marine drugs

Article

Extensive Tandem Duplication Events Drive the Expansion of the C1q-Domain-Containing Gene Family in Bivalves

Marco Gerdol [1,*], Samuele Greco [1] and Alberto Pallavicini [1,2]

[1] Department of Life Sciences, University of Trieste, 34127 Trieste, Italy; SAMUELE.GRECO@phd.units.it (S.G.); pallavic@units.it (A.P.)

[2] National Institute of Oceanography and Applied Geophysics, 34151 Trieste, Italy

* Correspondence: mgerdol@units.it; Tel.: +39-040-5588676

Received: 30 August 2019; Accepted: 12 October 2019; Published: 14 October 2019

Abstract: C1q-domain-containing (C1qDC) proteins are rapidly emerging as key players in the innate immune response of bivalve mollusks. Growing experimental evidence suggests that these highly abundant secretory proteins are involved in the recognition of microbe-associated molecular patterns, serving as lectin-like molecules in the bivalve proto-complement system. While a large amount of functional data concerning the binding specificity of the globular head C1q domain and on the regulation of these molecules in response to infection are quickly accumulating, the genetic mechanisms that have led to the extraordinary lineage-specific expansion of the C1qDC gene family in bivalves are still largely unknown. The analysis of the chromosome-scale genome assembly of the Eastern oyster *Crassostrea virginica* revealed that the 476 oyster C1qDC genes, far from being uniformly distributed along the genome, are located in large clusters of tandemly duplicated paralogs, mostly found on chromosomes 7 and 8. Our observations point out that the evolutionary process behind the development of a large arsenal of C1qDC lectin-like molecules in marine bivalves is still ongoing and likely based on an unequal crossing over.

Keywords: innate immunity; lectins; complement system; C1q; bivalve mollusks; tandem duplication; pattern recognition receptors

1. Introduction

A growing body of evidence supports the idea that a proto-complement system, composed of C3, factor B and complement receptors, has an ancient origin in the animal tree of life [1]. Although several lineage-specific gene losses and acquisitions are likely to have reshaped the organization of the proto-complement system during its evolution, leading to its complete loss in some major extant taxa (e.g., insects), the activation of this primary defense system often relies on the recognition of microbe-associated molecular patterns (MAMPs) by different types of soluble lectin-like molecules. In vertebrate animals, two different types of pattern recognition receptors (PRRs) are involved in the lectin pathway of the complement system: mannan-binding lectins (MBLs) and ficolins. These proteins can recognize carbohydrate moieties associated with pathogens, triggering the activation of the complement proteolytic cascade, and eventually leading to the opsonization and killing of invading microbes.

The vertebrate complement component C1q is a connecting link between innate and adaptive immunity, as it enables complement activation through the recognition of antigen-complexed immunoglobulins, in the second arm of the complement system—the so-called classical pathway. Due to the absence of immunoglobulins, a proper classical pathway does not exist in invertebrates, making the activation of the complement system only possible in response to MAMP recognition by lectins, to the spontaneous hydrolysis of C3 (i.e., the alternative pathway), or to the direct binding of MAMPs by C1q, which is also observed in vertebrates [2]. In addition to the three monomeric units of the vertebrate C1q complex, the globular head C1q domain is also found in a number of different proteins with non-complement related functions, such as cerebellin, adiponectin, emilin, multimerin and others [3], defining the C1q-domain-containing (C1qDC) protein family and highlighting the high versatility of this structural scaffold [2].

In spite of the absence of an adaptive immune system, a large number of C1qDC proteins are encoded by the genomes of numerous metazoans which lack an adaptive immune system [4]. Due to the presence of several hundred C1qDC genes, bivalve mollusks most certainly emerge as the prime example of an animal group in which C1qDC proteins have met significant evolutionary success. For example, in stark contrast with most other protostomes, including gastropod and cephalopod mollusks, which only produce less than a dozen different C1qDC proteins [5,6], the genome of the Pacific oyster *Crassostrea gigas* harbors 337 C1qDC genes. Multiple transcriptome [7,8] and genome sequencing efforts have confirmed that C1qDC genes contribute to 0.5–1.5% of the entire repertoire of protein-coding genes of most bivalve species (e.g., 296 genes in *Pinctada fucata* [9], 445 in *Modiolus philippinarum* [10], 554 in *Saccostrea glomerata* [11] and over 1200 in *Ruditapes philipinarum* [12]). Curiously, this massive gene family expansion has been inferred to have occurred quite recently in bivalve evolution, since it only targeted all Pteriomorphia and Heterodonta, regardless of the environmental niche, but not the two remaining basal classes of Palaeoheterodonta and Protobranchia [13].

While it is still unclear whether all bivalve C1qDC proteins are involved in immune recognition [13], functional studies indicate that many of them play an important role as lectin-like molecules. The binding properties of the C1q domain enable the recognition of a broad range of MAMPs, such as peptidoglycan (PGN) and lipopolysaccharide (LPS)—the major components of Gram-positive and negative bacterial cell walls respectively—but also of other sugars associated with invading microbes, such as mannan [14,15], beta-1-3-glucan and yeast-glucan [16–18]. The impressive molecular diversification of bivalve C1qDC proteins has been hypothesized to be linked with a parallel functional specialization [16], which may further extend the range of potentially recognized MAMPs [19,20].

Bivalve C1qDC proteins are expressed in different tissues [13] and, upon secretion in the extracellular environment, they might be released in the hemolymph [7], in the extrapallial fluid [21] or in the mucus that covers the gills [13], offering a first line of defense against invading microorganisms in different body districts. The recognition of MAMPs by bivalve C1qDC proteins, which is probably aided by additional humoral factors, promotes the agglutination of bacterial cells [15,22], also triggering the migration and phagocytic activity of hemocytes [22–24], which clearly indicate an opsin-like function for these important soluble PRRs.

In vertebrates, the activation of the complement proteolytic cascade by C1q is effected by the presence a collagen tail, which also enables trimerization and the formation of a typical bouquet structure and defines the C1q-like type I domain architecture [20]. However, collagen tails are extremely rare in bivalves, which seem to either rely on a functionally analogous coiled-coil region for the assembly of oligomeric complexes (C1q-like type II proteins) and often completely lack N-terminal extensions (sghC1q proteins) [13].

