ERIC ROBERTO GUIMARÃES ROCHA AGUIAR

Caracterização do viroma de animais e plantas através da análise do padrão e sequência de pequenos RNAs produzidos pela resposta do hospedeiro

BELO HORIZONTE - MG,
SETEMBRO DE 2015

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática

TESE DE DOUTORADO

Caracterização do viroma de animais e plantas através da análise do padrão e sequência de pequenos RNAs produzidos pela resposta do hospedeiro

Tese de doutorado apresentada ao Curso de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito para a obtenção do grau de Doutor em Bioinformática.

Orientado: Eric Roberto Guimarães Rocha Aguiar
Orientador: Dr. João Trindade Marques

Belo Horizonte,
Setembro de 2015

“A ciência consiste em substituir o saber que parecia seguro por uma teoria, ou seja, por algo problemático”.

José Ortega y Gasset

Agradecimentos

Agradeço ao Prof. João Marques por ter me aceitado no seu laboratório, pelos ensinamentos e paciência durante o meu doutoramento;

Agradeço à Profa. Fabíola pelas inúmeras contribuições e pelo agradável convívio no laboratório;

Não poderia deixar de agradecer aos meus colegas do Laboratório de RNAi pelas colaborações científicas e pelos laços de amizade formados ao longo destes 4 anos de convivência.

Agradeço ao Programa de Pós-Graduação em Bioinformática, principalmente o Prof. Vasco e as espetaculares Sheila e Natália, por sempre ajudarem tão prestativamente;

Agradeço ao grupo do CEBio que foi minha porta de entrada para a ciência e onde fui apresentado à Bioinformática, principalmente ao Guilherme, Ângela, Adhemar, Sara, Anderson, Fausto, Fabiano, Francislon e Laura.

Ao Prof. Jean-Luc, do CNRS-UPR9022 na França, que me recebeu e orientou durante o doutorado sanduíche, e à Carine, Akira, Simona, Francesco, Olivier, Estelle, Alice, Laurent que sempre estiveram dispostos a discutir e ajudar. Ao Laurent Trouxler pelas inúmeras discussões sobre Bioinformática. Agradeço também ao Florian, Magda, Christelle, Grabriela, e Valérie, pela agradável convivência e ajudas durante minha estadia na França.

Agradeço à minha família, especialmente a meu pais Seu Roberto e Dona Liu, principais incentivadores da minha jornada. A minha tia Rozeângela e meu irmão Roberto Filho juntamente com Mayzita que sempre foram grandes exemplos e estiveram sempre presentes me incentivando durante o andar do meu doutoramento. Ao meu sobrinho João que foi mais um incentivo na minha caminhada. Ao seu Rubem que sempre me apoiou e incentivou na minha carreira acadêmica. Agradeço carinhosamente a minha noiva Anna Rúbia, que foi extremamente compreensiva, me apoiando e incentivando durante esses quatro anos de ‘letrinhas verdes na tela preta parecendo com matrix’ passando no meu computador.

Sumário

Lista de figuras

Lista de tabelas

Lista de abreviaturas

AaDV2 – *Aedes aegypti densovírus 2*
Aae – *Aedes aegypti*
AIDS – Sindrome da Imunodeficiência Adquirida
cDNA – DNA complementar
CFAV – *Cell fusion agent virus*
CHIKV – *Chikungunya vírus*
DCV – *Drosophila C virus*
DENV – *Dengue virus*
DRV – *Drosophila reovirus*
DUV – *Drosophila uncharacterized virus*
Dme – *Drosophila melanogaster*
DNA – ácido desoxirribonucleico
dNTP – dideoxinucleotídeos
dsDNA – RNA de dupla fita
dsRBM – double stranded RNA binding
dsRNA – RNA de dupla fita
EMCV – *Encephalomyocarditis vírus*
FHV – *Flock house virus*
FPKM – fragmentos por kilobase de exons por milhão se sequências mapeadas
GLRaV-3 - *Grapevine leafroll-associated vírus*
HTV – *Humaitá-Tubiacanga vírus*
HPV – *Human Papillomavirus*
IFN – Interferon
IIV6 – *Insect iridescent vírus - 6*
ISGs – genes induzidos por interferon
Kb – kilobases
Llo – *Lutzomyia longipalpis*
LSV - *Laem Singh vírus*
LPRV1 – *Lutzomyia piaui reovírus 1*
LPRV2 – *Lutzomyia piaui reovírus 2*
LPNV – *Lutzomyia piaui nodavirus*
ML – máxima verossimilhança
MCV – *Mosquito caratinga virgavirus*
miRNA – microRNAs
MNV – *Mosquito nodavirus*
mRNA – RNA mensageiro
MXV – *Mosquito X virus*
NCBI – National Center for Biotechnology Information
nt – nucleotídeo(s)

ORF – *Open read frame*
PAGE – eletroforese em gel de poliacrilamida
PCR – reação em cadeia da polimerase
PCLV – *Phasi Charoen-like virus*
piRNA – piwi-*interacting* RNAs
RdRP – RNA polimerase dependente de RNA
RNA – ácido ribonucleico
RNAi – RNA de interferência
RPM – sequências por milhão
RNAse L – endoribonuclease L
rRNA – RNA ribossomal
RT-PCR – transcrição reversa seguida de PCR
SARS-CoV – *Severe acute respiratory syndrome coronavírus*
SGIV – *Singapore grouper iridovirus*
SINV - *Sindbis virus*
SINV-NoVB2 – *Sindbis vírus* condificante proteína B2 do *Nodamura virus*
SINV-GFP – *Sindbis vírus* recombinante com proteína GFP
siRNA – *small interfering RNAs*
TuMV – *Turnip mosaic virus*
WNV – *West Nile virus*
YFV – *Yellow Fever vírus*
vsiRNAs – viral *small interfering RNAs*
VSV – *Vesicular stomatitis virus*
ZIKV – Zika virus

Resumo

O monitoramento do conjunto de vírus, o viroma, presente em amostras biológicas é importante em estudos de biodiversidade e pode ter interesse para saúde humana e veterinária. O viroma de insetos vetores é de interesse especial para autoridades de saúde pública. A metagenômica a partir dos sequenciamentos em larga escala de pequenos e longos RNAs são estratégias utilizadas para detecção de vírus, embora as vantagens e desvantagens de cada uma delas ainda não tenham sido analisadas. Além disso, a identificação de sequências virais por metagenômica é limitada por análises de similaridade que utilizam bancos de dados de sequências previamente caracterizadas como referência. Neste projeto, foi desenvolvida uma estratégia para detecção de vírus a partir de bibliotecas de pequenos RNAs que utiliza não só a comparação de busca por similaridades contra bancos de dados, mas também um método que é independente de sequência. Este método utiliza o padrão de pequenos RNAs derivados de vírus e gerados pela resposta do hospedeiro, tais como a via de RNA de interferência. Através da utilização desta estratégia nós identificamos 7 vírus, dos quais seis representam novas espécies, que compõem o viroma de populações de laboratório de moscas da fruta e dois insetos vetores, mosquitos e flebotomíneos. Nós ainda comparamos diretamente o sequenciamento de pequenos e longos RNAs em amostras de mosquitos *Aedes* e observamos enriquecimento de sequências virais na fração de pequenos RNAs. Observamos também que o perfil de tamanho dos pequenos RNAs é uma assinatura única de cada vírus e pode ser utilizado para identificar sequências virais sem similaridade com sequências previamente depositadas nas bases de dados de referência. Além disso, observamos que o perfil de pequenos RNAs e assinaturas moleculares podem ser utilizadas para inferir tropismo viral nos ovários entre outros aspectos da biologia do vírus como a presença de inibidores para vias imunes. Nós também aplicamos a nossa estratégia para caracterização do viroma de mosquitos individuais coletados diretamente da natureza onde foi possível identificar 3 vírus, dois deles tendo sido identificados anteriormente nas populações de mosquitos de laboratório, PCLV e HTV, e o terceiro, MCV, representando uma nova espécie. Entretanto, enquanto o PCLV e HTV foram encontrados circulando nos mosquitos individuais em todas as estações, o MCV foi detectado somente no verão. A alta prevalência do PCLV e HTV sugere que estes vírus compõem o viroma residente dos mosquitos de campo enquanto o MCV seria um componente transiente. Com relação a ampla aplicabilidade da nossa estratégia, nós mostramos que a detecção de vírus utilizando pequenos RNAs pode ser aplicada a animais vertebrados, embora não de forma tão eficiente como para plantas e insetos.

Abstract

Surveillance of pool of virus, Virome, present in environmental samples are important to biodiversity studies and can greatly help to human and veterinary health. The virome of vector insects have special interest to public health authorities. Metagenomic from large-scale sequencing of small and long RNAs have been used to detect viruses, although advantages and disadvantages of each one have not been analyzed. Furthermore, the identification of viral sequences is limited by similarity searches against reference databases. Here, we developed a strategy to detect virus from small RNA libraries that take advantage of not only sequence similarity searches against reference database but also a sequence-independent strategy. This method relies on the production of virus-derived small RNAs by the host response such as the RNA interference pathway. Using this strategy, we identify 7 viruses, from which six likely represents new species that compose the virome of laboratory populations of fruit flies and two insect vectors, mosquitoes and sandflies. In Aedes, we compared sequencing of small and long RNAs and reveal that viral sequences are enriched in the small RNA fraction. We also noted that the small RNA size profile was a unique signature of each virus and could be used to identify novel viral sequences without known relatives in reference databases. Furthermore, we show that the small RNA profile could be used to infer viral tropism for ovaries among other aspects of virus biology. Additionally, we applied this strategy to individual mosquitoes caught directly from field where we identified 3 viruses, which two of them have been previously identified in *Aedes aegypti* laboratory populations, PCLV and HTV, and the third, MCV, likely representing a new specie. However, while PCLV and HTV where found circulating in individual mosquitoes in all seasons, MCV was restricted to summer. The high prevalence of PCLV and HTV suggests those virus are component of resident virome of wild mosquitoes while MCV would be a transient component. Regarding the broad application of our strategy, we showed that virus detection utilizing small RNAs could be applied to vertebrate animals although not as efficiently as to plants and insects.

1. Introdução

Os microrganismos são os seres vivos mais abundantes no meio terrestre. Apesar do seu tamanho microscópico, os microrganismos desempenham papel importante em diversos processos ecológicos, como a decomposição de matéria orgânica e inorgânica, e na produção de oxigênio. Os microrganismos também podem ser patogênicos e muitos são responsáveis por patologias que afetam plantas, animais e humanos. Em grande parte de estudos de biodiversidade de microrganismos presentes em diferentes ecossistemas, os vírus aparecem como entidades mais abundantes que bactérias e fungos (Carter and Saunders 2007).

1.1. Os vírus

Os vírus são extremamente abundantes e são caracterizados por uma diversidade genética extraordinária (Edwards and Rohwer 2005). Todos os vírus são parasitas intracelulares obrigatórios que dependem completamente do hospedeiro para se multiplicar ao contrário de bactérias e fungos que podem ter vida livre (Marques and Carthew 2007).

Diferente do que é visto para bactérias e fungos que armazenam o seu material genético na forma de DNA, os vírus podem armazenar suas informações genéticas tanto na forma de DNA quanto RNA. Além disso, nos vírus ambos os tipos de ácidos nucléicos podem ser encontrados na forma de fita simples ou dupla, segmentados ou não segmentados, e no caso do DNA podem ser circulares ou lineares (Baltimore 1971). Segundo Baltimore, ao considerar a síntese do RNA mensageiro do vírus e a replicação do genoma viral, podemos agrupar os vírus em sete classes distintas, mostradas na **Figura 1**.

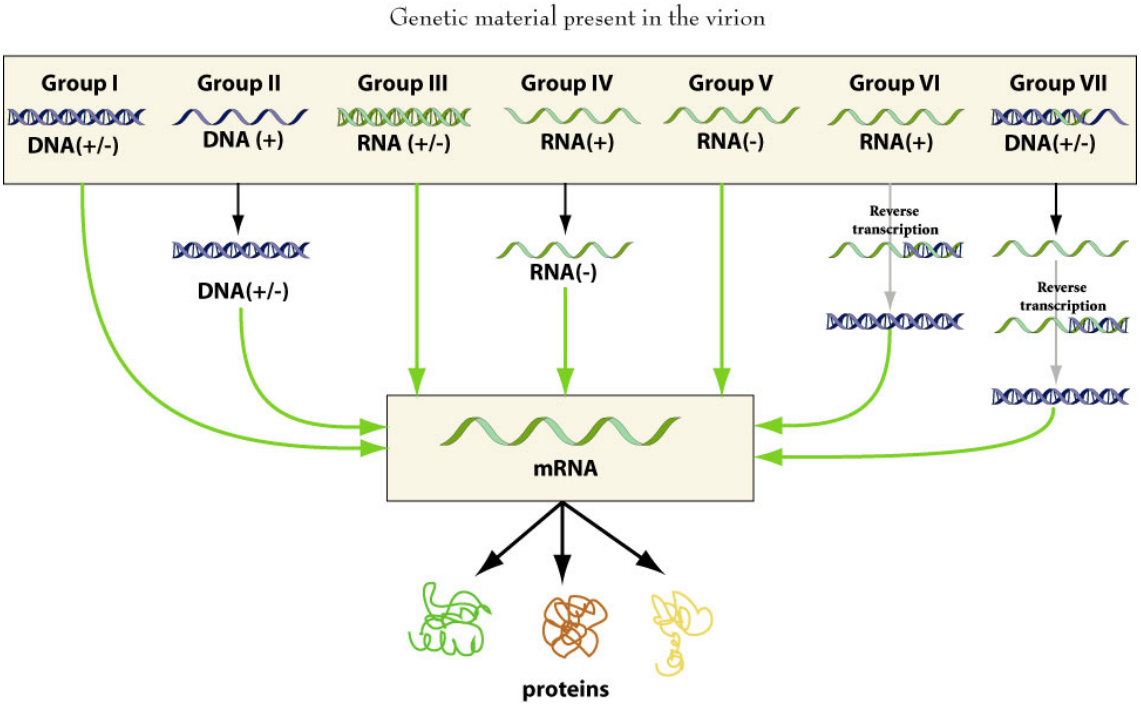

Figura 1. Classificação viral segundo Baltimore, baseada na síntese do mRNA viral e na replicação do genoma viral. (Fonte ViralZone – www.viralzone.org). De acordo com a classificação de Baltimore, o grupo I compreende os vírus de DNA de fita dupla, citando como exemplo os Adenovírus, Herpesvírus, Poxvírus e Rinovírus. No grupo II estão os vírus de DNA de fita simples seja de polaridade positiva ou negativa, como por exemplo os Parvovírus e Densovírus. No grupo III estão os vírus de RNA de fita dupla que compreendem as famílias *Reoviridae* e *Birnaviridae* por exemplo. No grupo IV se encontram os vírus de RNA de fita simples positiva, como os vírus das famílias *Picornaviridae*, *Nodaviridae*, *Flaviviridae*, *Coronaviridae, Potyviridae* e *Togaviridae*. O grupo V compreende os vírus de RNA de fita simples negativa, como os vírus das famílias *Bunyaviridae* e *Rhabdoviridae*. O grupo VI contém os vírus de RNA de fita simples que apresentam intermediário de DNA codificado através de transcriptase reversa, como por exemplo os Retrovírus. No grupo VII, classificado mais recentemente, estão contidos os vírus de DNA de fita simples que tem intermediário de RNA, como exemplo dos *Hepadnavírus*. Como característica comum, os vírus têm ou transcrevem, em algum momento da replicação viral, sequências de RNA que posteriormente vão ser traduzidas em proteína pela maquinaria celular.

Os vírus apresentam rápida taxa de multiplicação, que normalmente resulta em mutações, recombinações e rearranjos genéticos, que podem originar novas variantes virais. Estas variantes podem ainda ocorrer através da fusão de genomas de diferentes vírus quando em uma co-infecção. Esses eventos culminam com uma taxa maior de variações no genoma viral em comparação às variações no genoma do hospedeiro (Domingo, Escarmis et al. 1996, Sanjuan 2012). Como consequência da alta taxa de

mutação, alguns vírus podem rapidamente se adaptar e ganhar a capacidade de infectar novos hospedeiros (Gubler 2001, Parrish, Holmes et al. 2008, Longdon, Brockhurst et al. 2014). A variabilidade genética dos vírus pode explicar a ampla gama de organismos aos quais os vírus podem infectar, que vão desde hospedeiros unicelulares até plantas e mamíferos. Adicionalmente, alguns vírus que infectam plantas e mamíferos têm sido associados a inúmeras perdas econômicas na agricultura e na agropecuária, além causar grande impacto na saúde pública (Shepard, Undurraga et al. 2013, Cargnelutti, Olinda et al. 2014, Diez-Domingo, Perez-Yarza et al. 2014, Redinbaugh and Zambrano 2014).

1.1.1. Impacto econômico e social causado pelos vírus

Os vírus como parasitas intracelulares obrigatórios que ao invadirem as células de um organismo, prejudicam o funcionamento normal dessas células, alterando sua homeostase podendo levar o hospedeiro a patogênese. Em plantas, os vírus são responsáveis pela infecção e consequente diminuição da produção de culturas como a soja, sorgo, milho e arroz, levando a severas perdas econômicas (Nicaise 2014). Em animais de vida marinha de importância econômica como camarão, caranguejo e peixes, vírus de diversas famílias têm sido descritos causar infecção e patogênese aguda, levando a diminuição da produção (Bonami and Sri Widada 2011, Bonami and Zhang 2011). Na pecuária, exemplos como o *Vírus da estomatite vesicular* (VSV), *Bluetongue vírus sorotipo 8* e o recém descoberto *Schmallenberg vírus* tem preocupado pecuaristas na Europa devido sua disseminação e efeitos patogênicos em ruminantes (Cargnelutti, Olinda et al. 2014, Koenraadt, Balenghien et al. 2014). Em humanos, as doenças infeciosas, como infecções virais, são as principais causas de mortalidade ao redor do mundo. Entre as principais doenças virais que apresentam grande impacto na

saúde pública estão: gripe, AIDS, raiva, dengue, febre amarela, hepatite e herpes. Adicionalmente, existem relatos na literatura de associação entre o câncer cervical e o vírus do papiloma humano - HPV (Walboomers, Jacobs et al. 1999).

Alguns dos vírus causadores de doenças deletérias em humanos e plantas têm um ciclo de vida que envolvem mais de um hospedeiro, geralmente insetos vetores (Gubler 2001, Weaver and Barrett 2004, Bahder, Poojari et al. 2013, Nicaise 2014). O ciclo de transmissão destes vírus envolve a infecção produtiva de um hospedeiro intermediário, o inseto, e subsequente transmissão para o hospedeiro principal, reiniciando o ciclo. Como exemplo podemos citar o vírus *Grapevine leafroll-associated vírus 3* (GLRaV-3), vírus de maior incidência nos vinhedos do Brasil, e que causa o enrolamento das folhas das videiras (Vitis vinifera L.) (Fajardo, Eiras et al. 2005). Este vírus pode levar a diminuição de até 60% da produção de uvas, e pode ser transmitido pelo inseto *Pseudococcus maritimus* (Fajardo, Kuhn et al. 2002, Bahder, Poojari et al. 2013). Outro exemplo é a transmissão do *Dengue vírus* (DENV), *Chikungunya vírus* (CHIKV) e *Zika vírus* (ZIKV), transmitidos tanto no Brasil quanto em outros países tropicais e subtropicais principalmente pelo mosquito vetor *Aedes aegypti (Vijayakumar, George et al. 2013, Carrington and Simmons 2014, Diallo, Sall et al. 2014)*. Neste contexto, o estudo dos vírus, suas formas de transmissão, e sua interação com os respectivos e potenciais insetos vetores é de grande importância para o desenvolvimento de estratégias de prevenção e combate a vírus que causam impacto econômico e vírus com importância para saúde pública.

