Genomics 112 (2020) 4887–4896

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identification of variants associated with hard clam, *Mercenaria mercenaria*, resistance to Quahog Parasite Unknown disease

Sarah Farhat[a], Arnaud Tanguy[b], Emmanuelle Pales Espinosa[a], Ximing Guo[c], Isabelle Boutet[b], Roxanna Smolowitz[d], Diane Murphy[e], Gregg J. Rivara[f], Bassem Allam[a,*]

[a] Marine Animal Disease Laboratory, School of Marine and Atmospheric Sciences, Stony Brook University, 100 Nicolls Road, Stony Brook, NY 11794-5000, USA
[b] Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, Sorbonne Université, Place Georges Teissier, 29680 Roscoff, France
[c] Haskin Shellfish Research Laboratory, Department of Marine and Coastal Sciences, Rutgers University, 6959 Miller Avenue, Port Norris, NJ 08349, USA
[d] Roger Williams University, Department of Biology, Marine Biology, and Environmental Science, 1 Old Ferry Rd, Bristol, RI 02809, USA
[e] Cape Cod Cooperative Extension, 3195 Main St, Barnstable, MA 02630, NY 1197, USA
[f] Cornell University Cooperative Extension of Suffolk County, 3690 Cedar Beach Rd, Southold, NY 11971, USA

ARTICLE INFO

Keywords:
Hard clam
QPX
RADseq
Variants
Resistance
Parasite

ABSTRACT

Severe losses in aquacultured and wild hard clam (*Mercenaria mercenaria*) stocks have been previously reported in the northeastern United States due to a protistan parasite called QPX (Quahog Parasite Unknown). Previous work demonstrated that clam resistance to QPX is under genetic control. This study identifies single nucleotide polymorphism (SNP) associated with clam survivorship from two geographically segregated populations, both deployed in an enzootic site. The analysis contrasted samples collected before and after undergoing QPX-related mortalities and relied on a robust draft clam genome assembly. ~200 genes displayed significant variant enrichment at each sampling point in both populations, including 18 genes shared between both populations. Markers from both populations were identified in genes related to apoptosis pathways, protein-protein interaction, receptors, and signaling. This research begins to identify genetic markers associated with clam resistance to QPX disease, leading the way for the development of resistant clam stocks through marker-assisted selection.

1. Introduction

The hard clam, *Mercenaria mercenaria*, is one of the most important marine resources along the east coasts of the United States for its ecological services and commercial value. *M. mercenaria* is a relatively "hardy" species and only a small number of infectious diseases have population-scale impacts on the species, with the most prominent one being QPX disease. QPX (Quahog Parasite Unknown) disease was first detected in 1962 in New Brunswick, Canada, in a wild population of *M. mercenaria*, and in 1989, in aquacultured clams displaying a high mortality rate in Prince Edward Island, Canada [1]– [3]. Since then, QPX was found in multiple locations from Canada to Virginia [4,8,16,17]. QPX belongs to the phylum Labyrinthulomycota [2] and is considered to represent an opportunistic facultative parasite that is widespread in the environment [7,9] causing disease and mortality in certain groups of clams "*that may be disadvantaged in some way, perhaps from an unfavorable genotype-environment interaction*" [6].

Previous investigations demonstrated difference in disease development under different environmental conditions, particularly temperature [6,10,13,14,18]. Overall, QPX was shown to be well adapted to low temperature conditions, causing significantly higher disease levels at temperatures as low as 13 °C as compared to 21 or 27 °C [12,19]. Molecular investigations showed that QPX overexpresses chaperon genes and virulence-related genes under low temperature conditions, allowing the parasite to maintain high activity and virulence when clam immune responses are compromised [20,21,26]. Interestingly, genetic analyses supported QPX adaptation to local environmental conditions (e.g. temperature, salinity) throughout its geographical range [22]. Once infected, *M. mercenaria* usually initiates an intense inflammatory response and triggers a series of cellular and molecular processes [3]– [5,23,27]. Under optimal conditions for clams (e.g. > 21 °C), the immune response can lead to the neutralization of parasite cells and a complete healing of infected clams [15,28]. More important, prior work showed that clam resistance toward QPX is a genetic trait and depends upon the origin of the brood stock [6,10,11,19,24,29,30]. Although gene expression analyses have been previously performed, no population genetics studies were accomplished to contrast susceptible and resistant populations. Furthermore,

* Corresponding author.
E-mail address: bassem.allam@stonybrook.edu (B. Allam).

https://doi.org/10.1016/j.ygeno.2020.08.036
Received 29 June 2020; Received in revised form 12 August 2020; Accepted 28 August 2020
Available online 03 September 2020
0888-7543/

genetic information on *M. mercenaria*, and the whole *Veneridae* family for that matter, remains extremely limited with only the genome of the Manila clam (*Ruditapes philippinarum*) and the Venus clam (*Cyclina sinensis*) having been sequenced [31,33].

This study was designed to identify genetic features associated with clam resistance to QPX. First, a draft genome of *M. mercenaria* was generated. This genome was then used as a reference to investigate the distinctive genetic makers associated with QPX disease resistance in two geographically segregated populations deployed in an enzootic site in Massachusetts, USA. The overarching goal being the identification of markers that could be used to improve aquaculture populations.

2. Results and discussion

2.1. QPX disease development and clam mortality

The occurrence and severity of QPX infection and disease were dramatically different between the two groups of clams (named OYB and ARC in the manuscript) even though they had been cultured in plots adjacent to each other. At 18 months, 28 of 60 ARC clams (47%) examined histologically showed moderate (< 100 parasite cells in the tissue section) to extensive (> 100 parasite cells) infection in the mantle, gill, and in some animals also in the visceral mass (typically the connective tissue surrounding the digestive tract). These infections were active and usually showed many live QPX cells in the foci of infection. In contrast, only 10% of OYB clams (6 out of 60) showed signs of infection with only 3 of these displaying moderate to abundant numbers of QPX cells while the other three clams showing markedly less QPX cells in the tissues. Interestingly, the severity of the infection in both clam groups was reflected in clam mortality as mortality rate reached 50% in ARC, as opposed to only 20% in OYB, suggesting that mortality in deployed clams was mainly driven by QPX disease. Overall, results indicated that OYB clams were more efficient in preventing infection, and if they became infected, were more able to destroy the invading QPX cells.

2.2. Genome assembly and annotation

Using a combination of Illumina and 10× Genomics short-read and PacBio long-read sequencing technologies, we sequenced and assembled a draft genome for *M. mercenaria* (Table S1–S5). The genome assembly of 2.4 Gb (Table S4) was larger than the genome size estimation obtained by Bulk Fluorometric Assay of 2.00 pg (c-value) corresponding to ~1.956 Gb [37]. This genome was also larger than most Bivalvia genomes sequenced so far but is similar to the genome of the mussel *Modiolus philippinarum* [38]. The genome of *Ruditapes philippinarum*, the only other venerid clam sequenced [32], is 2.19 times smaller than that of *M. mercenaria* (Table 1). We predicted 42,214 protein-coding genes (Table 1) of which 91% had functional assignations: 60% assigned to a KEGG annotation, 82% with at least one domain and 85% with a hit on NR (Fig. S1). The Brite classification showed an enrichment in environmental information processing class A where signal transduction and signaling molecules and interaction were the most abundant in class B (Fig. S2). This genome assembly and gene repertoire were preliminary and likely inflated by haplotigs, and we are working on a chromosome-level assembly in an ongoing project (Farhat et al., unpublished). For that reason, the genome assembly and annotation will not be further discussed.

2.3. Overall variant detection

In order to reveal molecular features associated with clam resistance to QPX, we used dd-RAD sequencing techniques to contrast SNP profiles in clams derived from two different populations (OYB and ARC) and collected before and after QPX-related mortalities. After filtering and processing the raw reads by using Stacks software (see Materials and Methods), 353,766 and 244,005 SNPs in OYB and 224,061 and 211,508 SNPs in ARC were identified in each subpopulation (Time 0 and Time 18 months, respectively, Table S6). Fig. S3 shows the clustering of the samples according to the frequency of the detected variants and highlights a noticeable genetic segregation as a result of QPX-related mortalities. After keeping only variants significantly different between the two subpopulations (see Materials and Methods), the mean overall F_{st} between both subpopulations (Time 0 vs. Time 18 months) had a value of 0.072 in OYB and 0.118 in ARC showing a slightly larger genetic divergence in ARC than in OYB. The localization of the number of significant SNPs in the genome, i.e. in coding and noncoding regions, is provided in Table S7. The type of consequences of the variants in coding regions for the predicted proteins are described in Table S8. All significant variants are listed in Table S9.

Table 1
Assembly and annotation statistics: Global statistics of the genome assembly and annotation of *Mercenaria*. Statistics of the *Ruditapes philippinarum* genome [32] is given for comparison.

