

RESEARCH ARTICLE

The highly dynamic pangenome of basal chordates is enriched in defence and immunity genes and is inherited following the Mendelian law

Umberto Rosani^{1*}, Marco Gerdol², Mart Krupovic³

1 Department of Biology, University of Padova, Padova, Italy, **2** Department of Life Sciences, University of Trieste, Trieste, Italy, **3** Institut Pasteur, Université Paris Cité, CNRS UMR6047, Cell Biology and Virology of Archaea Unit, Paris, France

* umberto.rosani@unipd.it



OPEN ACCESS

Citation: Rosani U, Gerdol M, Krupovic M (2025) The highly dynamic pangenome of basal chordates is enriched in defence and immunity genes and is inherited following the Mendelian law. PLoS Genet 21(8): e1011833. <https://doi.org/10.1371/journal.pgen.1011833>

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, UNITED STATES OF AMERICA

Received: January 28, 2025

Accepted: August 5, 2025

Published: August 18, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pgen.1011833>

Copyright: © 2025 Rosani et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

Abstract

Pangenome analyses, which encompass the full genetic repertoire of a species, offer valuable insights into intraspecific diversity and phylogeographic gene patterns. While the taxonomic breadth and functional significance of animal pangenomes remain to be fully uncovered, recent findings (such as reports of open, bacterial-like pangenomes in bivalves) highlight the need to better understand the molecular mechanisms driving inter-haplotype structural variation. Genes affected by presence-absence variation (PAV), along with non-reference sequences (NRSs), represent evolutionary footprints that may shape genome architecture and plasticity, ultimately influencing the adaptability and long-term fitness of species. To investigate the pangenomic architecture of basal chordates, we analyzed available whole-genome resequencing data from *Branchiostoma belcheri* and *B. floridae*, examined the impact of structural genomic variation, and assessed the inheritance patterns of dispensable genes across generations. The pangenomes of both species include over a thousand of genes affected by PAV and exhibiting trans-generational Mendelian transmission from parents to offspring. We further demonstrate that 35 dispensable genes in *B. belcheri* are of exogenous origin, likely resulting from the integration of a malacoherpesvirus genome, thereby extending the known host range of *Malacoherpesviridae* from invertebrates to chordates. PAV preferentially targeted gene families involved in defense, immunity, and cell signaling, including GTPases of immunity-associated proteins (GIMAPs), caspases, toll-like receptors, and pattern recognition receptors containing apextrin C-terminal (APEC) domains. The dynamic nature of immunity genes in cephalochordates parallels patterns seen in open bacterial pangenomes, suggesting that fundamental principles of genome evolution and innovation across life domains are shaped by host–pathogen interactions.

and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All the data are either already available (as indicated in the M&M section) or included in the [Supporting information](#).

Funding: This work was supported by the Italian National Project PRIN PNRR 2022 (ID: Prot. P2022JEEMT - Developing a tool for the study of haplotype diversity in *Mytilus galloprovincialis*), which provides support to UR and MG. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Understanding how genetic diversity arises and is maintained within a species is key to reveal evolution and adaptation mechanisms. This study explores the full set of genes - or “pangenome” - in two species of lancelets (*Branchiostoma belcheri* and *B. floridae*), an early-diverging lineage of chordates. By analysing whole-genome sequences, we found that more than a thousand genes are not shared by all individuals. These dispensable genes were passed from parent to offspring like typical inherited traits and are enriched in domains involved in immunity and cell signaling. Strikingly, a block of 35 dispensable genes in *B. belcheri* represented an integrated genome of a malacoherpesvirus, revealing both the integration event and supporting the association between these viruses and lancelets. Overall, our findings revealed that immune-related genes are especially prone to be part of the dispensable gene set in both lancelet species, similar to what is observed in bacteria. This observation highlights a shared evolutionary strategy across life domains, where genetic flexibility - especially in defensive genes - may be crucial for survival in a changing environment.

Introduction

Biodiversity is a unifying concept in biology that can be studied on different levels, ranging from genes to ecosystems [1]. Indeed, although the definition of biodiversity is mostly discussed at the level of species, it is grounded in the genetic variation within species [2], and is tightly connected to the concept of function, which also spans the molecules-to-ecosystem range [3]. First proposed by Tetz in 2003, the term pangenome, according to its original definition, was intended as a unifying view of all genetic information, including the complete gene set of all living organisms, along with associated viruses and mobile genetic elements (MGE) [4,5]. However, in subsequent years, this term has been mostly used in the scientific literature with a much narrower scope, referring specifically to the full set of genes found across all strains of a given bacterial species. These genes can be classified as core, when shared by all individuals, or as dispensable, when subject to presence-absence variation (PAV) among individuals [6]. Dispensable genes, also referred to as ‘auxiliary’ genes [7], can be further classified as softcore, shell or cloud, based on their decreasing frequency in a population. The pangenome of each species can be defined as ‘open’ or ‘closed’, depending on the increase (or lack thereof) in the number of novel genes provided by sequencing of additional individuals [8]. Although this gene-centric definition derives from the fact that early studies on pangenomes were primarily conducted in prokaryotes [9], the subsequent identification of pangenomic architectures in eukaryotes prompted a broader reinterpretation. Today, the term pangenome is not limited to protein-coding genes but usually encompasses a wide range of genetic elements, more accurately capturing the full extent of intraspecific genomic diversity. Nevertheless, gene-centric approaches continue to be a key element in pangenomic

studies, because they allow obtaining a direct link between dispensable genes, their associated functions and, eventually, phenotypic traits [10]. Besides bacteria, pangenomes have been reported in plants [11], archaea [12], giant viruses [13], fungi [14] and animals, including humans [15], goat [16] and, among invertebrates, in the bivalve *Mytilus galloprovincialis* [17]. Although open pangenomes are more common in bacteria than eukaryotes [8,18,19], the relationships among pangenome type, genome size, repeat content, and species phylogeny remain poorly understood [10]. In bacterial pangenomes, the interplay between horizontal gene transfer (HGT) and homologous recombination likely shapes genome composition, with inter-species boundaries and ecological barriers to DNA exchange acting as key determinants [20]. Notably, the dispensable fraction of bacterial pangenomes is strongly enriched in genes involved in defence mechanisms and mobile genetic elements [21–23].

In sexual eukaryotes, where stronger evolutionary constraints typically limit genomic plasticity, pangenome evolution may be primarily shaped by sexual recombination, gene duplication, and domain reshuffling, with lateral gene transfers playing a comparatively minor role [24,25]. However, over long evolutionary timescales, mechanisms such as virus-mediated gene transfer, retrotransposon activity, homology-driven recombination, and non-homologous end joining may significantly contribute to the emergence and maintenance of dispensable genomic content [25,26]. Although the genomic mechanisms responsible for the emergence of dispensable genes are increasingly recognized, much less attention has been given to the transmission dynamics of these genes across generations. Dispensable genes can be modelled as biallelic loci, where one allele corresponds to gene presence and the other to gene absence, and are therefore assumed to follow Mendelian inheritance, segregating normally among offspring. However, this assumption may not always hold, because dispensable genes are frequently embedded within large structural variants that may impair homologous chromosome pairing during meiosis [27]. From a theoretical standpoint, such disruptions can lead to segregation distortion or non-random transmission, resulting in deviations from expected Mendelian ratios in progeny. Additionally, while each dispensable gene can be individually absent, the combined absence of multiple functionally redundant dispensable genes may not be equally tolerated. In such cases, selective constraints may limit the viable gene PAV configurations across individuals. In this context, although pangenomic studies in metazoans are still in their infancy, recent reports strongly suggest that transposable elements play a key role in driving the acquisition and maintenance of intraspecific structural variation [28]. Integration and endogenization of viral genomes within the host chromosomes further expands the functional landscape and complexity of pangenomes. This observation is supported by the identification of non-reference sequences (NRSs), likely originating from the combined activity of diverse MGEs, including both transposons and bona fide capsid-forming viruses, in human resequencing datasets [29].

Lancelets are marine benthic animals that display cosmopolitan distribution in temperate waters and occupy a basal taxonomic position within the phylum Chordata. They have been used as a model to study the evolution of early vertebrates [30], as these two lineages share a common ancestor that lived around half a billion years ago [31], but the intra-specific structural genomic variations in lancelets have never been investigated. The recent expansion of the genomic datasets for cephalochordates provides a solid basis for pangenome investigations. Lancelets have received considerable attention due to the so-called ‘amphioxus polymorphism paradox’: despite the remarkably high nucleotide variability (1 SNP every 34 nt), no apparent morphological differences are present among individuals, suggesting that in this species, SNPs do not impact phenotypic characteristics [32]. Chromosome-level genome assemblies have been generated for *B. belcheri*, *B. japonicum* and *B. floridae* [33], as well as for *B. lanceolatum* [34], and high-coverage resequencing data has been generated for *B. belcheri*.

Here, we investigate the pangenomic architectures in two lancelet species, *B. floridae* and *B. belcheri*. We report the widespread occurrence of dispensable genomic regions affected by PAV and show that these regions are enriched in genes involved in immunity and signalling. These observations highlight parallels between the pangenomes of prokaryotes and lancelets, suggesting that genome dynamics in these organisms are shaped, to a considerable extent, by antagonistic host–pathogen interactions. Finally, we traced the NRSs in lancelet pangenomes and discovered the integration of

a nearly complete malacoherpesvirus genome into the *B. belcheri* chromosome, possibly representing an active lancelet virus with ancient connections to mollusk herpesviruses.

Results

Lancelet genomes are characterized by widespread PAV

As we have previously demonstrated [17,35], aligning whole-genome resequencing data to an annotated reference genome represents a robust approach for detecting PAV. Unlike core genomic regions, which are consistently shared among all individuals of a species, regions affected by PAV exhibit inter-individual structural variation and are often entirely absent or present in a hemizygous state. These regions, along with the genes they contain, are therefore considered dispensable and constitute the variable fraction of the pangenome of a species. We applied this approach to 99 *B. belcheri* and 41 *B. floridae* whole genome resequencing samples (S1 Table), revealing that a considerable fraction of the two reference genome assemblies, i.e., 16.1% and 10.3% respectively, were dispensable. Most of these regions were softcore (i.e., dispensable, but present in > 90% of the samples), whereas shell/cloud regions, absent in more than 10% of the considered individuals, only accounted for 3.7% and 6.6% of the genome size in *B. belcheri* and *B. floridae*, respectively (Table 1).

The analysis of protein-coding genes confirmed these results, identifying 3,171 and 2,309 dispensable genes in *B. belcheri* and *B. floridae*, respectively. As in the case of genomic regions, a significant fraction of dispensable genes could be assigned to the softcore category, and only 1,112 (in *B. belcheri*) and 1,281 (in *B. floridae*) genes were classified into the shell/cloud category (accounting for 4.4% of the total in both species) (Table 1, S1 and S2 Files and S1 Fig). The number of missing genes per genome ranged from 350 to 550 in *B. belcheri* and from 450 to 600 in *B. floridae* (Fig 1a). By contrast, the two gonad samples used to generate the reference genome assembly of *B. belcheri* possesses a few missing genes and regions and provided a benchmark for the false positive rate of our pipeline, which was estimated to be ~0.23% for genomic regions and 0.05-0.09% for genes. In both species, the number of missing genomic regions and genes per resequenced genome were strongly correlated ($R=0.58$ in *B. belcheri*, $R=0.55$ in *B. floridae*, Fig 1b).

In line with the expected lower frequency of occurrence of dispensable genes in lancelet populations, the analysis of the transcriptomic datasets available for both species (S1 Table) showed that the shell and cloud genes were expressed at significantly lower levels than the core genes (Fig 1c-d), to the point that their transcription was frequently undetectable (Fig 1e-f).

Structural organization of lancelet pangenomes

Despite the high quality of the available genomic resources for *B. floridae* and *B. belcheri*, the two reference genome assemblies are not haplotype-resolved, which limits the ability to directly assess the spatial and size distribution of structural variants through the alignment between haplotypes. Nonetheless, we leveraged the information derived from the analysis of PAV patterns presented in the previous section to investigate some of these aspects. First, the mapping of shell and cloud genes in *B. floridae*, the only of the two considered species with a chromosome-scale genome assembly, allowed us to visualize the chromosomal distribution of dispensable genes (Fig 2). These genes were found to be distributed across all 19 chromosomes and, although some regions showed variation in density, this pattern was consistent with a random distribution of structural variants, as no chromosome exhibited a significantly higher concentration of PAV-affected genes than the genome-wide average.

Table 1. Results of the PAV analysis in the two lancelet species.