Although several functional aspects remain to be fully investigated, the past decade has witnessed significant progress in the study of bivalve C1qDC proteins. Although these reports have contributed to a better elucidation of their functional significance in the context of immune response, the unavailability of high-quality genome assemblies has so far prevented the study of the genetic and molecular mechanisms that have led to the generation of several hundred C1qDC genes in this class of aquatic filter-feeding metazoans. Here, through the analysis of a high-quality chromosome-scale genome assembly [25], we investigate the genomic organization of the 476 C1qDC genes found in the Eastern oyster *Crassostrea virginica*, revealing a highly inhomogeneous distribution across chromosomes and a still-ongoing gene family expansion process which is mostly based on tandem gene duplication and episodic positive selection acting on newly generated gene copies. This process may have a significant impact of the functional diversification of these lectin-like molecules, resulting in the extension of the range of recognized MAMPs.

2. Results

2.1. The Repertoire of the Eastern Oyster C1qDC Genes

In total, 476 C1qDC genes were identified in the genome of *C. virginica* (detailed in Supplementary File 1). This number is in line with the previous report of 337 C1qDC genes in the congeneric species *C. gigas* [13], whose genome is slightly smaller (558 Mb vs. 685 Mb) [26], and similar to other Pteriomorphia [9–11].

Following the classification scheme previously proposed in another publication [13], oyster C1qDC proteins were labeled as follows: (i) sghC1qDC proteins, i.e., proteins containing a signal peptide, immediately followed by the C1q domain; (ii) sC1q-like type I proteins, i.e., secreted proteins containing a collagen tail before the C1q domain; (iii) sC1q-like type II proteins, i.e., secreted proteins containing a coiled-coil tail before the C1q domain; (iv) smultiC1q, i.e., secreted proteins containing multiple C1q domains; (v) other/uncertain, i.e., proteins with different domain architectures, or those resulting from likely incomplete annotation.

As in the case of the Pacific oyster, the majority of the C1qDCgenes (262, 55%) belonged to the sC1q-like type II category. SghC1q proteins were the second most abundant type, with 111 genes (23%); 21 genes encoded proteins with multiple C1q domains (three in most cases), which may or may not include a coiled-coil region (Table 1). No C1q-like type I protein was found in the Eastern oyster, confirming the observation that the association between the C1q domain and *N*-terminal collagen regions, typical of vertebrates, rarely occurs in bivalves [7,13].

Table 1. Details about the number and type of C1qDC genes found in the *Crassostrea virginica* genome.

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	Total
Chromosome size (Mb)	65.67	61.72	77.06	59.69	98.70	51.26	57.83	75.94	104.16	32.65	684.68
Number of C1qDC genes	45	17	18	12	17	36	123	123	73	12	476
C1qDC density (genes/Mb)	0.67	0.28	0.23	0.28	0.17	0.70	2.13	1.63	0.70	0.37	0.70
sghC1q genes	8	0	4	3	1	2	18	63	12	0	111
sC1q-like type II genes	27	6	10	7	12	26	81	36	48	9	262
smultiC1q genes	3	1	0	2	1	0	1	14	0	0	21
sSUEL/C1q	0	10	0	0	0	0	6	0	0	0	16
Other C1qDc genes	7	0	4	0	3	8	17	10	13	3	63

A total of 79 genes were classified in the “other/uncertain” category. While in part these may correspond to truncated genes resulting from incorrect computational prediction (as previously shown for many C1qDC genes in *C. gigas* [13]), some may correspond to bona fide cytosolic or membrane-bound C1qDC proteins, which have been previously identified in other bivalve species [7,13]. An interesting subgroup of unusual C1qDC proteins was found to contain a D-galactoside/L-rhamnose-binding SUEL domain (acronym for sea urchin egg lectin), which is typically found in a number of lectins produced by marine echinoderms [27–29]. Although three sSUEL/C1q genes had been previously reported in the Pacific oyster, their number in the Eastern oyster (16) largely exceeds that of the congeneric species [13].

2.2. Chromosomal Distribution of Oyster C1qDC Genes

Far from being uniformly split among chromosomes, oyster C1qDC genes displayed a highly skewed distribution, with more than 50% genes being located on chromosomes 7 and 8. Indeed, while these two chromosomes contained 123 C1qDC genes each, others encoded as few as 12 C1qDC genes (i.e., chromosomes 4 and 10). This reflected a very uneven C1qDC gene density, ranging from 0.17 genes/Mb (chromosome 5) to 2.13 genes/Mb (chromosome 7), with the average genomic C1qDC gene density standing at 0.70 genes/Mb (Figure 1).

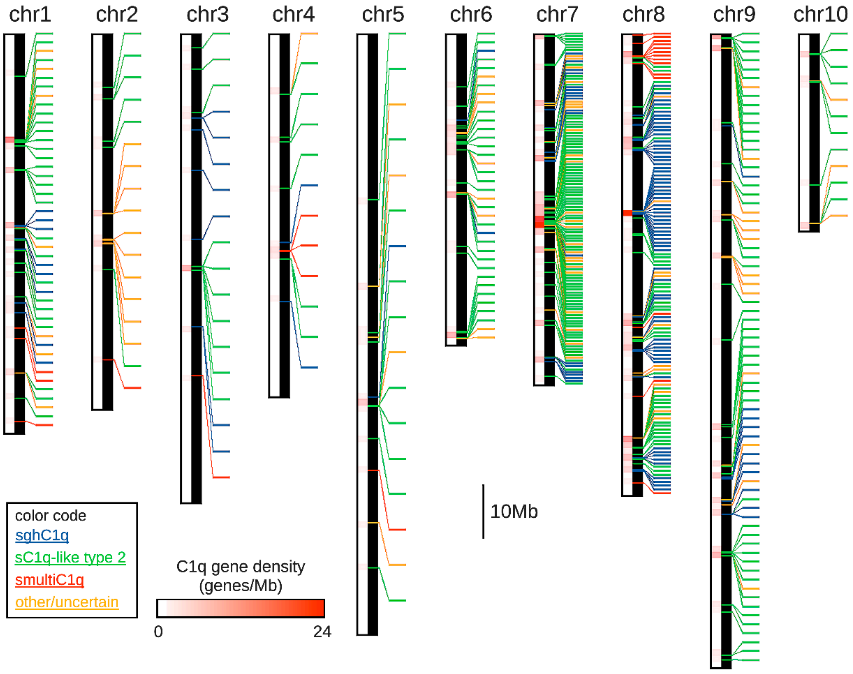

Figure 1. Chromosomal distribution of C1qDC genes in the *Crassostrea virginica* genome. The three main types of C1qDC genes (sghC1q, sC1q-like type 2, and smultiC1q) are indicated with different colors. C1qDC genes encoding proteins with a diverse domain organization, or of uncertain classification, were placed in the "other/uncertain" category. Chromosomes are drawn in scale.