1.1.2. Arbovirus e arboviroses

Os arbovirus (do inglês “artrophode borne-viruses”), são vírus transferidos a animais vertebrados por artrópodes vetores (Weaver and Barrett 2004). Mais de 500 arbovirus

diferentes já foram descritos até o momento e este número é, provavelmente, uma subestimativa (Gubler 2001). Dentre os artrópodes, os mosquitos são vetores de um grande número de arbovirus que ameaçam a saúde pública humana em todo o mundo, como exemplos o DENV e *Yellow Fever vírus* (YFV) (Hill, Kafatos et al. 2005, Beaty, Prager et al. 2009). Somente no caso do DENV, em torno de 2.5 bilhões de pessoas habitam áreas com risco de infecções, se espalhando por mais de 100 países na Ásia, no Pacífico, nas Américas, na África e no Caribe (Gubler 2001, Weaver and Barrett 2004). O impacto econômico das arboviroses é enorme. Considerando somente as Américas entre os anos 2000 e 2007, as estimativas de gastos médios com a Dengue foram de 2.1 bilhões de dólares anuais (Shepard, Coudeville et al. 2011). Outro grande desafio é a emergência de arboviroses desconhecidas que poderiam, inesperadamente, adquirir a capacidade de infectar humanos e causar pandemias (Gubler 2001).

A maioria dos arbovirus têm um ciclo silvestre, envolvendo hospedeiros invertebrados e vertebrados não humanos, embora, em alguns casos, um ciclo urbano envolvendo transmissões humano-mosquito-humano ocorram (Gubler 2001, Weaver and Barrett 2004). Nas últimas décadas, o crescimento populacional, a urbanização e as mudanças no comportamento humano aumentaram a exposição da humanidade aos arbovirus. A incidência de arboviroses, previamente restrita a algumas áreas e estações, tem aumentado e se tornado menos previsível em todo o mundo (Gubler 2001, Shepard, Coudeville et al. 2011, Shepard, Undurraga et al. 2013). O Brasil é um bom exemplo do aumento da incidência de arboviroses causado, pelo menos em parte, pelo desmatamento e a crescente urbanização nas últimas décadas (Figueiredo 2007).

Nos anos 1950 e 1960, nas Américas, houve sucesso na redução da incidência de arboviroses, como Dengue e Febre Amarela. Essa redução temporária foi atingida através de estratégias de uso de inseticidas que tinham como alvo o mosquito vetor. Contudo, questões envolvendo toxicidades e resistências levaram ao abandono dessa

estratégia, o que causou retorno do mosquito e o restabelecimento do ciclo de transmissão dos arbovirus (Gubler 2001). Apesar dos problemas, essa experiência mostrou que o controle do vetor é parte importante no controle das arboviroses, embora estratégias mais sustentáveis devam ser desenvolvidas.

Na última década, um grande número de estudos, centralizados na biologia de mosquitos vetores tem permitido o desenvolvimento de novas estratégias como a modificação do comportamento ou alteração da interação parasito-hospedeiro através da manipulação genética ou intervenção química (Nene, Wortman et al. 2007, Beaty, Prager et al. 2009, Arensburger, Megy et al. 2010). A combinação de diferentes estratégias é certamente necessária para o controle de arboviroses em todo o mundo, já que nenhuma delas, isoladamente e em todos os casos, seria capaz de alcançar resultados positivos. Outro fator importante é que não existem vacinas eficazes para a vasta maioria das doenças causadas por arbovirus, o que torna a vigilância e prevenção elementos fundamentais no controle desses agentes patogênicos.

Assim, a detecção e isolamento desses vírus a partir dos insetos vetores é de suma importância para o desenvolvimento de estratégias de controle e combate. Contudo, o isolamento de vírus em laboratório é difícil, devagar, ineficiente e caro, o que faz a detecção direta de material genético em amostras ambientais por sequenciamento, metagenômica, a única estratégia para se estudar a diversidade genética dos vírus, ou o Viroma, em determinado ambiente (Quan, Briese et al. 2008).

1.2. Metagenômica

Inicialmente a microbiologia tradicional e o sequenciamento de genomas de microrganismos eram baseados no cultivo de culturas clonais e sequenciamento de marcadores gênicos de organismos presentes no ambiente, normalmente o gene 16S

rRNA, para analisar o perfil da diversidade numa amostra natural (Lane, Pace et al. 1985). Contudo, estudos mostraram que a ampla maioria da biodiversidade microbiana é perdida devido a métodos baseados em cultivo (Hugenholtz, Goebel et al. 1998). Assim, a metagenômica surgiu como uma importante estratégia que permite o estudo da diversidade genética sem a necessidade de se isolar os organismos em laboratório (Edwards and Rohwer 2005, Bishop-Lilly, Turell et al. 2010). Os primeiros relatos de clonagem de DNA diretamente extraído e purificado de amostras ambientais foi feito em comunidades microbiais marinhas por (Schmidt, DeLong et al. 1991), que fizeram o sequenciamento de clones do gene 16S para análise da diversidade de picoplâncton marinho. O pioneirismo desta pesquisa abriu a possibilidade de estudar a diversidade microbiana que poderia incluir organismos não cultiváveis.

O termo metagenômica foi primeiramente cunhado por Jo Handelsman, da Universidade de Winsconsin, nos Estados Unidos, que sugeriu uma série procedimentos para avaliar os organismos e seus metabólitos presentes no solo (Handelsman, Rondon et al. 1998). Sua proposta utilizava-se da clonagem de metagenomas como forma de avaliar a diversidade genética da microbiota do solo. Desde o trabalho do Dr, Handelsman, diversas estratégias para o estudo de metagenômica foram desenvolvidas e aprimoradas, como a fragmentação dos ácidos nucléicos para sequenciamento, utilização do sequenciamento de RNAs, além do desenvolvimento de novas tecnologias de sequenciamento em larga escala.

1.2.1. Estratégias de metagenômica

Os estudos iniciais da microbiota eram realizados a partir do sequenciamento de organismos cultivados em laboratório a partir de amostras da natureza, e as análises feitas através de Reações em Cadeia da Polimerase (PCR) (Lane, Pace et al. 1985).

Com a utilização de métodos independentes de cultivo, feitos a partir da extração, amplificação, e clonagem de genes diretamente de amostras ambientais, novas sequências que não pertenciam a organismos cultiváveis começaram a ser descobertas, indicando novas espécies (Schmidt, DeLong et al. 1991). Com o avanço da tecnologia e o desenvolvimento de novas estratégias, como uso da fragmentação aleatória do DNA e sequenciamento 'shotgun', além dos sequenciadores de nova geração, houve um aumento considerável no número de sequências geradas e da descoberta de novos organismos através da metagenômica (Tyson, Chapman et al. 2004, Venter, Remington et al. 2004).

É importante ressaltar que o material genético, que pode ser tanto DNA quanto RNA, pode ser isolado de qualquer tipo de amostra do meio ambiente para ser utilizado na construção de bibliotecas, representando compreensivamente a diversidade genética de todos os organismos presentes (Tringe, von Mering et al. 2005, Djikeng, Kuzmickas et al. 2009, Willner, Furlan et al. 2009). Diversos estudos têm aplicado com sucesso a estratégia de metagenômica para a caracterização da diversidade genética de ambientes como solo, oceanos e avaliação de microbiota de organismos como plantas, insetos e humanos (Tringe, von Mering et al. 2005, Djikeng, Kuzmickas et al. 2009, Kreuze, Perez et al. 2009, Willner, Furlan et al. 2009, Roossinck, Martin et al. 2015, Webster, Waldron et al. 2015). As análises metagenômica de vários tipos de amostras ambientais mostraram que uma percentagem significativa do total da variabilidade genética se deve a sequências de origem viral (Edwards and Rohwer 2005).

Semelhante aos estudos para a caracterização do microbioma, o sequenciamento massivo de DNA e RNA tem sido aplicado com sucesso para caracterização da diversidade genética dos vírus, chamado de viroma (Breitbart, Salamon et al. 2002). Entretanto, a maioria dos trabalhos para caracterização do viroma utiliza-se de algum tipo de manipulação das amostras antes do sequenciamento, tais

como centrifugação e filtragem por coluna, que são utilizados para propiciar o enriquecimento das sequências virais, apesar de algumas vezes estes procedimentos levarem a contaminação das amostras (Naccache, Greninger et al. 2013, Oude Munnink, Jazaeri Farsani et al. 2013, Tokarz, Williams et al. 2014). Assim, a direta extração de ácidos nucléicos de amostras com pouca ou nenhuma manipulação poderia diminuir a chance de contaminação, apesar de, em alguns casos, levar ao sequenciamento de uma maioria de sequências não virais na biblioteca (Li, Shi et al. 2015). Somado a isso, trabalhos recentes utilizando técnicas de sequenciamento de nova geração empregaram, para a descoberta de novos vírus, o sequenciamento de RNA tanto da fração de alto peso molecular quanto de baixo peso molecular selecionado a partir do RNA total (Willner, Furlan et al. 2009, Coffey, Page et al. 2014, Webster, Waldron et al. 2015).

Em geral, os trabalhos que se utilizam do sequenciamento de RNA empregam a estratégia de 'shotgun' para fragmentação do RNA que é depois montado em '*contigs'*, sequências contíguas geradas a partir da sobreposição de sequências menores, utilizados para caracterização baseada em análises de similaridade de sequências contra bancos de dados de referência (Kreuze, Perez et al. 2009, Wu, Luo et al. 2010, Zhuang, Zhang et al. 2014). Para a construção das bibliotecas utilizadas no sequenciamento, a ampla maioria dos trabalhos feitos até o momento utilizam RNAs de alto peso molecular – ou longos RNAs –, que compreendem RNAs do genoma viral, mRNAs virais, ou RNAs produzidos durante a replicação viral.

Mais recentemente, foi demonstrado o potencial da utilização da fração de pequenos RNAs, comumente produtos da resposta imune inata do hospedeiro, para descoberta de novos vírus em plantas e insetos, mas não em mamíferos e vertebrados (Kreuze, Perez et al. 2009, Wu, Luo et al. 2010). Contudo, já se sabe que em mamíferos as vias de resposta imune também podem gerar pequenos RNAs, como é o caso da via

de RNAi, que é conservada na maioria dos organismos vertebrados e invertebrados (Li, Lu et al. 2013, Maillard, Ciaudo et al. 2013). Adicionalmente, em mamíferos também já foi mostrado que a ação de algumas nucleases podem acarretar na degradação do genoma viral em pequenos fragmentos de RNAs (Girardi, Chane-Woon-Ming et al. 2013). Assim, sugerindo que a estratégia baseada no sequenciamento dos pequenos RNAs derivados de vírus poderia ser estendida também a mamíferos e possivelmente vertebrados.

1.3. Origem dos pequenos RNAs

Em insetos e na maioria dos animais existem pelo menos três diferentes vias de RNAi que estão envolvidas na biogênese de tipos distintos de pequenos RNAs, nomeados microRNAs (miRNAs), piwi-*interacting* RNAs (piRNAs) e pequenos RNAs de interferência (siRNAs) **(Ding 2010)**. Cada tipo de pequenos RNAs tem distribuição de tamanho e preferências de nucleotídeo únicas, relacionados com as suas respectivas vias de origem **Figura 2**. Vale destacar que o papel da via de RNAi na defesa antiviral não se limita aos insetos, se estendendo a diversos organismos, como plantas e mamíferos.

1.3.1 RNAi

Em invertebrados, as vias de RNAi, principalmente a via de siRNAs, desempenham um papel majoritário na resposta antiviral, respondendo tanto à infecção por vírus de RNA quanto DNA (Kemp, Mueller et al. 2013). Além disso, foi demostrado que siRNAs são gerados durante a infeção tanto por vírus fita simples ou dupla de RNA ou DNA (Kemp, Mueller et al. 2013, Webster, Waldron et al. 2015). A via de siRNAs é

ativada pelo reconhecimento de RNAs de fita dupla (dsRNA – *double stranded RNA*), que podem se originar tanto de intermediários de replicação em vírus com genoma de RNA quanto do próprio hospedeiro a partir mRNAs transcritos em pares antisenso ou longos grampos (**Figura 2B e C**). Adicionalmente, já foi descrito que siRNAs também são gerados durante a infeção por vírus de DNA, o que sugere a existência de um intermediário de fita dupla de RNA, apesar de o mecanismo de biogênese do dsRNA ser pouco claro (Kemp, Mueller et al. 2013).

O resultado da ativação da via de siRNA é a clivagem processiva dos dsRNAs, gerando pequenos RNAs com tamanho de 21 nucleotídeos (nt), que quando de origem viral são chamados de vsiRNAs (*viral small interfering RNAs*). Somado a isso, diversos estudos descreveram vírus que codificam proteínas que são capazes de inibir a via de siRNAs, como a proteína B2 sintetizada pelo *Flock house vírus* (Chao, Lee et al. 2005). Contudo, apesar da inibição, ainda são encontrados pequenos RNAs derivados do vírus, que podem ser produtos da via de siRNAs quando esta é inibida parcialmente, ou oriundos da degradação aleatória do genoma viral devido o acumulo de RNA viral (Han, Luo et al. 2011, Li, Lu et al. 2013).

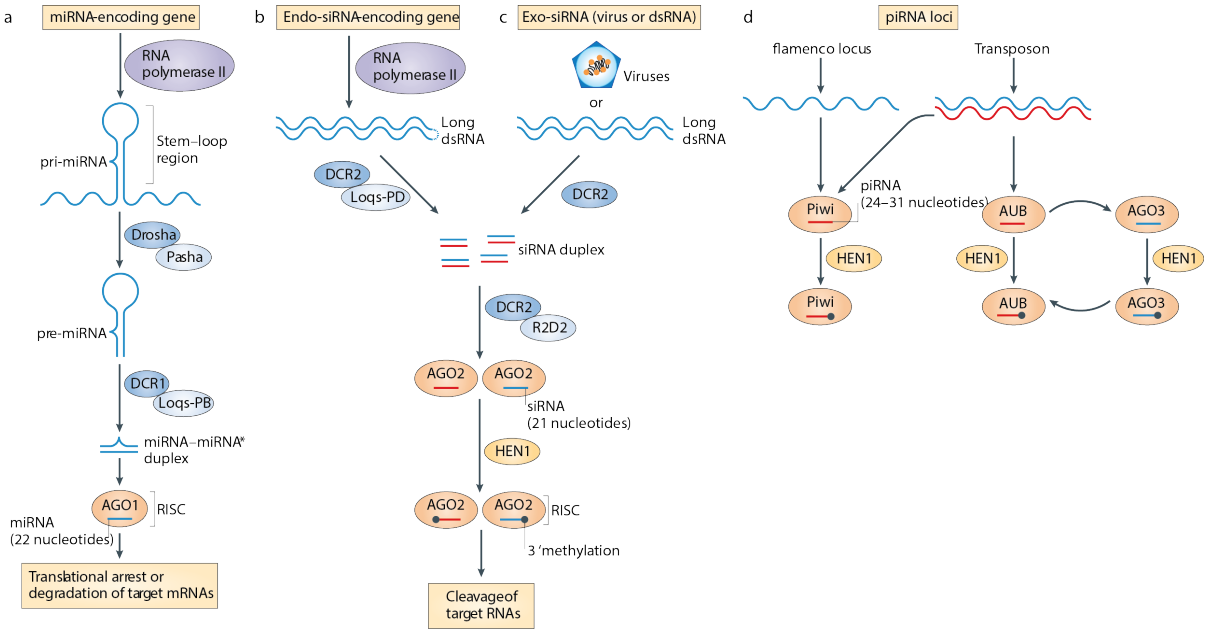

Figura 2 – Vias de RNA de interferência no modelo *D. melanogaster*. Em **A**, na via de microRNA após a transcrição pela RNA polimerase II de um gene codificador de miRNA, o transcrito primário é processado sequencialmente pelas RNAses tipo III Drosha no núcleo e Dicer-1 (DCR1) no citoplasma. Este processamento forma o precursor do miRNA (pre-miRNA) e então os miRNAs maduros que canonicamente possuem 22 nt de tamanho. O miRNA então liga-se na proteína Argonauta-1 (AGO1) e é formado o complexo de silenciamento induzido por RNA (RISC) que vai mediar o arraste traducional ou a degradação do mRNA alvo. Em **B** e **C**, a produção dos pequenos RNAs de interferência (siRNAs) a partir de longos transcritos auto-complementares ou transcritos complementares (via de endo-siRNA), ou dsRNAs exógenos ou virais (via de exo-siRNAs) são clivados pela RNAse tipo III processiva Dicer-2 (DCR2) gerando os pequenos RNAs com 21 nt de tamanho, que são então carregados na proteína Argonauta-2, metilados pela enzima HEN1, e então montado o complexo de silenciamento RISC. Em **C**, a biogênese dos piwiRNAs possivelmente não depende de uma RNAse tipo III. Os piRNAs ligam-se nas proteínas Piwi ou Aubergine (AUB), esta última colaborando com a Argonauta-3 (AGO3) para a amplificação dos piRNAs através de um mecanismo conhecido com ping-pong. Adaptado de Ding (2010).

Diferente da via de siRNAs, ainda se sabe pouco sobre os mecanismos de reconhecimento e ativação da via de piRNAs. Em insetos, a via de piRNA é principalmente ativada no controle de transposons em linhagens celulares germinativas, onde piRNAs primários e secundários podem ser gerados (Malone, Brennecke et al. 2009). No organismo modelo *Drosophila melanogaster*, os piRNAs primários são gerados a partir da clivagem de longos transcritos através de um processo endonucleolitico mediado pela proteína Zucchini. Posteriormente os piRNAs primários

são carregados nas proteína da família PIWI Aub/Ago3, dando origem ao ciclo de amplificação, referido como 'ping-pong', gerando os piRNAs secundários **Figura 2D** (Han, Wang et al. 2015). De forma interessante, em insetos os pequenos RNAs derivados da via de piRNA apresentam forte enriquecimento de Uracila na primeira base do piRNA primário e também um forte enriquecimento de Adenina na décima base do piRNA secundário, além da distância de 10 nt entre as posições 5' finais. É importante notar que alguns autores têm demonstrado que a via de piRNA também é ativada durante infecção viral em mosquitos e moscas, aonde especulam sobre a possibilidade de um papel antiviral nestes organismos (Wu, Luo et al. 2010, Morazzani, Wiley et al. 2012, Vodovar, Bronkhorst et al. 2012, Aguiar, Olmo et al. 2015).

Vale destacar que o papel da via de RNAi na defesa antiviral não se limita aos insetos e se estende a diversos organismos. O reconhecimento de dsRNA e geração de pequenos RNAs derivados do vírus foi mostrado ser conservado em insetos, plantas e mais recentemente também foi demonstrado em mamíferos (Wang, Wu et al. 2010, Li, Lu et al. 2013, Merkling and van Rij 2013). Contudo, na resposta à infecção viral em mamíferos, além das vias clássicas de RNAi, outras vias alternativas ou a ação de enzimas específicas têm sido descritas como geradoras de pequenos RNAs que podem ter função antiviral, como é o caso da ribonuclease L (RNAse L).