Species	No. of samples	Dispensable regions (No./Percentage)	Shell/cloud regions (No./Percentage)	Dispensable genes (No./Percentage)	Shell/cloud genes (No./Percentage)
<i>B. belcheri</i>	99	13,932 (16.1%)	3,228 (3.7%)	3,171 (12.6%)	1,112 (4.4%)
<i>B. floridae</i>	41	10,625 (10.3%)	6,805 (6.6%)	2,309 (7.9%)	1,281 (4.4%)

<https://doi.org/10.1371/journal.pgen.1011833.t001>

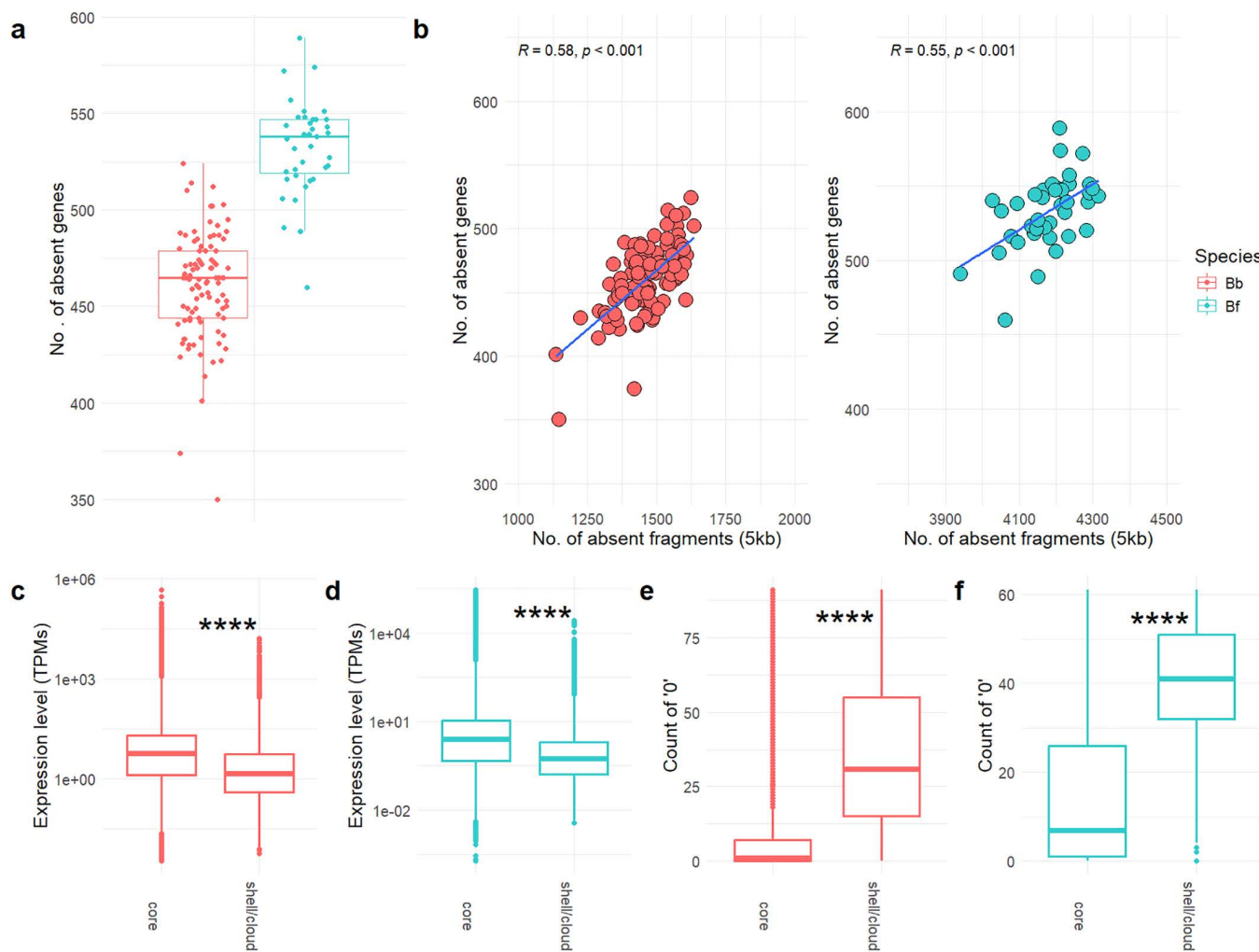


Fig 1. PAV analysis of *B. belcheri* (red) and *B. floridae* (blue) whole genome resequencing datasets. **a**. Number of absent genes per sample. **b**. Correlation between the number of absent genomic regions vs. the number of absent genes per sample, divided by species. The two DNA samples used to build the reference genome have been removed from the plot. **c** and **d**. Comparison between the expression levels of shell/cloud and core genes observed in the available RNA-seq datasets of *B. belcheri* (**c**) and *B. floridae* (**d**). **e** and **f**. Fraction of shell/cloud and core genes showing null expression values in the available RNA-seq datasets of *B. belcheri* (**e**) and *B. floridae* (**f**). ****, p -value < 0.001 .

<https://doi.org/10.1371/journal.pgen.1011833.g001>

We also analysed the distribution of contiguous shell/core gene block sizes in both species, finding that most of these consisted of singletons (80.4% in *B. belcheri*, 74.9% in *B. floridae*) or contained fewer than 5 genes (99.0% in *B. belcheri*, 99.1% in *B. floridae*) (S2 Table). This result aligns with the findings obtained from the correlation analysis of concordant PAV patterns and pairwise distances between pairs of dispensable genes placed on the same chromosome or scaffold. In fact, this correlation was significantly higher than random association only for distances below approximately 30 kb in *B. belcheri* and 25 kb in *B. floridae*, even though significant correlations were occasionally also observed across greater distances (S2 and S3 Figs). Thus, in both species of amphioxus, most of the genomic regions subjected to PAV were relatively small in size, with only rare instances of multiple neighbouring dispensable genes.

Nevertheless, an exception to this general pattern was observed in *B. belcheri*, which contains a block of 35 contiguous dispensable genes located on the same scaffold, spanning ~150 kb of sequence. Interestingly, these genes showed

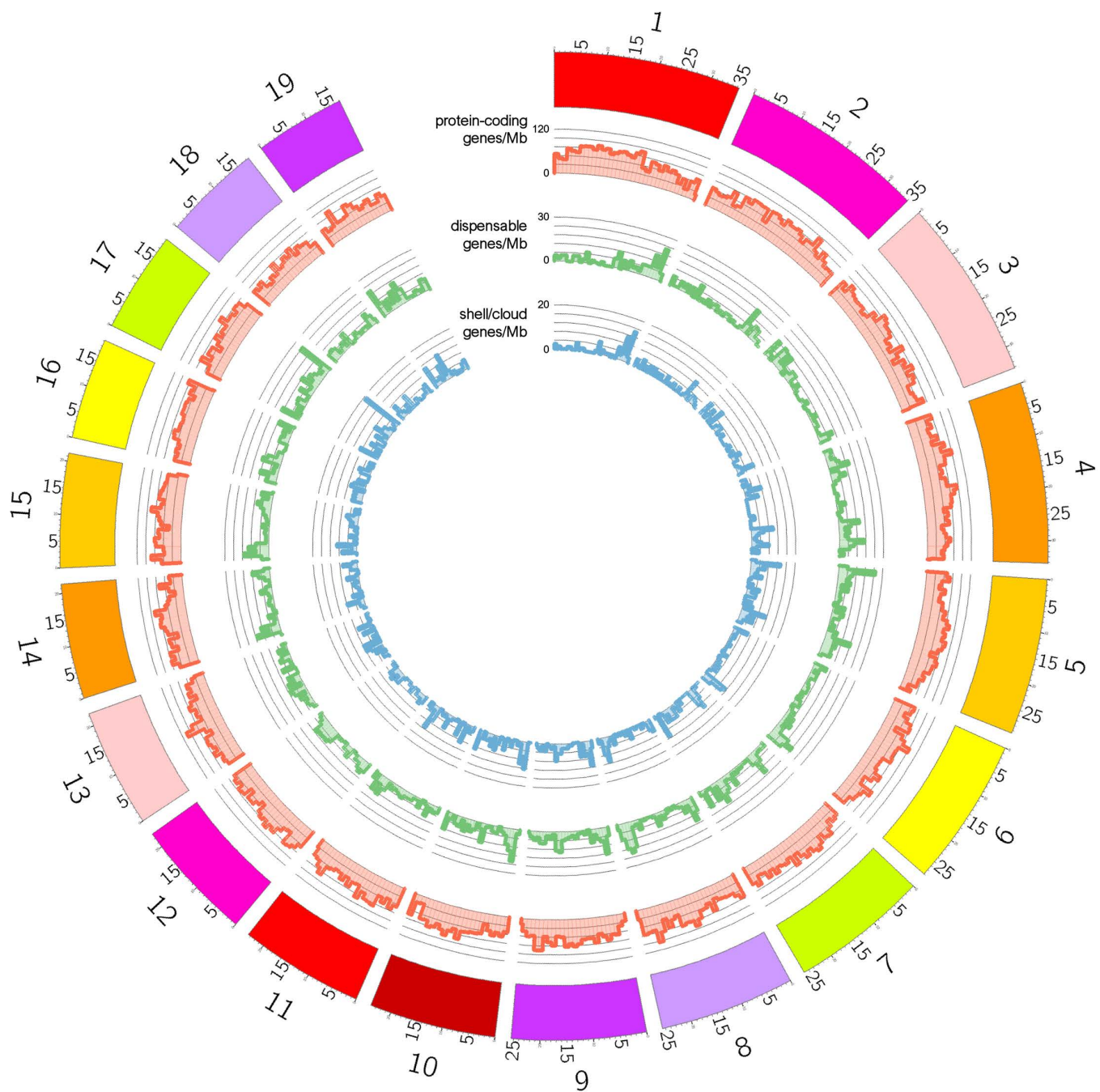


Fig 2. Spatial distribution of shell/cloud genes, dispensable genes and protein coding genes in the genome of *B. floridae*, reported as gene density per Mb. Note that unplaced genomic scaffolds and the mitochondrial genome were ignored.

<https://doi.org/10.1371/journal.pgen.1011833.g002>

highly concordant PAV patterns, being simultaneously absent in 79 out of the 99 resequenced individuals (S1 Fig), which resulted in a clear clustering of the resequenced genomes into two groups (S4 Fig). The nature of this unusually large dispensable region identified in the *B. belcheri* genome will be examined in detail in the following section.

Considering recent data supporting the key role played by transposable elements in establishing structural variation in the pangenome of bivalve mollusks [28], we investigated the potential enrichment of specific classes of repetitive elements in the regions flanking the shell and cloud gene blocks, as compared to the core genes. Although this analysis did not reveal a robust association between dispensable genomic regions and flanking MGEs, a significant enrichment of Gypsy-like LTR retrotransposons/reverse-transcribing viruses (family *Metaviridae*) and PIF/Harbinger-family DNA transposons was detected in *B. floridae* and *B. belcheri*, respectively, even though they were only associated with a very small fraction of all shell/cloud genes (i.e., ~1%) (S3 Table).

An integrated malacoherpesvirus-like genome and lancelet NRS

As mentioned in the previous section, the in-depth analysis of gene PAV patterns led to the identification of an anomalous large block of dispensable genes in the *B. belcheri* reference genome, which was also detectable in 20 of the analysed resequenced samples. These genes were found to be part of the integrated genome of a herpes-like virus spanning 149,856 bp in length. The viral genome displayed a GC content of 45.7%, slightly higher than the average GC% of *B. belcheri* (41.3%), and was flanked by regions rich in repetitive elements, both at the 3'- (4.7 kb with 43% of repeats) and at the 5'-end (28.8 kb with 65.5% of repeats, Fig 3a and S3 File).

The presence of a putative bona fide virus was further supported by the reannotation of this region using criteria valid for viral genomes, yielding a total of 62 open reading frames (ORFs), 27 of which showed homology to malacoherpesviruses and four to vertebrate herpesviruses. Overall, 72.7% of the provirus region length was covered by protein-coding genes (Tables 2 and S4). Hereinafter, we refer to this integrated herpesvirus (HV) sequence in *B. belcheri* as BelcheriHV-1. The viral genome consists of three discontinuous gene blocks. The first block of 46 kb is characterized by tightly clustered viral genes (coding density: 82.8%) and was identified in 20/20 samples. The second block (21 kb) showed a sparser distribution of genes that were not recognizably similar to genes of known malacoherpesviruses (coding density: 37.1%). The latter gene block was covered by multiple assembled scaffolds in each resequenced sample, likely because of the presence of repeated regions. The last block, originally spanning 81.5 kb in length, showed a deletion of 5 kb in all the 20 samples, resulting in a length of 75.5 kb, with a coding density of 75.6% (Table 2). Within this block, malacoherpesvirus genes are intermixed with genes of unknown origin. Unfortunately, the junctions between blocks cannot be resolved based on the available data. Interestingly, the SNPs located on BelcheriHV-1 genome were mostly detected either with a frequency of 100% (fixed variations) or 50%, possibly indicating the presence of two BelcheriHV-1 variants in some of the tested genomes (Fig 3b). Overall, BelcheriHV-1 was 24 × less impacted by nucleotide variations than the *B. belcheri* genome, with an average of one SNP per 5,958 nt, vs one SNP per 174 nt, respectively (Fig 3c). This result strongly suggests that BelcheriHV-1 has evaded the 'amphioxus polymorphism paradox'.

The average sequencing coverage of BelcheriHV-1 in the 20 positive lancelet samples was slightly lower than the genome coverage (i.e., 0.7×), which is in line with the differences observed between the average coverage of the 35 predicted viral genes compared with the average coverage of the 23,868 lancelet genes (Fig 3d). Only a single muscle sample (SRR8324690) showed a coverage as high, i.e., 1.2×, whereas a gonad sample showed 0.9× coverage based on genome regions (0.8× considering the gene coverages, Fig 3d).

We further screened the taxonomic composition of individual de novo assembled lancelet NRSs, which allowed the identification of three contigs showing similarity with BelcheriHV-1 obtained from the assembly of the genome of a single male *Belcheri* individual (SRR14996628), which tested negative for BelcheriHV-1. These contigs cumulatively amounted to 132.5 kb, encoded 72 ORFs (Tables 2 and S4 and S4 File) and showed homogenous coverage, which suggested that they were part of the same viral genome, called BelcheriHV-2. Both GC% content and coding density were similar to those of BelcheriHV-1 as well as to the gastropod-infecting malacoherpesvirus (HaHV-1, Table 2). The proviral genome was not present in any of the 41 available resequenced *B. floridae* samples or in any of the lancelet genome assemblies.

