



UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

CAYRO DE MACÊDO MENDES

**CARACTERIZAÇÃO IN SILICO DE ORFs VARIÁVEIS E DE REGIÕES
REGULATÓRIAS NO GENOMA DO VÍRUS DA SÍNDROME DA MANCHA
BRANCA (WSSV)**

NATAL - RN

2018

CAYRO DE MACÊDO MENDES

**CARACTERIZAÇÃO IN SILICO DE ORFs VARIÁVEIS E DE REGIÕES
REGULATÓRIAS NO GENOMA DO VÍRUS DA SÍNDROME DA MANCHA
BRANCA (WSSV)**

Defesa de Mestrado apresentada ao Programa
de Pós-Graduação em Bioinformática da
Universidade Federal do Rio Grande do Norte

Área de concentração: Bioinformática

Linha de Pesquisa: Desenvolvimento de
Produtos e Processos

Orientador: Prof. Dr. Daniel Carlos Ferreira
Lanza

Co-orientador: Prof. Dr. João Paulo Matos
Santos Lima

NATAL-RN

2018

Ficha catalográfica

Universidade Federal do Rio Grande do Norte - UFRN

Sistema de Bibliotecas - SISBI

Catálogo de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Leopoldo Nelson - -Centro de Biociências - CB

Mendes, Cayro de Macêdo.

Caracterização in silico de orfs variáveis e de regiões regulatórias no genoma do vírus da síndrome da mancha branca (WSSV) / Cayro de Macêdo Mendes. - Natal, 2019.

80 f.: il.

Dissertação (Mestrado) - Universidade Federal do Rio Grande do Norte. Centro de Biociências. Programa de Pós-Graduação em Bioinformática.

Orientador: Prof. Dr. Daniel Carlos Ferreira Lanza.

Coorientador: Prof. Dr. João Paulo Matos Santos Lima.

1. Penaeus vannamei - Dissertação. 2. Camarão - Dissertação. 3. Whispovirus - Dissertação. 4. ORF94 - Dissertação. 5. ORF75 - Dissertação. 6. ORF125 - Dissertação. I. Lanza, Daniel Carlos Ferreira. II. Lima, João Paulo Matos Santos. III. Universidade Federal do Rio Grande do Norte. IV. Título.

CAYRO DE MACÊDO MENDES

**CARACTERIZAÇÃO IN SILICO DE ORFs VARIÁVEIS E DE REGIÕES
REGULATÓRIAS NO GENOMA DO VÍRUS DA SÍNDROME DA MANCHA
BRANCA (WSSV)**

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em
Bioinformática da Universidade Federal do Rio Grande do Norte

Área de concentração: Bioinformática

Linha de Pesquisa: Desenvolvimento de Produtos e Processos

Orientador: Prof. Dr. Daniel Carlos Ferreira Lanza

Co-orientador: Prof. João Paulo Matos Santos Lima

Natal, 19 de novembro de 2018.

BANCA EXAMINADORA

Prof. Dr. Daniel Carlos Ferreira Lanza
Universidade Federal do Rio Grande do Norte
(Presidente)

Prof. Dr. Euzébio Guimarães Barbosa
Universidade Federal do Rio Grande do Norte
(Examinador Interno do Programa)

Prof. Dr. Sávio Torres de Farias
Universidade Federal da Paraíba
(Examinador Externo à Instituição)

AGRADECIMENTOS

Agradeço aos meus pais por todo carinho, amor e suporte ao longo desses vinte e quatro anos, principalmente nesta nova etapa da minha vida na qual estou longe de casa.

Ao Professor Daniel Lanza pela confiança em depositar em minhas mãos um projeto ambicioso e que me fez entender que a ciência é um trabalho contínuo e que exige autonomia por parte do pesquisador. Levarei este ensinamento adiante e me tornarei melhor a cada dia.

Ao Professor João Paulo pela paciência e atenção diante dos meus questionamentos repentinos, o que me permitiu ter uma base mais sólida e aprender a trilhar novas soluções diante de um problema.

Aos colegas do LAPLIC e do BioME pelos momentos descontraídos e troca de ideias que contribuíram de forma significativa durante estes dois anos.

*“A soul in tension that's learning to fly
Condition grounded but determined to try”
(Pink Floyd)*

RESUMO

O vírus causador da síndrome da mancha branca (WSSV) é um dos maiores problemas enfrentados pela carcinicultura mundial, causando consideráveis danos econômicos. O genoma do WSSV apresenta algumas regiões codificantes que variam entre os diferentes isolados. Essas regiões denominadas wsv129 (ORF75), wsv178 (ORF94), wsv249 (ORF125), wsv461/464 (ORFs14/15) e wsv477/502 (ORFs23/24) possivelmente estão envolvidas em mecanismos de virulência, mas não foram totalmente caracterizadas funcionalmente até o momento. A caracterização *in silico* vem sendo empregada como uma alternativa mais acessível para predição e estudo da estrutura de proteínas que não tem a estrutura cristalográfica disponível. Esse trabalho teve como objetivo a caracterização *in silico* das proteínas putativas codificadas pelas regiões variáveis do genoma do WSSV, no intuito de identificar possíveis funções. As regiões codificantes das ORFs wsv129, wsv178, wsv249 e os clusters formados pelas ORFs wsv461/464 e wsv477/502 foram analisados filogeneticamente, e estruturalmente. As sequências de aminoácidos foram submetidas a buscas por homólogos remotos, motivos, domínios conservados, reconhecimento de *fold* e predição estruturas secundárias e terciárias. Foi possível modelar estruturas terciárias de domínios proteicos e inferir possíveis funções para as ORFs wsv463a (*formina like* - regulação de filamentos de actina), wsv477 (interação com RNA - processos pós - transcricionais), wsv479 e wsv497 (*XPD like* - helicase), wsv 492 (hemaglutinina *like* - sinalização, *docking* viral) e wsv249 (domínios *ankyrin repeat* e RING-H2 - modulação da proteólise dependente de Ubiquitina), além de propor funções estruturais para as ORFs wsv 129 e wsv 178 em decorrência suas características estruturais e carga. Também foi possível detectar assinaturas associadas a sinais de localização nuclear dentro das unidades de repetição das sequências codificadas pelas ORFs wsv129 e wsv178. Além das análises de estrutura proteica foi realizada a prospecção de algumas regiões regulatórias 100 e 200 nt upstream às regiões codificantes e foi possível detectar alguns motivos, incluindo um sítio de ligação de “Zinc-Finger”, sugerindo a interação entre possíveis fatores de transcrição. A partir dos resultados foi proposto um modelo de atuação para cada uma das proteínas estudadas.

Palavras-chave: *Penaeus vannamei*, Camarão, Whispovirus, ORF94, ORF75, ORF125, VNTRs

ABSTRACT

The white spot syndrome virus (WSSV) is one of the biggest problems facing global shrimp farming, causing considerable economic losses. The WSSV genome presents some coding regions that vary between the different isolates. Those regions named wsv129 (ORF75), wsv178 (ORF94), wsv249 (ORF125), wsv461 / 464 (ORFs 14/15) and wsv477 / 502 (ORFs 23/24) are possibly involved in virulence mechanisms but have not been fully characterized functionally so far. The *in silico* characterization has been used as a more accessible alternative for prediction and study of the protein structure that does not have the crystallographic structure available. This work aimed to characterize *in silico* the putative proteins encoded by the variable regions of the WSSV genome, in order to identify possible functions. The variable regions of the ORs wsv129, wsv178, wsv249 and the clusters formed by the ORs wsv461/464 and wsv477/502 were analyzed phylogenetically, and structurally. The amino acid sequences were subjected to remote homologous searches of motifs, conserved domains, fold recognition and prediction of secondary and tertiary structures. It was possible to model tertiary structures of protein domains and infer possible functions for the ORs wsv463a (formin like - regulation of actin filaments), wsv477 (interaction with RNA - posttranscriptional processes), wsv479 and wsv497 (XPD like - helicase), wsv 492 (hemagglutinin like signaling, viral docking) and wsv249 (ankyrin repeat and RING-H2 domains - modulation of Ubiquitin-dependent proteolysis), besides proposing structural functions for the ORs wsv129 and wsv178 due to their structural characteristics and charge. It was also possible to detect signatures associated with nuclear localization signals within the repeating units of the sequences encoded by the ORs wsv129 and wsv178. In addition to protein structure analyzes, some 100 and 200 nt upstream regulatory regions were prospected for the coding regions and it was possible to detect some motifs, including a Zinc-Finger binding site, suggesting the interaction between possible transcription factors. From the results it was proposed a model of performance for each of the proteins studied.

Keywords: *Penaeus vannamei*, Shrimp, Whispovirus, ORF94, ORF75, ORF125, VNTRs

Lista de Figuras

Figura 1 – Camarão apresentando os principais sintomas da síndrome da mancha branca. As setas vermelhas apontam as manchas brancas calcificadas em seu exoesqueleto. Fonte: Aquanetviet. Disponível em: <<http://aquanetviet.com/post/133900619760/ph%C3%A1t-tri%E1%BB%83n-c%C3%B4ng-c%E1%BB%A5-duplex-gi%C3%B0p-pcr-ph%C3%A1t-hi%E1%BB%87n-th%E1%BB%9D>> Acessado em 13/12/2018.16

Figura 2 – Representação esquemática da partícula viral do WSSV com suas respectivas dimensões. Fonte: Sanchez-Paz, 2010.17

Figura 3 – Influência das regiões variáveis na reconstrução filogenética do WSSV. Dendrogramas foram obtidos a partir do alinhamento de genomas completos do WSSV isolados de diferentes regiões (A), na ausência das regiões variáveis (B) e na ausência das regiões homólogas (C). Para o cálculo de cada dendrograma foi utilizada a abordagem bayesiana utilizando o modelo GTR com distribuição gama, relógio molecular estrito, Cadeia Markov com 10^7 gerações e consenso de árvores com exclusão de 10% das árvores iniciais.30

Figura 4 – Alinhamento das regiões correspondentes ao cluster das ORFs wsv461/464 (ORF14/15). As sequências presentes no alinhamento representam variantes virais que foram isoladas de diferentes regiões do mundo. O alinhamento e anotação das sequências codificantes putativas foi realizado com auxílio do programa Geneious versão 11.0.3. As setas em cinza correspondem as ORFs putativas presentes dentro das inserções.33

Figura 5 – Alinhamento das regiões correspondentes ao cluster das ORFs wsv477/502 (ORF23/24). As sequências presentes no alinhamento representam variantes virais que foram isoladas de diferentes regiões do mundo. O alinhamento e anotação das sequências codificantes putativas foi realizado com auxílio do programa Geneious versão 11.0.3. As setas em cinza correspondem as ORFs putativas presentes dentro das inserções.34

Figura 6 – Caracterização estrutural da proteína putativa codificada pela wsv463a. (A) Representação esquemática da sequência proteica codificada pela wsv463a e seus respectivos domínios conservados. O domínio FH1, composto por resíduos de poliprolina está representado em vermelho, enquanto o domínio FH2 está representado em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv463a correspondente ao domínio *Formin Homology 2* (FH2). Subdomínios conservados (*Lasso*, *Linker*, *Knob*, *Coiled Coil* e *Post*) e aminoácidos do domínio FH2 típico são indicados. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no PyRAMA. (E) Escores globais gerados no QMEAN.38

Figura 7 – Caracterização estrutural da proteína putativa codificada pela wsv477. (A) Representação esquemática da sequência proteica codificada pela wsv477 e seus

respectivos domínios conservados. O motivo de reconhecimento de RNA está representado em laranja, um sítio de ligação de ATP na região N-terminal em azul, enquanto o domínio *Zinc-Finger* do tipo Cys2/Cys2 está representado em vermelho. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv477 correspondente ao motivo de reconhecimento de RNA (RRM). (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no PyRAMA. (E) Escores globais gerados no QMEAN.41

Figura 8 – Caracterização estrutural da proteína putativa codificada pela wsv479. (A) Representação esquemática da sequência proteica codificada pela wsv479 e seus respectivos domínios conservados. A região correspondente a XPD Helicase está representada em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv479 correspondente a XPD Helicase. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.44

Figura 9 – Caracterização estrutural da proteína putativa codificada pela wsv497. (A) Representação esquemática da sequência proteica codificada pela wsv497 e seus respectivos domínios conservados. A região correspondente a XPD Helicase está representada em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv497 correspondente a XPD Helicase. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.46

Figura 10 – Caracterização estrutural da proteína putativa codificada pela wsv492. (A) Representação esquemática da sequência proteica codificada pela wsv492 e seus respectivos domínios conservados. A região correspondente a subunidade HA2 da Hemaglutinina do Influenzavirus está representada em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv492 correspondente a subunidade HA2 da Hemaglutinina. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.48

Figura 11 – Representação esquemática da região variável wsv178 (ORF94). (A) Alinhamento múltiplo das sequências de nucleotídeos obtido através do Geneious versão 11.0.3. As setas em cinza representam as unidades de repetição de 54pb. (B) Representação da unidade de repetição através do WebLogo mostrando os sítios de polimorfismo.....49

Figura 12 – Representação esquemática da região variável da wsv249 (ORF125). (A) Alinhamento múltiplo das sequências de nucleotídeos obtido através do Geneious versão 11.0.3. As setas em cinza representam as unidades de repetição de 69pb. (B)

Representação da unidade de repetição através do WebLogo mostrando os sítios de polimorfismo.....50

Figura 13 – Representação esquemática da região variável da wsv129 (ORF75). (A) Alinhamento múltiplo das sequências de nucleotídeos obtido através do Geneious versão 11.0.3. As setas em cinza representam as unidades de repetição de 45pb, enquanto as setas em azul claro representam as unidades de 57pb. (B) Representação da unidade de repetição de 45pb através do WebLogo mostrando os sítios de polimorfismo. (C) Representação da unidade de repetição de 57pb através do WebLogo mostrando os sítios de polimorfismo.51

Figura 14 – Representação esquemática da sequência proteica codificada pela wsv129 (ORF75) e seus respectivos domínios conservados. (A) A região transmembrana (TM) está representada em rosa, as unidades de repetição correspondentes as sequências codificante de 45pb estão em vermelho, enquanto a repetições intercaladas correspondentes as unidades de repetição de 57pb estão marcadas em verde. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Gráfico representando a probabilidade de ocorrência de *Coiled coils* dentro da sequência proteica codificada pela wsv129.54

Figura 15 – Representação esquemática da sequência proteica codificada pela wsv178 (ORF94) e seus respectivos domínios conservados. A região transmembrana (TM) está representada em rosa, o sítio de clivagem se apresenta como o quadrado amarelo, enquanto as unidades de repetição correspondentes as sequências codificante de 54pb estão em vermelho. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade.56

Figura 16 – Caracterização estrutural da proteína putativa codificada pela wsv249 (ORF125) na posição 86-293. (A) Representação esquemática da sequência proteica codificada pela wsv249 e seus respectivos domínios conservados. A região correspondente a subunidade ao motivo ANK está representada em laranja, o domínio “Ring-H2” em azul, unidades de repetição caracterizados como motivos de interação com ubiquitina (UIM) em vermelho e uma região *Coiled coil* em verde. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv249 correspondente ao motivo ANK. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.58

Figura 17 – Caracterização estrutural da proteína putativa codificada pela wsv249 (ORF125) na posição 310-357. (A) Modelo tridimensional da sequência codificada pela wsv249 correspondente ao domínio RING-H2. (B) Escores de validação da qualidade do modelo obtido através do Molprobit. (C) *Plots* de Ramachandram gerados no no PyRAMA. (D) Escores globais gerados no QMEAN.60

Lista de Tabelas

Tabela 1 – Dados das sequências codificantes putativas encontradas entre os indels da wsv461/464(ORF14/15) e wsv477/502(ORF23/24) obtidos a partir do genoma do WSSV-CN01 (Genbank Accession: KT995472. 1).	19
Tabela 2 – Sequências de repetições <i>in tandem</i> presentes nas wsv129 (ORF75), wsv178 (ORF94) e wsv249(ORF125) obtidas a partir do genoma do WSSV-CN01 (Genbank Accession: (KT995472. 1).	20
Tabela 3 – Caracterização de sequências correspondentes ao cluster wsv461/464 (ORF14/15) obtidos a partir de diferentes isolados de WSSV.	32
Tabela 4 – Caracterização das sequências correspondentes ao cluster das ORF wsv477/wsv502 (ORF23/24).	35
Tabela 5 – Relação das unidades de repetição das VNTRs presentes no WSSV.	52
Tabela 6 – Perfís de sequências motivos de maior ocorrência dentro das regiões 200nt upstream das sequências codificantes do genoma do WSSV obtidos através do software MEME.	62

Lista de Abreviaturas e Siglas

ANK – *Ankyrim repeat*

DNA – *Deoxyribonucleic acid*

FH1 – *Formin Homology 1*

FH2 – *Formin Homology 2*

GBV-C – *GB Virus C*

IBDV – *Infectious Bursal Disease Virus*

ICTV – *International Committee on Taxonomy of Viruses*

Msa – *Modulator of staphylococcal accessory regulator*

NLS – *Nuclear Localization Signal*

ORF – *Open Reading Frames*

PCR – *Polymerase Chain Reaction*

PDB – *Protein Data Bank*

RdRp – *RNA-dependent RNA Polymerase*

RNP – *Ribonucleoprotein*

RRM – *RNA Recognition Motif*

RU – *Repeat Unit*

SNP – *Single Nucleotide Polymorphism*

UIM – *Ubiquitin Interacting Motif*

VNTR – *Variable Number of Tandem Repeat*

WSD – *White Spot Disease*

WSSV – *White Spot Syndrome Virus*

XPB – *Xeroderma Pigmentosum group D*

Sumário

1.INTRODUÇÃO	15
1.1 Doenças virais na carcinicultura.....	15
1.2 Vírus causador da Síndrome da Mancha Branca - WSSV.....	16
1.3 Variantes Virais	18
1.4 Caracterização <i>in silico</i> de sequências biológicas de função desconhecida.....	20
1.5 Caracterizações de sequências regulatórias	22
1.6 Justificativa.....	24
2 OBJETIVOS	25
2.1 Objetivo geral	25
2.2 Objetivos específicos	25
3 METODOLOGIA	26
3.1 Análise Filogenética.....	26
3.2 Alinhamento das VNTRs	26
3.3 Predição das estruturas secundárias e terciárias	26
3.4 Análise de motivos promotores nos genomas do WSSV	28
4 RESULTADOS E DISCUSSÃO	28
4.1 Análise filogenética e alinhamento das regiões variáveis do WSSV	28
4.2 - Caracterizações dos <i>clusters</i> wsv461/464 e wsv477/502	31
4.2.1 - ORF wsv463a	36
4.2.2 - ORF wsv477.....	39
4.2.3 - ORFs wsv479 e wsv497	42
4.2.4 - ORF wsv492.....	46
4.3 - Caracterização de ORFs contendo VNTRs.	48
4.3.1 - ORF wsv129.....	53
4.3.2 - ORF wsv178.....	55
4.3.3 - ORF wsv249.....	56
4.4 Predição de regiões regulatórias no genoma do WSSV.....	61
5 CONCLUSÃO	63
REFERÊNCIAS	64
APÊNDICES	71
APÊNDICE A – Motivos encontrados 200nt upstream das ORFs do WSSV	71
APÊNDICE B– Motivos encontrados 200nt upstream das ORFs do WSSV	73
APÊNDICE C – Motivos encontrados 200nt upstream das ORFs do WSSV	74
APÊNDICE D – Motivos encontrados 200nt upstream das ORFs do WSSV	75
APÊNDICE E – Motivos encontrados 200nt upstream das ORFs do WSSV.....	77
APÊNDICE F – Manuscrito submetido BMC Bioinformatics	80

1.INTRODUÇÃO

1.1 Doenças virais na carcinicultura

A pesca e cultivo de camarões têm contribuído significativamente para economia mundial. Grande parte da criação e comércio de camarão concentra-se em países asiáticos como Taiwan, Tailândia, China, Índia, Vietnã e Indonésia, sendo que países da América Latina como Equador, Brasil, Honduras e México (LIGHTNER, 2011) e do Oriente Médio (Iran e Arábia Saudita) vem ganhando destaque na carcinicultura nos últimos anos (FLEGEL, 2012; STENTIFORD et al., 2012).

Na aquicultura, os camarões marinhos dominam a produção mundial de crustáceos, sendo o *Penaeus vannamei* a espécie mais produzida, compondo cerca de 50% da produção anual de crustáceos em 2016 (FAO, 2018). Contudo, grande parte das espécies de crustáceos selvagens e cultiváveis é hospedeira para patógenos virais, os quais são responsáveis por cerca de 60% das doenças infecciosas que comprometem a sobrevivência e o desenvolvimento dos camarões em cultivo (FLEGEL et al., 2008; LIGHTNER, 2011). A dispersão e estabelecimento dos patógenos virais ao redor do globo se devem principalmente ao manejo precário durante o cultivo e o transporte de *commodities*, sejam espécimes vivos ou congelados (KARUNASAGAR; ABABOUC, 2012; TANG et al., 2012).

A síndrome da mancha branca (WSD), causada pelo *White Spot Syndrome Virus* (WSSV) é uma das principais enfermidades que acometem a carcinicultura mundial e consiste no principal problema enfrentado pela carcinicultura brasileira atualmente (Figura 1). Os primeiros relatos de surtos da síndrome mancha branca foram em 1992 no sudoeste asiático e desde então o vírus tem se dispersado por diversas fazendas de camarão localizadas no continente asiático (CHOU HSIN-YIU et al., 1995). No ano de 1995 foram relatados casos de mancha branca no sul dos Estados Unidos (WANG; NUNAN; LIGHTNER, 2000).

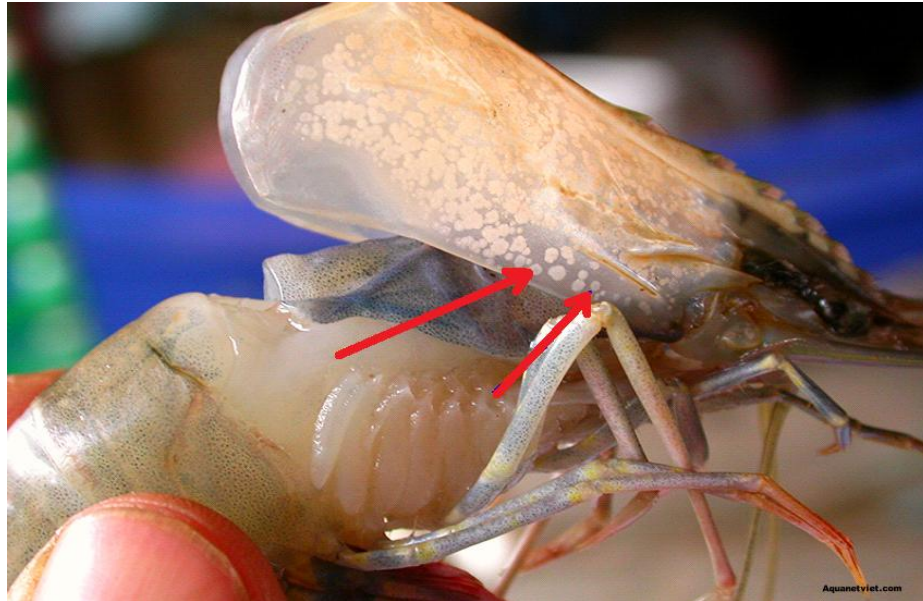


Figura 1 - Camarão apresentando os principais sintomas da síndrome da mancha branca. As setas vermelhas apontam as manchas brancas calcificadas em seu exoesqueleto. Fonte: disponível em: <http://aquanetviet.com/post/133900619760/ph%C3%A1t-tri%E1%BB%83n-c%C3%B4ng-c%E1%BB%A5-duplex-gi%C3%BAp-pcr-ph%C3%A1t-hi%E1%BB%87n-th%E1%BB%9Dj>. Acessado em 13/12/2018.