	M. mercenaria	*R. philippinarum*
Genome length	2.45 Gb	1.12 Gb
N50	610 Kb	56.5 Mb
Number of genes (protein coding)	42,214	27,652
Number of monoexonic	11,686	ND
Repeat elements (%)	3	26.38
Mean genes length (bp)	11,261	12,875
Mean CDS length (bp)	1059	ND
Mean exons length (bp)	198	232
Mean introns length	1891	1230
Mean number of exons per gene	5.35	4.17
Gene coverage	19.6%	21.6%
CDS coverage	1.8%	ND
Gene BUSCO assessment	88.12%	91.0%

Genes enriched with variants in survivors represented a total of 180 and 173 genes in OYB and ARC clams, respectively, including 100 and 120 genes displaying non-synonymous variants. All genes enriched in survivor clams were categorized using BRITE and are represented in Fig. 1. Most of the genes were related to genetic information processing. Particularly, genes related to ubiquitin system, chromosome conformation, membrane trafficking and messenger RNA biogenesis were overrepresented in terms of number of genes and mean frequency of the variants. Categories known for organism defense against stress were also overrepresented, such as genes related to exosomal proteins, receptor proteins or cell adhesion. Two categories of enzymes were also overrepresented: hydrolases and oxidoreductases. Among these genes, some variants induced the presence of a stop codon.

From all genes, 163 and 221 had variants depleted in clams collected at Time 18 months in OYB and ARC, respectively. Among these, 92 and 147 had non-synonymous variants in OYB and ARC, respectively. Most of the categories were the same as the categories highlighted for genes enriched at Time 18 (Fig. 2).

2.4. Variants enriched at time 18 months

2.4.1. Common in both populations (OYB and ARC)

Among variants identified to be differentially represented between Time 0 and Time 18 months clams, 308 variants were found at the same positions in the genome for both OYB and ARC (Table S7). Among these, only 11 variants were found in a CDS (Table S7). Four of these variants had opposite frequencies in OYB compared to ARC (i.e. higher frequency for OYB but lower frequency for ARC at Time 18 compared to Time 0 or vice versa, Table S9 “common genes” A). Three variants were found to encode synonymous variations in the translated genes (Table S9 “common genes” B). Finally, three variants leading to a missense were found in three genes (Table S9 “common genes” C): mRNA.scaffold2179.16.1: “titin [EC:2.7.11.1]” (2 variants at same

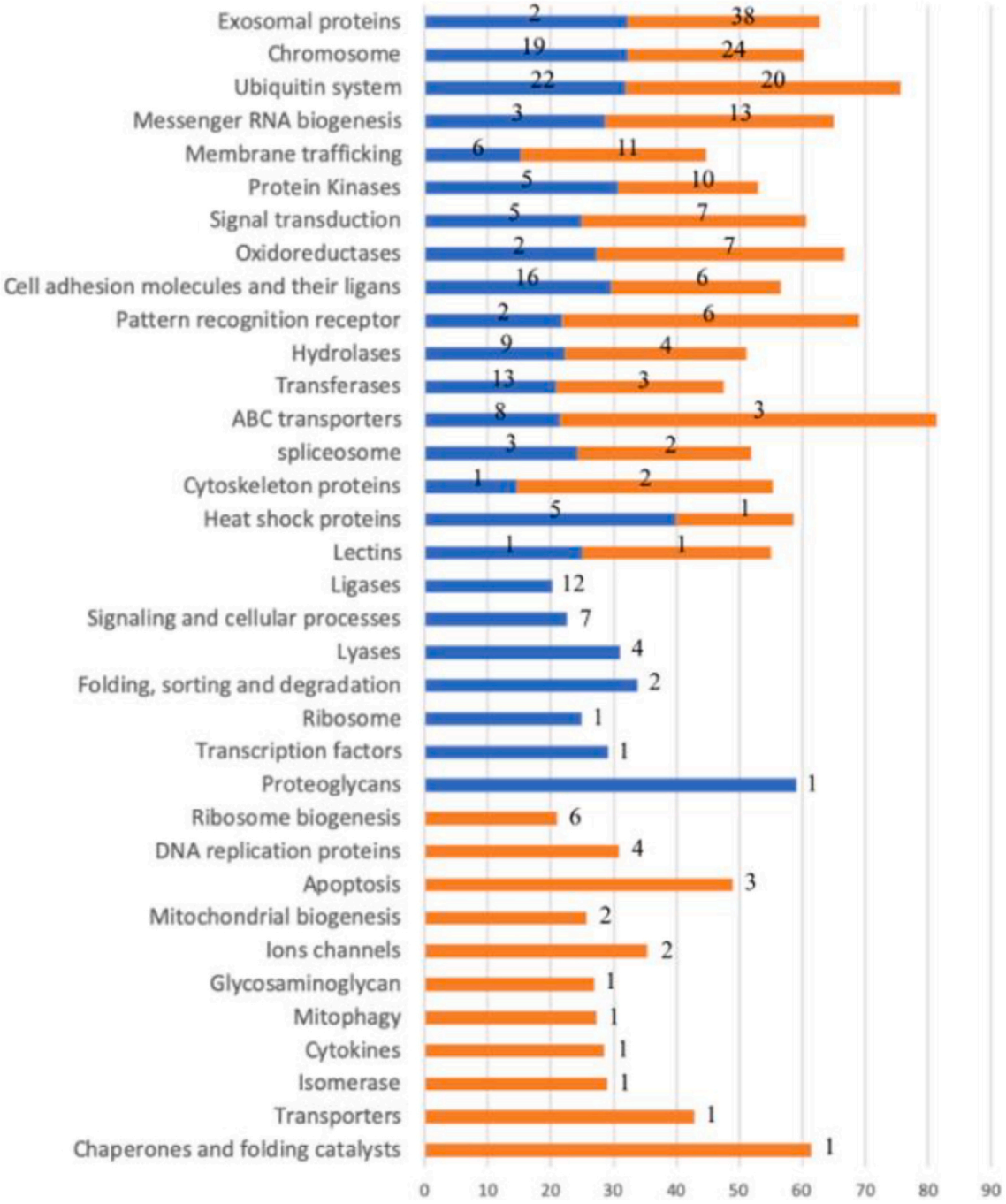

Fig. 1. Global functional categories having variants enriched at Time 18 months: Mean variant frequency per BRITE and KEGG functional categories for variants enriched in genes at Time 18 months in OYB (blue) and ARC (orange). Numbers on barplot represent the number of genes found in that category. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

positions) with a lower frequency in clams from both populations at Time 18 (discussed in "Variants depleted at Time 18"), mRNA.scaffold790.9.4: "Ankyrin" and mRNA.scaffold607.147.3: "beta-galactosidase [EC:3.2.1.23]"-related gene with a higher frequency at Time 18 months.

mRNA.scaffold790.9.4 has two domains, the vacuolar protein sorting-associated domain (VPS9) followed by an Ankyrin-repeats domain. Ankyrin is known to link membrane proteins to the cytoskeleton of a cell in eukaryotes [39]. However, the addition of VPS9 domain was found in the ankyrin repeat domain-containing protein 27, also named Varp (VPS9-ankyrin repeat domain-containing protein), which regulates endosome dynamics [40] and a guanine nucleotide exchange factor by interacting with different Rab proteins [41]. The regulation of the endosome is important for different processes like protein transportation, membrane trafficking and signaling pathways directly related to immune response [42]. As this variant was enriched in both populations at Time 18 months as compared to Time 0, it may suggest that endosomal transport and endocytosis performance is probably associated with a better performance of clams against QPX.

mRNA.scaffold607.147.3 is a beta-galactosidase with Glycoside Hydrolase Family 35 domain that hydrolyzes lactose into glucose and galactose. This enzyme was well studied in plants and its functions include degradation of cell walls for flower senescence or fruit ripening [43–45]. Its functions in animals remain unclear but might be involved