C1qDC gene density was also largely non-homogeneous along chromosomes, as the genes were often placed in packed clusters, which often included more than a dozen units. This resulted in local peaks of very high gene density (i.e., >20 C1qDC genes/Mb in some regions of chromosome 7 and 8), as opposed to very large regions (>10 Mb) completely devoid of C1qDC genes (e.g., chromosome 3, 5 and 10) (Figure 1). Overall, 1.38% of the genes encoded by the *C. virginica* genome pertain to the C1qDC gene family. The relative abundance of C1qDC genes was however much higher for the two aforementioned chromosomes, standing at 4.78% and 3.16% for chromosome 7 and 8, respectively.

At the same time, a complete view of the localization of C1qDC genes in the genome (Figure 1) clearly shows that the different types of C1qDC genes were not evenly distributed. Type II C1q-like genes were the most abundant in all chromosomes, with the exception of chromosome 8, where the sghC1q type was the predominant one (Table 1). The high abundance of sghC1qDC genes in chromosome 8 was particularly notable, as 63 out of the 111 C1qDC genes of this type (57%) were found in this chromosome. Similarly, the majority of smultiC1qDC genes (14 out of 21) were found on chromosome 8, located in a dense cluster (Figure 1). On the other hand, most sSUELC1q genes (10 out of 16) were located on chromosome 2.

2.3. Most Oyster C1qDC Genes Are Tandemly Duplicated

The organization of oyster C1qDC genes in monotypic gene clusters (Figure 1) clearly suggest that these sequences are paralogous, which may have been generated by tandem gene duplication. We analyzed this evolutionary aspect in more depth by (i) inspecting the pairwise sequence similarity of the encoded proteins, and (ii) investigating their relative position. The clustering approach evidenced that 421 C1qDC proteins (88.44%) were grouped in clusters of at least two putative paralogs, and that 172 genes (36.13%) were grouped in large clusters (i.e., including ≥10 putative paralogous genes).

The C1qDC genes included in such large clusters were, for the most part, both evolutionarily closely related and spatially close to each other, as exemplified by one of the largest clusters identified in chromosome 7 (Figure 2a), which included 34 C1qDC genes, mostly of the C1q-like type II group. The phylogenetic analysis revealed that, of all C1qDC genes encoded by chromosome 7, these 34 genes created a well-supported monophyletic group (97% posterior probability), and they were organized in two distinct gene subclusters, containing 6 and 28 genes, respectively. The two subclusters were found to be separated by ~1.8 Mb sequence (Figure 2a) and, for the most part, included tandemly duplicated genes encoded on the same strand.

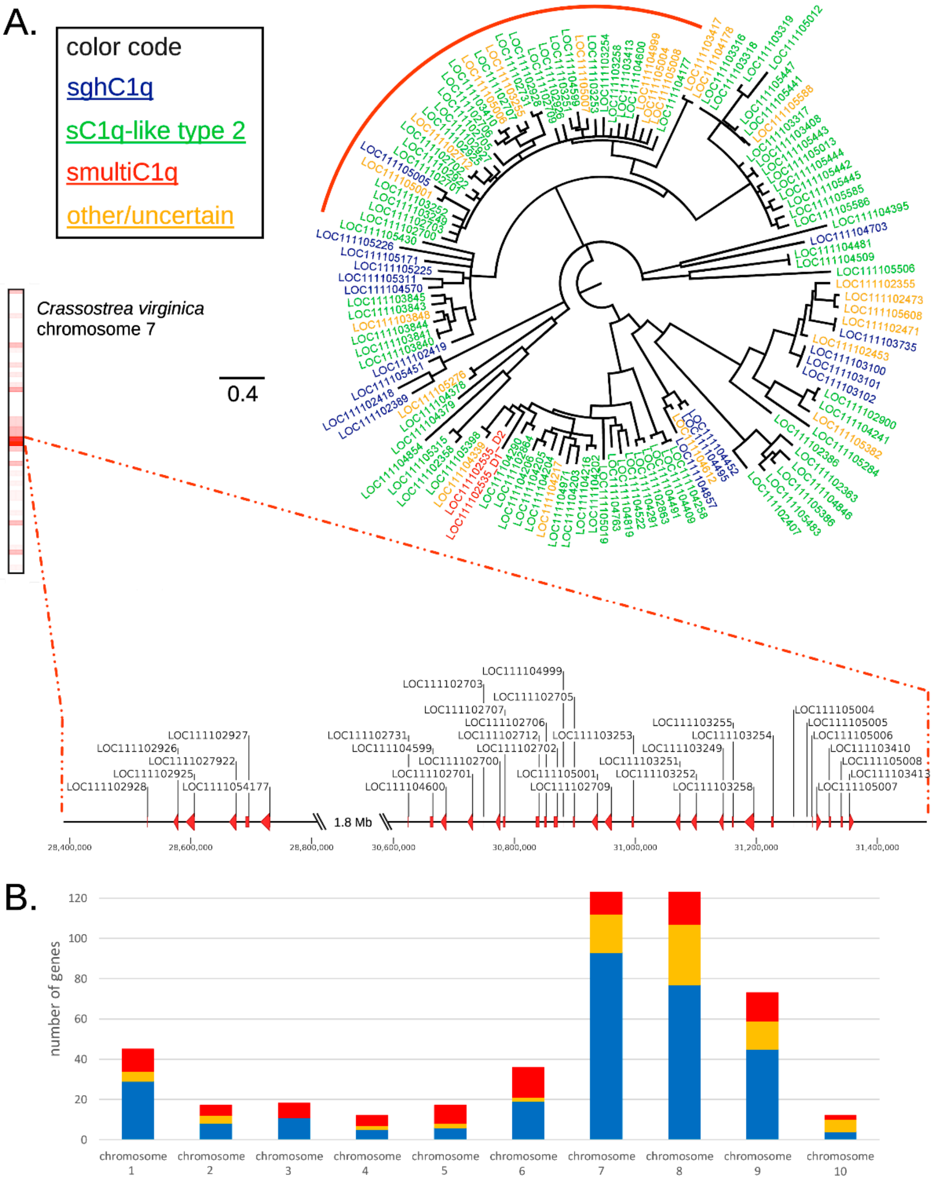

Figure 2. (**A**) Bayesian phylogenetic tree of the C1qDC proteins encoded by *Crassostrea virginica* chromosome 7. The three main types of C1qDC genes (sghC1q, sC1q-like type 2 and smultiC1q) are indicated with different colors. The location of the C1qDC genes pertaining to the monophyletic group indicated with a red circle is shown in detail in the zoomed-in chromosomal region. Nodes supported by posterior probability values <50% were collapsed. (**B**) Number of tandemly duplicated, proximally duplicated and dispersed C1qDC genes in each *C. virginica* chromosome.