1.3.2. RNAse L

A ribonuclease L é uma endoribonuclease que tem sido amplamente estudada dado o seu papel na resposta antiviral contra diferentes vírus em mamíferos (Chakrabarti, Jha et al. 2011). Através do reconhecimento de padrões moleculares diferentes do que são encontrados no hospedeiro, como no caso de RNAs virais, os interferons são ativados culminando na expressão dos genes estimulados por interferon

(ISGs). Dentre os genes estimulados está o 2',5'-oligoadenylate synthetase 1 (OAS1), que é ativado através do reconhecimento de RNA de fita dupla de origem viral, levando a ativação da enzima RNAse L. Após sua ativação, a RNAse L, através do seu domínio ribonuclease, cliva todos os RNAs de fita simples presentes na célula, sejam eles de origem viral ou não viral, gerando pequenos fragmentos de RNA (Thompson, Kaminski et al. 2011, Nan, Nan et al. 2014), **Figura 3**.

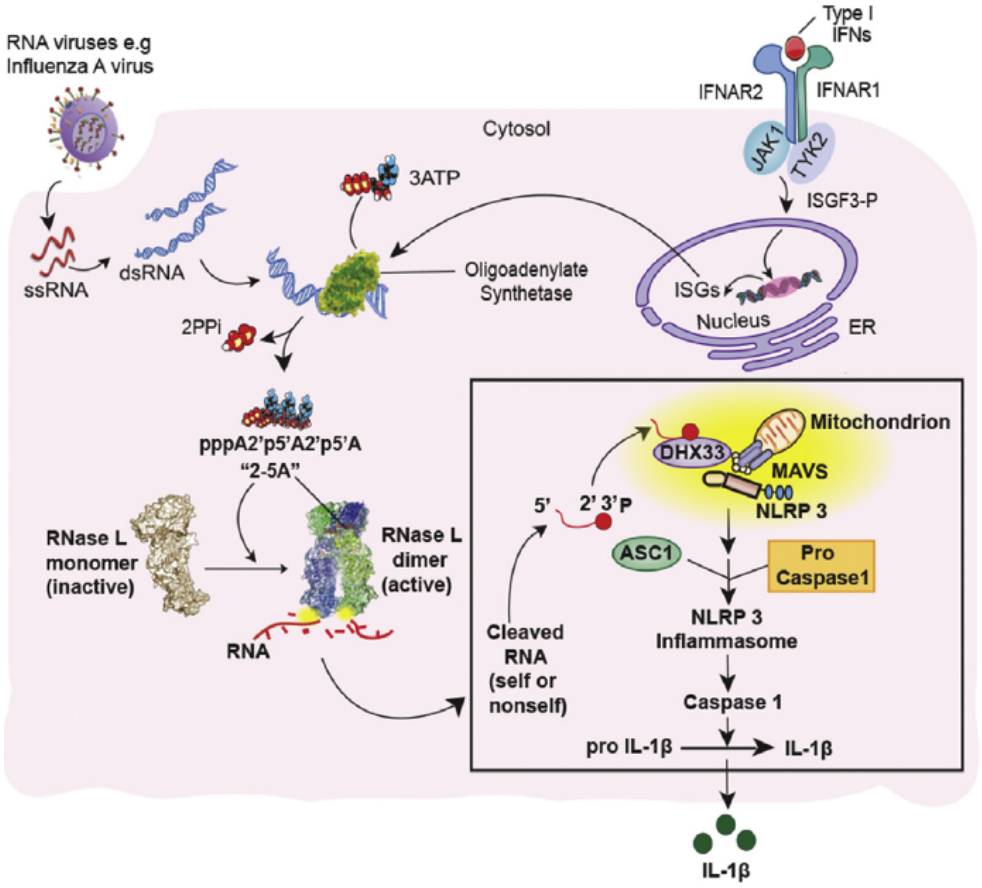

Figura 3. Mecanismo de ativação da RNAse L. Proteínas de reconhecimento de padrão do hospedeiro reconhecem padrões exógenos presentes em ácidos nucleicos de micróbios que levam a liberação de interferon. A liberação de interferon induz a transcrição de genes induzidos por interferon (ISGs), entre elas a OAS que é responsável pela ativação da RNAse L. A RNAse L ativada cliva todos RNAs presentes nas células gerando pequenos fragmentos de RNAs que podem ser reconhecidos por proteínas inflamatórias podendo levar a autofagia e apoptose. Adaptado de (Chakrabarti, Banerjee et al. 2015).

Adicionalmente, os pequenos RNAs resultantes da fragmentação do RNA viral podem ser reconhecidos por proteínas inflamatórias, levando a ativação de processos como apoptose e a autofagia, que compõe uma resposta imune complexa responsável pela restrição e combate a infecção viral (Zhou, Paranjape et al. 1997, Chakrabarti, Banerjee et al. 2015).

Assim, os pequenos RNAs gerados pelas vias de resposta imune do hospedeiro têm um papel fundamental em diversos organismos, desde insetos que utilizam a via de RNAi como principal componente da resposta imune, até os mamíferos que, além da via de RNAi conservada, contam com outros mecanismos de defesa que tem como intermediário pequenos RNAs derivados do vírus. Nesse contexto, o sequenciamento de pequenos RNAs pode ser utilizado para a detecção de infecções virais tanto em insetos e plantas quanto em mamíferos, que através de mecanismos compartilhados e distintos geram pequenos RNAs derivados do vírus. Estes pequenos RNAs, se clonados e sequenciados, podem ser utilizados como repórteres para a identificação dos vírus contidos numa amostra.

1.4. Utilização do sequenciamento de pequenos RNAs para detecção de vírus

Grande parte da resposta antiviral gerada pelas vias de resposta imune em insetos, plantas e vertebrados é passível de ser detectada através do sequenciamento de pequenos RNAs (Wang, Wu et al. 2010, Li, Lu et al. 2013, Marques, Wang et al. 2013). Contudo, além da estratégia de sequenciamento, existem outras etapas fundamentais para a identificação e caracterização das sequências virais, como a reconstrução do genoma viral através da montagem de sequências contíguas, estratégias de caracterização das sequências identificadas, dentre outras.

Devido a natureza da resposta da via de siRNAs, os vsiRNAs representam

sequências derivadas tanto da fita senso como da antisenso do genoma viral, independente da polaridade deste genoma (Myles, Wiley et al. 2008, Mueller, Gausson et al. 2010). Além disto, os vsiRNAs também são gerados por processamento relativamente uniforme de toda a extensão do genoma viral (Myles, Wiley et al. 2008, Marques, Kim et al. 2010, Mueller, Gausson et al. 2010). Estas características permitem a montagem de *contigs* correspondentes às sequências virais a partir dos vsiRNAs identificados por sequenciamento, demonstrado a partir do sequenciamento de bibliotecas de pequenos RNAs derivados de insetos e plantas infectados (Kreuze, Perez et al. 2009, Marques, Kim et al. 2010). Adicionalmente, já foi visto que além dos pequenos RNAs canônicos derivados das vias de RNAi, uma ampla gama de RNAs com tamanho variado, como os derivados da degradação aleatória do genoma viral, podem permitir a montagem de *contigs* derivados de vírus (Webster, Waldron et al. 2015). Contudo, apesar da possibilidade de montagem de *contigs*, métricas importantes que avaliam a qualidade da montagem como N50 (valor o qual todos os *contigs* com esse tamanho ou maiores contém no mínimo metade das bases de todos os *contigs* somados), tamanho do maior *contig* e número de *contigs* gerados são importantes para garantir o mínimo de confiabilidade das sequências montadas, e tornar possível a utilização de estratégias de caracterização da sequências (O'Neil and Emrich 2013).

Ainda é importante lembrar que a maioria absoluta dos trabalhos que utilizam o sequenciamento de pequenos ou longos RNAs aplicam análises baseadas em similaridade de sequência contra bancos de dados de referência para a caracterização das sequências encontradas (Kreuze, Perez et al. 2009, Wu, Luo et al. 2010, Tokarz, Williams et al. 2014, Zhuang, Zhang et al. 2014). Assim, o passo da montagem de *contigs* é extremamente importante para a caracterização das sequências, visto que o tamanho do *contigs* está diretamente relacionado com a possibilidade de encontrar

sequências relacionadas à vírus, e com a possibilidade de identificar falsos positivos (Gonzalez and Pearson 2010).

A identificação e caracterização de novas sequências virais é importante para a prevenção de epidemias na área de saúde pública, e também devido os genes virais serem únicos e potencialmente úteis para novas aplicações biológicas. As estratégias utilizadas até o momento, na maioria das vezes, é limitada à uns poucos vírus conhecidos (Edwards and Rohwer 2005). Estratégias complementares vêm sendo desenvolvidas para superar essa limitação, contudo nenhuma delas foi ainda padronizada e extensivamente avaliada para diferentes espécies. Assim, novas estratégias são necessárias para aprimorar a detecção de vírus e ajudar a caracterizar novas sequências desconhecidas, que comumente são encontradas em estudos de sequenciamento em larga escala, normalmente chamadas de matéria negra da metagenômica (Wu, Wu et al. 2011, Oh, Byrd et al. 2014).

2. Justificativa

O estudo da biodiversidade é importante ecológico, para podermos entender a evolução e adaptação das espécies, e do ponto de vista biotecnológico e médico para permitir o desenvolvimento de estratégias de uso ou controle de organismos. Os vírus contribuem com uma grande parte da diversidade genética observada em diversos ambientes. O estudo da biodiversidade é limitado pela capacidade de identificação de microrganismos em laboratório. Nesse contexto, a metagenômica permite o estudo da diversidade genética coletiva presente em um ecossistema através do sequenciamento direto do material genético isolado do meio ambiente. A metagenômica tem sido empregada em estudos para identificação do conjunto de vírus, o viroma, em amostras biológicas. A caracterização do viroma de insetos vetores tem uma particular significância para a saúde humana, visto que mosquitos e outros insetos podem transmitir diversos vírus, como o *Dengue vírus* e *Chikungunya vírus*, para a população humana. Contudo, a identificação de vírus através de estratégias de metagenômica têm uma limitação inerente, visto que a identificação da diversidade genética é feita através da busca de similaridade contra sequências depositadas em bancos de dados de referência. Desta forma, são perdidas as sequências virais ou de outra origem que apresentam alta divergência comparadas às referências caracterizadas e depositadas em bancos de dados. Novas estratégias são necessárias para melhorar a detecção de vírus e ajudar a caracterizar sequências desconhecidas comumente encontradas em estudos de sequenciamento em larga escala, chamadas também de ‘matéria negra’ da metagenômica. O desenvolvimento de novas estratégias que sejam mais eficientes para análise do viroma em amostras complexas seria de grande ajuda na identificação e caracterização de vírus circulantes na natureza.

3. Objetivos

3.1. Objetivo Geral

Desenvolver um pipeline de análise e caracterização do conjunto de vírus, o Viroma, presentes em diferentes organismos.

3.2. Objetivos específicos

1- Desenvolver um pipeline para a montagem e caracterização de sequências virais a partir de bibliotecas de RNA comparando a eficiência das frações de pequenos e longos RNAs;

2- Caracterizar o viroma de diferentes populações de insetos e avaliar como o padrão de pequenos RNAs pode ser utilizado para identificação de novas sequências virais;

3- Aplicar o pipeline desenvolvido para avaliar a dinâmica temporal/espacial do viroma em mosquitos de campo;

4- Avaliar a aplicabilidade da nossa estratégia de detecção de vírus em bibliotecas de pequenos RNAs derivadas de plantas e vertebrados;

4. Materiais e Métodos

4.1. Avaliação do melhor *k-mer* para montagem de *contigs*

Para avaliar qual seria o melhor *k-mer*, tamanho utilizado para dividir as sequências em pequenos blocos de sequência para posterior análise de sobreposição, as sequências após remoção de adaptadores e bases indefinidas foram mapeadas contra o genoma do hospedeiro permitindo 2 *mismatches*. As sequências que não apresentavam similaridade com o genoma do hospedeiro foram selecionadas e filtradas por tamanho de 21 nt. As sequências filtradas de 21 nt foram então montadas em sequências maiores obtidas a partir da sobreposição das *sequências* (*contigs*) utilizando diversos tamanho de *k-mer* entre 11 e 19 nt. Após a montagem, os tamanhos de *k-mer* foram avaliados segundo N50, tamanho do maior *contigs*, número de *contigs* derivados do vírus e cobertura do genoma viral, e o tamanho de *k-mer* com melhores métricas foi utilizado na montagem de *contigs*.

4.2. Montagem de *contigs*

As *sequências* oriundas do sequenciamento, após remoção de adaptadores e bases indefinidas, foram mapeadas contra os genomas de referência, utilizando *Bowtie* permitindo 2 *mismatches* para identificar as *sequências* que "não tinham origem conhecida". Em seguida, as *sequências* que não mapearam contra nenhuma referência foram utilizadas para a montagem *contigs* utilizando o *Velvet* (Zerbino and Birney 2008) utilizando *k-mer* 15, definido segundo avaliação do melhor *k-mer* na seção anterior. Para que as estratégias pudessem ser mensuradas, métricas como tamanho médio de *contigs*, cobertura e N50 foram utilizadas como sugerido por (Miller, Koren et al. 2010).

4.3. Definição de limiar de detecção de *contigs* virais

Para avaliar a influência das sequências virais e não virais na detecção de *contigs* virais foram separadas as *sequências* de 21 nt, após pré-processamento, que não apresentaram similaridade com sequências derivadas do hospedeiro. Essas sequências foram separadas em dois grupos distintos: aquelas *sequências* que apresentaram similaridade com o vírus, referidas como grupo sinal, e as *sequências* que não apresentaram similaridade com o vírus, referidas como grupo ruído. Foram realizadas diluições da quantidade de *sequências* de cada grupo nas proporções de (100, 80, 40, 20, 10, 5 e 2,5 %) e montados *contigs* os quais foram avaliados segundo métricas comuns de análise de montagem de *contigs*: N50, maior *contig*, média, *contigs* com similaridade com o genoma viral e cobertura do genoma viral. Após a montagem, os *contigs* eram mapeados contra o genoma do vírus utilizando o *Bowtie* permitindo 1 *mismatche* e a cobertura avaliada utilizando a suíte *BEDtools* (Quinlan and Hall 2010) e scripts próprios desenvolvidos em Perl e R.

Para buscar embasamento estatístico para as comparações, para cada diluição foram executadas 10 rodadas de montagem de *contigs* aonde em cada uma das rodadas as *sequências* utilizadas eram extraídas randomicamente do grupo sinal ou ruído. Os resultados eram consolidados utilizando a média das rodadas. Inicialmente foram mantidas constantes a quantidade de *sequências* do grupo sinal e diminuídas as *sequências* do grupo ruído. Posteriormente, foram mantidas as *sequências* do grupo ruído e decrementadas as *sequências* do grupo sinal. Assim pôde-se avaliar a influência do sinal e do ruído na montagem de *contigs*.

4.4. Processamento e extração de ácidos nucleicos

Mosquitos *Aedes aegypti* usados nos experimentos foram obtidos de colônias de laboratório estabelecidas de ovos coletados em três diferentes regiões do Rio de Janeiro, Humaitá, Tubiacanga e Belford Roxo, no sudoeste do Brasil. Os mosquitos *Aedes* capturados da natureza analisados individualmente foram coletados na cidade de Caratinga, Minas Gerais. As Colônias de laboratório de *Lutzomyia longipalpis* foram obtidas de animais capturados na natureza na cidade de Teresina – Piaui. As bibliotecas de *Drosophila melanogaster* foram preparadas a partir de moscas selvagens de estoques de laboratório que foram infectadas com *Vesicular stomatitis vírus*, *Drosophila C vírus* ou *Sindbis vírus* como descrito anteriormente (Galiana-Arnoux, Dostert et al. 2006). Insetos individuais ou grupos de insetos foram anestesiados com monóxido de carbono e diretamente macerados no Trizol utilizando '*beads'* de vidro. Ovários foram dissecados de mosquitos fêmeas e diretamente homogeneizados em Trizol usando pipetas. DNA ou RNA total foram extraídos usando Trizol de acordo com protocolo do fabricante (Invitrogen).

4.5. Construção das bibliotecas de RNAs

RNA total foi extraído de 3 distintos grupos de *Aedes*, *Lutzomyia* e *Drosophila* que foram utilizadas para a construção de bibliotecas independentes de pequenos RNAs. No caso de Aedes, o mesmo total de RNA foi utilizado para a construção de 3 bibliotecas de RNAs longos. Os pequenos RNAs foram selecionados por tamanho (~18-30 nt) em PAGE desnaturante antes de ser utilizado para construção das bibliotecas, como previamente descrito (Pfeffer, Zavolan et al. 2004). Para os mosquitos de campo, o RNA foi extraído de cada individuo independentemente, o qual foi utilizado para o

sequenciamento utilizando a mesma estratégia utilizada para as populações de insetos. As bibliotecas de RNAs longos foram construídas a partir de RNA total que foi enriquecido para polyA e realizado depleção de RNA ribosomal (rRNA) utilizando o kit '*TruSeq Stranded Total RNA kit*' de acordo com o protocol do fabricante (Illumina). O sequenciamento das bibliotecas foi realizado pelo 'IGBMC Microarray e Sequencing platform'. A estratégia de sequenciamento dos pequenos RNAs foi 1 X 50 pares de base (pb) e 2 X 100 pb (sequenciamento senso e antisenso) resultando em uma média de tamanho das sequências de 190 nt.

4.6. Pré-processamento das bibliotecas

4.6.1 SOLiD

A biblioteca de pequenos RNAs foi submetida a filtro de correção de erros utilizando o *SAET*, software disponibilizado pela Applied® Biosystems. Em seguida, a biblioteca foi submetida a filtro de qualidade utilizando o script *csfasta_quality_filter*, também disponibilizado pela Applied®, empregando os parâmetros –l 15, –t 2, –m 20 e –r 1 onde –l define o tamanho mínimo aceitável das sequência obtidas do sequenciamento (*sequências)* após a remoção de bases com qualidade ruim no final da sequência, -t número máximo de *sequências* ruins nas primeiras 10 bases, –m média de qualidade na read e –r define a possibilidade de remoção de bases na região terminal das *sequências* (trimagem) caso a qualidade esteja abaixo da aceitável segundo parâmetro -m. As *sequências* que passaram nos filtros de qualidade foram submetidas a filtro de *gaps* onde somente as *sequências* que não continham bases indefinidas (*gap)* foram utilizadas nas análises posteriores. Devido ao tamanho dos pequenos RNAs ,18-35 nucleotídeos (nt), a região terminal das *sequências* pode conter os adaptadores utilizados no processo de sequenciamento. Assim, inicialmente as bibliotecas de

pequenos RNAs foram submetidas a remoção de adaptadores utilizando o software *cutadapt* (Martin 2011) tomando como base os adaptadores descritos no *kit* utilizado para construção da biblioteca sequenciada.

4.6.2 Illumina

As sequências brutas originadas do sequenciamento de pequenos RNAs foram submetidas a filtro de qualidade de trimagem de adaptadores utilizando os scripts fastx_quality_filter e fastq_clipper respectivamente, programas inclusos no pacote fastx-toolkit (versão 0.0.14) (http://hannonlab.cshl.edu/fastx_toolkit/index.html). As sequências das bibliotecas de pequenos RNAs com baixa qualidade, menor phred 20, menores que 15 nt, após remoção dos adaptadores ou contendo bases ambíguas foram descartadas. No caso das bibliotecas de RNAs longos, as sequências brutas foram submetidas a filtro de qualidade utilizando o script fastx_quality_filter. Sequências com baixa qualidade, menor phred 20, ou contendo bases ambíguas foram descartadas. As sequências remanescente de pequenos ou longos RNAs foram então mapeadas contra sequências de referência oriundas de elementos transponíveis (TE), bactéria (2739 genomas completos depositados no Genbank) e genomas do hospedeiro (Lutzomyia longipalpis, Aedes aegypti, e Drosophila melanogaster) utilizando Bowtie (versão 1.1.1) para as bibliotecas de pequenos RNAs ou Bowtie2 (versão 2.2.4) para bibliotecas de RNAs longos ambos permitindo 1 mismatche (Langmead, Trapnell et al. 2009, Langmead and Salzberg 2012). O genoma de Drosophila (versão 5.44) foi obtido dos bancos de dados do flybase.org. A última versão do genomas de Aedes (Liverpool cepa L3) e Lutzomyia (Jacobina cepa J1) foram obtidos do VectorBase (www.vectorbase.org). Sequências de elementos transponíveis foram obtidos do TEfam (http://tefam.biochem.vt.edu/tefam). As sequências remanescentes, referidas como sequências processadas, foram utilizadas para montagem de *contigs* e análises posteriores.