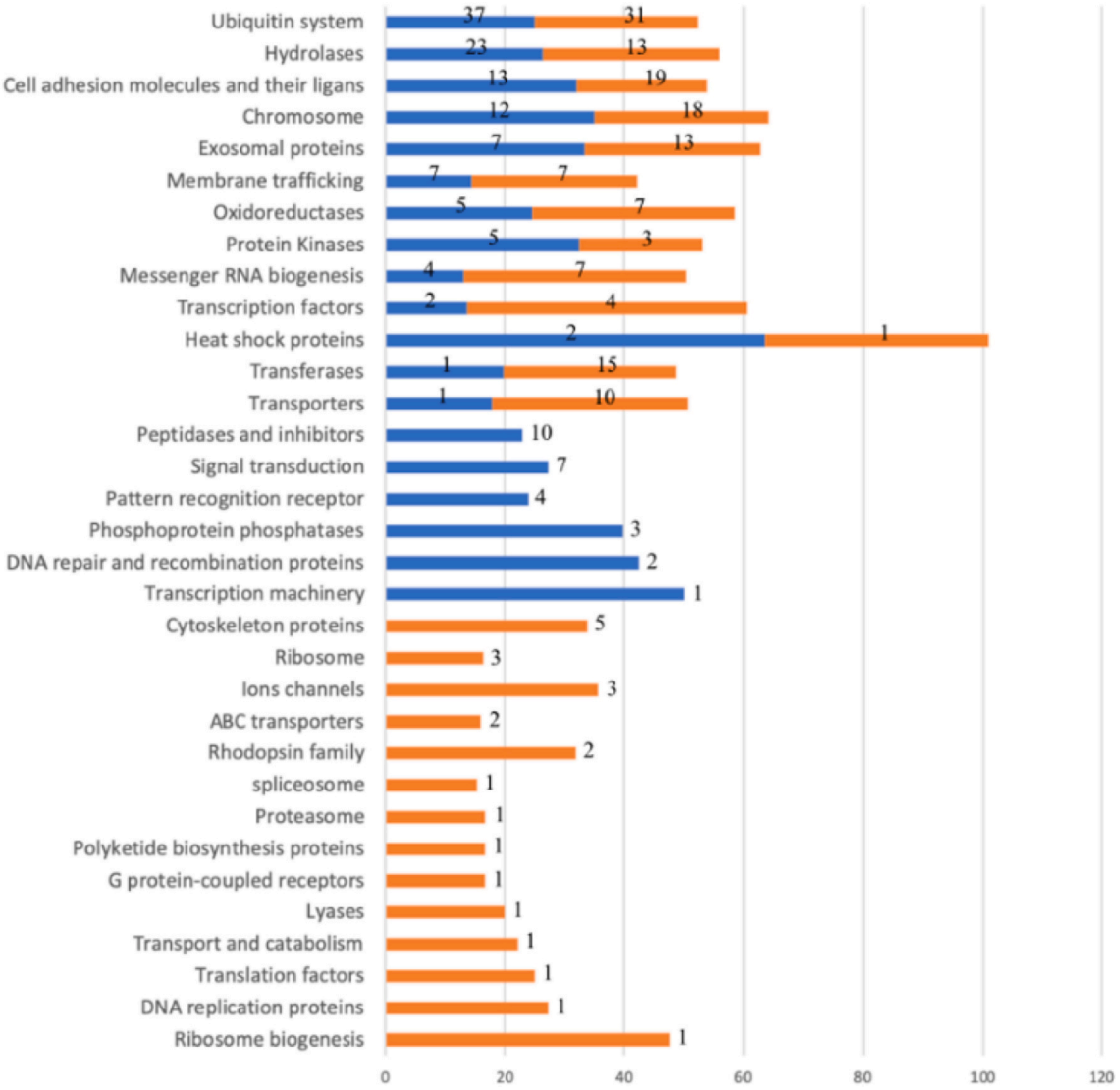

Fig. 2. Global functional categories having variants depleted at Time 18 months: Mean variant frequency per BRITE and KEGG functional categories for variants depleted in genes at Time 18 in OYB (blue) and ARC (orange). Numbers on barplot represent the number of genes found in that category. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in energy metabolism (suggesting that selection in response to QPX may be related to energetic performance) or degradation of microbial cell walls. It should be noted that QPX is known to produce a thick layer of mucin-like glycoproteins that contributes to its pathogenesis [46], while beta-galactosidase is capable of degrading glycoproteins as described for example in *Streptococcus pneumoniae* [47]. Therefore, the enrichment of this SNP among survivor clams may be related to a better performance of this enzyme in the degradation of QPX-secreted products.

Besides the variants found at the same position in a gene, we also analyzed variants found in the same genes (but at different positions) in both OYB and ARC. Eight genes were found to be enriched at Time 18 months in both populations, but at different positions, including 2 genes having synonymous variants in both populations and one gene having a synonymous variant in OYB and a missense variant in ARC (Table S9 "common genes" D). The 5 other genes displayed missense variants in both populations and included mRNA.scaffold1003.9.1: "fibrillin 2/3", mRNA.scaffold1111.79.2: "NLR family CARD domain-containing protein 3", mRNA.scaffold1158.52.1: "pre-mRNA-splicing factor CWC26", mRNA.scaffold23.145.2: "PERQ amino acid-rich with GYF domain-containing protein" and one gene mRNA.scaffold1630.56.1 with unknown function (Table S9 "common genes" D).

mRNA.scaffold1003.9.1, the fibrillin-related gene annotated by the best match in KEGG and NR database was composed of EGF-like domain, EGF-CA domain, growth factor receptor cystein-rich domain and a transmembrane domain. Fibrillin is a cell matrix protein known to contain these domains with the alternate of two other domains TB and TGF-β [48]. Because none of these domains was detected, this protein may not be a fibrillin even though the presence of the EGF domains supports a role in the cell matrix or in protein-protein interactions. Further, the variant was not found in a domain, and as a consequence the impact of this missense variant on the protein remains unclear.

The nucleotide-binding domain, leucine-rich-containing (NLR) family CARD domain-containing proteins are a class of cytoplasmic pattern-recognition receptors. Most of these proteins are involved in immunity. The NLR protein 3 was described to be involved in pathogen or damage sensing by activating the inflammasome [49,50]. This protein was found to play an important role in animal immunity as demonstrated for canid cells infected by *Neospora caninum* or mice exposed to *Leishmania amazonensis* [51,52]. The protein in *M. mercenaria* (mRNA.scaffold1111.79.2) did not share all the domains present in the

mice NLRP3 but had NACHT_NTPase, HEPN_DZIP3 domains and a transmembrane domain. The NACHT nucleoside triphosphate (NTPase) domain was described as having a role in apoptosis and transcription activation [53]. Overall, it is most likely that this protein is indeed involved in clam immunity even though variants were found outside all predicted domain.

mRNA.scaffold1158.52.1 gene perfectly matches the transcript "comp181525_c0_seq3" shown by [25] to be significantly upregulated during the inflammatory response of clams against QPX (Table S10). The domains found in the protein are "large tegument protein UL36, provisional", "DNA translocase FtsK" and "Kazal type serine protease inhibitors". This gene has an unknown function. However, the importance of this protein appears to be crucial for clam resistance given the enrichment of the variant in survivor clams from both lines (OYB and ARC) and its upregulation during infection [25]. In fact, QPX is known to produce and secrete a range of serine proteases that are considered key for successful infection [46,54]. In this context, the ability of the clam to inhibit these proteases is likely essential to limit disease development and enhance survivorship.

mRNA.scaffold23.145.2, the PERQ amino acid-rich with GYF domain-containing protein contains one domain annotated as "patched family". The "patched family" domain is a family of receptors known in the hedgehog signaling pathway permitting the growth of tissues or organs described in different organisms [55].

2.4.2. Stop-gained variants

Variants inducing a stop codon were found in both population but only one gene shared the same variant in both populations (mRNA.scaffold427.20.1). This gene was annotated as "mitogen-activated protein kinase 4 [EC:2.7.11.25]" using KEGG database and the best match on NR was "ubiquitin-conjugating enzyme E2 Q2-like isoform X2 [*Mizuhopecten yessoensis*]". The related protein contains conserved "ubiquitin-conjugating enzyme E2 domain", "ubiquitin-conjugating enzyme/RWD-like domain", "protein kinase domain" and "serine/threonine protein kinase" domains. This protein is likely to be involved in a signaling pathway from stimulus-activated receptors responding to a stress [56]. The variant detected shortened the protein to 112aa instead of 506aa cutting in half the kinase domain and losing the ubiquitin-conjugating enzyme domain that could inhibit the production of this protein and disable the signaling of a possible stress (Fig. 3). In ARC, this variant was enriched at Time 18 months while it was depleted in OYB. Finding the same variant in both geographic populations but enriched in opposing subpopulations (Time 0 vs. Time 18 months) may be caused by contrasting genotype-environment interactions.