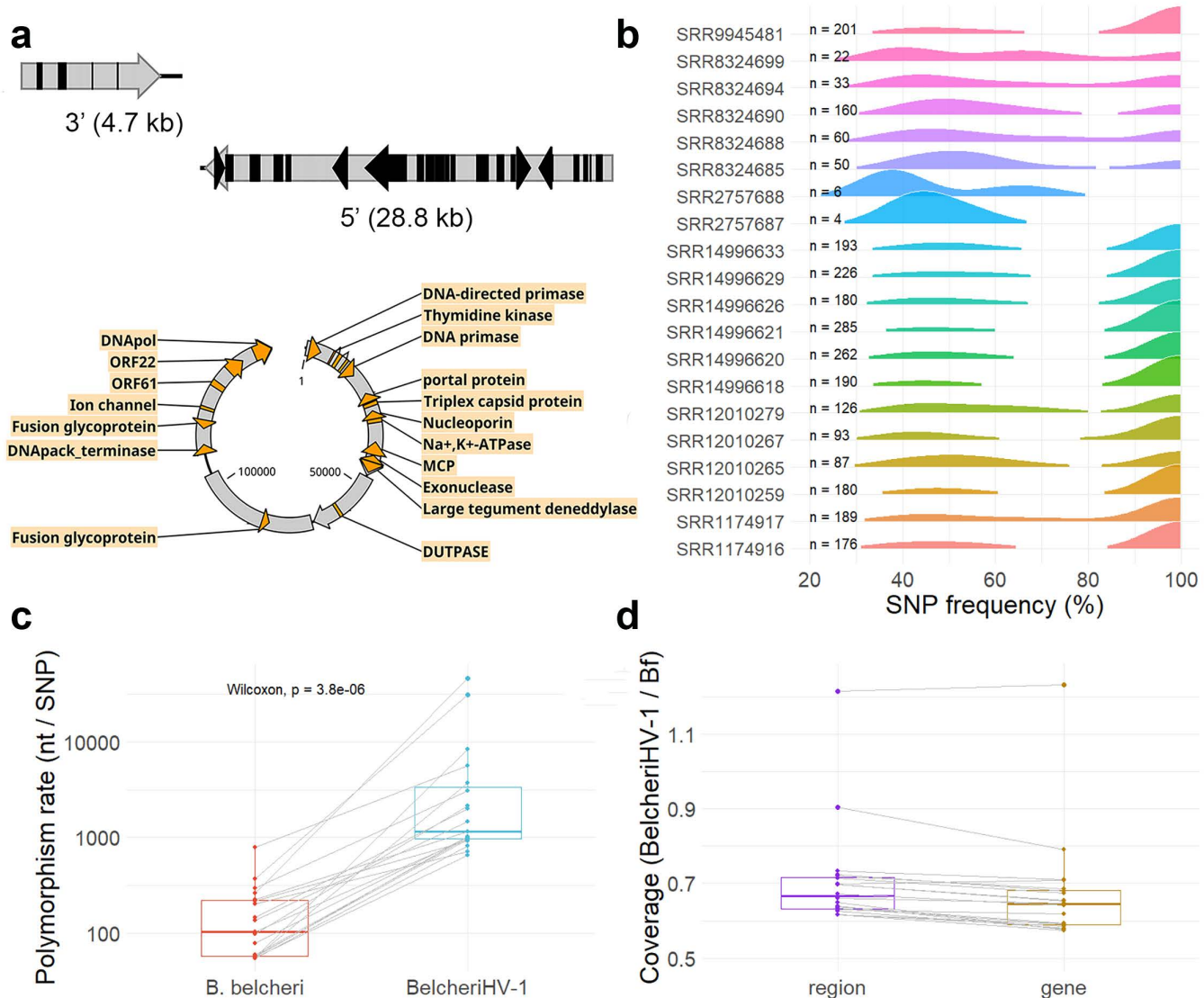


Fig 3. The integration of BelcheriHV-1 into the genome of *B. belcheri*. **a.** Circular visualization of the BelcheriHV-1 genome is provided, indicating its putative integration site between two genomic regions rich in interspersed repeats (reported as black arrows or boxes). In the BelcheriHV-1 genomes, the genes showing similarity with herpesviruses are reported as orange arrows, together with their putative annotations. More detailed annotations are reported in [Tables 3](#) and [S4](#). **b.** The distribution frequencies of the BelcheriHV-1 SNPs detected in the 20 positive samples. The total number of SNPs is also reported for each sample. **c.** The polymorphism rates, computed as the distance between SNPs, of *B. belcheri* and of BelcheriHV-1 in the 20 positive samples (the same samples are matched by grey lines). **d.** The ratio between the BelcheriHV-1 and the *B. belcheri* coverages computed based on non-overlapping sliding windows (region) or genes (the same samples are matched by grey lines).

<https://doi.org/10.1371/journal.pgen.1011833.g003>

Dispensable genes are enriched in specific conserved domains and functions

To investigate the possible functional implications of lancelet dispensable genes, we analysed the predicted functions of the shell/cloud dispensable genes using an enrichment test based on the Pfam conserved domains and the Gene Ontology (GO) terms. A total of 46 and 26 conserved domains were significantly enriched among the *B. belcheri* and *B. floridae* shell/cloud dispensable genes, respectively ([Fig 4a-b](#)), with 17 domains shared by both species ([Fig 4c](#) and [S5 Table](#)).

Table 2. Belcheri herpesviruses. The table provided a comparison between the genome length, the number of predicted genes, the GC content and the coding density of belcheri herpesviruses in comparison with other two malacoherpesviruses (OsHV-1 and HaHV-1). For BelcheriHV1 the details regarding the three identified genome blocks are also reported.

Virus name	Genome length	Host species	Status	No. of predicted genes	GC content (%)	Coding density (%)
BelcheriHV-1	149,856	Lancelet	Integrated	62	45.7	70.4
-block 1	46,000	/	/	19	46.5	82.8
-block 2	21,000	/	/	13	45.4	37.1
-block 3	75,500	/	/	30	45.5	75.6
BelcheriHV-2	132,488	Lancelet	Integrated/fragmented	74	44.9	85.1
OsHV-1	203,893	Bivalve	Free living	127	38.6	83.8
HaHV-1	210,993	Gastropod	Free living	114	46.9	81.6

<https://doi.org/10.1371/journal.pgen.1011833.t002>

In both species, the most over-represented was the Death domain (PF00531), present in 35 and 34 dispensable genes in *B. floridae* and *B. belcheri*, respectively. Several domains, typically implicated in the recognition of microbe-associated molecular patterns, such as the *lectin C-type domain* (PF00059), the *apextrin C-terminal domain* (PF16977) and the *TIR domain* (PF13676), as well as with intracellular immune signalling, such as *AIG1* (PF04548), were also significantly over-represented in both lancelet species.

The same analysis performed on the GO terms revealed 17 biological processes, 27 molecular functions and two cellular components (kinesin complex and microtubule) enriched in *B. belcheri*, whereas 16 biological processes, 13 molecular functions and one cellular component (plasma membrane) were enriched in *B. floridae* (Fig 4d-e and S5 Table). Notably, nine biological functions (signal transduction, regulation of apoptotic process, DNA integration, toll-like receptor signalling pathway, immune response, DNA recombination, positive regulation of interleukin-1 production, proteolysis and cell adhesion) and five molecular functions (GTP binding, cysteine-type endopeptidase activity, channel activity, cysteine-type peptidase activity, and protein tyrosine kinase activity) were shared between the two species (Fig 4f). Considering specific functional domains, the *lectin C-type domain*, responsible for carbohydrate binding and functioning as a receptor in various contexts, including adhesion and signalling, and the *death domain*, involved in apoptosis and inflammation, were the most enriched domains in *B. belcheri*. Similarly, *death* and *death effector* domains were the most enriched domains in *B. floridae*. The *MAC/perforin-like domain* found in cytolytic proteins that form pores in the target membranes, the *AIG1 family domain* present in the GTPase of the immunity associated protein (GIMAPs) family, a group of proteins involved in immunity in plants and invertebrates, and the *apextrin C-terminal domain* (APEC), associated with pattern recognition receptors, were all enriched in both lancelet species. The dispensability rates of the shell/cloud genes encoding selected Pfam domains were similar in the two species, except for the AIG1 family domain, which showed a higher dispensability in *B. belcheri* (Fig 4g). Since the size of most de novo assembled NRSs obtained from whole genome resequencing datasets was significantly smaller than the average gene length of both species (i.e., ~6 kb), most of the predicted genes associated with these sequences were expected to be fragmented. Despite this limitation, the analysis of the NRS-encoded proteins confirmed, in both species, the high representation of the same Pfam domains enriched among the shell/cloud dispensable genes (Fig 4h-i).

We further investigated the evolutionary history of the shell/cloud dispensable genes annotated in the reference genomes and those associated with NRSs by determining their orthology based on 11 selected proteomes including three lancelet, one colonial ascidia, three vertebrate and three invertebrate species (Fig 5). We retrieved 1,021 orthogroups that included dispensable genes of *B. belcheri* (N=564) or *B. floridae* (N=699). Dispensable genes were mostly present in different combinations among lancelet species, and only in 4.4% of the cases they belonged to orthogroups with a broad taxonomic distribution, covering all the other considered species (Fig 5a). Notably, 243 of orthogroups are shared between *B. belcheri* and *B. floridae* (Fig 5b). On average 49% and 33% of the paralogs per orthogroup were dispensable in *B.*

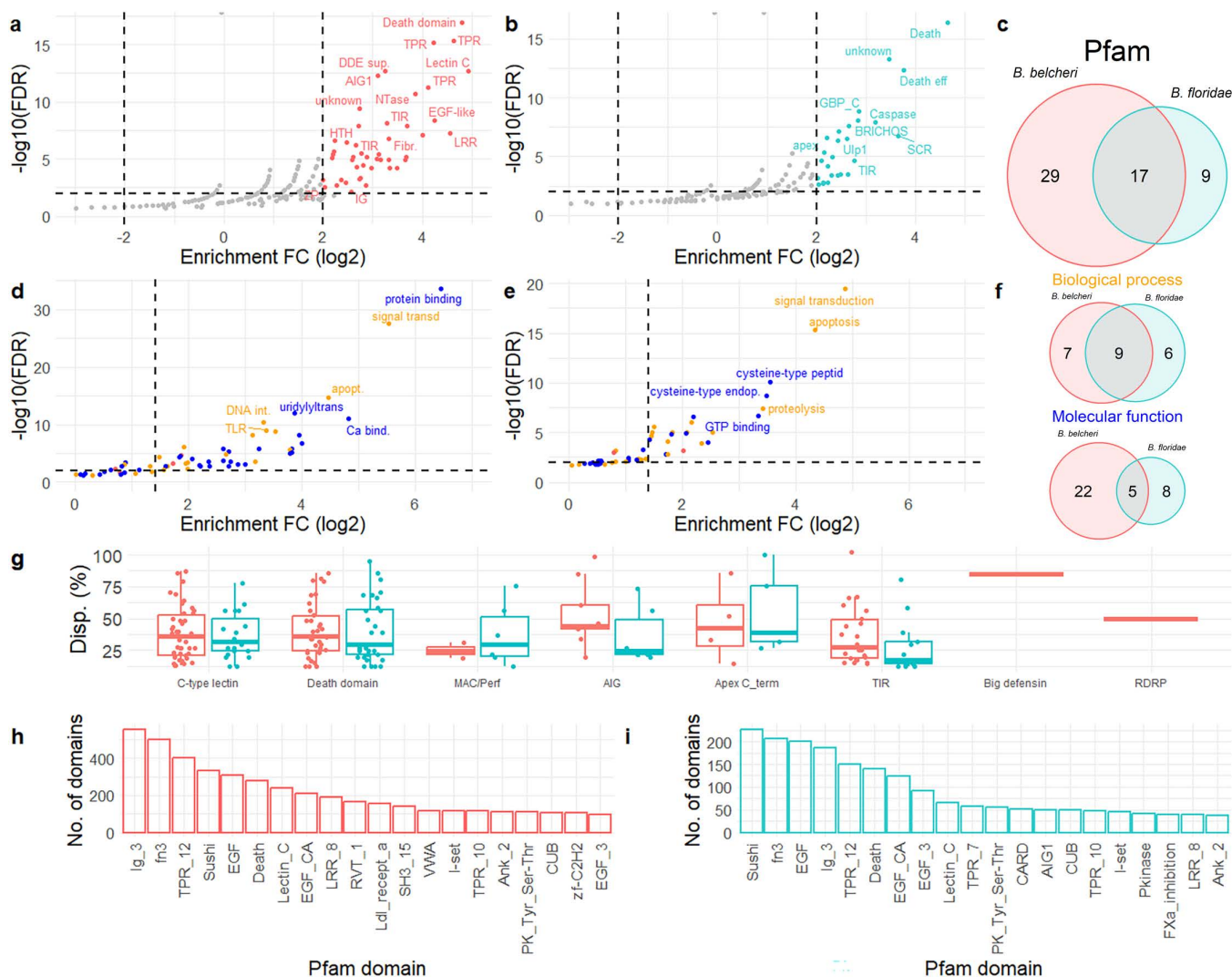


Fig 4. Functional enrichment of dispensable genes in lancelets. Volcano plot depicting the enriched Pfam domains encoded by the shell/cloud dispensable genes in *B. belcheri* (a) and *B. floridae* (b). The domains distribution is based on the false discovery rate (FDR) reported in a logarithm scale ($-\log_{10}$) on the y-axis and the enrichment fold change reported as $\log_2$ on the x-axis. Significantly enriched domains are coloured and named with abbreviations. c. Venn diagram depicting the enriched Pfam domains found exclusively in either *B. belcheri* or *B. floridae* or shared by the two species. Volcano plot depicting the enriched GO terms retrieved from the shell/cloud dispensable genes in *B. belcheri* (d) and *B. floridae* (e). The GO term distribution is reported as reported above for Pfam domains in panel (a); GO terms are coloured according to their classification into 'biological process' (orange), 'molecular function' (blue) or 'cellular component' (red). f. Venn diagram depicting the enriched biological process and molecular function GO terms found exclusively in either *B. belcheri* or *B. floridae* or shared by the two species. g. Dispensability fractions of selected gene families, reported as the number of individuals where the gene is absent over the total number of tested individuals. The abundance of the different Pfam domains encoded by the dispensable genes associated with de novo assembled NRSs is plotted for *B. belcheri* (h) and for *B. floridae* (i). The 20 most abundant domains are reported.

