Completion Report

Award number: NA11NMF4570217

Awardee: Maryland Department of Natural Resources

Award period: July 1, 2011 – June 30, 2014

Report period: July 1, 2011 – June 30, 2014

Report date: September 28, 2014

Project Title:

Development, validation, and application of molecular detection tools for a virus infecting *Mya arenaria* softshell clams in Chesapeake Bay, for evaluations of disease pathology, virus transmission, and disease distribution

Investigators:

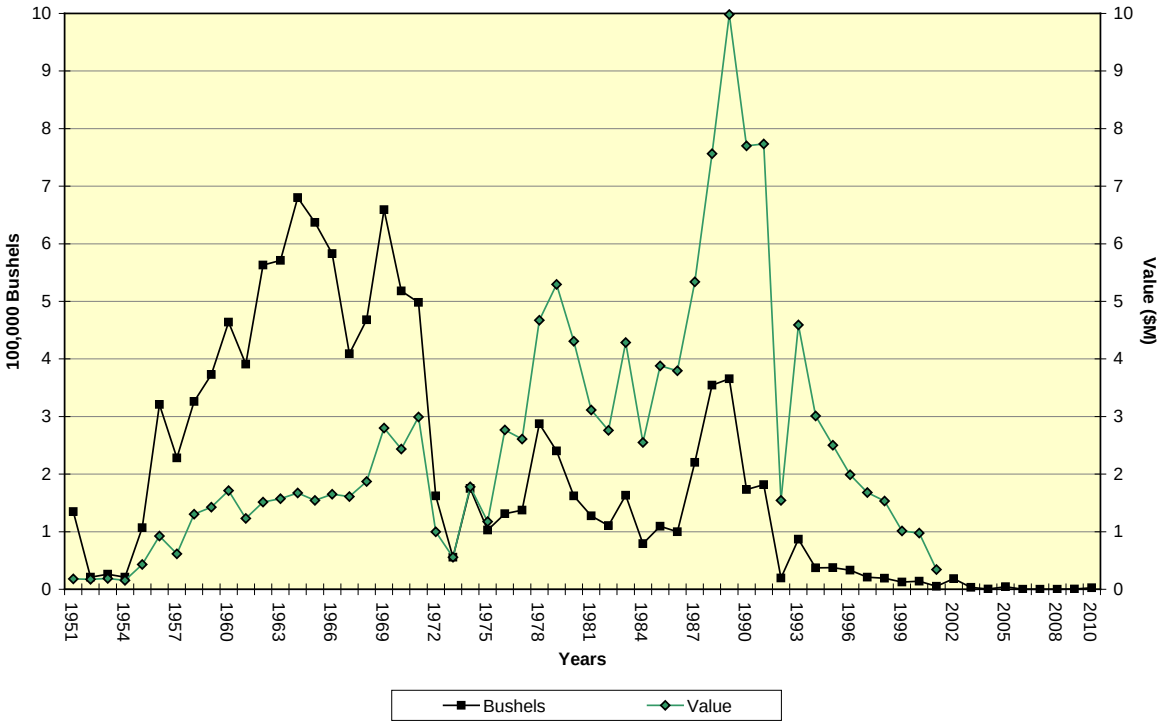
Christopher F. Dungan
Maryland Department of Natural Resources
Cooperative Oxford Laboratory
904 S. Morris St.
Oxford, MD 21645
chris.dungan@maryland.gov, 410-226-5193

Eric J. Schott, PhD
University of Maryland Center for Environmental Science
Institute of Marine and Environmental Technology
701 East Pratt St
Baltimore, MD 21202
schott@umces.edu, 410-234-8881

Introduction

Harvests of once-abundant and commercially valuable *Mya arenaria* softshell clams from Maryland Chesapeake Bay waters have radically declined from peak levels that commonly exceeded 500,000 bushels annually when the virgin fishery was fully capitalized during the 1960s and early 1970s (Fig. 1). The dramatic loss of formerly abundant *Mya arenaria* clams in Chesapeake Bay may reflect effects of overfishing, habitat degradation, and expanding disease impacts. Filter-feeding softshell clams render valuable ecological services to Chesapeake Bay water quality through consumption and assimilation of organic nitrogen and carbon compounds in their phytoplankton diet, as oysters are also famously credited. Unlike oysters that require relatively scarce hard benthic substrates in Chesapeake Bay, softshell clams occupy abundant soft benthic sediments as habitats from which to render their ecological services.

Fig. 1 Maryland Softshell Clam Landings and Ex-vessel Values



During nearly five recent decades of consistent and dramatic softshell clam resource decline in Chesapeake Bay, the following diseases have been documented to affect *Mya arenaria* clams at high prevalences:

1983 - Disseminated neoplasia (DN), a lethal, leukemia-like disorder of unknown etiology

1990 - *Perkinsus chesapeaki* infections by a protozoan pathogen similar to the lethal agent of oyster dermo disease

2000 - Pathogenic infections of gill epithelial cell nuclei (GENH) by an unknown virus.

Histopathological analyses during NOAA-funded surveys of diseases among Maryland Chesapeake Bay clams revealed prevalent virus-like lesions among softshell clam gill tissues, in which the nuclei of some gill epithelial cells were greatly enlarged and structurally abnormal (Fig. 2). The syndrome was provisionally termed gill epithelial cell nuclear hypertrophy (GENH) (Dungan et al. 2007, Dungan et al. 2009).

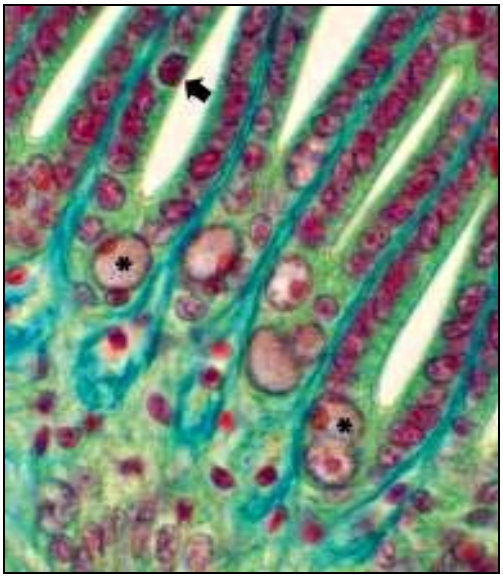


Fig. 2. *Mya arenaria* clam gill histological section stained to differentiate nucleic acid types (DNA = red) associated with viral inclusion bodies present in hypertrophied epithelial cell nuclei (*). Note the extreme enlargement (to 10 µm here) and abnormal staining of virus-infected nuclei, compared to the morphological characteristics of normal gill epithelial cell nuclei, one of which is undergoing mitotic division (arrow). Cell cytoplasm is green, and structural collagen fibers are blue.

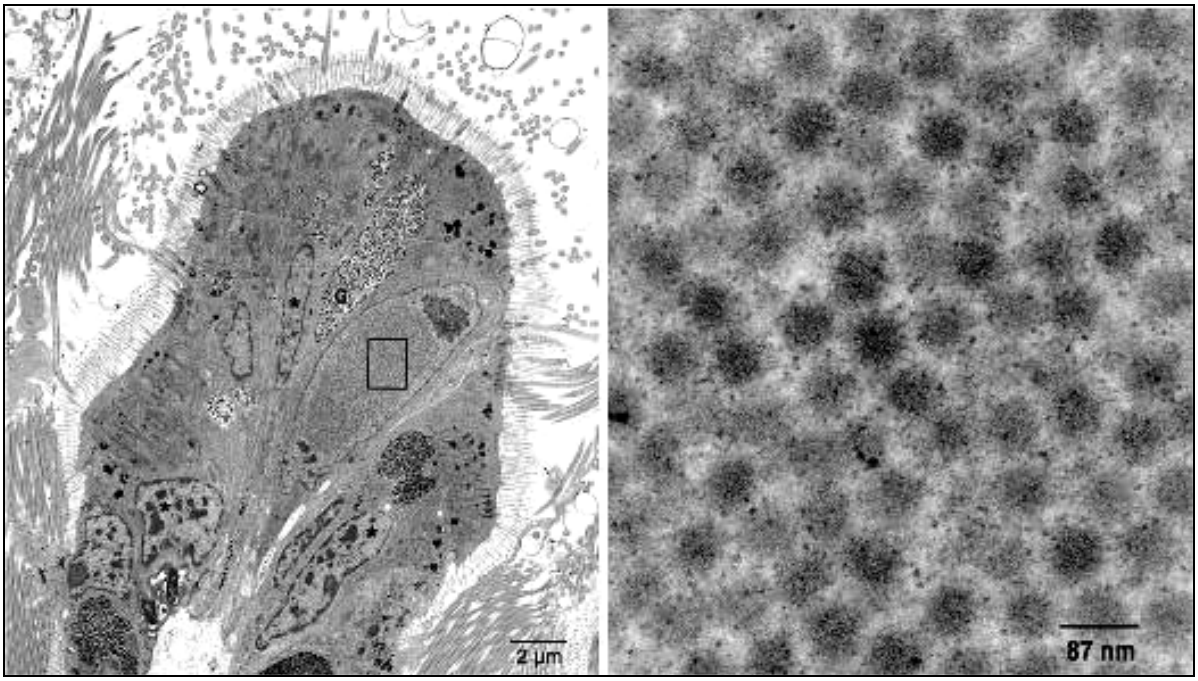
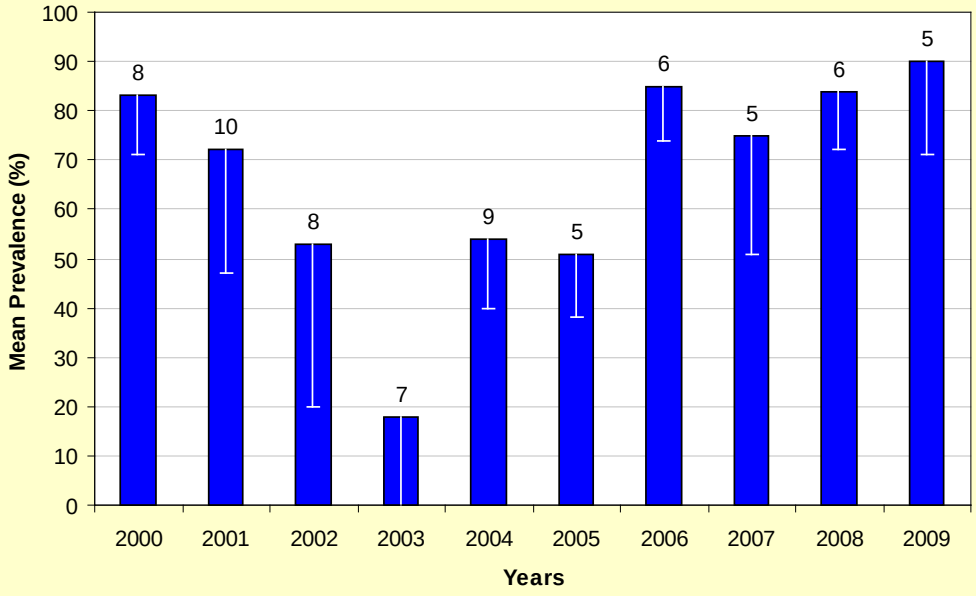


Fig. 3. Transmission electron microscopy of the distal tip of a *Mya arenaria* gill filament (left) in which at least three normal, uninfected epithelial cell nuclei (asterisks) occur adjacent to an enlarged and abnormal nucleus that has been commandeered for virus replication and assembly (inset rectangle). At high magnification (right), some of the abundant virus particles within an infected gill cell nucleus appear to bear radial surface projections.