Three other genes presented stop-gained variants in OYB only and enriched at Time 18: mRNA.scaffold1485.16.2, mRNA.scaffold24.27.1 and mRNA.scaffold239.36.2 (Fig. 3). mRNA.scaffold1485.16.2 was annotated as "deleted in malignant brain tumors 1 protein" by KEGG while the best match on NR was "PREDICTED: MAM and LDL-receptor class A domain-containing protein 2-like [*Crassostrea gigas*]". The gene predicted in the oyster *C. gigas* encodes a conserved protein containing, among others, multiple MAM (meprin-A5-protein tyrosine phosphatase mu), SRCR (scavenger receptor cysteine-rich, part of a family of soluble or membrane-bound receptors [57]), LDLR (low density lipoprotein receptor), Trypsin and Frizzled domains that were conserved in *M. mercenaria* but other domains were found in the clam only and were not detected in the oyster and inversely. For instance, EGF-like, EGF-CA (EGF-like calcium binding) and Ig-like (immunoglobulin like) domain were found in *M. mercenaria* but not in *C. gigas*, while SEA (Sperm protein, Enterokinase and Agrin), CUB and a transmembrane domain were detected in *C. gigas* but not in *M. mercenaria*. Moreover, EGF-CA domain is a resistant structure described to have a role in spacer unit, in protein–protein interactions, or in structural stabilization [58,59] and the superfamily of Immunoglobulin-like domain was described as a heterogenic group of proteins sharing a common fold and often involved in immunorecognition [60]. While the exact function of this gene is unclear, it most likely encodes a receptor protein. The variant shortened the protein from 4782aa to 3269aa, removing the trypsin domain and 4 MAM and SRCR domains and might induce a non-functional protein. The functional annotation (KEGG) of the second gene (mRNA.scaffold24.27.1) was "receptor-type tyrosine-protein phosphatase delta [EC:3.1.3.48]" and the best match against NR was "uncharacterized protein LOC111119286 isoform X2 [*Crassostrea virginica*]". It contains a protein-tyrosine phosphatase PTP domain, a laminin EGF domain, EGF-like domains, a growth factor receptor cysteine-rich domain and a transmembrane domain with a cytoplasmic and extracellular region. Receptor-type tyrosine-protein phosphatase is known to be a signaling molecule regulating a range of different cellular processes including cell growth, or differentiation in human cells [61]. As a consequence of this variant, the protein is truncated to only 20aa instead of 1137aa likely inferring a loss-of-function. The third and last gene having a stop-gained variant, mRNA.scaffold239.36.2, had no similarities with KEGG or NR databases and no conserved domain.

Three genes had stop-gained variant only in ARC population and enriched at Time 18: mRNA.scaffold1946.65.1, mRNA.scaffold1338.85.1 and mRNA.scaffold5889.5.1 (Fig. 3). The first gene, mRNA.scaffold1946.65.1, had no similarities with KEGG database but had a best match on NR with "52 kDa repressor of the inhibitor of the protein kinase-like" (P52rIPK), containing a DUF domain, a zinc finger domain TTF-type and a ribonuclease H-like domain. The P52rIPK described in humans contains also the same DUF domain but has a different type of zinc finger domain. The function of this gene was described as involved in stress signaling [62]. mRNA.scaffold1338.85.1 is annotated as "pre-rRNA-processing protein TSR1 homolog isoform X1" by KEGG and had a best match on NR with "pre-rRNA-processing protein TSR1 homolog isoform X1 [*C. virginica*]" and contained the 40s ribosome biogenesis protein Tsr1 and BMS1 C-terminal domain. This protein is involved in the complex forming the ribosome. The variant reduces the protein from 357aa to 225aa inducing the loss of the 3-last aa of the domain. Finally, mRNA.scaffold5889.5.1 had no similarities in neither databases and had no conserved domains.

2.5. Variants depleted at time 18

2.5.1. Common variants in both populations OYB ARC

Two variants were found depleted at Time 18 in the same gene and at the same positions in both populations. This gene, mRNA.scaffold2179.16.1, annotated as "titin [EC:2.7.11.1]" by alignment on KEGG database, is not likely to be a titin [63] as it contains not only Ig-like domain but also Roc domain, Ras of Complex, domain of DAPkinase, C-terminal of Roc (COR) domain, P-loop containing nucleoside triphosphate hydrolase domain, and Death domain that were not described in titin proteins. Variants were found in each population in Ig-like domain, in Roc domain and P-loop containing nucleoside triphosphate hydrolase domain. Proteins with these later domains were previously described. For example, the ROCO proteins [64], particularly DAPk, are composed of Roc, Cor and death domain [65] with an addition of ankyrin and kinase domain. While these later domains were not detected in mRNA.scaffold2179.16.1, the Ig-like domain was present. These ROCO proteins were shown to play a role in apoptosis [66,67], and could play a similar role in clams. The fact that the same variant was depleted in clams from both lines (OYB and ARC) sampled at Time 18 suggests that the detected mutation results in a higher susceptibility to QPX.

Overall, variants in 8 genes were found with a lower frequency at Time 18 from both populations (OYB and ARC) compared to Time 0. Among these, 3 genes included missense variants in one of the populations but only synonymous variants in the second one. The 5 other genes were found having missense variants in both populations. These included mRNA.scaffold5874.7.1: "inhibitor of apoptosis 1 [Hyriopsis schlegelii]", mRNA.scaffold156.57.3: "Baculoviral IAP repeat-containing protein 2/3", mRNA.scaffold146.46.1: "NACHT, LRR and PYD

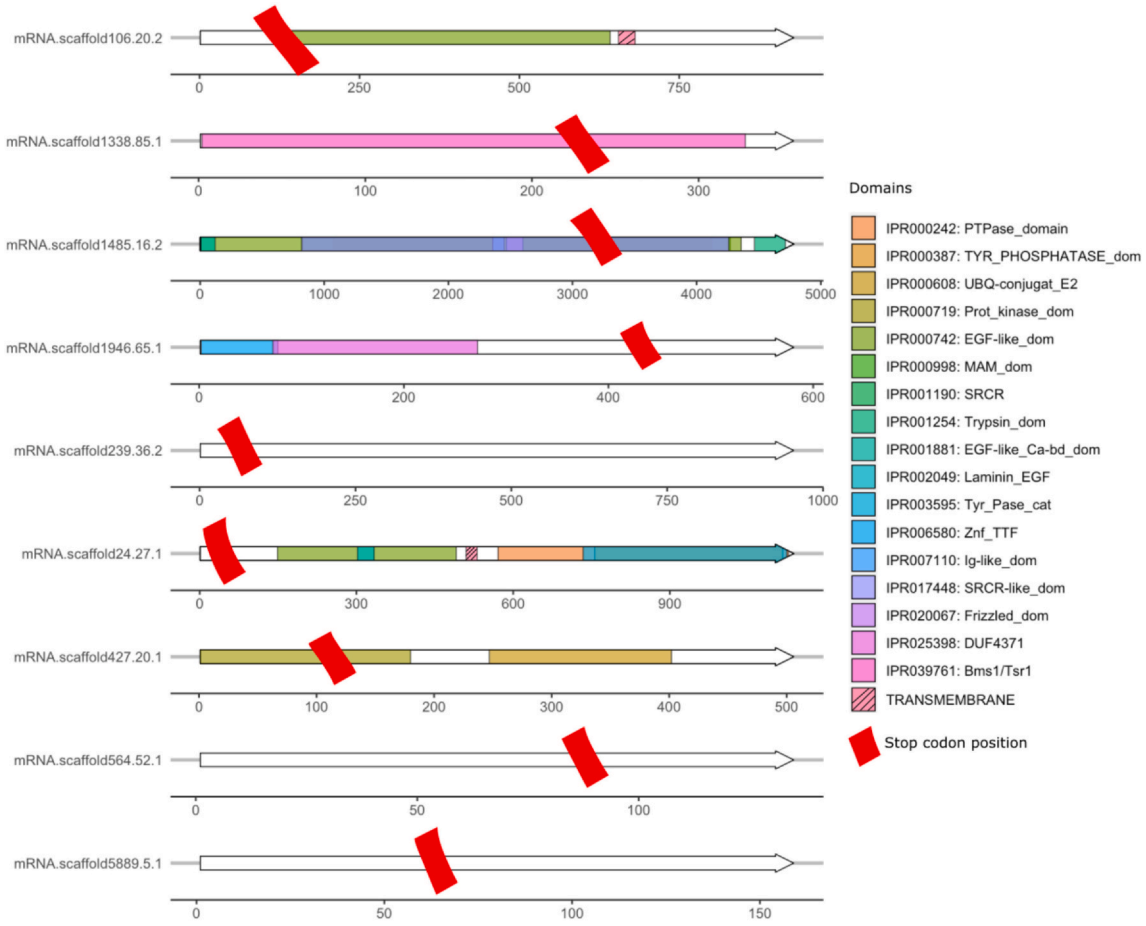

Fig. 3. Proteins having stop codon variant: Representation of all proteins having a variant inducing a stop codon. Each box represents a domain predicted using Interproscan with the IPR id and the short IPR description. The position of the stop codon is shown in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

domains-containing protein 3", mRNA.scaffold1427.119.1: "mucin-2", and mRNA.scaffold2179.16.1: "Titin" related genes (Table S9 "common genes" D). The three first genes are involved in apoptosis pathways. Apoptosis is a major host immune mechanism that contributes to the prevention of the spread of pathogens in a broad range of animals, including molluscs [68–71]. For example, regulation of apoptosis-related genes was suggested to represent a main mechanism for the resistance of the flat oyster (*Ostrea edulis*) to the protozoan parasite *Bonamia ostreae* [72]. Similarly, our previous investigations showed a marked regulation of apoptotic pathways in clams during infection with QPX [24]. The fact that we found these variants in the populations at Time 0 suggests that the outcome resulting from these mutations leads to an increased susceptibility to the disease. The two other genes (mRNA.scaffold1427.119.1: "mucin-2", and mRNA.scaffold2179.16.1 also shown in Table S9 "common genes" C) contains Ig-like domain as described above for mRNA.scaffold2179.16.1 (which had a variant at the same position in OYB and ARC populations). mRNA.scaffold1427.119.1 had only this domain but the variants were not found in the domain region. The best match of this gene on NR database was with the predicted gene "Polycystic kidney disease protein 1-like 3" from *Mizuhopecten yessoensis*, sharing only 30% of identity and the KEGG prediction was "mucin-2". Mucin-2 is characterized by the presence of multiple domains tandem repeats rich in threonine and proline which is not the case for this protein. As none of those predicted functions seems to represent an exact match and that members of the immunoglobulin superfamily are found in several proteins with different functions, the function of this particular gene is unclear.