<https://doi.org/10.1371/journal.pgen.1011833.g004>

belcheri and *B. floridae*, respectively, even though some orthogroups were exclusively composed of dispensable genes (Fig 5c). Notably, the dispensable genes annotated in NRSs were distributed in 6,183 orthogroups, of which 55% were lancelet-specific.

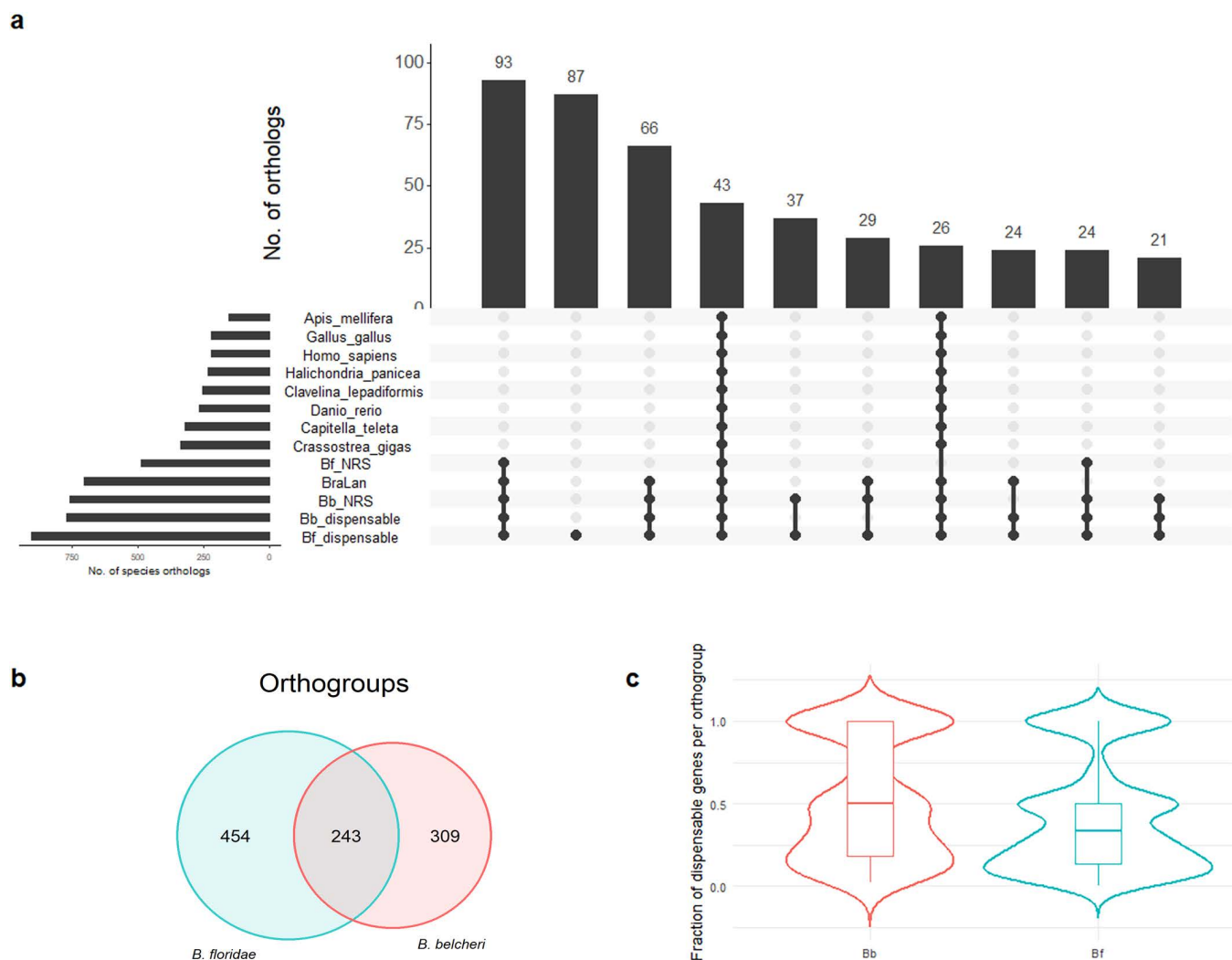


Fig 5. Orthology analysis of lancelet dispensable genes. **a.** Upset plot showing the distribution of the orthogroups that included lancelet dispensable genes and de-novo gene (NRS) among the 11 selected species. The graph reports the number of orthogroups detected for each species combination. **b.** Number of orthogroups exclusively found in *B. floridae*, *B. belcheri*, or in both species, among those including lancelet dispensable genes. **c.** Distribution of the fraction of dispensable genes per orthogroup in the two lancelet species.

<https://doi.org/10.1371/journal.pgen.1011833.g005>

Dispensable defence genes in lancelets: GIMAP and APEC proteins as case studies

Our functional enrichment analysis revealed that the AIG1 domain was one of the most significantly over-represented Pfam domains among the dispensable genes in both lancelet species. This domain characterizes a large and heterogeneous group of P-loop GTPases, which include animal GTPases of immunity-associated protein (GIMAP) family and plant immune-associated nucleotide-binding proteins (IAN) [36]. Indeed, 9 out of 13 GIMAP genes annotated in *B. belcheri* and 6 out of 12 GIMAP genes annotated in *B. floridae* were shell/cloud dispensable genes. We investigated in more detail the evolution of this gene family in lancelet, by recovering additional GIMAP genes from lancelet de novo assembled NRSs and comparing their sequences with those of the other species included in the orthology analysis reported in the previous section. Consistent with previous reports [37–39], GIMAPs have been subjected to multiple independent gene family expansion events during evolution, leading to the emergence of several lineage-specific clades (Fig 6a). This was also the case with

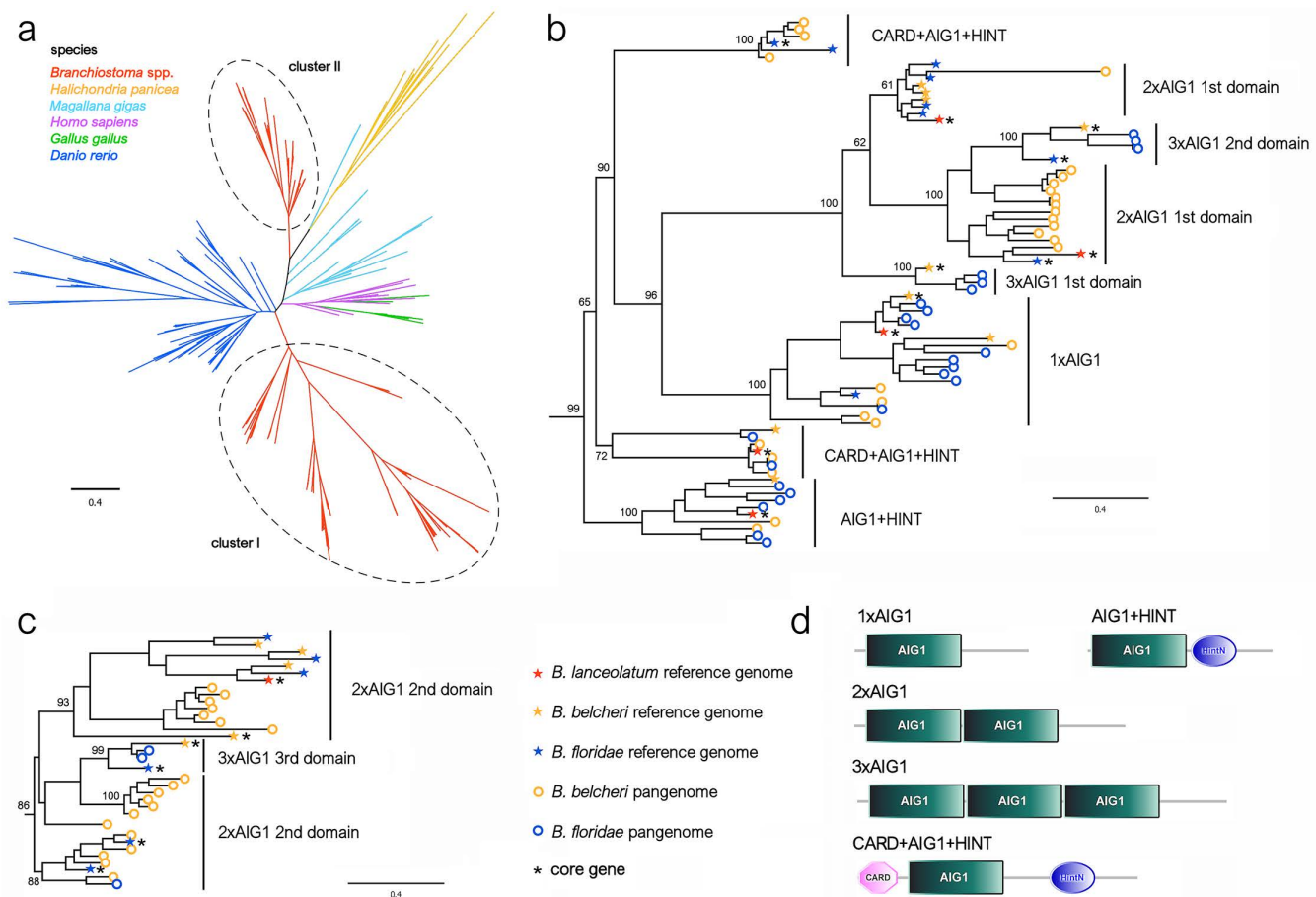


Fig 6. The AIG1 domain. a. Unrooted ML phylogeny of representative metazoan AIG1 family proteins. The lancelet proteins are highlighted in red and classified within two major groups, i.e., clade I and clade II, which are detailed in panels b and c, respectively. Here, the bootstrap support values for the major nodes of the tree are shown, along with the core or dispensable status of the sequences and an indication of the domain architecture. d. Exemplified domain architecture of lancelet AIG1-domain containing proteins.

<https://doi.org/10.1371/journal.pgen.1011833.g006>

the lancelet sequences. In a phylogenetic tree based on the individual AIG1 domains present in multidomain proteins, the sequences formed two distinct clusters, distantly related to homologs from vertebrates, *M. gigas* and *H. panicea*. In particular, the two clusters represent: (i) a large group of AIG1 domains associated with different architectures (cluster I, Fig 6b); (ii) the C-terminal domain of proteins displaying either two or three consecutive AIG1 domains (cluster II, Fig 6c). The most recurrent architectures, which involved the presence of the N-terminal caspase activation and recruitment domain (CARD) and C-terminal HINT domains (Fig 6c), were unique to amphioxus among animals. However, some architectures, such as the CARD+AIG1+HINT domain combination, marked highly interesting cases of convergent evolution, being only present in a few multicellular fungi. Albeit unusual, similar occurrences have been previously documented for GIMAP-like genes in nematodes, where they have been interpreted as the probable result of horizontal gene transfer [38]. The coupling of the CARD and AIG1 domains further emphasizes the involvement of these proteins in apoptosis and inflammation.