O primeiro relato da mancha branca no Brasil ocorreu em 2005 no estado de Santa Catarina, seguido dos estados do Ceará e Bahia (MULLER et al., 2010) e desde então tem gerado grande impacto econômico para a indústria da carcinicultura. Desde seu surgimento, o prejuízo causado pela mancha branca é estimado entre \$8-\$15 bilhões e vem aumentando a cada ano (FLEGEL et al., 2008; VERBRUGGEN et al., 2016).

O estado do Rio grande do Norte, atualmente o maior produtor de camarão no Brasil, vem sofrendo prejuízos severos decorrentes da mancha branca. As primeiras notificações da mancha branca no estado foram em meados de 2011, sendo que entre os anos de 2015 e 2016, houve um decréscimo de 17,80% de sua produção em toneladas (FILHO, 2017).

1.2 Vírus causador da Síndrome da Mancha Branca - WSSV

O WSSV é um vírus envelopado de DNA dupla-fita circular, que infecta diversas espécies de camarões cultiváveis, podendo causar a mortalidade de 90-100% da população de um viveiro entre 3-10 dias de infecção (VAN HULTEN et al., 2001; YANG et al., 2001). Um dos principais sintomas clínicos reportados em

camarões infectados pelo WSSV em diferentes regiões do mundo é o surgimento de pequenas manchas brancas calcificadas de 0.5-2.0 mm de diâmetro no cefalotórax do animal infectado. No Brasil o sintoma mais comum é o avermelhamento do cefalotórax (cor de telha) associada a expansão dos cromatóforos cuticulares. Os sintomas também incluem letargia, redução do consumo de alimento e de atividades motoras, perda da cutícula e descoloração do hepatopâncreas culminando na morte do animal (CHOU HSIN-YIU et al., 1995; PRADEEP et al., 2012; VAN HULTEN et al., 2000).

Originalmente classificado na família *Baculoviridae*, o WSSV foi reclassificado pelo Comitê Internacional em Taxonomia de Vírus (*International Committee on Taxonomy of Viruses* - ICTV) como único representante da família *Nimaviridae* e gênero *Whispovirus* (VERBRUGGEN et al., 2016). O vírion do WSSV possui uma estrutura muito particular que o diferencia dos demais vírus caracterizados até então (Figura 2), que consiste de uma partícula baciliforme de aproximadamente 70-167 x 210-420 nm, composta por um nucleocapsídeo revestido de um envelope trilaminar, seguido de uma estrutura semelhante a um flagelo na sua região terminal que possui função até então desconhecida (SÁNCHEZ-PAZ, 2010).

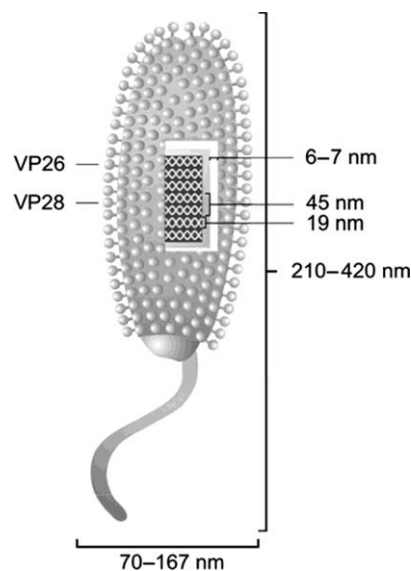


Figura 2 – Representação esquemática da partícula viral do WSSV com suas respectivas dimensões.
Fonte: Sanchez-Paz, 2010.

O WSSV possui um genoma de aproximadamente 300kbp, sendo um dos maiores vírus que infectam animais já descritos, contendo dentre 180-185 possíveis janelas de leitura aberta (*Open Reading Frames* - ORFs) (VAN HULTEN et al., 2001). Até o momento, pouco se sabe sobre a função da maioria das ORFs putativas presentes no genoma, sobretudo devido à impossibilidade de cultivar o vírus em linhagens celulares, o que dificulta a clonagem e avaliação funcional de proteínas e regiões regulatórias. Além disso, a maioria dessas proteínas preditas não apresenta homologia com as sequências conhecidas nos repositórios. (LI et al., 2017; OAKEY; SMITH, 2018; VAN HULTEN et al., 2001; YANG et al., 2001). Das poucas sequências caracterizadas, sabe-se que as proteínas codificantes atuam no processo de replicação viral, metabolismo de nucleotídeo e modulação da resposta imune do hospedeiro (CHEN et al., 2002; LIU et al., 2001; TZENG et al., 2002; VAN HULTEN et al., 2001; WANG et al., 2005).

1.3 Variantes Virais

A comparação de sequências de diferentes isolados do WSSV revelou que vírus de diferentes regiões são morfologicamente e geneticamente similares, apresentando identidade entre 92-99% em seu genoma (MULLER et al., 2010; PRADEEP et al., 2008; SHEKAR; PRADEEP; KARUNASAGAR, 2012). As diferenças que ocorrem entre os isolados do WSSV estão associadas a algumas regiões variáveis que existem no genoma. Essas regiões têm sido utilizadas para determinação de variantes virais, bem como para o estudo de sua origem evolutiva e dispersão. (DIEU, 2004; GUDKOV; MURWANTOKO; WALKER, 2014; RODRIGUEZ ANAY et al., 2016).

As principais regiões variáveis no genoma do WSSV consistem em inserções que ocorrem entre as ORFs wsv461/464 e wsv477/502 (também denominadas ORF14/15 e ORF23/24 - Tabela 1), em três ORFs que possuem repetições *in tandem* de número variável (*variable number of tandem repeats* - VNTRs) denominadas wsv129, wsv178 e wsv249 (também denominadas ORF 75, 94 e 125 – Tabela 2) e em nove regiões homólogas não codificantes (DIEU et al., 2010; GALAVÍZ-SILVA et al., 2004).

Tabela 1 – Dados das sequências codificantes putativas encontradas entre os *indels* da wsv461/464 (ORF14/15) e wsv477/502 (ORF23/24) obtidos a partir do genoma do WSSV-CN01 (*Genbank Accession*: KT995472. 1).

ORFs	Comprimento de Nucleotídeos (pb)	Janela de leitura	Comprimento de aminoácidos	Características conhecidas na literatura (Marks et al., 2005; Yang et al., 2001)
wsv461	897	-1	298	-
wsv463a	1530	-1	509	Região transmembranar
wsv463b	876	-1	291	-
wsv463c	915	-2	304	-
wsv463d	1338	-3	445	Região transmembranar (2x)
wsv464	867	-3	288	-
wsv477	627	1	208	Sítio de ligação de ATP, Zn-finger
wsv479	1530	-1	509	VP9, DUF1335
wsv482	540	1	179	Região transmembranar
wsv484	651	-3	216	Região transmembranar (4x)
wsv486	2340	1	779	Possível elemento transponível
wsv489	738	-2	245	-
wsv490	339	-2	112	-
wsv492	411	3	136	-
wsv493	687	-2	228	-
wsv495	1317	3	438	Domínio Colágeno
wsv497	1698	-2	565	VP9, DUF1335
wsv500	630	-2	209	Sítio de ligação de ATP, Zn-finger
wsv502	3030	2	1009	Sinal, Sítio de ligação de ATP, Zn-finger,IBR

As VNTRs das ORFs wsv129, wsv178 e wsv249 possuem unidades de repetições (RUs) de 45bp, 54bp, 69bp respectivamente, sendo que a wsv129 apresenta unidades de 57pb intercaladas. Além da variação no número de RUs, essas regiões VNTR apresentam polimorfismos de nucleotídeo único (SNPs) que podem ser utilizados como padrões para identificação (SHEKAR; PRADEEP; KARUNASAGAR, 2012).

Tabela 2 – Sequências de repetições *in tandem* presentes nas wsv129 (ORF75), wsv178 (ORF94) e wsv249 (ORF125) obtidas a partir do genoma do WSSV-CN01 (Genbank Accession: (KT995472. 1).

ORFs	Comprimento de Nucleotídeos (pb)	Comprimento de repetições	Janela de Leitura	Comprimento de Aminoácidos	Características conhecidas na literatura (Yang et al., 2001)
wsv129	1515	45bp, 57bp	+2	504	Região de repetição
wsv178	909	54bp	+2	302	Região de repetição
wsv249	2283	69bp	+1	760	E3 Ubiquitina-ligase

Alguns estudos já indicaram que essas regiões variáveis podem ter relação com o fenótipo da doença (GAO et al., 2014; HOA et al., 2005; LI et al., 2016). No entanto, uma correlação direta entre função e fenótipo de infecção ainda não foi possível. Até o momento existem poucas evidências de como essas proteínas atuam funcionalmente. Através de abordagens alternativas como a caracterização *in silico* é possível obter novas informações a respeito destas sequências.

1.4 Caracterização *in silico* de sequências biológicas de função desconhecida

Ao longo dos anos, as tecnologias de sequenciamento de DNA têm evoluído em sua eficácia, permitindo o sequenciamento de genomas em um menor espaço de tempo e a um menor preço, e consequentemente aumentando a quantidade de dados a serem depositados nos repositórios públicos de maneira exponencial.

Contudo, grande parte desses dados depositados correspondentes a sequências codificantes e regulatórias não são caracterizados na mesma velocidade, sobretudo em decorrência de limitações nas etapas experimentais (MULDER et al., 2008). A título de exemplo, o banco de dados UniprotKB (<http://www.uniprot.org/>) possui atualmente depositadas 558.898 sequências de proteínas anotadas manualmente e 137.213.158 anotadas automaticamente e não revisadas. Grande parte dessas sequências possui função desconhecida, sendo na maioria das vezes identificadas como proteínas hipotéticas.

O aumento nesse volume de dados tem favorecido cada vez mais a caracterização *in silico*, que se apresenta como uma alternativa para predição estrutural e funcional de moléculas biológicas. Em muitos casos a abordagem *in silico* fornece dados que complementam as informações adquiridas em abordagens experimentais, e em algumas situações possibilita a geração de informações que

não poderiam ser obtidas por abordagens *in vitro* (KHOR et al., 2015; PETREY; HONIG, 2005)

No campo da bioinformática aplicada ao estudo de proteínas, a caracterização funcional de sequências proteicas envolve o uso de ferramentas de busca por similaridade entre sequências (KHOR et al., 2015). A busca por similaridade segue o pressuposto de que sequências de proteínas possuem relações evolutivas que podem ser analisadas a partir da similaridade contida na sequência dos resíduos de aminoácidos, o que permite associar determinados padrões (ou sequências) a funções biológicas similares (YONA; LEVITT, 2002).

A modelagem comparativa consiste no uso de estruturas tridimensionais já elucidadas experimentalmente como molde (*template*) na construção do modelo estrutural da sequência-alvo. Essa abordagem é uma das alternativas mais empregadas para predição de estruturas de proteínas não caracterizadas experimentalmente, possibilitando, em alguns casos, a geração de um bom modelo com apenas 30% de identidade entre a sequência-alvo e o *template* (FLOUDAS, 2007). Para geração de modelos a partir da comparação de sequências com menos de 30% de identidade é recomendado o uso de abordagens de *fold recognition*, também conhecidas como *threading* (BUCHAN et al., 2013; LOBLEY; SADOWSKI; JONES, 2009). Abordagens de *threading* permitem comparar o *fold* (enovelamento) assumido por uma determinada sequência-alvo contra os bancos de dados estruturais, possibilitando posterior modelagem (FLOUDAS, 2007).

Nos casos em que não é possível a detecção de uma estrutura conservada, utiliza-se a modelagem *ab initio*, que dispensa o uso de *templates* conhecidos. Nessa abordagem a sequência proteica é modelada com base na natureza físico-química de seus resíduos de aminoácidos. Valores de similaridade muito inferiores a 30% compreendem uma faixa onde não é possível inferir homologia entre sequências biológicas conhecida como *Twilight Zone* (KHOR et al., 2015). A *Twilight Zone* impõe abordagens mais específicas para detecção de homologia remota, incluindo buscas iterativas através de modelos probabilísticos construídos a partir de perfis oriundos de alinhamentos múltiplos de sequência ou Modelos Ocultos de Markov (*Hidden Markov Model - HMM*) (ALTSCHUL et al., 1997; REMMERT et al., 2012; SÖDING, 2005). Em certos casos não é possível inferir homologia ou função baseando-se apenas nas sequências de aminoácidos.

A combinação de duas ou mais abordagens para obtenção modelo estrutural de uma proteína geralmente propiciam maior assertividade estatística. A título de exemplo, abordagens de modelagem combinadas já foram usadas para predição do regulador acessório estafilocócico (Msa) de *Staphylococcus aureus* (NAGARAJAN; ELASRI, 2007) e para caracterização do receptor de cobre dependente de TonB de *Acinetobacter baumannii* envolvido na ligação com a célula hospedeira via ligação à fibronectina (ABDOLLAHI; RASOOLI; MOUSAVI GARGARI, 2018).

Abordagens *in silico* também têm sido amplamente usadas na caracterização de proteínas virais. Um exemplo foi a modelagem comparativa e por *threading* da RNA polimerase dependente de RNA (RdRp) do “GB” vírus C (GBV-C) realizada com base nas coordenadas atômicas da RdRp do vírus da hepatite C (FERRON et al., 2005). Em outro trabalho, foi realizada a caracterização estrutural do VP5 do vírus da Doença da Bursite Infecciosa (IBDV) onde a análise funcional demonstrou similaridade com o receptor do tipo *Toll* do hospedeiro (GANGULY; RASTOGI, 2018). Em uma abordagem em que foi utilizada homologia remota, foi possível detectar que a suposta proteína de capsídeo do vírus do tomate do sudeste (STV Amalgavirus) possui similaridade com proteínas do nucleocapsídeo de tenuivírus (*Bunyaviridae*) embora a RNA polimerase dependente de RNA (RdRp) do amalgavirus tenha homologia com a RdRp partitivírus (*Partitiviridae*), sugerindo que os amalgavirus surgiram de um processo de recombinação entre o partitivírus e o tenuivírus (KRUPOVIC; DOLJA; KOONIN, 2015).

1.5 Caracterizações de sequências regulatórias

A expressão de um gene está estritamente associada a um conjunto de proteínas que se ligam ao DNA denominadas fatores de transcrição que podem atuar aumentando ou suprimindo o recrutamento da RNA polimerase. Os fatores de transcrição interagem com motivos de nucleotídeos que correspondem a pequenos segmentos genômicos de 5-20 pb que ocorrem com grande frequência, antecedendo as sequências codificantes, em regiões correspondentes a promotores ou em regiões mais distantes como *enhancers* ou silenciadores (DAS; DAI, 2007; PAVESI; MAURI; PESOLE, 2004), podendo ocorrer em ambas às fitas de DNA. A expressão ou supressão de um gene dependem diretamente da ligação destes fatores de transcrição aos motivos promotores e, em alguns casos, a ligação do fator de

transcrição pode ocorrer indiretamente, associando-se a outro fator de transcrição (BOEVA, 2016).

A caracterização de regiões regulatórias ou promotores é baseada na busca e identificação desses motivos, que apresentam padrões que se repetem dentro do genoma. Diversos algoritmos são usados para quantificar a ocorrência de padrões super-representados. Dentre as abordagens usadas para detecção de motivos regulatórios existe abordagens enumerativas que consistem em uma busca exaustiva que realiza a contagem de ocorrências de um padrão exato de nucleotídeos de um dado comprimento dentro de todo o genoma, contabilizando todos os padrões possíveis da sequência-alvo.

Apesar do grande alcance das abordagens enumerativas, podem ocorrer variações no padrão identificado, uma vez que o DNA está sujeito a mutações (inserções, deleções e substituições). Além disso, abordagens desse tipo podem retornar falsos positivos (D'HAESELEER, 2006). Outra abordagem muito utilizada que consegue contornar este problema consiste no uso de modelos probabilísticos que representam sequências motivo, por meio de uma matriz de peso de posição (*Position weight matrix* - PWM). Esta abordagem possibilita a cobertura de motivos que apresentam variações que normalmente não seriam detectadas através de métodos enumerativos. A representação gráfica deste modelo normalmente consiste em uma sequência de letras empilhadas em que os pesos são proporcionais ao conteúdo de informação da posição (CROOKS et al., 2004).

A presença de motivos regulatórios nos genomas virais é de significativa importância para modulação do seu ciclo de replicação. A regulação da expressão de genes virais normalmente é dividida em etapas temporais estritamente relacionadas ao perfil de seus motivos promotores. Geralmente a co-expressão dos genes se dá a partir de uma co-regulação transcricional, ou seja, genes que possuem mecanismos de regulação semelhantes, provavelmente possuem um conjunto de sequências motivos semelhantes (DAS; DAI, 2007). Geralmente o perfil de expressão gênica em vírus de DNA dupla-fita é classificado em imediatamente precoce, precoce e tardio (LI et al., 2009; LIU et al., 2015).

Genes de expressão imediatamente precoce consistem na primeira classe de genes que são expressos após a entrada do vírus na célula. Geralmente essas proteínas estão envolvidas na evasão do sistema de defesa do hospedeiro e modulação do ciclo celular. Os genes de expressão precoce estão envolvidos na

síntese de proteínas envolvidas no processo de replicação e metabolismo de nucleotídeos, enquanto os genes de expressão tardia codificam proteínas estruturais e para a montagem do vírion (LI et al., 2009; LIN et al., 2011, 2013).

Um dos maiores desafios encontrados dentro da caracterização *in silico* de sítios de ligação de fatores de transcrição é inferir de maneira significativa se um conjunto de sequências muito representados dentro do genoma possui relevância biológica ou se ocorre meramente ao acaso, e apesar de já existir sítios de ligação bem elucidados e depositados em repositórios (VLIEGHE, 2006), ainda existem muitos que ainda possuem função desconhecida. A busca por motivos putativos de sequências virais através abordagens *in silico* pode fornecer mais pistas sobre a dinâmica de expressão destes genes e dos possíveis fatores de transcrição que modulam estas etapas.

1.6 Justificativa

O WSSV tem sido considerado o vírus de maior importância econômica para carcinicultura mundial, e tem gerado um grande impacto negativo na produção nacional de camarão nos últimos anos. Até o momento pouco se sabe sobre o seu mecanismo de infecção e replicação, pelo fato de que grande parte de suas sequências codificantes não apresentaram homologia com as sequências conhecidas depositadas nos repositórios públicos. Além disso, apesar de existirem diferenças entre os fenótipos da infecção observados em campo, as possíveis correlações existentes entre as variações no genoma e os diferentes fenótipos em campo são escassas. Este trabalho teve como principal proposta esgotar os recursos de bioinformática disponíveis para caracterização de proteínas putativas codificadas pelas regiões variáveis do WSSV e de regiões regulatórias no genoma (promotores), no intuito de identificar evidências que auxiliem no entendimento das funções das proteínas virais e correlações entre genótipo e fenótipo.

2 OBJETIVOS

2.1 Objetivo geral

Caracterizar *in silico* as proteínas putativas codificadas pelas regiões variáveis do genoma do WSSV e possíveis sítios de ligação de fatores de transcrição que ocorrem *upstream* às sequências codificantes.

2.2 Objetivos específicos

- Analisar o efeito das regiões variáveis do genoma nas relações filogenéticas entre diferentes isolados do WSSV.
- Aprimorar a caracterização das regiões VNTR das ORFs wsv129, wsv178 e wsv249, com relação ao número de unidades de repetição e sítios polimórficos, considerando um número maior de sequências de diferentes isolados do WSSV.
- Aprimorar a caracterização das inserções existentes entre as ORFs wsv461/464 e wsv477/502, considerando um número maior de sequências de diferentes isolados do WSSV.
- Realizar predições estruturais e inferir possíveis funções para as proteínas putativas codificadas pelas ORFs wsv129, wsv178 e wsv249.
- Determinar regiões codificantes, realizar predições estruturais e inferir possíveis funções para algumas das proteínas putativas existentes nas inserções entre as ORFs wsv461/464 e wsv477/502.
- Predizer potenciais regiões regulatórias e motivos promotores que ocorrem no genoma do WSSV.

3 METODOLOGIA

3.1 Análise Filogenética

Para análise filogenética, foram gerados alinhamentos a partir de dez sequências genômicas de WSSV obtidas no *GenBank* (BENSON et al., 2005). A inferência evolutiva Bayesiana foi realizada através do software BEAST versão 2.4.3 (DRUMMOND et al., 2012) utilizando o modelo GTR com distribuição gama, relógio molecular estrito, cadeia de comprimento de dez milhões de gerações e consenso de árvores com eliminação de 10% das árvores iniciais. A ESS e a convergência estatística foram analisadas usando o programa Tracer 1.6. A visualização das árvores foi realizada pelo software FigTree.

3.2 Alinhamento das VNTRs

Sequências nucleotídicas disponíveis correspondentes às regiões variáveis dos diferentes isolados do WSSV foram obtidas no *GenBank* e submetidas a um alinhamento múltiplo através do MAFFT utilizando as configurações padrão (KATO; ROZEWICKI; YAMADA, 2017), exigindo também inspeção e edição manual no Geneious versão 11.0.3 (KEARSE et al., 2012). As unidades de repetição de cada variante foram alinhadas contra uma unidade de repetição de referência para uma anotação dos polimorfismos de nucleotídeo único (SNPs). Os sítios de polimorfismo foram visualizados por meio do WebLogo (CROOKS et al., 2004).

3.3 Predição das estruturas secundárias e terciárias

Para detecção de homólogos remotos, as buscas foram realizadas com auxílio das ferramentas do BLAST (BLASTp, PSIBLAST e PHIBLAST) contra bancos de dados não redundantes(nr), Swiss-Prot, refseq e PDB (ALTSCHUL et al., 1990, 1997) excluindo da busca sequências pertencentes ao WSSV (taxid:342409) na intenção de avaliar a existência sequências proteicas conservadas em outras espécies.

Buscas baseadas em perfis Ocultos de Markov também foram realizadas através de JACKHMMER, hmmscan, HHBlits e HHPred (EDDY, 1996, 2011;

JOHNSON; EDDY; PORTUGALY, 2010; REMMERT; BIEGERT; HAUSER, 2012; SÖDING; BIEGERT; LUPAS, 2005).

No hmmscan foram feitas buscas contra as coleções de perfis HMM de proteínas que incluem Pfam, *SuperFamily*, TIGRfam e CATH. As buscas no JACKHMMER foram conduzidas com até três iterações contra o banco de dados de sequências nr e Swiss-prot. Os filtros de busca foram desligados na intenção de tornar a busca ainda mais sensível em ambos os programas do HMMER.

As sequências de aminoácidos foram submetidas em uma busca iterativa usando HHBlits (2-4 iterações) contra banco de dados Uniclust30 (MIRDITA et al., 2017) versão de outubro de 2017. A partir da saída gerada, os melhores resultados foram selecionados e submetidos no HHPred contra o banco de dados do PDB70.

Os melhores hits do HHPred foram usados como *templates* para gerar modelos da estrutura terciária. As estruturas terciárias foram modeladas no PHYRE2, ITASSER, Swiss-Model e Modeller (BIASINI et al., 2014; ESWAR et al., 2006; KELLY et al., 2015; ZHANG, 2008). A predição de estruturas secundárias e reconhecimento de *fold* foi realizada através do *webserver* PSIPRED (BUCHAN et al., 2013; LOBLEY; SADOWSKI; JONES, 2009) para a busca de *templates* para modelagem.

Para estimar a qualidade dos modelos estruturais gerados, foram usados o Molprobit e o QMEAN (CHEN et al., 2010). As avaliações realizadas no Molprobit incluem avaliação de sobreposições estéricas (*Clashes*) e geometria da proteína (Comprimento e ângulos de ligações covalentes, rotâmeros e *Ramachandran*). Os *plots* de Ramachandran foram gerados pelo PyRAMA, uma implementação do Python 3 (LOVELL et al., 2003). Todas as imagens dos modelos foram geradas no UCSF Chimera (PETTERSEN et al., 2004).

