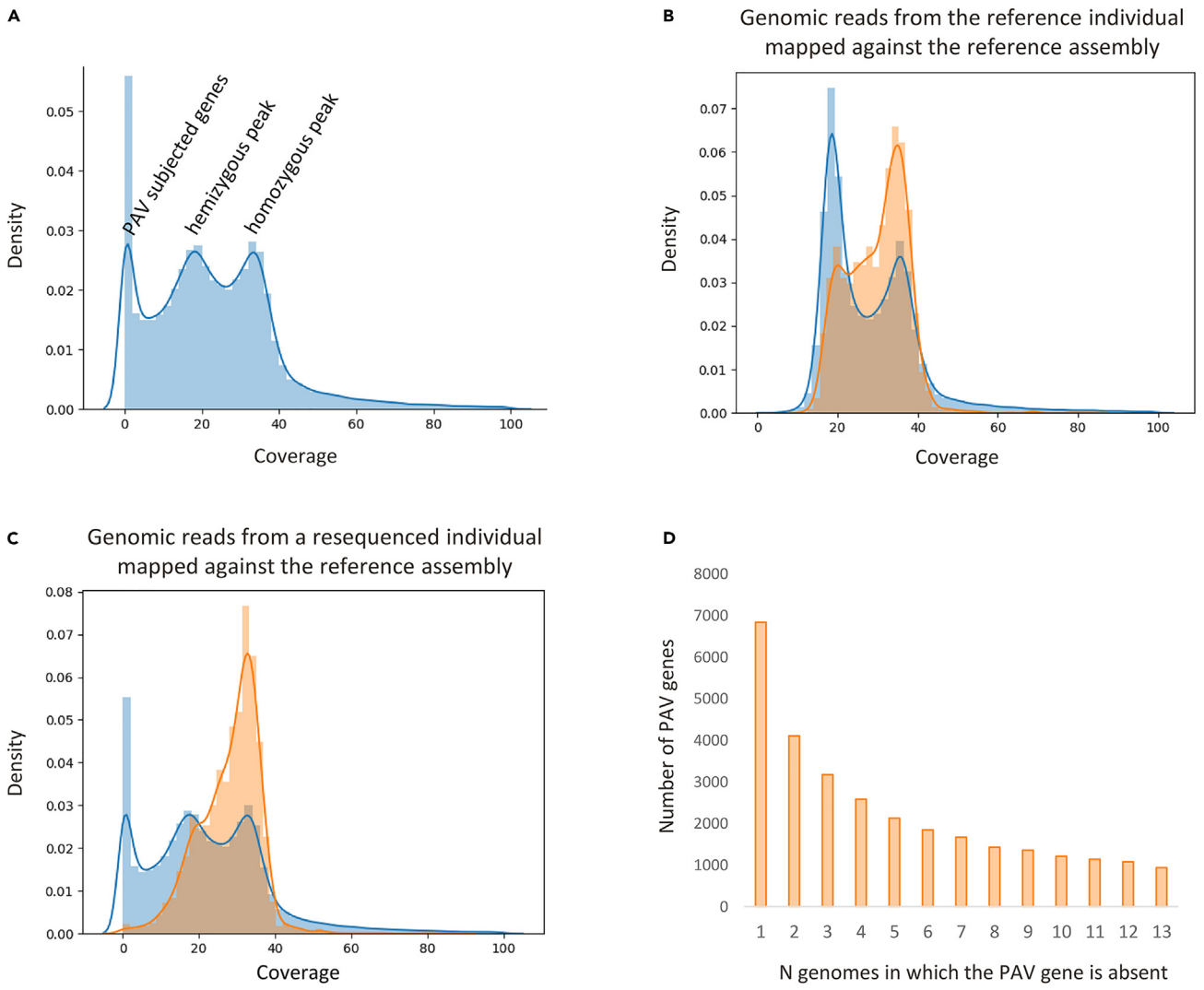


Figure 2. Gene PAV detection

(A) Mapping example of genomic reads from a resequenced individual against the reference genome. Three peaks can be observed based on different coverage values: homozygous genes (presenting the expected sequencing coverage), hemizygous genes (half coverage) and genes without coverage, that is, absent in the resequenced individual (PAV subjected genes).

(B) Mapping of genomic reads from the reference genome individual against the reference assembly. The lack of the zero coverage peak can be observed. Blue curves represent coverage for all genes and orange curves represent coverage for the BUSCO complete genes only.

(C) Mapping of genomic reads from a resequenced individual against the reference assembly. Blue curves represent coverage for all genes and orange curves represent coverage for the BUSCO complete genes only.

(D) Distribution of absent genes retrieved from all resequenced individuals. Genes were considered dispensable if they were missing in at least one individual.

samples, and "cloud" genes, which were mostly expressed in less than 10% of all samples. This direct relation between genomic PAV and the absence of transcriptomic detection reaffirmed the results obtained in the genomic analysis in a much larger set of individuals, confirming a non-casual trace of dispensability, most likely linked with the likelihood of a given dispensable gene to be present in the gene repertoire of each individual subjected to RNA-sequencing. Similar analyses were also performed for specific gene families, either enriched in core or dispensable genes[21] (Figure S1).

The degree of expression variability (calculated as the standard deviation from the average expression of each gene in the whole transcriptomic dataset) was calculated for the core genes and for dispensable genes, subdivided into different groups based on their level of dispensability (Figure 4D). Core genes were subdivided into BUSCO-conserved core genes and non-BUSCO core genes, revealing that housekeeping genes were less variable. However, regardless of the fact of being housekeeping or not, core genes showed the lowest amount of variation among samples, whereas expression variability increased progressively and significantly with the degree of dispensability. The same trend was observed by separately analyzing the transcriptomic data available for different tissues (Figure 4E). The linear