The investigation carried out at the whole-genome scale confirmed the widespread occurrence of tandem duplications, which impacted 297 C1qDC genes (62.39%). In total, 84 genes (17.65%) were proximally duplicated (i.e., they were not directly flanked by C1qDC genes, but placed within 100 Kb distance from the closest one), and thus were the likely product of similar unequal crossing over processes [30] (Figure 2b). As previously mentioned, only a tiny fraction of the Eastern oyster C1qDC genes (i.e., <12%) were single-copy (i.e., they encoded proteins sharing <50% sequence similarity with other C1qDC proteins) and, in accordance with this observation, the occurrence of dispersed C1qDC genes (95 genes, 19.96%) was only slightly higher, showing the low prevalence of gene duplications driven by the activity of transposable elements. Chromosomes 7 and 8—i.e., the two chromosomes with the highest C1qDC gene density—were also those containing the highest proportion of tandemly or proximally duplicated genes (91.06% and 86.99%, respectively), confirming the fundamental importance of these processes in the bivalve C1qDC gene family expansion.

2.4. The Gene Duplication Process Is Still Ongoing and Is Paired with Diversifying Selection

Multiple lines of evidence suggest that the development of a very large repertoire of C1qDC genes in the Eastern oyster is the product of an evolutionary process that is still presently ongoing. Evidence in support of the very recent origin of paralogous C1qDC gene copies by tandem duplication is provided, for example, by the existence of eight nearly-identical genes, which encode proteins sharing 100% sequence identity at the amino acid level. In particular, three paralogous sghC1qDC genes located on chromosome 7 (i.e., LOC111103100, LOC111103101 and LOC111103102) share 100% sequence identity at the nucleotide level within the coding sequence, and only display small intronic indels due to the presence of microsatellites. A total of 65 C1qDC genes (13.63%) encode proteins with >95% sequence identity with the closest paralog, and this number rises to 89 genes (18.66%) for a sequence similarity threshold equal to 90%.

An additional example of this evolutionary process is given by a group of six highly similar paralogous sC1q-like type 2 genes found in chromosome 8 (i.e., LOC111105959, LOC111105960, LOC111105962, LOC111106427, LOC111106428 and LOC111109313), which all encode proteins of very similar length (249–251 aa) and share >90% sequence identity (Figure 3a). The six genes retain the same gene architecture, with conserved splicing donor and acceptor sites. Similarly, they display a nearly identical size of exons (with the only two exceptions of a 3nt-long in-frame deletion in LOC111105962 exon 1, and a 1nt-long deletion at the end of the coding sequence of LOC1111105960, which resulted in the acquisition of an additional codon at the 3′ end of the ORF). As expected, the intron was subject to more relaxed selective constraints compared with the two exons, displaying a higher evolutionary rate, as evidenced by the higher number of indels and SNPs (Figure 3a).

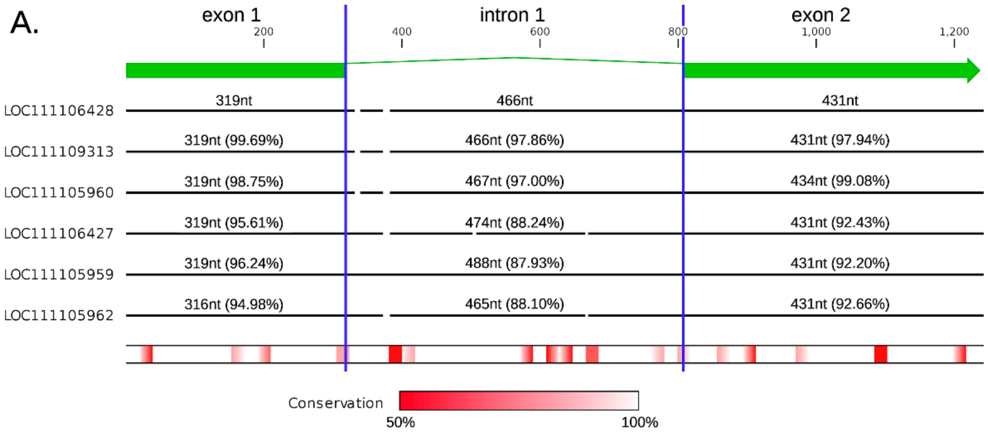

Figure 3. *Cont.*

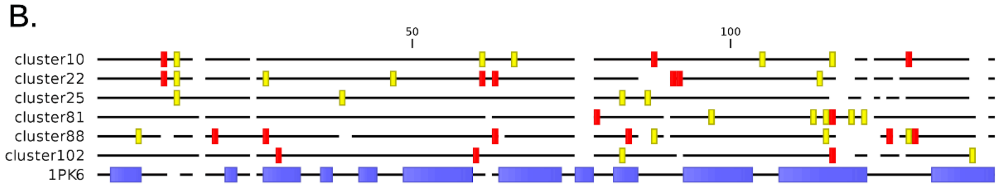

Figure 3. (**A**) Sequence comparison among the six sC1q-like type 2 paralogous genes LOC111105959, LOC111105960, LOC111105962, LOC111106427, LOC111106428 and LOC111109313, all found in chromosome 8. Identity percentages are shown for exon1, intron 1 and exon 2 separately, using the LOC111106428 sequence as a reference. Sequence conservation is shown with a heat map and indels are shown as line breaks. (**B**) Position of sites subject to positive selection in the six largest clusters of paralogous C1qDC genes identified in *C. virginica*. Strongly supported sites (i.e., detected with at least 2 different prediction methods) are shown in red. Sites with moderate support (i.e., detected with just one prediction method) are shown in yellow. The position of the beta strands in the human C1q chain A (PDB: 1PK6) are indicated by blue blocks.

BUSTED revealed strong evidence (p-value < 0.05) of episodic positive selection in the six largest clusters of paralogous C1qDC genes identified with a genome-wide analysis, suggesting that the gene duplication process is rapidly followed by molecular (and possibly functional) diversification through positive selection acting on a limited number of specific sites. In contrast with these hypervariable sites, the vast majority of the sites included in the globular C1q domain were found to evolve under negative selection, likely due to selective constraints linked with the maintenance of the typical 10-strand jelly-roll fold structure of the C1q domain itself, as well as of the contact surfaces between subunits required for oligomerization. The number of sites subject to significant diversifying selection, and their statistical support, largely varied from one cluster to the other. The four clusters that included C1q-like type 2 genes (i.e., cluster 10: 18 genes, cluster 22: 26 genes, cluster 81: 15 genes, and cluster 88: 19 genes) displayed between seven and 11 positively selected sites; i.e., roughly 5–10% of the sites included in the globular C1q head. The two clusters comprising sequences of different types (i.e., cluster 25: 16 sSUELC1q genes; cluster 102: 21 sghC1q genes) displayed a somewhat lower number of positively selected sites compared to the four aforementioned cases (four and five, respectively) (Figure 3b).