During 2006-2013, clam samples were specifically processed to preserve high-quality tissue samples for transmission electron microscopy (TEM), in order to visualize potential viruses. Gill tissue samples from those and other clams were also preserved for subsequent virus and virus nucleic acid extractions. Extensive analyses by TEM of virus-infected gill epithelial cells confirmed and imaged abundant virus-like particles within the greatly enlarged nuclei of those cells. Virus particles are icosahedral, appear to bear surface projections, and have diameters of 60-87 nm, with 60-nm virions typical in paracrystalline virus arrays that occur in nuclei of infected gill epithelial cells (Fig. 3). The fact that the virus appears to be completely assembled within nuclei of infected clam cells suggests that the virus genome is composed of DNA rather than RNA.

GENH virus lesions occurred among gill epithelial cells of up to 100% of clams in samples from sites throughout Maryland and Virginia waters of Chesapeake Bay that have been evaluated histologically for the disease during 2000-2009 (Dungan et al. 2012) (Fig. 4). Therefore, this virus disease of the physiologically and nutritionally critical gill tissues of softshell clams is both prevalent and regionally widespread (Dungan et al. 2007, Homer et al. 2011). Current histological and electron microscopic methods for detection and monitoring of the virus and its effects are both labor-intensive and expensive. Development and validation of complementary, molecular-genetic methods to identify and detect the virus in clam tissues and environmental samples will empower research efforts of sufficient scope to accurately document the epidemiology and pathological effects at clam population levels of such virus infections among *Mya arenaria* softshell clams in Chesapeake Bay and elsewhere.

Fig. 4 Mean (- SD) annual GENH disease prevalences among *Mya arenaria* clams in Chesapeake Bay, 2000-2009. Numbers of 30-clam samples that were used to calculate annual means are shown.



Nothing is currently known about the genome of the GENH virus, including a critical lack of genomic nucleic acid sequence data that are essential for design and development of PCR- and DNA probe-based molecular-genetic assays for virus detection and quantification. Established methods are available for sequencing the complete genomes of viruses that may be purified from homogenized, infected tissues. We used those methods to fully sequence the putative genome of the GENH virus toward development, validation, and application of molecular-genetic assays for detecting the GENH virus in clam tissues and environmental samples.

Project Objectives

Project objectives include development and validation of quantitative molecular-genetic tools for detection and quantification of virus pathogen particles in *Mya arenaria* clam host tissues, tissues of potential vector animals, and in environmental water and sediment samples. Specific tasks include the following:

1. Sequence the entire genome of an unidentified virus that pathologically infects nuclei of *Mya arenaria* softshell clam gill epithelial cells at high prevalences in Chesapeake Bay.

2. Use the virus genome sequence information to identify the virus, and to develop and validate DNA probes and assays for detection and localization of virus lesions among *Mya arenaria* clams from Chesapeake Bay and elsewhere, whose histological tissue samples have been archived over 50 recent years.

3. Use the virus genome sequence information to develop and validate quantitative PCR assays for detection and enumeration of the agent-virus in clam tissues, and in environmental water and sediment samples.

Materials and Methods

Virus enrichment for PFGE. Viruses and clam cells from frozen or ethanol-preserved gill tissues of infected clams were embedded into agarose, and their nucleic acids were directly extracted before subsequent separation of small virus genome nucleic acids by pulsed field gel electrophoresis (PFGE). Nucleic acid bands of virus genome-sized dimensions were then excised from PFGE gels for DNA extractions and sequencing.

Virus nucleic acids were also enriched using other established methods (e.g, Walz et al. 1998, Ozbun et al. 2002), including methods in development for blue crab virus study in the Schott lab. (Bowers et al. 2010). Frozen or ethanol-preserved clam gill tissues (20-100 mg samples of pooled tissue) were homogenized with an Ultra-Turrax instrument and clarified of debris by centrifugation at 400 x g. Clarified preparations were fractionated on sucrose density step gradients, and fractions assessed for viral DNA by agarose gel electrophoresis. Prior to electrophoresis, selected subsamples were also treated with RNAase and DNAase to degrade nucleic acids external to virions. As needed, separate virus enrichments will be attempted by differential density centrifugation in cesium chloride (CsCl).

Clam gill tissues used for virus enrichments were from *M. arenaria* that were examined histologically to evaluate several pathological conditions, including the presence of pathologically hypertrophied nuclei among gill epithelial cells (GENH) caused by virus replication in those nuclei. Sample clams were categorized histologically as apparently free of hypertrophic nuclei (GENH), or to show GENH lesions at relative abundances that were ranked on a categorical infection intensity scale (1-5). DNAs extracted from frozen or ethanol-preserved gill tissues from clams with high-intensity infections (ranks 3-5) were compared to those from apparent uninfected glam gills (rank 0), to determine if virus enrichments from putative virus-free versus virus-infected clams showed that the virus-infected samples have distinguishing bands following PFGE, or within CsCl centrifugal cushions.

Nucleic acid preparation. Enriched virions were suspended in STE (NaCl, Tris, EDTA, pH 8) extracted with phenol–chloroform–isoamyl alcohol (25:24:1) and centrifuged. The aqueous phase were re-extracted with chloroform and precipitated with NaCl and isopropanol. The precipitate was washed in 75% ethanol, air dried, and re-suspended in TE buffer. Quantification of DNA was spectrophotometric (Nanodrop UV spectrophotometer at 260 nm), and assessment of DNA integrity and size will be by agarose gel electrophoresis. If the virus genome is characteristic of members of either the Polyomaviridae or Papillomaviridae families, we expect a dominant band of 5-8 kb. Additional isoforms of the circular DNA genome would be generated from single and double strand breaks, visible as secondary or tertiary bands on agarose gels. If the nucleic acids from enriched GENHV reflect a taxonomic affinity with the Adenoviridae family, a larger DNA mass of >20 kb will be evident, likely as a smear of >10 kb. The TEM image presented in Fig. 3B shows projections on the virions that are characteristic of the Adenoviridae, and not the other two candidate virus families considered for GENHV.

Virus enrichment for direct nucleic acid extraction. Tissue was first minced with a single edge blade, then homogenized in pH 7.4 Tris-buffered saline (TBS) with a Ultra-Turrax homogenizer. The typical scale of each preparation was less than 100 mg fresh tissue homogenized in 3 ml of buffer. Homogenates were cleared of large debris by centrifugation at 2,000 x g for 5 minutes; then layered over a step gradient of CsCl of 1.24 gm/ml and 1.44 gm/ml in an ultracentrifuge tube. Centrifugation was conducted at 92,000 x g, 2 h, 15 ˚C. The putative virus layers were removed from the interfaces atop the 1.24 gm/ml and 1.44 gm/ml CsCl steps. Debris is visible at the bottom of the tube was also collected. All fractions were diluted 1:10 in TBS; then concentrated by pelleting at 100,000 x g, 1 h, 15 ˚C.

DNA was extracted from the putative virus fractions by gentle SDS/proteinase K (to avoid shearing DNA), followed by phenol/chloroform clean up and ethanol precipitation. Once dissolved into nuclease-free water, quantification of DNA will be by spectrophotometry (Nanodrop). Putative virus DNA was analyzed by agarose gel electrophoresis of both untreated DNA (200 ng/lane) and restriction-digested DNA (XhoI, EcoRI separately). Undigested, virus-derived DNA was expected to create a band or tight smear of greater than 20 kb on standard 1% agarose gels. When digested with restriction enzymes, high molecular weight virus DNA should create distinct bands that sum up to ~50 kb, whereas if the preponderance of DNA is clam genomic DNA it will create a smear.

Virus genome sequencing. Full GENH virus genome sequence was sought by state-of-the-art pyrosequencing methods, first using the Roche 454 Genome Sequencer FLX instrument with titanium chemistry, and later using a high-throughput Illumina MiSeq sequencer. Purified viral template DNA (after whole genome amplification, see results) was prepared by shearing to the proper size, and then ligated to adaptors. An overview of the pyrosequencing process is available online (http://www.454.com/products-solutions/how-it-works/index.asp).

Each Roche 454 sequencing plate has the capacity to produce 1.25 million sequencing reads of 400 bases, or approximately 500 million bases of total sequence. It is possible to divide the plate into halves or quadrants; for the GENH virus genome determination, ½ plate was used. With the rule of thumb that 10 fold sequence coverage is needed to fully assemble genomes, the potential 250 million bases of sequence is far more than adequate to determine the largest contemplated possible GENH genome of 35,000 bp. Genomic sequences were subjected to contig assembly using the computers and algorithms in use at the Genomics Resource Center of the Institute for Genome Sciences at the University of Maryland School of Medicine, and later at the IMET BioAnalytical Services (BAS) lab. These tools allow sequencing facilities to deliver contigs, scaffolds, and metrics in a variety of standard formats. It is also possible to simultaneously sequence independent libraries (DNA samples) on one 545 or MiSeq by encoding each library with unique end-label sequences that are deconvoluted by the sequence assembler, resulting in independent sequence assemblies. This multiplexing feature was employed using at least two independent preparations of GENH virus DNA, originating from clams collected in different years or locations. Each virus genome will be independently assembled (computationally). Only limited sequence variability is expected from the GENH virus' dsDNA genome. However, the library and sequencing redundancy should provide robust certainty about the genome sequence.