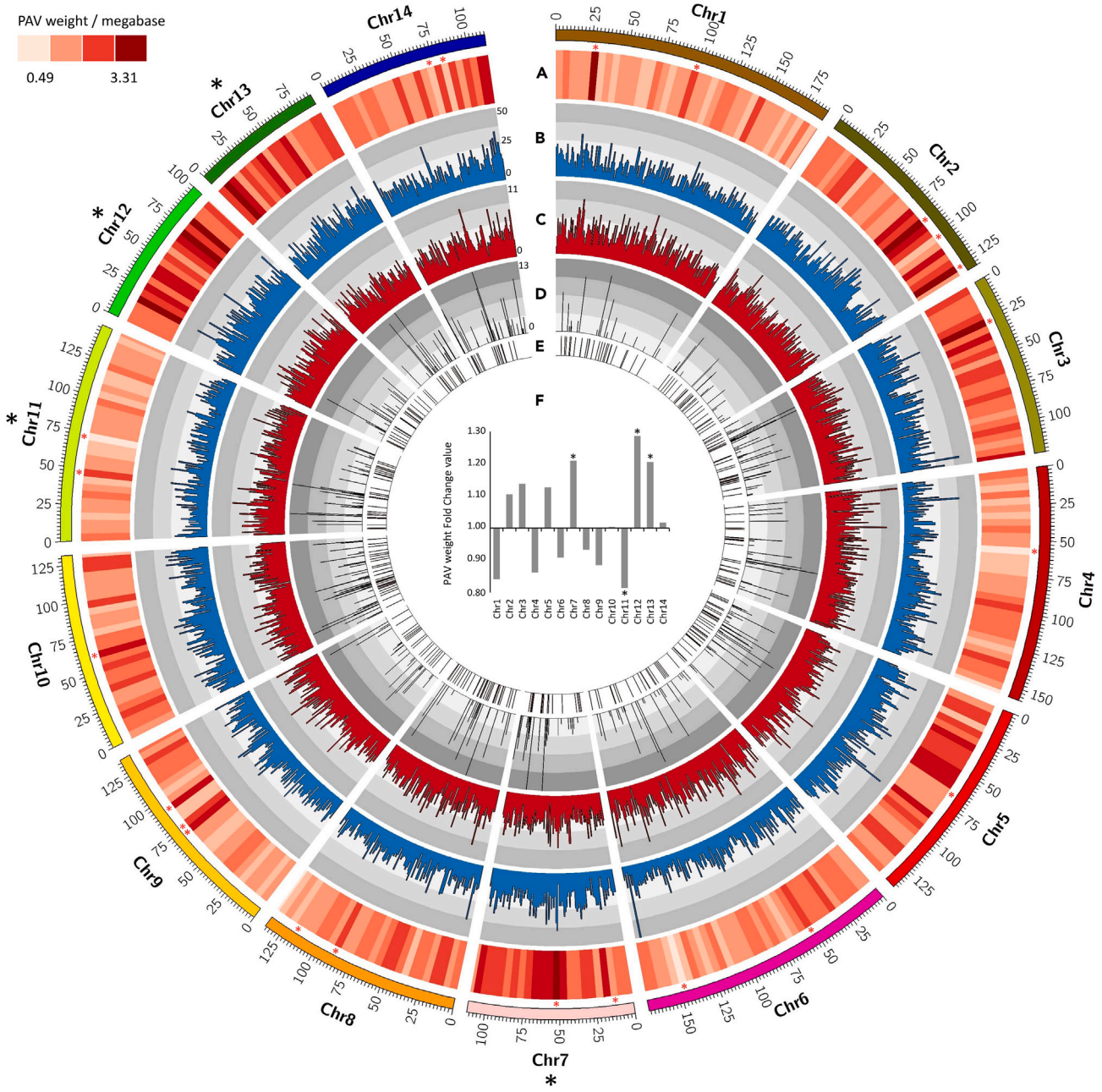


Figure 3. Distribution of the presence/absence variation (PAV) across the *Mytilus galloprovincialis* chromosomes

(A) Heatmap showing the different weights of the PAV phenomenon across chromosomal regions of 5 megabases.

(B) Proportion of dispensable genes with respect to the total number of coding genes in each megabase of genomic sequence (ranging 1–48%).

(C) Average dispensability (the number of individuals in which any given gene was missing among all analyzed resequenced genomes), calculated for the dispensable genes located in each megabase of genomic sequence (ranging 1–11).

(D) Distribution of dispensable C1q genes across the genome. The height of the histogram corresponds to the dispensability of each gene (ranging 1–13).

(E) Distribution of C1q core genes.

(F) Histogram showing fold change values calculated for the average dispensability of each chromosome with respect to the whole genome. Statistically significant differences, calculated with ANOVA-Welch F test using the dispensability data in the 5 megabase windows, are shown with black asterisks. The same black asterisks are shown outside the Circos plot, along the name of the PAV-enriched chromosomes. Additionally, for each chromosome, red asterisks indicate 5-megabase windows with PAV weight significantly different from chromosomal average.

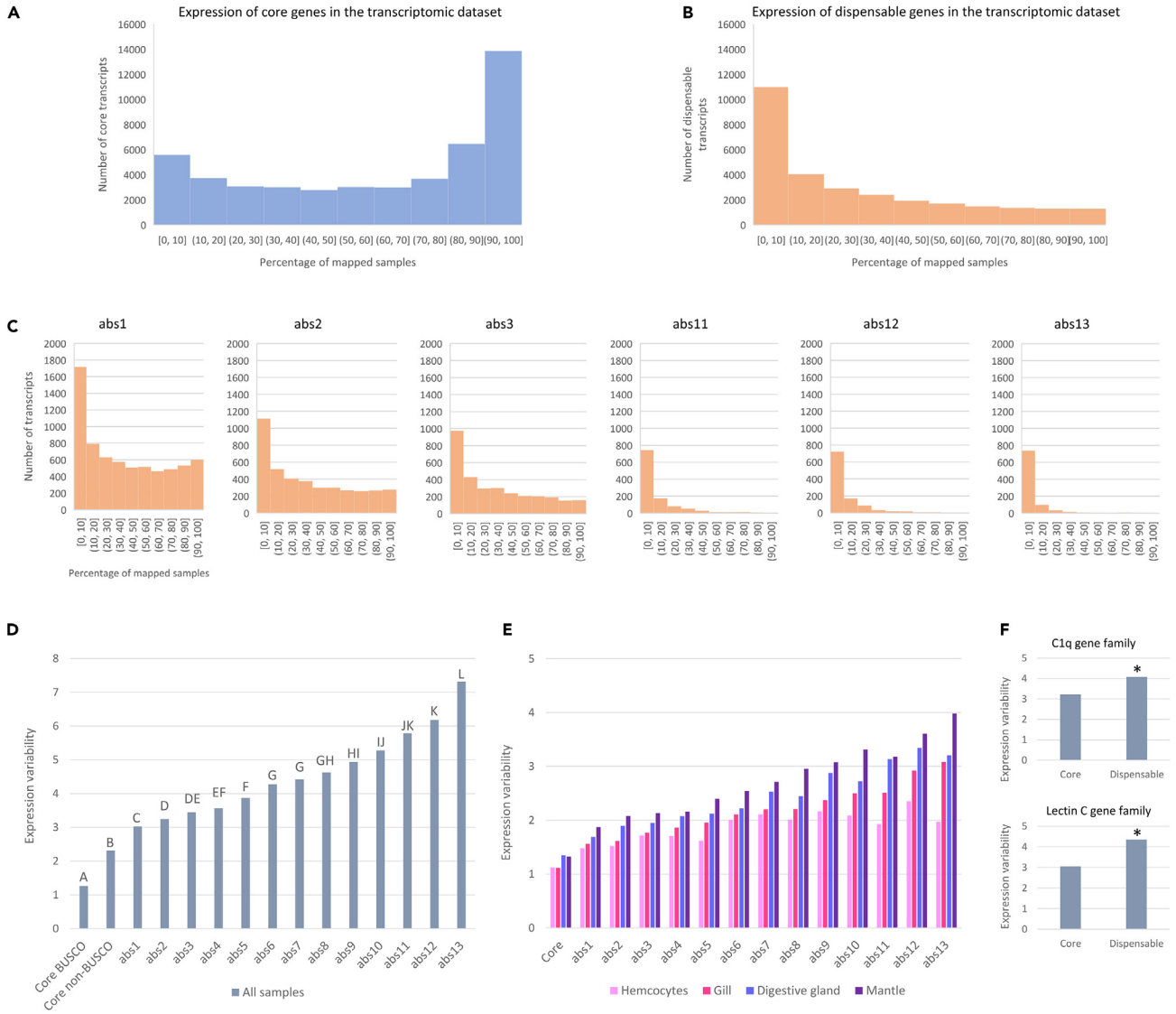


Figure 4. Link between gene expression and gene PAV

(A) Histograms that represent the number of core genes (Y axis) expressed in a given fraction of the 235 analyzed samples in the transcriptomic dataset (X axis).

(B) The same result is shown for the dispensable genes.

(C) A breakdown of the data shown in panel B, based on the dispensability of each gene (each category is labeled absN, where N indicates the number of individuals where a given gene was absent). The graph represents the top 3 groups of genes characterized by the lowest (abs1–abs3) and highest (abs11–abs13) dispensability.