As regiões de desordem, *Coiled coils*, topologia transmembranar e regiões contendo peptídeos sinais foram obtidas através dos softwares Foldindex, Coils, TMHMM e SignalP, (KROGH et al., 2001; PETERSEN et al., 2011; PRILUSKY et al., 2005), respectivamente. *SMART*, CDD, ScanProsite e *Eucariotic Linear Motif* (ELM) foram usados para detectar domínios, padrões e motivos conservados (DE CASTRO et al., 2006; DINKEL et al., 2012; LETUNIC; BORK, 2018; MARCHLER-BAUER et al., 2015).

3.4 Análise de motivos promotores nos genomas do WSSV

Para a caracterização dos motivos promotores dentro do genoma do WSSV, foi usado um *script* escrito na linguagem Python em conjunto com módulos do Biopython (COCK et al., 2009) com o qual é realizada a extração das regiões de 200nt upstream de cada ORF em ambas as fitas, sendo as sequências correspondentes a fita negativa convencionadas para o sentido positivo da fita.

Em seguida foi gerando um arquivo multifasta contendo todas as sequências *upstream* às ORFs preditas. Para evitar falsos motivos dentro da análise, as sequências onde se encontravam as regiões homólogas (hrs) do genoma do WSSV foram retiradas da análise. A busca dos motivos de comprimento entre 4-12nt foi empregada através do software MEME (*Multiple Em for Motif Elicitation*) (BAILEY; ELKAN, 1994). A ocorrência de um motivo específico por sequência individual foi definida em zero a um. Os motivos que ocorreram em maior frequência foram comparados contra motivos conhecidos em bancos de dados através do TOMTOM (GUPTA et al., 2007).

4 RESULTADOS E DISCUSSÃO

4.1 Análise filogenética e alinhamento das regiões variáveis do WSSV

A influência de regiões variáveis na reconstrução filogenética do WSSV foi avaliada comparando dendrogramas gerados a partir de genomas completos (Figura 3 A), de genomas que tiveram as regiões variáveis das ORFs wsv129, wsv178 e wsv249 e as regiões de inserção entre as ORFs wsv461/464 e wsv477/502 retiradas (Figura 3 B) e de genomas que tiveram as regiões homólogas retiradas (Figura 3 C). Para essa análise considerou-se como regiões variáveis as VNTRs nas ORFs wsv129, wsv178 e wsv249 e as regiões de inserção entre as ORFs wsv461/464 e wsv477/502.

A ausência das regiões variáveis resultou em uma mudança na distribuição de grupos asiáticos e no encurtamento do comprimento dos ramos no dendrograma, como mostrado na Figura 3 B. Essa mudança na topologia da árvore observada na Figura 3 B também foi superior à mudança observada no dendrograma calculado a partir de genomas sem as regiões homólogas (Figura 3 C). Os grupos que

apresentaram maiores diferenças na topologia da árvore por efeito da excisão das regiões variáveis foram aqueles que apresentam isolados com inserções entre as regiões da wsv461/464 e wsv477/502, indicando que essas inserções influenciam mais no agrupamento destes isolados.

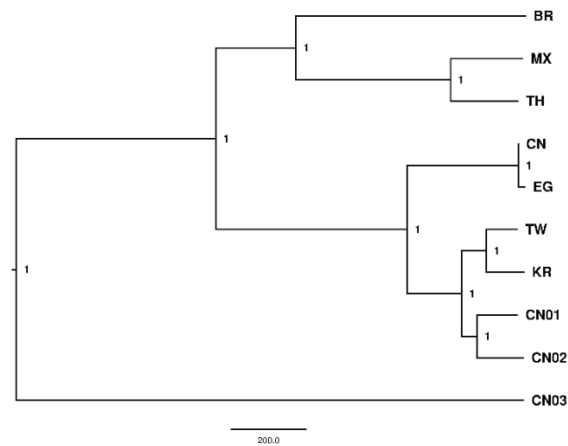
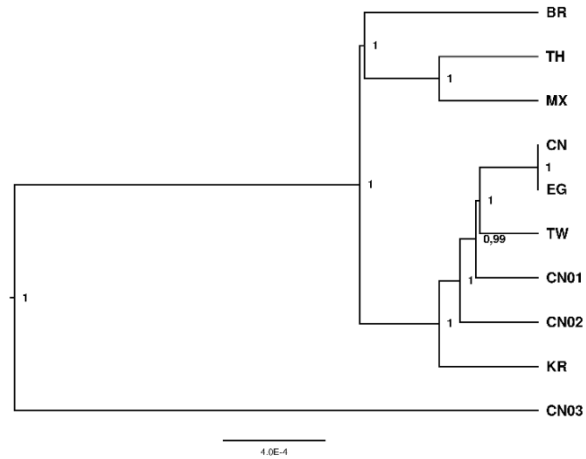
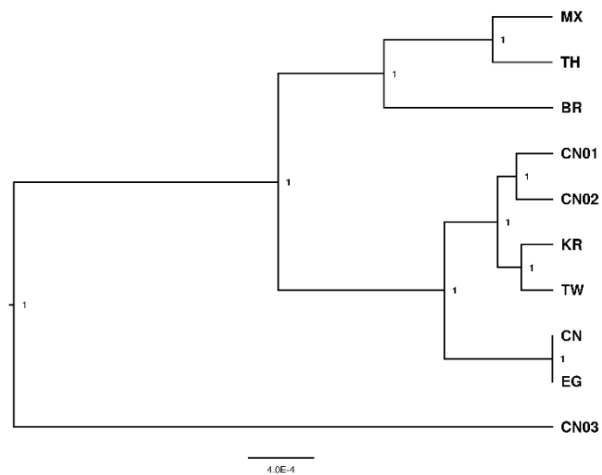
A**B****C**

Figura 3 - Influência das regiões variáveis na reconstrução filogenética do WSSV. Dendrogramas foram obtidos a partir do alinhamento de genomas completos do WSSV isolados de diferentes regiões (A), na ausência das regiões variáveis (B) e na ausência das regiões homólogas (C). Para o cálculo de cada dendrograma foi utilizada a abordagem bayesiana utilizando o modelo GTR com distribuição gama, relógio molecular estrito, Cadeia Markov com 10^7 gerações e consenso de árvores com exclusão de 10% das árvores iniciais.

A partir da análise da árvore contendo os genomas completos, foi possível observar o padrão de relações típico, semelhante ao observado em estudos anteriores (DANTAS et al., 2018; LI et al., 2016; OAKLEY; SMITH, 2018). Em todas as árvores, o WSSV-CN03 (KT995471.1) se agrupou distante das demais variantes. Resultados anteriores do sequenciamento dos genomas dos isolados WSSV-CN01 (KT995472.1), WSSV-CN02 (KT995470.1) e WSSV-CN03 (KT995471.1), mostram que as sequências correspondentes à wsv129 e wsv178 estão ausentes no WSSV-CN03.

Já foi descrito que no WSSV-CN03 ocorre a translocação da wsv486 para a região onde o wsv195 é normalmente encontrada em outros genomas, sugerindo que essa sequência possa atuar como um elemento transponível. Tanto o isolado WSSV-CN02 quanto do WSSV-CN03 possuem deleções entre as regiões correspondentes a wsv231/242, sendo essas deleções de 4Kb e 3.3Kb respectivamente (LI et al., 2016). A deleção de 2160pb existente entre as ORFs wsv332/340 no isolado WSSV-CN03 juntamente com as demais deleções e translocações já reportadas (LI et al., 2016), explica apenas parcialmente a maior distância desse isolado em relação aos outros isolados, uma vez que, conforme o dendrograma da Figura 1B, mesmo desconsiderando o efeito das regiões variáveis esse isolado se manteve distante. O papel dessas inserções/deleções nos mecanismos de adaptação viral ainda é obscuro, uma vez que elas não existem na maioria dos isolados do WSSV. De toda forma, a quantidade e a qualidade das sequências utilizadas para a construção dos dendrogramas apresentados aqui ainda não é suficiente para definição de um padrão evolutivo, uma vez que, além do número de sequências ser pequeno, as excisões encontradas podem ser artefato de sequenciamento.

De forma a aumentar os conhecimentos sobre a função dessas proteínas e sobre o seu papel na infecção viral foi realizada a caracterização das regiões variáveis do genoma do WSSV e a modelagem estrutural de algumas proteínas putativas existentes nessas regiões.

4.2 - Caracterizações dos *clusters* wsv461/464 e wsv477/502

A caracterização das sequências que ocorrem nas regiões flanqueadas pelas ORFs wsv461/464 e wsv477/502 (Figuras 4 e 5) revelaram inserções de até 5Kb e

13 Kb, respectivamente. O *cluster* wsv461/464 contém até 6 ORFs putativas (wsv461, wsv463a, wsv463b, wsv463c, wsv463d, wsv464) como detalhado na Tabela 3, entretanto, na maioria dos isolados este cluster é truncado e desprovido destas ORFs. As inserções dos isolados WSSV-CN02 e WSSV-TW possuem a wsv461 e a wsv463 fusionados em uma única ORF putativa (Figura 4).

Tabela 3 - Caracterização de sequências correspondentes ao cluster wsv461/464 (ORF14/15) obtidos a partir de diferentes isolados de WSSV.

Número de acesso	País/Descrição		Inserção	Comprimento da inserção (pb)	Nº de ORFs putativas	ORFs putativas que ocorrem no cluster
AF440570.1	Taiwan	1994	Sim	506	2	(wsv461 f, wsv464)
AF332093.3	China	1996	Sim	257	3	(wsv461*, wsv463 f, wsv464)
AF369029.2	Tailândia	1996	Não	-	2	(wsv461, wsv464*)
JX515788.1	Coréia do Sul	2011	Não	-	1	(wsv464)
KR083866.1	Egito	2014	Sim	257	3	(wsv461*, wsv463 f, wsv464)
KT995472.1	China 01	1994	Sim	4999	6	(wsv461, wsv463a, wsv463b, wsv463c, wsv463d, wsv464)
KT995470.1	China 02	2010	Sim	506	2	(wsv461 f, wsv464)
KT995471.1	China 03	2010	Não	-	2	(wsv461*, wsv464*)
KY827813.1	China 04	2012	Não	-	2	(wsv461*, wsv464*)
MF784752.1	Brasil	2015	Não	-	2	(wsv461, wsv464*)
KR083844.1	Egito 09	2014	Sim	4894	6	(wsv461, wsv463a, wsv463b, wsv463c, wsv463d ,wsv464)
KF771904.1	EUA Texas	1997	Sim	4894	6	(wsv461, wsv463a, wsv463b, wsv463c, wsv463d ,wsv464)
AY753327.1	Tailândia 96 II	1996	Sim	5025	7	(wsv461, wsv463a t, wsv463b, wsv463c, wsv463d ,wsv464)
KR083865.1	Egito 14	2014	Sim	137	2	(wsv461*, wsv464*)
KF771899.1	EUA DC97	1997	Sim	137	2	(wsv461*, wsv464*)
KR083846.1	Egito 11	2014	Sim	1136	2	(wsv461*, wsv464*)
KF771902.1	EUA Texas 2	2002	Sim	1136	2	(wsv461*, wsv464*)
KX501221.1	Tailândia 12	2012	Não	-	2	(wsv461*, wsv464*)
EU327501.1	Índia IN-05	2005	Não	-	2	(wsv461*, wsv464*)
MG702567.1	Índia AP4RU	2013	Não	-	2	(wsv461*, wsv464*)
HQ257380.1	México Mx-F	2004	Sim	335	2	(wsv461*, wsv464*)
HQ257382.1	México Mx-C	2005	Sim	335	2	(wsv461*, wsv464*)
HQ257383.1	México Mx-G	2004	Sim	335	2	(wsv461*, wsv464*)
MF768985.1	Australia	2016	Sim	171	2	(wsv461*, wsv464*)
KX501220.1	Tailândia TH-14	2012	Não	-	2	(wsv461*, wsv464*)
KU746817.1	China FCG2-14	2014	Não	-	2	(wsv461*, wsv464*)
EF468499.1	Índia IN-07	2005	Sim	335	2	(wsv461*, wsv464*)

* Sequências parciais, f sequências formadas a partir da fusão de duas sequências e t duas ORFs formadas a partir da wsv463a (ocorre exclusivamente na WSSV TH-96-II).

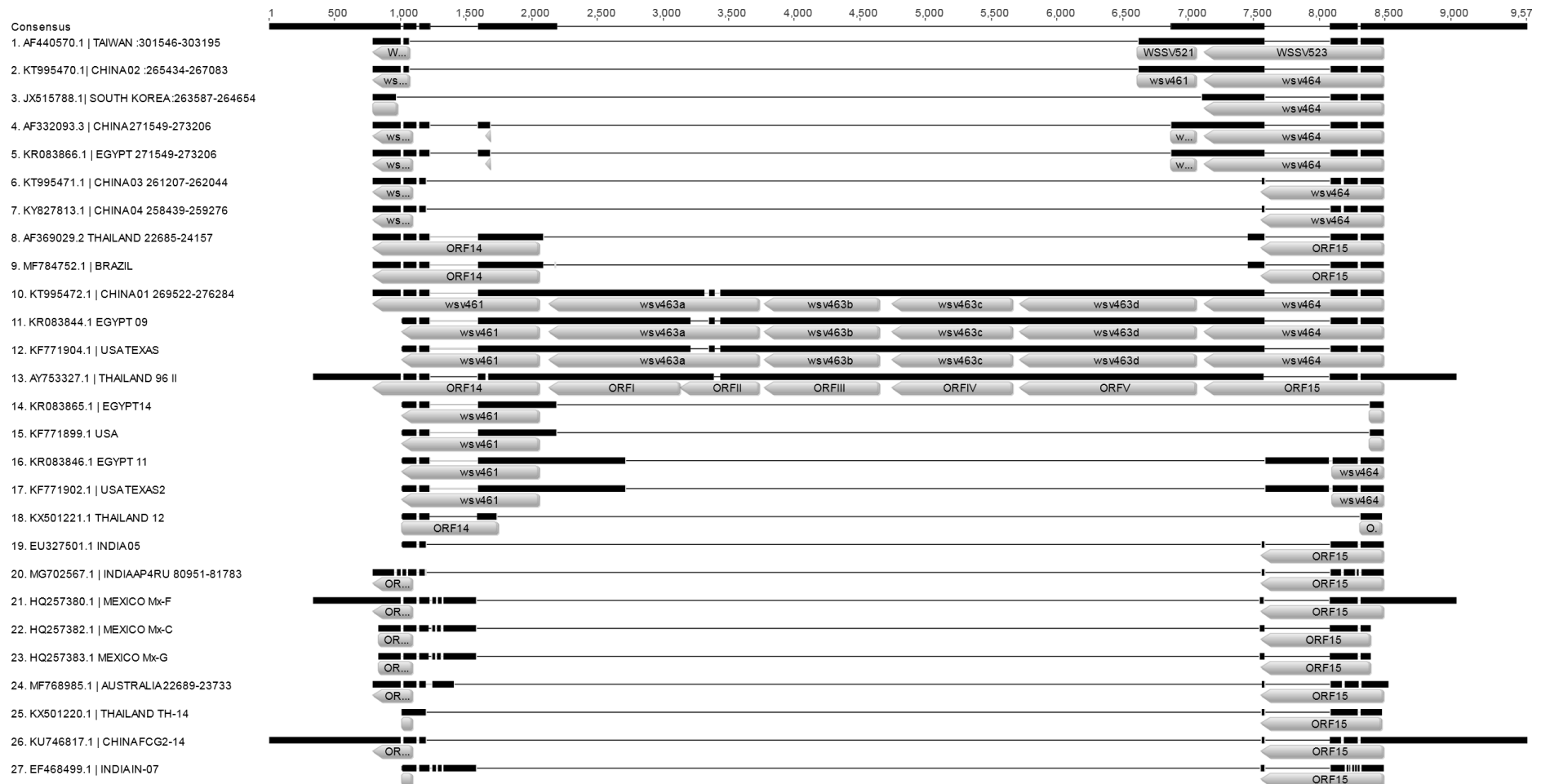


Figura 4 - Alinhamento das regiões correspondentes ao cluster das ORFs wsv461/464 (ORF14/15). As sequências presentes no alinhamento representam variantes virais que foram isoladas de diferentes regiões do mundo. O alinhamento e anotação das sequências codificantes putativas foi realizado com auxílio do programa Geneious versão 11.0.3. As setas em cinza correspondem as ORFs putativas presentes dentro das inserções.

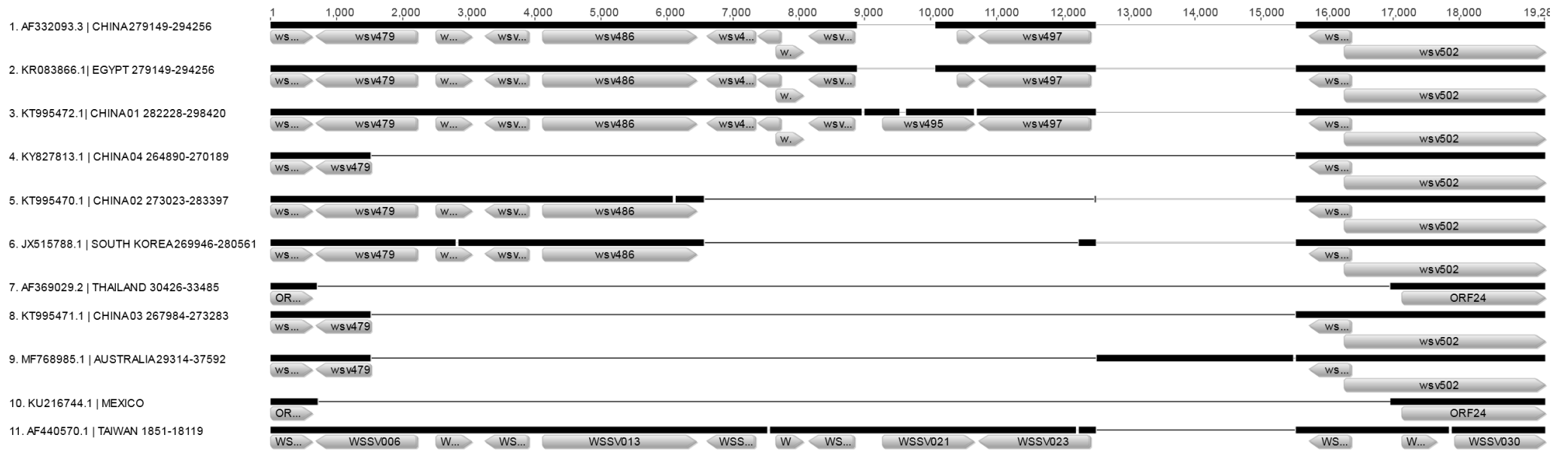


Figura 5 - Alinhamento das regiões correspondentes ao cluster das ORFs wsv477/502 (ORF23/24). As sequências presentes no alinhamento representam variantes virais que foram isoladas de diferentes regiões do mundo. O alinhamento e anotação das sequências codificantes putativas foi realizado com auxílio do programa Geneious versão 11.0.3. As setas em cinza correspondem as ORFs putativas presentes dentro das inserções.

O número de regiões codificantes putativas no cluster wsv477/502 é bem superior, compreendendo até 13 ORFs (wsv477, wsv479, wsv482, wsv484, wsv486, wsv489, wsv490, wsv492, wsv493, wsv495, wsv497, wsv500 e wsv502) como detalhado na Tabela 4. Curiosamente, para este *cluster*, o número de sequências com inserções truncadas é menor (Tabela 4 e Figura 5) quando comparado ao cluster wsv461/464.

Tabela 4 - Caracterização das sequências correspondentes ao cluster das ORF wsv477/wsv502 (ORF23/24).

Número de acesso	País/Descrição	Ano	Inserção	Comprimento da inserção (pb)	Nº de ORFs putativas	ORFs putativas
AF440570.1	Taiwan	1994	Sim	12613	13	(wsv477, wsv479, wsv482, wsv484, wsv486, wsv489, wsv490, wsv492, wsv493, wsv495, wsv497, wsv500, wsv502)
AF332093.3	China	1996	Sim	11451	13	(wsv477, wsv479, wsv482, wsv484, wsv486, wsv489, wsv490, wsv492, wsv493, wsv495*, wsv497, wsv500, wsv502)
AF369029.2	Tailândia	1996	Sim	81	2	(wsv477, wsv502)
JX515788.1	Coréia do Sul	2011	Sim	6959	8	(wsv477, wsv479, wsv482, wsv484, wsv486, wsv500, wsv502)
KR083866.1	Egito	2014	Sim	11451	13	(wsv477, wsv479, wsv482, wsv484, wsv486, wsv489, wsv490, wsv492, wsv493, wsv495*, wsv497, wsv500, wsv502)
KT995472.1	China 01	1994	Sim	12536	13	(wsv477, wsv479, wsv482, wsv484, wsv486, wsv489, wsv490, wsv492, wsv493, wsv495*, wsv497, wsv500, wsv502)
KT995470.1	China 02	2010	Sim	6718	8	(wsv477, wsv479, wsv482, wsv484, wsv486, wsv500, wsv502)
KT995471.1	China 03	2010	Sim	1643	4	(wsv477, wsv479*, wsv500, wsv502)
KU216744.1	México	2008	Sim	98	2	(wsv477, wsv502)
KY827813.1	China 04	2012	Sim	1643	4	(wsv477, wsv479*, wsv500, wsv502)
MF768985.1	Austrália	2016	Sim		4	(wsv477, wsv479*, wsv500, wsv502)

* Sequências parciais.

Os resultados apresentados nesse estudo estão de acordo com informações já publicadas, e com as caracterizações preliminares que já existiam dos dois clusters analisados. Esses estudos anteriores já indicavam que possivelmente

algumas proteínas codificadas por estas regiões possuem domínios funcionais (LIN et al., 2011; YANG et al., 2001). Entretanto, essa foi a primeira abordagem em que todos os genomas já sequenciados foram avaliados simultaneamente. Foi interessante observar que, nas sequências mais recentes, o número de inserções é inferior ou essas inserções são menores, o que condiz a proposição de que o genoma do WSSV tem diminuído em tamanho com o tempo (LI et al., 2017; OAKLEY; SMITH, 2018).

Em um ensaio com as variantes WSSV-TH, que é desprovida destas inserções, e WSSV-TH-96-II que apresenta estas inserções, foi observado que a WSSV-TH foi mais virulenta que a WSSV-TH-96-II (MARKS et al., 2005a). Um dos poucos estudos do perfil de expressão de alguns desses genes *in vivo* sugerem que estes *indels* atuam em uma via funcional mas não essencial para estes genes durante a infecção (MARKS et al., 2005b) e que um genoma menor poderia favorecer a replicação viral.

Li et al, por outro lado, sequenciou os genomas das linhagens de WSSV-CN01 de virulência alta, WSSV-CN02 de virulência intermediária e WSSV-CN03 de virulência baixa, demonstrando que o isolado WSSV-CN01 possui o maior genoma em relação aos outros dois isolados, sugerindo que a presença destas regiões variáveis pode estar relacionada a codificação de fatores de virulência. Nesse contexto o tamanho do genoma não teria associação com a virulência. Nesse caso os isolados menos virulentos poderiam ter perdido fatores regulatórios ou a própria região codificante com perda de função (LI et al., 2016).