Gross genome structure. The microscopic investigations of GENHv, presented above, leave open the possibility that the virus is in one of three major groups of dsDNA viruses. If the virus is polyomavirus-like, its genome will assemble into a circular dsDNA molecule of ~5 kb encoding 4-7 major polypeptides (Figure 5A). If it is from a member of the Papillomaviridae, a ~8 kb circular genome would be anticipated, with an organization roughly like that in Figure 5B, which includes early genes that encode non-structural proteins, and late proteins that comprise the capsid (L1 nd L2). Also notable is the fact that both Papillomaviridae and Polyomaviridae are associated with epithelial infections, a characteristic of the GENH virus. The third possible virus type, Adenoviridae (Figure 5C), would be expected to produce a linear dsDNA genome assembly of >30 kb, encoding at least a dozen proteins. Of special interest in the Adenoviridae is the possibility of a gene encoding a predicted DNA polymerase that's characteristic of that virus family. On the other hand, papilloma-like viruses have been reported in many in bivalves species (Garcia et al. 2006), and are associated with nuclear hypertrophy (Elston 1997).

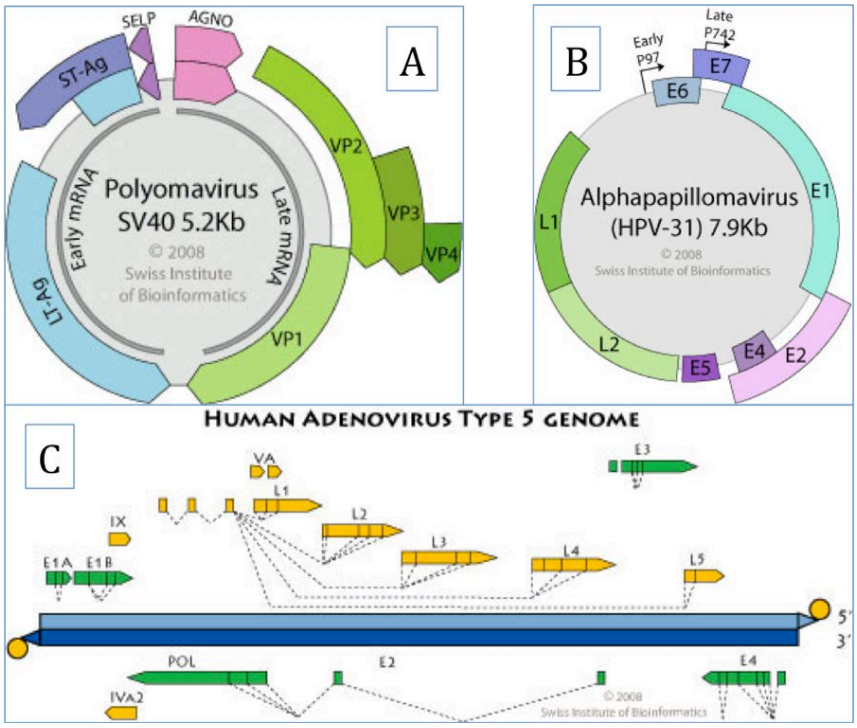


Fig. 5. Genome organizations of representative dsDNA viruses.
A) Polyomavirus, B) Papillomavirus, C) Adenovirus. Images © Expasy, Viralzone

Bioinformatics. Obtaining a full virus genome is the only way to be assured of having the best possible understanding of the taxonomic relationship of GENHv to other viruses. The need for a complete genome (rather than selected genes) for phylogenetic comparisons derives from the fact that, unlike in all cellular organisms, there are no universally-shared genes among DNA viruses (such as the ribosomal RNA genes among cellular organisms). Virus genomes also display a very high degree of sequence variation between 'species', even among 'conserved' genes. Although there are now thousands of viruses sequenced, nucleotide databases are still relatively depauperate of virus genomes, and most sequenced genomes are from viruses that infect vertebrate hosts.

Phylogenetic relationships based on alignment of protein encoding regions. The sequences of all predicted proteins encoded in the completed GENHv sequence were individually analyzed using BLAST tools (blast.ncbi.nlm.nih.gov/Blast.cgi) to identify the most similar predicted proteins in GenBank. Two searches were used: one against the non-redundant GenBank database that includes all nucleic acids from any source (cellular, viral, and environmental), as well as a targeted BLAST search against all known virus genomes. The non-redundant search is likely to uncover many more matches with a low similarity. Homologs of other viruses can also be specifically searched for at low stringency by limiting a BLAST search to one or another family of virus (e.g., only Polyoma-, Papillo-, Adeno-).

Virus sequences/genes with the most similarity to the GENHv (identified by BLAST analysis) were aligned to cognate sequences of GENHv using CLC Workbench and CLC Genomics software (http://www.clcbio.com/products/clc-genomics-server). If relationships are detcted, preliminary phylograms can be drawn using the graphical output options (neighbor-joining, phylip, or distance (Chenna et al. 2003)). For example, if the sequence is that of an adenovirus, it may include a DNA polymerase gene. Alignment of adenoviral polymerases has been used to define relationships within that group (Harrach and Benko 2007).

PCR assay development and validation. Based on the BLAST analyses and locations of predicted protein-coding genes, two categories of primers were produced to permit amplification of GENHV from field or research material.

1) **Diagnostic** - these primers will be designed to amplify short (<200 bp) stretches of GENHV, and target regions of the GENHV genome that do not display polymorphisms in the sequence assembly. These primers are checked to ensure that they are not similar to known *M. arenaria* genomic sequences in GenBank or sequence produced in this study.

2) **Investigative** - designed to span large (>1 kb) stretches of the virus genome. These can be used to query GENHV diversity in DNA extracted from infections obtained across time and space.

Diagnostic primers above can be applied in a quantitative real time PCR (qPCR) format (ABI FAST7500), as we have done with PCR assays for other pathogens (Nagle et al. 2009, Schott et al. 2010 World Aquaculture Society conference presentation). Primer pairs that generate unique PCR amplicons from virus genomic DNA were designed, and their reaction parameters validated using preserved gill tissues whose infection status is inferred by histological results from adjacent gill tissues. Clinical validation of the developed PCR assay can later be accomplished using fresh gill tissues from new clam samples.

Because there are no cell culture systems available for titering viruses of molluscs, this ability to approximate virus genome copy numbers using qPCR will provide the best measure of virus load in clam tissues or environmental samples. Anticipated applications for real time quantitative PCR enumeration of GENHV DNA includes surveys for virus in apparently healthy clams, quantification of virus in different tissues, investigation of whether GENHV can be detected in water, experimental transmission studies, and identification of specific pathogen free clams.

In situ hybridization assay development and validation. Based on genome sequence bioinformatic analyses and PCR-based evaluations for specificity to DNA extracted from GENHV-infected clams, labeled DNA oligonucleotide probes that hybridize to unique GENH virus genome sequences will be designed. Labeled oligonucleotides will be validated by in situ hybridization (ISH) assays on histological sections from clams whose gill epithelia have been confirmed as infected or uninfected by existing TEM or histological results (Reece & Dungan 2006, Reece et al. 2008). Briefly, we designed (3) digoxigenin-labeled oligonucleotide probes of 24 bp that will specifically hybridize with unique segments of the putative GENH virus genome,

or to genetic targets in nuclei and cytoplasmic (mitochondrial) compartments of both GENHV-infected and uninfected clam cells. These labeled probes will be hybridized to sections of target tissues according to established methods for in situ hybridization (Arzul et al. 2002), and detection of the oligonucleotide will be achieved by an alkaline phosphatase-conjugated mouse IgG antibody against digoxigenin and a BCIP/nitroblue tetrazolium chromogen. Appropriate negative controls will include hybridizations of the specific oligo lacking digoxigenin, and hybridization with a dig-labeled probe against irrelevant targets.

Additional fluorescent probes will be designed for direct fluorescent detection of GENH virus DNAs by FISH assays. This method has been used recently to visualize a DNA virus in spiny lobster (Li et al. 2007). Once validated, both types of ISH probes can be used for retrospective analyses of archived histological specimens from historic clam samples, in order to assess the geographic and temporal distributions of GENH virus infections among *Mya arenaria* clams. ISH studies may also elucidate the mechanism of virus infection in experimental studies, addressing for example whether the lesions seen in gills arise from virus that is filtered from the water, as suspected, or whether the virus enters the clam through an alimentary route.

Results

Sample analyses and clam tissue selections for virus enrichment or isolation. Histological slides from 529 clams whose separate gill tissue samples had been previously preserved for extraction of virus particles or nucleic acids by freezing or by ethanol dehydration, were analyzed to identify archived gill tissues were optimal for virus isolation or extraction procedures. Among the analyzed clams, 36 were identified to harbor heavy GENH virus loads (lesions) in their gill tissues without other complicating infections, and 6 were identified as useful negative control-tissues from clams without GENH virus lesions or other infections. From the inventory of promising preserved clam gill tissues, preserved gill tissues from 18 clams with heavy infections were selected and transferred for virus enrichment, isolation, and nucleic acid extraction procedures at UMCES-IMET, as were negative-control gill tissues from 6 apparently uninfected clams.

Additional clam samples were processed to secure and identify DNA and histological samples for future diagnostic PCR and ISH assays. From two Chesapeake Bay sites, 73 clams were collected, processed to preserve gill tissue DNA samples, and processed to produce histological samples that were analyzed and scored for abundances of gill GENH lesions. A sample of Oregon clams was secured, processed for DNA and histological samples, and the latter analyzed to reveal a 53% prevalence of GENH-like lesions. Archived histological samples from 60 clams of a 1972 Massachusetts sample from which GENH-like were reported, were accessed, sectioned, and analyzed to reveal a 24% prevalence of GENH-like lesions.

Virus enrichment, isolation and nucleic acid extractions. Histological analyses identified a number of GENH virus-infected *M. arenaria* gills that have been preserved and archived in ethanol during recent years; and which each contain approximately 10 mg of fresh gill tissue wet weight. Although it might be difficult to recover enough intact virus particles from 10 mg of infected gill tissue to produce enough virus nucleic acid for purification and sequencing, efficient recovery of intact DNA from ethanol-preserved samples is routinely successful. Therefore, we tested a strategy of separating viral genetic material from host genetic material without first attempting to isolate and purify virus particles.