(D) Expression variability (calculated as the standard deviation from the average expression of each transcript in the whole dataset), calculated for all the genes sharing the same level of dispensability. Only genes with an expression value of a minimum of 10 TPM in at least one sample were used. Statistically significant differences calculated with an ANOVA-Welch F test are shown above the bars by using different letters. Core genes were divided between BUSCO-conserved and non-conserved genes in order to reveal the behavior of housekeeping genes.

(E) Expression variability, calculated separately for each of the four main tissues present in the transcriptomic dataset. For all tissues, every fraction of dispensable genes was significantly different from the core fraction.

(F) Analysis of expression variability in specific gene families. Significant differences between dispensable and core genes are indicated with an asterisk (see also Figure S1).

relationship between dispensability and variability of expression was evident in gill, digestive gland, and mantle samples, but it was less clear in hemocytes. However, for the four sample types, dispensable genes presented a significant increase in expression variability with respect to the core genes. Additionally, to overcome possible differences caused by the inclusion of different genes, the core and dispensable genes belonging to the same gene families were analyzed, supporting the greater variability of expression in dispensable genes (Figure 4F).

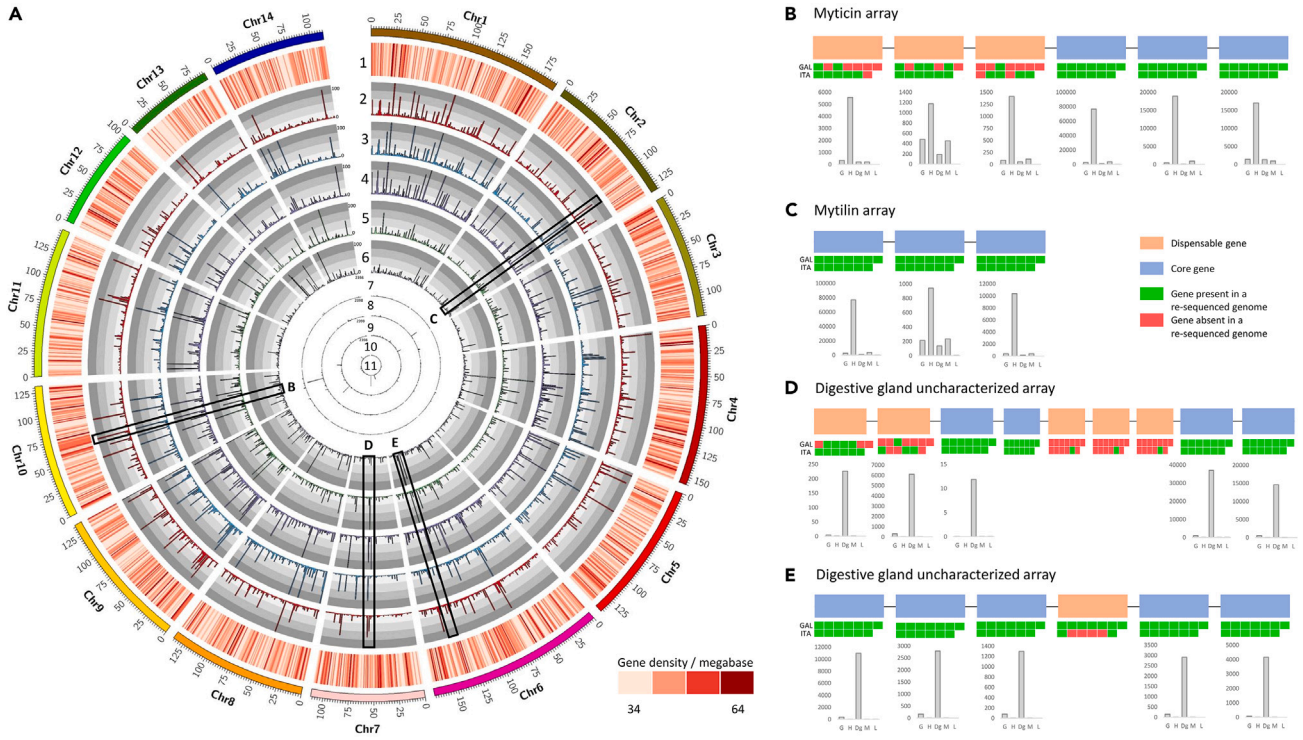


Figure 5. Most highly expressed regions in the mussel genome

(A) The average expression of all *Mytilus galloprovincialis* SRA transcriptomic samples was calculated for different mussel samples. (1) Gene density per megabase in the mussel genome. (2–6) Average expression per megabase in the mussel genome, capped at 100 TPMs to prevent the regions displaying extremely high expression levels from masking other highly expressed regions. Expression values are shown for gills (2), hemocytes (3), mantle (4), digestive gland (5), and larvae (6). Uncapped expression levels are also shown for gills (7), hemocytes (8), mantle (9), and digestive gland (10) (see also Figure S2; Table S4). (B–E) The most highly expressed regions are shown, using average gene-level TPM expression levels for each sample type (G: gills, H: hemocytes, Dg: digestive gland, M: mantle, L: larvae) along with dispensability data.

Chromosomal gene expression: Genome hotspots and tissue-specific regions

The average expression per megabase was calculated for all *M. galloprovincialis* chromosomes, providing a detailed map of the genomic regions with the higher expression levels in different mussel tissues (Figure 5A). This approach allowed the identification of both regions displaying high generalized expression in all tissues and regions displaying strong tissue specificity (and thereby most likely under the control of potent transcription factors). The genomic regions showing the highest expression levels included arrays of genes encoding antimicrobial peptides, which were strongly expressed in hemocytes, and arrays of uncharacterized genes highly expressed in the digestive gland (Figures 5B–5E). Although the strong expression of these gene arrays was mostly due to the contribution of core genes, associated dispensable genes were highly expressed as well, mainly in the myticin array.

Among the many chromosomal regions displaying strong tissue specificity (Figure S2), an interesting discovery was represented by an array of genes encoding several endonucleases, which were specifically expressed in the digestive gland. As shown in Table S4, most of these tissue-specific arrays contained intermixed dispensable and core genes, consistent with the observation that mussel dispensable genes are frequently associated with recently expanded gene families, through processes that likely involved tandem gene duplication.[31] Uncharacterized genes with strong expression were not uncommon, evidencing the large knowledge gaps that persist regarding the functional role of lineage-specific gene families, which may carry out fundamental functions for the physiology of an organism despite the lack of detectable homology.[40]

Implications of dispensable genes in adaptation

The list of absent genes associated with each resequenced individual was analyzed with the aim of investigating the possible presence of patterns linked with adaptation, with particular reference to the geographical origin of the samples. The repertoire of dispensable genes that were absent in each resequenced individual was used to perform a principal component analysis (PCA), which allowed us to obtain a clear grouping of individuals based on their geographical origin (i.e., Galicia or Italy) (Figure 6A). This indicated that the repertoires of dispensable genes shared by mussels belonging to the same population showed higher overlaps than those shared by mussels sampled in different geographical locations. This prompted us to investigate the presence of dispensable genes exclusively associated with either one or the other