2.5.2. Stop-gained variants

One gene with variant inducing a stop codon was found in one gene in OYB only and depleted at Time 18. It (mRNA.scaffold564.52.1) had no similarities with KEGG or NR databases and had no conserved domains, which makes it difficult to infer a function in silico (Fig. 3). In ARC, one stop-gained variant was found in mRNA.scaffold106.20.2 annotated as "receptor-type tyrosine-protein phosphatase kappa [EC:3.1.3.48]" by KEGG and the best match against NR was "Receptor-type tyrosine-protein phosphatase T [*Crassostrea gigas*]". The conserved domains found are 9 EGF-like (epidermal growth factor) domain and 4 Furin-like repeats. EGF-like domains are commonly found, in tandem, in extracellular or transmembrane proteins that are involved in different functions like blood coagulation, fibrinolysis, cell surface receptors and cell matrix proteins [73]. This protein has a transmembrane domain with a cytoplasmic and non-cytoplasmic part that indicate a role at the interface of the cell membrane. The presence of the stop codon reduces the protein to 633aa (instead of 929aa, Fig. 3) removing the transmembrane domain and one of the EGF-like domains. Interestingly, this gene is thought to be involved in clam immune response to QPX as it was found to be up-regulated in infected clams [25]. Therefore, the mutation detected here might translate into an absence of this protein at the cell surface and a failure of activation of underlying immune pathways.

2.6. Cross mapping with differential expression

As discussed above, the transcriptome used for the annotation of the clam genome derives from the analysis of *M. mercenaria* RNAseq

samples during QPX infection done by Wang et al. [24]. This study characterized the immune response of *M. mercenaria* during QPX infection comparing transcripts from inflammatory clam tissues surrounding parasite cells from infected clams against control samples collected from healthy clams [25]. For this analysis, we cross-mapped the genes found to be differentially expressed by Wang et al. [24] with the genes having variant enrichment (in CDS regions) using a BLASTN alignment [74], keeping the best reciprocal hit (BRH) with more than 50% of the transcriptome aligned. The concordance of variant enrichment with gene regulation during infection suggests a strong relation to QPX resistance. Nonsynonymous variants were found in a total of 10 genes that were also shown to be differentially-regulated during QPX infection, including 2 genes in OYB, 7 genes in ARC and one common to both populations. mRNA.scaffold1158.52.1 was found in both OYB and ARC and up regulated in infected clams (discussed above). The two genes found in OYB (ATP-binding cassette, subfamily A (ABC1), member 3 and an uncharacterized gene) had variants enriched at Time 18 months and corresponded to genes repressed during the infection. In ARC, one gene annotated as uncharacterized had variants enriched at Time 18 months and was found to be down regulated during the infection. The last six genes were found with variant depleted at Time 18 in ARC. The first 3 genes were found down regulated during the infection including mRNA.scaffold106.20.2 (already discussed above), mRNA.scaffold1046.19.1, annotated as Frizzled 1/7 protein and mRNA.scaffold424.15.2 annotated as hydroxymethylglutaryl-CoA reductase (NADPH). mRNA.scaffold1046.19.1-related protein is a transmembrane receptor protein described as part of the Wnt pathway which is a signaling pathway involved in the regulation of multiple immune functions in animals [75–79]. mRNA.scaffold424.15.2 enzyme is known to participate in biosynthesis of cholesterol. The three last genes were found to be up regulated during infection. First, mRNA.scaffold12.47.1 was annotated as Arachidonate 5-lipoxygenase, an enzyme transforming fatty acids. Second, mRNA.scaffold821.18.1, annotated as "Apoptotic chromatin condensation inducer in the nucleus", had no conserved domain unlike the corresponding proteins in human, oyster (*C. gigas*) or scallop (*Mizuhopecten yessoensis*) (Q9UKV3, K1QEG1 and KP79_PYT14214 in Uniprot database, respectively) which have conserved domains as Acinus_RRM or SAP domains. As no conserved domains were detected, it is difficult to infer a function to this gene. Finally, mRNA.scaffold1739.15.2, annotated as "Limbic system-associated membrane protein" (LSAMP) had conserved Ig-like and transmembrane domains. This protein is a neuronal surface glycoprotein involved in the immune system [80]. The gene associated was described to be overexpressed in the oyster *Pinctada martensii* in response to the gram-negative bacteria *Vibrio alginolyticus* [81].

3. Conclusion

Here, we generated a first assembly of the *M. mercenaria* genome and identified genetic markers associated with clam susceptibility or resistance to QPX. Since we analyzed two clam populations derived from two separate hatcheries, one cannot rule out the possibility of a carryover effect from early life stages that could modulate resistance to QPX disease in adult clams, although our results more strongly support a genetic basis for disease resistance. Susceptible populations showed markers in proteins implicated in apoptosis and in receptor proteins often associated with gained stop codons. Receptors are important in inducing signaling pathway making the cell able to react to a potential stress and apoptosis regulation is one of the most common immune mechanisms in animals. If variants alter the production of proteins implicated in these processes, the host could become susceptible to stress and/or infection. Resistant populations displayed specific variants in receptor proteins and in proteins having a role in stress signaling, hydrolase activity immunorecognition and cell matrix proteins. Nevertheless, many proteins displaying significant variants did not have any similarity in public databases making function inference impossible. Overall, the study allowed the identification of a large set of markers that can be used as solid basis for evaluating marker-assisted or genomic selection of hard clams for resistance against QPX.

4. Materials and methods

All bioinformatic analyses were performed on the high-performance computing server (Bridges) of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 [82].

4.1. Collection of Mercenaria mercenaria

For the draft genome, DNA was extracted using classical phenol-chloroform methods from freshly-collected adductor muscle of an adult clam derived from an aquaculture line bred and grown at Frank M. Flower and Sons in Oyster Bay, New York.

For RAD sequencing, phenol-chloroform extraction (detailed below) was also used to generate DNA samples from clams collected before and after undergoing QPX-related mortalities. Briefly, 1-year old juvenile clams were produced from two geographically segregated populations: juveniles from the OYB clam line derived from Oyster Bay were provided by Hatchery 1 located in Oyster Bay, NY, and juveniles from the ARC clam line were provided by Hatchery 2 located in Dennis, Massachusetts. Samples submitted to RAD sequencing consisted of randomly selected clams (48/population/time point, totaling 192 clams) collected at Time 0 (before deployment in the field, confirmed to be free of QPX using histopathological techniques) or after 18 months of deployment in an enzootic site in Massachusetts. The selected field site was located in Barnstable Harbor (Latitude: 41.7167761, Longitude: −70.2661323) where QPX has been continuously present since its first detection in the 1990's. Salinity in that harbor during the deployment period ranged from 23.0 to 30.5 ppt and temperature from 1.9 to 24.5 °C. Each clam strain was deployed in 3 replicate plots (1.5 m × 1.5 m netted plots) in a general randomized block design with about 1000 clams per plot. Deployed clams were monitored over time for mortality and were sampled after 18 months for QPX disease assessment and genotyping. For disease diagnostics, a cross section including mantle, gill, kidney/heart, digestive gland, gonad, and foot was sampled and submitted to standard histopathology techniques [5]. An aliquot of mantle tissue was fixed in ethanol for genotyping (detailed below).

4.2. Genome sequencing

4.2.1. Illumina Hiseq

The library was generated using the NxSeq® AmpFREE Low DNA Library Kit Library Preparation Kit (Lucigen) according to the manufacturer's recommendations. Dual-indices adaptors were purchased from IDT. The library was quantified using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were normalized, pooled, then denatured in 0.05 N NaOH and neutralized using HT1 buffer. ExAMP was added to the mix following the manufacturer's instructions. The pool was loaded at 200pM on an Illumina cBot and the flowcell was ran on a HiSeq X for 2 × 151 cycles (paired-end mode). A phiX library was used as a control and mixed with libraries at 1% level. The Illumina control software was HCS HD 3.4.0.38, the real-time analysis program was RTA v. 2.7.7. Program bcl2fastq2 v2.20 was finally used to demultiplex samples and generate fastq reads.