Electron microscopic analyses indicated that the GENH virus has a diameter of 60-85 nm, and also suggested that the virions may also bear radial surface spikes. Based on analysis by electron microscopy, the size and morphology of GENH virus indicated it is most likely in one of three classes of virus: adeno-, polyoma-, or papillomavirus. The genomes of members of each of these virus families are double-stranded DNAs of ~35 kb (linear) or 5-8 kb (circular).

An effective way to separate and visualize DNA molecules of a wide range of sizes (several kb to hundreds of kb) is pulsed field gel electrophoresis (PFGE). The method relies on the *in situ* liberation of DNA from a suspension of cells (and/or viruses) within an agarose block. This maintains the integrity of large fragments of DNA or even chromosomes, which would normally shear during manipulations in solution. We used PFGE to investigate the DNA of archived specimens of *M. arenaria* gill tissues containing GENH virus lesions, based on analyses of histological sections from the same clams.

We first confirmed that PFGE serves to separate episomal DNA from bivalve genomic DNA, by comparing 1) *Mytilus edulis* gill tissue that had been archived in 75% ethanol, 2) bacteria carrying a 5 kb plasmid, 3) a mixture of *M. edulis* gill and bacteria/plasmid. The separation of the 5 kb plasmid from genomic DNA was clear (data not shown). The resulting suspension was embedded in agarose, subjected to SDS and proteinase K digestion, and then incorporated into a PFGE gel. Agarose blocks were prepared with three different densities of tissue (estimated at 6, 3, and 1.5 mg equivalents).

In Figure 6 can be seen the results of conducting PFGE separations of DNA from archived *M. arenaria* gill fragments that were scored as either GENH positive or GENH negative by histological analysis. In both lanes from the *M. arenaria* sample with GENH-virus lesions, a clear band of about 50 kb can be seen, and a more faint band of about 20 kb is also evident. Adenoviruses are reported to be on the order of 35-42 kb, but virtually all adenovirus studies are from those infecting mammalian hosts (www.virology.net/big virology/BVDNAadeno.html). Fowl adenoviruses D is in fact 45 kb in length (Ojkic & Nagy, 2000). The gel image was also digitally manipulated (not shown) to enhance fluorescence in the 5-10 kb range, where possible papilloma or polyoma virus genomes would appear, but no trace of bands was observed between 4.4 and 20 kb.

Total DNA extracts of *M. arenaria* may display mitochondrial genomic DNA, and this may give rise to an episome on PFGE gels. In virtually all molluscs (including heterodonts like *Mya arenaria*, Dryer & Steiner, 2006), the mitochondrial genome is approximately 16 kb. We did not observe any prominent banding at 16 kb, even when that region of the gel was enhanced. It is possible that the mitochondrial DNA is not abundant enough to be visualized by this method

from the few milligrams of tissue analyzed. We may investigate the possibility that mitochondrial DNA is present in different regions of the gel by conducting PCR amplifications with conserved primers for the COX1 gene. These primers have been used in *Mya arenaria* previously (Strasser & Barber, 2009), and have a long history as a tool in 'DNA barcoding' diverse organisms (Folmer et al. 1994).

Amplification of PFGE-isolated virus DNAs by whole genome amplification (WGA).
The PFGE separation of putative GENH virus DNA from host *M. arenaria* DNA (as in Fig. 6) was repeated three additional times. Each PFGE separation was followed by extraction of DNA from the section of the gel showing the 50 kb and 20 kb bands. In total, approximately 100 nanograms of DNA were recovered, of which we estimate 10-20% is genome of the putative virus. A subsample of this DNA was then subjected to whole genome amplification (WGA), using a relatively recent technique that employs Phi polymerase (Genomiphi, GE Healthcare). The yield of amplified DNA was treated with S1 nuclease to resolve branched DNA structures that are produced during the WGA process. In all, a yield of nearly 1 microgram of DNA was achieved.

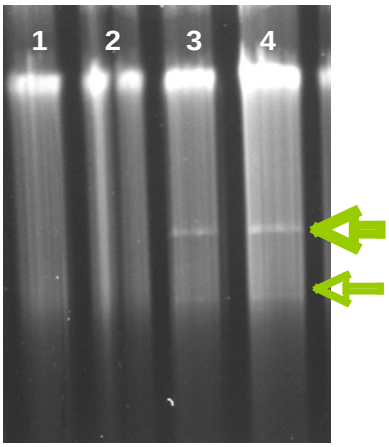


Fig 6. Preparative PFGE gel to enrich GENH virus DNA. Large and small arrows point to 50-kb and 20-kb bands that were excised for sequencing. Lanes 1, 2 are uninfected *M. arenaria*, lanes 3, 4 from clams with high histology scores for GENH disease.

Illumina sequencing of virus DNAs.
The amplified DNA was re-purified by ethanol precipitation, and taken to the Genomics Resource Center (GRC) at the Institute for Genome Sciences at the Univ. of Maryland School of Medicine. Staff at the GRC conducts three steps to produce the millions of sequences that result from a single Illumina run. First, input DNA is sheared and ligated to end-adapters. These molecules are then size-selected for optimum sequencing in downstream steps. Second, individual DNA molecules are hybridized to a solid matrix where millions of them are amplified *in situ*. Clusters of identical molecules then constitute the template for the third step, in which a sequencing primer is extended with nucleotide precursors that generate base-specific flashes of fluorescence. Optical readers in the Illumina sequencing machine read the fluorescent flashes, and transfer information to a computer for storage and later analysis.

The DNA submitted for sequencing successfully generated 30 million different sequencing reads, for a total of 30 billion bases of DNA sequence. According to GRC staff, the quality of the sequence passed their QA/QC, and qualified for further bioinformatic analyses.

Bioinformatic analyses: GenBank screening and strategy for creating alignments of contiguous virus genome segments (contigs).

The preliminary analysis by GRC staff consisted of a nucleotide-based search of GenBank using 250,000 of the sequencing runs, which returned no significant virus-related matches. However, in our estimation, this level of analysis would not be expected to return any strong hits. Virus sequences are very diverse, and there are relatively few invertebrate virus sequences in GenBank.

As an illustration of the great difference between a nucleotide–based and protein-based search, if the *pol* gene of the plum headed parakeet adenovirus is used to search GenBank using a standard nucleotide algorithm, no significant matches are returned. If, however, the same nucleotide sequence is translated and compared to the predicted protein products of GenBank entries, a list of over 100 adenovirus genes from diverse tetrapod vertebrates is returned. An alignment of some of those matches is shown in Fig. 7.

```
parakeet SALTHPLPFGIPVGKTEKEQEIEKFQRLLDTKEKISYF-ENIKPMIININAFPPPKEKLDCLPPLCSRKSGKLCWTNEPLKDEVITSID
snake    SALSHPMPYGPTLSPFDSAVAIAEFQRKLDGQSELSYF-DDIFPMIVVADAFPPSLHCLDVLPPLCSKRSGKLCWTNEPLLGEVLTTVD
raptor   SALTHPMPYGIPVGEKERLEEIKKFTNLLSRRDKISYF-QGIKPMIVTVNAFPPPTELLDPLPPLCSKKSGKLCWTNEPLNNEVVTSID
human    SALTHPMPWGTPLNPYERALAVREWQMTLDDPATISYFDKDLLPGIFTIDADPPDEFMLDPLPPFCSRKGGRLCWTNEPLRGEVATSVD
```

Fig. 7. An alignment of a part of the protein encoded by the adenovirus *pol* gene from tetrapod vertebrates. This is an illustration of a candidate virus gene product likely to be conserved between adenoviruses of vertebrates and invertebrates. Bioinformatics technicians at GRC are searching for homologs to this gene in the Illumina sequence data.

Dr. Schott met with GRC staff to discuss the strategy for sequence assembly and discovery of virus sequences within a preponderance of clam genomic DNA sequence (recall, only 10-20% of sequencing reads are expected to be virus). It was important to emphasize to GRC bioinformaticians that the goal of the project was to detect potentially highly represented sequences that would have only very weak similarity to any sequence in GenBank, and to therefore conduct protein-based (not nucleotide-based) searches.

Based on information that indicates that the GENH virus may be an adenovirus (morphology in TEM and apparent genome size), we provided the GRC staff an alignment of DNA polymerase gene (*pol*) sequences from vertebrate adenoviruses (there are no invertebrate adenoviruses reported in the literature). The protein encoded by the *pol* gene is the only adenovirus product with a level of conservation that may be extended to invertebrate species of adenovirus.

Many sequence assemblers discard highly represented sequences because, in large genomes, such data may arise from microsatellite sequences or transposon sequences dispersed throughout the genome; which may interfere with reliable contig assembly. We have therefore asked GRC staff to suppress automated computational steps that remove repetitive sequences, and to conduct an assembly of a subset of that unfiltered sequence.

Absence of promising sequences in Illumina-derived data, January 2013.
After another set of meetings with the scientists at the GRC, it was clear that there are no promising sequences in the data generated by Illumina sequencing of whole genome amplified DNA isolated from the gel in Figure 6. CRG bioinformaticians used the polymerase (*pol*) gene alignments as well as alignments of other genes conserved in adenoviruses to search for weak homology in the 30 billion bases of sequence. Technicians also assembled a list of BLAST hits that had homology to virus genomes, and still found no promising sequences. Instead, the BLAST hits were dominated by sequences representative of bacteria genomes. A small number of sequences aligned to a short stretch of infectious salmon anemia virus (ISAV, an RNA genome), but were determined to be insignificant because they were isolated reads and did not belong to a contig.

GCR staff let us know that the absence of useful sequence is indeed one of the risks of using a whole genome amplification method, because it can exaggerate contamination by DNA from exogenous organisms, and because the Genomiphi polymerase creates sequence artifacts due to the nature of the branching DNA replication. Possible sources of DNA contamination for our samples were many, starting with the DNA extracted from the gills themselves. Although clean when they were preserved, clam gills are not an axenic system. They will have scores of endogenous microbial flora present in and on them. Experimentally, there are many additional sources of bacterial contamination. The one most prominent in our minds is the electrophoresis system on which PFGE gels were run. While measures were taken to clean all of the equipment before using it, surfaces of gel boxes and dust incorporated into buffers and agarose could have introduced bacteria from the two labs used for this work. One of the labs was, alas, a marine microbiology lab.