3. Discussion

In eukaryotes, small-scale gene duplication (i.e., independent from whole-genome duplication) can occur through a number of different mechanisms [31], which may or may not include the activity of transposable elements. Tandem gene duplications result in the creation of two adjacent paralogous genes, which are usually separated by a few Kb of intergenic sequence. Proximal gene duplications similarly result in the creation of two paralogous genes, which are placed at slightly longer distances compared with tandem duplications, and usually separated from each other by one or more genes. Both types of duplication are generally thought to arise from an unequal crossing over; i.e., from the misalignment of homologous regions of sister chromatids during meiosis [30]. An alternative process behind the creation of paralogous gene copies depends on the activity of transposable elements (TEs), such as DNA transposons [32] or retrotransposons [33]. In this case, new gene copies may either retain the original exon/intron architecture and regulatory features (DNA transposons), or completely lack introns, being detached from the promoter region of the original paralogous copy (retrotransposons). Even though the activity of TEs is sometimes spatially localized [34], in most cases, the new gene copies are dispersed; i.e., they are placed in genomic regions distant from their paralogs, often in different chromosomes.

Our investigation demonstrates that the predominant evolutionary process behind the massive expansion of the C1qDC gene family in bivalves is tandem gene duplication. This conclusion is supported by the observation that 62.39% of paralogous C1qDC genes are tandemly duplicated and that an additional 17.65% are proximally duplicated (Figure 2b). Hence, the widespread presence of C1qDC gene clusters in the Eastern oyster genome suggests that an unequal crossing over might be considered as the primary driving force behind this gene family expansion event. Our study further suggests that this process is still actively ongoing, since numerous identical or nearly-identical gene copies, inferred to have a very recent evolutionary origin, are present in dense gene clusters, which shows a highly inhomogeneous distribution along the genome.

However, this remarkable gene family expansion, with the extensive retention of newly generated gene copies, requires some additional explanation, as it seemingly contradicts the observation that the majority of duplicated genes with redundant function evolve under more relaxed selective constrains, being more prone to loss-of-function mutations, pseudogenization and consequent loss [35,36]. Over the years, multiple and partially overlapping theoretical models have been proposed to explain the fixation of duplicated genes in populations, which depend on a complex combination between the evolutionary dynamics of a given gene family and functional properties of the encoded proteins [37,38]. The retention of hundreds of C1qDc genes in the Eastern oyster, as well as in other bivalve species [5,7,13], most certainly indicates that the generation of tandemly and—in a minor way—proximally duplicated C1qDC genes represents an important evolutionary advantage in these filter-feeding organisms. Based on the functional information collected so far in oysters, scallops, mussels, clams and other bivalves, this advantage could arise from the neofunctionalization or subfunctionalization of new gene copies [39], which may acquire the ability to recognize additional MAMPs, extending the range of pathogenic microorganisms recognized, through the mutation of key sites involved in glycan recognition [16].

One key aspect that remained to be investigated was whether such molecular diversification is the result of relaxed selective constraints on redundant gene copies, or of positive selection acting on selected sites potentially involved in MAMP-binding, which could significantly affect the lectin-like function of oyster C1qDC proteins. The analysis of the six largest clusters of paralogous genes found in the *C. virginica* genome revealed highly supported traces of episodic positive selection, suggesting that duplicated gene copies undergo rapid molecular, and possibly functional, diversification. This process only seemed to act on a limited number of localized residues (4–11 per cluster, Figure 3b), which were in part shared by the different clusters and displayed a higher frequency in the hypervariable and gap-rich C-terminal part of the C1q domain (Figure 3b). On the other hand, the vast majority of the sites included in the C1q domain was subject to strong purifying selection, likely due to structural constraints. Our current knowledge of the sites involved in MAMP recognition in bivalve C1qDC proteins is virtually non-existent, and most of the information available concerning the residues involved in C1q ligand binding derives from very distantly related model organisms, such as mouse and human [40,41]. Consequently, the possibility to infer whether the position of positively selected sites in the 3D structure of the globular head C1q domain is paired with a functional specialization and to the acquisition of the ability to recognize novel ligands in tandemly duplicated C1qDC genes falls beyond our current reach.

Another interesting aspect revealed by this study was the complete lack of C1q-like type I proteins in *C. virginica*. In light of the recent report of the presence of this domain architecture in early-branching metazoans [42], its absence in the Eastern oyster suggests that bivalves do not necessarily require collagen tails for the activation of the proto-complement system. The presence of a very few non-orthologous C1q-like type I genes in other bivalve species [7,13] further reinforces the idea that alternative N-terminal structures (i.e., coiled-coil regions) are the predominant structure used for C1qDC protein oligomerization and the subsequent activation of the complement proteolytic cascade. The finding that 16 sSUEL/C1q genes were present in the Eastern oyster genome was another relevant result, since this domain combination, unique to Bivalvia, is highly reminiscent of the recently reported case of C1q-related proteins (QREPs) found in a few Caenogastropoda and Heterobranchia gastropod species [42]. As with sSUEL/C1q, these unusual proteins combine the C-terminal globular C1q domain with N-terminal domains with marked binding properties (i.e., two immunoglobulin-like domains). Altogether, these findings stimulate further research towards the functional characterization of these potentially multifunctional lectin-like proteins.

In summary, this study provides, for the first time, evidence supporting the important role of tandem gene duplication and an unequal crossing over in the expansion and molecular evolution of the C1qDC gene family in bivalves. Based on preliminary results, this rapid and still-ongoing process is likely paired with episodic diversifying selection, which only acts on a limited number of spatially localized residues, whose functional significance in MAMP binding should be investigated in depth in the future. Evolutionary processes analogous to those we have described for the Eastern oyster C1qDC genes may have targeted several other gene families involved in immune response, either as receptors or as effectors, that have similarly reportedly undergone massive lineage-specific expansion and molecular diversification [5,43,44].

4. Materials and Methods

4.1. Identification of C1qDC Genes

The annotated chromosome-scale nuclear genome assembly v.3.0 of *Crassostrea virginica* (Gmelin, 1791) [25] was downloaded from the GenBank repository (GCA_002022765.4). Protein-coding genes were virtually translated and the resulting amino acid sequences were screened for the presence of one or more C1q domains with HMMER v.3.2.1 [45], based on a 0.05 e-value threshold, using the PFAM profile of the C1q domain (PF00386) as a query. Positive hits were further refined by removing sequences with partial domains, taking into account the possibility of incorrect annotations.