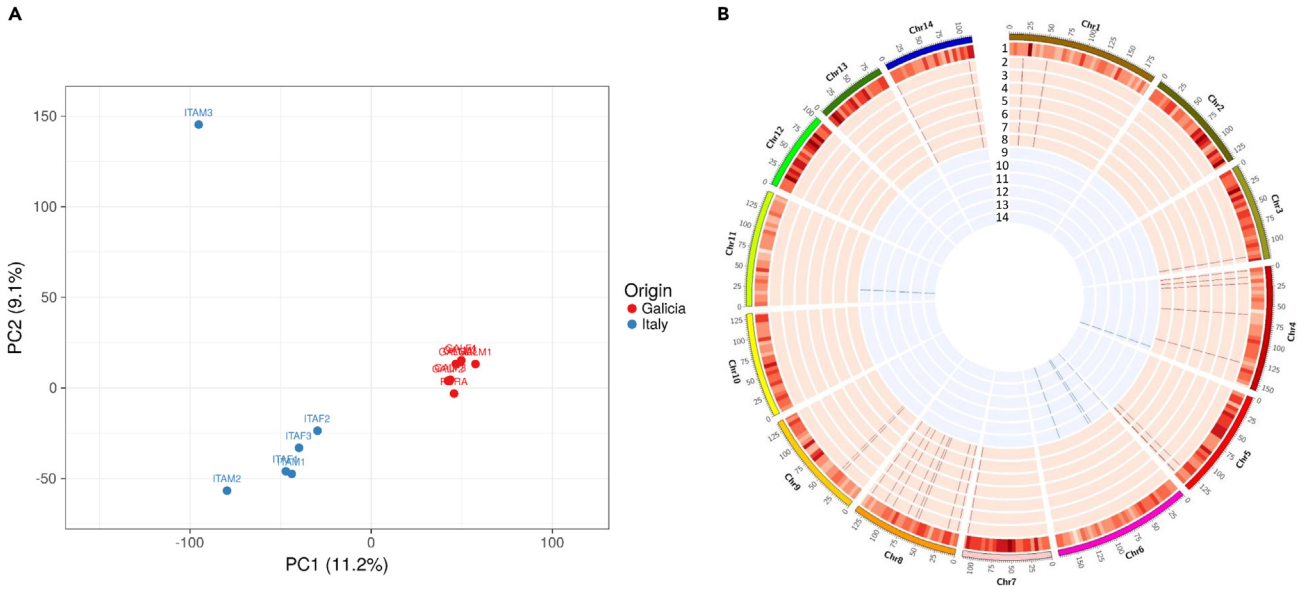


Figure 6. Geographical differences associated with the dispensable gene repertoire

(A) PCA performed with the repertoire of dispensable genes that were absent in each resequenced individual. The clustering of individuals belonging to the same geographical location highlights the larger overlap between the dispensable gene repertoire of individuals from the same region.

(B) Chromosomal distribution of the dispensable genes showing "extreme" geographical distribution (dispensable genes that were present in all individuals from one region and absent in all individuals from the other region). (1) Heatmap representing PAV weight in 5-megabase windows in the reference genome; (2–8) Genomic location in the reference assembly of dispensable genes that were present only in the resequenced individuals from Galicia; (9–14) Genomic location in the reference assembly of dispensable genes that were present only in the resequenced individuals from Italy (see also Table S5).

population, even though all analyzed genes were, by definition, present in the reference individual (sampled in Galicia). The dispensable genes displaying the most skewed distributions included members from important immune-related families as well as calmodulins, low-density lipoproteins or cytochrome-related genes (Table S5). Figure 6B shows the chromosomal distribution of the dispensable genes characterized by the most "extreme" distribution, present in all individuals from one population and absent in all individuals from the other one. In detail, chromosome 8 contained several dispensable genes only found in Galician mussels, and chromosome 6 included several genes only found in Italian mussels (except Lola, the reference individual). Most of these genes were uncharacterized, and would be therefore good targets for future functional characterization efforts.

DISCUSSION

Despite the use of a different, more redundant assembly version (mg3 instead of mg10) that provided a more comprehensive view of the hemizygous regions of the mussel genome, the analysis of gene PAV in this study identified the same fraction of dispensable genes (38%) as that previously reported in the reference genome.[21] As shown in Figure 2D, most dispensable genes were "soft core", being absent in a few individuals of the dataset (for example, 23% of the dispensable genes were absent in just one resequenced individual). These observations highlighted the highly dynamic nature of the mussel pangenome, pointing out the need to include a very high number of resequenced individuals, possibly belonging to multiple populations, to delineate the core and dispensable fractions of the mussel genome in a reliable manner.

The rates of hemizygosity and homozygosity are key indicators of gene dispensability, allowing estimation of the impact of gene PAV in a species and the "openness" of its pangenome. The gene coverage graphs obtained from the reference individual and from resequenced individuals belonging to the same species display major differences. Such graphs display only two peaks in the reference genome, since all genes are either found in one or both homologous chromosomes and are presented either in a hemizygous or homozygous state. On the other hand, a fraction of the genes associated with hemizygous regions in the reference genome might be entirely missing in the resequenced individuals, leading to a third peak that corresponds to genes lacking detectable coverage. This is the direct consequence of the Mendelian mode of inheritance of such genes, which, depending on the combination between the two parental haplotypes, could lead to nullizygosity in some offspring. At the population level, this would lead to the observation of varying individuals either showing the presence or absence of a given gene, which would therefore be defined as dispensable and associated with PAV. Although we have previously described this phenomenon in *M. galloprovincialis*,[21] the same hemizygosity patterns have been recently observed in other mussel species,[41,42] as well as in other bivalves,[23] suggesting that gene PAV could be considerably more common than originally thought in these organisms. The high dispensable:core gene ratio found in mussels points toward complex open pangenomes, with several thousand dispensable genes that could provide beneficial accessory functions.[21,23]

In this study, we investigated the biological implications of gene PAV in *M. galloprovincialis* in greater detail, exploiting a new chromosome-scale assembly obtained by scaffolding in the genome of *Mytilus chilensis*.[43] Since the two species are congeneric and closely related,[44] this type of approach was appropriate for filling in the gaps that exist in the Mediterranean mussel fragmented reference genome, allowing us to reorder the scaffolds based on their inferred chromosomal location, as previously done for other assemblies within the same species complex.[45]

This strategy, combined with the mapping of dispensable genes, allowed us to investigate the distribution of hemizygous genomic regions, highlighting similarities but also some discrepancies with data previously reported in other molluskan species. In detail, hemizygous regions have been described to be broadly and nearly uniformly distributed along the chromosomes of other mollusks, which nevertheless did not display a level of hemizygosity as high as in mussels.[23] The impact of hemizygosity on gene PAV in other mollusk species was significantly inferior to that observed in mussels, with the highest interindividual variations in gene content (11–16%) being reported in the clam *Mercenaria mercenaria*.[24] Although the distribution of dispensable mussel genes was generally consistent with this view, we highlighted the presence of chromosomal hotspots of dispensability. This observation is consistent with the data recently reported in *Pinctada fucata*, where a large megabase-scale hemizygous genomic region was specifically located in a single chromosomal scaffold, where most reported SVs of this species were found, perhaps as a result of high transposon activity.[25] The association of this region, present in only one of the two haplotypes of the sequenced individual, with protein-coding genes would in turn imply the possibility that the same region is subject to PAV at the population level.[25] Similar to the case of *Mytilus galloprovincialis*,[21] the major hemizygous genomic region of *P. fucata* was enriched in immune-related genes, such as those encoding immunoglobulin-like domains. Overall, the chromosomal distribution of hemizygous regions in mussels was characterized by a broad distribution of dispensable genes and the presence of some hotspots, similar to those described in other bivalves, such as *Crassostrea virginica*.[26]