After the failure to obtain quality sequence for the presumed GENH virus using the WGA approach, we embarked on two tracks to obtain GENH virus DNA in quantities and quality sufficient for direct sequencing or cloning. One approach is to re-run multiple PFGE gels to obtain more target DNA than in the previous attempt. The other approach is to use a simplified virus particle isolation method.

Enhanced PFGE yields of virus DNA, January - June 2013.
By conducting a larger scale extraction and repeating the procedure that produced the results shown in Figure 7, we anticipated being to be able to recover 10x more DNA for direct sequencing on the Illumina system. Unfortunately, after a single attempt to use approximately 3 times more clam material for each lane of the gel, we found that it was not possible to discern the putative GENH virus bands from what appears to be an overloaded PFGE gel. Thus, to recover the amount of material needed for a direct cloning or sequencing, at least 10 PFGE gels would need to be run. This is a straightforward task, and could be accomplished in about 3 weeks. However, the downstream steps, of excising the putative band, purifying the DNA from the agarose, and removing the traces of agarose from the recovered DNA, pose a challenge to the integrity of the DNA itself. Our experience with these manipulations to date is that each step reduces the quantity of the DNA, and we were never able to remove all of the agarose from the DNA excised from gels. A method that avoids handling of the DNA, and can concentrate it in a buffer system that does not include agarose, is preferable.

Centrifugation and filtration for GENH virus enrichment, January - June 2014.
As an alternative to PFGE, we have tested a method to recover small amounts of virus particles from frozen and ethanol preserved tissues. The method is relatively simple, and consists of a filtration step followed by digestion of host RNA and DNA prior to lysis of virus particles. The method's success relies on the integrity of the virus particles to protect the virus genetic material from enzymes outside the capsid (Ng et al. 2009).

Figure 8 shows a schematic of the filter/nuclease enrichment process, as well as a demonstration that the process is effective for isolation of the genomic RNA of the crab virus, RLV. In the example shown in Figure 9B, an RNA genome was visualized; obviously for GENH virus, the process is expected to reveal a DNA genome. The RNA in Fig. 9B was derived from 50 milligrams of frozen crab tissue, which is similar to the quantity of putative GENH virus-infected clam gill tissue available from individual clams already determined by histology to be GENH-positive.

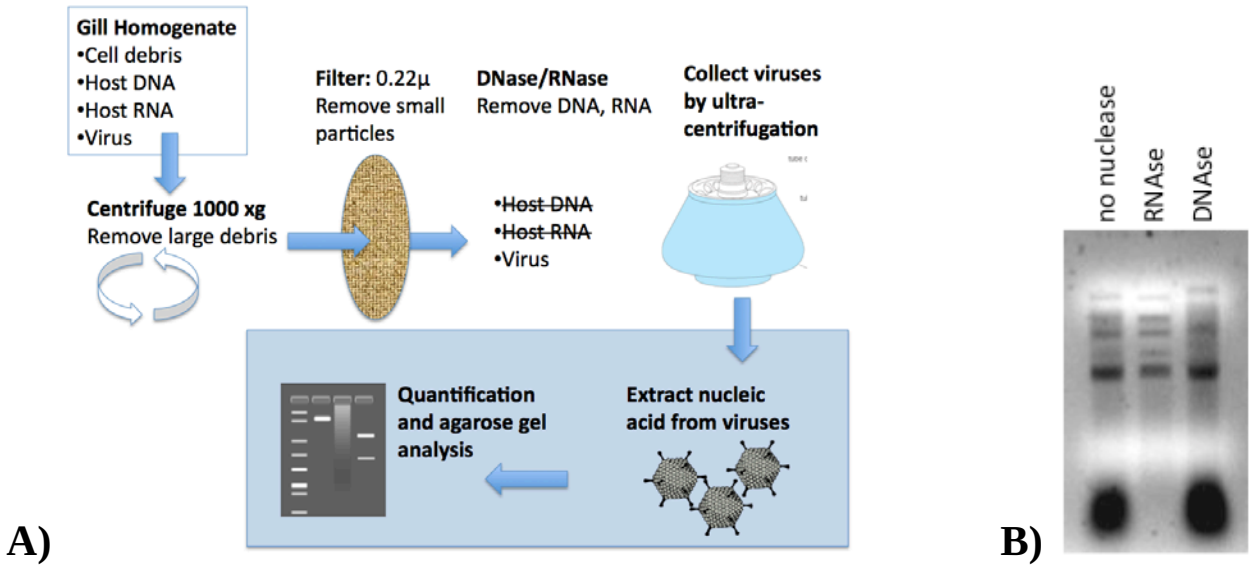


Fig. 8. Enrichment of virus particles using filtration and nuclease treatment.
A) Schematic of the process.
B) Demonstrating the effectiveness of the methodology with the blue crab virus, RLV. The blue crab virus has an RNA genome; therefore, RNA was extracted in the final step of the scheme shown in (A). Note that the RNAse treated sample lacks the low molecular weight RNA that is present in control (no nuclease) and DNAse treatments, yet the virus genomic RNA is still intact.

The demonstration of this method using frozen archived material was successful. We also will conduct this trial run with ethanol-preserved material (crab with RLV). If useful for ethanol-preserved samples, this will expand our pool of material available for virus isolations, as there are numerous ethanol-preserved (GENH) clam gill samples at DNR-Oxford. It is expected that if the GENH virus is indeed an adeno-like virus, it will be stable to ethanol, since adenoviruses are both stable in the environment and resistant to disinfection by ethanol (University of Kentucky Department of Biological Safety).

The filter/nuclease method was applied to ~50 mg of GENH-positive (frozen) archived clam gill (100108 CRBSMa-11). As a control, the same method was applied to archived crab tissue known to be carrying the RLV virus. Homogenization was accomplished by bead-beating (0.5 mm and 3 mm glass beads), and the remaining procedure followed the scheme depicted in Figure 9 with the following modifications. The putative GENH filtrate was split into two aliquots, and one was treated with DNAse I, and the other was incubated without DNAse. Samples were then treated with EDTA to arrest nuclease activity, and centrifuged at 65,000 x g for 3 hours to produce glassy pellets of putative virus-enriched material. DNA was extracted from the pellets with by the proteinase K-SDS methods and extracted with phenol/chloroform, and analyzed by gel electrophoresis. Faint bands of 25-50 kb were evident in the GENH sample, and faint bands of RLV genome in the positive control. The GENHV DNA was digested with restriction enzymes SpeI and EcoRI to attempt to fragment the genome into discrete and clone-able fragments.

Unfortunately, there was no evidence of digestion of DNA after 1 hour of digestion. A second digest was attempted with two enzymes that target G/C-rich sites: BamHI and XhoI. These digests resulted in the disappearance of the large DNA band, but there were no obvious smaller bands appearing in the digests. The lack of smaller bands is a likely result of the limit of sensitivity of the ethidium bromide (EtBr) stain, and the very low amount of DNA used in the digests (i.e., one >25 kb band was visible, but the same mass of DNA, when divided into 3 or more fragments are individually undetectable by EtBr staining). These results suggest that this approach will generate GENH DNA for eventual sequencing. Clearly there is a need for further optimization of the method. The amount of virus DNA recovered was too little for cloning or direct sequencing by the Nextera process (see below). Based on the RLV-only test of the method (Fig. 8), nucleic acid recovery from the nuclease treated GENH material was lower than expected. It is also possible that virus DNA was less abundant than anticipated, in the clam gills used in this trial.

Additional planned iterations of the process were replaced with a modified process that did not involve the nuclease and filtration steps, and instead relied solely on differential density centrifugation to enrich putative virus DNA. Those efforts are reported below for cesium chloride gradient enrichment and high-throughput sequencing of putative virus DNA.

Following the repeat of the filter-nuclease treatment, recovered DNA was analyzed by gel electrophoresis and quantified by UV spectrophotometry. Of the several options for converting this DNA into sequencing templates, the best and most elegant is transposon-mediated tagging and fragmentation. Outlined in Figure 10, this method has been formulated as a kit that requires only 50 ng DNA or less. It was anticipated that the outcome of the filter/nuclease enrichment will provide at least 50 ng of template, and that we will be able to use the Nextera method to produce a sequencing template for Illumina technology.

The virus enrichment work described above was conducted with a table top ultracentrifuge (Beckman TL-100), with a sample capacity of just 3 ml per tube. While this configuration was suitable for pelleting virus particles following the filtration/nuclease treatments depicted in Figure 8, the small tubes are impractical for producing gradients of sucrose or cesium chloride. It would be preferable to conduct the virus preparations on a preparative scale instrument. In

November, IMET acquired a new ultracentrifuge which facilitates the enrichment or purification of virus particles from GENHV-positive samples. The new centrifuge is a WX Ultra80 (Thermo Scientific), with fixed-angle (Fiberlite™ F50L-8 x 39), and swinging bucket (TH641) rotors. Prior to the arrival of this instrument, we had been hesitant to conduct the ultracentrifugation steps with the preparative scale 1980s-era centrifuges available at IMET. A recent independent inspection of all of the department's ultracentrifuges did indeed confirm that all of the rotors were past due for replacement, and should be removed from service. The larger tube capacity and swinging bucket feature of the TH641 allowed us to produce sucrose and cesium gradients to enrich viruses from both ethanol-preserved and frozen biopsies of GENHV-positive gills. In addition to the many archived samples in -80 ˚C and ethanol storage, additional GENHV-positive gill biopsies were obtained in 2012 and 2014, following the start of this project. Therefore, ample material was available for preparative scale virus enrichments following methods established for preparative scale adenovirus production (Scott & Chamberlain 2003).

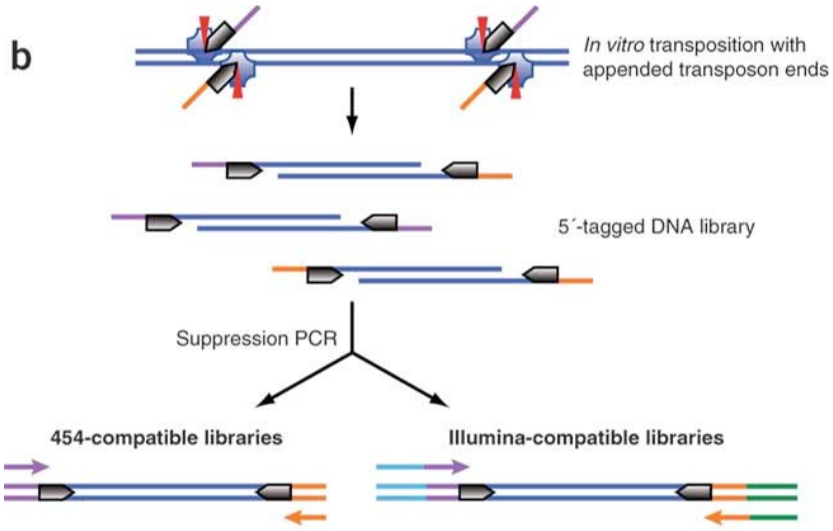


Fig. 9. Nextera "tagmentation" approach to creating sequencing templates. The products can be sequenced by either Illumina or Roche 454 technology.