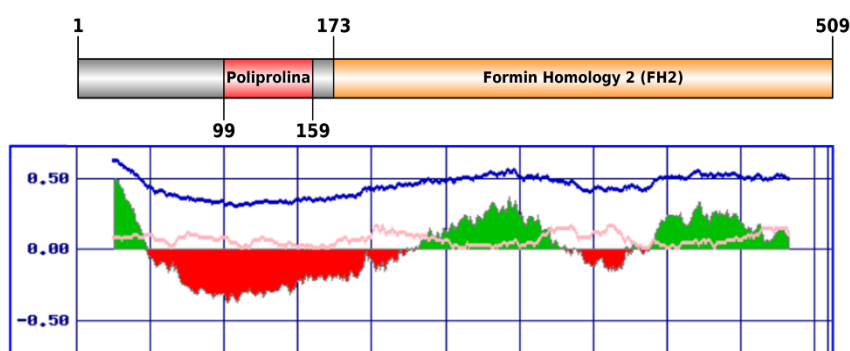
A relação destas inserções com o *fitness* do WSSV ainda é controversa, pois não se sabe o que grande parte das ORFs putativas presentes nestas inserções codificam. Os resultados da caracterização de cada região codificante estão apresentados e discutidos a seguir. Não foi possível obter o modelo tridimensional de todas as sequências. Apenas os resultados com alta confiança estatística estão apresentados, baseados na validação dos modelos/predição de proteínas propostos.

4.2.1 - ORF wsv463a

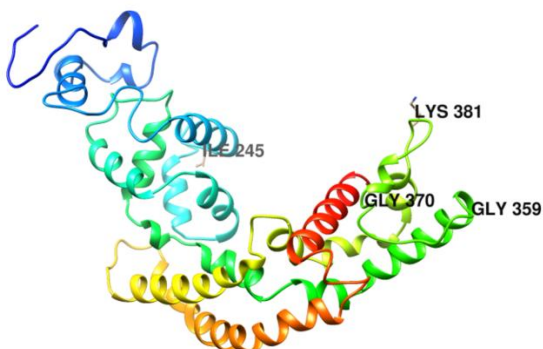
A ORF wsv463a, uma sequência codificante putativa localizada dentro do cluster wsv461/464, apresenta um domínio rico em prolina localizado na porção desordenada da proteína predita (99-159) e um grande domínio similar ao domínio

Formin Homology 2 (FH2) entre as posições 173-509 (Figura 6 A). O modelo tridimensional do domínio estruturado maior, obtido a partir da estrutura PDB: 3OBV, confirma uma estrutura tridimensional composta de alfa-hélices similar ao domínio FH2, que são divididas em cinco subdomínios (*Lasso*, *Linker*, *Knob*, *Coiled coil*, *Post*) demonstrados na Figura 6 B. Os testes de validação, escores e *plots* de Ramachandran correspondentes ao modelo tridimensional são apresentados abaixo do modelo (Figura 6 C, D, E).

A



B



C

WSV463a Formin Homology 2 (FH2)			
Clashscore, todos os átomos:	103.57	0 th percentil [^] (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	7	2.40%	Ideal: <0.3%
Rotâmeros favoráveis	266	91.10%	Ideal: >98%
Ramachandran aberrantes	5	1.48%	Ideal: <0.05%
Ramachandran favoráveis	309	91.69%	Ideal: >98%
Escore do MolProbity [^]	3.26	14 th percentil [^] (N=27675, 0Å - 99Å)	
Desvios do Cβ >0.25Å	3	0.96%	Ideal: 0
Ligações ruins:	0 / 2663	0.00%	Ideal: 0%
Ângulos ruins:	49 / 3594	1.36%	Ideal: <0.1%
Prolinas Cis:	0 / 11	0.00%	Esperado: ≤1 por cadeia, ou ≤5%

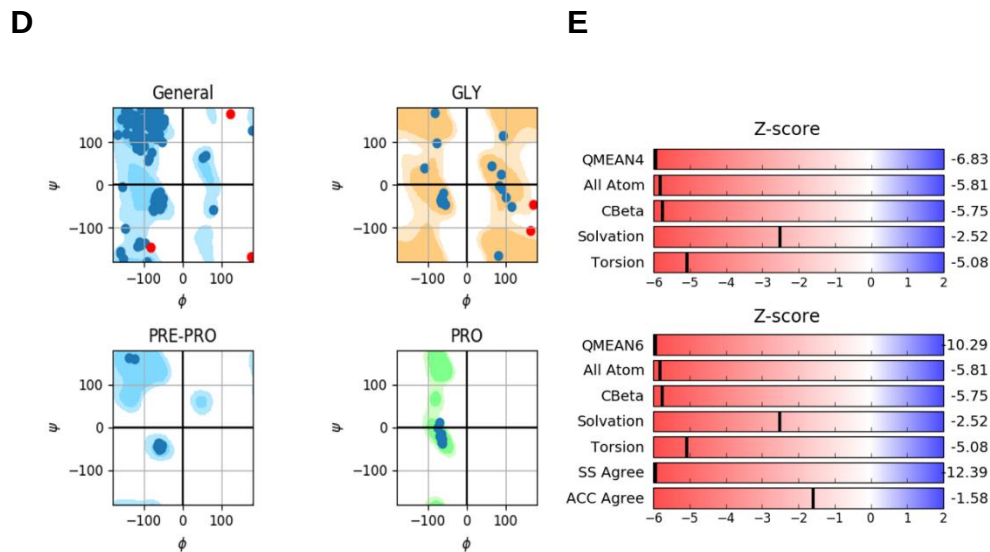


Figura 6 - Caracterização estrutural da proteína putativa codificada pela wsv463a. (A) Representação esquemática da sequência proteica codificada pela wsv463a e seus respectivos domínios conservados. O domínio FH1, composto por resíduos de poliprolina está representado em vermelho, enquanto o domínio FH2 está representado em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv463a correspondente ao domínio *Formin Homology 2* (FH2). Subdomínios conservados (*Lasso*, *Linker*, *Knob*, *Coiled Coil* e *Post*) e aminoácidos do domínio FH2 típico são indicados. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) *Plots* de Ramachandram gerados no PyRAMA. (E) Escores globais gerados no QMEAN.

Forminas consistem de uma família de proteínas que regula o alongamento dos filamentos de actina que são importantes e muitos processos celulares incluindo formação de filamentos de actina, anel citocinético, filopodia e fibras de estresse (ASPENSTRÖM, 2010; OTOMO et al., 2005). Estes processos são mediados pelo domínio FH2 localizado na região C-terminal, que se dimeriza com outro domínio FH2 de outra formina, formando um anel hidrofóbico no que se liga a extremidade do filamento de actina, os protegendo das proteínas de capeamento (OTOMO et al., 2010; XU et al., 2004).

Cada uma das proteínas que compõem o dímero possui resíduos conservados que estão diretamente relacionados a ligação com a actina que consistem da Isoleucina que está localizada no subdomínio *Knob* e da Lisina na interface *Lasso/Post*. Mutações que ocorrem dentro destes resíduos podem comprometer a atividade de nucleação da actina (SCHÖNICHEN; GEYER, 2010). Os resíduos de glicina nas posições 359 e 370 da proteína codificada pela wsv463a

atuam na dimerização, assim como a Ile245 e a Lys381 que atuam na ligação com a actina permanecem conservados (Figura 6 B).

Forminas também possuem um domínio FH1 composto por resíduos de poliprolina similar ao observado no esquema da proteína da wsv463a (Figura 6 A). O domínio FH1 está diretamente relacionado com a interação das proteínas profilinas durante o processo de alongamento da actina. As profilinas, que estão ligadas aos monômeros de actina estimulam a montagem dos monômeros de actina através da interação com o domínio FH1 e FH2, aumentando a velocidade de alongamento da actina (SCHÖNICHEN; GEYER, 2010; THOMPSON et al., 2013).

Como componente essencial do citoesqueleto, a actina pode ser utilizada pelo vírus dentro de células hospedeiras em diversos estágios do seu ciclo de vida, incluindo entrada replicação e montagem do vírion por meio da modulação das proteínas ligantes a actina (SPEAR; WU, 2014). Filamentos de actina fornecem força mecânica para o transporte intracelular dos vírus, resultando em mudanças nos padrões na célula (NEWSOME; MARZOOK, 2015). O vírus da imunodeficiência humana (*Human Immunodeficiency Virus* - HIV) pode ser transportado no interior de células dendríticas através da filopodia produzida pelas forminhas (AGGARWAL et al., 2012). A formina FHOD1 em conjunto com a GTPase Rac1 do *Vaccinia* vírus estão associados a formação da cauda de actina, atuando de maneira integrada com a via da N-WASP-APR2/3, assim sendo essencial para motilidade e dispersão do *Vaccinia* vírus (ALVAREZ; AGAISSE, 2013). O fato do WSSV codificar forminas pode estar relacionado a regulação das proteínas fibrosas dos hospedeiros envolvidos no empacotamento viral e/ou transporte intracelular.

4.2.2 - ORF wsv477

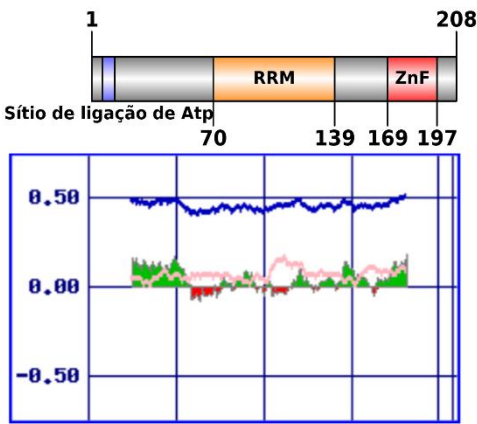
A ORF wsv477 é a primeira região codificante no cluster wsv477/502 (Figura 5 A). O produto putativo da wsv477 apresenta um domínio homólogo a um motivo de reconhecimento de RNA (*RNA Recognition Motif* - RRM) entre as posições 70-150 e um domínio *Zinc-Finger* em sua região C-terminal (Figura 7 A).

O modelo 3D do domínio RRM, modelado a partir da estrutura de uma pequena ribonucleoproteína (U1 snRNP, PDB: 4PKD), revelou uma estrutura terciária com duas alfa-hélices e três folhas beta dispostas de forma anti-paralela seguindo um padrão $\beta 1\alpha 2\beta 2\beta 3\alpha 2$, o que corresponde à típica estrutura

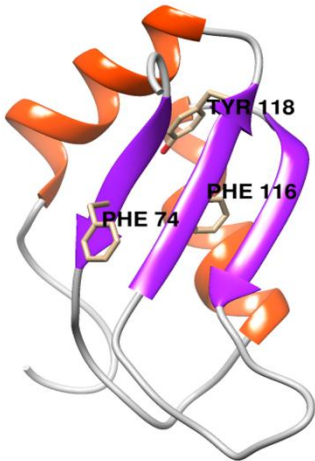
$\beta 1\alpha 2\beta 2\beta 3\alpha 2\beta 4$ observada no motivo RRM (Figura 7 B). Os resíduos aromáticos envolvidos na ligação ao RNA permanecem conservados nas cadeias centrais da proteína proposta, que são compostas por Phe74 na posição 2 do $\beta 1$, Tyr118 na posição 5 de $\beta 3$ e Phe116 na posição 3 de $\beta 3$ (Figura 7 B).

A RRM pode ser encontrada em todos os organismos, com maior abundância em proteínas codificadas por eucariotos em múltiplas cópias ou em conjunto com outros domínios como os domínios *Zinc-Finger* dos tipos CCCH ou CCHC, que podem se ligar ao RNA. O RRM está envolvido em eventos pós-transcricionais, incluindo *splicing*, processamento de pré-mRNA, edição de RNA, regulação de tradução e degradação de RNA (CONTE, 2000; KONDO et al., 2015; MARIS; DOMINGUEZ; ALLAIN, 2005; MARTIN-TUMASZ et al., 2011; WEBER et al., 2010).

A



B



C

WSV477(ORF23) Motivo de Reconhecimento de RNA (RRM)			
Clashscore, todos os átomos:	94.83	0 th percentil ⁺ (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	3	4.69%	Ideal: <0.3%
Rotâmeros favoráveis	56	87.50%	Ideal: >98%
Ramachandran aberrantes	0	0.00%	Ideal: <0.05%
Ramachandran favoráveis	67	97.10%	Ideal: >98%
Escore do MolProbity [^]	3.11	19 th percentil ⁺ (N=27675, 0Å - 99Å)	
desvios do Cβ >0.25Å	0	0.00%	Ideal: 0
Ligações ruins:	0 / 606	0.00%	Ideal: 0%
Ângulos ruins:	11 / 819	1.34%	Ideal: <0.1%
Prolinas Cis:	0 / 1	0.00%	Esperado: ≤1 por cadeia, ou ≤5%

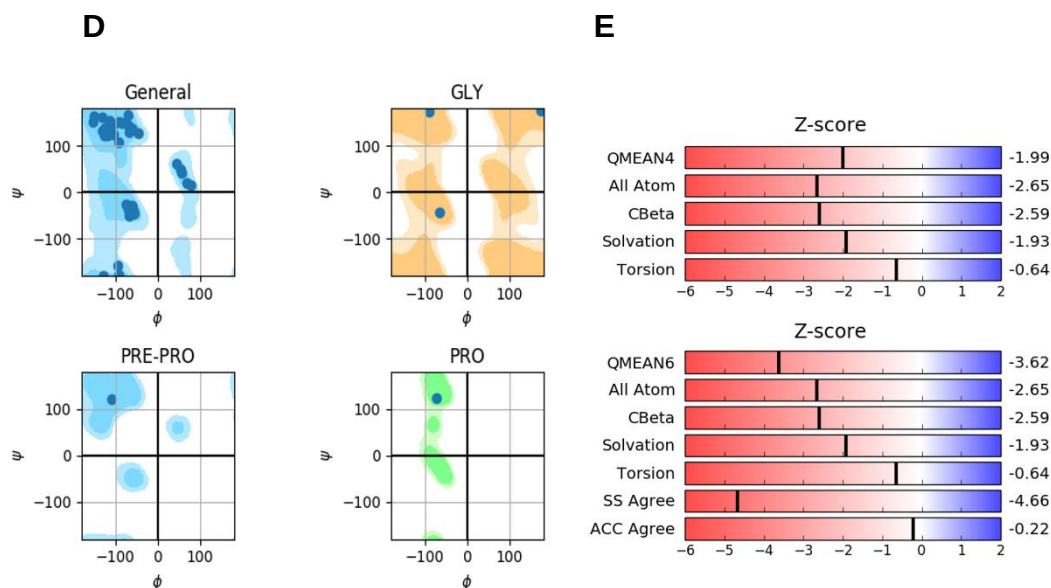


Figura 7 - Caracterização estrutural da proteína putativa codificada pela wsv477. (A) Representação esquemática da sequência proteica codificada pela wsv477 e seus respectivos domínios conservados. O motivo de reconhecimento de RNA está representado em laranja, um sítio de ligação de ATP na região N-terminal em azul, enquanto o domínio *Zinc-Finger* do tipo Cys2/Cys2 está representado em vermelho. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv477 correspondente ao motivo de reconhecimento de RNA (RRM). (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) Plots de Ramachandram gerados no PyRAMA. (E) Escores globais gerados no QMEAN.

A ORF wsv477 foi previamente caracterizada como um gene imediatamente precoce de 624 pb codificando uma proteína putativa de 208 aminoácidos, com um sítio de ligação ATP/GTP entre as posições 7-14 e um domínio de *Zinc-finger* tipo Cys2/Cys2 entre os resíduos 169-197 (HAN; XU; ZHANG, 2007). Além disso, foi observado que a injeção de miR-7 em camarões infectados com WSSV poderia reduzir a expressão do wsv477 e diminuir o número de cópias do genoma do WSSV entre 12 e 96 horas após a infecção (HUANG; XU; ZHANG, 2012; HUANG; ZHANG, 2012). Os resultados aqui apresentados corroboram que o domínio RRM na proteína wsv477 pode estar relacionado a etapas pós-transcricionais que determinam a eficiência da replicação viral.

Algumas proteínas RRM estão envolvidas na replicação de vírus de RNA, incluindo a ribonucleoproteína nuclear heterogênea A1 (hnRNP A1) no vírus da

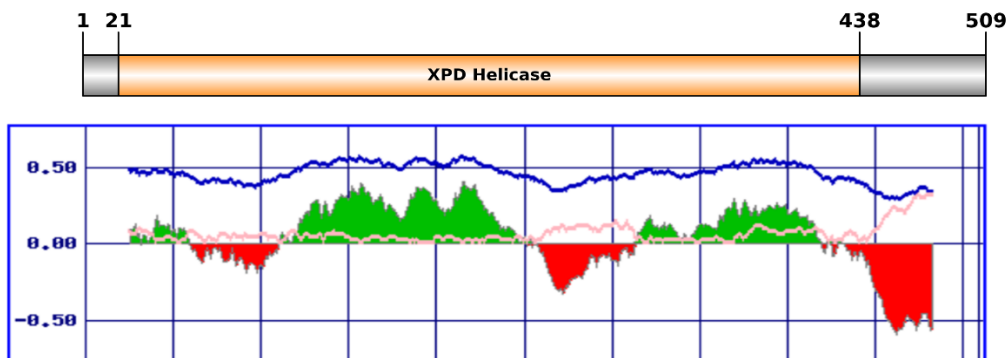
Hepatite C, que interage com uma RNA polimerase dependente de RNA (NS5b) e septina 6, formando um complexo de replicação (KIM et al., 2007).

A interface entre RNA e RRM ocorre através de quatro resíduos conservados localizados nas fitas centrais $\beta 1$ e $\beta 3$ (chamadas RNP2 e RNP1, respectivamente) do RRM, onde as bases nitrogenadas do RNA se ligam à cadeia lateral dos aminoácidos aromáticos localizados na fita $\beta 1$ (posição 2 do RNP2) e na $\beta 3$ (posição 5 do RNP1). O terceiro resíduo aromático na posição $\beta 3$ (posição 3 do RNP1) interage hidrofobicamente com os dois anéis de pentose de cada nucleotídeo (CLÉRY; BLATTER; ALLAIN, 2008; MARIS; DOMINGUEZ, 2005).

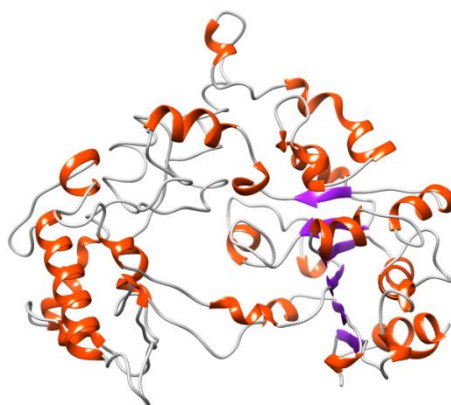
4.2.3 - ORFs wsv479 e wsv497

As ORFs wsv479 e wsv497 também estão localizadas no cluster wsv477/502. Estas ORFs produzem proteínas putativas que são similares em sua composição, tamanho e enovelamento dos aminoácidos, sugerindo que elas possuem funções similares. Não foi possível inferir qualquer função por homologia de sequência. Através de abordagens de *threading*, foi possível determinar que a estrutura terciária de ambas as ORFs apresentam *fold* semelhante ao da helicase de reparo de *Xeroderma Pigmentosum* do grupo D (XPD, PDB: 4A15) (Figuras 8 e 9).

A



B



C

WSV479 XPD Helicase			
Clashscore, todos os átomos:	10.8	67 th percentil* (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	13	3.48%	Ideal: <0.3%
Rotâmeros favoráveis	326	87.17%	Ideal: >98%
Ramachandran aberrantes	22	5.29%	Ideal: <0.05%
Ramachandran favoráveis	337	81.01%	Ideal: >98%
Escore do MolProbity [™]	2.68	37 th percentil* (N=27675, 0Å - 99Å)	
Desvios do C β >0.25Å	16	4.07%	Ideal: 0
Ligações ruins:	1 / 3411	0.03%	Ideal: 0%
Ângulos ruins:	83 / 4604	1.80%	Ideal: <0.1%
Prolinas Cis:	0 / 18	0.00%	Esperado: ≤ 1 por cadeia, ou $\leq 5\%$
Cis não-Prolinas:	2 / 399	0.50%	Ideal: <0.05%
Peptídeos torcidos:	6 / 417	1.44%	Ideal: 0

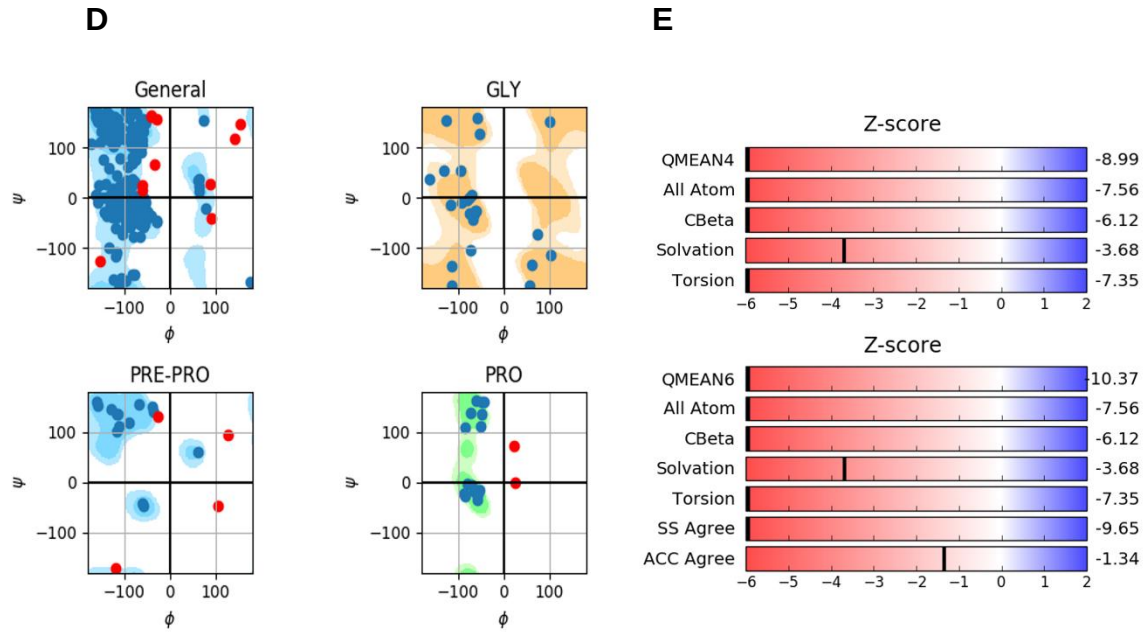
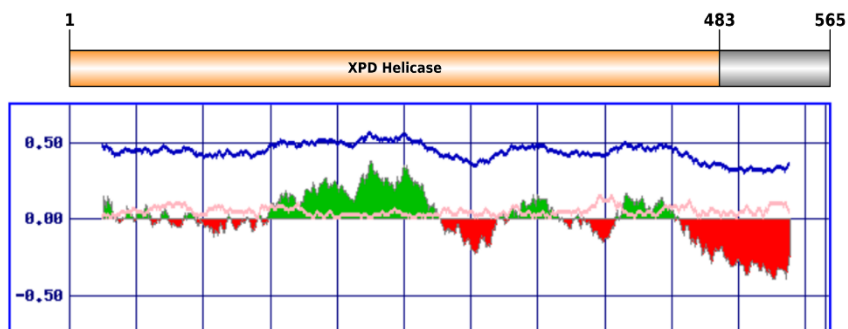


Figura 8 - Caracterização estrutural da proteína putativa codificada pela wsv479. (A) Representação esquemática da sequência proteica codificada pela wsv479 e seus respectivos domínios conservados. A região correspondente a XPD Helicase está representada em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv479 correspondente a XPD Helicase. (C) Escores de validação da qualidade do modelo obtido através do Molprobability. (D) Plots de Ramachandram gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.