The apexrin C-terminal (APEC) domain associated with pattern recognition receptors was also one of the most significantly enriched domains in the set of dispensable genes from lancelet, with 5 out of 13 annotated genes being subjected to PAV in both *B. belcheri* and *B. floridae*. We investigated the evolutionary history of these genes by evaluating their relationships with APEC-domain containing proteins from other metazoan lineages. In line with the results of a recent

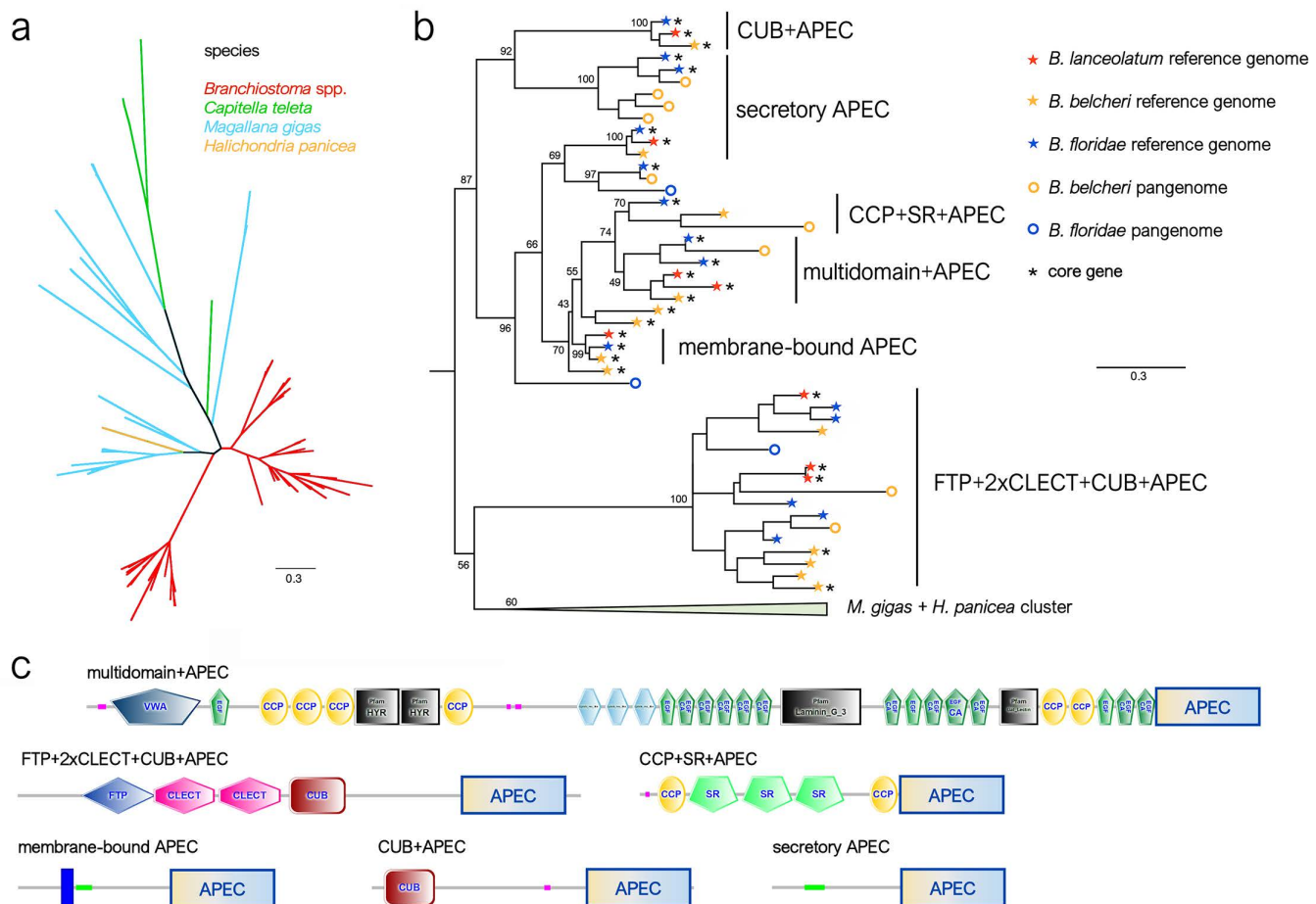


Fig 7. The APEC domain. **a.** Unrooted ML phylogeny of representative metazoan APEC family proteins. The lancelet proteins are highlighted in red. **b.** Detailed view of the phylogenetic relationship among lancelet APEC proteins. The bootstrap support values for the major nodes of the tree are shown, along with the core or dispensable status of the sequences and an indication of the domain architecture. **c.** Exemplified domain architecture of lancelet APEC-domain containing proteins.

<https://doi.org/10.1371/journal.pgen.1011833.g007>

evolutionary study, we could not detect any APEC domain containing proteins in either vertebrates or arthropods, and the sequences identified in the species selected for our comparative analyses were mostly grouped in lineage-specific clades, consistent with the previous report [40]. The lancelet APEC domain-containing proteins were grouped in two well-supported clades (Fig 7). The first clade included proteins carrying multiple carbohydrate recognition domains, namely, a fucoselectin and two consecutive C-type lectin domains at the N-terminus, followed by a CUB domain (which, despite not being directly involved in pathogen recognition, is frequently associated with complement factors), and an APEC domain (Fig 7c). Strikingly, most of the sequences belonging to this clade were dispensable. The second clade was highly heterogeneous and included both simple secretory proteins [41,42] and more complex multidomain proteins. The identification of complex multidomain APEC domain-containing proteins was consistent with previous reports from mollusks [43], even though the domain combinations identified in amphioxus were different from those of mollusks (Fig 7b).

Beside these statistically enriched domains, we have also considered certain genes of interest. We note that one of the six cellular RNA-directed RNA polymerase (RdRP) genes reported in *B. belcheri* was also identified as dispensable in 50% of the tested individuals. In the same species, the big defensin gene was absent in 77% of the individuals, whereas the two arthropod-like defensin genes were not subjected to PAV (Fig 4g).

Transgenerational dispensable gene transmission follows Mendelian inheritance rules

To gain insight into possible mechanisms of dispensable genes' transmission from parental genotypes to the offspring genomes, we analysed a dataset of 98 *B. floridae* samples, which included high coverage resequencing of two parents (namely M and F genotypes, ~60×) and low coverage resequencing of 96 offspring generated from these parents (~7×).

First, we investigated the M and F genotypes for the presence/absence status of the shell/core dispensable genes previously identified in *B. floridae* (Table 1). Using coverage cut-offs, we could establish the PAV status for 798 out of the 1,281 shell/core dispensable genes, whereas the remaining genes could not be reliably classified due to intermediate coverage values (S5 Fig).

A total of 282 genes annotated in the reference genome were absent in both parents (class I), whereas 311 genes found in hemizyosity in only one parental haplotype were missing in the other one (class II) and 137 genes were in hemizyosity in both F and M parents (class III). Moreover, 45 genes were present in only one parent (in association with both haplotypes) and 21 genes are present in one haplotype and in hemizyosity in the other one (Table 3). We calculated the fraction of genes missing in the offspring genomes for the three most abundant classes of genes, as well as for core genes. The finding that a median of 97% of the class I gene (i.e., genes absent in both parents, n=282) were confirmed as missing in offspring samples confirmed the reliability of this approach (Fig 8). The other classes of genes were observed in offspring samples with frequencies close to those expected by Mendelian inheritance. In particular, those found in hemizyosity in only one of the two parental genotypes (class II) were absent in 40% of the offspring (expected 50%) and those found in hemizyosity in both parental genomes (class III) were absent in 19% of the offspring (expected 25%, Fig 8). The slight discrepancies between observed and expected PAV rates could be explained by the fact that read mapping was carried out against the reference genome assembly, which most certainly does not include a full collection of all the dispensable genes associated with *B. floridae*, thereby leading to the cross-mapping of reads deriving from similar dispensable paralogs.

Discussion

It is becoming increasingly clear that multicellular eukaryotes, including animals, possess dynamic pangenomes, some of them with an architecture resembling that of the open pangenomes of bacteria [10,44]. Nevertheless, to date, animal pangenomes have been largely understudied, and the genomic mechanisms promoting large-scale structural variations,

Table 3. Presence-Absence status of the *B. floridae* shell/cloud dispensable genes in F and M haplotypes and in the genomes of the progenies. The analysis was carried out on 798 out of 1,281 shell/core dispensable genes previously identified in *B. floridae*. The gene class and its description in the parental genotypes, the expected offspring genotypes and absence percentages are reported together with the number of genes per category and the median of absence in offsprings.

Gene class	Parental genotypes		Expected offspring genotypes			Observed genes (No.)	Observed median absence rate in offspring
	F	M	++	+-	--(PAV)		
core genes	++	++	100%	/	/	2	/
I	--	--	/	/	100%	282	97.9%
II	+-	--	/	50%	50%	146	39.6%
	--	+-				165	
III	+-	+-	25%	50%	25%	137	17.7%
not tested	--	++	/	100%	/	27	/
	++	--				18	
not tested	+-	++	50%	50%	/	7	/
not tested	++	+-				14	
undetermined	/	/	/	/	/	420	/

<https://doi.org/10.1371/journal.pgen.1011833.t003>

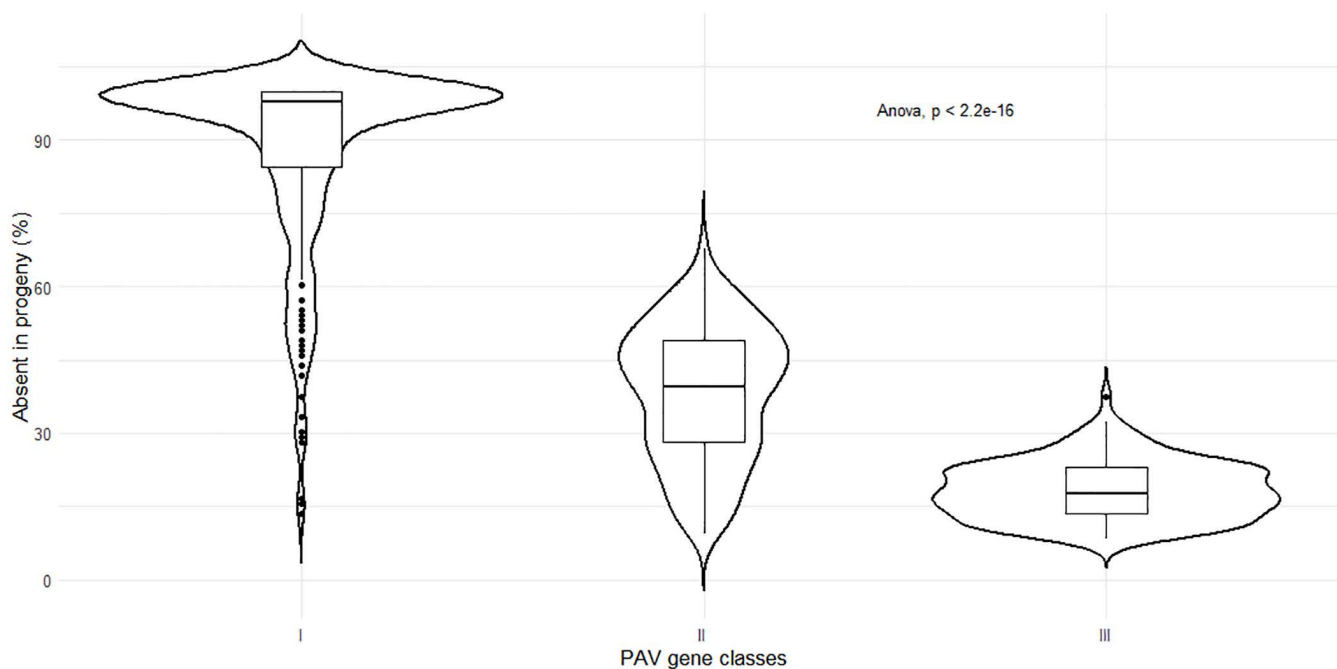


Fig 8. Transmission of PAV genes from parents to offsprings. The plot depicts the fractions of absent genes per sample (i.e., resequenced offspring) according to the four considered classes of *B. floridae* dispensable genes (see Table 2).

<https://doi.org/10.1371/journal.pgen.1011833.g008>

as well as the functional role played by the dispensable fraction of pangenomes are still debated [45,46]. Bivalves represent one of the best examples of animals with open bacterial-like pangenomes studied so far. In fact, following the first report of an open pangenome in the Mediterranean mussel [17], similar genomic organizations have been also reported in a number of other bivalve species [28,44]. Although still in their infancy, comparative pangenomic studies in metazoans indicate that the openness of vertebrate pangenomes is far lower than that of bivalves, which makes the investigation of pangenomes of early-branching chordate lineages particularly interesting in light of their evolutionary relatedness to vertebrates [31].

Surprisingly, our analysis of whole genome resequencing data revealed that hundreds of genes annotated in the *B. belcheri* and *B. floridae* reference genomes were dispensable, emphasizing the prevalence of structural variants subjected to PAV in lancelets. By analysing the genotypes of paired parent/offspring data in *B. belcheri*, we further demonstrated that the inheritance of lancelet dispensable genes conformed to Mendelian expectations, supporting the notion that, at least in this system, these genes behave as typical biallelic loci with presence/absence alleles. This result indicates that, despite theoretical possibilities for segregation distortion due to their association with structural variants, such effects may be minimal or absent in amphioxus. This would be consistent with the relatively small size of inferred indels associated with PAV in both lancelet species. The significant over-representation of Pfam domains and GO terms linked with intracellular pattern recognition, intracellular signalling, apoptotic and inflammatory processes in the dispensable fraction of lancelet pangenomes, is consistent with this interpretation. The strong representation of defence and immunity-related genes in lancelet pangenomes is further supported by a few key examples, such as the big defensin and eukaryotic RdRP gene families in *B. belcheri*. The single big defensin gene annotated in the reference genome, which encodes an antimicrobial peptide with broad-spectrum activity [47], was absent in 88% of *B. belcheri* samples, mirroring the findings previously reported for the same gene family in the Pacific oyster [48]. Although eukaryotic RdRPs (not to be confused with the non-homologous viral RdRPs), sporadically distributed in animals, are primarily known for their role in RNA interference,

this does not seem to be the case of lancelet orthologs [49], which may function in an as-yet undiscovered RNA-guided defence mechanism [50]. Strikingly, one of the six annotated RdRP genes of *B. belcheri* was absent in 50% of the resequenced individuals.