Cesium chloride gradient enrichment and high throughput sequencing of putative virus DNA, **January - June 2014**.

The WX Ultra80 ultracentrifuge and swinging bucket (TH641) rotor were used to fractionate homogenized clam gill tissue following an adaptation of published protocols for adenovirus purification (Scott & Chamberlain 2003). An initial fractionation and validation of the method was first conducted with homogenates of blue crab muscle infected with RLV, which, as a reovirus, is predicted to have a density similar to that of adenoviruses. The results confirmed that the CsCl method is effective in producing a fraction that is highly enriched for the RLV genome (Fig. 10). The ultracentrifugation method was then used to enrich putative virus from preserved clam gills that had been scored as highly infected with GENH.

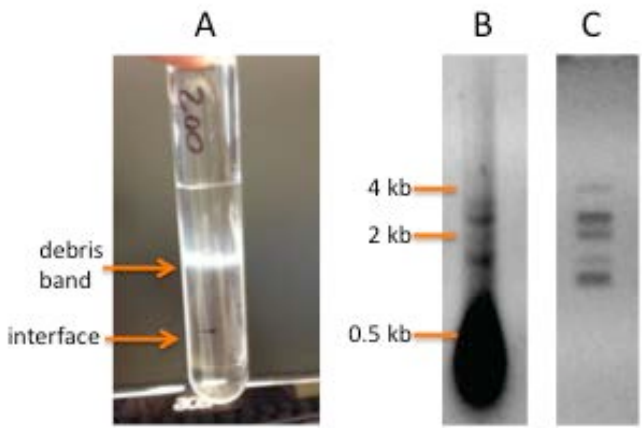


Fig. 10. Validation of CsCl gradient virus enrichment method using RLV-infected crab muscle. **A.** CsCl gradient after centrifugation, showing locations of interfaces harvested; **B.** RNA from debris at the top of the 1.24 g/ml layer; **C.** RNA from interface between 1.24 and 1.40 g/ml.

Clam gill tissues (14 to 70 mg) were homogenized in isotonic buffer using a rotor-stator instrument (Ultra-Turrax). Homogenates were clarified by centrifugation at 500 x g and the supernatant was layered onto a CsCl gradient of 1.24 and 1.40 g/ml density. Ultracentrifugation at 45,000 x g for 2 hours was conducted at 15 ˚C. Following centrifugation, three fractions were collected from each sample: the debris at the top of the 1.24 g/ml layer, the interface between 1.24 and 1.4 g/ml, and the bottom of the tube. DNA was extracted from the fractions by an SDS/proteinase K procedure, followed by chloroform/isoamyl alcohol extraction and precipitation with Na-OAc and ethanol. DNA was suspended in nuclease free water and assessed by spectrophotometry (Nanodrop) and gel electrophoresis. Both analyses confirmed that the DNA from the interface between 1.24 and 1.40 g/ml had the highest quantity and quality of DNA.

Based on gel and Nanodrop assessments, DNAs from two fractionated samples (080930 CHBRMa-4 and 120413 EBLPMa-24) were then subjected to the 'tagmentation' process (Fig. 9) to prepare templates for high throughput sequencing on IMET's MiSeq instrument. An additional DNA extract was prepared from a homogenized, but not CsCl-fractionated, subsample of clam 080930 CHBRMa-4. The sequencing of all three DNA extracts was successful, returning 1.4 - 4.2 million sequencing reads per sample, with read lengths of over 100 bases, for a minimum of 140 million bases of data per sample.

The rationale for using MiSeq technology to find the GENHV genome.
The sequencing capacity of the MiSeq system is not high enough to produce a full genome sequence of a metazoan, such as the clam, *M. arenaria*. Based on the genome size of other bivalves (oyster genome; Zhang et al. 2012), the *M. arenaria* genome is likely to be on the order of 2 gigabases. A single MiSeq run would produce up to 1-2 gb of sequence in 100 base increments, which when assembled into contigs would represent less than 1 gb. Thus, it will produce an incomplete clam genome data set. However, in an infected tissue sample, the number of viruses (and their small genomes) may out-number the host cells by several orders of magnitude. Thus, even though the MiSeq analysis would be unsuccessful in producing a full host genome, the representation of virus genomes would be more than sufficient to produce a full length genome.

Initial assembly of the sequencing reads produced over 20,000 contigs per sample with varying average depth of coverage, but most contigs with over 8-fold coverage, which is good. One encouraging signal is that the distribution of contig sizes ranges from a few hundred bases to approximately 25,000 bases. Most encouraging is that there is a discontinuous jump in contig length between 8.5 kb and a small number of contigs 14-25 kb, as might be expected if a DNA sample contained a large amount of host genomic DNA and a fraction of viral DNA (Fig. 11). If the DNA sequenced consisted of a substantial amount of clam DNA with an enrichment of virus DNA (which is highly represented on a molar basis but still a minority of the DNA by mass), the prediction for MiSeq data is similar to what we have obtained.

Preliminary BLASTx analysis (searching a protein database with all possible protein translations of the DNA sequences) indicates the two largest contigs have no homology to any protein-coding genes in GenBank. This may be expected of a novel virus, because GenBank consists mostly of

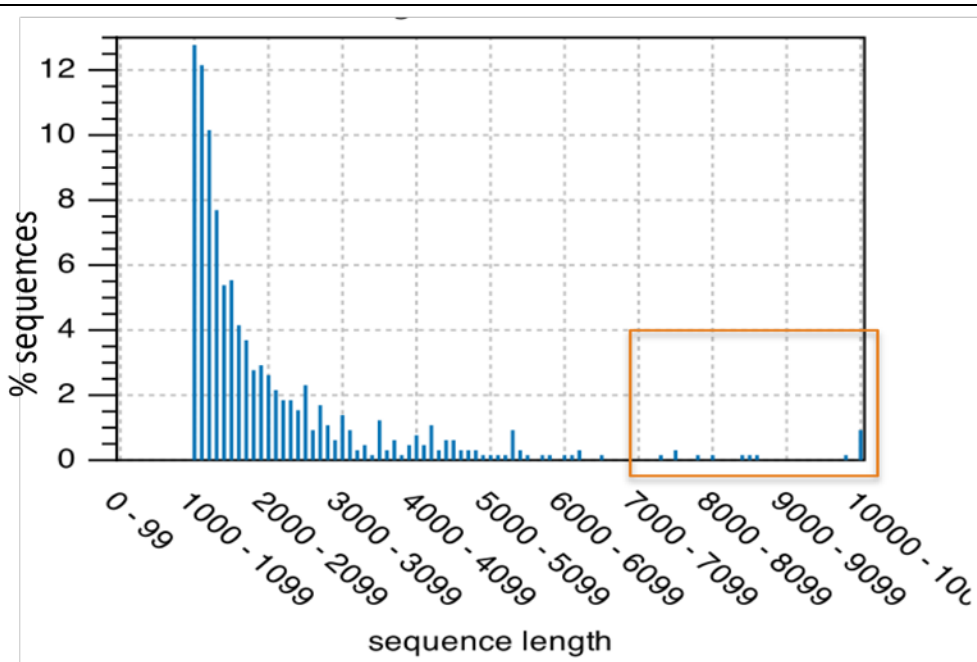


Fig. 11. Size distribution of contigs generated by assembly of individual sequences from DNA enriched for virus-like particles from clam 080930 CHBRMa-4. The inset-box encompasses larger contigs that will be investigated first, as possible parts of the GENH virus genome.

metazoan and bacterial sequences, and viruses from invertebrate hosts are poorly represented. All documented adenoviruses are from tetrapod hosts, and it is generally true that virus genomes diverge rapidly, such that even within a virus family, there is little sequence conservation between related viruses infecting hosts from different taxa.

Next steps are to look in detail at the comparison of the large contigs obtained from the three DNA samples. One prediction is that the un-enriched DNA of clam 080930 CHBRMa-4 will have less of the GENH virus DNA than the enriched subsample. This would be manifest as a contig that has fewer reads/less depth in the un-enriched DNA. Another prediction is that the largest contigs from DNA isolated from different infected clams should be overlapping if not almost identical.

The samples of infected clam tissue obtained for this study had the highest histological score for GENHV. However, the virus infects only the epithelial tissue and only a subset of those cells carry heavy virus loads, so on a total tissue basis (10-40 mg preserved in ethanol or -80), the virus may not be highly over-represented in the total DNA. If the virus enrichment depicted in Fig. 11 was successful, then the DNA extracted from the putative virus fraction should have a higher proportion of virus sequences than would be found in un-enriched DNA. Optimally, these sequences would assemble into a large contig that represents some or all of the GENHV genome. The DNA isolated from clam tissue that did not undergo virus enrichment on CsCl would also have the potential to contain virus sequences, but also might not.

<u>An in-depth inspection of contigs for characteristics expected of a virus genome.</u>
For each DNA preparation (two enriched and one unenriched) the set of BLASTx results from contigs exceeding 1000 bases in length and with at least 10-fold coverage were examined in detail by visual inspection of each BLAST result. Over 1/3 of the sequences had no similarity to GenBank proteins. Of the remaining contigs, most encoded proteins associated with transposable elements, which are frequently found in multicopy in most metazoan genomes. The hallmark of transposons is the gag-pol polyprotein, often also referred to as RNA-directed DNA polymerase. Other transposon-related proteins are helicases and endonucleases. This finding, that much of the *M. arenaria* genome is made up of transposons is interesting in that transposons and viruses have shared ancestry, but not immediately relevant to the search for GENHV sequences.