All C1qDC proteins were further characterized as follows: the presence of *N*-terminal signal peptides and transmembrane regions was inspected with Phobius [46]; additional conserved protein domains were predicted with InterProScan v.5 [47] and coiled-coil regions were predicted with COILS [48], with a window length of 14, 21 and 28 amino acids, based on a probability threshold > 0.5. Protein nomenclature followed the scheme proposed in a previous publication [13] (see Section 2.1 for details).

4.2. Sequence Analysis of C1qDC Genes

The coordinates of each C1qDC gene were obtained from the genome GFF (General Feature Format) annotation file, and their positions were indicated on chromosomes, drawn to scale. The four gene categories indicated in Section 4.1 (with the exception of type I C1q-like genes, as no sequence of this type was identified) were indicated with different colors. The C1qDC gene density per Mb of genomic DNA sequence was computed using a 1 Mb sequence window length, and represented as a heat map on the side of each chromosome.

The amino acid sequences were clustered by similarity using CD-HIT v.4.8.1 [49], with different sequence identity thresholds to identify the products of putative gene duplication events. Sequences sharing > 50% pairwise identify were considered to be part of paralogous gene clusters and further classified as (i) tandemly duplicated genes, if they were flanked by at least one paralogous C1qDC genes, either at the 5′ or 3′ side; (ii) proximally duplicated genes, if the flanking genes were not pertaining to the C1qDC family, but the closest paralogous C1qDC was located within 100 Kb of distance; and (iii) dispersed genes, if the closest C1qDC gene was not located within 100 Kb of distance.

The evolutionary history of the C1qDC genes encoded by chromosome 7 (i.e., the oyster chromosome displaying the highest C1qDC gene density) was further investigated by generating a multiple sequence alignment (MSA) of all the encoded amino acid sequences. The MSA file, obtained with MUSCLE v.6.0 [50], was trimmed to only include the C1q domain; i.e., the region shared by all proteins. For smultiC1q proteins, each domain was separately added to the MSA. ModelTest-NG v.0.1.5 [51] estimated the WAG (Whelan And Goldman) model of molecular evolution [52], with a proportion of invariable sites and a gamma-shaped rate of variation across sites (WAG+G+I), as the best-fitting for the dataset analyzed. Phylogenetic inference was carried out with

MrBayes v.3.2.7a [53], running two parallel MCMC analyses with four chains each. The convergence of all estimated parameters was assessed with Tracer v.1.7.1 [54], based on the reaching of an effective sample size > 200; i.e., with 400,000 generations. The resulting phylogenetic tree was represented as a 50% majority rule consensus tree (i.e., nodes supported by posterior probability values <50% were collapsed).

4.3. Positive Selection Analysis

The six largest clusters of paralogous C1qDC genes, identified as described in Section 2.2, all containing 15 genes or more, were analyzed to detect the signatures of positive and negative selection as follows. First, the nucleotide sequences of the open reading frames were aligned with MEGA X [55], using MUSCLE [50], preserving the integrity of codon triplets. The obtained MSA files were trimmed to remove positions containing gaps in the alignment, as well as regions not included in the globular C1q domain. The resulting processed MSA files were processed with BUSTED [56] to detect domain-wise signatures of episodic positive selection, based on a p-value threshold < 0.05. The position of sites subject to positive selection was predicted with FUBAR, SLAC, FEL and MEME [57–59], based on a p-value threshold < 0.05 (or posterior probability > 0.95). All these tools were implemented with the use of the online Datamonkey 2.0 platform (https://datamonkey.org) [60]. Sites evolving under positive selection were categorized either as strongly supported (i.e., detected with at least two different methods) or as moderately supported (i.e., detected with just one method). Sites subject to positive selection were graphically represented in the alignment between the consensus sequences of the six selected clusters, using the experimentally determined structure of the chain A of the human C1q (PDB: 1PK6) [19] as a reference. The multiple sequence alignment was implemented with structural information using Expresso [61].

Supplementary Materials: The following are available online at http://www.mdpi.com/1660-3397/17/10/583/s1, Table S1: list of C1qDC genes identified in the *Crassostrea virginica* genome.

Author Contributions: Conceptualization, M.G. and A.P.; methodology, M.G. and S.G.; formal analysis, M.G. and S.G.; writing—original draft preparation, M.G.; writing—review and editing, M.G., S.G. and A.P.; visualization, M.G. and S.G.; supervision, A.P.; funding acquisition, A.P.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 678589. Samuele Greco is supported by a grant from the Italian National Program for Antarctic Research, project PNRA16_00099.

Acknowledgments: We thank Davide Fracarossi for bioinformatics assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Gorbushin, A.M. Immune repertoire in the transcriptome of Littorina littorea reveals new trends in lophotrochozoan proto-complement evolution. *Dev. Comp. Immunol.* **2018**, *84*, 250–263. [CrossRef] [PubMed]
2. Nayak, A.; Pednekar, L.; Reid, K.B.M.; Kishore, U. Complement and non-complement activating functions of C1q: A prototypical innate immune molecule. *Innate Immun.* **2012**, *18*, 350–363. [CrossRef] [PubMed]
3. Tom Tang, Y.; Hu, T.; Arterburn, M.; Boyle, B.; Bright, J.M.; Palencia, S.; Emtage, P.C.; Funk, W.D. The complete complement of C1q-domain-containing proteins in Homo sapiens. *Genomics* **2005**, *86*, 100–111. [CrossRef] [PubMed]
4. Carland, T.M.; Gerwick, L. The C1q domain containing proteins: Where do they come from and what do they do? *Dev. Comp. Immunol.* **2010**, *34*, 785–790. [CrossRef] [PubMed]
5. Zhang, L.; Li, L.; Guo, X.; Litman, G.W.; Dishaw, L.J.; Zhang, G. Massive expansion and functional divergence of innate immune genes in a protostome. *Sci. Rep.* **2015**, *5*, 8693. [CrossRef]
6. 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. [CrossRef]