4.7. Otimização da estratégia de montagem de *contigs*

As sequências processadas foram utilizadas para montagem de *contigs* utilizando o software Velvet (versão 1.0.13) (Zerbino and Birney 2008). A montagem de *contigs* foi realizada em paralelo utilizando estratégias diferentes para cada biblioteca. Para as bibliotecas de pequenos RNAs foi realizado a montagem utilizando diferentes tamanhos de sequências de pequenos RNAs (20-23, 24-30 e 20-30 nt). Em cada caso, estratégias paralelas de montagem foram empregadas onde foram utilizados valor de *k-mer* fixo (*k-mer* 15) com demais parâmetros padrão ou valor de *k-mer* (*k-mers* de 15 a 31 foram avaliados) automaticamente definido através do script VelvetOptimser (versão 2.2.5) (http://bioinformatics.net.au/software.velvetoptimiser.shtml). Para as bibliotecas de RNAs longos, a montagem de *contigs* foi realizada utilizando valor de *k-mer* fixo (*k-mer* 31) ou automaticamente definido (*k-mers* de 15 a 91 foram avaliados). Para cada biblioteca, os resultados de cada montagem executada em paralelo foram consolidadas utilizando CAP3 (data da versão 21-12-07) com os parâmetros *max gap length in overlap* definido para 2 e *overlap length cutoff* definido para 20 (Huang and Madan 1999). A remoção de *contigs* redundantes foi realizada utilizando o script BLASTClust que é distribuído no pacote BLAST (versão 4.0d) (Altschul, Gish et al. 1990), requerendo que no mínimo 50% do tamanho da sequência apresentasse no mínimo 50% de identidade com a sequência comparada. Os *contigs* não redundantes maiores que 50 nt receberam identificadores específicos para identificar sua origem e foram posteriormente caracterizados.

4.8. Caracterização de *contigs* baseada em sequência

Os *contigs* montados foram caracterizados segundo similaridade de sequência (nucleotídeo ou proteína), presença de ORFs e análise de domínios conservados. Foi utilizado o software BLAST para as buscas por similaridade de sequência contra os bancos de dados não redundantes de sequências (nucleotídeo e proteína) do National Center for Biotechnology Information (NCBI). Os softwares InterproScan (versão 5.3-46.0) (Mulder and Apweiler 2007) e HMMer (version 3.0) (Eddy 2009) foram utilizados para verificar a presença de ORFs e domínios conservados e o banco de dados do Pfam (versão 27.0) (Finn, Bateman et al. 2014) empregado na análise de domínios proteicos. *Contigs* que apresentassem valor de *e-value*, menor que 1e-5 para comparação em nucleotídeo ou 1e-3 para comparação em proteínas, foram considerados significantes. Os segmentos genômicos virais foram classificados como sugerido por (Ladner, Beitzel et al. 2014).

4.9. Análise do perfil de pequenos RNAs

Para as análises baseadas em padrão, as sequências processadas de pequenos RNAs foram mapeadas contra *contigs* ou genomas de referência utilizando o software Bowtie permitindo 1 *mismatche*. O perfil de tamanho dos pequenos RNAs foi calculado como a frequência do tamanho das sequências de pequeno RNA de 15-35 nt que foram mapeadas no genoma de referência ou *contig* avaliando cada polaridade separadamente. Foi utilizado Z-score para normalizar o perfil de tamanho de pequenos RNAs e produção de mapas de calor *heatmaps'* para cada *contig* ou sequência de referência através do R (versão 3.0.3) com o pacote gplots (versão 2.16.0). Para avaliar a relação entre os perfis de pequenos RNAs de diferentes *contigs* ou sequências de referência, foi calculada o valor de correlação de Pearson, com intervalo de confiança de 95%, dos valores de Z-score. As similaridades entre os perfis de pequenos RNAs

foram definidas utilizando agrupamento hierárquico com o emprego de UPGMA como critério de acoplamento. Grupos de sequências com mais de um elemento que apresentassem no mínimo 0,8 de valor de correlação entre todos os elementos do grupo eram definidos como clusters. A densidade de cobertura de pequenos RNAs foi calculada como o número de vezes que uma sequência de pequeno RNA cobriu cada nucleotídeo no genoma de referência ou *contig*. O perfil de tamanho e densidade de cobertura de pequenos RNAs foram calculados utilizando Perl (versão 5.12.4) juntamente com a biblioteca BioPerl (versão 1.6.923) e gráficos gerados utilizando R com o pacote ggplot2 (versão 1.0.1).

4.10. Análises filogenéticas e frequência de di-nucleotídeos

Sequências de nucleotídeo ou proteína foram escolhidas de acordo com similaridade de sequências utilizando BLAST e foram alinhadas utilizando Muscle (Edgar 2004), implementado no MEGA (Tamura, Stecher et al. 2013). O melhor modelo de substituição foi estimado utilizando métodos de máxima verossimilhança (ML) e as árvores de ML construídas utilizando 100 de replicatas de 'bootstrap'. As árvores consenso tiveram a sua raiz definida através do métodos 'mid-point'. Os valores de *bootstrap* acima de 70% de confiança foram indicados nas árvores. As arvores ML baseadas em sequências de nucleotídeos foram construídas no MEGA utilizando modelo de substituição Tamura 3 (T92+Γ) (Tamura 1992). As arvores ML baseadas em sequências de proteínas foram construídas no MEGA utilizando modelo de substituição de Poisson (em um caso, o modelo de substituição de WAG foi utilizado como indicado) (Zuckerkandl and Pauling 1965, Whelan and Goldman 2001). Análises de frequência de di-nucleotídeo em cada *contig* ou sequência de referência foi calculado e os resultados clusterizados baseado na correlação de Spearman para a construção de dendogramas,

essencialmente como descrito em (Lobo, Mota et al. 2009). Os cálculos foram realizados utilizando scripts em Perl e R utilizado para gerar os gráficos.

Tabela 1. Sequências utilizadas nas análises de filogenia e frequência de di-nucleotídeos.

Família	Gênero	Vírus	Identificador
Bunyaviridae	*Hantavírus*	*Hantaan vírus (HanV)*	P23456
		Seoul vírus Hantavírus (SeV)	P27314
		Tula vírus (Tula)	AJ005637.1
	Nairovírus	*Crimean-Congo hemorrhagic fever (CrCHFV)*	Q6TQR6
		Dugbe vírus isolate ArD44313 (DuV)	Q66431
	Orthobunyavírus	*Bunyamwera vírus (BunV)*	P20470
		La crosse (LCV)	Q8JPR2
	Phlebovírus	*Rift Valley fever vírus (RVFV)*	P27316
		Uukuniemi vírus (strain S23) (UkV)	P33453
	Tospovírus	*Tomato spotted (strain Brazilian Br-01)*	P28976
		Melon yellow spot vírus (MYSV)	AB061774.1
		Bean necrotic mosaic vírus (BNMV)	NC_018070.1
		Groundnut bud necrosis vírus (GBNV)	NC_003614.1
	Tenuivírus	*Rice stripe vírus (RSTV)*	Q85431
		Rice grassy (RGTV)	NC_002323.1
	Unclassified	*Phasi Charoen-like vírus (PCLV)*	KM001085.1
Reoviridae	*Aquareovírus*	*Aquareovírus*	Q8JU61
		Aquareovírus G	B2BNE0
		Aquareovírus	Q8VA42
	Coltivírus	*Colorado tick fever vírus (strain USA/Florio N-7180)*	Q9DSQ0
	Cypovírus	*Bombyx cypovírus 1 (BMCV)*	AF323782.1
	Dinovernavírus	*Aedes pseudoscutellaris reovírus (isolate France)*	Q2Y0E9
	Fijivírus	*Fiji disease vírus (Fijivírus)*	Q8JYK1
		Nilaparvata lugens reovírus (NFV)	NC_003654.1
	Mycoreovírus	*Cryphonectria parasitica mycoreovírus 1 (strain 9B21)*	Q7TDB6
	Orthoreovírus	*Reovírus type 1 (strain Lang)*	P0CK32
		Reovírus type 2 (strain D5/Jones)	P17377
		Reovírus type 3 (strain Dearing)	P0CK31
	Oryzavírus	*Rice ragged stunt vírus (isolate Thailand)*	O92604
	Cardoreovírus	*Eriocheir sinensis reovírus (isolate China/905)*	Q698V5
	Mimoreovírus	*Micromonas pusilla reovírus*	Q1I0V0
	Orbivírus	*Bluetongue vírus 10*	P13840
	Phytoreovírus	*Rice gall dwarf vírus (isolate Fujian)*	Q98631
		Rice dwarf vírus (isolate Fujian)	Q98631
	Rotavírus	*Rotavírus A (isolate SI/South Africa/H96/58)*	A2T3S0
		Rotavírus B	Q45UG0
		Rotavírus C	Q91E95
	Seadornavírus	*Banna vírus (strain Indonesia/JKT-6423/1980)*	Q9INJ1
Nodaviridae	*Betanodavírus*	*Striped jack nervous necrosis vírus (STNV)*	Q9QAZ8
		Tiger puffer nervous necrosis vírus (TPNV)	NC_013460
		Senegalese sole Iberian betanodavírus (SBIV)	NC_024492.1
	Alphanodavírus	*Flock house vírus (FHV)*	Q66929
		Nodamura vírus (NoV)	Q9IMM4
		Macrobrachium rosenbergii nodavírus (MrNV)	Q6XNL5
		Penaeus vannamei nodavírus (PVNV)	NC_014978.1
		Drosophila mel. American nodavírus (DmANV)	GQ342965.1
Unclassified	*Unasigned*	*Ixodes scapularis associated vírus 2 (Ixodes2)*	KM048319.1
		Ixodes scapularis associated vírus 1 (Ixodes1)	KM048318.1
Luteoviridae	*Polerovírus*	*Potato leafroll vírus (PLRV)*	KC456054.1
		Cucurbit aphid-borne yellows vírus (CABYV)	NC_003688.1
	Enamovírus	*Pea enation mosaic vírus-1 (strain WSG) (PEMV-1)*	P29154
		Citrus vein enation vírus (CEMV)	YP_008130302
	Luteovírus	*Bean_leafroll_vírus*	AAL66233.1
		Barley yellow dwarf vírus (isolate PAV) (BYDV)	P09505
	Sobemovírus	*Soybean yellow common mosaic vírus (SYCV)*	AEO16607
		Sesbania mosaic vírus (SMV)	YP_007697678
	Tombusviridae	*Carnation mottle vírus (CARMV)*	NC_001265.2
		Melon necrotic spot vírus (MNSV)	AB232925.1
		Tobacco necrosis vírus (TNV)	M33002.1

Lista de organismos, acrônimos e identificadores das sequências para cada banco de dados, Genbank ou Uniprot, onde foram extraídas são mostrados na **Tabela 1**.

4.11. RT-PCR e sequenciamento de Sanger

200 ng de total RNA extraído dos insetos foram reversamente transcritos em cDNA utilizando a transcriptase reversa MMLV. cDNA ou DNA foram submetidos a reação de PCR utilizando *primers* específicos. Oligonucleotideo *primers*, listados na **Tabela 2,** foram desenhados de acordo com as sequências dos *contigs* obtidos através do nosso pipeline de montagem. Os produtos de PCR foram diretamente submetidos a sequenciamento de Sanger.

Tabela 2. Oligonucleotideos utilizados para reações de PCR.

Alvo	Tm (^{o}C)	5'	3'
Vírus			
PCLV	55	CTATTATTGGCACCCCTGAA	CCAGATCCTAGCATTGGTTT
HTV	55	GTATACGCGTTGGTGAGTAT	CCGACTCAGCATAATTACGA
Aae.92	55	GTCTGATTTGCCCAACTCTA	CAGCATCGCAGGTTATAGTA
LPRV1	55	CCATGATCCAGCAATTCAAC	GTGCACACATATCATAAGCG
LPRV2	55	CTGGAAGATCAATGGTGTGA	TAATGGCGATGGACGATAAG
LPNV	55	GTGTTAATTGTGTGCGTTCC	GGTGACTCAATCAATGAACG
DUV	55	GAACTATCGCACCGTTTAAC	GTTGTGTCGTGTCTAGAAGT
MCV	55	TCGGACAAGGTTAAAGACTG	TCTTCGTATCGTGAAGAACC
DRV	55	GTGTGGTCTACATGTCAAGT	GGTAACAGCGTGTACCATAT
Segmentos do LPRV1			
3330/3331	55	ACACGTCGTTAATACCTCAG	GTTGTGAAGTAACTGGCAAC
3332/3333	55	AAGTAAACCCAGACCACATC	GTGTAGAGTATATGCGTGCA
3310/3311	55	ATTCCAGTCAGCGTAAAGTT	AGGTGTGATGGCATTGTAAT
3312/3313	55	TAATAGTTGTAGCCATGGCC	TCTCCACAGAGCAATCAATC
Segmentos do LPRV2			
3336/3337	55	TGCTACTCTAGTTCTCGTCA	CCATCTAAGTGTCAGCGTTA
3338/3339	55	AATATAGCCTATGCGACGAC	ACCACATGTATAATCGACGG
3314/3315	55	TTCCTGACGGGTAGACATAT	GAGTGCAAGCATATGACAAC
3318/3319	55	GGTATAACACGGTTTCCTGT	TACTACACTGCGGCTAGTT
3316/3317	55	CATGCAAGGAACATGATGTC	TATGTCAATGTGCGCATCTA

4.12. Códigos de acesso

As bibliotecas de *pools* de insetos sequenciadas foram depositadas no '*Small Read Archive of the National Center for Biotechnology Information'* sob números de acesso descritos na **Tabela 6**. Todas as sequências virais geradas derivadas dos *pools* de insetos neste projeto foram depositadas no banco de dados '*GenBank'* sob números de acesso KR003784-KR003824.

5. Resultados e discussão

5.1. Padronização do pipeline para detecção de vírus a partir de bibliotecas de pequenos RNAs

Com o intuito inicial de analisar a diversidade viral existente em mosquitos capturados da natureza nós desenvolvemos um pipeline de montagem de genomas virais utilizando os pequenos RNAs originários da via de RNAi derivados dos vírus (vsiRNAs). Contudo, apesar de trabalhos anteriores demonstrarem com sucesso a montagem de sequências virais a partir dos vsiRNAs de insetos e plantas (Kreuze, Perez et al. 2009, Wu, Luo et al. 2010), não se tem na literatura informações conclusivas sobre os melhores parâmetros e condições para a montagem de *contigs* a partir do sequenciamento de pequenos RNAs. Assim, nós decidimos avaliar parâmetros e condições necessárias para a eficiente montagem de *contigs* virais através da utilização de pequenos RNAs derivados de inseto.

Inicialmente, para padronizar a nossa estratégia de montagem de *contigs*, nós utilizamos uma biblioteca de pequenos RNAs preparada a partir de moscas *Drosophila melanogaster* deficientes para o gene *R2D2* que foram infectadas artificialmente em laboratório com o vírus VSV (Marques, Wang et al. 2013). Na **Tabela 3** pode-se ver um resumo dos dados após pré-processamento (ver Matérias e Métodos). Nós estabelecemos algumas etapas dentro do pipeline de detecção de vírus em bibliotecas de pequenos RNAs que serão discutidas individualmente nas seções subsequentes.

Tabela 3. Resumo da biblioteca de mosca mutante para R2D2 infectada com VSV.

# sequências totais	# sequências após remoção de adaptadores	# sequências 21 nt após remoção sequências D. mel	# sequências mapeadas VSV 21 nt	# sequências mapeadas VSV 21nt (+)	# sequências mapeadas SINV 21nt (-)
70.200.097	36.788.766	4.536.467	27.106	13.723	13.383

5.1.1. Determinação de parâmetros para enriquecimento de sequências de origem viral

Sequenciamento de pequenos RNAs em diferentes estudos têm demonstrado que uma fração considerável das sequências derivadas do sequenciamento de pequenos RNAs é de origem do hospedeiro (Wang, Wu et al. 2010, Kemp, Mueller et al. 2013, Marques, Wang et al. 2013). Assim para tentar enriquecer os nossos dados para sequências virais nós decidimos remover todas as sequências que apresentavam similaridade com o genoma da mosca. Para isso nós realizamos o mapeamento das sequências no genoma da *Drosophila* onde foi observado que de um total de 36.787.341 sequências, 6.143.364 sequências foram mapeadas. A partir das 30.643.977 sequências não mapeadas foram selecionamos 4.560.202 sequências que apresentaram tamanho de 21 nt, tamanho canônico dos siRNAs oriundos do processamento da Dicer-2 em insetos (Wu, Luo et al. 2010, Han, Luo et al. 2011, Marques, Wang et al. 2013, van Cleef, van Mierlo et al. 2014), para a etapa de montagem de *contigs*.

5.1.2. Determinação dos parâmetros para a montagem de *contigs* a partir de pequenos RNAs

Afim de montar *contigs* que representassem o genoma do VSV nós utilizamos o software de montagem *Velvet*, conhecido por ser otimizado para montagem de sequências pequenas (Zerbino and Birney 2008, Miller, Koren et al. 2010). Como a escolha do *k-mer,* tamanho o qual as sequências serão decompostas e comparadas

entre se com o intuito de achar sobreposição, pode influenciar na qualidade e confiança dos *contigs* montados como revisado por (Miller, Koren et al. 2010), nós avaliamos diferentes tamanhos de *k-mer* entre 11 e 19, visto que o tamanho do siRNA canônico derivado de insetos é 21 nt e o *k-mer* utilizado pelo Velvet tem de ser um número impar para se evitar palíndromos (**Tabela 4**) (Zerbino and Birney 2008).

Nós observamos que quando utilizado os *k-mers* de tamanho 11 ou 19, nós montamos poucos ou nenhum *contig*, provavelmente devido à baixa estringência no caso do *k-mer* 11, que permite a sobreposição de sequência por mero acaso, ou por alta estringência no caso do *k-mer* 19 (**Tabela 4**).

Tabela 4. Avaliação do melhor *k-mer* para montagem de *contigs* virais a partir dos siRNAs.

	Todos os *contigs*				*Contigs* VSV			
k-mer	***#contigs***	**N50**	**média de tamanho**	**maior *contig***	***#contigs***	**média de tamanho**	**maior *contig***	**cobertura VSV (%)**
11	0	0	0	0	0	0	0	0
13	5428	27	27.89	56	39	27.9	36	9.47
15	238	72	71.74	241	81	76.1	187	52.97
17	78	65	67.55	149	39	65.69	114	22.28
19	3	56	58.33	63	2	56.5	57	0.98

Os *k-mers* intermediários, 13,15 e 17 apresentaram os melhores resultados, onde obteve-se 9,47, 52,97 e 22,28 % de cobertura do genoma viral a partir dos *contigs* montados, **Tabela 4**. Contudo, o *k-mer* 15 apresentou os melhores valores tanto para o número de *contigs* virais, quanto para o tamanho do maior *contig* viral e respectiva cobertura do genoma do VSV. Esse resultado aliado a intermediária quantidade de *contigs* totais montados, 238 contra 5.428 gerados com *k-mer* de 13 e 78 gerados com *k-mer* de 17 nos levou a decidir pela utilização do *k-mer* de 15 para tentar montar *contigs* virais a partir dos siRNAs.