One 16 kb contig in a virus-enriched DNA sample showed homology to all of the genes in the mitochondrial genome of bivalves. This is a reassuring confirmation that an episomal (organellar) DNA circle can be sequenced using the methodology chosen. The fact that the intact 16 kb mitochondrial genome was not found in all three DNA sequences also serves as an illustration that episomal DNA (such as a virus) may not fully assemble into an unbroken contig even if present. The full *M. arenaria* mitochondrial genome was recently deposited into GenBank (Aug. 2014) by the University of Maine (accession YP_009054293.1).

The largest contigs in each data set are particularly interesting because they could represent major parts of the GENHV genome:

First, the largest contig in enriched 080930 CHBRMa-4, and 120413 EBLPMa-24 is approximately 34 kilobases. The sequence corresponds to two contigs of 16 kb and 17 kb (summing to 33 kb) in the unenriched 080930 CHBRMa-4. This contig encodes no similarity to any protein in GenBank.

Second, one contig of about 24 kb, present in all three DNA preparations, has a gene encoding a histone-like protein and is represented by very high coverage (300-700 fold, while other contigs have 30-50 fold coverage) in both enriched and unenriched samples from 080930 CHBRMa-4. Adenoviruses also encode histone-like proteins (Lischwe and Sung 1977), which are involved in packaging virus DNA into capsids. This contig is more abundant in the DNA extracts of CsCl-enriched material, consistent with it being associated with a virus-like particulate. Because of the extraordinary abundance of this sequence in the MiSeq dataset, we refer to this contig as "300x". The histone-like genes in 300x has over 60% identity, across 99 amino acids, to histone H3 from a variety of organisms, from protists and plants to tetrapods.

Third, there is one ~24 kb contig that appears to be present only in the DNA from virus-enriched preparations. Even when segments of this 24 kb are used to search the smaller contigs (albeit those still above 1 kb in length) from the un-enriched DNA, no similarity is found. It is surprising that this sequence appears totally absent from the DNA of 080930 CHBRMa-4 extracts that had not been subjected to CsCl enrichment for virus. Given that the virus was present in this explant, it was expected that virus-like sequences would be present in unenriched DNA, but just at lower copy or in incomplete contigs.

Table 1. List of PCR product sizes expected for each primer pair, and the results anticipated if each primer set is specific for GENHV genome sequences. For comparison, a hypothetical primer set for the actin gene (clam nuclear, not viral) is listed.

Primer Set	**Expected target**	**PCR product size**	**Expected abundance in uninfected clam DNA**	**Expected abundance in GENHV clam DNA**
MyaMito-1	Mitochondria	157 bp	++	++
MyaMito-2		125 bp	++	++
GENH34-1	GENHV	168 bp	-	++
GENH34-2		117 bp	-	++
GENH24-1	GENHV	168 bp	-	++
GENH24-2		111 bp	-	++
GENHV300x	GENHV	144 bp	-	++
Mya Actin	Nuclear	hypothetical	+	+

<u>Success in identifying the GENHV genome</u>

To develop experimental evidence that one or all of the Table 1 contigs is truly the GENHV genome, we designed oligos to PCR amplify short segments of each contig. We also designed primers specific for the apparent mitochondrial genome. These primers were used in amplifications of DNA extracted from *M. arenaria* that had scored positive and negative for the GENH virus by histology. DNA was extracted from the frozen GENHV+ and GENHV- material using a commercial spin column kit (MoBio). An equal amount of DNA was used in PCR amplifications with each primer set, using a 96-well PCR format, and visualizing the PCR products on an agarose gel.

Experimental findings show that the 300x primer set appears to identify GENHV-infected DNA but not DNA of GENHV-free *M. arenaria* (Fig. 12). Of the four putative virus genome contigs, only the primer set designed to "300x" produced results that are consistent with GENHV-specificity. However, the correlation is not exact: it is seen that DNA of clam EB39 produced the same size PCR product as seen in the for GENHV+ clams. It is important to point out that the histological score of GENHV-negative was derived from one part of the clam gill, while the DNA was extracted from a different part of the clam gill. Thus, there is likelihood that the histological score and the PCR result are both accurate with respect to the presence of GENHV. Additionally, the minimum amount of virus detectable by histology is far higher than the amount needed to detect the virus using PCR methods.

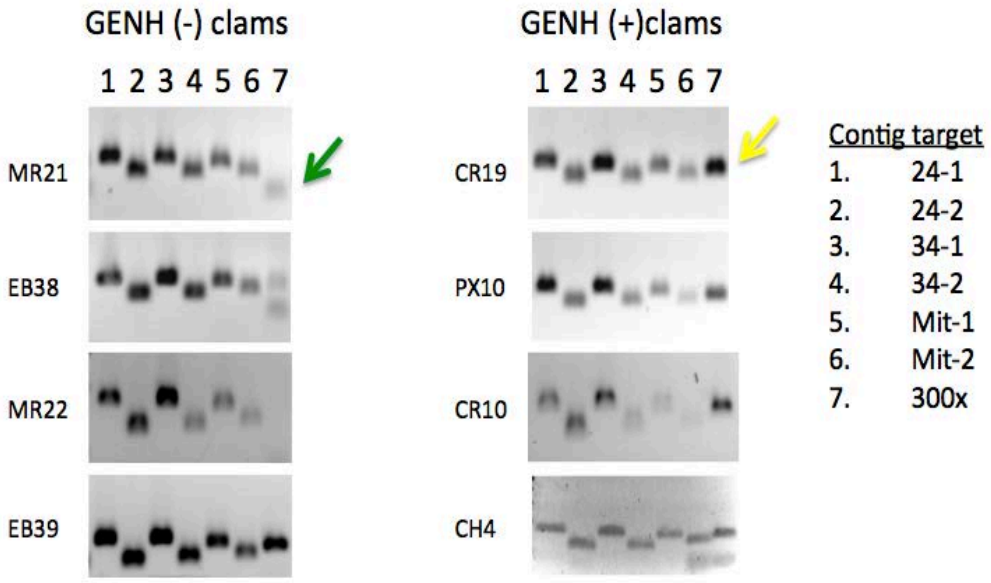


Fig. 12. Screening PCR primers for correlation with GENHV histological scores. Primers designed to the 24 kb, 34 kb, mitochondrial, and "300x" contigs were applied to DNA extracted from clam gills that scored as GENHV negative and GENHV positive. As can be seen in the agarose gel images of PCR products, the primer pair designed against the 300x contig was specific for GENHV. The green arrow points to a primer-dimer, a non-specific product; the yellow arrow points to the specific product of expected size for the 300x primer set. Two primer pairs were tested for all targets except for 300x.

Predicted GENHV genome sequence

Figure 13 depicts the open reading frames (ORFs) of the putative GENHV genome. The prediction of a few large ORFs and many smaller ORFs is at first confusing for a virus genome, because many viruses have simple intronless gene organization. However, this pattern is entirely consistent with the structure of adenovirus genomes, which are complex and contain genes with many introns and alternative splice variants (see also Fig. 5). Indeed, the first, Nobel Prize-winning discovery of introns in 1977, was made in adenoviruses (Chow et al 1977; Berget et al. 1977). Given the morphology, cell type affected, and genome characteristics of GENHv, it

appears that the sequence of the 300x contig may represent a first adeno-like virus genome from an invertebrate host.

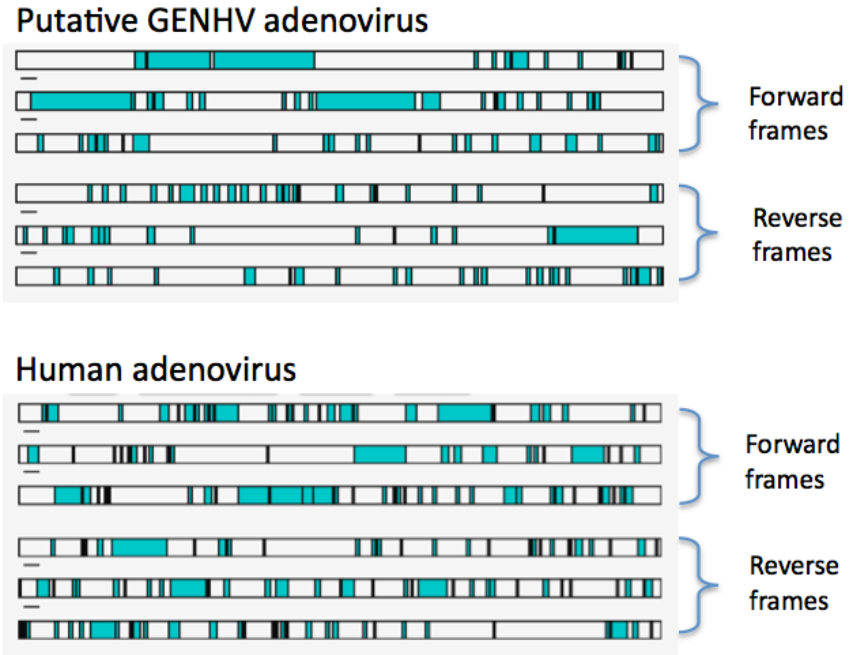


Fig. 13. Open reading frame maps of putative GENHV and human adenovirus C. Blue-green bars represent open reading frames (ORFs) that begin with ATG and end with a standard stop codon. **Top**: Open reading frames in the 23,861 bp sequence of GENHV. **Bottom**: ORFs encoded in the 35,934 bp human adenovirus C sequence. Unlike many other dsDNA viruses, adenoviruses encode genes with introns and many alternative transcripts. This results in the pattern of multiple short ORFs displayed by both genomes depicted here.

Future studies needed to confirm that contig 300x is indeed the genome of GENHV and develop a molecular probe for the GENH virus.

The goal of this research proposal was to develop tools that will allow rapid, sensitive, and accurate detection of the GENH virus in clam tissues and environmental samples. The success identifying a putative GENHV genome sequence and confirmation that the PCR amplification of 300x-derived sequence correlates with the GENH phenotype is very encouraging. The critical study that needs to be conducted next is outlined in Figure 14: it is to develop and apply an in situ hybridization probe that is specific for GENHV. Using the genomic information obtained in this study, it is envisioned that a series of oligonucleotide probes will be designed to anneal to either the 300x (GENHV) sequences or to other sequences that are predicted to be present on the mitochondrial or nuclear genome of *M. arenaria*.

The apparent success in identifying the GENHV genome will now open many avenues of research, from wider prevalence studies, including retrospective examination of archived material, to investigations of where else in the clam or in the environment the GENHV may occur.