The fact that many of the dispensable genes are implicated in cellular defence and signalling is reminiscent of the “gene-accordion” of orthopoxviruses, a mechanism of rapid genome evolution through waves of gene duplication and diversification, followed by reduction in gene copy number, which allows these viruses to overcome host defences [51]. This process, apparently operating in most eukaryotic DNA viruses of the phylum *Nucleocytoviricota*, generates higher genomic entropy than simple SNPs, as reported in monkeypox virus [52,53]. Here we report that a similar “gene-accordion” operates in lancelets, with immunity-related genes undergoing dynamic gene copy number expansions, presumably in response to differential exposure to viral and/or bacterial pathogens. More broadly, the dynamic nature and high inter-individual variability of these genes appear to reflect a fundamental principle shared across all domains of life, from bacteria and archaea to eukaryotes, including cephalochordates, as demonstrated in this study. Indeed, prokaryotic immunity genes, including CRISPR-Cas systems, restriction-modification, Argonautes and many more, are usually part of the dispensable part of the pangenome, clustered in the hypervariable genomic loci [21,23]. Thus, in both prokaryotes and eukaryotes with open pangenomes, genomic innovations could be primarily stimulated by the interactions with diverse pathogens.

As instructive case studies, we focused our attention GIMAPs and APEC domain-containing proteins, both involved in defence mechanisms and subjected to extensive PAV in both lancelet species. Overall, we identified 21 novel GIMAP genes associated with dispensable NRSs in *B. floridae*, and 25 in *B. belcheri*. This would expand the complete repertoire of lancelet GIMAP genes to 38 and 33 in *B. belcheri* and *B. floridae*, respectively. Notably, GIMAPs have a complex evolutionary history, which has only very recently been investigated in a few metazoan phyla [37–39], with the overwhelming amount of functional information currently available only from human, where seven functional paralogous genes are present [54]. In humans, GIMAP genes appear to be mainly expressed in circulating immune cells, being crucial regulators of T-cell survival and differentiation [55,56], and their dysregulation is linked to several autoimmune diseases and cancer [57,58]. Although GIMAPs have been investigated in only a handful of other animal phyla, multiple lines of evidence support their involvement in immunity and stress tolerance. For example, GIMAPs have been hypothesized to play a key role in regulating hemocyte survival in oysters [59], which may be also supported by their strong association with PAV in mussels [17]. Another molluscan species, the snail *Biomphalaria glabrata*, has a largely expanded repertoire of GIMAPs, often expressed in the amebocyte-producing organ [60] and displaying marked differences in expression between individuals susceptible and resistant to *Schistosoma mansoni* infection [61]. The immune role of GIMAPs in animals has been proposed to be an ancestral feature, based on the evidence that these proteins retain their inducibility following PAMP stimulation in some cnidarian species [39,62].

APEC domain-containing proteins are a family of lectin-like secretory pattern recognition receptors (PRR), first described in *Branchiostoma japonicum* [41] and later also reported in *B. floridae*, where their involvement in microbial recognition, as well as in the modulation of immune response via the TRAF6/NF- κ B axis, were confirmed [42]. Interestingly, recent studies have demonstrated that APEC domain-containing proteins with different structural organizations can act as PRRs in mollusks and echinoderms [43,63–65], and the presence of genes encoding similar proteins in the genomes of other metazoan phyla suggests that the APEC domain may have been recruited on multiple occasions as a carbohydrate recognition domain during evolution [66].

Phylogenetic evidence and gene PAV pattern analyses of GIMAPs and APEC domain-containing proteins suggest that these two gene families underwent rapid evolution in lancelets, leading to a high molecular diversification and contributing to the creation of novel protein forms most likely involved in immune recognition and signaling. We showed that most of the dispensable genes of lancelets belong to species- or family-specific clusters or orthologs, suggesting that PAV mostly impacted the recently acquired taxonomically-restricted gene families, or families that underwent lineage-specific

expansions, a phenomenon previously reported also for mussels [17]. Importantly, dispensable genes were, on average, expressed at lower levels than core genes, matching the observation collected for duplicated genes in *B. lanceolatum* [34], thereby suggesting that they can play a tissue-specific role, as expected for rapidly-evolving, young genes. Several orthogroups exclusively included dispensable genes, highlighting their functional significance for the establishment of individual gene complements and possibly also in the generation novel multi-domain proteins. Notably, a comparative genomic analysis of multi-domain proteins has revealed that *B. floridae* displays the highest number of novel domain combinations among eukaryotes [67], an observation which can now be at least in part explained as the widespread occurrence of PAV in this species.

Our pangenomic investigations also revealed the presence of a complete herpes-like virus genome (called BelcheriHV-1) integrated in the reference genome of *B. belcheri* and shared within 20% of the resequenced samples. A similar partial viral genome, called BelcheriHV-2, was found to be associated with a single *B. belcheri* sample. The integration of malacoherpes-like virus genomes has been previously reported in both lancelet species, but so far lacked validation [68–71]. Notably, mammalian herpesviruses (family *Orthoherpesviridae*) also can integrate into host genomes, as shown for the endogenized herpesvirus-6 found in ~1% of the human population, where it occupies different chromosomal positions and is associated with distinct polymorphisms [72]. Here, based on the comparison of proviral and host genome coverages, we suggest that, whenever present, BelcheriHV-1 may be found in either one or two copies per genome and seems to be poorly impacted by SNPs suggesting a recent integration. Furthermore, the identification of BelcheriHV-2 suggests that these viruses actively circulate in the lancelet populations. It has been previously suggested that *Alloherpesviridae* and *Malacoherpesviridae* shared a common ancestor [73,74], and thus the lancelet herpesviruses described herein could represent the missing link between invertebrate and vertebrate herpesviruses.

Conclusions

Here, we reported the first pangenome analysis of lancelet species, revealing that the high heterozygosity typical of these species mirrors an extended open pangenome, with features resembling bacterial pangenomes. Although the high heterozygosity of these species and the lack of haplotype-resolved reference genome assemblies are limiting factors that could have, to some extent, decreased the sensitivity of our analyses and added biases related to inaccurate haplotype discrimination and/or overestimation of duplicated genomic regions [33,34], our investigation revealed the dynamic nature of the dispensable fraction of lancelet pangenomes. We demonstrated that dispensable protein-coding genes usually belong to highly variable multigene families involved in immune response and signalling processes. The proliferation of such genes appears to occur through duplication events, followed by the rapid molecular diversification of paralogs, yielding genomic regions with structural variation. Gene PAV can be seen as the readout of these processes or as a further mechanism to promote adaptation to changing environments and the exposure to pathogens. Further studies are needed to explore how specific combinations of dispensable genes may influence organismal fitness in lancelet species at both individual and population levels.

Materials and methods

Data download and preliminary analyses

The reference genome assemblies of *B. floridae* (GCF_000003815.2) and *B. belcheri* (GCF_001625305.1) and their associated gene annotations were recovered from NCBI on December 1st 2023. Available whole genome resequencing data and related metadata of *B. floridae* (N=41, plus N=98 parental-offspring sequencing datasets) and *B. belcheri* (N=100) were downloaded from the NCBI SRA database using the *srahunter* tool [75] (S1 Table). However, upon further inspection, one of the *B. belcheri* samples (i.e., SRR12010277) was identified as a misclassified *B. floridae* individual and thereby discarded, reducing the total number of suitable samples for this species to 99. A total of 91 and 61 RNA sequencing datasets of *B. belcheri* and *B. floridae* were also downloaded from the SRA archive (S1 Table). Raw reads were trimmed for quality and to remove adapter sequences using *fastp* v0.23.1 [76] with the following parameters: -V

--detect_adapter_for_pe -x -g -n 2 -5 -3 -p -l 75. The lancelet genomes were subject to a repeat prediction/classification and soft-masking step using *RepeatModeler* v.2.0.5 and *RepeatMasker* v4.1.2 (<http://www.repeatmasker.org/RepeatMasker/>). The quality-trimmed DNA reads were mapped on the corresponding reference genome using *bwa mem* [77] allowing read multi-mapping and the resulting files were used for both SNP identification and for the subsequent PAV analysis. To identify SNPs, the mapping files were refined by removing duplicated reads, reads with more than 10 variants per 100-nt windows, and unaligned read ends (> 20 nt) using the tools implemented in the CLC Genomics Workbench v.23 suite (Qiagen, Hilden, Germany). SNP were called applying a minimal frequency of 35%, a minimum coverage of 10×, and subsequently annotated with overlapping information of genes and CDS to predict their functional consequences.

Identification of dispensable regions and genes in *B. floridae* and *B. belcheri* genomes

To identify dispensable regions and genes in resequenced lancelet genomes we applied a previously developed PAV detection pipeline (https://github.com/Carmen-Tuc/PAV_pipeline) [78], with minor modifications. The average coverage of target regions (i.e., non-overlapping windows of genomic sequence spanning 5kb) or genes (only considering exons), was calculated using *samtools depth* [79] by analyzing the bam mapping files obtained from each sample. Observed coverage values were compared to those expected for core regions, inferred from the analysis of single-copy complete Benchmarking Universal Single-Copy Orthologs (BUSCO) genes, identified using BUSCO v5.2.2 [80] with the metazoa_odb10 dataset. Absent regions or genes were defined as those with coverage below one-eighth of the expected value. An arbitrary threshold, set at 10% of the samples analysed for each of the two species, was applied to identify shell and cloud dispensable genes and regions to be subject to further analysis, thereby excluding softcore elements present in the majority (i.e., ≥ 90%) of individuals. The percentage of mapped reads per sample was determined with *samtools stats* and used to detect possible outliers. The presence/absence matrix of dispensable regions was used to run a Principal Component Analysis (PCA) using the R packages *FactoMineR*, *ggcorrplot*, *corr* and *factoextra*, and to build a heatmap with the *heatmaply* package. *B. belcheri* and *B. floridae* proteomes were functionally annotated using *InterProScan* v.5.57-90.0 [81]. An enrichment analysis was computed with a hypergeometric test based on the Pfam domains and on the GO terms associated with shell and cloud genes using an in-house python script available at https://gitlab.com/54mu/enrichment_test [82]. Annotations were considered significantly enriched if they exhibited an observed-to-expected count ratio greater than 4 for Pfam domains or greater than 2 for GO terms, and if the associated False Discovery Rate (FDR)-corrected P-value [83] was lower than 0.01.

The genomic distribution of dispensable genes was further inspected in *B. floridae*, by calculating the density of dispensable genes per Mb and graphically representing the data with *Circos* [84]. To assess the correlation between the proximity between dispensable pairs and their PAV patterns, Wilcoxon rank-sum tests were used to evaluate statistically significant differences (i.e., FDR-corrected p-value < 0.05) between intra- and inter-chromosomal gene pairs (with the latter used as a background, due to the expected lack of correlation), computed as the average value for 5 Kb-long intervals of distance.

In addition, the possible association between dispensable regions and flanking repeated elements was investigated by comparing the repeat content in the upstream and downstream regions of each dispensable gene block (obtained from the libraries generated by *RepeatModeler*) to that of all core genes. The significant enrichment (i.e., FDR-corrected p-value < 0.05) for each category of repeats found in upstream and downstream flanking regions (spanning 5, 10 or 20 Kb of genomic sequence) was assessed with a hypergeometric test.

The *B. floridae* parental-offspring sequencing dataset was further analyzed to determine the heterozygosity or homozygosity status of shell/cloud dispensable genes in the female (F) and male (M) parental genotypes. Coverage-based thresholds were established using the average read depth of genes not affected by presence/absence variation (PAV). Genes in hemizygous state were defined as those with coverage within ±50% of half the average depth (i.e., 15–45× for the F genotype and 14–42× for the M genotype). All statistical analyses, data manipulation and visualisation steps were performed in R 4.2.3 [85] using the *tidyverse*, *ggplot2*, *ggpubr*, *ggrepel*, *VennDiagram*, *UpSetR*, *smplot2* and *data.table* packages. The script is available at https://github.com/umberos/Lancelet_pangenome.

Analysis of the de-novo assembled pangenome and of associated non-reference sequences

For each resequenced lancelet individual, the DNA reads that could not be mapped on the reference genomes were recovered from the mapping files using *samtools* and de novo assembled with *SPAdes v3.15.5* [86], applying the *–isolate* flag. Assembled non-redundant contigs longer than 2.5kb (obtained with *cd-hit-est* 0.8 [87]) were considered as NRSs and were therefore subject to gene prediction using two different approaches, namely *Prodigal v2.6.3*, run in metagenomic mode, and *BRAKER v3.0.8* [88], using the *B. floridae* gene models as training set. The non-redundant coding sequences predicted by *BRAKER3* were translated into proteins, subjected to another round of redundancy removal (*cd-hit* 0.8), and taxonomically classified using *diamond v. 2.0.6 blastp* searches against the NCBI *nr* database [89], only keeping those assigned to *Metazoa* for subsequent analyses. Viral hits among the assembled contigs were detected using *geNomad v1.7.4* [90], based on *prodigal* gene predictions and viral marker protein profiles. The hits showing relevant viral matches were extracted and further analysed. *Blastp* searches using the relevant translated ORFs as queries were performed against the NCBI *nr* database (downloaded February 2023) to assign a best match to each hit.

Gene expression analysis

Quality-trimmed RNA reads were mapped on the corresponding reference genomes using the *CLC mapper* of the CLC Genomics Workbench, applying the following parameters: mismatch cost=2; insertion cost=3; deletion cost=3; length fraction=0.8; similarity fraction=0.8 and expression values were counted as Transcripts Per Million (TPM). The computed expression levels, as well as the fraction of samples without detectable transcription, were compared between the sets of core and shell/cloud genes.