Para que pudéssemos utilizar métricas para avaliar a cobertura, número de *contigs* gerados, influência do sinal (sequências derivadas do vírus) e ruído (sequências não derivadas do vírus) na montagem de *contigs* de virais, foram extraídas todas as

sequências que, segundo mapeamento contra genoma do vírus, eram de origem do VSV. Nós observamos que 27.106 das 4.560.202 sequências de 21 nt não mapeadas no genoma da mosca eram derivadas do VSV, referidas como sinal, enquanto as sequências não mapeadas, 4.533.096, foram utilizadas como ruído nas análises de montagem de *contigs*. Para definir a condição ótima, foi realizada montagem de *contigs* utilizando somente sequências de 21 nt derivadas do vírus VSV, onde não haveria interferência do ruído. A partir dos *contigs* montados, foram calculadas as métricas de montagem e cobertura do genoma viral utilizadas para avaliar a eficiência da montagem como pode ser visto na **Tabela 5**.

Tabela 5. Resumo da montagem utilizando somente as sequências mapeadas no genoma viral.

# sequências	***# contigs***	**N50**	**maior *contig* (nt)**	**cobertura VSV (%)**
27.106	97	87	259	71,45

Nós observamos que na condição ótima todos os parâmetros de montagem assim como a cobertura do genoma viral apresentaram melhores valores quando em comparação com as condições convencionais, onde existe a presença de sequências não oriundas do VSV. Por exemplo, o tamanho do maior *contig* e cobertura do genoma viral foram 38 e 18,48% menor na condição convencional em comparação com a condição ótima (comparação **Tabela 4** *k-mer* 15 e **Tabela 5**). Assim, nós percebemos uma diminuição razoável nas métricas importantes para a eficiência do nosso pipeline. Tendo em vista esse fato nós decidimos investigar em detalhes a influência da quantidade de sequências virais e não virais na detecção de *contigs* oriundos do VSV.

5.1.3. Análise da influência da proporção de sequências virais e não virais para a montagem de derivados do vírus

Inicialmente foi avaliada a influência que as sequências não virais, o ruído, exerce na detecção de *contigs* virais e consequentemente na cobertura do genoma. As

sequências consideradas ruído, 4.533.096, foram diluídas continuamente nas proporções de (100, 80, 40, 20, 10, 5 e 2,5 %) enquanto as sequências consideradas sinal, 27.106, foram mantidas constantes. Assim, nós executamos o *pipeline* para que fosse possível avaliar o comportamento das métricas de montagem de *contigs* em cada diluição com relação à detecção dos *contigs* derivados do VSV. A compilação dos resultados pode ser visualizada na **Figura 4A**.

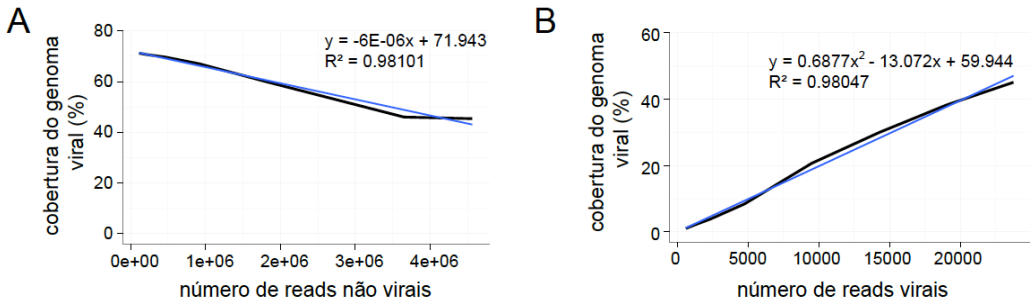

Figura 4. Avaliação da diminuição do ruído na montagem de *contigs* e cobertura do genoma viral. A cobertura diminui até um ponto onde a quantidade de ruído não é tão influente na cobertura do genoma viral.

A partir desta análise foi possível perceber que a quantidade de sequências não virais diminui a eficiência de detecção de *contigs* do VSV, mas não exerce papel fundamental na detecção de *contigs* do vírus (**Figura 4A)**. Em contrapartida, a diluição das sequências derivadas do VSV influenciou fortemente no decaimento da cobertura do genoma viral como pode ser visto na **Figura 4B**, sugerindo que as sequências derivadas do VSV exercem uma influência mais direta na montagem de *contigs* virais.

Ainda a partir da diluição das sequências virais foi possível chegar a uma sugestão do montante de sequências derivadas do VSV necessárias para a detecção de *contigs*, cerca de 1.200. É importante ressaltar que essa é uma informação importante que poderia ser usada posteriormente para definir quais insetos sequenciar, visto que poderíamos correlacionar os resultados de PCR da análise da carga viral de vírus de interesse com a quantidade de sequências virais encontradas.

De forma interessante, quando avaliada as diluições das sequências derivadas do VSV nós percebemos que mesmo na completa ausência das sequências derivadas do vírus, *contigs* com similaridade com o vírus são montados, sendo assim falsos positivos. Assim, nós resolvemos investigar as métricas da montagem de *contigs* para avaliar se existia alguma correlação com a identificação de falsos positivos. Através dessa avaliação foi possível observar uma forte relação entre o tamanho do maior *contig* com similaridade com o VSV e o N50 da montagem de *contigs*, **Figura 5**.

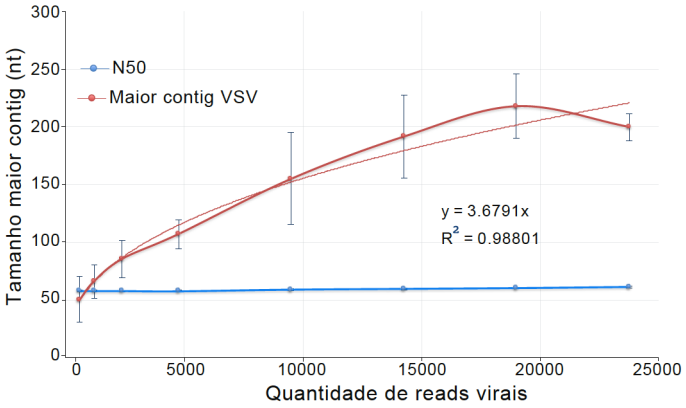

Figura 5. Relação entre N50 da montagem e tamanho do maior *contig* do VSV sugere limiar de detecção de *contigs* virais. O número de sequências virais foi diluído seriadamente e realizada montagem de *contigs* para cada diluição onde as métricas de montagem foram avaliadas. Mesmo na ausência completa de sequências derivadas do vírus, *contigs* com similaridade com o VSV são encontrados.

Desta forma nós ponderamos que o N50 poderia ser utilizado como filtro de falso positivos para definição do tamanho mínimo dos *contigs* que seriam utilizados para as análises de caracterização.

Por fim, com os resultados da padronização da estratégia utilizando a biblioteca de pequenos RNAs de *Drosophila* infectada com VSV foi possível avaliar parâmetros importantes para a montagem e caracterização de sequências virais como o melhor *K-mer* a se utilizar na montagem de *contigs*, o efeito da quantidade de sequências virais e não virais na detecção de *contigs* virais e sugerir um limiar de detecção de *contigs* virais

para se evitar falsos positivos. Assim nós nos utilizamos dessas informações para o ajuste do pipeline desenvolvido que pode ser visualizado na **Figura 6**.

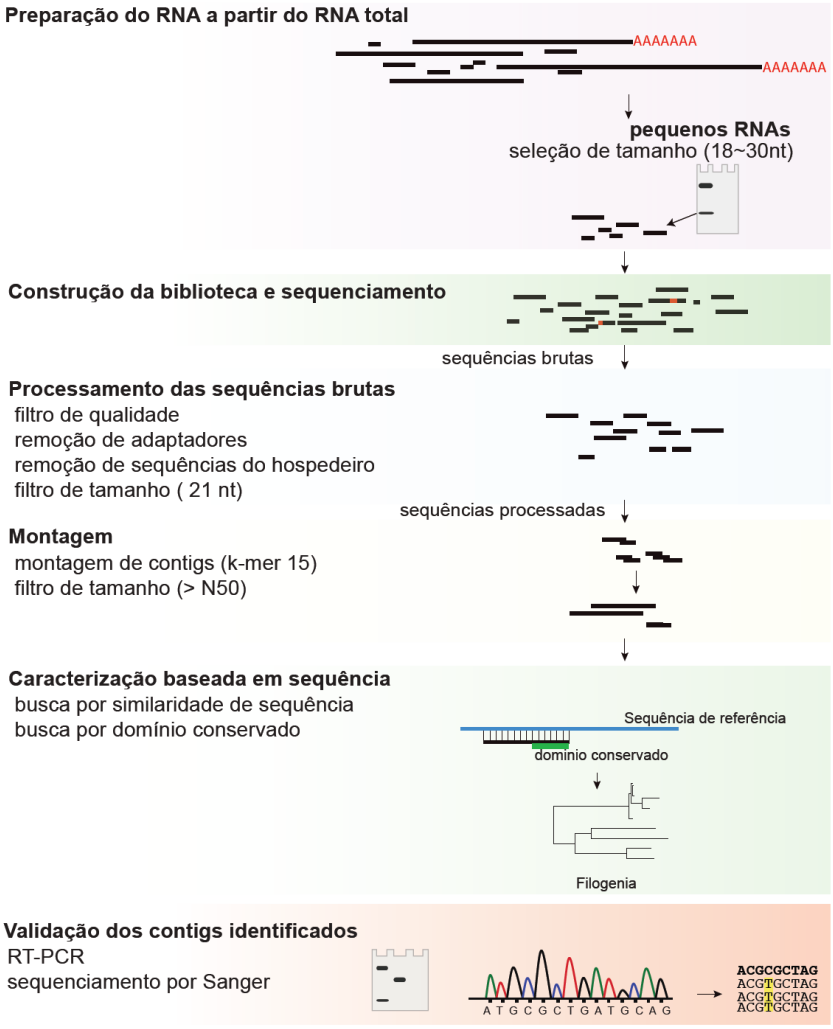

Figura 6. Pipeline de detecção de vírus em bibliotecas de pequenos RNAs. A partir das bibliotecas de pequenos RNAs, as sequências brutas serão processadas, para tentar enriquecer para sequências virais, e utilizadas para montagem de *contigs*. Os *contigs* serão caracterizadas utilizando estratégias baseadas em similaridade de sequências e validadas através de RT-PCR e sequenciamento de Sanger.

A partir da definição dos melhores parâmetros e condições para identificação de sequências virais baseado no sequenciamento de pequenos RNAs nós decidimos

avaliar o comportamento do nosso pipeline em amostras de populações de três diferentes insetos de laboratório com o intuito de otimizar a nossa estratégia.

5.2. Caracterização do viroma de populações da mosca *Drosophila melanogaster* e dos insetos vetores *Aedes aegypti* e *Lutzomyia longipalpis*

Já foi descrito na literatura que através do sequenciamento de pequenos RNAs é possível a identificação de novos vírus em insetos e plantas (Kreuze, Perez et al. 2009, Wu, Luo et al. 2010). Assim, como forma de avaliar melhor nossa estratégia de montagem e detecção de *contigs* virais a partir de bibliotecas de pequenos RNAs foram construídas bibliotecas de pequeno RNAs de populações de laboratório de moscas *Drosophila melanogaster* de laboratório e populações de mosquitos *Aedes aegypti* e flebotomíneos *Lutzomyia longipalpis*, dois importantes vetores para patógenos de humanos (**Tabela 6**).

Tabela 6. Resumos das bibliotecas de RNAs sequenciadas nesse trabalho.

Biblioteca	Número de indivíduos no pool	ID SRA	Infecção artificial	Número total de sequências	Número sequências mapeadas hospedeiro	Número de sequências processadas	Número de *contigs*	N50 (nt)	Tamanho maior *contig* (nt)	Número *contigs* similarida de vírus
Drosophila melanogaster										
Dme_1	6	SRR1803381	SINV	4,234,079	3,194,745	596,761	327	79	4,765	41
Dme_2	6	SRR1803382	DCV	15,786,440	11,483,909	2,018,666	343	137	3,496	5
Dme_3	6	SRR1803383	VSV	24,474,261	21,701,073	1,039,581	171	288	3,482	37
***Aedes aegypti* – Pequenos RNA**										
Aae_1	6	SRR1803377	-	9,081,151	8,076,206	891,983	1,686	67	2,301	16
Aae_2	6	SRR1803378	-	12,183,902	10,827,108	999,843	1,658	66	1,611	17
Aae_3	6	SRR1803379	-	9,253,941	8,010,814	1,158,379	2,722	68	5,122	12
***Aedes aegypti* – Longos RNA**										
Aae_1	6	SRR1813817	-	39,488,681	35,767,399	3,721,282	295,760	136	2,070	12
Aae_2	6	SRR1813823	-	57,584,234	52,130,913	5,453,321	358,323	136	1,988	17
Aae_3	6	SRR1813824	-	62,302,651	56,383,441	5,919,210	357,264	138	2,334	9
Lutzomya longipalpis										
Llo_1	8	SRR1803384	-	12,297,884	10,852,586	483,139	1,207	69	1,345	14
Llo_2	7	SRR1803385	-	9,463,241	8,162,975	449,4327	2,151	63	980	12
Llo_3	7	SRR1803386	-	8,109,613	7,285,842	659,215	1,541	78	1,113	31

5.2.1. Otimização da montagem de *contigs* a partir de bibliotecas de pequenos RNAs

Bibliotecas de *Drosophila* foram preparadas a partir de linhagens de laboratório infectadas com três vírus diferentes, *Drosophila C vírus* (DCV), *Sindbis vírus* (SINV) e *Vesicular stomatitis vírus* (VSV) que foram utilizados para ajudar a otimizar nosso pipeline de detecção de vírus a partir de sequências de pequenos RNAs (**Tabela 6**). As Bibliotecas de pequenos RNAs foram preparados a partir de insetos inteiros sem nenhuma manipulação das amostras antes da extração de RNA, para minimizar os riscos de contaminação da amostra no laboratório. Após o sequenciamento das bibliotecas de pequenos RNAs, os dados foram processados para enriquecer para potenciais sequências virais através da remoção de sequências derivadas do hospedeiro e de genomas bacterianos. Sequências derivadas dos hospedeiros corresponderam à grande maioria de pequenos RNAs (73-92%) das bibliotecas, mas uma porcentagem substancial de sequências (3,4-14% das bibliotecas) permaneceu após estas etapas de processamento (**Figura 7**).

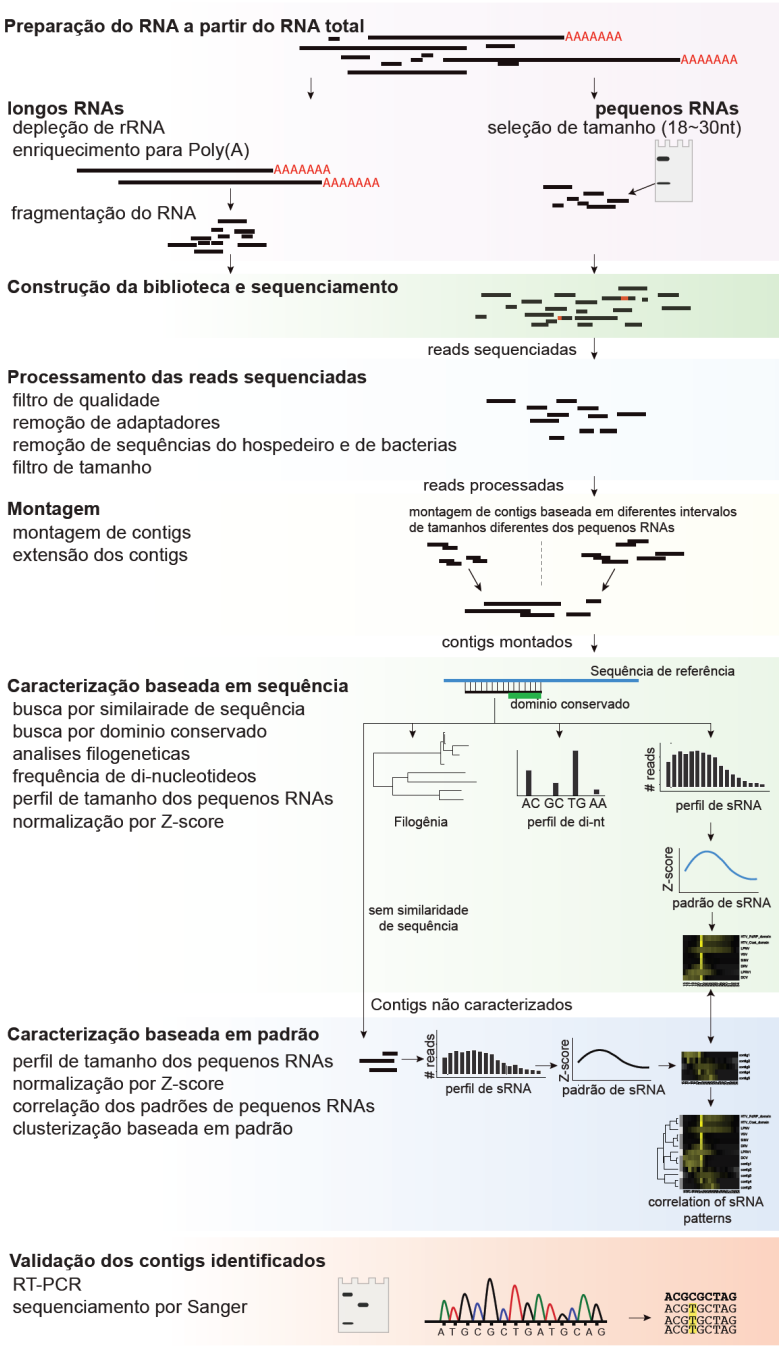

Figura 7. Pipeline otimizado de detecção de vírus baseado em pequenos e longos RNAs. Diferentes frações de RNA foram utilizadas para a construção de bibliotecas de pequenos e longos RNAs. As sequências brutas foram processadas para enriquecer para potenciais sequencias virais. As sequências processadas foram então submetidas a montagem de *contigs* e posterior extensão. Os *contigs* montados foram caracterizados utilizando estratégias baseadas em sequência e padrão. Os *contigs* virais foram validados por RT-PCR e sequenciamento de Sanger. Mais detalhes no texto da seção.

Inúmeros estudos tem mostrado que os pequenos RNAs derivados de vírus apresentam majoritariamente 21 nt de tamanho (siRNAs) (Mueller, Gausson et al. 2010, Kemp, Mueller et al. 2013, Marques, Wang et al. 2013). Apesar desses resultados

sugerirem que siRNAs otimizaria a montagem de *contigs*, nós ponderado que focar somente nas sequências de 21 nt poderia ter um efeito restringente na montagem de *contigs*, visto que já foi descrito na literatura a via de piRNAs ou degradação por outras nucleases podem gerar pequenos RNAs derivados de vírus em insetos (Wu, Luo et al. 2010, Morazzani, Wiley et al. 2012, Vodovar, Bronkhorst et al. 2012). Nesse contexto, nós levamos em consideração que esses pequenos RNAs de diferentes origens poderiam contribuir para a montagem de mais *contigs* e com tamanho maiores. Assim, nós testamos o uso de diferentes tamanhos de pequenos RNAs para a montagem de *contigs* nas amostras derivadas dos diferentes insetos (**Figura 8A**). O maior número e tamanho de *contigs* foram obtidos quando utilizando pequenos RNAs de tamanho de 20-23 nt e 24-30 nt na montagem de *contigs* separadamente e os resultados combinados posteriormente (**Figura 8A**).