After obtaining a detailed cartography of the dispensable gene locations across the chromosomes of *M. galloprovincialis*, we tried to elucidate their possible functional implications. Mussel dispensable genes are enriched in young immune-related gene families that have been subjected to large lineage-specific expansions,[21] marking a feature shared by nearly all bivalves, but uncommon in other mollusks.[46] Strikingly, such large expanded immune gene families are also frequently associated with hemizygous genomic regions in several bivalves.[23] If we assume that the molecular diversification of immune receptors and effectors has important functional implications in the arms race between the host and the pathogen, then the strong association between gene PAV and the recent expansion of these immune gene families might provide a significant evolutionary benefit to bivalves in their natural habitat. Indeed, gene duplications can lead to the neofunctionalization of new gene copies,[47] whose signatures are particularly evident in several bivalve immune gene families.[31,34] The high complexity of these diversified immune repertoires is further amplified by an additional layer of interindividual diversity, provided by gene PAV, which endows each individual with a unique molecular arsenal that might prevent a pathogenic or environmental challenge from affecting all individuals in high-density mussel beds. It is becoming increasingly clear that these genomic factors also have profound effects at the transcriptomic level, as highlighted by the remarkable differences in the individual response displayed by mussels to the same stimulus.[48]

Even though the dispensable gene repertoire of mussels from the same population differed between individuals, the results of this study indicated that within-population interindividual differences in PAV patterns were significantly lower than those observed between populations. While this would be consistent with the shared ancestry of the mussels found in the same geographical location, i.e., Galicia and the Adriatic Sea, and in line with the proposed mode of Mendelian inheritance of dispensable genes, these results may be linked with local adaptation to the different environmental conditions found in the two sampling areas. Curiously, a recent study carried out in oysters that focused on copy number variations within expanded gene families, did not report clear differences between sampling locations, revealing that inbred individuals had a lower number of copy variants.[26] Although the results of our comparative study need to be validated for additional populations and with analyses of a greater number of individuals, the correlation between PAV patterns and the geographical origin of samples offers some interesting cues about the potential usefulness of dispensable genes as molecular markers to be used in conjunction with classical SNP-based genotyping approaches. Additionally, dispensable genes exclusively present in a specific population were identified; most of them being uncharacterized and which would constitute interesting targets for future characterization.

The expression data analyzed in this study strongly reinforce the idea that gene PAV in mussels is not a phenomenon merely linked with genome architecture but also has important implications on the transcriptional landscape and, consequently, on phenotype. Indeed, mussels expressing a given set of dispensable genes providing advantageous accessory functions may display enhanced fitness in certain habitats, being able to adapt to certain environmental conditions that would not have been tolerated by individuals expressing only the core part of the genome.[21] This interpretation is fully consistent with the data recently reported in *Mytilus chilensis*, where different chromosomal clusters of genes were specifically expressed in different populations adapted to largely different environmental conditions.[43]

The dispensable fraction of the mussel genome emerges as the "main culprit" in explaining the important interindividual and interpopulation differences in gene expression that have been frequently described in different settings in the mussel scientific literature over the past few decades. In fact, the lack of expression from certain genes, which has always been traditionally interpreted as evidence of tight transcriptional control, is most often explained by the lack of the gene itself in a given individual or population rather than by simple transcriptional repression. This interpretation was strongly supported by the linear correlation we observed between the levels of dispensability and expression variability, which could be consistently replicated across a broad range of tissues. The presence or absence of expression has been previously used as a proxy to estimate gene PAV in other species.[14] Although we found a strong relationship between transcriptomic and genomic PAV results, such inferences would always need genomic verification, as in our study.

In addition to having profound biological significance, our findings also have important practical implications. We demonstrated that the dispensable fraction of the mussel pangenome is a strong source of noise in transcriptomic experiments due to the inherent variability of the

expression of genes with PAV across individuals. Such noise might reach a point where the differential expression of core genes that are responsive to a given stimulus is entirely masked by spurious data linked with apparent gene alterations of gene expression that are not due to a modulation of expression but due to PAV. Therefore, limiting gene expression analyses to the core genome fraction only would provide, in principle, more reliable and reproducible results than analyses carried out using the full gene repertoire as a reference. On the other hand, our results also showed that dispensable genes are of major importance in this species, often displaying strong tissue specificity and being expressed at high levels in the individuals who carry them. Hence, disregarding dispensable genes in transcriptomic analyses might be detrimental for the interpretation of results, as this would prevent the detection of several biologically meaningful signals. Altogether, our data indicated that the most appropriate approach to follow while analyzing gene expression data of species characterized by open pangenomes would probably lie in the construction of *de novo* assemblies using the individual samples from each experiment. Although the number of high-quality genomic resources available for mussels is quickly growing,[21,41–43,49] the extent to which gene PAV affects individual gene repertoires highlights the importance of selecting the most appropriate transcriptomic reference to obtain a reliable biological interpretation. Although the impact of the PAV gene in other bivalve species remains to be fully understood, preliminary data seem to point in the same direction.[25,26,35]

The genomic regions associated with the highest transcriptional activity were identified in different tissues, highlighting the strong expression of some dispensable genes, such as myticins. This gene family provides an interesting example for the importance of PAV in this species, since myticin genes are among the most highly expressed in hemocytes, cumulatively accounting for over 100,000 average TPMs, i.e., more than 10% of the total transcriptional effort of this tissue. Myticins are implicated in different immune functions, have antiviral and antimicrobial properties but also display cytokine-like activity, and they are affected by PAV to the point that each individual mussel possesses its own unique repertoire of myticin variants.[33,50,51] Additional uncharacterized gene clusters of similar importance for other tissues have been identified in the current study and could be the target of functional studies in the near future.

Overall, this study provides additional evidence supporting the relationships between PAV, the local adaptation of mussel populations to different marine coastal environments and the high invasiveness of this species, as indicated by the differences between the dispensable repertoires shown by different populations and their enrichment in immune and stress-response functions. The great relevance of PAV for the interpretation of gene expression data was also confirmed, evidencing the importance of considering the dispensable gene repertoire of each individual to discern transcriptional noise from biologically meaningful signals. In summary, the success of mussels in colonizing different environments and their high resilience to biotic and abiotic stress are supported by a complex open pangenome with large lineage-specific gene family expansions and widespread PAV, whose effect on gene expression profiles can explain the great adaptability of these organisms to their challenging and changing habitats.

Limitations of the study

This study represents a first incursion into the PAV phenomenon to try to explain the biological and immune reality of the mussel. A relevant aspect to be achieved in the future would be to improve this analysis by increasing the number of individuals as well as their geographical origins. The greater the number of sequenced genomes, the better the conclusions that can be drawn. Fully sequenced haplotypes and PAV studies in other bivalve species would be of great interest as well.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Mg3 chromosome-level assembly
  - Presence/absence variation and dispensability analyses
  - Gene expression data
  - Data visualization
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107827.

ACKNOWLEDGMENTS

This research was funded by Ministerio de Ciencia e Innovación (PID2021-124955OB-I00), Agencia Galega de Innovación (IN607B 2022/13), Agencia Estatal de Investigación, and European Research Funds AEI/EU-FSE (PRE2019-090760).