4.2.2. Pacbio sequel

The DNA library was prepared following the Pacific Biosciences 20 kb Template Preparation Using BluePippin Size-Selection System

protocol. 7.5 μg of high molecular weight genomic DNA (final volume of 100 μl) was sheared using the Covaris g-TUBES (Covaris Inc., Woburn, Massachusetts, USA) at 4000 RPM for 60 s on each side, on an Eppendorf centrifuge 5424 (Eppendorf, Hamburg, Germany). The DNA damage repair, end repair and SMRT bell ligation steps were performed as described in the template preparation protocol with the SMRTbell Template Prep Kit 1.0 reagents (Pacific Biosciences, Menlo Park, CA, USA). The DNA library was size selected on a BluePippin system (Sage Science Inc., Beverly, MA, USA) using a cutoff range of 12 kb to 50 kb. The sequencing primer was annealed with sequencing primer v3 at a final concentration of 0.83 nM and the Sequel 2.1 polymerase was bound at 0.5 nM. The libraries were sequenced on a PacBio Sequel instrument at a loading concentration (on-plate) of 6 pM using the diffusion loading protocol, Sequel Sequencing Plate 2.1, SMRT cells 1 M v2 and 10 h movies.

4.2.3. 10× Chromium

The gDNA was size selected on a BluePippin system (Sage Science Inc., Beverly, MA, USA) using a cutoff range of 40 kb to 80 kb. The 10× Chromium shotgun libraries were prepared following the Chromium Genome Reagent kits v2 User Guide RevB protocol, using the Chromium™ Genome Library & Gel Bead Kit v2, Chromium™ Genome Chip Kit and Chromium™ i7 Multiplex Kit (10× Genomics Inc., Pleasanton, CA, USA). The 10× Chromium shotgun libraries were quantified using the Quant-iT(tm) PicoGreen(r) dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The sequencing was performed on HiSeq X following the same method described previously.

4.3. Genome assembly

A de-novo draft assembly was generated using sequences derived from three sequencing technologies, PacBio, Illumina PE and 10× Chromium. First, the Masurca hybrid assembler v3.2.8 [83] was used to assemble ~69-fold raw coverage of Illumina PE shotgun reads along with ~15-fold raw coverage of Pacbio Sequel reads (8 cells), assuming a 2 Gb estimated genome size. This produced a first assembly (Table S1, S2). Next, the 10× Chromium reads (~67 fold raw coverage) were used by the ARCS v1.0.4 (https://github.com/bcgsc/arcs) together with LINKs pipeline v1.8.5 (https://github.com/bcgsc/LINKS) to help scaffold and make the Masurca assembly more contiguous (Table S3, S4). Then, a BLASTN version 2.7.1 [74] of the scaffolds against NT was computed removing the scaffold with no match and scaffold having only bacterial hits (Table S4). BUSCO analysis (https://busco.ezlab.org/, using the eukaryota_odb9 database) was run on the final scaffolded assembly (Table S5).

4.4. Genome annotation

The genome annotation of *M. mercenaria* was done using the following pipeline. First, low complexity regions were masked using dustmasker from BLAST version 2.7.1 [74]. RepeatMasker version open-4.0.9 [84] was used with Dfam 3.0 database in order to mask repeated elements in the genome. Then, the transcriptome of *M. mercenaria* [25] and proteins from UniProt Mollusca [85] were mapped on the repeated-masked genome with blat to rapidly identify the position of the sequences. To refine the alignments, only matches with more than 80% identity were kept and given to exonerate version 2.4.0 [86] using est2genome model and protein2genome model for the transcriptome and proteins mapping, respectively. Transcriptome mapping was filtered with at least 90% of identity and at least 85% of the transcript length matching the genome. Proteins mapping were filtered with at least 50% of identity and at least 50% of the protein length matching the genome. An ab-initio prediction was done using default workflow of maker version 2.31.9 [87] with a first round of training of SNAP [88] on the transcriptome sequences [25]. A second round was done using the resulting prediction. Finally, Gmove [89] combined all different resources listed above to build the final gene set.

Once identified, the genes were translated into proteins in order to unveil their function. Each protein was aligned to the KEGG database [90] using BLAST [74] with a minimum e-value of 10^{-5} and keeping the best match per protein and those with a score greater than 90% of the best match. Domains were defined using InterProScan 5.36–75.0 [91] with the default parameters. Finally, a BLASTp [74] alignment of the predicted proteins was done against NR [92] by keeping the best three matches.

4.5. Library construction and RAD sequencing

For each of the 192 individuals, DNA was extracted from alcohol-preserved clam tissues (mantle) using a standard phenol-chloroform-isoamyl alcohol (PCI 25:24:1) protocol. After two washes with PCI, DNA was precipitated overnight with absolute ethanol at −20 °C, then centrifuged, washed with 70% ethanol, dried and suspended in PCR-grade water. DNA was then purified using Genomic DNA Clean-up Kit (Macherey-Nagel, Bethlehem, PA) using manufacturer's instructions. All DNA samples were run in a 1% agarose 1× TBE gel and quantified using Qubit spectrometer with High sensitivity dsDNA quantification kit (Invitrogen) according to the manufacturer's instructions. Double-digest RAD-seq (ddRADSeq) individual libraries were produced following [93]. Briefly, for each individual, 200 ng of genomic DNA was digested with *Pst*I and *Mse*I restriction enzymes (New England Biolabs, Ipswich, MA). Barcoded adaptors were then ligated to the digested DNA fragments and purified using Nucleo Mag NGS Clean-up and Size Select Kit (Macherey-Nagel, Bethlehem, PA). Eight microliters of the purified template were used for enrichment and Illumina indexing by PCR using Q5 hot start DNA polymerase (New England Biolabs) (98 °C 30s, 15 cycles 98 °C 10s, 60 °C 20s, 72 °C 30s). A final elongation was done by added buffer, dNTP and primers for 15 min at 72 °C. PCR products were run in a 1% agarose 1× TBE gel, quantified using Qubit spectrometer with High sensitivity dsDNA quantification kit (Invitrogen) according to the manufacturer's instructions and then pooled in equal proportions in two separate libraries (ARC and OYB). A 300–800 base pair size selection of fragments was performed using a 1.5% agarose cassette in a pippin prep equipment (Sage Science). Each fraction was run through a DNA chip (Agilent) in a Bioanalyser (Agilent) to determine mean size of the fraction and quantify using Qubit. Each library (4 nM) was sequenced separately on one lane of HiSeq4000.

4.6. Variant detection

First, reads from RADSeq were demultiplexed and filtered using process_radtags from stacks version 2.3 [94] with -b -c -q -r parameters. Second, reads were mapped onto the genome assembly using CLC Genomics Workbench version 11.0.1 (https://digitalinsights.qiagen.com/) with default parameters. The mapping generated was sorted and given as input to stacks [94] under the ref_map mode with a minimum allele frequency of 0.05 (−-min-maf), a F_{st} correction with a *p*-value cutoff of 0.05, a minimum percentage of samples per population of 20 and 50 (−r) for OYB and ARC dataset respectively (according to QPX-related mortality rate in each population) and a p-value less than 0.05 calculated on the divergence from Hardy-Weinberg equilibrium (−-hwe). SNPs associated with resistance and susceptibility were identified as those showing significant elevated FST values between the time 0 and the time 18 months subpopulation (Corrected AMOVA FST $>$ 0.05 and $P < 0.05$) and with variation/allele frequency shifting in the same direction in both populations.

The clustering of samples according to variant frequency (Fig. S3) was done following this tutorial (https://grunwaldlab.github.io/Population_Genetics_in_R/gbs_analysis.html). Using the gene

annotation prediction, a homemade script in python allowed the detection of the position of each variant in the genome (intergenic, UTR, CDS or intron) and the type of consequence the variant has on the gene (Missense, Synonymous, stop gained). Genes carrying the resistant SNPs were identified and studied for possible functions in immune response or resistance, along with their expression profile under QPX challenge.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.08.036.

Funding

This research was primarily supported by a grant from the United States Department of Agriculture (Award 2016-70007-25759), with additional support from the National Oceanic and Atmospheric Administration's Aquaculture Program via the New York Sea Grant (NA18OAR4170096, East Coast Hard Clam Selective Breeding Collaborative) and the National Science Foundation (IOS-1656753).

Acknowledgments

The authors would like to thank farmers (Dave Relyea, Richard Kraus) for providing experimental clams, and personnel of the Marine Animal Disease Laboratory for help with sample processing.