7. Gerdol, M.; Manfrin, C.; De Moro, G.; Figueras, A.; Novoa, B.; Venier, P.; Pallavicini, A. The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: A widespread and diverse family of immune-related molecules. *Dev. Comp. Immunol.* **2011**, *35*, 635–643. [CrossRef]
8. Zhang, L.; Li, L.; Zhu, Y.; Zhang, G.; Guo, X. Transcriptome analysis reveals a rich gene set related to innate immunity in the Eastern oyster (*Crassostrea virginica*). *Mar. Biotechnol.* **2014**, *16*, 17–33. [CrossRef]
9. Takeuchi, T.; Koyanagi, R.; Gyoja, F.; Kanda, M.; Hisata, K.; Fujie, M.; Goto, H.; Yamasaki, S.; Nagai, K.; Morino, Y.; et al. Bivalve-specific gene expansion in the pearl oyster genome: Implications of adaptation to a sessile lifestyle. *Zool. Lett.* **2016**, *2*, 3. [CrossRef]
10. Sun, J.; Zhang, Y.; Xu, T.; Zhang, Y.; Mu, H.; Zhang, Y.; Lan, Y.; Fields, C.J.; Hui, J.H.L.; Zhang, W.; et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* **2017**, *1*, 0121. [CrossRef]
11. Powell, D.; Subramanian, S.; Suwansa-Ard, S.; Zhao, M.; O'Connor, W.; Raftos, D.; Elizur, A. The genome of the oyster Saccostrea offers insight into the environmental resilience of bivalves. *DNA Res. Int. J. Rapid Publ. Rep. Genes. Genomes* **2018**, *25*, 655–665. [CrossRef] [PubMed]
12. Mun, S.; Kim, Y.-J.; Markkandan, K.; Shin, W.; Oh, S.; Woo, J.; Yoo, J.; An, H.; Han, K. The Whole-Genome and Transcriptome of the Manila Clam (*Ruditapes philippinarum*). *Genome Biol. Evol.* **2017**, *9*, 1487–1498. [CrossRef]
13. Gerdol, M.; Venier, P.; Pallavicini, A. The genome of the Pacific oyster *Crassostrea gigas* brings new insights on the massive expansion of the C1q gene family in Bivalvia. *Dev. Comp. Immunol.* **2015**, *49*, 59–71. [CrossRef] [PubMed]
14. Zong, Y.; Liu, Z.; Wu, Z.; Han, Z.; Wang, L.; Song, L. A novel globular C1q domain containing protein (C1qDC-7) from *Crassostrea gigas* acts as pattern recognition receptor with broad recognition spectrum. *Fish. Shellfish. Immunol.* **2019**, *84*, 920–926. [CrossRef] [PubMed]
15. Cui, Y.; Wei, Z.; Shen, Y.; Li, C.; Shao, Y.; Zhang, W.; Zhao, X. A novel C1q-domain-containing protein from razor clam *Sinonovacula* constricta mediates G-bacterial agglutination as a pattern recognition receptor. *Dev. Comp. Immunol.* **2018**, *79*, 166–174. [CrossRef] [PubMed]
16. Wang, L.; Zhang, H.; Wang, M.; Zhou, Z.; Wang, W.; Liu, R.; Huang, M.; Yang, C.; Qiu, L.; Song, L. The transcriptomic expression of pattern recognition receptors: Insight into molecular recognition of various invading pathogens in Oyster *Crassostrea gigas*. *Dev. Comp. Immunol.* **2019**, *91*, 1–7. [CrossRef] [PubMed]
17. Wang, L.; Wang, L.; Kong, P.; Yang, J.; Zhang, H.; Wang, M.; Zhou, Z.; Qiu, L.; Song, L. A novel C1qDC protein acting as pattern recognition receptor in scallop *Argopecten irradians*. *Fish. Shellfish. Immunol.* **2012**, *33*, 427–435. [CrossRef]
18. Wang, L.; Wang, L.; Zhang, D.; Jiang, Q.; Sun, R.; Wang, H.; Zhang, H.; Song, L. A novel multi-domain C1qDC protein from Zhikong scallop Chlamys farreri provides new insights into the function of invertebrate C1qDC proteins. *Dev. Comp. Immunol.* **2015**, *52*, 202–214. [CrossRef]
19. Gaboriaud, C.; Juanhuix, J.; Gruez, A.; Lacroix, M.; Darnault, C.; Pignol, D.; Verger, D.; Fontecilla-Camps, J.C.; Arlaud, G.J. The Crystal Structure of the Globular Head of Complement Protein C1q Provides a Basis for Its Versatile Recognition Properties. *J. Biol. Chem.* **2003**, *278*, 46974–46982. [CrossRef]
20. Reid, K.B.M. Complement Component C1q: Historical Perspective of a Functionally Versatile, and Structurally Unusual, Serum Protein. *Front. Immunol.* **2018**, *9*, 764. [CrossRef]
21. Yin, Y.; Huang, J.; Paine, M.L.; Reinhold, V.N.; Chasteen, N.D. Structural characterization of the major extrapallial fluid protein of the mollusc *Mytilus edulis*: Implications for function. *Biochemistry* **2005**, *44*, 10720–10731. [CrossRef] [PubMed]
22. Li, S.; Ruan, Z.; Yang, X.; Li, M.; Yang, D. Immune recognition, antimicrobial and opsonic activities mediated by a sialic acid binding lectin from *Ruditapes philippinarum*. *Fish. Shellfish. Immunol.* **2019**, *93*, 66–72. [CrossRef] [PubMed]
23. Jiang, S.; Li, H.; Zhang, D.; Zhang, H.; Wang, L.; Sun, J.; Song, L. A C1q domain containing protein from *Crassostrea gigas* serves as pattern recognition receptor and opsonin with high binding affinity to LPS. *Fish. Shellfish. Immunol.* **2015**, *45*, 583–591. [CrossRef] [PubMed]
24. Li, H.; Kong, N.; Sun, J.; Wang, W.; Li, M.; Gong, C.; Dong, M.; Wang, M.; Wang, L.; Song, L. A C1qDC (CgC1qDC-6) with a collagen-like domain mediates hemocyte phagocytosis and migration in oysters. *Dev. Comp. Immunol.* **2019**, *98*, 157–165. [CrossRef]