AUTHOR CONTRIBUTIONS

Conceptualization: A.F., B.N., C.G.E., and M.G. Methodology: M.G., A.S., M.R.-C., and C.G.E. Investigation: A.S. and M.R.-C. Visualization: A.S. and M.R.-C. Supervision: A.F. and BN. Funding acquisition: A.F. and B.N. Writing – Original draft: A.S. Writing – Review & Editing: A.S., B.N., A.F., M.R.-C., C.G.E., and M.G.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: May 16, 2023
Revised: July 12, 2023
Accepted: September 1, 2023
Published: September 4, 2023

REFERENCES

1. Feuk, L., Marshall, C.R., Wintle, R.F., and Scherer, S.W. (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. Hum. Mol. Genet. *15*, R57–R66. https://doi.org/10.1093/hmg/ddl057.
2. Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. *7*, 85–97. https://doi.org/10.1038/nrg1767.
3. Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. Curr. Opin. Microbiol. *23*, 148–154. https://doi.org/10.1016/j.mib.2014.11.016.
4. Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589–594. https://doi.org/10.1016/j.gde.2005.09.006.
5. McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have pangenomes. Nat. Microbiol. *2*, 17040. https://doi.org/10.1038/nmicrobiol.2017.40.
6. Golicz, A.A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. Plant Biotechnol. J. *14*, 1099–1105. https://doi.org/10.1111/pbi.12499.
7. Hübner, S., Bercovich, N., Todesco, M., Mandel, J.R., Odenheimer, J., Ziegler, E., Lee, J.S., Baute, G.J., Owens, G.L., Grassa, C.J., et al. (2019). Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat. Plants *5*, 54–62. https://doi.org/10.1038/s41477-018-0329-0.
8. Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature *499*, 209–213. https://doi.org/10.1038/nature12221.
9. McCarthy, C.G.P., and Fitzpatrick, D.A. (2019). Pan-genome analyses of model fungal species. Microb. Genom. *5*, e000243. https://doi.org/10.1099/mgen.0.000243.
10. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81. https://doi.org/10.1038/nature15394.
11. Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet. *51*, 30–35. https://doi.org/10.1038/s41588-018-0273-y.
12. Tian, X., Li, R., Fu, W., Li, Y., Wang, X., Li, M., Du, D., Tang, Q., Cai, Y., Long, Y., et al. (2020). Building a sequence map of the pig pan-genome from multiple *de novo* assemblies and Hi-C data. Sci. China Life Sci. *63*, 750–763. https://doi.org/10.1007/s11427-019-9551-7.
13. Secomandi, S., Gallo, G.R., Sozzoni, M., Iannucci, A., Galati, E., Abueg, L., Balacco, J., Caprioli, M., Chow, W., Ciofi, C., et al. (2023). A chromosome-level reference genome and pangenome for barn swallow population genomics. Cell Rep. *42*, 111992. https://doi.org/10.1016/j.celrep.2023.111992.
14. Huang, Y.-S., Lin, C.-Y., and Cheng, W.-C. (2021). Investigating the Transcriptomic and Expression Presence-Absence Variation Exist in Japanese Eel (Anguilla japonica), a Primitive Teleost. Mar. Biotechnol. *23*, 943–954. https://doi.org/10.1007/s10126-021-10077-w.
15. Berg, P.R., Star, B., Pampoulie, C., Sodeland, M., Barth, J.M.I., Knutsen, H., Jakobsen, K.S., and Jentoft, S. (2016). Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. Sci. Rep. *6*, 23246. https://doi.org/10.1038/srep23246.
16. Sodeland, M., Jorde, P.E., Lien, S., Jentoft, S., Berg, P.R., Grove, H., Kent, M.P., Arnyasi, M., Olsen, E.M., and Knutsen, H. (2016). "Islands of Divergence" in the Atlantic Cod Genome Represent Polymorphic Chromosomal Rearrangements. Genome Biol. Evol. *8*, 1012–1022. https://doi.org/10.1093/gbe/evw057.
17. Small, K.S., Brudno, M., Hill, M.M., and Sidow, A. (2007). Extreme genomic variation in a natural population. Proc. Natl. Acad. Sci. *104*, 5698–5703. https://doi.org/10.1073/pnas.0700890104.
18. Hollenbeck, C.M., Portnoy, D.S., Garcia de la serrana, D., Magnesen, T., Matejusova, I., and Johnston, I.A. (2022). Temperature-associated selection linked to putative chromosomal inversions in king scallop (Pecten maximus). Proc. Biol. Sci. *289*, 20221573. https://doi.org/10.1098/rspb.2022.1573.
19. Crombie, T.A., Zdraljevic, S., Cook, D.E., Tanny, R.E., Brady, S.C., Wang, Y., Evans, K.S., Hahnel, S., Lee, D., Rodriguez, B.C., et al. (2019). Deep sampling of Hawaiian Caenorhabditis elegans reveals high genetic diversity and admixture with global populations. Elife *8*, e50465. https://doi.org/10.7554/eLife.50465.
20. Dey, A., Chan, C.K.W., Thomas, C.G., and Cutter, A.D. (2013). Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri. Proc. Natl. Acad. Sci. *110*, 11056–11060. https://doi.org/10.1073/pnas.1303057110.
21. Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-Ortiz, M.A., Murgarella, M., Greco, S., et al. (2020). Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol. *21*, 275. https://doi.org/10.1186/s13059-020-02180-3.
22. Tong, X., Han, M.-J., Lu, K., Tai, S., Liang, S., Liu, Y., Hu, H., Shen, J., Long, A., Zhan, C., et al. (2022). High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. Nat. Commun. *13*, 5619. https://doi.org/10.1038/s41467-022-33366-x.
23. Calcino, A.D., Kenny, N.J., and Gerdol, M. (2021). Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philos. Trans. R. Soc. Lond. B Biol. Sci. *376*, 20200153. https://doi.org/10.1098/rstb.2020.0153.
24. Farhat, S., Bonnivard, E., Pales Espinosa, E., Tanguy, A., Boutet, I., Guiglielmoni, N., Flot, J.-F., and Allam, B. (2022). Comparative analysis of the Mercenaria mercenaria genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks. BMC Genom. *23*, 192. https://doi.org/10.1186/s12864-021-08262-1.
25. Takeuchi, T., Suzuki, Y., Watabe, S., Nagai, K., Masaoka, T., Fujie, M., Kawamitsu, M., Satoh, N., and Myers, E.W. (2022). A high-quality, haplotype-phased genome reconstruction reveals unexpected haplotype diversity in a pearl oyster. DNA Res. *29*, dsac035. https://doi.org/10.1093/dnares/dsac035.
26. Modak, T.H., Literman, R., Puritz, J.B., Johnson, K.M., Roberts, E.M., Proestou, D., Guo, X., Gomez-Chiarri, M., and Schwartz, R.S. (2021). Extensive genome-wide duplications in the eastern oyster (Crassostrea virginica). Philos. Trans. R. Soc. Lond. B Biol. Sci. *376*, 20200164. https://doi.org/10.1098/rstb.2020.0164.
27. Gosling, E. (2015). In Marine bivalve molluscs, 2nd ed., N. Hoboken, ed. (John Wiley & Sons).