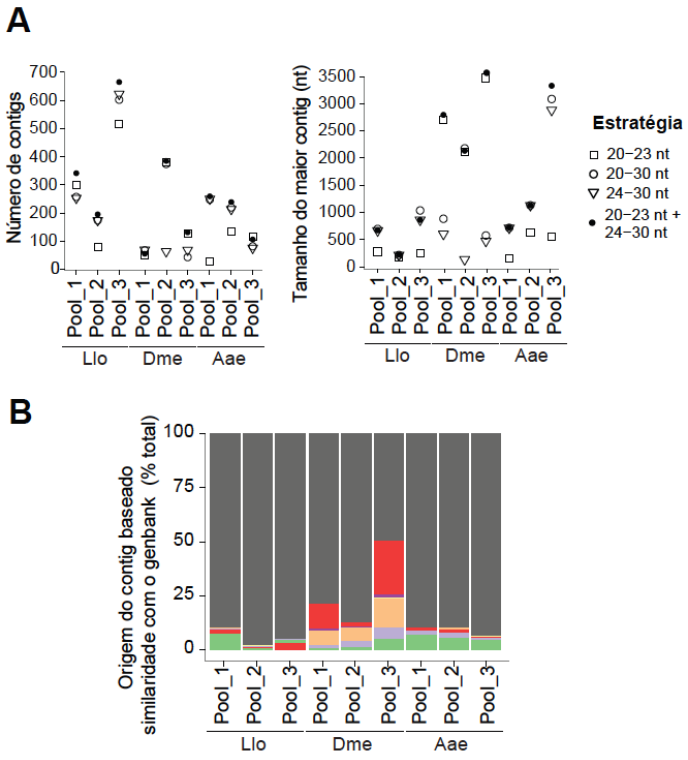

Figura 8. Estratégia de montagem e caracterização de sequências virais a partir do sequenciamento de pequenos RNAs é capaz de identificar vírus circulantes. (**A**) Comparativo do número de *contigs* e tamanho do maior *contig* em cada biblioteca de pequenos RNAs utilizando diferentes intervalos de tamanho no passo de montagem de *contigs*. (**B**) Proporção de *contigs* montados em cada biblioteca com similaridade significante com sequências de referência classificadas por táxon, incluindo sequências desconhecidas.

A montagem de *contigs* utilizando outros tamanhos de pequenos RNAs incluindo 20-23, 24-30 ou 20-30 nt resultou em métricas variáveis dependendo da biblioteca analisada. Assim, foi decidido que a combinação das montagens de *contigs* utilizando os pequenos RNAs de tamanho 20-23 e 24-30 nt separadamente seria uma estratégia que se adequaria melhor a bibliotecas as quais não se tem nenhum conhecimento prévio sobre o perfil de pequenos RNAs.

Para tentar avaliar a robustez da nossa estratégia de detecção de novos vírus desenvolvida nós utilizamos as bibliotecas de *Drosophila melanogaster* artificialmente infectadas com diferentes vírus como 'controle positivo' da identificação de vírus (**Tabela 6**). Nestas bibliotecas foram identificados 42, 40 e 1 *contigs* que mostraram similaridade de sequência com o VSV, SINV e DCV respectivamente (**Figura 9**). Com a identificação dos respectivos vírus utilizados na infecção artificial foi determinado que a estratégia poderia detectar com sucesso vírus sabidamente presentes nas amostras, apesar da detecção ser limitada pela quantidade de pequenos RNAs derivados do vírus já tinha sido visto nas análises anteriores realizadas para a padronização da estratégia e no caso do DCV. Foram observados 1.572 pequenos RNAs derivados do DCV, os quais permitiram a montagem de somente um *contig* cobrindo cerca de 0,8% do genoma viral (**Figura 9A**). Em contrapartida, 53.620 e 9.588 pequenos RNAs derivados do VSV e SINV, respectivamente, permitiram a montagem de múltiplos *contigs* únicos que cobriram 81,1 e 23,4% dos respectivos genomas virais (**Figura 9A**). Assim, confirmou-se que alta cobertura do genoma viral é característica importante para permitir a montagem de *contigs* a partir da sobreposição de pequenos RNAs.

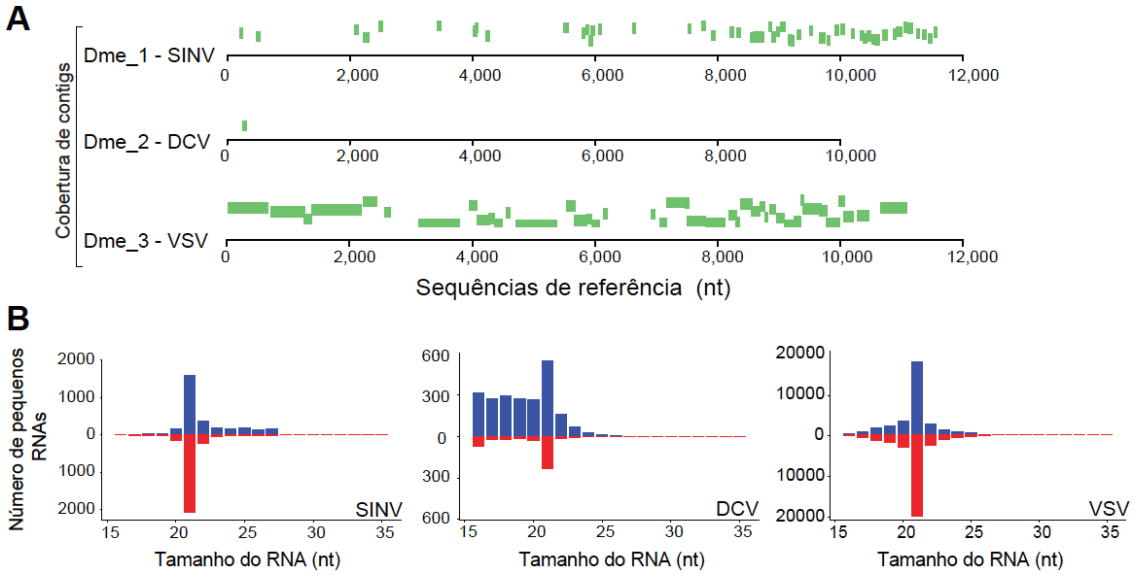

Figura 9. Pequenos RNAs derivados de vírus podem ser montados em *contigs*. (A) Distribuição dos *contigs* montados utilizando nossa estratégia de montagem baseada em pequenos RNAs ao longo dos genomas de referência do DCV, VSV e SINV. (B) Perfil de tamanho dos pequenos RNAs derivados do DCV, VSV e SINV utilizados para infecção de laboratório da *Drosophila*.

Em seguida, todos os *contigs* montados a partir das bibliotecas de pequenos RNAs de *Drosophila melanogaster, Aedes aegypti* e *Lutzomyia longipalpis* foram utilizados para buscas de similaridade baseada em sequência contra os bancos de dados não redundantes do NCBI (nucleotídeo e proteína). A grande maioria dos *contigs* montados em todas as 9 bibliotecas de pequenos RNAs (10.577 de um total de 11.806) não apresentaram similaridade de sequência significativa e foram referidas como "desconhecidas" (**Figura 8B**). A grande Maioria desses *contigs* (92%) apresentaram tamanho menor que 100 nt e foram descartados de analises mais detalhadas.

Não obstante, a clusterização de nossas bibliotecas baseado na similaridade dos *contigs* desconhecidos conseguiu separar as amostras advindas de *Drosophila*, *Aedes* e *Lutzomyia*, o que sugere que aqueles *contigs* são hospedeiro-específicos (**Figura 10**). Os 1.229 *contigs* não redundantes remanescentes foram classificados de acordo com o táxon designado pelo resultado mais significante de BLAST (**Figura 8B**). Diversos *contigs* apresentaram similaridade com sequencias derivadas do reino Animália, em especial *contigs* vindos das bibliotecas de mosquitos e flebotomíneos (**Figura 8B**). Esses *contigs*

provavelmente pertencem ao genoma do inseto, mas não foram removidas na etapa de pré-processamento. Isto reflete o fato de que os genomas de *Aedes aegypti* e *Lutzomyia longipalpis* não estarem tão bem curados como o de *Drosophila melanogaster* (Adams, Celniker et al. 2000, Nene, Wortman et al. 2007). Muitos dos *contigs* remanescentes são derivadas de bactérias e fungos que poderiam ser parte do microbioma do inseto.

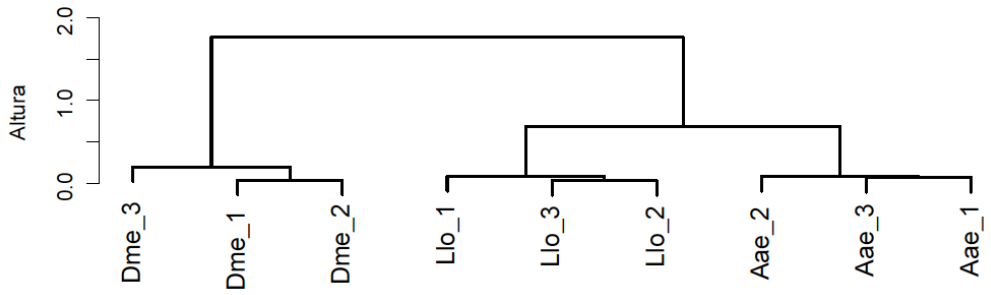

Figura 10. *Contigs* não caracterizados classificam as bibliotecas de pequenos RNAs de maneira hospedeiro-especifica. Clusterização das bibliotecas baseada na similaridade de sequências dos *contigs* desconhecidos agrupa as bibliotecas de forma inseto-especifica.

5.2.2. Comparação entre estratégias utilizando pequenos ou longos RNAs para a detecção de sequências virais

Através da utilização de bibliotecas de pequenos RNAs nós observamos que foi possível montar *contigs* em todas as bibliotecas avaliadas. Adicionalmente, através de análises de similaridade nós observamos que parte dos *contigs* montados apresentavam similaridade significante com sequências virais (**Figura 8**). Contudo, apesar da identificação dessas sequências virais, nós não tínhamos informação suficiente para comparar como esta estratégia se comportaria quando em comparação com outras alternativas para detecção de sequências virais. A maioria das estratégias já investigadas utilizam algum tipo de manipulação das amostras com o intuito de enriquecer para sequências virais antes da extração dos ácidos nucleicos, apesar disso poder resultar em contaminação (Tokarz, Williams et al. 2014, Li, Shi et al. 2015). Nesse contexto, o direto sequenciamento de RNAs longos é também utilizado, mas pode ser limitado pela

abundância de RNAs ribossomais do hospedeiro que representam a vasta maioria das sequências em bibliotecas de RNAs longos (Vivancos, Guell et al. 2010).

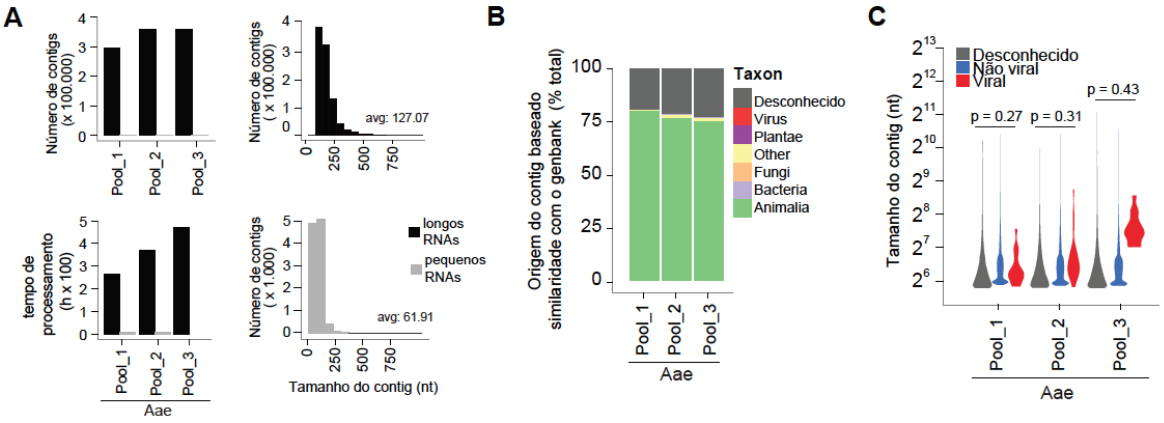

Figura 11. Análise comparativa dos *contigs* virais montados a partir de bibliotecas de pequenos e longos RNAs. (A) Comparativo do tempo de processamento, número e distribuição de frequência dos tamanhos de *contigs* para as bibliotecas de *Aedes aegypti* de pequenos e longos RNAs mostrados em cinza e preto, respectivamente. (B) Proporção de contigs montados em cada biblioteca com similaridade significante com sequências de referência classificadas por táxon, incluindo sequências desconhecidas. (C) Distribuição de tamanho dos *contigs* correspondentes a viral (vermelho), não viral (azul) ou desconhecidas (cinza) em cada biblioteca. P-*values* das diferenças entre os tamanhos dos *contigs* foram calculados utilizando *test t* de *Student* e indicados na figura.

Como alternativa, nós construímos bibliotecas de RNAs longos após a depleção de RNA ribossomal e enriquecimento para polyA a partir do mesmo RNA total das populações de *Aedes aegypti* utilizadas no preparo das bibliotecas de pequenos RNAs (**Tabela 6**). Isto nos permitiu a comparação direta dos resultados do sequenciamento em larga escala dos pequenos e longos RNAs das mesmas amostras sem qualquer manipulação anterior à extração de RNA. O número e tamanho das sequências obtidas com o sequenciamento de RNAs longos resultou em 10,4 vezes mais dados em comparação às bibliotecas de pequenos RNAs (**Tabela 6**). Como resultado, as bibliotecas de RNAs longos propiciaram a montagem de um total de 1.001.347 *contigs* com um N50 de ~136 nt comparados a 6.066 *contigs* com um N50 de ~48 nas bibliotecas de pequenos RNAs (**Figura 11A e Tabela 6**). O grande número de *contigs* resultou em 43 a 72 vezes mais requerimento de tempo de processamento para as bibliotecas de longos RNAs para análises de busca por similaridade contra bancos de dados quando

em comparação dos longos e pequenos RNAs (**Figura 11B**). A maioria do *contigs* montados a partir das bibliotecas de RNAs longos (> 60%) apresentaram similaridade a sequências derivadas do reino Animal que provavelmente representam partes ainda não montadas do genoma do *Aedes aegypti* **(Figura 11B)**. A bibliotecas de RNAs longos também apresentaram uma grande quantidade de sequências sem nenhuma similaridade significante a sequências presentes nos bancos de dados de referência. Entretanto, eles não representam a maioria dos *contigs* como nas bibliotecas de pequenos RNAs (**Figura 11B**). Esses resultados sugeririam que as bibliotecas de RNAs longos são mais indicadas para a identificação de vírus, visto que elas têm significativamente mais *contigs*. Contudo, o número total de *contigs* virais foi muito similar nas bibliotecas de longos e pequenos RNAs (**Figura 11**). Adicionalmente, a média de tamanho dos *contigs* virais foi maior nas bibliotecas de pequenos RNAs, o que favorece a caracterização do *contig* através de análises baseadas em similaridade de sequências. (**Figura 11C**).

Finalmente, nossos resultados indicaram que as bibliotecas de pequenos RNAs foram enriquecidas e naturalmente favorecem a montagem de *contigs* virais em comparação com as bibliotecas de RNAs longos. O mecanismo de biogênese dos pequenos RNAs pelas vias do hospedeiro aparentemente favorecem a geração de sequências com sobreposição que são importantes para a montagem e extensão dos *contigs* comparado ao sequenciamento de RNAs longos. É possível que os RNAs virais poderiam ser enriquecidos em RNAs longos se nós não tivéssemos limitado nosso sequenciamento a RNAs poliadenilados. Contudo, a depleção de RNAs ribossomais por si só poderia ainda enviesar os nossos resultados de sequenciamento. Adicionalmente, é importante salientar que os pequenos RNAs apresentam enriquecimento natural para sequencias virais sem que haja a necessidade de nenhum passo extensivo de processamento antes da construção da biblioteca. Assim, nós decidimos pela utilização da estratégia de sequenciamento de pequenos RNAs para avaliar o viroma de insetos.

5.2.3. Detecção de vírus baseada em similaridade de sequência

A partir da análise dos *contigs* baseado em similaridade de sequência a bancos de dados de referência, nós observamos que dos 1.229 *contigs* não redundantes, 223 (~18%) apresentaram similaridade significante com sequências virais em bancos de dados de referência (**Figura 8B e Tabela 7**). A media de tamanho dos *contigs* virais foi significantemente maior do que todos os outros *contigs* montados de diferentes origens (**Figura 12**). Esses resultados sugeriram que a estratégia baseada em pequenos RNAs favorece a montagem de *contigs* virais de tamanhos maiores quando em comparação com *contigs* de outras origens. Foram removidos 83 *contigs* derivados de DCV, SINV e VSV que estavam entre os 223 *contigs* virais. Nós consolidamos os *contigs* baseado em similaridade de sequência onde os 140 *contigs* remanescentes foram filtrados para eliminar sequências similares detectadas em mais de uma biblioteca da mesma espécie de inseto. Nós ainda utilizamos uma estratégia baseada em sobreposição de sequências para tentar estender as sequências virais utilizando *contigs* com similaridade a vírus da mesma família viral. Esses passos permitiram gerar resultados consolidados a partir das 3 bibliotecas independentes de pequenos RNAs oriundas de cada população de insetos, *Drosophila*, *Aedes* e *Lutzomyia*. As etapas de remoção de redundância, consolidação e extensão de *contigs* resultaram em uma redução de 140 *contigs* para 34 sequências não redundantes que foram designadas para, ao menos, 7 vírus baseadas no resultado mais significante de BLAST em comparação a sequências depositadas em bancos de dados de referência (**Tabela 7**). Analises filogenéticas sugeriram que 6 dos 7 vírus representavam novas espécies de vírus (**Figura 13**). Com relação ao viroma de cada inseto, 2 vírus foram detectados em mosquitos, 3 em flebotomíneos e 2 em moscas da fruta.

Em mosquitos foram detectados *contigs* que pertenciam a uma nova cepa do *Phasi Charoen Like-vírus* (PCLV), um bunyavirus previamente identificado em mosquitos da Tailândia (**Figura 13A**) (Yamao, Eshita et al. 2009). Em adição, nós também

identificamos *contigs* oriundos de um novo vírus relacionado ao *Laem Singh vírus* (LSV) e 2 outros vírus recentemente descritos em carrapatos, *Ixodes scapularis associated vírus* 1 e 2 (**Figura 13B**) (Tokarz, Williams et al. 2014). Este vírus foi nomeado *Humaita-Tubiacanga vírus* (HTV), refletindo a origem do hospedeiro onde ele foi identificado.