References

[1] R.E. Drinnan, E.B. Henderson, 1962 Mortalities and Possible Disease Organisms in Nequac Quahogs, New Brunswick, Canada (1963).
[2] S.K. Whyte, R.J. Cawthorn, S.E. McGladdery, QPX (Quahaug Parasite X), a pathogen of northern quahaug *Mercenaria mercenaria* from the Gulf of St. Lawrence, Canada, Dis. Aquat. Org. 19 (2) (1994) 129–316.
[3] R. Smolowitz, A review of QPX disease in Thenorthernquahog (=hard clam) *Mercenaria mercenaria*, J. Shellfish Res. 37 (4) (Oct. 2018) 807–819.
[4] L. Ragone Calvo, J. Walker, E. Burreson, Prevalence and distribution of QPX, Quahog Parasite Unknown, in hard clams *Mercenaria mercenaria* in Virginia, USA, Dis. Aquat. Org. 33 (3) (Jul. 1998) 209–219.
[5] R. Smolowitz, D. Leavitt, F. Perkins, Observations of a Protistan disease similar to QPX in *Mercenaria mercenaria* (hard clams) from the coast of Massachusetts, J. Invertebr. Pathol. 71 (1) (Jan. 1998) 9–25.
[6] S.E. Ford, J.N. Kraeuter, R.D. Barber, G. Mathis, Aquaculture-associated factors in QPX disease of hard clams: density and seed source, Aquaculture 208 (1–2) (May 2002) 23–38.
[7] R. Gast, et al., Environmental distribution and persistence of Quahog Parasite Unknown (QPX), Dis. Aquat. Org. 81 (3) (Sep. 2008) 219–229.
[8] Q. Liu, B. Allam, J.L. Collier, Quantitative real-time PCR assay for QPX (Thraustochytriidae), a parasite of the hard clam (*Mercenaria mercenaria*), Appl. Environ. Microbiol. 75 (14) (Jul. 2009) 4913–4918.
[9] S. Geraci-Yee, J. Collier, B. Allam, World Aquaculture Society: Aquaculture 2019 – Presentation Abstract, (2019).
[10] L. Calvo, S. Ford, J. Kraeuter, D. Leavitt, R. Smolowitz, Influence of host genetic origin and geographic location on Qpx disease in northern Quahogs (=hard clams), *Mercenaria mercenaria*, J. Shellfish Res. 26 (1) (2007) 109–119.
[11] S.F. Dahl, M. Perrigault, B. Allam, Laboratory transmission studies of QPX disease in the hard clam: interactions between different host strains and pathogen isolates, Aquaculture 280 (1) (2008) 64–70.
[12] M. Perrigault, D.M. Buggé, B. Allam, Effect of environmental factors on survival and growth of quahog parasite unknown (QPX) in vitro, J. Invertebr. Pathol. 104 (2) (2010) 83–89.
[13] M. Perrigault, S.F. Dahl, E.P. Espinosa, B. Allam, Effects of salinity on hard clam (*Mercenaria mercenaria*) defense parameters and QPX disease dynamics, J. Invertebr. Pathol. 110 (1) (May 2012) 73–82.
[14] M. Perrigault, S.F. Dahl, E.P. Espinosa, L. Gambino, B. Allam, Effects of temperature on hard clam (*Mercenaria mercenaria*) immunity and QPX (Quahog Parasite Unknown) disease development: II. Defense parameters, J. Invertebr. Pathol. 106 (2) (Feb. 2011) 322–332.
[15] S.F. Dahl, M. Perrigault, Q. Liu, J.L. Collier, D.A. Barnes, B. Allam, Effects of temperature on hard clam (*Mercenaria mercenaria*) immunity and QPX (Quahog Parasite Unknown) disease development: I. Dynamics of QPX disease, J. Invertebr. Pathol. 106 (2) (Feb. 2011) 314–321.
[16] M.M. Lyons, J.E. Ward, R. Smolowitz, K.R. Uhlinger, R.J. Gast, Lethal marine snow: pathogen of bivalve mollusc concealed in marine aggregates, Limonol. Oceanography 50 (6) (2005).
[17] Q. Liu, J.L. Collier, B. Allam, Seasonality of QPX disease in the Raritan Bay (NY) wild hard clam (*Mercenaria mercenaria*) population, Aquac. Res. 48 (2017) 1269–1278.
[18] C. Brothers, E. Marks, R. Smolowitz, Conditions affecting the growth and zoosporulation of the protistan parasite QPX in culture, Biol. Bull. 199 (2) (2000) 200–201.
[19] J.N. Kraeuter, et al., Evaluation of three northern Quahog (=hard clam) *Mercenaria mercenaria* (Linnaeus) strains grown in Massachusetts and New Jersey for QPX-resistance, J. Shellfish Res. 30 (3) (Dec. 2011) 805–812.
[20] E. Rubin, A. Tanguy, M. Perrigault, E. Pales Espinosa, B. Allam, Characterization of the transcriptome and temperature-induced differential gene expression in QPX, the thraustochytrid parasite of hard clams, BMC Genomics 28 (2014) 15–245.
[21] E. Rubin, A. Tanguy, E. Pales Espinosa, B. Allam, Differential gene expression in five isolates of the clam pathogen, Quahog Parasite Unknown (QPX), J. Eukaryot. Microbiol. 64 (5) (2017).
[22] S. Bassim, B. Allam, SNP hot-spots in the clam parasite QPX, BMC Genomics 19 (2018) 486.
[23] M. Perrigault, B. Allam, Differential immune response in the hard clam (*Mercenaria mercenaria*) against bacteria and the protistan pathogen QPX (quahog parasite unknown), Fish Shellfish Immunol. 32 (2012) 1124–1134.
[24] K. Wang, P. Espinosa, A. Tanguy, B. Allam, Alterations of the immune transcriptome in resistant and susceptible hard clams (*Mercenaria mercenaria*) in response to Quahog Parasite Unknown (QPX) and temperature, Fish Shellfish Immunol. 49 (2016) 163–176.
[25] K. Wang, C. Del Castillo, E. Corre, P. Espinosa, B. Allam, Clam focal and systemic immune responses to QPX infection revealed by RNA-seq technology, BMC Genomics 17 (1) (2016) 146.
[26] K. Wang, P. Espinosa, B. Allam, Effect of 'heat shock' treatments on QPX disease and stress response in the hard clam, *Mercenaria mercenaria*, J. Invertebr. Pathol. 138 (2016).
[27] A.D.M. Dove, P.R. Bowser, R.M. Cerrato, Histological analysis of an outbreak of QPX disease in wild hard clams *Mercenaria mercenaria* in New York, J. Aquat. Anim. Health 16 (4) (Dec. 2004) 246–250.
[28] S.F. Dahl, B. Allam, Laboratory transmission studies of qpx disease in the northern quahog (=hard clam): development of an infection procedure, J. Shellfish Res. 26 (2007) 383–389.
[29] L.M. Ragone Calvo, E.M. Burreson, QPX susceptibility in hard clams varies with geographic origin of brood stock, Dis. Aquat. Org. 208 (2002).
[30] S.F. Dahl, J. Thiel, B. Allam, Field performance and QPX disease progress in cultured and wild-type strains of *Mercenaria mercenaria* in New York waters, J. Shellfish Res. 29 (1) (2010) 83–90.
[31] S. Mun, et al., The whole-genome and transcriptome of the manila clam (*Ruditapes philippinarum*), Genome Biol. Evol. 9 (6) (2017) 1487–1498.
[32] X. Yan, et al., Clam genome sequence clarifies the molecular basis of its benthic adaptation and extraordinary shell color diversity, iScience 19 (Sep. 2019) 1225–1237.
[33] M. Wei, et al., Chromosome-level clam genome helps elucidate the molecular basis of adaptation to a buried lifestyle, iScience 23 (6) (May 2020) 101148.
[37] R. Hinegardner, Cellular DNA content of the mollusca, Comp. Biochem. Physiol. Part A Physiol. 47 (2) (Feb. 1974) 447–460.
[38] J. Sun, et al., Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes, Nat. Ecol. Evol. 1 (5) (Apr. 2017).
[39] S.R. Cunha, P.J. Mohler, Ankyrin protein networks in membrane formation and stabilization, J. Cell. Mol. Med. 13 (11–12) (Nov. 2009) 4364–4376.
[40] X. Zhang, X. He, X.Y. Fu, Z. Chang, Varp is a Rab21 guanine nucleotide exchange factor and regulates endosome dynamics, J. Cell Sci. 119 (6) (Mar. 2006) 1053–1062.
[41] F. Wang, et al., Varp interacts with Rab38 and functions as its potential effector, Biochem. Biophys. Res. Commun. 372 (1) (Jul. 2008) 162–167.
[42] P.A. Gleeson, The role of endosomes in innate and adaptive immunity, Seminars in Cell and Developmental Biology, 31 Elsevier Ltd, Jul. 2014, pp. 64–72.
[43] E.M. O'Donoghue, S.D. Somerfield, L.M. Watson, D.A. Brummell, D.A. Hunter, Galactose metabolism in cell walls of opening and senescing petunia petals, Planta 229 (3) (Feb. 2009) 709–721.
[44] Y.O. Ahn, et al., Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 35, Phytochemistry 68 (11) (Jun. 2007) 1510–1520.
[45] W. Tanthanuch, M. Chantarangsee, J. Maneesan, J. Ketudat-Cairns, Genomic and expression analysis of glycosyl hydrolase family 35 genes from rice (*Oryza sativa* L.), BMC Plant Biol. 8 (2008) 84.
[46] E. Rubin, E. Pales Espinosa, A. Koller, B. Allam, Characterisation of the secretome of the clam parasite, QPX, Int. J. Parasitol. 45 (2–3) (Feb. 2015) 187–196.
[47] V.S. Terra, K.A. Homer, S.G. Rao, P.W. Andrew, H. Yesilkaya, Characterization of novel β-galactosidase activity that contributes to glycoprotein degradation and virulence in *Streptococcus pneumoniae*, Infect. Immun. 78 (1) (Jan. 2010) 348–357.
[48] A. Piha-Gossack, W. Sossin, D.P. Reinhardt, The evolution of extracellular fibrillins and their functional domains, PLoS One 7 (3) (Mar. 2012).
[49] C. Lupfer, T.D. Kanneganti, Unsolved mysteries in NLR biology, Front. Immunol. 4 (2013) no. SEP.
[50] Z. Huang, et al., NACHT, LRR and PYD domains-containing protein 3 inflammasome is activated and inhibited by berberine via toll-like receptor 4/myeloid differentiation primary response gene 88/nuclear factor-κB pathway, in phorbol 12-myristate 13-acetate-induced macrophages, Mol. Med. Rep. 17 (2) (Feb. 2018) 2673–2680.
[51] X. Wang, et al., NLRP3 inflammasome participates in host response to *Neospora caninum* infection, Front. Immunol. 9 (2018) 1791.
[52] L. Bird, Inflammasome: inflammasome is involved in parasite resistance, Nat. Rev. Immunol. 13 (8) (2013) 548.
[53] E.V. Koonin, L. Aravind, The NACHT family – a new group of predicted NTPases implicated in apoptosis and MHC transcription activation, Trends Biochem. Sci. 25 (5) (May 2000) 223–224.
[54] E. Rubin, G.T. Werneburg, E.P. Espinosa, D.G. Thanassi, B. Allam, Identification and characterization of peptidases secreted by Quahog Parasite Unknown (QPX), the