28. Suttle, C.A. (2007). Marine viruses — major players in the global ecosystem. Nat. Rev. Microbiol. *5*, 801–812. https://doi.org/10.1038/nrmicro1750.
29. Azam, F., and Malfatti, F. (2007). Microbial structuring of marine ecosystems. Nat. Rev. Microbiol. *5*, 782–791. https://doi.org/10.1038/nrmicro1747.
30. Zannella, C., Mosca, F., Mariani, F., Franci, G., Folliero, V., Galdiero, M., Tiscar, P.G., and Galdiero, M. (2017). Microbial Diseases of Bivalve Mollusks: Infections, Immunology and Antimicrobial Defense. Mar. Drugs *15*, 182. https://doi.org/10.3390/md15060182.
31. Saco, A., Rey-Campos, M., Rosani, U., Novoa, B., and Figueras, A. (2021). The Evolution and Diversity of Interleukin-17 Highlight an Expansion in Marine Invertebrates and Its Conserved Role in Mucosal Immunity. Front. Immunol. *12*, 692997. https://doi.org/10.3389/fimmu.2021.692997.
32. Gerdol, M., Manfrin, C., De Moro, G., Figueras, A., Novoa, B., Venier, P., and Pallavicini, A. (2011). The C1q domain containing proteins of the Mediterranean mussel Mytilus galloprovincialis: A widespread and diverse family of immune-related molecules. Dev. Comp. Immunol. *35*, 635–643. https://doi.org/10.1016/j.dci.2011.01.018.
33. Rey-Campos, M., Novoa, B., Pallavicini, A., Gerdol, M., and Figueras, A. (2020). Comparative genomics reveals a significant sequence variability of myticin genes in mytilus galloprovincialis. Biomolecules *10*, 943. https://doi.org/10.3390/biom10060943.
34. Gerdol, M., Schmitt, P., Venier, P., Rocha, G., Rosa, R.D., and Destoumieux-Garzón, D. (2020). Functional Insights From the Evolutionary Diversification of Big Defensins. Front. Immunol. *11*, 758.
35. Rosa, R.D., Alonso, P., Santini, A., Vergnes, A., and Bachère, E. (2015). High polymorphism in big defensin gene expression reveals presence–absence gene variability (PAV) in the oyster Crassostrea gigas. Dev. Comp. Immunol. *49*, 231–238. https://doi.org/10.1016/j.dci.2014.12.002.
36. Sollitto, M., Kenny, N.J., Greco, S., Tucci, C.F., Calcino, A.D., and Gerdol, M. (2022). In Detecting Structural VariantsStructural variants and Associated Gene Presence–Absence Variation Phenomena in the Genomes of Marine Organisms BT - Marine Genomics: Methods and Protocols, C. Verde and D. Giordano, eds. (Springer US), pp. 53–76. https://doi.org/10.1007/978-1-0716-2313-8_4.
37. Kaas, R.S., Friis, C., Ussery, D.W., and Aarestrup, F.M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genom. *13*, 577. https://doi.org/10.1186/1471-2164-13-577.
38. Acevedo-Rocha, C.G., Fang, G., Schmidt, M., Ussery, D.W., and Danchin, A. (2013). From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet. *29*, 273–279. https://doi.org/10.1016/j.tig.2012.11.001.
39. Contreras-Moreira, B., and Vinuesa, P. (2013). GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. Appl. Environ. Microbiol. *79*, 7696–7701. https://doi.org/10.1128/AEM.02411-13.
40. Johnson, B.R. (2018). Taxonomically Restricted Genes Are Fundamental to Biology and Evolution. Front. Genet. *9*, 407. https://doi.org/10.3389/fgene.2018.00407.
41. Yang, J.-L., Feng, D.-D., Liu, J., Xu, J.-K., Chen, K., Li, Y.-F., Zhu, Y.-T., Liang, X., and Lu, Y. (2021). Chromosome-level genome assembly of the hard-shelled mussel Mytilus coruscus, a widely distributed species from the temperate areas of East Asia. GigaScience *10*, giab024. https://doi.org/10.1093/gigascience/giab024.
42. Paggeot, L.X., DeBiasse, M.B., Escalona, M., Fairbairn, C., Marimuthu, M.P.A., Nguyen, O., Sahasrabudhe, R., and Dawson, M.N. (2022). Reference genome for the California ribbed mussel, Mytilus californianus, an ecosystem engineer. J. Hered. *113*, 681–688. https://doi.org/10.1093/jhered/esac041.
43. Gallardo-Escárate, C., Valenzuela-Muñoz, V., Nuñez-Acuña, G., Valenzuela-Miranda, D., Tapia, F., Yévenes, M., Gajardo, G., Toro, J.E., Oyarzún, P.A., Arriagada, G., et al. (2022). The native mussel Mytilus chilensis genome reveals adaptative molecular signatures facing the marine environment. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.06.506863.
44. Zbawicka, M., Trucco, M.I., and Wenne, R. (2018). Single nucleotide polymorphisms in native South American Atlantic coast populations of smooth shelled mussels: hybridization with invasive European Mytilus galloprovincialis. Genet. Sel. Evol. *50*, 5. https://doi.org/10.1186/s12711-018-0376-z.
45. Simon, A. (2022). Three new genome assemblies of blue mussel lineages: North and South European Mytilus edulis and Mediterranean Mytilus galloprovincialis. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.02.506387.
46. Regan, T., Stevens, L., Peñaloza, C., Houston, R.D., Robledo, D., and Bean, T.P. (2021). Ancestral Physical Stress and Later Immune Gene Family Expansions Shaped Bivalve Mollusc Evolution. Genome Biol. Evol. *13*, evab177. https://doi.org/10.1093/gbe/evab177.
47. Rastogi, S., and Liberles, D.A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. *5*, 28. https://doi.org/10.1186/1471-2148-5-28.
48. Rey-Campos, M., Moreira, R., Valenzuela-Muñoz, V., Gallardo-Escárate, C., Novoa, B., and Figueras, A. (2019). High individual variability in the transcriptomic response of Mediterranean mussels to Vibrio reveals the involvement of myticins in tissue injury. Sci. Rep. *9*, 3569. https://doi.org/10.1038/s41598-019-39870-3.
49. Regan, T., Hori, T.S., and Bean, T.P. (2022). A blue mussel chromosome-scale assembly and genomic resources for aquaculture, marine ecology and evolution. Preprint at bioRxiv. https://doi.org/10.1101/2022.11.17.516937.
50. Balseiro, P., Falcó, A., Romero, A., Dios, S., Martínez-López, A., Figueras, A., Estepa, A., and Novoa, B. (2011). Mytilus galloprovincialis Myticin C: A chemotactic molecule with antiviral activity and immunoregulatory properties. PLoS One *6*, 1–14. https://doi.org/10.1371/journal.pone.0023140.
51. Novoa, B., Romero, A., Álvarez, Á.L., Moreira, R., Pereiro, P., Costa, M.M., Dios, S., Estepa, A., Parra, F., and Figueras, A. (2016). Antiviral activity of myticin C peptide from mussel: an ancient defense against herpesviruses. J. Virol. *90*, 7692–7702. https://doi.org/10.1128/JVI.00591-16.
52. Chu, C., Li, X., and Wu, Y. (2019). GAPPadder: a sensitive approach for closing gaps on draft genomes with short sequence reads. BMC Genom. *20*, 426. https://doi.org/10.1186/s12864-019-5703-4.
53. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
54. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
55. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
56. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
57. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. F1000Research *10*, 33. https://doi.org/10.12688/f1000research.29032.2.
58. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.
59. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645. https://doi.org/10.1101/gr.092759.109.
60. Blighe, K., and Lun, A. (2022). PCAtools: Everything Principal Components Analysis. R package version 2.0.
61. Fox, J. (2005). The R Commander: A Basic-Statistics Graphical User Interface to R. J. Stat. Softw. *14*, 1–42. https://doi.org/10.18637/jss.v014.i09.
62. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Chromosome level assembly of *Mytilus galloprovincialis*	This paper	Figshare: https://doi.org/10.6084/m9.figshare.22759397
Mytilus galloprovincialis Assembly annotation	This paper	Figshare: https://doi.org/10.6084/m9.figshare.22759397
Mussel core and dispensable genes	This paper	Figshare: https://doi.org/10.6084/m9.figshare.22759397
Mussel transcriptomic dataset	This paper	Figshare: https://doi.org/10.6084/m9.figshare.22759397
Mussel RNA-seq data	NCBI - SRA	Accessions numbers listed in Table S6
Genomic data from resequenced mussels	Gerdol et al[21]	NCBI Bioproject: PRJEB25106
Code for PAV analysis	Sollito et al.[36]	https://github.com/Carmen-Tuc/PAV_pipeline
Software and algorithms		
Minimap2	CLC Genomics Workbench 23.0 (QIAGEN, Denmark)	https://digitalinsights.qiagen.com/
Long read mapping tool	CLC Genomics Workbench 23.0 (QIAGEN, Denmark)	https://digitalinsights.qiagen.com/
GAPPadder	Chu et al.[52]	https://github.com/simoncchu/GAPPadder
fastp	Chen et al.[53]	https://github.com/OpenGene/fastp.
BWA	Li et al.[54]	https://bio-bwa.sourceforge.net/
SAMtools	Danecek et al.[55]	https://github.com/samtools/samtools
BUSCO	Simão et al.[56]	https://github.com/metashot/busco
Snakemake	Mölder et al.[57]	https://github.com/snakemake
Salmon	Patro et al.[58]	https://github.com/COMBINE-lab/salmon
Circos	Krzywinski et al.[59]	http://www.circos.ca/software/
R package PCA tools	Blighe et al.[60]	https://www.rdocumentation.org/packages/PCAtools/versions/2.5.13
'Rcmdr' package	Fox[61]	https://www.rdocumentation.org/packages/Rcmdr/versions/2.8-0/topics/Rcmdr-package