Gene orthology analysis

The proteomes of 11 species including the vertebrates *Homo sapiens*, *Danio rerio* and *Gallus gallus*, the ascidian *Clavelina lepadiformis*, the invertebrates *Crassostrea gigas*, *Capitella teleta*, *Apis mellifera* and *Halichondria panicea*, three lancelet species (*B. belcheri*, *B. floridae* and *B. lanceolatum*) and the proteins encoded by the de novo assembled NRSs of lancelets were processed to identify ortholog genes using *orthofinder v2.5.4*, with *mafft* and *FastTree* used to infer maximum likelihood trees from multiple sequence alignments [91]. The orthogroups that included lancelet dispensable genes were identified and the number and origin of orthologs in these clusters were counted. To investigate specific domains, all protein sequences derived from the annotation of the three lancelet reference genomes that included at least one AIG1 (PF04548) or APEC (PF16977) domain were extracted, together with the homologous sequences placed in the same orthogroups from other selected metazoan species. Due to the highly variable architectures of the protein sequences, only the region comprising the characterizing domain was selected for phylogenetic inference analysis. In detail, all accessory portions of sequences present either at the N-terminal or at the C-terminal end of the target domains were removed, and only sequences covering at least 50% of the expected length were kept for subsequent analysis. Whenever multiple domains were detected, all of them were extracted and added to the list. Trimmed sequences were aligned with *MUSCLE* [92] and, upon manual curation, which removed all the positions of the alignment with gaps in > 50% entries, the multiple sequence alignments were subjected to maximum likelihood phylogenetic analyses with *Iq-Tree* [93]. The best-fitting molecular models of evolution, according to the corrected Bayesian information criterion, were WAG+F+R5 and WAG+F+R3 for the AIG1 and APEC datasets, respectively. The reliability of the trees was evaluated using an ultrafast bootstrap test with 1000 replicates.

Analysis of the integrated BelcheriHV-1 genome

To confirm the presence of the BelcheriHV-1 genome in the 20 samples showing the presence of the 35 contiguous viral genes, *de novo* assemblies were performed for each sample, and the resulting contigs were subsequently mapped to the *B. belcheri* reference genome using the *large gap mapper* tool implemented in CLC Genomics Workbench. The mapping files were manually inspected to determine the coverage of the BelcheriHV-1 genome.

Supporting information

S1 Table. Metadata of the datasets analysed for this study. The list included the NCBI IDs and relevant metadata for the genomic and transcriptomic datasets analysed in this study.
(XLSX)

S2 Table. Size distribution of the blocks of contiguous shell/cloud dispensable genes in *B. belcheri* and *B. floridae*. *This gene block corresponds to the integrated BelcheriHV-1 genome.
(XLSX)

S3 Table. Significantly enriched (hypergeometric test, FDR-corrected p-value) repeated elements associated with shell/core gene blocks, compared with core genes. Enrichments tests were carried out for flanking regions of different sizes (i.e., 5, 10 and 20 Kb).
(XLSX)

S4 Table. Annotation file of the BelcheriHV-1 and BelcheriHV-2 predicted proteomes. The table reported the *blastp* best hit against the NCBI non-redundant protein database (nr) and against the SWISS-prot database.
(XLSX)

S5 Table. Enrichment test performed for *B. belcheri* and *B. floridae* shell/cloud dispensable genes. For both Pfam domains and GO terms, the ID and description, the total counts in the predicted proteome, the observed counts among dispensable genes, the FDR p-value and the enrichment value are reported. For GO terms, the classification into biological and molecular functions are also reported.
(XLSX)

S1 Fig. Correlation between gene dispensability and the number of observed PAV occurrences among the resequenced samples of *B. belcheri* (a) and *B. floridae* (b).
(TIFF)

S2 Fig. Average concordance ($\pm$ SD) between the presence/absence status of intrachromosomal pairs of dispensable genes in *B. belcheri*, for bins of pairwise distance ranging 5Kb each. Statistically significant differences (Wilcoxon rank-sum tests, FDR-corrected p-value < 0.05) compared with inter-chromosomal dispensable gene pairs (dashed red line) are highlighted by an asterisk. Note that the genes included in the integrated BelcheriHV-1 were omitted.
(TIF)

S3 Fig. Average concordance ($\pm$ SD) between the presence/absence status of intrachromosomal pairs of dispensable genes in *B. floridae*, for bins of pairwise distance ranging 5Kb each. Statistically significant differences (Wilcoxon rank-sum tests, FDR-corrected p-value < 0.05) compared with inter-chromosomal dispensable gene pairs (dashed red line) are highlighted by an asterisk.
(TIF)

S4 Fig. Principal Component Analysis and sample correlation plots based on the presence-absence patterns of shell/cloud dispensable genes for *B. belcheri* (a-b) and *B. floridae* (c-d). The *B. belcheri* samples in the PCA are coloured according to the presence of the integrated BelcheriHV-1 genome (YES: blue, NO: red). Instead, the *B. floridae* samples in the PCA are coloured by experiment, with the blue ones referring to the parental + offspring dataset.
(TIF)

S5 Fig. Genome coverage plot of the female (a) and male (c) genotypes used to produce the offsprings. The orange dotted lines indicated the average coverage computed based on core genes. Gene coverage plots for the

dispensable genes in the female (b) and male (d) genotypes. The pink dotted lines indicate the cut-off used to determine the absence of a given gene, whereas the two blue lines indicate the range used to determine the genes in hemizyosity, and the orange lines refer to the expected coverages of core genes.

(TIFF)

S1 File. The data matrix reporting the presence-absence of *B. belcheri* genes in the tested DNA samples. “1” refers to an absent gene, “0” to a present gene.

(CSV)

S2 File. The data matrix reporting the presence-absence of *B. floridae* genes in the tested DNA samples. “1” refers to an absent gene, “0” to a present gene. The matrix included also the dataset of the offspring, for a total of 140 samples.

(CSV)

S3 File. BelcheriHV-1 genome (fasta format).

(FA)

S4 File. BelcheriHV-2 genome (fasta format).

(FA)

Acknowledgments

Computational resources were provided by the University of Padova Strategic Research Infrastructure Grant 2017: *CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione*.

Author contributions

Conceptualization: Umberto Rosani, Marco Gerdol, Mart Krupovic.

Data curation: Umberto Rosani.

Formal analysis: Umberto Rosani.

Funding acquisition: Umberto Rosani, Marco Gerdol.

Methodology: Umberto Rosani, Marco Gerdol.

Supervision: Marco Gerdol, Mart Krupovic.

Writing – original draft: Umberto Rosani.

Writing – review & editing: Umberto Rosani, Marco Gerdol, Mart Krupovic.

References

1. Gaggiotti OE, Chao A, Peres-Neto P, Chiu C, Edwards C, Fortin M, et al. Diversity from genes to ecosystems: A unifying framework to study variation across biological metrics and scales. *Evol Appl*. 2018;11(7):1176–93.
2. Ramel C. Biodiversity and intraspecific genetic variation. *Pure Appl Chem*. 1998;70(11):2079–84. <https://doi.org/10.1351/pac199870112079>
3. Farnsworth KD, Albantakis L, Caruso T. Unifying concepts of biological function from molecules to ecosystems. *Oikos*. 2017;126(10):1367–76. <https://doi.org/10.1111/oik.04171>
4. Tets VV. Pangenome. *Tsitologiia*. 2003;45(5):526–31.
5. Tetz VV. The pangenome concept: a unifying view of genetic information. *Med Sci Monit*. 2005;11(7):HY24–9. PMID: [15990697](https://pubmed.ncbi.nlm.nih.gov/15990697/)
6. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev*. 2005;15(6):589–94. <https://doi.org/10.1016/j.gde.2005.09.006> PMID: [16185861](https://pubmed.ncbi.nlm.nih.gov/16185861/)
7. Yocca AE, Platts A, Alger E, Teresi S, Mengist MF, Benevenuto J, et al. Blueberry and cranberry pangenomes as a resource for future genetic studies and breeding efforts. *bioRxiv [Internet]*. 2023 Aug 2 [cited 2024 Oct 23];2023.07.31.551392. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10418200/>

8. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* 2008;11(5):472–7. <https://doi.org/10.1016/j.mib.2008.09.006> PMID: [19086349](#)
9. Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci U S A.* 2005;102(39):13950–5. <https://doi.org/10.1073/pnas.0506758102> PMID: [16172379](#)
10. Richard GF. Eukaryotic pangenomes. In: Tettelin H, Medini D, editors. *The pangenome: diversity, dynamics and evolution of genomes.* Cham (CH): Springer. 2020.
11. Morgante M, De Paoli E, Radovic S. Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol.* 2007;10(2):149–55. <https://doi.org/10.1016/j.pbi.2007.02.001> PMID: [17300983](#)
12. Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D’Auria G, Zhaxybayeva O, Papke RT, et al. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* 2007;1(3):235–45. <https://doi.org/10.1038/ismej.2007.35> PMID: [18043634](#)
13. Colson P, Raoult D. Gene repertoire of amoeba-associated giant viruses. *Intervirology.* 2010;53(5):330–43. <https://doi.org/10.1159/000312918> PMID: [20551685](#)
14. McCarthy CGP, Fitzpatrick DA. Pan-genome analyses of model fungal species. *Microb Genom.* 2019;5(2):e000243. <https://doi.org/10.1099/mgen.0.000243> PMID: [30714895](#)
15. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet.* 2019;51(1):30–5.
16. Li R, Fu W, Su R, Tian X, Du D, Zhao Y, et al. Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. *Front Genet.* 2019;10:1169. <https://doi.org/10.3389/fgene.2019.01169> PMID: [31803240](#)
17. Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol.* 2020;21(1):275. <https://doi.org/10.1186/s13059-020-02180-3> PMID: [33168033](#)
18. Romero-Calle DX, Pedrosa-Silva F, Tomé LMR, Sousa TJ, de Oliveira Santos LTS, de Carvalho Azevedo VA, et al. Hybrid Genomic Analysis of *Salmonella enterica* Serovar Enteritidis SE3 Isolated from Polluted Soil in Brazil. *Microorganisms.* 2022;11(1):111. <https://doi.org/10.3390/microor-ganisms11010111> PMID: [36677403](#)
19. Hu J, Chen L, Li G, Pan Y, Lu Y, Chen J, et al. Prevalence and genetic characteristics of *fosB*-positive *Staphylococcus aureus* in duck farms in Guangdong, China in 2020. *J Antimicrob Chemother.* 2023;78(3):802–9.
20. Lawrence JG, Retchless AC. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods Mol Biol.* 2009;532:29–53. https://doi.org/10.1007/978-1-60327-853-9_3 PMID: [19271178](#)
21. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science.* 2018;359(6379):eaar4120. <https://doi.org/10.1126/science.aar4120>
22. Hochhauser D, Millman A, Sorek R. The defense island repertoire of the *Escherichia coli* pan-genome. *PLoS Genet.* 2023;19(4):e1010694. <https://doi.org/10.1371/journal.pgen.1010694> PMID: [37023146](#)
23. Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science.* 2021;374(6566):488–92.
24. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* 2005;33(14):4626–38. <https://doi.org/10.1093/nar/gki775> PMID: [16106042](#)
25. Sibbald SJ, Eme L, Archibald JM, Roger AJ. Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends Parasitol.* 2020;36(11):927–41. <https://doi.org/10.1016/j.pt.2020.07.014> PMID: [32828660](#)
26. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet.* 2012;13(4):283–96. <https://doi.org/10.1038/nrg3199> PMID: [22421730](#)
27. Zhang L, Reifová R, Halenková Z, Gompert Z. How Important Are Structural Variants for Speciation?. *Genes (Basel).* 2021;12(7):1084. <https://doi.org/10.3390/genes12071084> PMID: [34356100](#)
28. Liu S, Shi C, Chen C, Tan Y, Tian Y, Macqueen DJ, et al. Haplotype-resolved genomes provide insights into the origins and functional significance of genome diversity in bivalves. *Cell Rep.* 2025;44(5):115697. <https://doi.org/10.1016/j.celrep.2025.115697> PMID: [40349337](#)
29. Wu Z, Li T, Jiang Z, Zheng J, Gu Y, Liu Y, et al. Human pangenome analysis of sequences missing from the reference genome reveals their widespread evolutionary, phenotypic, and functional roles. *Nucleic Acids Res.* 2024;52(5):2212–30. <https://doi.org/10.1093/nar/gkae086> PMID: [38364871](#)
30. Holland LZ, Holland ND. The invertebrate chordate amphioxus gives clues to vertebrate origins. *Curr Top Dev Biol.* 2022;147:563–94. <https://doi.org/10.1016/bs.ctdb.2021.12.011> PMID: [35337463](#)
31. D’Aniello S, Bertrand S, Escriva H. Amphioxus as a model to study the evolution of development in chordates. *Elife.* 2023;12:e87028. <https://doi.org/10.7554/eLife.87028> PMID: [37721204](#)
32. Zhu Y, Lu N, Chen J-Y, He C, Huang Z, Lu Z. Deep whole-genome resequencing sheds light on the distribution and effect of amphioxus SNPs. *BMC Genom Data.* 2022;23(1):26. <https://doi.org/10.1186/s12863-022-01038-w> PMID: [35395709](#)