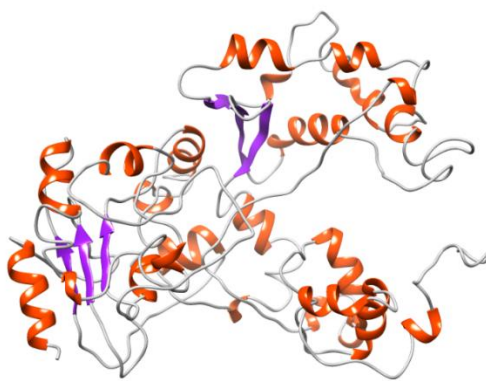
XPD pertence à superfamília helicase 2 e um componente do fator de transcrição IIH (TFIIH), que está associado à via de reparo de excisão nuclear (NER), catalisando a abertura da dupla hélice ao redor do local danificado, fornecendo acesso a fatores NER em um processo dependente de ATP (KUPER et al., 2012). A helicase XPD consiste de dois domínios motores chamados HDR1 e HDR2, um domínio *Arch* e um cluster de ferro-enxofre (FeS) sobreposto na HDR1. O processo de ligação do XPD ao DNA é através do domínio HDR2 (LIU et al., 2012).

Foi descrito anteriormente que as sequências wsv479 e wsv497 têm um domínio VP9 conservado (também conhecido como ICP11) localizado na região N-terminal que possui um *fold* semelhante ao da ferredoxina sugerido a função de reconhecimento de DNA (LIU et al., 2006). Considerando os resultados apresentados aqui, essas proteínas provavelmente possuem funções relacionadas ao processamento do genoma e eventos de recombinação.

A



B



C

WSV497 XPD Helicase			
Clashscore, todos os átomos:	2.97	98 th percentil [†] (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	6	1.38%	Ideal: <0.3%
Rotâmeros favoráveis	398	91.49%	Ideal: >98%
Ramachandran aberrantes	26	5.43%	Ideal: <0.05%
Ramachandran favoráveis	388	81.00%	Ideal: >98%
Escore do MolProbity ^{††}	1.92	80 th percentil [†] (N=27675, 0Å - 99Å)	
Desvios do C β >0.25Å	8	1.77%	Ideal: 0
Ligações ruins:	0 / 3938	0.00%	Ideal: 0%
Ângulos ruins:	66 / 5298	1.25%	Ideal: <0.1%
Prolinas Cis:	0 / 16	0.00%	Esperado: ≤ 1 por cadeia, ou $\leq 5\%$
Cis não-Prolinas:	8 / 464	1.72%	Ideal: <0.05%
Peptídeos torcidos:	2 / 480	0.42%	Ideal: 0

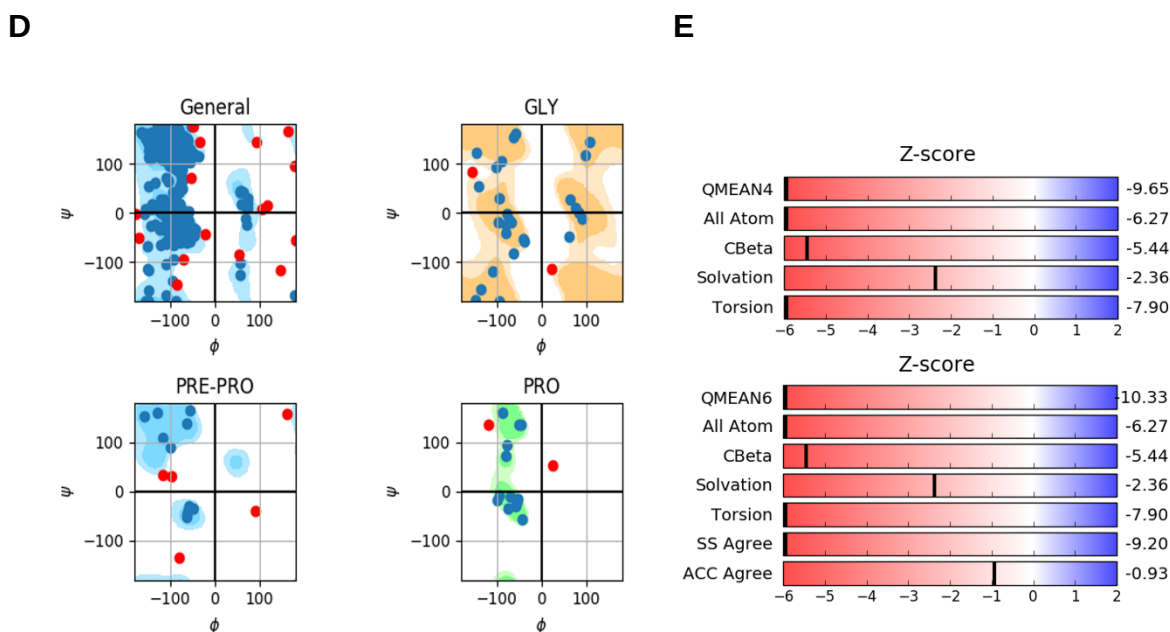


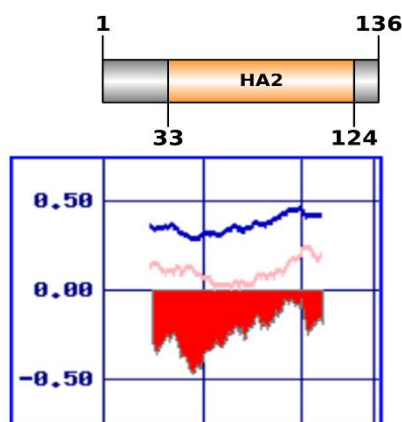
Figura 9 - Caracterização estrutural da proteína putativa codificada pela wsv497. (A) Representação esquemática da sequência proteica codificada pela wsv497 e seus respectivos domínios conservados. A região correspondente a XPD Helicase está representada em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv497 correspondente a XPD Helicase. (C) Escores de validação da qualidade do modelo obtido através do Molprobability. (D) Plots de Ramachandram gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.

4.2.4 - ORF wsv492

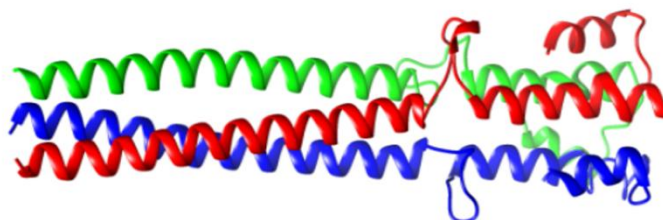
O produto putativo da ORF wsv492 demonstrou alta similaridade com a subunidade HA2 da hemaglutinina do vírus influenza (Figura 10 A). O modelo 3D mostra uma subunidade semelhante a HA2 (PDB: 1QU1) formada por um por três α -hélices (Figura 10 B).

As hemaglutininas consistem em glicoproteínas que permanecem ancoradas no envelope viral e medeiam a entrada viral (CHEN; SKEHEL; WILEY, 1999). Após a síntese proteica, a sequência que codifica a hemaglutinina (HA0) sofre uma clivagem pós-translacional, produzindo as subunidades HA1 e HA2 que formam uma estrutura homotrimérica. A subunidade HA1 interage com o ácido siálico, um monossacarídeo presente na superfície da membrana mediando a endocitose da partícula viral, enquanto a subunidade HA2 consiste em uma estrutura hidrofóbica em forma de bastão associada ao processo de fusão de membrana induzido por pH (WANG et al., 2015), mecanismo com o qual o vírus liberta-se do endossomo e entra em contato com o citosol da célula hospedeira.

A



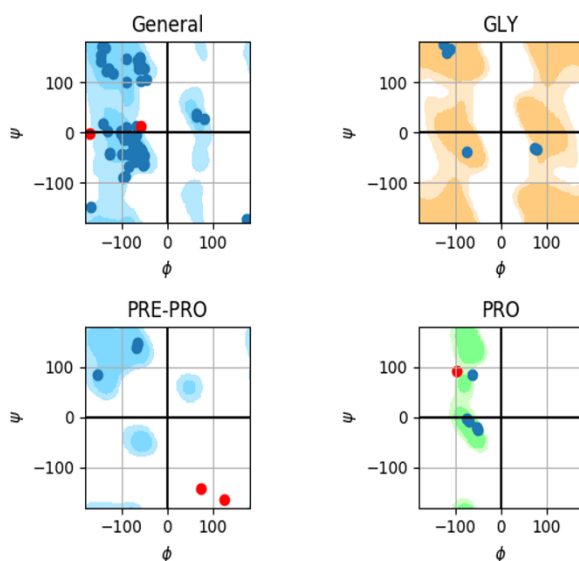
B



C

WSV492 Hemaglutinina subunidade HA2			
Clashscore, todos os átomos:	6.9	87 th percentil [†] (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	12	4.71%	Ideal: <0.3%
Rotâmeros favoráveis	227	89.02%	Ideal: >98%
Ramachandran aberrantes	7	2.59%	Ideal: <0.05%
Ramachandran favoráveis	249	92.22%	Ideal: >98%
Escore do MolProbity [^]	2.37	55 th percentil [†] (N=27675, 0Å - 99Å)	
Desvios do Cβ >0.25Å	8	2.96%	Ideal: 0
Ligações ruins:	1 / 2349	0.04%	Ideal: 0%
Ângulos ruins:	46 / 3138	1.47%	Ideal: <0.1%
Prolinas Cis:	0 / 6	0.00%	Esperado: ≤1 por cadeia, ou ≤5%

D



E

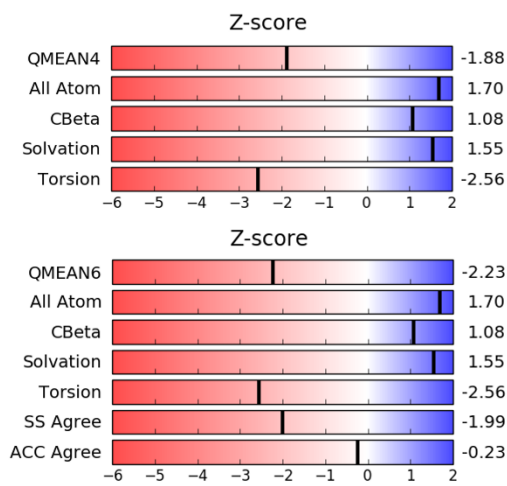


Figura 10 - Caracterização estrutural da proteína putativa codificada pela wsv492. (A) Representação esquemática da sequência proteica codificada pela wsv492 e seus respectivos domínios conservados. A região correspondente a subunidade HA2 da Hemaglutinina do Influenzavirus está representada em laranja. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv492 correspondente a subunidade HA2 da Hemaglutinina. (C) Escores de validação da qualidade do modelo obtido através do Molprobit. (D) Plots de Ramachandram gerados no PyRAMA. (E) Escores globais gerados no QMEAN.

Estes resultados indicam que a proteína putativa codificada por wsv492 tem funções que se assemelham a uma proteína de fusão induzida por mudança de pH, que poderia atuar na etapa de escape do endossomo, corroborando o ciclo de replicação proposto para o WSSV (ESCOBEDO-BONILLA et al., 2008; VERBRUGGEN et al., 2016).

4.3 - Caracterização de ORFs contendo VNTRs.

O alinhamento das ORFs wsv129, wsv178 e wsv249 (Figuras 11, 12 e 13 respectivamente) revelou que as regiões VNTR mais e menos variáveis são as regiões das ORFs wsv129 e wsv249, respectivamente, considerando o valor normalizado pelo número total de sequências (Tabela 5). Observou-se unidades de repetição (RUs) contendo poucos sítios polimórficos em todos os casos. Na ORF wsv178 foi observada uma RU típica já descrita de 54pb, e também substituições ocorrendo nas posições 1, 36 e 48 da RU (Figura 11 B).

A ORF wsv249 (Figura 12 A) possui um número de unidades de repetição bem menor comparada as outras duas VNTR e possui substituições nas posições 2, 9, 12, 27, 50, 53 e 61 de suas RUs, sendo as três últimas ocorrendo com maior frequência (Figura 12 B). A ORF wsv129 apresenta dois tipos de RU, a mais frequente com 45 bp e uma repetição de 57 bp intercalando as RU de 45 bp (Figura 13 A). A presença de *indels* no interior das unidades de repetição da wsv129 ocorrem em com maior frequência em relação as wsv178 e wsv249 (Figura 13 B e C).

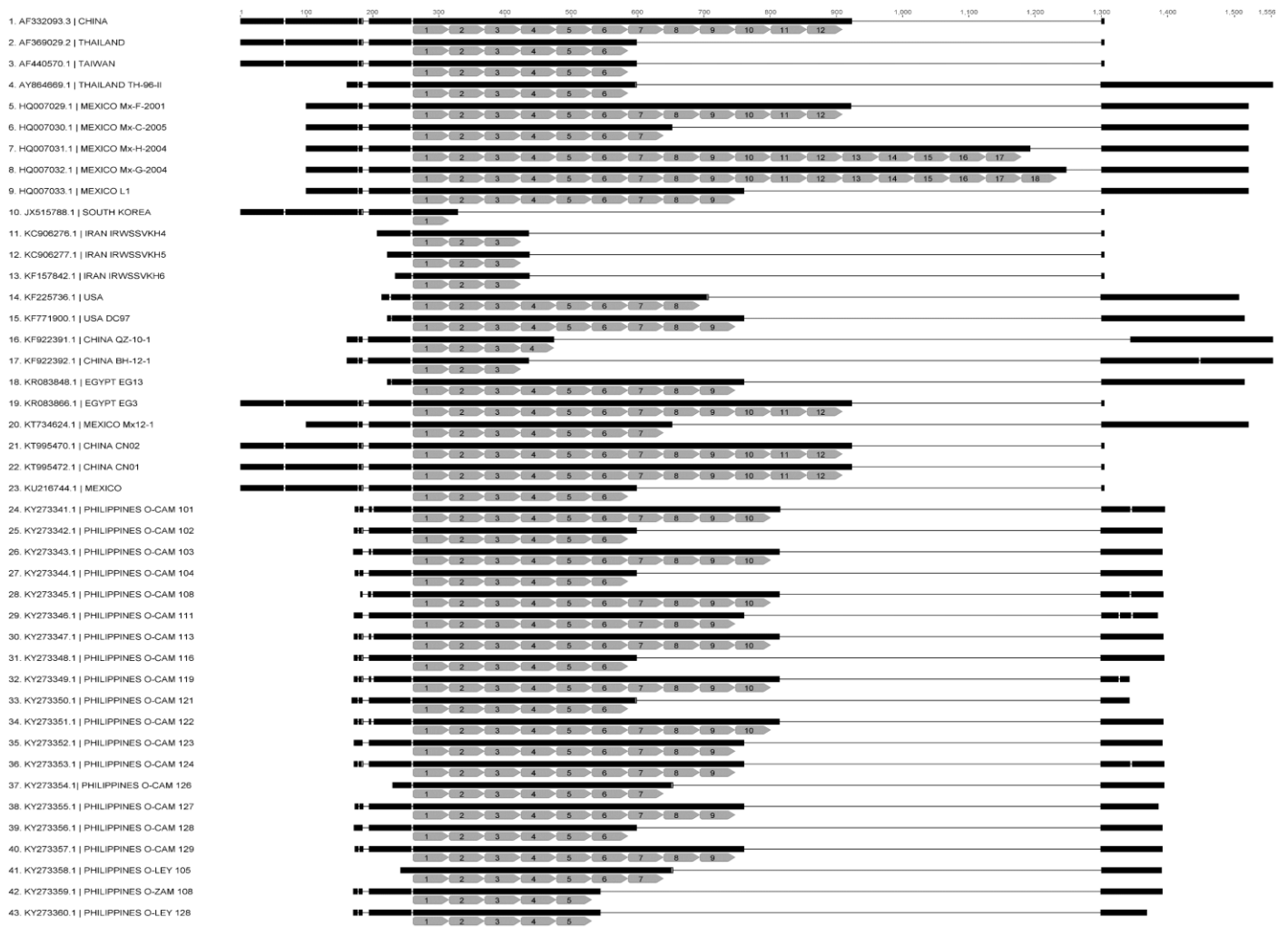
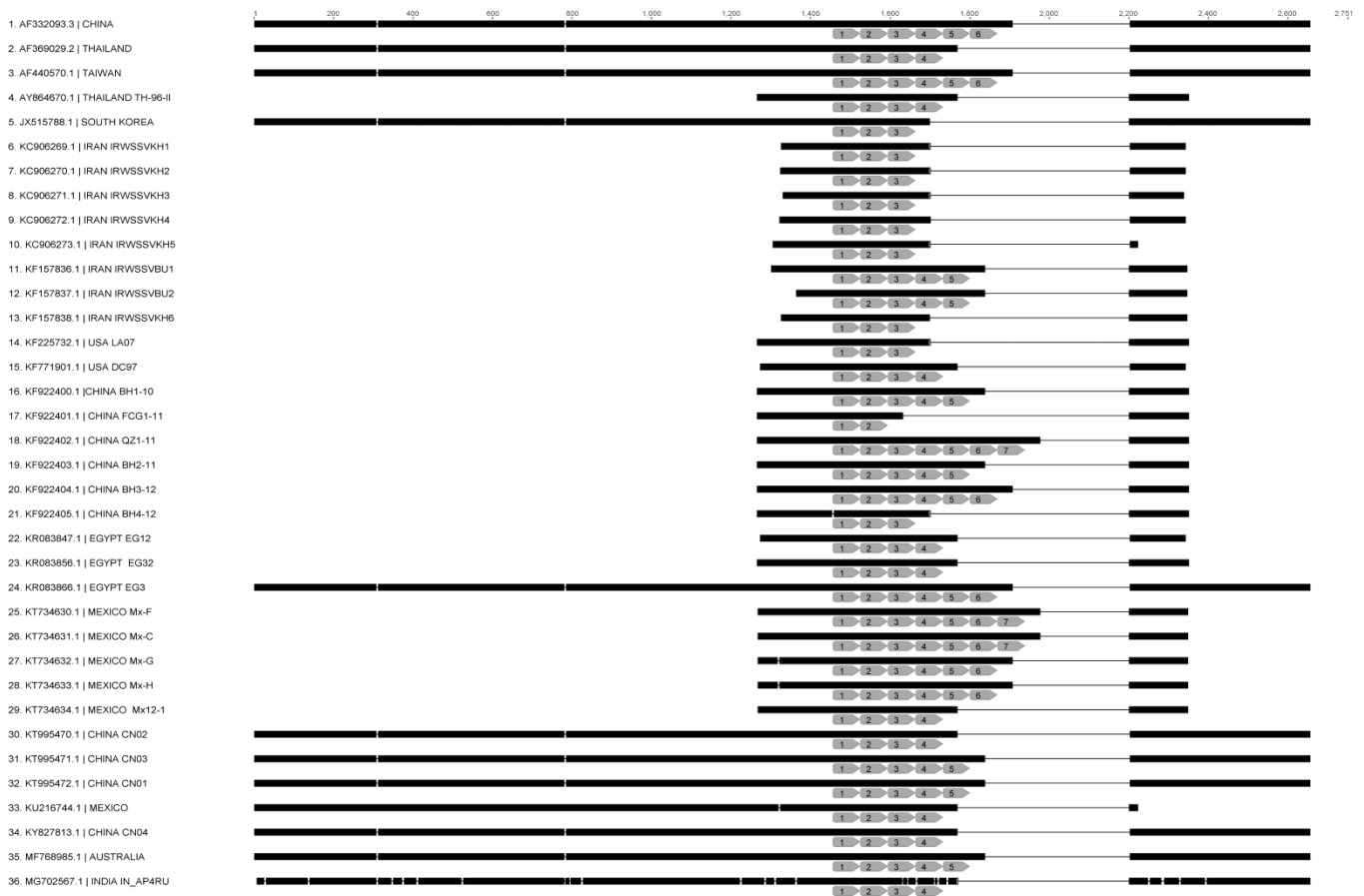
A**B**

Figura 11 - Representação esquemática da região variável wsv178 (ORF94). (A) Alinhamento múltiplo das sequências de nucleotídeos obtido através do Geneious versão 11.0.3. As setas em cinza representam as unidades de repetição de 54pb. (B) Representação da unidade de repetição através do WebLogo mostrando os sítios de polimorfismo

A



B

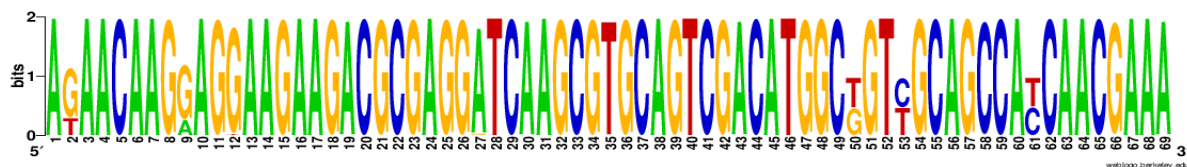
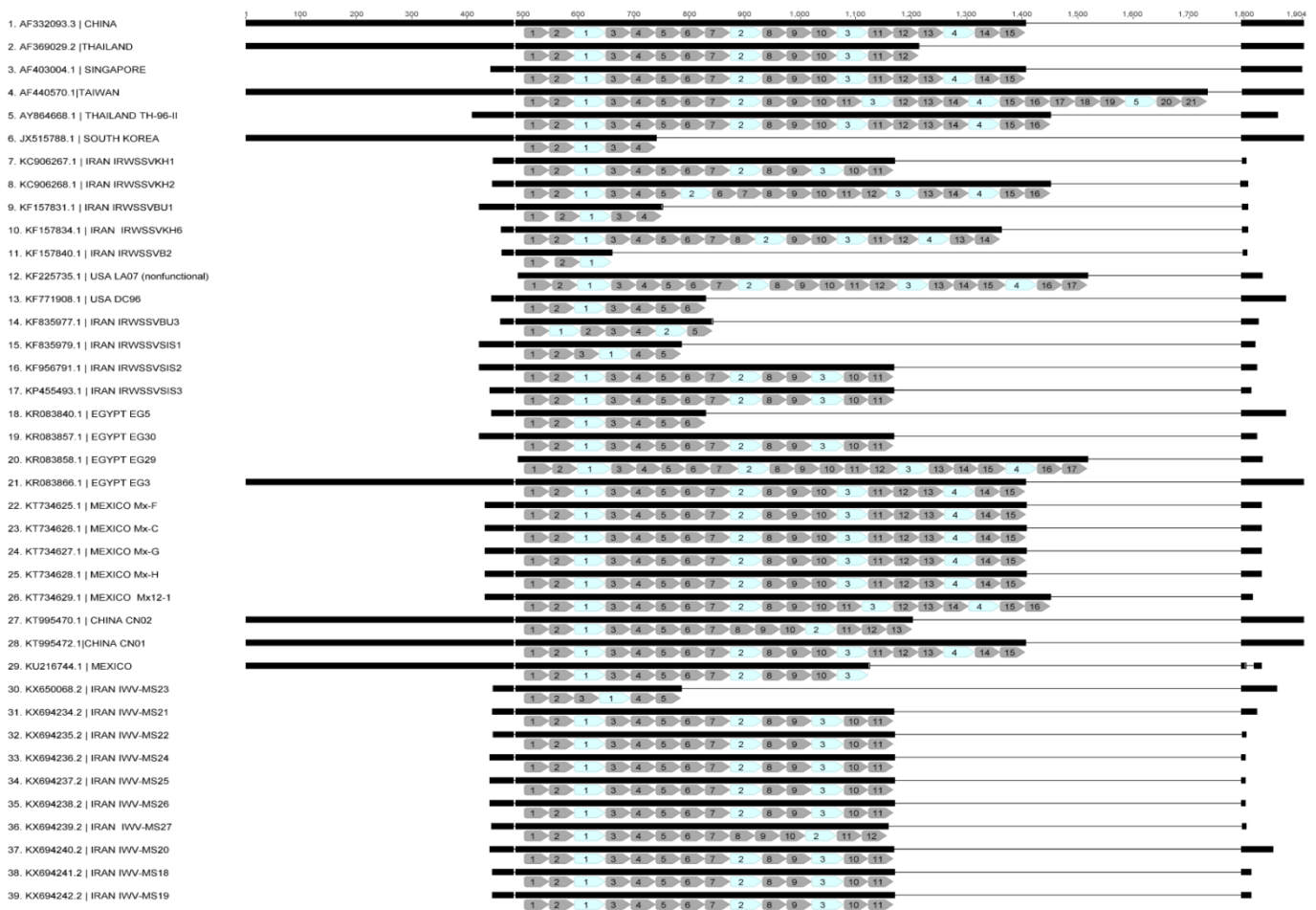


Figura 12 - Representação esquemática da região variável da wsv249 (ORF125). (A) Alinhamento múltiplo das sequências de nucleotídeos obtido através do Geneious versão 11.0.3. As setas em cinza representam as unidades de repetição de 69pb. (B) Representação da unidade de repetição através do WebLogo mostrando os sítios de polimorfismo.