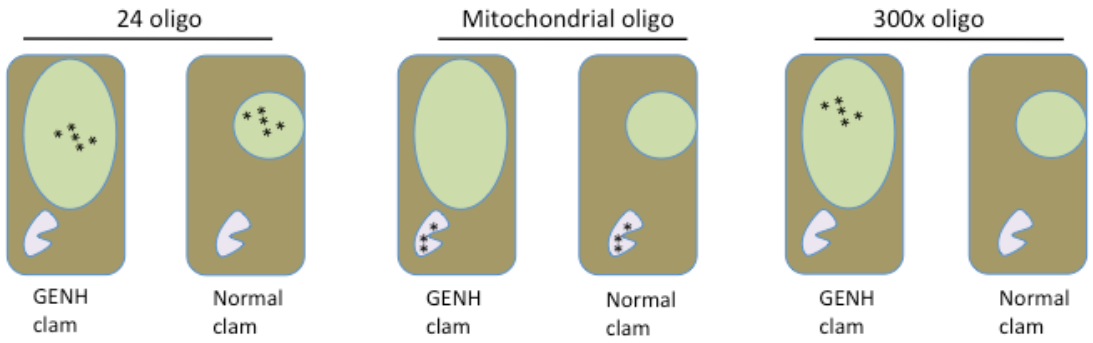


Fig. 14. Predicted results of in situ hybridizations with oligonucleotide probes designed to the GENHV sequences and other contigs identified in MiSeq data (Table 1). Cells depicted are hypothetical GENH afflicted and normal clam gill epithelial cells. Asterisks represent *in situ* hybridization signal with oligonucleotides designed to anneal to sequences in contigs Clam24, Mito, or 300x-virus.

Summary

Despite unanticipated technical difficulties that delayed them, significant probable achievements were ultimately made toward the proposed project goals of acquiring extensive or complete genome sequences for the virus that causes lesions in softshell clam gill tissues, and toward developing molecular-genetic tools to determine the mechanisms and impacts of that disease on severely depleted *Mya arenaria* populations of Chesapeake Bay.

Although we lack complete confirmation yet, we believe that we have sequences representing the entire genome of the unidentified virus that pathologically infects nuclei of *Mya arenaria* softshell clam gill epithelia at high prevalences and intensities. Further efforts with additional funding, should allow us to completely construct the organized genome of the GENH virus.

We expect to use the pending virus genome information to test our tentative hypothesis that the GENH virus may be the first adeno-like virus known to infect an invertebrate host.

We plan to use virus genome sequence information to develop and validate PCR assays for detection and enumeration of the agent in clam tissues, and in environmental water and sediment samples, in order to document the transmission and pathogenesis of the new virus disease among *Mya arenaria* clams in Chesapeake Bay and elsewhere.

We also plan to also use the virus genome sequence information to develop and validate DNA probes and assays to detect and localize virus lesions among *Mya arenaria* clams from Chesapeake Bay and elsewhere, whose histological tissue samples have been archived during 50 recent years.

Acknowledgements

The PIs gratefully acknowledge the expert assemblies and bioinformatic analyses of MiSeq DNA sequencing data by Ms. Emily Flowers, MS student at UMCES-IMET. The expert insight of Esther C. Peters, PhD is also applauded, for her pioneering 2004 recognition of GENH as a viral pathology of softshell clam gills.

Literature cited

Arzul, I., T. Renault, A. Thébault and A. Gérard 2002. Detection of oyster herpesvirus DNA and proteins in asymptomatic *Crassostrea gigas* adults. *Virus Research* 84:151-160

Berget SM, Moore C, Sharp PA (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA". Proc. Natl. Acad. Sci. U.S.A. 74 (8): 3171–5

Bowers, H.A., G.A. Messick, A. Hanif, R. Jagus, L. Carrion, O. Zmora, and E.J. Schott 2010. Physicochemical properties of double-stranded RNA used to discover a reo-like virus from blue crab *Callinectes sapidus*. *Dis. Aquat. Org*. 93:17-29.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, and J.D. Thompson 2003. Multiple sequence alignment with the Clustal series of programs *Nucl. Acids Res*. 31(13): 3497-3500.

Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA". Cell 12 (1): 1–8.

Dryer, H. and G. Steiner 2006. The complete sequences and gene organization of the mitochondrial genomes of the heterodont bivalves Acanthocardia tuberculata and Hiatella arctica – and the first record for a putative ATPase subunit 8 gene in marine bivalves. Frontiers in Zoology 3:13.

Dungan, C.F., E.C. Peters, and J. Marcino 2012. Three diseases of softshell clams, Mya arenaria, in Chesapeake Bay: 2000-2009. J. Shellfish Res. 31:278 (abstract).

Dungan, C.F, M.L. Homer, M. Tarnowski, E.C. Peters, K.S. Reece, and E.M. Burreson 2009 (presentation). Chesapeake Bay clam fishery collapse: probable disease impacts. NOAA Chesapeake Bay Office Fisheries Science Symposium, Laurel, MD, April 2009. http://chesapeakebay.noaa.gov/images/stories/pdf/fss09program.pdf

Dungan, C.F., E.C. Peters, and R.M. Hamilton 2007 (presentation). Virus infection associated with nuclear hypertrophy in gill epithelial cells of Chesapeake Bay softshell clams, *Mya arenaria*. World Aquaculture Society 2007 meeting, San Antonio, TX, March 2007.

Dungan, C.F., K.S. Reece, R.M. Hamilton, N.A. Stokes, and E.M. Burreson 2007. Experimental cross-infections by *Perkinsus marinus* and *P. chesapeaki* in three sympatric species of Chesapeake Bay oysters and clams. *Dis. Aquat. Org*. 76:67-75.

Dungan, C.F., R.M. Hamilton, K.L. Hudson, C.B. McCollough and K.S. Reece 2002. Two epizootic diseases in Chesapeake Bay commercial clams *Mya arenaria* and *Tagelus plebeius*. *Dis. Aquat. Org.* 50:67-78.

Elston, R.A. 1997. Special topic review: bivalve mollusc viruses. *World J. Microbiol. Biotechnol.* 13:393–403.

Folmer O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit 1 from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3:294–299.

Garcia, C., M. Robert, I. Arzul, B. Chollet, J.-P. Joly, L. Miossec, T. Comtet, and F. Berthe 2006. Viral gametocytic hypertrophy of *Crassostrea gigas* in France: from occasional records to disease emergence? *Dis. Aquat. Org.* 70:193–199.

Harrach, B. and M. Benko 2007. Phylogenetic Analysis of Adenovirus Sequences. In: W.S.M. Wolfe and A.E. Tlooefson (ed.) Methods in Molecular Medicine, Vol. 131: Adenovirus Methods and Protocols, Second Edition, vol. 2: Ad Proteins, RNA, Lifecycle, Host Interactions, and Phylogenetics. Humana Press Inc., Totowa, NJ.

Homer, M.L., C.F. Dungan, and M.L. Tarnowski 2011. Assessment of Chesapeake Bay commercial clams *Mya arenaria* and *Tagelus plebeius*, with emphasis on abundance and disease status. Completion report for project NA07NMF4570326, November 28, 2011, the NOAA Chesapeake Bay Office Fisheries Research Program, Annapolis, Maryland, 28 pp.

Li, C., J.D. Shields, H. J. Small, K.S. Reece, L. Carmony, R.A, Hartwig, R.A. Cooper, R.E. Ratzlaff 2007. Detection of *Panulirus argus* Virus 1 (PaV1) in the Caribbean spiny lobster using fluorescence in situ hybridization (FISH). *Dis. Aquat. Org*. 72:185–192.

Lischwe M.A., Sung M.T. 1977. A histone-like protein from adenovirus chromatin. *Nature*. 267:552–554.

Nagle, L., A.R. Place, E.J. Schott, R. Jagus, G. Messick, and J.S. Pitula 2009. A real-time quantitative PCR-based assay for enhanced detection of *Hematodinium* sp. infection in the blue crab (*Callinectes sapidus*). *Dis. Aquat. Org.* 84:79-87

Ng, T. F., W.K. Suedmeyer, E. Wheeler, F. Gulland, and M. Breitbart. 2009. Novel anellovirus discovered from a mortality event of captive California sea lions. *J. Gen. Virol.* 90:1256-1261

Ojkic, D. & Nagy, E. 2000. The complete nucleotide sequence of fowl adenovirus type 8. *J. Gen. Virol.* 81:1833–1837.

Ozbun, M.A. 2002. Infectious human papillomavirus type 31b: purification and infection of an immortalized human keratinocyte cell line. *J. Gen. Virol.* 83:2753-63.

Reece, K.S., C.F. Dungan and E.M. Burreson 2008. Molecular epizootiology of *Perkinsus marinus* and *P. chesapeaki* infections among wild oysters and clams in Chesapeake Bay, USA. *Dis. Aquat. Org*. 82:237-248.

Reece, K. and C. Dungan 2006. *Perkinsus* sp. infections of marine molluscs. *In*: AFS-FHS blue book: suggested procedures for detection and identification of certain finfish and shellfish pathogens, 2005 edition, Chapter 5.2.1-17, Fish Health Section, American Fisheries Society, Bethesda, MD

Schott, E.J., H.A. Bowers, A. Hanif, G.A. Messick, D. McIntosh, J.S. Pitula, and R. Jagus 2010. Development and testing of a molecular tool for environmental monitoring of *Hematodinium* sp., a lethal parasite of the blue crab *Callinectes sapidus*. Oral presentation, World Aquaculture Society meeting, March 1-4, 2010. San Diego, CA.

Scott, J.M. and J.S. Chamberlain 2003. Gutted Adenoviral Vectors for Gene Transfer to Muscle in: Methods in Molecular Biology, vol. 219: Cardiac Cell and Gene Transfer. J.M. Metzger (ed.) Humana Press Inc., Totowa, NJ. pp 19-28.

Strasser, C.A. and P.H. Barber 2009. Limited genetic variation and structure in softshell clams (*Mya arenaria*) across their native and introduced range. *Conserv. Genet.* 10:803–814.

University of Kentucky Department of Biological Safety data sheet Guidelines for Research Involving Viral Vectors. ehs.uky.edu/docs/pdf/bio_viral_vectors_0001.pdf

Walz, C., M. Anisi, T.R. Schlehofer, J.R.Gissmann, L. Schneider, and M. Muller. 1998. Detection of infectious adeno-associated virus particles in human cervical biopsies. *Virology* 247:97–105.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490:49-54.