protistan parasite of hard clams, Dis. Aquat. Org. 122 (1) (Nov. 2016) 21–33.
[55] L. Quijada, A. Callejo, C. Torroja, I. Guerrero, The Patched Receptor: Switching On/Off the Hedgehog Signaling Pathway, (2013).
[56] K.J. Cowan, K.B. Storey, Mitogen-activated protein kinases: new signaling pathways functioning in cellular responses to environmental stress, J. Exp. Biol. 206 (7) (Apr-2003) 1107–1115.
[57] M.R. Sarrias, J. Grønlund, O. Padilla, J. Madsen, U. Holmskov, F. Lozano, The scavenger receptor cysteine-rich (SRCR) domain: an ancient and highly conserved protein module of the innate immune system, Crit. Rev. Immunol. 24 (1) (2004) 1–37.
[58] E.J. Boswell, N.D. Kurniawan, A.K. Downing, Calcium-binding EGF-like domains, Encyclopedia of Inorganic and Bioinorganic Chemistry, John Wiley & Sons Ltd, Chichester, UK, 2011.
[59] J. Stenflo, Y. Stenberg, A. Muranyi, Calcium-binding EGF-like modules in coagulation proteinases: function of the calcium ion in module interactions, Biochim. Biophys. Acta Protein Struct. Mol. Enzymol. 1477 (1–2) (07-Mar-2000) 51–63.
[60] D.M. Halaby, J.P.E. Mornon, The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity, J. Mol. Evol. 46 (4) (1998) 389–400.
[61] E.H. Fischer, H. Charbonneau, N.K. Tonks, Protein tyrosine phosphatases: a diverse family of intracellular and transmembrane enzymes, Science (80-) 253 (5018) (1991) 401–406.
[62] M. Gale, C.M. Blakely, A. Darveau, P.R. Romano, M.J. Korth, M.G. Katze, P52$^{r\ IPK}$ regulates the molecular cochaperone P58 ^{IPK} to mediate control of the RNA-dependent protein kinase in response to cytoplasmic stress †, Biochemistry 41 (39) (Oct. 2002) 11878–11887.
[63] L. Tskhovrebova, J. Trinick, Properties of titin immunoglobulin and fibronectin-3 domains, J. Biol. Chem. 279 (45) (Nov. 2004) 46351–46354.
[64] L. Bosgraaf, P.J.M. Van Haastert, Roc, a Ras/GTPase domain in complex proteins, Biochim. Biophys. Acta, Mol. Cell Res. 1643 (1–3) (Dec. 2003) 5–10.
[65] S. Bialik, A. Kimchi, The DAP-kinase interactome, Apoptosis 19 (2013) 316–328.
[66] L. Aravind, V.M. Dixit, E.V. Koonin, Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons, Science (80-) 291 (5507) (Feb. 2001) 1279–1284.
[67] O. Cohen, DAP-kinase is a Ca2+/calmodulin-dependent, cytoskeletal-associated protein kinase, with cell death-inducing functions that depend on its catalytic activity, EMBO J. 16 (5) (Mar. 1997) 998–1008.
[68] A. Romero, B. Novoa, A. Figueras, The complexity of apoptotic cell death in mollusks: an update, Fish and Shellfish Immunol. 46 (1) (Sep. 2015) 79–87.
[69] E.R. James, D.R. Green, Manipulation of apoptosis in the host-parasite interaction, Trends Parasitol. 20 (6) (Jun. 2004) 280–287.
[70] K. Terahara, K. Takahashi, Mechanisms and immunological roles of apoptosis in molluscs, Curr. Pharm. Des. 14 (2) (Jan. 2008) 131–137.
[71] I. Sokolova, Apoptosis in molluscan immune defense, Invertebr. Surviv. J. 6 (2009) 49–58.
[72] B. Morga, T. Renault, N. Faury, I. Arzul, New insights in flat oyster *Ostrea edulis* resistance against the parasite *Bonamia ostreae*, Fish Shellfish Immunol. 32 (6) (Jun. 2012) 958–968.
[73] M.A. Wouters, I. Rigoutsos, C.K. Chu, L.L. Feng, D.B. Sparrow, S.L. Dunwoodie, Evolution of distinct EGF domains with specific functions, Protein Sci. 14 (4) (Apr. 2005) 1091–1103.
[74] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (Oct. 1990) 403–410.
[75] C. Fathke, et al., Wnt signaling induces epithelial differentiation during cutaneous wound healing, BMC Cell Biol. 7 (Jan. 2006).
[76] Z. Zhang, et al., Secreted frizzled related protein 2 protects cells from apoptosis by blocking the effect of canonical Wnt3a, J. Mol. Cell. Cardiol. 46 (3) (Mar. 2009) 370–377.
[77] T.W. Holstein, The evolution of the wnt pathway, Cold Spring Harb. Perspect. Biol. 4 (7) (Jul. 2012) 1–17.
[78] F.J.T. Staal, T.C. Luis, M.M. Tiemessen, WNT signalling in the immune system: WNT is spreading its wings, Nat. Rev. Immunol. 8 (8) (Aug-2008) 581–593.
[79] M. Kahn, Can we safely target the WNT pathway? Nat. Rev. Drug Discov. 13 (7) (2014) 513–532 Nature Publishing Group.
[80] A.F. Pimenta, P. Levitt, Characterization of the genomic structure of the mouse limbic system-associated membrane protein (Lsamp) gene, Genomics 83 (5) (May 2004) 790–801.
[81] Y. Wang, D. Fu, P. Luo, X. He, Identification of the immune expressed sequence tags of pearl oyster (*Pinctada martensii*, dunker 1850) responding to *Vibrio alginolyticus* challenge by suppression subtractive hybridization, Comp. Biochem. Physiol. Part D Genomics Proteomics 7 (3) (Sep. 2012) 243–247.
[82] J. Towns, et al., XSEDE: accelerating scientific discovery, Comput. Sci. Eng. 16 (5) (Sep. 2014) 62–74.
[83] A.V. Zimin, et al., Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm, Genome Res. 27 (5) (May 2017) 787–792.
[84] A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0, (2015).
[85] T. UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (D1) (Jan. 2019) D506–D515.
[86] G. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, BMC Bioinformatics 6 (1) (Feb. 2005) 31.
[87] B.L. Cantarel, et al., MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Res. 18 (1) (Jan. 2008) 188–196.
[88] I. Korf, Gene finding in novel genomes, BMC Bioinformatics 5 (May 2004) 59.
[89] D. Marion, et al., Gmove a tool for Eukaryotic Gene Predictions using Various Evidence, (2017).
[90] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation, Nucleic Acids Res. 44 (D1) (Jan. 2016) D457–D462.
[91] P. Jones, et al., InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (9) (May 2014) 1236–1240.
[92] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 33 (Database issue) (Jan. 2005) D501–D504.
[93] A. Brelsford, C. Dufresnes, N. Perrin, High-density sex-specific linkage maps of a European tree frog (*Hyla arborea*) identify the sex chromosome without information on offspring sex, Heredity (Edinb). 116 (2) (Feb. 2016) 177–181.
[94] N.C. Rochette, A.G. Rivera-Colón, J.M. Catchen, Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics, Mol. Ecol. 28 (21) (Nov. 2019) 4737–4754.