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Antonio Figueras (antoniofigueras@iim.csic.es).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The genomic and transcriptomic data generated in this paper, namely the chromosome-level assembly, the PAV data and the transcriptomic dataset, has been deposited at Figshare and is publicly available as of the date of publication. The corresponding DOI is listed in the key resources table. This paper also analyzes existing, publicly available RNA-seq data and genome resequencing data. The SRA accession numbers for the RNA-seq data are listed in Table S6 and the accession number for the resequencing data is listed in the key resources table.

- All code needed for the PAV analysis is publicly available as of the date of publication and the accession is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Mg3 chromosome-level assembly

Although the published reference genome of *Mytilus galloprovincialis* corresponds to the mg10 assembly version (GenBank: GCA_900618805.1), we used the mg3 assembly for all the analyses reported in this study. This decision was justified by the fact that this version included all the uncollapsed allelic variants found in the reference individual, thereby providing a comprehensive view of hemizygous genomic regions, which were partly removed to attempt sequence redundancy reduction in mg10.[21]

The mg3 assembly was scaffolded to a chromosome level using the chromosome-level *M. chilensis* genome as a ref.[43] Briefly, the mg3 contigs were mapped to the *M. chilensis* genome using Minimap2,[62] implemented in CLC Genomics Workbench 23.0 (QIAGEN, Denmark). Mg3 comprised 22,290 scaffolds with 77,787 and 85,594 annotated genes and CDS, respectively. A total of 99.44% of these scaffolds were mapped successfully to the 14 *M. chilensis* chromosomes, while 130 scaffolds were unmapped (Table S1). To polish and improve the assembly, PacBio long reads from the reference individual (ERR4327922, NCBI Bioproject: PRJEB24883) were also mapped against the reference genome of *M. chilensis*. The bioinformatic analysis was conducted using the "Long read mapping" tool in CLC using the following parameters: match score = 2; mismatch cost = 4; gap open cost = 4; gap extend cost = 2; long gap open cost = 24; long gap extend cost = 1; score bonus for global alignment = 0. Most scaffolds showed continuous orientation and overlapping and GAPPadder[52] was used to close the remaining gaps. Finally, the consensus sequence from the mapped mg3 scaffolds and PacBio long reads was extracted. The 22,160 mapped mg3 contigs were oriented and assembled in 14 pseudochromosomes, reaching a genome size of 1.881 Gbases for *M. galloprovincialis* (Table S2) containing 77,769 protein-coding genes (Table S3).

Presence/absence variation and dispensability analyses

The presence/absence variation detection pipeline, which used this version of the genome assembly as a reference, was run as previously described, analyzing whole genome resequencing data obtained from multiple individuals.[21,36] Resequencing data was available for individual genomes from Galicia (3 males and 4 females), namely, GALF1 (ERR2772606), GALF2 (ERR2772607), GALF3 (ERR2772608), GALM1 (ERR3169425), GALM2 (ERR3159549), GALM3 (ERR2772611), and PURA (SRR1598987), and from Italy (3 males and 3 females), namely, ITAF1 (ERR2772612), ITAF2 (ERR2772613), ITAF3 (ERR2772614), ITAM1 (ERR3169424), ITAM2 (ERR2772616), and ITAM3 (ERR2772617). Sequencing data obtained from the reference individual (ERR3169426) were used as a control. All sequencing data are publicly available from the NCBI SRA database, associated with NCBI BioProject: PRJEB25106 and NCBI BioProject: PRJNA262617. Reads were trimmed with fastp[53] and mapped with bwa mem[54] against the reference genomic sequence. Then, the BAM files were sorted with samtools by leftmost coordinates, and samtools depth[55] was used to compute the observed sequencing coverage at each position of the genome.

The transcripts that were included in the mg3 annotation (85,612) were filtered to keep only the longest transcript per gene (77,787), and their exon coordinates were retrieved. A BUSCO[56] analysis was performed at the phylum level (Mollusca database, OrthoDB v.10) with the longest transcript per gene with the aim of retrieving complete single copy genes expected to be shared by all mollusks. Coverage files from each resequenced individual were used to estimate the average coverage for each coding gene and for the BUSCO genes based on exon sequences only. The coverage threshold to identify the genes likely to be absent in a given individual was calculated by dividing the median coverage for the BUSCO genes by eight, an arbitrary threshold that in our previous studies allowed us to reach an excellent correlation between *in silico* gene presence/absence calls and PCR confirmation.[21,36] Per-transcript coverage graphs were prepared for each individual to visually demonstrate the two distinct peaks associated with genes found in a homozygous or hemizygous condition, as well as those lacking any coverage (Figure 2A).

Finally, the genes named absent in at least one resequenced individual were merged to build a complete list of all the dispensable genes annotated in the reference genome.

For each gene in the mussel genome, a parameter named "dispensability" was calculated as the number of individual resequenced genomes in which that particular gene was missing. Therefore, core genes would present a dispensability level of zero, while dispensable genes would have higher or lower dispensability depending if they were absent in just one individual or more rarely found among the analyzed genomic data.

Gene expression data

A gene expression analysis was carried out by mapping *M. galloprovincialis* transcriptomic samples against the aforementioned mg3 gene annotations (considering the longest transcript per gene). The IDs of all the SRA data employed are provided in Table S6 (these included all publicly available *M. galloprovincialis* SRA transcriptomic samples obtained using Illumina sequencing platforms that presented a minimum number of 2.5 million spots per run).

Mussel transcriptomic data were analyzed with a Snakemake pipeline[57] that included sample trimming with fastp of both single-end and paired-end reads,[53] mapping them against the longest transcript per gene in the mg3 assembly using salmon[58] and the generation of mapping saturation curves using a python script. Samples with a mapping rate lower than 30% or with highly saturated mapping curves were discarded.

The expression levels of each transcript were calculated using the transcript per million (TPM) metrics reported by Salmon. The average TPM of each transcript was calculated for each tissue/sample type, allowing the identification of tissue-specific genes, defined as those showing an average TPM value >50 in a single tissue/sample type and an average TPM value <5 in all the other tissues/sample types.

Data visualization

Circos[59] was used to visualize chromosomal data, and the R package PCA tools[60] was used for the principal component analyses.

QUANTIFICATION AND STATISTICAL ANALYSIS

Several analyses required statistical tests. ANOVA-Welch F tests were run, in order to prevent assumption problems with unequal variances, using the Rcmdr package[61] for the analyses included in Figures 3 and 4. The significance was defined by a p-value < 0.05 and statistical significance was indicated with asterisks in both figures. For the analyses in Figure 3, input data was the dispensability value per megabase in each chromosome and the statistical test reported the chromosomes or the 5-megabase windows enriched in dispensability (see figure legend for details). For the analysis in Figure 4, input data were expression deviation values across the transcriptomic dataset for each group of genes, classified according to their dispensability.