25. Gómez-Chiarri, M.; Warren, W.C.; Guo, X.; Proestou, D. Developing tools for the study of molluscan immunity: The sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish. Shellfish. Immunol.* **2015**, *46*, 2–4. [CrossRef]
26. Zhang, G.; Fang, X.; Guo, X.; Li, L.; Luo, R.; Xu, F.; Yang, P.; Zhang, L.; Wang, X.; Qi, H.; et al. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **2012**, *490*, 49–54. [CrossRef]
27. Hatakeyama, T.; Ichise, A.; Unno, H.; Goda, S.; Oda, T.; Tateno, H.; Hirabayashi, J.; Sakai, H.; Nakagawa, H. Carbohydrate recognition by the rhamnose-binding lectin SUL-I with a novel three-domain structure isolated from the venom of globiferous pedicellariae of the flower sea urchin Toxopneustes pileolus. *Protein Sci. Publ. Protein Soc.* **2017**, *26*, 1574–1583. [CrossRef]
28. Ozeki, Y.; Matsui, T.; Suzuki, M.; Titani, K. Amino acid sequence and molecular characterization of a D-galactoside-specific lectin purified from sea urchin (*Anthocidaris crassispina*) eggs. *Biochemistry* **1991**, *30*, 2391–2394. [CrossRef]
29. Carneiro, R.F.; Teixeira, C.S.; de Melo, A.A.; de Almeida, A.S.; Cavada, B.S.; de Sousa, O.V.; da Rocha, B.A.M.; Nagano, C.S.; Sampaio, A.H. L-Rhamnose-binding lectin from eggs of the *Echinometra lucunter*: Amino acid sequence and molecular modeling. *Int. J. Biol. Macromol.* **2015**, *78*, 180–188. [CrossRef]
30. Freeling, M. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **2009**, *60*, 433–453. [CrossRef]
31. Maere, S.; De Bodt, S.; Raes, J.; Casneuf, T.; Van Montagu, M.; Kuiper, M.; Van de Peer, Y. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA.* **2005**, *102*, 5454–5459. [CrossRef] [PubMed]
32. Cusack, B.P.; Wolfe, K.H. Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol. Biol. Evol.* **2007**, *24*, 679–686. [CrossRef] [PubMed]
33. Kaessmann, H.; Vinckenbosch, N.; Long, M. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **2009**, *10*, 19–31. [CrossRef] [PubMed]
34. Zhang, J.; Zuo, T.; Peterson, T. Generation of tandem direct duplications by reversed-ends transposition of maize ac elements. *PLoS Genet.* **2013**, *9*, e1003691. [CrossRef]
35. Wagner, A. The fate of duplicated genes: Loss or new function? *BioEssays News Rev. Mol. Cell. Dev. Biol.* **1998**, *20*, 785–788. [CrossRef]
36. Lynch, M.; Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **2000**, *290*, 1151–1155. [CrossRef]
37. Innan, H.; Kondrashov, F. The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* **2010**, *11*, 97–108. [CrossRef]
38. Hahn, M.W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* **2009**, *100*, 605–617. [CrossRef]
39. Rastogi, S.; Liberles, D.A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* **2005**, *5*, 28. [CrossRef]
40. Gaboriaud, C.; Frachet, P.; Thielens, N.M.; Arlaud, G.J. The Human C1q Globular Domain: Structure and Recognition of Non-Immune Self Ligands. *Front. Immunol.* **2012**, *2*, 92. [CrossRef]
41. Agrawal, A.; Shrive, A.K.; Greenhough, T.J.; Volanakis, J.E. Topology and structure of the C1q-binding site on C-reactive protein. *J. Immunol. Baltim. MD 1950* **2001**, *166*, 3998–4004. [CrossRef] [PubMed]
42. Gorbushin, A.M. Derivatives of the lectin complement pathway in Lophotrochozoa. *Dev. Comp. Immunol.* **2019**, *94*, 35–58. [CrossRef] [PubMed]
43. McDowell, I.C.; Modak, T.H.; Lane, C.E.; Gomez-Chiarri, M. Multi-species protein similarity clustering reveals novel expanded immune gene families in the eastern oyster Crassostrea virginica. *Fish. Shellfish. Immunol.* **2016**, *53*, 13–23. [CrossRef] [PubMed]
44. Gerdol, M.; Moreira, R.; Cruz, F.; Gómez-Garrido, J.; Vlasova, A.; Rosani, U.; Venier, P.; Naranjo-Ortiz, M.A.; Murgarella, M.; Balseiro, P.; et al. Massive gene presence/absence variation in the mussel genome as an adaptive strategy: First evidence of a pan-genome in Metazoa. *BioRxiv* **2019**, 781377. [CrossRef]
45. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39*, W29–W37. [CrossRef]
46. Käll, L.; Krogh, A.; Sonnhammer, E.L.L. Advantages of combined transmembrane topology and signal peptide prediction—The Phobius web server. *Nucleic Acids Res.* **2007**, *35*, W429–W432. [CrossRef]

47. Finn, R.D.; Attwood, T.K.; Babbitt, P.C.; Bateman, A.; Bork, P.; Bridge, A.J.; Chang, H.-Y.; Dosztányi, Z.; El-Gebali, S.; Fraser, M.; et al. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* **2017**, *45*, D190–D199. [CrossRef]
48. Lupas, A.; Van Dyke, M.; Stock, J. Predicting coiled coils from protein sequences. *Science* **1991**, *252*, 1162–1164. [CrossRef]
49. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
50. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
51. Darriba, D.; Posada, D.; Kozlov, A.M.; Stamatakis, A.; Morel, B.; Flouri, T. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *BioRxiv* **2019**, 612903. [CrossRef] [PubMed]
52. Whelan, S.; Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699. [CrossRef] [PubMed]
53. Huelsenbeck, J.P.; Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinf. Oxf. Engl.* **2001**, *17*, 754–755. [CrossRef] [PubMed]
54. Rambaut, A.; Drummond, A.J.; Xie, D.; Baele, G.; Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **2018**, *67*, 901–904. [CrossRef] [PubMed]
55. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef] [PubMed]
56. Murrell, B.; Weaver, S.; Smith, M.D.; Wertheim, J.O.; Murrell, S.; Aylward, A.; Eren, K.; Pollner, T.; Martin, D.P.; Smith, D.M.; et al. Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **2015**, *32*, 1365–1371. [CrossRef] [PubMed]
57. Kosakovsky Pond, S.L.; Frost, S.D.W. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **2005**, *22*, 1208–1222. [CrossRef]
58. Murrell, B.; Wertheim, J.O.; Moola, S.; Weighill, T.; Scheffler, K.; Kosakovsky Pond, S.L. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **2012**, *8*, e1002764. [CrossRef]
59. Murrell, B.; Moola, S.; Mabona, A.; Weighill, T.; Sheward, D.; Kosakovsky Pond, S.L.; Scheffler, K. FUBAR: A fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.* **2013**, *30*, 1196–1205. [CrossRef]
60. Weaver, S.; Shank, S.D.; Spielman, S.J.; Li, M.; Muse, S.V.; Kosakovsky Pond, S.L. Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. *Mol. Biol. Evol.* **2018**, *35*, 773–777. [CrossRef]
61. Armougom, F.; Moretti, S.; Poirot, O.; Audic, S.; Dumas, P.; Schaeli, B.; Keduas, V.; Notredame, C. Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **2006**, *34*, W604–W608. [CrossRef] [PubMed]