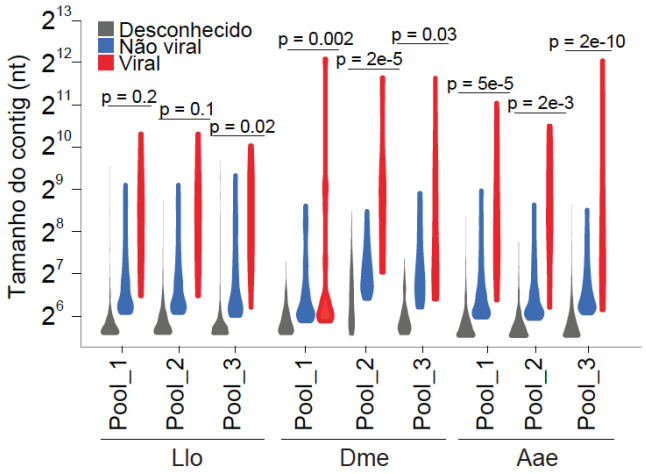

Figura 12. O tamanho dos *contigs* virais é significativamente maior que dos *contigs* não virais. Distribuição de tamanho dos *contigs* correspondentes a viral (vermelho), não viral (azul) ou desconhecidas (cinza) em cada biblioteca. P-*values* das diferenças entre os tamanhos dos *contigs* foram calculados utilizando *test t* de *Student* e indicados na figura.

Em flebotomíneos, nós observamos diversos *contigs* não redundantes apresentando similaridade com reovírus e nodavírus (**Tabela 7**). Especificamente, 23 *contigs* não redundantes apresentaram similaridade a vírus do gênero *Cypovírus*, da família *Reoviridae* (**Tabela 7**).

Esse numero de *contigs* virais únicos é alto, mesmo considerando o fato que reovírus podem apresentar até 12 segmentos genômicos. Baseado em análises filogenéticas dos segmentos genômicos que codificam as polimerases virais dependentes de RNA (RdRPs), nós identificamos com sucesso duas sequências distintas oriundas de reovírus pertencentes ao gênero *Cypovírus* (**Figura 13C**). Estes vírus foram nomeados *Lutzomyia Piaui reovírus* 1 (LPRV1) e *Lutzomyia Piaui reovírus 2* (LPRV2), refletindo a localização geográfica onde o seu hospedeiro foi obtido. A partir da analise dos outros *contigs* virais montados a partir das bibliotecas de flebotomíneos, foram identificados

contigs que apresentaram similaridade com nodavírus sugerindo que eles pertencem a um novo vírus relacionado com o *Nodamura vírus* e pertencente ao gênero *Alphanodavírus* (**Figura 13D**). Esse novo vírus foi nomeado *Lutzomyia Piaui nodavírus* (LPNV).

Em moscas da fruta, nos identificamos *contigs* que mostraram similaridade a duas famílias virais não relacionada a vírus utilizados nas infecções experimentais. Essa descoberta sugeriu que o estoque de *Drosophila* de laboratório que foram utilizadas já carregavam infecções virais não conhecidas previamente (**Figura 14**). Um conjunto de *contigs* mostrou similaridade a reovírus. Analises filogenética sugeriu que este vírus pertencia a vírus do gênero *Fijivírus*, da família *Reoviridae* (**Figura 13C**). Este vírus foi nomeado *Drosophila reovírus* (DRV). Outro *contig* viral mostrou similaridade com *Acyrthosiphon pisum vírus* mas não pode ser designado a nenhuma família viral através de análises filogenéticas (**Figura 13E**). Por este motivo, este vírus foi nomeado *Drosophila uncharacterized vírus* (DUV).

Sequências correspondendo a todos os potenciais 7 vírus foram amplificados com sucesso através de reações de PCR do RNA transcrito, mas não do DNA (**Figura 14 e dados não mostrados**). Isto indicou que estes vírus estão presentes somente na forma de RNA, o que é consistente com a observação que eles presumidamente pertencem a famílias virais com genomas de RNA (**Figura 13 e Figura 14**). Para avaliar a qualidade da nossa montagem de sequências a partir dos pequenos RNAs também foram sequenciados, através do método Sanger, os produtos de PCR. O sequenciamento mostrou que as sequências derivadas do sequenciamento de Sanger em comparação com os *contigs* montados tem entre 99 a 100% de identidade (**Figura 14**).

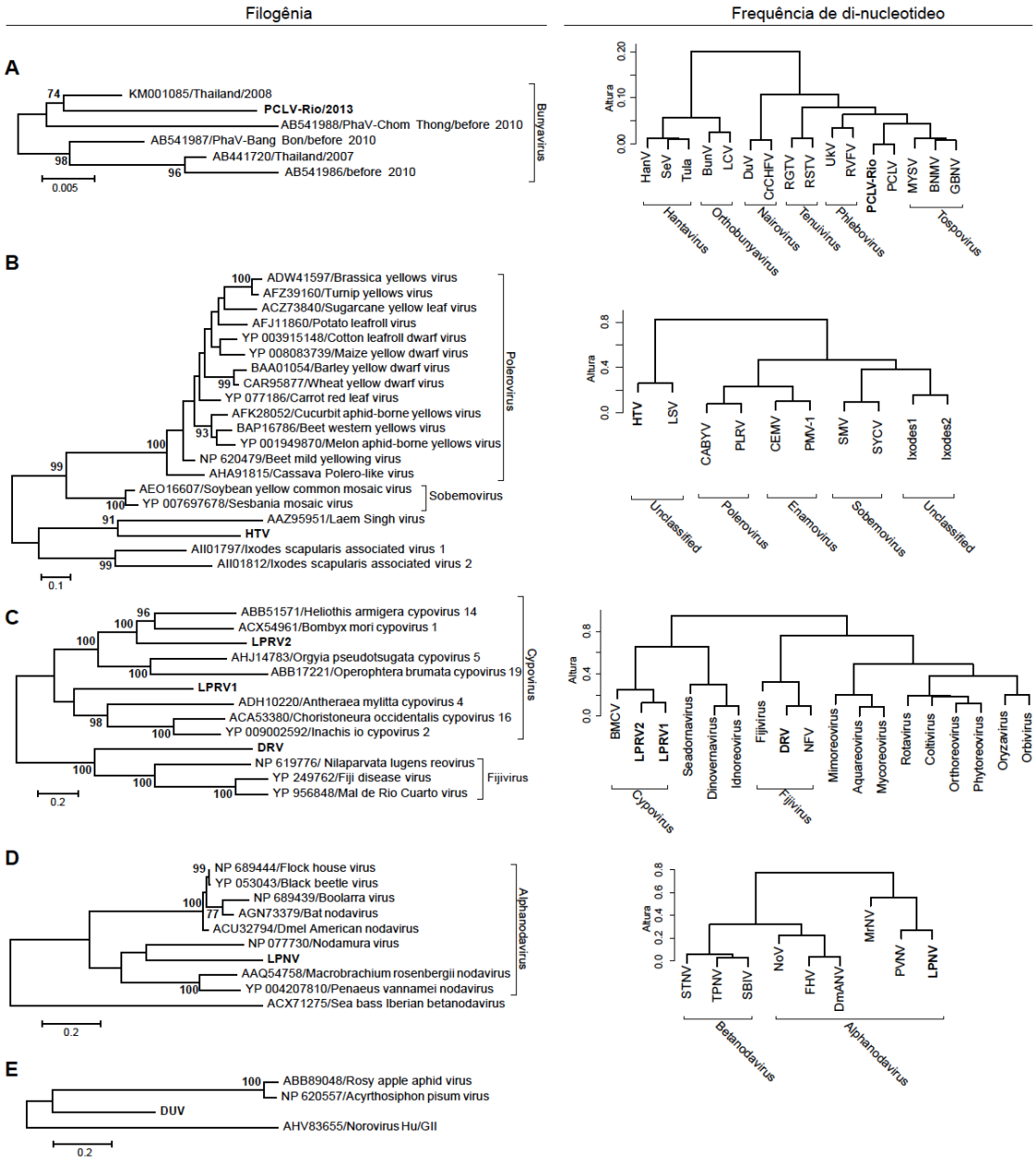

Figura 13. Vírus identificados através do sequenciamento de pequenos RNAs pertencem a diversas famílias de acordo com análises filogenéticas e frequência de di-nucleotídeos. (A) PCLV em mosquitos clusteriza com outras cepas virais da Tailândia com confiança segundo *bootstrap* de 74%, relação que é confirmada pela análise de di-nucleotídeos. (B) HTV também identificado em mosquitos é clusterizado com o vírus não classificado *Laem Singh vírus* (LSV), também vírus mais similar segundo blastP, com confiança segundo *bootstrap* de 91%. Não foi possível designar um gênero ao HTV, visto que os vírus relacionados não são classificados. Analise de di-nucleotídeo também clusterizou juntos o HTV e LSV separados dos vírus caracterizados. (C) Os dois reovírus identificados em flebotomíneos, LPRV1 e LPRV2 clusterizaram no mesmo cluster do gênero *Cypovírus* enquanto o reovírus encontrado em moscas, DRV, aparenta ser um membro distante do gênero *Fijivírus*. Nós designamos os três vírus a família *Reoviridae*, subfamília *Spinareovirinae*. As analises de di-nucleotídeo também corroboraram a classificação sugerida pela filogenia. (D) O vírus de flebotomíneos LPNV clusteriza com vírus do gênero *Alphanodavírus* da família *Nodaviridae*, resultado que é corroborado com a análise de di-nucleotídeos. (E) Norovírus é utilizado como *outgroup* na análise filogenética do DUV. Os vírus mais próximos ao DUV *Rosy apple aphid vírus* e *Acyrthosiphon pisum vírus* não tem táxon definido não tornando possível designar família ao DUV.

É importante ressaltar que todos as diferenças pontuais em nucleotídeos estavam também presentes nas sequências de pequenos RNAs das bibliotecas individuais antes da montagem dos *contigs*, o que sugeriu que essas diferenças poderiam vir da variação natural presente em populações virais (**dado não mostrado**).

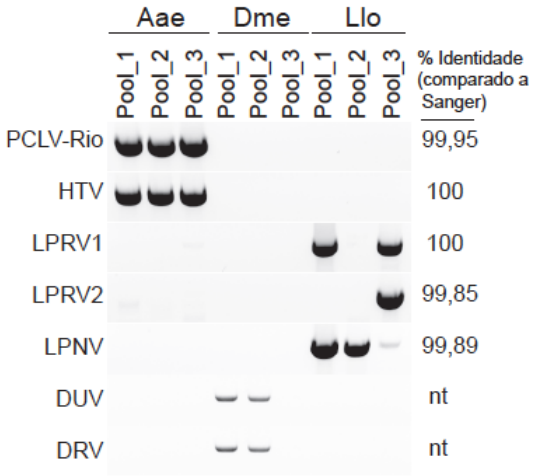

Figura 14. As sequências dos novos vírus identificados foram confirmadas por RT-PCR. Sequências virais de RNA foram detectadas através de RT-PCR do RNA total extraído de 3 pools separados de populações de *Drosophila*, *Aedes* e *Lutzomyia*. O sequenciamento de Sanger dos produtos de PCR mostrou alta identidade com as sequências montadas através do nosso pipeline de metagenômica como mostrado na coluna da direita (não testado– nt).

Notavelmente, a presença dos 7 vírus foi somente detectada nas populações de insetos correspondentes utilizadas para a construção das bibliotecas de pequenos RNAs onde eles foram identificados. Assim, os resultados indicaram que nossa estratégia baseada no sequenciamento de pequenos RNAs é robusta, confiável e não propicia a geração de artefatos.

Tabela 7. Sumário dos vírus identificados nas populações de laboratório de *Drosophila melanogaster*, Aedes *aegypti* e *Lutzomyia longipalpis*.

hospedeiro[a]	família viral	vírus	maior *contig* (nt)	Status[b] do segmento	# *contig*[c]	ID estratégia	melhor hit	E-value	Número de acesso (referência em nt)
Aae	*Bunyaviridae*	PCLV	3,936	CC	4	blastx	glycoprotein precursor [Phasi Charoen-like vírus]	0E+00	AIF71031.1 (3,852)
		PCLV	6,807	CC	23	blastx	RdRP [Phasi Charoen-like vírus]	0E+00	AIF71030.1 (6,783)
		PCLV	1,332	CC	3	blastx	nucleocapsid [Phasi Charoen-like vírus]	2E-72	AIF71032.1 (1,398)
	Unassigned	HTV	1,609	CC	8	blastx	structural protein precursor [Drosophila A vírus]	2E-65	YP_003038596.1 (1,326)
		HTV	2,793	CC	13	blastx	putative RdRP [Laem Singh vírus]	8E-34	AAZ95951.1 (507)
Llo	*Reoviridae*	LPRV1	3,762	CC	11	blastx	RdRP [Choristoneura occidentalis cypovírus 16]	3E-173	ACA53380.1 (3,675)
		LPRV1	3,687	CC	5	blastx	VP3 [Inachis io cypovírus 2]	1E-81	YP_009002593.1 (3,450)
		LPRV1	3,200	CC	2	blastx	VP4 [Inachis io cypovírus 2]	4E-63	YP_009002588.1 (3,201)
		LPRV1	1,842	CC	2	blastx	VP5 [Inachis io cypovírus 2]	2E-16	YP_009002589.1 (1,899)
		LPRV1	841	CC	1	blastx	polyhedrin [Simulium ubiquitum cypovírus]	6E-69	ABH85367.1 (836)
		LPRV1	3,685	CC	1	blastx	VP2 [Inachis io cypovírus 2]	5E-24	YP_009002587.1 (3,649)
		LPRV1	1,547	HQ	2	phmmer	unknown [Choristoneura occidentalis cypovírus 16]	9,00E-03	ABW87641.1 (1,946)
		LPRV1	2,237	CC	3	blastx	unknown [Choristoneura occidentalis cypovírus 16]	2,00E-01	ABW87640.1 (2,214)
		LPRV1	2,231	CC	1	padrão	-	-	-
		LPRV1	1,345	CC	1	padrão	-	-	-
		LPRV1	688	HQ	1	padrão	-	-	-
		LPRV1	680	HQ	1	padrão	-	-	-
		LPRV2	3,680	CC	1	blastx	RdRP [Bombyx mori cypovírus 1]	0E+00	AAK20302.1 (3,854)
		LPRV2	1,116	CC	1	blastx	polyhedrin [Heliothis armigera cypovírus 14]	4E-11	AAY34355.1 (956)
		LPRV2	2,043 +779 + 1,392	SD	3	blastx	VP1 protein [Dendrolimus punctatus cypovírus 1]	4E-70	AAN84544.1 (4,164)
		LPRV2	964	HQ	1	blastx	hypothetical protein LdcV14s9gp1 [Cypovírus 14]	2E-09	NP_149143.1 (1,141)
		LPRV2	678 +1,035 + 1,617	SD	3	blastx	VP3 [Bombyx mori cypovírus 1]	5E-14	ADB95943.1 (3,262)
		LPRV2	443 +579+769	SD	3	blastx	viral structural protein 4 [Bombyx mori cypovírus 1]	2E-10	ACT78457.1 (1,796)
		LPRV2	1,516	HQ	1	blastx	VP2 protein [Dendrolimus punctatus cypovírus 1]	8E-53	AAN86620.1 (3,846)
		LPRV2	599	HQ	4	blastx	unknown [Operophtera brumata cypovírus 18]	4E-10	ABB17215.1 (2,883)
		LPRV2	286	HQ	2	blastx	putative VP5 [Dendrolimus punctatus cypovírus 1]	3E-02	AAO61786.1 (1,501)
		LPRV2	641	SD	1	padrão	-	-	-
		LPRV2	1,212	SD	1	padrão	-	-	-
		LPRV2	1,174	CC	1	padrão	-	-	-
		LPRV2	976	SD	1	padrão	-	-	-
		LPRV2	535	SD	1	padrão	-	-	-
	Nodaviridae	LPNV	2,054	CC	5	blastx	capsid protein [Nudaurelia capensis beta vírus]	1E-42	NP_048060.1 (1,836)
		LPNV	3,189	CC	23	blastx	RdRP [Nodamura vírus]	9E-82	NP_077730.1 (3,129)
Dme	Indefinida	DUV	1,905+452	SD	2	blastx	protein P1 (RdRP) [Acyrthosiphon pisum vírus]	2E-63	NP_620557.1 (10,035)
	Reoviridae	DRV	635+175	SD	2	blastx	RdRP [Fiji disease vírus]	8E-05	YP_249762.1 (4,532)

a – *Aedes aegypti* (Aae), *Drosophila melanogaster* (Dme) e *Lutzomyia longipalpis* (Llo); **b** - Status do segmentos como descrito por Ladner et al (39): SD:

5.2.4. Classificação de sequencias virais utilizando o padrão de pequenos RNAs

Nossos resultados indicaram que as bibliotecas de pequenos RNAs favorecem a detecção de vírus quando em comparação com as bibliotecas de RNAs longos. Contudo, a maioria dos *contigs* montados a partir dos pequenos RNAs não foram passiveis de identificação através de analises baseadas em similaridade de sequências contra bancos de dados de referência (**Figura 8B**). Assim, nós percebemos a necessidade do desenvolvimento de uma estratégia que tornaria possível a caracterização de sequências virais com alta divergência com sequências previamente caracterizadas.

Analisando o perfil de tamanho dos pequenos RNAs produzidos pelas vias de resposta imune do hospedeiro percebemos que cada vírus analisado neste trabalho apresentou um padrão único, incluindo PCLV, SINV, VSV, DCV, HTV, LPNV, LPRV1, LPRV2, DRV e DUV **(Figura 9B** e **Figura 15A**). Adicionalmente, os perfis de pequenos RNAs observados para *contigs* derivados de outros organismos como fungos e bactéria apresentaram perfil tão único quanto os vistos nos vírus descritos (**Figura 15A**). No caso dos vírus de genoma multissegmentados, o perfil de tamanho dos pequenos RNAs foi notavelmente similar para os segmentos de origem do mesmo vírus como o PCLV e LPNV (**Figura 15A**).

Os perfis de tamanho dos pequenos RNAs foram consistentes com tamanhos característicos de diversas origens dos pequenos RNAs incluindo siRNAs (pico de tamanho em 21 nt), piRNAs (pico de tamanho em 27-28 nt) ou degradação do RNA viral (sem enriquecimento de tamanho, forte enviesamento de geração de pequenos RNAs correspondente a fita genômica) que são difíceis de se caracterizar visualmente (**Figura 9B e Figura 15A**). Dessa forma, nós decidimos desenvolver uma estratégia que tornasse possível a comparação entre estes diferentes perfis dos pequenos RNAs que levasse em consideração possíveis diferenças de expressão do mesmo RNA em bibliotecas diferentes. Para isso, nós utilizamos o calculo de *Z-score* para normalizar o perfil de tamanhos dos pequenos RNAs, que transforma o perfil de tamanho em padrão de

comportamento, método que já utilizado com sucesso na normalização de dados de expressão (Cheadle, Vawter et al. 2003). Contudo, apesar de normalizar os dados, ainda era necessário correlacionar os padrões do perfil de tamanho dos *contigs* e sequências de referencia para que fosse possível definir numericamente uma relação mínima entre eles. Para isso, nós utilizamos a correlação de Pearson que basea-se na relação linear entre os padrões comparados, que é o esperado para os padrões analisados. Assim, espera-se que o padrão de pequenos RNAs em um *contig* não caracterizado seja o mesmo do visto em um *contig* caracterizado, considerando o fato que eles tenham origem do mesmo vírus (por exemplo ambos tenham um pico no tamanho de 21 nt vindos de ambas as fitas). Nós utilizamos então o Z-score para gerar *heatmaps* (mapas de calor) para cada sequência que foi subsequentemente submetida a clusterização hierárquica baseada na correlação de *Pearson* da comparação par a par entre elas (**Figura 15B**). Através da utilização dessa estratégia, foi possível perceber que o perfil de tamanho dos pequenos RNAs oriundos de cada vírus apresentaram baixa correlação entre elas. Em contraste, os segmentos do PCLV, S, L e M, foram agrupados em um mesmo cluster (cluster 7) assim como os segmentos RdRP e o capsídeo do LPNV (cluster 5) (**Figura 15B**). Visto que os *contigs* representando segmentos genômicos do mesmo vírus foram agrupados juntos com alta correlação, > 0,92, nós testamos de que maneira a correlação do perfil de tamanho dos pequenos RNAs poderia nos ajudar na classificação de outras sequências que apresentaram similaridade suficiente a sequências derivadas de vírus, mas não puderam ser caracterizadas com uma maior riqueza de detalhes.