A



B



C

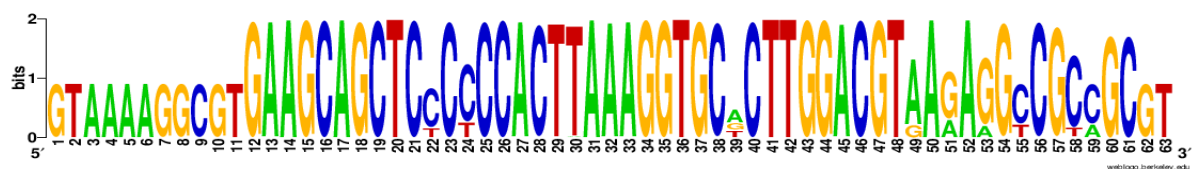


Figura 13 - Representação esquemática da região variável da wsv129 (ORF75). (A) Alinhamento múltiplo das seqüências de nucleotídeos obtido através do Geneious versão 11.0.3. As setas em cinza representam as unidades de repetição de 45pb, enquanto as setas em azul claro representam as unidades de 57pb. (B) Representação da unidade de repetição de 45pb através do WebLogo mostrando os sítios de polimorfismo. (C) Representação da unidade de repetição de 57pb através do WebLogo mostrando os sítios de polimorfismo.

Tabela 5 - Relação das unidades de repetição das VNTRs presentes no WSSV.

VNTRs	Comprimento das RU (pb)	Nº de sequências alinhadas	Nº de diferentes comprimentos de VNTR	Nº mínimo de RU dentro de uma VNTR	Nº máximo de RU dentro de uma VNTR	Nº de RU mais frequentes dentro das VNTR
wsv129.a	45	39	13	2	21	11
wsv129.b	57	39	5	1	5	3, 4
wsv178	54	43	12	1	18	6
wsv249	69	34	5	2	7	4

Apesar do considerável número de conjuntos de sequências de origens comuns no banco de dados, a variação é bem pronunciada nas três VNTRs. Um ponto importante a ser considerado nessa análise é a falta sequências oriundas de outras regiões onde o WSSV é endêmico. Além disso há alguns casos em que as sequências depositadas não apresentam o início e final das unidades de repetição, ou seja, as regiões conservadas, impossibilitando inferir o número real de unidades de repetição destas que por sua vez, foram excluídas da análise.

Os perfis das RU observadas coincidem com aqueles já descritos anteriormente (RAMOS-PAREDES et al., 2012; SHEKAR; PRADEEP; KARUNASAGAR, 2012). Curiosamente a janela de leitura é mantida independente da variação no número de RUs. Não foi possível obter modelos tridimensionais confiáveis para os produtos putativos das ORFs wsv129 e wsv178, uma vez que elas possuem uma grande porção não estruturada e muitos aminoácidos carregados.

As VNTRs do WSSV também têm sido usadas para os estudos epidemiológicos e a wsv178 pode ser considerada como o marcador mais usado para determinar variantes (GUDKOV; MURWANTOKO; WALKER, 2014; SIMROUNI et al., 2014).

Tem sido sugerido que os genótipos com menor número de unidades de repetição estão diretamente relacionados com uma alta virulência WSSV. Foram observados cerca de 12 genótipos relacionados a ORF wsv178 de amostras coletadas de fazendas de camarão onde houve surtos de Mancha Branca entre os anos de 2000 e 2002 Tailândia contendo entre 6 – 20 RU na ORF wsv178 sendo o genótipo que apresentou 8 unidades de repetição o mais prevalente dentro da população (WONGTEERASUPAYA et al., 2003).

A partir de tecidos de camarão isolados de fazendas entre os anos de 2005-2013, foi observado que os genótipos das amostras isoladas de regiões de surto apresentavam um número de unidades de repetição bem menor na wsv178 (entre 3 e 5 RU) comparado às regiões que não havia uma alta mortalidade, que possuíam entre 9-13 RU (RODRIGUEZ ANAY et al., 2016).

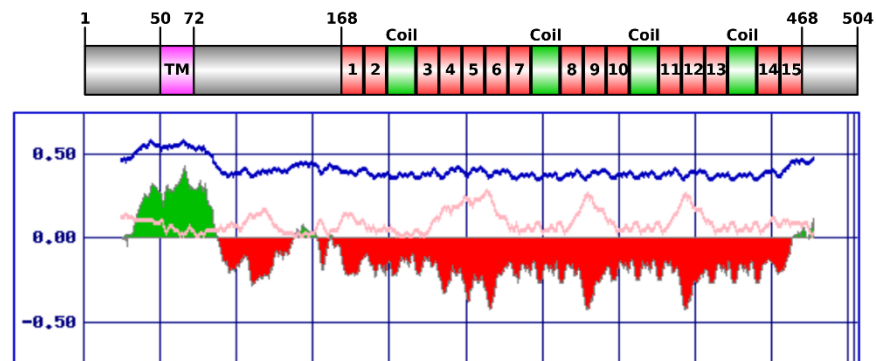
A virulência de três genótipos encontrados na Índia foi caracterizada após infectarem camarões (*Penaeus monodon*) e caranguejos (*Scylla serrata*) com o WSSV. Foi sugerido que todas as variantes apresentam a mesma virulência, sendo que, aquelas que possuem menor genoma apresentam maior vantagem na replicação (PRADEEP et al., 2012). Hoa et al., sugeriu que genótipos misturados estão diretamente relacionados com baixa virulência se comparados a uma infecção conduzida por apenas um genótipo (HOA et al., 2011).

Existem casos em que as regiões VNTR não são detectadas nos isolados. Temos como exemplos dados de genotipagem de variantes virais da Arábia Saudita, Madagascar e Moçambique, bem como dados de sequenciamento do genoma da Austrália desprovidos da sequência total ou parcial da wsv178 (OAKEY; SMITH, 2018; TANG; LE GROUMELLE; LIGHTNER, 2013). Contudo, foi observado que os *primers* usados foram desenvolvidos em 2003 (WONGTEERASUPAYA et al., 2003) e por se tratar de variantes mais recentes, é possível que os *primers* não tenham apresentado eficiência para genotipagem em decorrência de alterações na sequência alvo.

4.3.1 - ORF wsv129

As predições da estrutura secundária do produto codificado pela ORF wsv129 revelaram pequenas hélices transmembrana na região N-terminal (posições 50-72) que coincidem com uma região estruturada da proteína predita (Figura 14 A). Curiosamente, vários *Coiled coils* codificados pelas unidades de repetição de 57pb e sinais de localização nuclear (NLS) são preditos no final de cada repetição de 45pb na região VNTR (Figura 12 B).

A



B

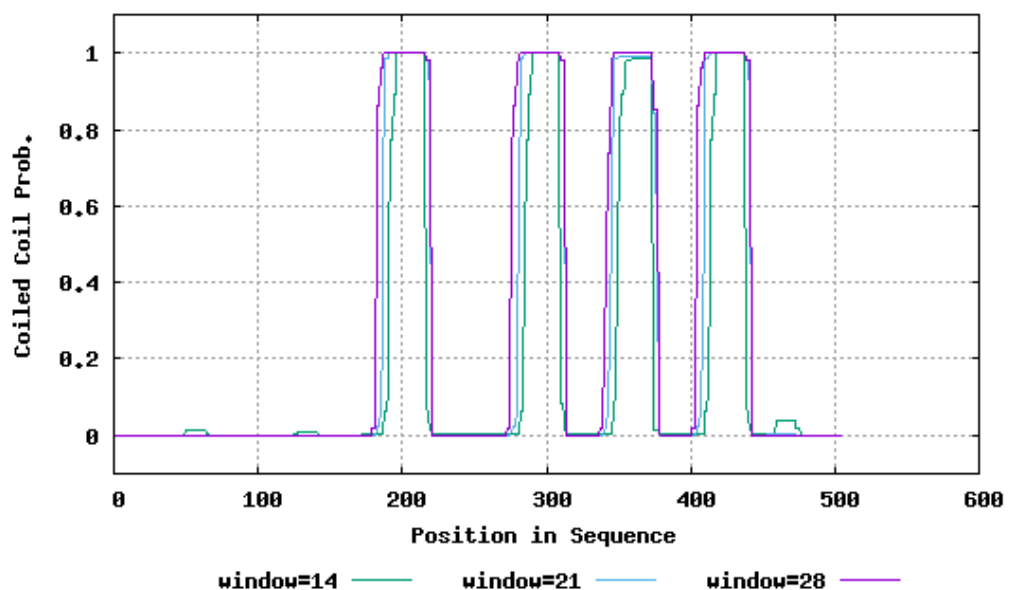


Figura 14 – Representação esquemática da sequência proteica codificada pela wsv129 (ORF75) e seus respectivos domínios conservados. (A) A região transmembrana (TM) está representada em rosa, as unidades de repetição correspondentes às sequências codificantes de 45pb estão em vermelho, enquanto as repetições intercaladas correspondentes às unidades de repetição de 57pb estão marcadas em verde. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Gráfico representando a probabilidade de ocorrência de *Coiled coils* dentro da sequência proteica codificada pela wsv129.

A predição de uma grande porção não estruturada rica em *Coiled coils* em associação com o domínio transmembrana indica uma função estrutural. De fato, dezoito proteínas estruturais localizadas no vírion do WSSV, incluindo a proteína codificada por wsv129, foram detectadas previamente, por análise proteômica (HUANG, 2002). Neste mesmo trabalho, uma análise temporal mostrou que o produto wsv129 é expresso tardiamente, pelo menos 6 horas após a infecção pelo WSSV.

O NLS age direcionando proteínas para o núcleo da célula e pode ser subdividido em duas subclasses: monopartida, formada por um grupo que consiste de uma sequência K (K/R) X (K/R) ou bipartida, formada por dois grupos de resíduos básicos separados por um ligante de 10-12 aminoácidos que podem variar (Lange et al., 2007). O transporte de macromoléculas através do núcleo ocorre através de um complexo proteico de nucleoporina chamado complexo nuclear de poros (NPC). Proteínas acima de 40kDa requerem um sinal específico que permitirá interagir com proteínas transportadoras que facilitarão sua entrada no núcleo (COKOL; NAIR; ROST, 2000; MARFORI et al., 2011; RIDDICK; MACARA, 2007).

A ligação da proteína transportadora da NLS é importante para sua liberação no núcleo. Vírus de DNA que infectam animais se replicam no núcleo da célula hospedeira. Para conduzir a entrada do genoma viral no núcleo, os vírus grandes normalmente liberam seu DNA associado a proteínas estruturais associadas a sinais de localização nuclear (SAGE; MOULAND, 2013). Apesar de apresentar o padrão, não foram encontrados na literatura proteínas virais contendo sinais de NLS em conjunto. Como o mecanismo de entrada no núcleo da célula hospedeira é desconhecido essa pode ser uma evidência interessante a ser trabalhada. Por outro lado, não é possível descartar a possibilidade de que esses sinais NLS sejam artefatos da análise.

4.3.2 - ORF wsv178

Pequenas hélices transmembrana também foram previstas na região N-terminal do produto putativo wsv178 (posições 7-29) coincidindo com a única região dobrada da proteína (Figura 15). Curiosamente, foi previsto um local de clivagem entre os resíduos 26 e 27, o qual poderia separar o produto codificado pela região VNTR da porção transmembrana. A presença de sinais putativos de localização nuclear também foi observada no início de cada unidade de repetição.

Estes dados estruturais estão de acordo com dados anteriores, sugerindo que wsv129 e wsv178 estão associados a funções similares (DIEU, 2004). Estas proteínas também podem ter alguma função adaptativa, considerando que estes dois VNTR são os mais variáveis, mesmo entre os isolados de uma mesma região.

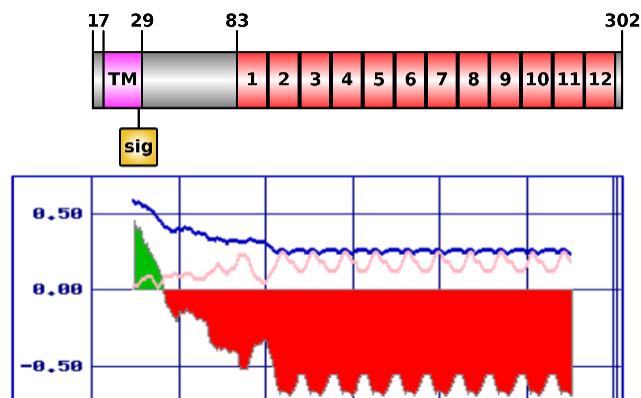


Figura 15 – Representação esquemática da sequência proteica codificada pela wsv178 (ORF94) e seus respectivos domínios conservados. A região transmembrana (TM) está representada em rosa, o sítio de clivagem se apresenta como o quadrado amarelo, enquanto as unidades de repetição correspondentes às sequências codificantes de 54pb estão em vermelho. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade.

4.3.3 - ORF wsv249

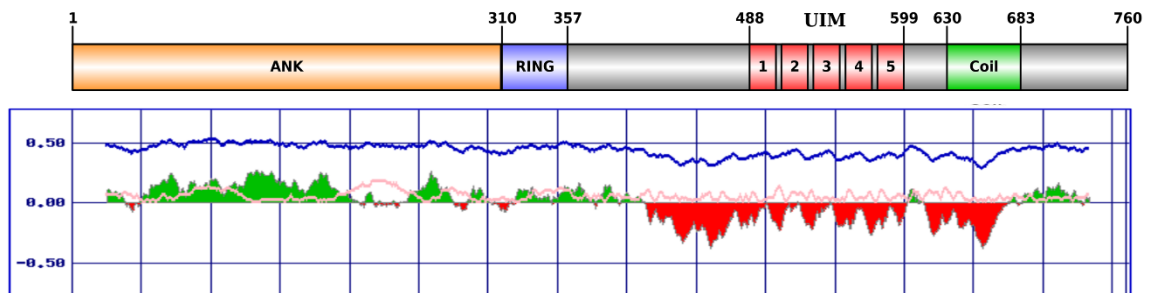
O produto gerado pelo wsv249 (ORF125) tem uma cadeia mais estruturada em comparação com os outros ORFs que contêm regiões VNTR (Figura 16). A caracterização por homologia remota e reconhecimento de fold revelou que os primeiros 300 resíduos da região do terminal N correspondem a um motivo de repetição de *Ankyrin* (ANK, PDB: 4HDB) (Figura 16 A, B). Também foi possível identificar um domínio RING-H2 imediatamente após a região do domínio ANK. O modelo 3D do domínio RING-H2, modelado a partir da estrutura do domínio *Ring-finger* (PDB: 4AP4), é mostrado na Figura 17 A, bem como os principais aminoácidos conservados. Os testes de validação deste modelo tridimensional são apresentados abaixo dos modelos.

Alguns resíduos de glicina (89, 171, 231) localizados na segunda folha beta, bem como resíduos de leucina localizados na segunda hélice alfa (105, 106, 109, 110, 148, 199, 247, 248, 250, 282) do motivo ANK permaneceram em posições conservadas no modelo proposto. O típico tetrapeptídeo TPLH não foi observado em sequência devido à divergência no nível da estrutura primária.

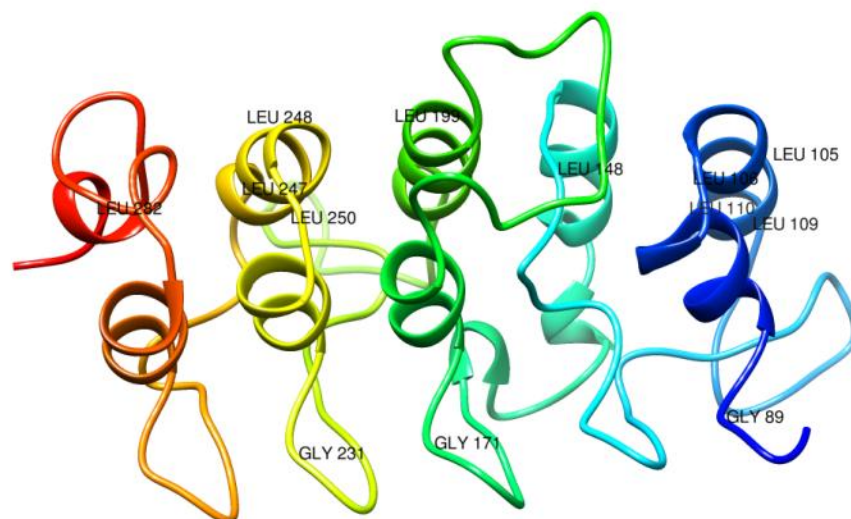
Dados *SMART* revelaram que as unidades de repetição *in tandem* em wsv249, apesar do alto valor E (devido ao pequeno comprimento da sequência),

correspondem a motivos de interação ubiquitina (UIM). Além disso, um RING-H2 foi detectado imediatamente ao lado do domínio Ankirin localizado entre as posições 310-357 da proteína (Figuras 14 A e 15) que tem o Cys-X₂-Cys-X (9-39)-Cys-X(1-3)-His-X (2-3)-His-X₂-Cys-X (4-48)-Cis-X₂-Cys, em que "X" compreende qualquer aminoácido.

A



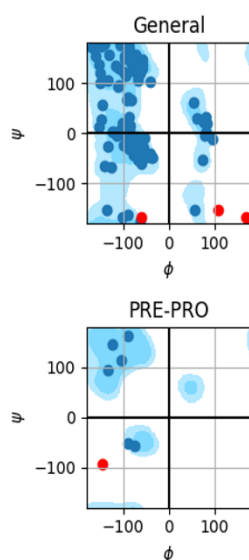
B



C

WSV249(ORF125) Motivo ANK			
Clashscore, todos os átomos:	9.45	74 th percentil [†] (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	5	2.82%	Ideal: <0.3%
Rotâmeros favoráveis	164	92.66%	Ideal: >98%
Ramachandran aberrantes	5	2.43%	Ideal: <0.05%
Ramachandran favoráveis	170	82.52%	Ideal: >98%
Escore do MolProbity [^]	2.54	45 th percentil [†] (N=27675, 0Å - 99Å)	
Desvios do Cβ >0.25Å	5	2.60%	Ideal: 0
Ligações ruins:	0 / 1676	0.00%	Ideal: 0%
Ângulos ruins:	28 / 2259	1.24%	Ideal: <0.1%
Prolinas Cis:	0 / 7	0.00%	Esperado: ≤1 por cadeia, ou ≤5%
Cis não-Prolinas:	1 / 200	0.50%	Ideal: <0.05%
Peptídeos torcidos:	2 / 207	0.97%	Ideal: 0

D



E

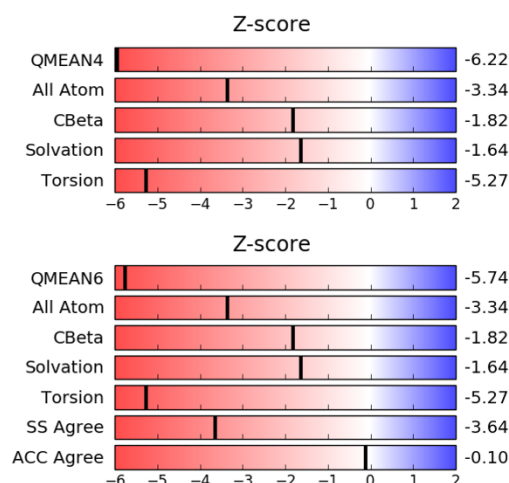


Figura 16 – Caracterização estrutural da proteína putativa codificada pela wsv249 (ORF125) na posição 86-293. (A) Representação esquemática da sequência proteica codificada pela wsv249 e seus respectivos domínios conservados. A região correspondente a subunidade ao motivo ANK está representada em laranja, o domínio “Ring-H2” em azul, unidades de repetição caracterizados como motivos de interação com ubiquitina (UIM) em vermelho e uma região *Coiled coil* em verde. O gráfico abaixo gerado no Foldindex descreve as regiões estruturadas (em verde) e não estruturadas (em vermelho). O eixo x indica as posições dos aminoácidos e o eixo y representa os escores de probabilidade. (B) Modelo tridimensional da sequência codificada pela wsv249 correspondente ao motivo ANK. (C) Escores de validação da qualidade do modelo obtido através do Molprobity. (D) Plots de Ramachandran gerados no no PyRAMA. (E) Escores globais gerados no QMEAN.

Repetição de *Ankyrin* (ANK) é um motivo composto de cerca de 33 resíduos de aminoácidos que são importantes na modulação de várias vias celulares mediando interações específicas proteína-proteína; a maioria das sequências de

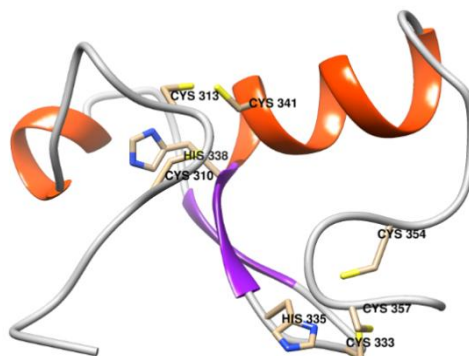
proteínas que exibem esses motivos geralmente consistem em reguladores transcricionais, moduladores do desenvolvimento e diferenciação celular (PARRA et al., 2015). O motivo ANK adota uma estrutura de hélice-alça-hélice na qual duas hélices alfa estão dispostas de uma forma anti-paralela e a alça sobressai para fora da armação para facilitar a formação de folha beta tipo gancho com laçadas vizinhas. O tetrapeptídeo conservado T-P-L-H (6-9) forma uma curva fechada começando na primeira hélice alfa do ANK (LI; MAHAJAN; TSAI, 2006). As ligações de hidrogênio entre os grupos hidroxila de treonina e o anel imidazólico de histidina contribuem para a estabilidade da ANK. O motivo Val/Ile-Val-Xxx (hipofílico)-Leu/Val-Leu-Leu (posições 17-22) localizado na segunda hélice alfa estabiliza a estrutura geral de uma proteína ANK. Os resíduos de glicina são armazenados nas posições 4 localizadas na segunda folha beta, 13 e 25 na extremidade da hélice alfa (HUANG et al., 2006; SEDGWICK; SMERDON, 1999).

ANK é um motivo encontrado em abundância em espécies eucarióticas, mas pouco se sabe sobre a presença de ANK em proteínas virais, exceto para espécies do gênero poxvírus que apresentam motivos ANK nas regiões terminais das proteínas envolvidas no processo de ubiquitinação (HERBERT; SQUIRE; MERCER, 2015; NOEL et al., 2014). Um PSI-BLAST do domínio ANK foi realizado contra sequências de poxvírus e foi possível observar acertos para o motivo ANK, apesar da baixa cobertura apresentada.

UIMs consistem em aproximadamente 20 resíduos de aminoácidos que obedecem um padrão X-Ac-Ac-Ac-Ac- ϕ -XX-Ala-XXX-Ser-XX-Ac-XXXX, onde " ϕ " corresponde a um resíduo hidrofóbico, "Ac" é um resíduo ácido e "X" pode ser qualquer resíduo de aminoácido (FISHER et al., 2003; HOFMANN; FALQUET, 2001). Estes motivos estão diretamente associados a reconhecimento de proteínas ubiquitinadas.