33. Huang Z, Xu L, Cai C, Zhou Y, Liu J, Xu Z, et al. Three amphioxus reference genomes reveal gene and chromosome evolution of chordates. *Proc Natl Acad Sci U S A*. 2023;120(10):e2201504120. <https://doi.org/10.1073/pnas.2201504120> PMID: 36867684
34. Brasó-Vives M, Marlétaz F, Echchiki A, Mantica F, Acemel RD, Gómez-Skarmeta JL, et al. Parallel evolution of amphioxus and vertebrate small-scale gene duplications. *Genome Biol*. 2022;23:243.
35. Rosani U, Sollitto M, Fogal N, Salata C. Comparative analysis of Presence-Absence gene Variations in five hard tick species: impact and functional considerations. *Int J Parasitol*. 2024;54(3–4):147–56. <https://doi.org/10.1016/j.ijpara.2023.08.004> PMID: 37806426
36. Leippe DD, Wolf YI, Koonin EV, Aravind L. Classification and evolution of P-loop GTPases and related ATPases. *J Mol Biol*. 2002;317(1):41–72. <https://doi.org/10.1006/jmbi.2001.5378> PMID: 11916378
37. Coelho JC, Calhoun ED, Calhoun GN, Poole AZ. Patchy Distribution of GTPases of Immunity-Associated Proteins (GIMAP) within Cnidarians and Dinoflagellates Suggests a Complex Evolutionary History. *Genome Biol Evol*. 2022;14(2):evac002. <https://doi.org/10.1093/gbe/evac002> PMID: 35015849
38. Weinstein DJ, Allen SE, Lau MCY, Erasmus M, Asalone KC, Walters-Conte K, et al. The genome of a subterrestrial nematode reveals adaptations to heat. *Nat Commun*. 2019;10(1):5268.
39. Weiss Y, Forêt S, Hayward DC, Ainsworth T, King R, Ball EE, et al. The acute transcriptional response of the coral *Acropora millepora* to immune challenge: expression of GiMAP/IAN genes links the innate immune responses of corals with those of mammals and plants. *BMC Genomics*. 2013;14:400. <https://doi.org/10.1186/1471-2164-14-400> PMID: 23768317
40. Li Y, Li J, Yan X, Chen S, Wu C, Huang H, et al. Broad distribution, high diversity and ancient origin of the ApeC-containing proteins. *Mol Phylogenet Evol*. 2021;155:107009. <https://doi.org/10.1016/j.ympev.2020.107009> PMID: 33186688
41. Huang G, Huang S, Yan X, Yang P, Li J, Xu W, et al. Two apextrin-like proteins mediate extracellular and intracellular bacterial recognition in amphioxus. *Proc Natl Acad Sci U S A*. 2014;111(37):13469–74. <https://doi.org/10.1073/pnas.1405414111> PMID: 25187559
42. Li J, Li Y, Fan Z, Chen S, Yan X, Yue Z, et al. Two amphioxus apeC-containing proteins bind to microbes and inhibit the TRAF6 pathway. *Front Immunol*. 2021.
43. Li J, Liu S, Zhang Y, Huang Q, Zhang H, OuYang J, et al. Two novel mollusk short-form ApeC-containing proteins act as pattern recognition proteins for peptidoglycan. *Front Immunol*. 2022;13:971883. <https://doi.org/10.3389/fimmu.2022.971883> PMID: 36275759
44. Calcino AD, Kenny NJ, Gerdol M. Single individual structural variant detection uncovers widespread hemizyosity in molluscs. *Philos Trans R Soc Lond B Biol Sci*. 2021;376(1825):20200153.
45. Miao J, Wei X, Cao C, Sun J, Xu Y, Zhang Z, et al. Pig pangenome graph reveals functional features of non-reference sequences. *J Anim Sci Biotechnol*. 2024;15(1):32. <https://doi.org/10.1186/s40104-023-00984-4> PMID: 38389084
46. Abondio P, Cilli E, Luiselli D. Human pangenomics: promises and challenges of a distributed genomic reference. *Life (Basel)*. 2023;13(6):1360.
47. Teng L, Gao B, Zhang S. The first chordate big defensin: identification, expression and bioactivity. *Fish Shellfish Immunol*. 2012;32(4):572–7.
48. Rosa RD, Alonso P, Santini A, Vergnes A, Bachère E. High polymorphism in big defensin gene expression reveals presence-absence gene variability (PAV) in the oyster *Crassostrea gigas*. *Dev Comp Immunol*. 2015;49(2):231–8. <https://doi.org/10.1016/j.dci.2014.12.002> PMID: 25482648
49. Pinzón N, Bertrand S, Subirana L, Busseau I, Escrivá H, Seitz H. Functional lability of RNA-dependent RNA polymerases in animals. *PLoS Genet*. 2019;15(2):e1007915. <https://doi.org/10.1371/journal.pgen.1007915> PMID: 30779744
50. Koonin EV, Krupovic M. New faces of prokaryotic mobile genetic elements: guide RNAs link transposition with host defense mechanisms. *Curr Opin Syst Biol*. 2023;36:100473. <https://doi.org/10.1016/j.coisb.2023.100473> PMID: 37779558
51. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, et al. Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses. *Cell*. 2012;150(4):831–41.
52. Filée J. Route of NCLDV evolution: the genomic accordion. *Curr Opin Virol*. 2013;3(5):595–9. <https://doi.org/10.1016/j.coviro.2013.07.003> PMID: 23896278
53. Monzón S, Varona S, Negredo A, Vidal-Freire S, Patiño-Galindo JA, Ferressini-Gerpe N, et al. Monkeypox virus genomic accordion strategies. *Nat Commun*. 2024;15(1):3059. <https://doi.org/10.1038/s41467-024-46949-7> PMID: 38637500
54. Krücken J, Schroetel RMU, Müller IU, Saïdani N, Marinovski P, Benten WPM, et al. Comparative analysis of the human gimap gene cluster encoding a novel GTPase family. *Gene*. 2004;341:291–304.
55. Ciucci T, Bosselut R. Gimap and T cells: A matter of life or death. *Eur J Immunol*. 2014;44(2):348–51.
56. Nitta T, Takahama Y. The lymphocyte guard-IANs: regulation of lymphocyte survival by IAN/GIMAP family proteins. *Trends Immunol*. 2007;28(2):58–65. <https://doi.org/10.1016/j.it.2006.12.002> PMID: 17196432
57. Filén S, Lahesmaa R. GIMAP Proteins in T-Lymphocytes. *J Signal Transduct*. 2010;2010:268589. <https://doi.org/10.1155/2010/268589> PMID: 21637352
58. Heinonen MT, Laine A-P, Söderhäll C, Gruziova O, Rautio S, Melén E, et al. GIMAP GTPase family genes: potential modifiers in autoimmune diabetes, asthma, and allergy. *J Immunol*. 2015;194(12):5885–94. <https://doi.org/10.4049/jimmunol.1500016> PMID: 25964488

59. McDowell IC, Nikapitiya C, Aguiar D, Lane CE, Istrail S, Gomez-Chiarri M. Transcriptome of American Oysters, *Crassostrea virginica*, in Response to Bacterial Challenge: Insights into Potential Mechanisms of Disease Resistance. *PLoS ONE*. 2014;9(8):e105097. <https://doi.org/10.1371/journal.pone.0105097>
60. Zhang S-M, Loker ES, Sullivan JT. Pathogen-associated molecular patterns activate expression of genes involved in cell proliferation, immunity and detoxification in the amebocyte-producing organ of the snail *Biomphalaria glabrata*. *Dev Comp Immunol*. 2016;56:25–36. <https://doi.org/10.1016/j.dci.2015.11.008> PMID: 26592964
61. Lu L, Loker ES, Zhang S-M, Buddenborg SK, Bu L. Genome-wide discovery, and computational and transcriptional characterization of an AIG gene family in the freshwater snail *Biomphalaria glabrata*, a vector for *Schistosoma mansoni*. *BMC Genomics*. 2020;21(1):190. <https://doi.org/10.1186/s12864-020-6534-z> PMID: 32122294
62. Bailey GF, Coelho JC, Poole AZ. Differential expression of *Exaipstasia pallida* GIMAP genes upon induction of apoptosis and autophagy suggests a potential role in cnidarian symbiosis and disease. *J Exp Biol*. 2020;223(Pt 21):jeb229906. <https://doi.org/10.1242/jeb.229906> PMID: 32978315
63. Jiang K, Nie H, Yin Z, Yan X, Li Q. Apextrin from *Ruditapes philippinarum* functions as pattern recognition receptor and modulates NF- κ B pathway. *Int J Biol Macromol*. 2022;214:33–44. <https://doi.org/10.1016/j.ijbiomac.2022.06.041> PMID: 35697169
64. Jiang K, Yin Z, Zhang Y, Xu Q, Yu Y, Cong W, et al. Genome-wide investigation and expression analysis of MACPF gene family reveals its immune role in response to bacterial challenge of Manila clam. *Genomics*. 2021;113(3):1136–45. <https://doi.org/10.1016/j.ygeno.2021.02.013> PMID: 33639237
65. Xiang Y, Shi Y, Sun L, Liang W, Chen K, Li C. Novel ApeC-containing protein mediates the recognition and internalization of *Vibrio splendidus* in *Apostichopus japonicus*. *Int J Biol Macromol*. 2024;275(Pt 2):133737. <https://doi.org/10.1016/j.ijbiomac.2024.133737> PMID: 38986992
66. Gerdol M, Luo Y-J, Satoh N, Pallavicini A. Genetic and molecular basis of the immune system in the brachiopod *Lingula anatina*. *Dev Comp Immunol*. 2018;82:7–30. <https://doi.org/10.1016/j.dci.2017.12.021> PMID: 29278680
67. Zmasek CM, Godzik A. This Déjà vu feeling—analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Comput Biol*. 2012;8(11):e1002701. <https://doi.org/10.1371/journal.pcbi.1002701> PMID: 23166479
68. Savin KW, Cocks BG, Wong F, Sawbridge T, Cogan N, Savage D, et al. A neurotropic herpesvirus infecting the gastropod, abalone, shares ancestry with oyster herpesvirus and a herpesvirus associated with the amphioxus genome. *Virology*. 2010;7:308. <https://doi.org/10.1186/1743-422X-7-308> PMID: 21062506
69. Rosani U, Venier P. Oyster RNA-seq Data Support the Development of Malacoherpesviridae Genomics. *Front Microbiol*. 2017;8:1515. <https://doi.org/10.3389/fmicb.2017.01515> PMID: 28848525
70. Mushegian A, Karin EL, Pupko T. Sequence analysis of malacoherpesvirus proteins: Pan-herpesvirus capsid module and replication enzymes with an ancient connection to “Megavirales”. *Virology*. 2018;513:114–28. <https://doi.org/10.1016/j.virol.2017.10.009> PMID: 29065352
71. Rosani U, Gaia M, Delmont TO, Krupovic M. Tracing the invertebrate herpesviruses in the global sequence datasets. *Front Mar Sci*. 2023;10:1159754. <https://doi.org/10.3389/fmars.2023.1159754>
72. Liu X, Kosugi S, Koide R, Kawamura Y, Ito J, Miura H, et al. Endogenization and excision of human herpesvirus 6 in human genomes. *PLoS Genet*. 2020;16(8):e1008915. <https://doi.org/10.1371/journal.pgen.1008915> PMID: 32776928
73. Brito AF, Baele G, Nahata KD, Grubaugh ND, Pinney JW. Intrahost speciations and host switches played an important role in the evolution of herpesviruses. *Virus Evol*. 2021;7(1):veab025. <https://doi.org/10.1093/ve/veab025> PMID: 33927887
74. Davison AJ, Trus BL, Cheng N, Steven AC, Watson MS, Cunningham C, et al. A novel class of herpesvirus with bivalve hosts. *J Gen Virol*. 2005;86(1):41–53.
75. Bortoletto E, Frizzo R, Rosani U, Venier P. sraHunter: A User-Friendly Tool for Efficient Retrieval and Management of SRA Data [Internet]. *bioRxiv*; 2024 [cited 2024 Mar 30]. 2024.03.19.585745 p. Available from: <https://www.biorxiv.org/content/10.1101/2024.03.19.585745v1>
76. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–90.
77. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324> PMID: 19451168
78. Sollitto M, Kenny NJ, Greco S, Tucci CF, Calcino AD, Gerdol M. Detecting Structural Variants and Associated Gene Presence-Absence Variation Phenomena in the Genomes of Marine Organisms. *Methods Mol Biol*. 2022;2498:53–76. https://doi.org/10.1007/978-1-0716-2313-8_4 PMID: 35727540
79. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943
80. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210–2. <https://doi.org/10.1093/bioinformatics/btv351> PMID: 26059717
81. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236–40. <https://doi.org/10.1093/bioinformatics/btu031> PMID: 24451626
82. Falcon S, Gentleman R. Hypergeometric Testing Used for Gene Set Enrichment Analysis. In: Hahne F, Huber W, Gentleman R, Falcon S, editors. *Bioconductor Case Studies* [Internet]. New York (NY): Springer; 2008 [cited 2022 Jul 22]. p. 207–20. (Use R!). Available from: https://doi.org/10.1007/978-0-387-77240-0_14

83. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1995;57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
84. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. *Genome Res* [Internet]. 2009 Jun 18 [cited 2025 May 18]; Available from: <http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109>
85. Giorgi FM, Ceraolo C, Mercatelli D. The R Language: An Engine for Bioinformatics and Data Science. *Life (Basel)*. 2022;12(5):648. <https://doi.org/10.3390/life12050648> PMID: [35629316](#)
86. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. 2017;27(5):824–34. <https://doi.org/10.1101/gr.213959.116> PMID: [28298430](#)
87. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150–2.
88. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 2008;24(5):637–44.
89. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59–60. <https://doi.org/10.1038/nmeth.3176> PMID: [25402007](#)
90. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* [Internet]. 2023 Sep 21 [cited 2024 Jan 15];1–10. Available from: <https://www.nature.com/articles/s41587-023-01953-y>
91. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):238. <https://doi.org/10.1186/s13059-019-1832-y> PMID: [31727128](#)
92. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7. <https://doi.org/10.1093/nar/gkh340> PMID: [15034147](#)
93. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*. 2020;37(5):1530–4. <https://doi.org/10.1093/molbev/msaa015> PMID: [32011700](#)