O produto da wsv249 foi previamente caracterizado como uma ubiquitina ligase, atuando como um elemento chave na modulação da abundância de proteínas dentro das células através do mecanismo de proteólise dependente de ubiquitina (WANG et al., 2005; ZHENG; SHABEK, 2017). A atividade da ubiquitina-ligase é governada pelo domínio RING-H2. A presença de ANK em conjunto com o RING-H2 reforça a função do wsv249 na modulação de proteínas hospedeiras.

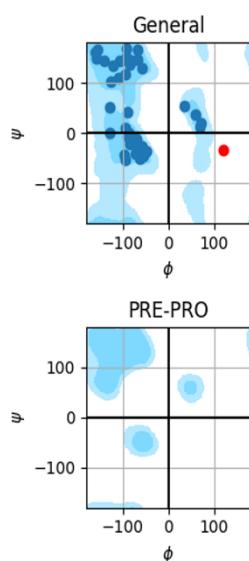
A



B

WSV249(ORF125) RING-H2			
Clashscore, todos os átomos:	2.24	99 th percentil [*] (N=1784, Todas as resoluções)	
Clashscore é o número de sobreposições estéricas graves (> 0.4 Å) por 1000 átomos.			
Rotâmeros pobres	1	1.89%	Ideal: <0.3%
Rotâmeros favoráveis	49	92.45%	Ideal: >98%
Ramachandran aberrantes	1	1.85%	Ideal: <0.05%
Ramachandran favoráveis	50	92.59%	Ideal: >98%
Escore do MolProbity [^]	1.67	90 th percentil [†] (N=27675, 0Å - 99Å)	
Desvios do C β >0.25Å	1	1.85%	Ideal: 0
Ligações ruins:	0 / 451	0.00%	Ideal: 0%
Ângulos ruins:	11 / 603	1.82%	Ideal: <0.1%
Prolinas Cís:	0 / 0	0.00%	Esperado: ≤ 1 por cadeia, ou $\leq 5\%$

C



D

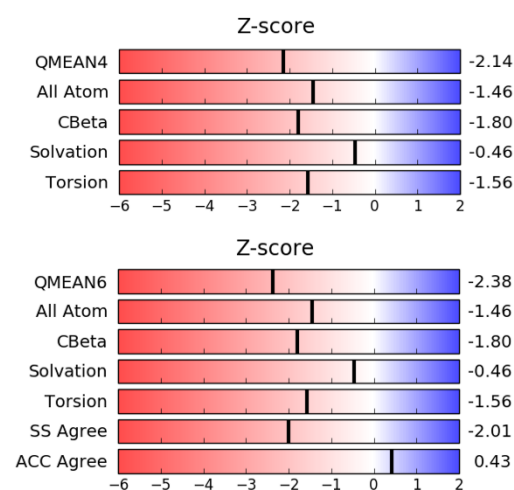


Figura 17 – Caracterização estrutural da proteína putativa codificada pela wsv249 (ORF125) na posição 310-357. (A) Modelo tridimensional da sequência codificada pela wsv249 correspondente ao domínio RING-H2. (B) Escores de validação da qualidade do modelo obtido através do Molprobity. (C) Plots de Ramachandram gerados no no PyRAMA. (D) Escores globais gerados no QMEAN.

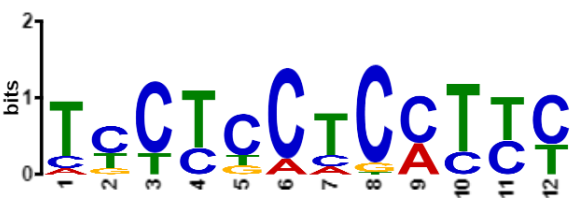
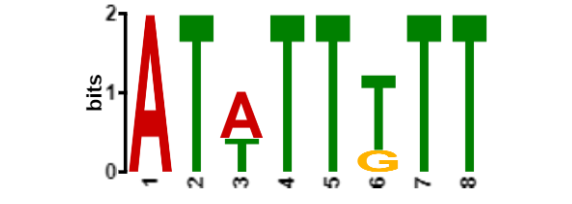
4.4 Predição de regiões regulatórias no genoma do WSSV

A busca realizada retornou 100 possíveis motivos que ocorrem dentro das regiões upstream a cada ORF predita do WSSV. Os motivos de maior ocorrência e mais significativos detectados pelo MEME estão descritos na Tabela 6.

Dentro das regiões entre 100 nucleotídeos upstream das sequências codificantes foi possível detectar uma grande ocorrência de motivos ricos em A e T. Os três primeiros motivos apresentados na tabela possuem maior ocorrência nas regiões 200nt upstream de sequências codificantes já caracterizadas tais como a DNA polimerase (ORF27), a subunidade menor da Ribonucleotideo redutase (ORF98), a dUTPase (ORF71) e a proteína de ligação responsiva ao AMPc (ORF66).

A partir da busca obtida no TOMTOM foi observado que os três primeiros hits foram preditos para sítio de ligação de fatores *zinc-finger* C2H2, que são motivos muito comuns entre proteínas que atuam como fatores de transcrição, sugerindo que estas regiões atuam como sítios de ligação de fatores de transcrição destas sequências codificantes. Variações mais acentuadas destes motivos também foram detectadas em menor ocorrência em outras regiões.

Tabela 6 - Perfis de seqüências motivos de maior ocorrência dentro das regiões 200nt upstream das seqüências codificantes do genoma do WSSV obtidos através do software MEME.

Motivo	E-Value	Número de ocorrências
	6.6e-013	48
	2.2e-002	25
	6.9e-001	22
	3.4e+003	32
	1.2e+000	85

Motivos ricos em A e T são muito comuns em muitas espécies de grandes vírus de DNA dupla-fita e tem sido demonstrado que são fundamentais para expressão de genes virais. Ensaios de mutação pontual e de imunofluorescência em marciellevirus demonstraram que mudanças no octâmero AAATATT influênciam de forma negativa na expressão (OLIVEIRA et al., 2017). Com relação às regiões regulatórias do WSSV, Marks et al., quantificou a abundância dos motivos promotores de 4 a 8 nt presentes a 100 e 200 upstream das ORFs do WSSV através

de uma abordagem enumerativa. Também foi usado o software MEME para caracterizar os elementos regulatórios presentes nas regiões upstream dos genes de expressão precoce e tardia que foram atribuídos através de microarray.

Através da abordagem enumerativa foi possível observar a abundância do TATA box a 100nt upstream das ORF, bem como o motivo AACC encontrados em clusters de expressão tardia e motivos de 4 e 5-mer ricos em C e G. A análise através do MEME demonstrou um resultado semelhante ao observado na abordagem de enumeração, identificando a presença do *TATA box* consenso e do sítio de iniciação da transcrição (ITS) em grande parte dos genes de expressão precoce e tardia (MARKS et al., 2006), bem como a partir do alinhamento das regiões promotoras dos genes experimentalmente caracterizados foi possível identificar os motivos (a/c)TCANT e ATNAC como sítios de iniciação da transcrição dos genes precoces e tardios, respectivamente. As ITS de expressão precoce possuem similaridade com o sítio de ligação de RNA polimerase II de *Drosophila*.

5 CONCLUSÃO

Com relação a filogenia, foram observados indícios de que as regiões onde ocorrem inserções são mais determinantes para a reconstrução filogenética do WSSV.

As proteínas codificadas pelas regiões de inserção no genoma do WSSV que puderam ter sua estrutura caracterizada por abordagens *in silico* possivelmente apresentam funções relacionadas à manutenção do citoesqueleto (wsv463a), clivagem e reparo do DNA (wsv479 e wsv497) e ancoragem e mecanismos relacionados ao estágio inicial da infecção viral (wsv492). Considerando que essas inserções ocorrem em poucos isolados, essas funções possivelmente não são essenciais para o mecanismo de infecção do WSSV. Por outro lado, a proteína wsv477, que ocorre em todos os isolados na região que precede a inserção, possivelmente está associada ao acondicionamento e/ou manutenção do genoma viral.

Duas das três proteínas que apresentam regiões VNTR são em grande parte não estruturadas e, dadas as suas características, possivelmente têm algum tipo de função estrutural na formação de envelope ou nucleocapsídeo. A existência de *Coiled coils* na região VNTR da wsv129 corrobora com uma estrutura que permite a

ligação de outras moléculas. A proteína wsv249 foi melhor caracterizada, a presença de ANK em conjunto com o RING-H2 na mesma proteína pode estar associada não somente ao processo de ubiquitinação, mas também à modulação de proteínas hospedeiras.

Foi também mostrada a presença de motivos de frequente ocorrência no genoma do WSSV que interagem com proteínas de domínio *Zinc-Finger* C2H2 e podem estar diretamente relacionados com expressão de genes através da interação com fatores de transcrição.

REFERÊNCIAS

- ABDOLLAHI, S.; RASOOLI, I.; MOUSAVI GARGARI, S. L. An in silico structural and physicochemical characterization of TonB-dependent copper receptor in *A. baumannii*. **Microbial Pathogenesis**, v. 118, n. October 2017, p. 18–31, 2018.
- AGGARWAL, A. et al. Mobilization of HIV spread by diaphanous 2 dependent filopodia in infected dendritic cells. **PLoS Pathogens**, v. 8, n. 6, 2012.
- ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403–10, 1990.
- ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389–3402, 1997.
- ALVAREZ, D. E.; AGAISSE, H. The formin FHOD1 and the small GTPase Rac1 promote vaccinia virus actin-based motility. **Journal of Cell Biology**, v. 202, n. 7, p. 1075–1090, 2013.
- ASPENSTRÖM, P. Formin-binding proteins: Modulators of formin-dependent actin polymerization. **Biochimica et Biophysica Acta - Molecular Cell Research**, v. 1803, n. 2, p. 174–182, 2010.
- BENSON, D. A. et al. GenBank. **Nucleic Acids Research**, v. 33, n. DATABASE ISS., p. 34–38, 2005.
- BIASINI, M. et al. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. **Nucleic Acids Research**, v. 42, n. W1, p. 252–258, 2014.
- BOEVA, V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. **Frontiers in Genetics**, v. 7, n. FEB, 2016.
- BUCHAN, D. W. A. et al. Scalable web services for the PSIPRED Protein Analysis Workbench. **Nucleic acids research**, v. 41, n. Web Server issue, p. 349–357, 2013.
- CHEN, J.; SKEHEL, J. J.; WILEY, D. C. N- and C-terminal residues combine in the fusion-pH influenza hemagglutinin HA2 subunit to form an N cap that terminates the triple-stranded coiled coil. **Proceedings of the National Academy of Sciences**, v. 96, n. 16, p. 8967–8972, 1999.
- CHEN, L. et al. Identification of a Nucleocapsid Protein (VP35) Gene of Shrimp White Spot Syndrome Virus and Characterization of the Motif Important for Targeting VP35 to the Nuclei of Transfected Insect Cells. v. 53, p. 44–53, 2002.
- CHEN, V. B. et al. MolProbity: All-atom structure validation for macromolecular crystallography. **Acta Crystallographica Section D: Biological Crystallography**, v. 66, n. 1, p. 12–21, 2010.
- CHOU HSIN-YIU et al. Pathogenicity of a baculovirus infection causing white spot syndrome in cultured penaeid shrimp in Taiwan. **Diseases of Aquatic Organisms**, v. 23, n. 3, p. 165–173, 1995.
- CLÉRY, A.; BLATTER, M.; ALLAIN, F. H. T. RNA recognition motifs: boring? Not quite. **Current**

Opinion in Structural Biology, v. 18, n. 3, p. 290–298, 2008.

COKOL, M.; NAIR, R.; ROST, B. Finding nuclear localization signals. **EMBO reports**, v. 1, n. 5, p. 411–5, 2000.

CONTE, M. R. Structure of tandem RNA recognition motifs from polypyrimidine tract binding protein reveals novel features of the RRM fold. **The EMBO Journal**, v. 19, n. 12, p. 3132–3141, 2000.

CROOKS, G. et al. NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. **Genome Res**, v. 14, p. 1188–1190, 2004.

DANTAS, M. D. A. et al. Direct sequencing of the white spot syndrome virus from Brazil: Genome assembly and new insights on phylogeny. **Virus Research**, v. 245, n. December 2017, p. 52–61, 2018.

DAS, M. K.; DAI, H. K. A survey of DNA motif finding algorithms. **BMC Bioinformatics**, v. 8, n. SUPPL. 7, p. 1–13, 2007.

DIEU, B. T. M. Molecular epidemiology of white spot syndrome virus within Vietnam. **Journal of General Virology**, v. 85, n. 12, p. 3607–3618, 2004.

DIEU, B. T. M. et al. Evaluation of white spot syndrome virus variable DNA loci as molecular markers of virus spread at intermediate spatiotemporal scales. **Journal of General Virology**, v. 91, n. 5, p. 1164–1172, 2010.

DRUMMOND, A. J. et al. Bayesian phylogenetics with BEAUti and the BEAST 1.7. **Molecular Biology and Evolution**, v. 29, n. 8, p. 1969–1973, 2012.

EDDY, S. R. Hidden Markov models. **Current Opinion in Structural Biology**, v. 6, n. 3, p. 361–365, 1996.

EDDY, S. R. Accelerated profile HMM searches. **PLoS Computational Biology**, v. 7, n. 10, 2011.

ESCOBEDO-BONILLA, C. M. et al. Review A review on the morphology, molecular characterization, morphogenesis and pathogenesis of white spot syndrome virus. **Journal of Fish Diseases**, v. 31, p. 1–18, 2008.

ESWAR, N. et al. **Comparative protein structure modeling using Modeller**. [s.l: s.n.]. v. Chapter 5
FAO. **The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals**. Rome: [s.n.].

FERRON, F. et al. The modeled structure of the RNA dependent RNA polymerase of GBV-C virus suggests a role for motif E in Flaviviridae RNA polymerases. **BMC Bioinformatics**, v. 6, p. 1–16, 2005.

FILHO, J. C. **Produção aquícola 2016 – Dados do IBGE mostram que a piscicultura cresceu 4,94% em 2016 enquanto a carcinicultura, afetada pela mancha branca, teve uma queda de 26,09%**. Disponível em: <<https://panoramadaaquicultura.com.br/producao-aquicola-2016-dados-do-ibge-mostram-que-a-piscicultura-cresceu-494-em-2016-enquanto-a-carcinicultura-afetada-pela-mancha-branca-teve-uma-queda-de-2609/>>.

FISHER, R. D. et al. Structure and ubiquitin binding of the ubiquitin-interacting motif. **Journal of Biological Chemistry**, v. 278, n. 31, p. 28976–28984, 2003.

FLEGEL, T. W. et al. Shrimp Disease Control : Past , Present and Future. **Diseases in Asian Aquaculture VI**, p. 355–378, 2008.

FLEGEL, T. W. Historic emergence, impact and current status of shrimp pathogens in Asia. **Journal of Invertebrate Pathology**, v. 110, n. 2, p. 166–173, 2012.

GALAVÍZ-SILVA, L. et al. White spot syndrome virus genetic variants detected in Mexico by a new multiplex PCR method. **Aquaculture**, v. 242, n. 1–4, p. 53–68, 2004.

GANGULY, B.; RASTOGI, S. K. Structural and functional modeling of viral protein 5 of Infectious Bursal Disease Virus. **Virus Research**, v. 247, n. January, p. 55–60, 2018.

GAO, M. et al. White spot syndrome virus strains of different virulence induce distinct immune response in *Cherax quadricarinatus*. **Fish & Shellfish Immunology**, v. 39, n. 1, p. 17–23, 2014.

GUDKOV, N.; MURWANTOKO, I.; WALKER, P. J. Stability of the WSSV ORF94 VNTR genotype

- marker during passage in marine shrimp, freshwater crayfish and freshwater prawns. **Diseases of Aquatic Organisms**, v. 111, n. 3, p. 249–257, 2014.
- HAN, F.; XU, J.; ZHANG, X. Characterization of an early gene (wsv477) from shrimp white spot syndrome virus (WSSV). **Virus Genes**, v. 34, n. 2, p. 193–198, 2007.
- HERBERT, M. H.; SQUIRE, C. J.; MERCER, A. A. Poxviral ankyrin proteins. **Viruses**, v. 7, n. 2, p. 709–738, 2015.
- HOA, T. T. et al. Genotypic variations in tandem repeat DNA segments between ribonucleotide reductase subunit genes of White Spot Syndrome Virus (WSSV) isolates from Vietnam. **Diseases in Asian Aquaculture V**, n. May 2014, p. 339–351, 2005.
- HOA, T. T. T. et al. Mixed-genotype white spot syndrome virus infections of shrimp are inversely correlated with disease outbreaks in ponds. **Journal of General Virology**, v. 92, n. 3, p. 675–680, 2011.
- HOFMANN, K.; FALQUET, L. A ubiquitin-interacting motif conserved in components of the proteasomal and lysosomal protein degradation systems. **Trends in Biochemical Sciences**, v. 26, n. 6, p. 347–350, 2001.
- HUANG, C. Proteomic Analysis of Shrimp White Spot Syndrome Viral Proteins and Characterization of a Novel Envelope Protein VP466. **Molecular & Cellular Proteomics**, v. 1, n. 3, p. 223–231, 2002.
- HUANG, J. et al. An ankyrin repeat-containing protein, characterized as a ubiquitin ligase, is closely associated with membrane-enclosed organelles and required for pollen germination and pollen tube growth in lily. **Plant physiology**, v. 140, n. 4, p. 1374–1383, 2006.
- HUANG, T.; XU, D.; ZHANG, X. Characterization of host microRNAs that respond to DNA virus infection in a crustacean. **BMC Genomics**, v. 13, n. 1, 2012.
- HUANG, T.; ZHANG, X. Functional Analysis of a Crustacean MicroRNA in Host-Virus Interactions. **Journal of Virology**, v. 86, n. 23, p. 12997–13004, 2012.
- JOHNSON, L. S.; EDDY, S. R.; PORTUGALY, E. Hidden Markov model speed heuristic and iterative HMM search procedure. **BMC Bioinformatics**, v. 11, 2010.
- KARUNASAGAR, I.; ABABOUC, L. Shrimp viral diseases, import risk assessment and international trade. **Indian Journal of Virology**, v. 23, n. 2, p. 141–148, 2012.
- KATOH, K.; ROZEWICKI, J.; YAMADA, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. **Briefings in Bioinformatics**, n. May, p. 1–7, 2017.
- KEARSE, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. **Bioinformatics**, v. 28, n. 12, p. 1647–1649, 2012.
- KELLY, L. A. et al. The Phyre2 web portal for protein modelling, prediction, and analysis. **Nature Protocols**, v. 10, n. 6, p. 845–858, 2015.
- KHOR, B. Y. et al. General overview on structure prediction of twilight-zone proteins. **Theoretical Biology and Medical Modelling**, v. 12, n. 1, p. 1–11, 2015.
- KIM, C. S. et al. An RNA-Binding Protein, hnRNP A1, and a Scaffold Protein, Septin 6, Facilitate Hepatitis C Virus Replication. **Journal of Virology**, v. 81, n. 8, p. 3852–3865, 2007.
- KONDO, Y. et al. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. **eLIFE**, v. 4, p. 1–19, 2015.
- KRUPOVIC, M.; DOLJA, V. V.; KOONIN, E. V. Plant viruses of the Amalgaviridae family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. **Biology Direct**, v. 10, n. 1, p. 12, 2015.
- KUPER, J. et al. Functional and structural studies of the nucleotide excision repair helicase XPD suggest a polarity for DNA translocation. **EMBO Journal**, v. 31, n. 2, p. 494–502, 2012.
- LI, F. et al. Identification of the immediate-early genes of white spot syndrome virus. **Virology**, v. 385, n. 1, p. 267–274, 2009.
- LI, F. et al. Comparative genomic analysis of three white spot syndrome virus isolates of different virulence. **Virus Genes**, 2016.

- LI, F. et al. Comparative genomic analysis of three white spot syndrome virus isolates of different virulence. **Virus Genes**, v. 53, n. 2, p. 249–258, 2017.
- LI, J.; MAHAJAN, A.; TSAI, M. D. Ankyrin repeat: A unique motif mediating protein-protein interactions. **Biochemistry**, v. 45, n. 51, p. 15168–15178, 2006.
- LIGHTNER, D. V. Virus diseases of farmed shrimp in the Western Hemisphere (the Americas): A review. **Journal of Invertebrate Pathology**, v. 106, n. 1, p. 110–130, 2011.
- LIN, F. et al. Identification of three immediate-early genes of white spot syndrome virus. **Archives of Virology**, v. 156, n. 9, p. 1611–1614, 2011.
- LIN, F. et al. Characterization of white spot syndrome virus immediate-early gene promoters. **Journal of General Virology**, v. 94, n. Pt 2, p. 387–392, 2013.
- LIU, H. et al. Europe PMC Funders Group Structure of the DNA repair helicase XPD. **Cell**, v. 133, n. 5, p. 801–812, 2012.
- LIU, W.-J. et al. Cloning, Characterization, and Phylogenetic Analysis of a Shrimp White Spot Syndrome Virus Gene That Encodes a Protein Kinase. **Virology**, v. 289, n. 2, p. 362–377, 2001.
- LIU, W. J. et al. The promoter of the white spot syndrome virus immediate-early gene WSSV108 is activated by the cellular KLF transcription factor. **Developmental and Comparative Immunology**, v. 49, n. 1, p. 7–18, 2015.
- LIU, Y. et al. Identification of a Novel Nonstructural Protein, VP9, from White Spot Syndrome Virus: Its Structure Reveals a Ferredoxin Fold with Specific Metal Binding Sites. **Journal of Virology**, v. 80, n. 21, p. 10419–10427, 2006.
- LOBLEY, A.; SADOWSKI, M. I.; JONES, D. T. pGenTHREADER and pDomTHREADER: New methods for improved protein fold recognition and superfamily discrimination. **Bioinformatics**, v. 25, n. 14, p. 1761–1767, 2009.
- LOVELL, S. C. et al. Structure validation by C alpha geometry: phi,psi and C beta deviation. **Proteins-Structure Function and Genetics**, v. 50, n. August 2002, p. 437–450, 2003.
- MARFORI, M. et al. Molecular basis for specificity of nuclear import and prediction of nuclear localization. **Biochimica et Biophysica Acta - Molecular Cell Research**, v. 1813, n. 9, p. 1562–1577, 2011.
- MARIS, C.; DOMINGUEZ, C. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. **Nucleic Acids Research**, v. 33, n. 12, p. 2118–2131, 2005.
- MARIS, C.; DOMINGUEZ, C.; ALLAIN, F. H. T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. **FEBS Journal**, v. 272, n. 9, p. 2118–2131, 2005.
- MARKS, H. et al. Fitness and virulence of an ancestral White Spot Syndrome Virus isolate from shrimp. **Virus Research**, v. 110, n. 1–2, p. 9–20, 2005a.
- MARKS, H. et al. Gene-expression profiling of White spot syndrome virus in vivo. **Journal of General Virology**, v. 86, n. 7, p. 2081–2100, 2005b.
- MARKS, H. et al. In silico identification of putative promoter motifs of White Spot Syndrome Virus. **BMC bioinformatics**, v. 7, p. 309, 2006.
- MARTIN-TUMASZ, S. et al. A novel occluded RNA recognition motif in Prp24 unwinds the U6 RNA internal stem loop. **Nucleic Acids Research**, v. 39, n. 17, p. 7837–7847, 2011.
- MIRDITA, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. **Nucleic Acids Research**, v. 45, n. D1, p. D170–D176, 2017.
- MULDER, N. J. et al. In silico characterization of proteins: UniProt, InterPro and Integr8. **Molecular Biotechnology**, v. 38, n. 2, p. 165–177, 2008.
- MULLER, I. C. et al. Genotyping of White spot syndrome virus (WSSV) geographical isolates from Brazil and comparison to other isolates from the Americas. **Diseases of Aquatic Organisms**, v. 88, n. 2, p. 91–98, 2010.
- NAGARAJAN, V.; ELASRI, M. O. Structure and function predictions of the Msa protein in

Staphylococcus aureus. **BMC Bioinformatics**, v. 8, n. SUPPL. 7, p. 1–9, 2007.

NEWSOME, T. P.; MARZOOK, N. B. Viruses that ride on the coat-tails of actin nucleation. **Seminars in Cell and Developmental Biology**, v. 46, p. 155–163, 2015.

NOEL, E. A. et al. Chlorovirus Skp1-Binding Ankyrin Repeat Protein Interplay and Mimicry of Cellular Ubiquitin Ligase Machinery. **Journal of Virology**, v. 88, n. 23, p. 13798–13810, 2014.

Oakey, H. J.; Smith, C. S. Complete genome sequence of a white spot syndrome virus associated with a disease incursion in Australia. **Aquaculture**, v. 484, n. August 2017, p. 152–159, 2018.

OLIVEIRA, G. P. et al. The Investigation of Promoter Sequences of Marseilleviruses Highlights a Remarkable Abundance of the AAATATTT Motif in Intergenic Regions Oliveira, G. P., Lima, M. T., Arantes, T. S., Assis, F. L., Rodrigues, R. A. L., da Fonseca, F. G., ... Abrahão, J. S. (. **Journal of Virology**, v. 91, n. 21, p. e01088-17, 2017.

OTOMO, T. et al. Structural basis of actin filament nucleation and processive capping by a formin homology 2 domain. **Nature**, v. 433, n. 7025, p. 488–494, 2005.

OTOMO, T. et al. Crystal structure of the formin mDia1 in autoinhibited conformation. **PLoS ONE**, v. 5, n. 9, p. 1–13, 2010.

PARRA, R. G. et al. Structural and Energetic Characterization of the Ankyrin Repeat Protein Family. **PLoS Computational Biology**, v. 11, n. 12, p. 1–20, 2015.

PAVESI, G.; MAURI, G.; PESOLE, G. In silico representation and discovery of transcription factor binding sites. **Briefings in bioinformatics**, v. 5, n. 3, p. 217–236, 2004.

PETREY, D.; HONIG, B. Protein structure prediction: Inroads to biology. **Molecular Cell**, v. 20, n. 6, p. 811–819, 2005.

PRADEEP, B. et al. Genotyping of white spot syndrome virus prevalent in shrimp farms of India. **Diseases of Aquatic Organisms**, v. 78, n. 3, p. 189–198, 2008.

PRADEEP, B. et al. Biology, host range, pathogenesis and diagnosis of white spot syndrome virus. **Indian Journal of Virology**, v. 23, n. 2, p. 161–174, 2012.

RAMOS-PAREDES, J. et al. New genetic recombination in hypervariable regions of the white spot syndrome virus isolated from *Litopenaeus vannamei* (Boone) in northwest Mexico. **Aquaculture Research**, v. 43, n. 3, p. 339–348, fev. 2012.

REMMERT, M. et al. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. **Nature Methods**, v. 9, n. 2, p. 173–175, 2012.

REMMERT, M.; BIEGERT, A.; HAUSER, A. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. v. 9, n. 2, 2012.

RIDDICK, G.; MACARA, I. G. The adapter importin- α provides flexible control of nuclear import at the expense of efficiency. **Molecular Systems Biology**, v. 3, n. 118, p. 1–7, 2007.

RODRIGUEZ ANAY, L. et al. Analysis of repeated compound units in ORF94 of white spot syndrome virus isolated from *Litopenaeus vannamei* from outbreak and non outbreak shrimp farms in Sonora, Mexico. **Latin American Journal of Aquatic Research**, v. 44, n. 4, p. 850–854, 2016.

SAGE, V. LE; MOULAND, A. J. Viral Subversion of the Nuclear Pore Complex. p. 2019–2042, 2013.

SÁNCHEZ-PAZ, A. White spot syndrome virus: An overview on an emergent concern. **Veterinary Research**, v. 41, n. 6, 2010.

SCHÖNICHEN, A.; GEYER, M. Fifteen formins for an actin filament: A molecular view on the regulation of human formins. **Biochimica et Biophysica Acta - Molecular Cell Research**, v. 1803, n. 2, p. 152–163, 2010.

SEDGWICK, S. G.; SMERDON, S. J. The ankyrin repeat: A diversity of interactions on a common structural framework. **Trends in Biochemical Sciences**, v. 24, n. 8, p. 311–316, 1999.

SHEKAR, M.; PRADEEP, B.; KARUNASAGAR, I. White spot syndrome virus: Genotypes, epidemiology and evolutionary studies. **Indian Journal of Virology**, v. 23, n. 2, p. 175–183, 2012.

SIMROUNI, M. M. et al. Sequence analysis of ORF94 in different White Spot Syndrome Virus (WSSV) isolates of Iran. **Iranian Journal of Fisheries Sciences**, v. 13, n. 2, p. 492–502, 2014.

- SÖDING, J. Protein homology detection by HMM-HMM comparison. **Bioinformatics**, v. 21, n. 7, p. 951–960, 2005.
- SÖDING, J.; BIEGERT, A.; LUPAS, A. N. The HHpred interactive server for protein homology detection and structure prediction. **Nucleic Acids Research**, v. 33, n. SUPPL. 2, p. 244–248, 2005.
- SPEAR, M.; WU, Y. Viral exploitation of actin: Force-generation and scaffolding functions in viral infection. **Virologica Sinica**, v. 29, n. 3, p. 139–147, 2014.
- STENTIFORD, G. D. et al. Disease will limit future food supply from the global crustacean fishery and aquaculture sectors. **Journal of Invertebrate Pathology**, v. 110, n. 2, p. 141–157, 2012.
- TANG, K. F. J. et al. New genotypes of white spot syndrome virus (WSSV) and Taura syndrome virus (TSV) from the Kingdom of Saudi Arabia. **Diseases of Aquatic Organisms**, v. 99, n. 3, p. 179–185, 2012.
- TANG, K. F. J.; LE GROUMELLE, M.; LIGHTNER, D. V. Novel, closely related, white spot syndrome virus (WSSV) genotypes from Madagascar, Mozambique and the Kingdom of Saudi Arabia. **Diseases of Aquatic Organisms**, v. 106, n. 1, p. 1–6, 2013.
- THOMPSON, M. E. et al. FMNL3 FH2-actin structure gives insight into formin-mediated actin nucleation and elongation. **Nature Structural and Molecular Biology**, v. 20, n. 1, p. 111–118, 2013.
- TZENG, H. F. et al. Chimeric polypeptide of thymidine kinase and thymidylate kinase of shrimp white spot syndrome virus: Thymidine kinase activity of the recombinant protein expressed in a baculovirus/insect cell system. **Virology**, v. 299, n. 2, p. 248–255, 2002.
- VAN HULTEN, M. C. W. et al. Analysis of a genomic segment of white spot syndrome virus of shrimp containing ribonucleotide reductase genes and repeat regions. **Journal of General Virology**, v. 81, n. 2, p. 307–316, 2000.
- VAN HULTEN, M. C. W. et al. The White Spot Syndrome Virus DNA Genome Sequence. **Virology**, v. 286, n. 1, p. 7–22, 2001.
- VERBRUGGEN, B. et al. Molecular mechanisms of white spot syndrome virus infection and perspectives on treatments. **Viruses**, v. 8, n. 1, p. 1–29, 2016.
- VLIEGHE, D. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. **Nucleic Acids Research**, v. 34, n. 90001, p. D95–D97, 2006.
- WANG, Q.; NUNAN, L. M.; LIGHTNER, D. V. Identification of genomic variations among geographic isolates of white spot syndrome virus using restriction analysis and Southern blot hybridization. **Dis. Aquat. Org.**, v. 43, n. 3, p. 175–181, 2000.
- WANG, W. et al. Intermonomer Interactions in Hemagglutinin Subunits HA1 and HA2 Affecting Hemagglutinin Stability and Influenza Virus Infectivity. **Journal of Virology**, v. 89, n. 20, p. 10602–10611, 2015.
- WANG, Z. et al. RING-H2 Protein WSSV249 from White Spot Syndrome Virus Sequesters a Shrimp Ubiquitin-Conjugating Enzyme, P_{Ubc}, for Viral Pathogenesis RING-H2 Protein WSSV249 from White Spot Syndrome Virus Sequesters a Shrimp Ubiquitin-Conjugating Enzyme, P_{Ubc}, for . v. 79, n. 14, p. 8764–8772, 2005.
- WEBER, G. et al. Functional organization of the Sm core in the crystal structure of human U1 snRNP. **EMBO Journal**, v. 29, n. 24, p. 4172–4184, 2010.
- WONGTEERASUPAYA, C. et al. High variation in repetitive DNA fragment length for white spot syndrome virus (WSSV) isolates in Thailand. **Diseases of Aquatic Organisms**, v. 54, n. 3, p. 253–257, 2003.
- XU, Y. et al. Crystal structures of a formin homology-2 domain reveal a tethered dimer architecture. **Cell**, v. 116, n. 5, p. 711–723, 2004.
- YANG, F. et al. Complete Genome Sequence of the Shrimp White Spot Bacilliform Virus Complete Genome Sequence of the Shrimp White Spot Bacilliform Virus. **Journal of Virology**, v. 75, n. 23, p. 11811–11820, 2001.
- YONA, G.; LEVITT, M. Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. **Journal of Molecular Biology**, v. 315, n. 5, p. 1257–1275, 2002.

ZHANG, Y. I-TASSER server for protein 3D structure prediction. **BMC Bioinformatics**, v. 9, p. 1–8, 2008.

ZHENG, N.; SHABEK, N. Ubiquitin Ligases: Structure, Function, and Regulation. **<https://doi.org/10.1146/annurev-biochem-060815-014922>**, 2017.

APÊNDICES

APÊNDICE A – Motivos encontrados 200nt upstream das ORFs do WSSV

Nome	Início	<i>p-value</i>	Sítios
ORF96	63	2.30e-8	TCCTCCTCCTCC
ORF95	100	2.30e-8	TCCTCCTCCTCC
ORF78	77	2.30e-8	TCCTCCTCCTCC
ORF127	46	5.70e-8	TCCTCCTCCTTC
ORF90	53	2.53e-7	TCCTCCTCCTCT
ORF43	83	7.68e-7	TCCTCCTCACCC
ORF140	135	1.36e-6	TCCTCCTCCCTT
ORF162	187	1.69e-6	TCCTCCTCATTT
ORF68	119	1.69e-6	TTCTCCTCATTC
ORF66	111	1.69e-6	TCCTCCCCATTC
ORF114	115	2.01e-6	CCCTCCTCCTCT
ORF98	73	3.43e-6	CCCCCCCCCTCC
ORF89	61	4.66e-6	TCCCGCTCCCCC
ORF82	96	6.20e-6	TCCTCCCCATTT
ORF99	68	7.72e-6	TCCTCATCATCC
ORF63	7	7.72e-6	CCCCCCTCACCC
ORF142	77	9.83e-6	TCTCCCTCCTTT
ORF156	149	1.25e-5	ACCCCTCATCC
ORF91	20	1.56e-5	TTCTTCTCCTCT
ORF70	56	1.56e-5	TCCCGCACCTCC
ORF158	41	2.38e-5	TGCTGCTCCTTT
ORF34	49	2.38e-5	TGCTCCACCCTC
ORF29	61	2.38e-5	TCTCCACCTTC
ORF153	22	2.67e-5	TCCCCATCACCC
ORF64	152	2.67e-5	TCTCTCTCCTTC
ORF11	179	2.93e-5	TTTTCTCCCT
ORF30	64	3.23e-5	TTTGCTCCTTC
ORF120	21	3.54e-5	TTCTCCCCACTT
ORF71	106	3.54e-5	TTCCACCTCC
ORF121	51	3.87e-5	CCCTCCCGCTTC

ORF69	67	3.87e-5	TTTTCCTCATTT
ORF143	187	4.60e-5	AGCTCCTCCTTT
ORF171	185	5.06e-5	TTCTCATCATCT
ORF6	36	5.06e-5	TGCTGCCCATTC
ORF131	80	6.04e-5	TTTCTCTCCTTC
ORF147	133	6.52e-5	TTTTCCTGCTTC
ORF163	26	7.07e-5	TTCTTCTGCTTC
ORF27	97	7.07e-5	TGCTCACCCCCC
ORF72	114	8.22e-5	CCCCGCACCCCC
ORF24	16	8.87e-5	CTCCCCTGATCC
ORF100	98	9.53e-5	CTCTCCACATTT
ORF76	20	1.19e-4	TCCTTATCATTT
ORF113	151	1.36e-4	CTCTCAACCTTC
ORF139	18	1.45e-4	TCCTTCTTCTTC
ORF83	168	1.45e-4	AGTTCCTCCTCT
ORF54	166	1.55e-4	ACCTGCTCACTT
ORF181	34	1.75e-4	TCTCTCTCACTT
ORF149	68	1.98e-4	TTTCTCCCATTC

APÊNDICE B– Motivos encontrados 200nt upstream das ORFs do WSSV

Nome	Início	<i>p-value</i>	Sítios
ORF139	108	1.34e-7	GAGGAGGAAGA
ORF11	101	1.34e-7	GAGGAGGAAGA
ORF43	122	3.56e-7	GAGGAGGGGGA
ORF81	173	6.72e-7	GAGGAGCAGGA
ORF156	109	1.22e-6	GAGGAGGAACA
ORF119	66	1.44e-6	GAGGAGCGAGA
ORF141	66	1.95e-6	GAGGATGAAGA
ORF84	32	1.95e-6	AAGGAGGAAGA
ORF56	176	3.23e-6	GAGGAGGTGGA
ORF92	45	7.88e-6	GAAGGGGAGGA
ORF127	5	1.04e-5	GAAGATGAAGA
ORF159	47	1.17e-5	GAGGAAGAAGG
ORF150	181	1.17e-5	AAGGAGGAACA
ORF65	14	1.17e-5	GAGGAGCACGA
ORF38	169	1.60e-5	GAGGATGGAGG
ORF172	128	1.77e-5	GAGGAGGTGCA
ORF60	164	1.91e-5	AAAGAGGGAGA
ORF71	24	2.08e-5	AAGGAGGAAGT
ORF40	12	2.08e-5	GAGGACCAAGA
ORF180	18	2.36e-5	GAAGAAGGAGA
ORF121	129	2.64e-5	GAGGGGGGGGT
ORF64	131	2.64e-5	GAAGAGCAACA
ORF78	133	2.98e-5	GAAGGGGAGGG
ORF136	4	4.71e-5	GATGATGAAGA
ORF102	41	5.53e-5	AAGGAACAGGA

APÊNDICE C – Motivos encontrados 200nt upstream das ORFs do WSSV

Nome	Início	<i>p-value</i>	Sítios
ORF156	134	4.27e-7	CAACAACAACA
ORF140	184	4.27e-7	CAACAACAACA
ORF116	177	1.07e-6	CAACAACAAAA
ORF61	72	1.07e-6	CAACAACAAAA
ORF163	166	1.77e-6	CAACTACAACA
ORF79	179	1.77e-6	CAACTACAACA
ORF77	55	2.82e-6	CAACTACAAAA
ORF91	168	5.15e-6	CAACAAAAAAA
ORF75	49	5.15e-6	CAACAAAAAAA
ORF130	41	8.16e-6	CAACAACAAAC
ORF29	39	8.16e-6	CAACAACAAAC
ORF133	106	9.95e-6	CAACGACAACA
ORF98	187	9.95e-6	CAACAACCCCA
ORF93	132	1.26e-5	CAACTAAAAAA
ORF60	10	1.52e-5	AAACTACAACA
ORF178	35	1.82e-5	AAACAACCAAA
ORF74	138	1.82e-5	AAACAACCAAA
ORF43	2	1.86e-5	CAACATCAACA
ORF146	167	2.11e-5	CAACTAACAAA
ORF66	36	2.72e-5	AAACAACACAA
ORF99	8	2.91e-5	CAACAAACCCA
ORF129	187	3.54e-5	CAACAAAAATA

APÊNDICE D – Motivos encontrados 200nt upstream das ORFs do WSSV

Nome	Início	p-value	Sítios
ORF120	186	2.15e-6	AAAAAATAAAA
ORF107	107	2.15e-6	AAAAAATAAAA
ORF28	172	2.15e-6	AAAAAATAAAA
ORF13	122	3.58e-6	ACAAAATAAAA
ORF136	111	5.68e-6	ATAAAATAAAA
ORF126	27	5.68e-6	ATAAAATAAAA
ORF15	132	5.68e-6	ATAAAATAAAA
ORF115	3	9.25e-6	AAATAATAAAA
ORF113	8	1.14e-5	TAAAAATAAAA
ORF118	137	1.51e-5	ACATAATAAAA
ORF35	58	1.51e-5	ACAAAACAAAA
ORF150	46	1.65e-5	TCAAATAAAA
ORF149	148	2.08e-5	ATAAACAAAA
ORF114	76	2.84e-5	AAACAATAAAA
ORF90	162	3.19e-5	TAAAACAAAA
ORF112	41	3.60e-5	AAATAATAGAA
ORF152	190	3.97e-5	TCATAATAAAA
ORF148	139	3.97e-5	TCAAACAAAA
ORF96	113	4.48e-5	ATACAATAAAA
ORF74	117	4.48e-5	ATACAATAAAA
ORF54	117	4.48e-5	ATATAACAAAA
ORF184	186	5.50e-5	ATATAATAGAA
ORF132	132	6.45e-5	AGAAAATAGAA
ORF108	132	6.86e-5	TAAAACAGAA
ORF56	187	6.86e-5	TAAAACAGAA
ORF155	114	7.26e-5	CAAAAATAAAA
ORF7	179	7.26e-5	TCATAATAGAA
ORF163	134	8.44e-5	AAATAATAAAG
ORF111	143	8.44e-5	AAAAACAAAG
ORF40	44	9.24e-5	TGATAATAAAA
ORF4	121	9.88e-5	AAAAATATAA

ORF27 61 1.05e-4 **ACAACATAAAA**

APÊNDICE E – Motivos encontrados 200nt upstream das ORFs do WSSV

Nome	Início	<i>p-value</i>	Sítios
ORF178	68	1.17e-4	ATATTGTT
ORF173	73	1.17e-4	ATATTTTT
ORF172	65	1.17e-4	ATATTGTT
ORF162	34	1.17e-4	ATATTTTT
ORF155	24	1.17e-4	ATATTTTT
ORF127	52	1.17e-4	ATATTTTT
ORF118	75	1.17e-4	ATATTTTT
ORF99	84	1.17e-4	ATATTTTT
ORF96	36	1.17e-4	ATATTTTT
ORF68	74	1.17e-4	ATATTTTT
ORF61	24	1.17e-4	ATATTTTT
ORF54	6	1.17e-4	ATATTTTT
ORF43	53	1.17e-4	ATATTTTT
ORF38	54	1.17e-4	ATATTTTT
ORF31	42	1.17e-4	ATATTGTT
ORF2	82	1.17e-4	ATATTTTT
ORF147	30	1.94e-4	ATATTTTC
ORF71	1	1.94e-4	ATATTTTC
ORF70	3	1.94e-4	ATATTTTC
ORF53	43	1.94e-4	ATATTTTC
ORF32	83	1.94e-4	ATATTTTC
ORF16	9	1.94e-4	ATATTTTC
ORF9	50	1.94e-4	ATATTTTC
ORF182	45	4.18e-4	ATTTTGTT
ORF153	87	4.18e-4	ATTTTGTT
ORF119	20	4.18e-4	TTATTGTT
ORF110	77	4.18e-4	TTATTTTT
ORF106	15	4.18e-4	ATTTTTTT
ORF101	51	4.18e-4	ATTTTTTT
ORF97	54	4.18e-4	TTATTTTT

ORF92	71	4.18e-4	TTATTGTT
ORF84	68	4.18e-4	ATTTTTTT
ORF65	7	4.18e-4	ATTTTGTT
ORF40	23	4.18e-4	ATTTTTTT
ORF39	45	4.18e-4	ATTTTTTT
ORF36	89	4.18e-4	ATTTTTTT
ORF28	44	4.18e-4	TTATTTTT
ORF13	53	4.18e-4	TTATTGTT
ORF164	2	5.65e-4	TTATTTTC
ORF134	31	5.65e-4	ATTTTTTC
ORF130	59	5.65e-4	TTATTGTC
ORF120	76	5.65e-4	TTATTTTC
ORF100	6	5.65e-4	ATTTTGTC
ORF90	22	5.65e-4	ATTTTGTC
ORF59	30	5.65e-4	TTATTTTC
ORF30	38	5.65e-4	TTATTTTC
ORF7	6	5.65e-4	ATTTTGTC
ORF6	28	5.65e-4	ATTTTGTC
ORF83	19	6.39e-4	ATATTTTG
ORF62	39	6.39e-4	ATATTGTG
ORF159	68	7.09e-4	ATGTTTTT
ORF145	47	7.09e-4	ATGTTGTT
ORF114	4	7.09e-4	ATGTTTTT
ORF34	10	7.09e-4	ATGTTGTT
ORF166	62	8.63e-4	TTTTTGTT
ORF158	39	8.63e-4	TTTTTGTT
ORF148	89	8.63e-4	ATGTTTTC
ORF136	32	8.63e-4	ATGTTGTC
ORF183	30	9.33e-4	TTTTTGTC
ORF169	88	9.33e-4	TTTTTGTC
ORF168	57	1.20e-3	ATTTTG TG
ORF132	24	1.20e-3	TTATTG TG
ORF126	53	1.20e-3	ATTTTTTG
ORF113	23	1.20e-3	ATTTTG TG

ORF105	84	1.20e-3	ATTTTTTG
ORF76	66	1.20e-3	ATCTTTTC
ORF15	5	1.20e-3	TTATTTTG
ORF115	30	1.26e-3	TTGTTTTT
ORF35	48	1.26e-3	TTGTTTTT
ORF11	75	1.26e-3	TTGTTTTT
ORF109	71	1.31e-3	TTGTTGTC
ORF128	74	1.42e-3	TTCTTTTT
ORF179	27	1.54e-3	TTCTTGTC
ORF177	3	1.54e-3	TTTTTGTC
ORF139	68	1.54e-3	TTCTTGTC
ORF111	85	1.54e-3	TTCTTGTC
ORF82	5	1.54e-3	TTTTTGTC
ORF152	67	1.58e-3	ATCTTTTG
ORF131	13	1.58e-3	ATCTTTTG
ORF116	59	1.67e-3	TTCTTTTG
ORF117	52	1.74e-3	GTATTTT
ORF55	10	1.86e-3	ATATTTGT
ORF26	19	1.86e-3	GTATTTTC
ORF150	35	2.16e-3	TTATTGGT
ORF42	15	2.25e-3	TTATTGTC

APÊNDICE F – Manuscrito submetido BMC Bioinformatics**BMC Bioinformatics**
FEDERAL UNIVERSITY OF RIO GRANDE DO NORTE**--Manuscript Draft--**

Manuscript Number: BINF-D-18-00740

Full Title: FEDERAL UNIVERSITY OF RIO GRANDE DO NORTE

Article Type: Research article

Section/Category: Structural analysis

Funding Information: CNPq

(409378/2016-0)

Prof. Daniel Lanza

Abstract:

Corresponding Author: Daniel Lanza

Universidade Federal do Rio Grande do Norte

BRAZIL

Corresponding Author Secondary

Information:

Corresponding Author's Institution: Universidade Federal do Rio Grande do Norte

Corresponding Author's Secondary

Institution:

First Author: Cayro Mendes

First Author Secondary Information:

Order of Authors: Cayro Mendes

Diego Teixeira

João Paulo Lima

Daniel Lanza

Order of Authors Secondary Information:

*Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation***NOTA: O ARTIGO DO TRABALHO NÃO FOI ANEXADO POR RAZÕES DE DIREITOS AUTORAIS**