No caso dos flebotomíneos, baseado na análise das sequências codificadoras das RdRPs virais, nós identificamos com sucesso dois reovírus distintos, nomeados LPRV1 e LPRV2. Contudo, nós observamos outros 21 *contigs* não redundantes que apresentaram similaridade a reovírus que não puderam ser designados ao LPRV1 ou LPRV2 somente baseados na busca por similaridade. Como o perfil de tamanho dos pequenos RNAs da RdRP do LPRV1 e LPRV2 foram claramente distintos (**Figura 15A**), nós hipotetizamos que nós poderíamos utilizar o padrão apresentado por estes segmentos para designar a

origem das 21 sequências remanescentes que apresentaram similaridade com vírus da família *Reoviridea*.

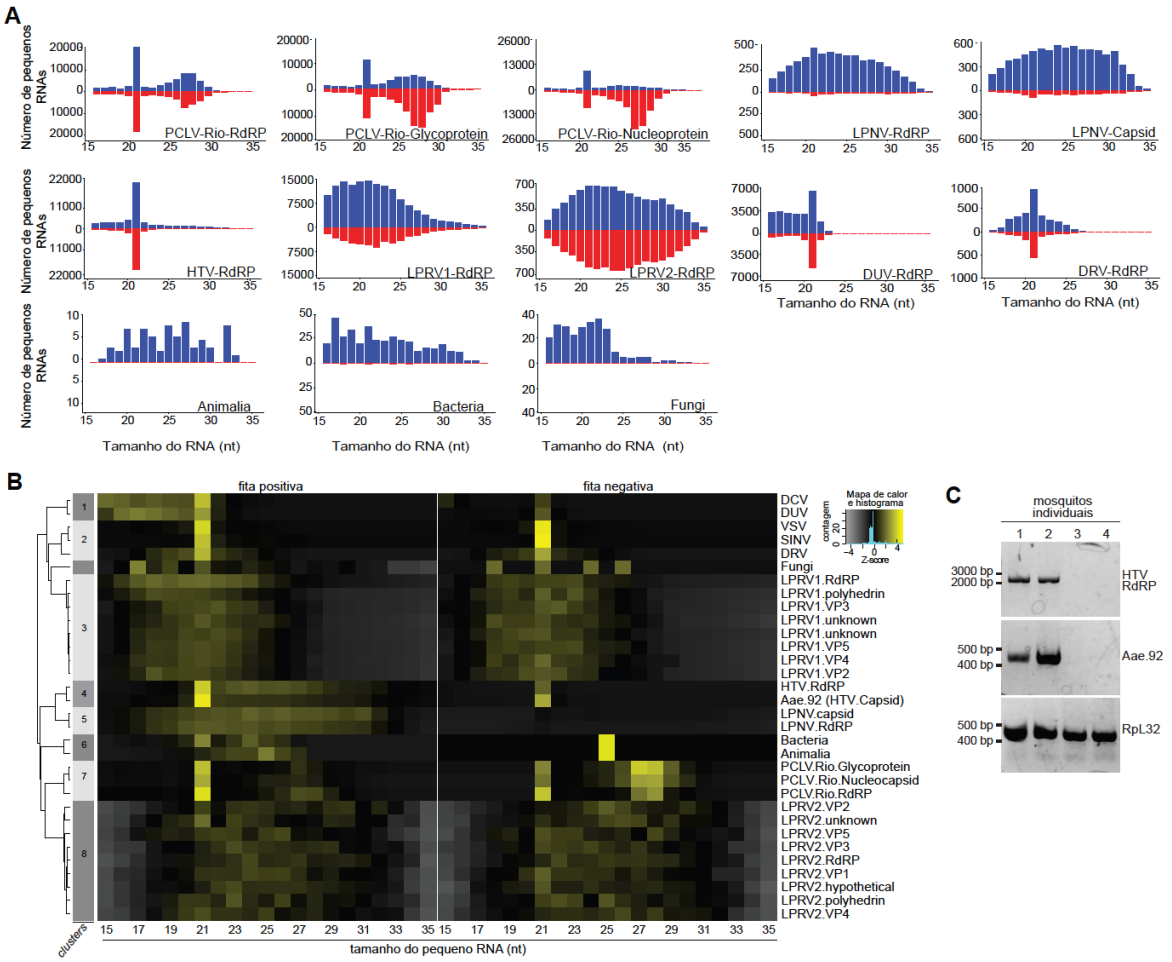

Figura 15. O perfil de tamanho dos pequenos RNAs é capaz de classificar *contigs* virais não caracterizados. (A) Perfil de tamanho dos pequenos RNAs de segmentos virais e *contigs* representativos dos reinos Animália, Bactéria e Fungi caracterizados identificados através de buscas por similaridade de sequência. Azul e vermelho representam pequenos RNAs derivados da fita positiva e negativa, respectivamente. (B) Clusterização hierárquica dos *contigs* virais montados em bibliotecas de moscas da fruta, mosquitos e flebotomíneos. A clusterização foi baseada em correlação de Pearson do perfil de tamanho dos pequenos RNAs mostrado como forma de mapa de calor '*heatmap*'. Clusters com mais de um elemento foram indicados na barra vertical da esquerda e numerados de acordo com a ordem com que eles apareceram de cima para baixo. Os clusters foram definidos através de correlação de Pearson acima de 0.8. (C) O *contig* Aae.92 e o segmento correspondente à RdRP do HTV que agruparam juntos nas análises perfil de tamanho dos pequenos RNAs no painel B mostraram perfeita correlação de expressão em mosquitos individuais como determinado por RT-PCR. Os resultados são representativos de mais de 20 mosquitos individuais que foram analisados. O gene *RpL32* foi usado como controle no RT-PCR.

Através do uso desta estratégia, nós observamos que, baseado na similaridade dos perfis de tamanho dos pequenos RNAs, 7 dos remanescente *contigs* agruparam junto com a RdRP do LPRV1 (cluster 3) enquanto que 14 dos remanescente *contigs*

formaram um cluster com a RdRP do LPRV2 (cluster 8) (**Tabela 7 e Figura 15B)**. Adicionalmente, a menor correlação observada entre os segmentos e a respectiva RdRP caracterizada foi 0,86, similar ao que foi visto para os segmentos do PCLV. Assim, como prova de conceito de funcionamento da nossa estratégia nós decidimos analisar a expressão dos *contigs* nos clusters 3 e 8 em comparação as RdRPs do LPRV1 e LPRV2 através de RT-PCR.

Consistente com o comportamento visto no perfil de tamanho dos pequenos RNAs, os *contigs* no cluster 3 foram detectados nas mesmas bibliotecas as quais foram identificados a RdRP do LPRV1 enquanto que os *contigs* contidos no cluster 8 seguiram o padrão de expressão visto na RdRP do LPRV2 (**dados não mostrados**).

Em mosquitos, nós identificamos um *contig* viral (Aae.92) com tamanho de 1.609 nt predito codificar uma proteína com domínio de capsídeo (PF00729). Este *contig* apresentou similaridade com a proteína de capsídeo do *Drosophila A vírus* (DAV), mas analises filogenéticas sugerem que os dois vírus são consideravelmente distintos (**Figura 13B e Figura 16)**.

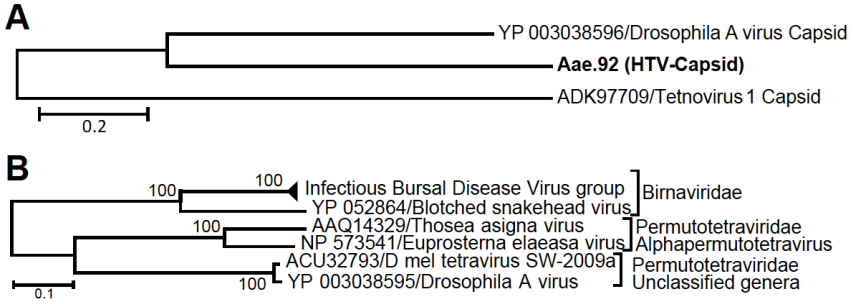

Figura 16. RdRP e capsídeo do HTV e DAV apresentam similaridade com vírus de diferentes famílias. (A) Análise filogenética do capsídeo do HTV e vírus mais próximos segundo similaridade de proteínas sugeri maior similaridade com DAV enquanto a RdRP do HTV é mais similar a outros vírus não caracterizados (**Figura 13**B). (B) Análises filogenéticas da RdRP do DAV e os vírus mais próximos segundo similaridade de sequência em proteína sugeri maior similaridade com vírus da família *Permutotetraviridae* com 100% de *bootstrap*.

As analises filogenéticas ainda sugeriram que o *contig* Aae.92 pertence a uma outra família viral distinta dos dois vírus encontrados nos mosquitos, PCLV e HTV. Contudo, nós observamos que o DAV é um vírus incomum, o qual apresenta as

sequências proteicas da RdRP e capsídeo com similaridade com famílias virais diferentes (**Figura 16**) (Ambrose, Lander et al. 2009). Notavelmente, o perfil de tamanho dos pequenos RNAs da RdRP do HTV e do *contig* Aae.92 apresentaram extraordinária similaridade e foram clusterizados juntos baseados na correlação do perfil de tamanho dos pequenos RNAs, com correlação > 0,98 (**Figura 15**). Desse modo, nós hipotetizamos que o *contig* Aae.92 codifica a proteína de capsídeo do HTV, desde que nós somente caracterizamos o segmento correspondente a RdRP do vírus. Em concordância com a nossa hipótese, nós observamos 100% de correlação entre a detecção por RT-PCR do *contig* Aae.92 e a RdRP do HTV avaliados em mosquitos individuais (**Figura 15C**). Assim, nossos resultados sugeriram a análise do perfil de tamanho dos pequenos RNAs pode ser utilizada como uma ferramenta importante na classificação se sequencias virais.

5.2.5. Identificação de *contigs* virais utilizando uma estratégia baseada no padrão de pequenos RNAs

Nós observamos que os *contigs* derivados de vírus apresentaram perfil de tamanho dos pequenos RNAs único para cada vírus e que isso pode ser utilizado para designar a origem de sequências a vírus específicos em nossas amostras. Baseado nesse fato, nós levantamos a hipótese que nossa estratégia baseada no padrão dos pequenos RNAs derivados dos vírus poderia também ajudar na identificação de *contigs* não caracterizados independente de buscas por similaridade de sequência.

Para selecionar *contigs* não caracterizados candidatos para posterior análise por padrão nós observamos métricas obtidas dos *contigs* virais montados em nossas bibliotecas de pequenos RNAs e percebemos que os *contigs* virais são significantemente maiores do que os *contigs* não virais (N50 de 208 nt em comparação a 63 nt para *contigs* não virais). Além disso, nós já havíamos observado que na etapa de padronização da estratégia que o N50 é uma boa métrica para se detectar falsos positivos (**Figura 5**). Dessa forma, nós utilizamos o N50 como forma de filtrar os 10.577 *contigs* não caracterizados, resultando na seleção de 106 *contigs* candidatos maiores que 208 nt. Nós

eliminamos a redundância entre os *contigs* selecionados, o que resultou em 79 *contigs* únicos que foram nomeados de acordo com a sua biblioteca de origem, *Lutzomyia* (Llo), *Drosophila* (Dme) ou *Aedes* (Aae).

Nós avaliamos o perfil de pequenos RNAs dos 79 *contigs* não caracterizados os quais foram comparados a *contigs* virais previamente caracterizados utilizando clusterização hierárquica. Como visto na seção anterior, os *contigs* de origem do mesmo visto tendem a apresentar uma alta correlação do padrão de pequenos RNAs. Assim, nós utilizamos a correlação mínima de 0,8 para definição dos clusters de similaridade. A análise de clusterização resultou em 17 clusters contendo mais de um *contig*, os quais foram nomeados sequencialmente de acordo com sua posição no *heatmap*. Nós observamos que 72 dos 79 *contigs* únicos foram agrupados em 11 clusters de similaridade diferentes (**Figura 17**). De forma interessante, os clusters agruparam exclusivamente de forma organismo-especifica nas bibliotecas do mesmo inseto, *Lutzomyia*, *Drosophila* ou *Aedes*.

Os *contigs* não caracterizados encontrados nas bibliotecas de *Lutzomyia* foram agrupados em 3 clusters separados que claramente mostraram distintos padrões de pequenos RNAs (Clusters 2, 6 e 7 na **Figura 17**). O cluster 6 continha *contigs* com o perfil de pequenos RNAs consistente com piRNAs de insetos (pico entre 27-28 nt) sugerindo que eles poderiam representar elementos transponíveis (**Figura 17**). O cluster 2 continha 4 *contigs* não caracterizados que foram agrupados e mostraram alta correlação com segmentos do LPRV1 previamente caracterizados (destacados em vermelho). O cluster 17 continha 19 *contigs* não caraterizados que apresentaram razoável correlação com segmentos do LPRV2 (**Figura 17**). Notavelmente, no cluster 7, nós observamos que 5 dos 19 *contigs* formaram um subgrupo com correlação maior que 0,93 com a RdRP do LPRV2 (destacado em vermelho). Esses resultados nos indicaram que alguns desses *contigs* não caracterizados nas bibliotecas de *Lutzomyia* poderiam representar segmentos adicionais do LPRV1 e LPRV2.

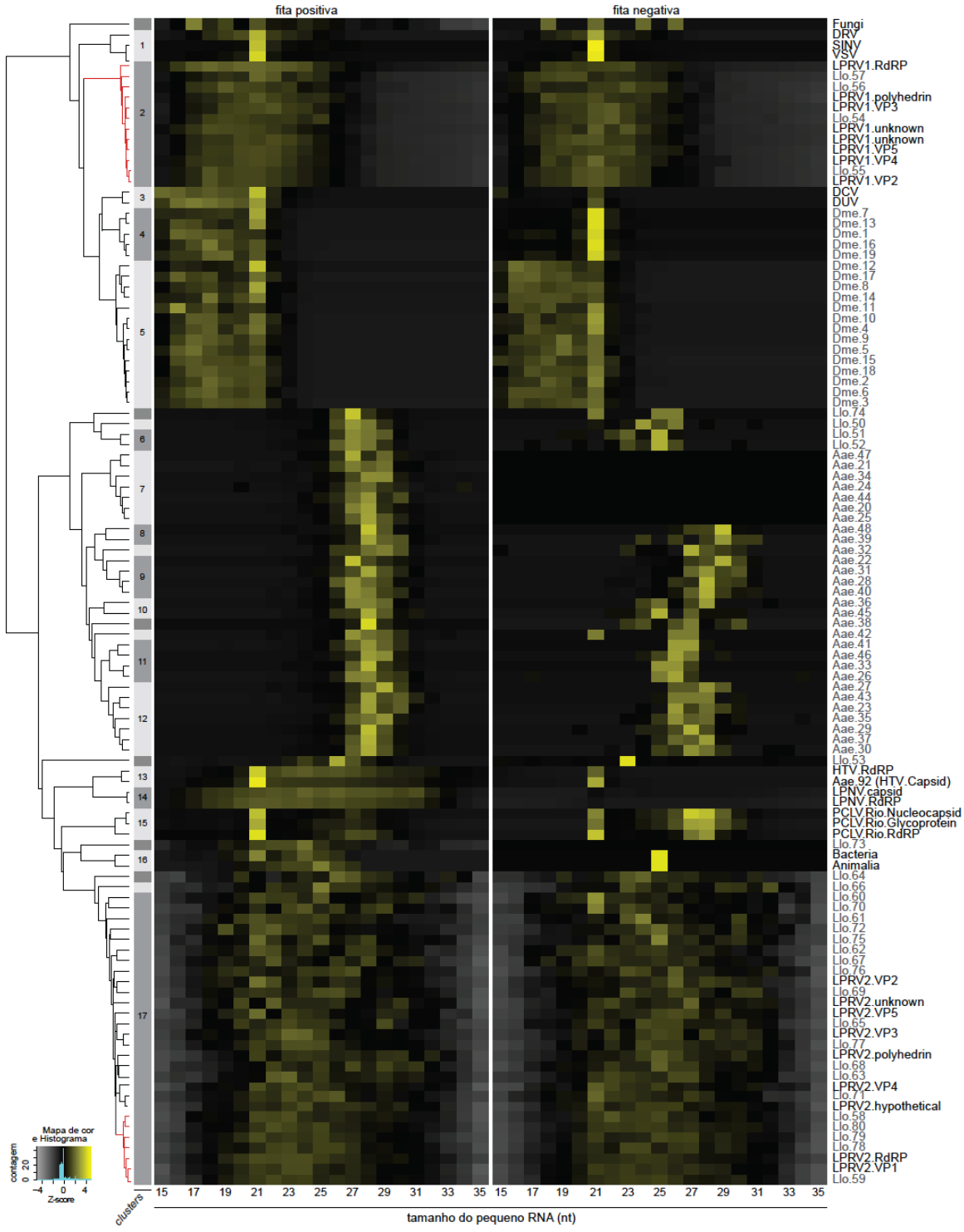

Figura 17. Análise baseada no padrão de pequenos RNAs é capaz de identificar sequências virais independente de similaridade de sequência em bancos de dados de referência. Clusterização hierárquica dos *contigs* virais e não caracterizados montados a partir das bibliotecas de pequenos RNAs de moscas, mosquitos e flebotomíneos. A clusterização foi baseada em correlação de Pearson do perfil de tamanho dos pequenos RNAs mostrado como forma de mapa de calor '*heatmap*'. Clusters com mais de um elemento foram indicados na barra vertical da esquerda e numerados de acordo com a ordem com que eles apareceram de cima para baixo. Os clusters foram definidos através de correlação de Pearson acima de 0.8.

É importante ressaltar que, devido a natureza multi-segmentada do genoma dos reovírus, nós esperávamos encontrar mais segmentos de ambos os LPRVs do que os detectados através de análises baseadas em similaridade de sequência (**Tabela 7**). Com o objetivo de investigar essa possibilidade, nós analisamos a expressão de *contigs* não caracterizados destacados no cluster 2 e cluster 17 que apresentaram as maiores correlações com perfil de pequenos RNAs dos segmentos da RdRP do LPRV1 e LPRV2, respectivamente.

Todos os 4 *contigs* não caraterizados no cluster 2 seguiram o mesmo padrão de expressão da RdRP do LPRV1 enquanto os 5 *contigs* não caraterizados do cluster 17 copiaram a expressão da RdRP do LPRV2 (**Figura 18**). Nenhum dos 9 novos *contigs* dos LPRVs apresentaram similaridade suficiente com outras sequências de reovírus depositadas em bancos de dados de referência, o que sugere que eles são segmentos menos conservados do genoma do vírus. Somente um dos nove *contigs* não caracterizados, Llo.58, designado ao LPRV2, apresentou uma ORF completa que foi predita codificar uma proteína de 361 aminoácidos contendo dois possíveis domínios (**Figura 19**).

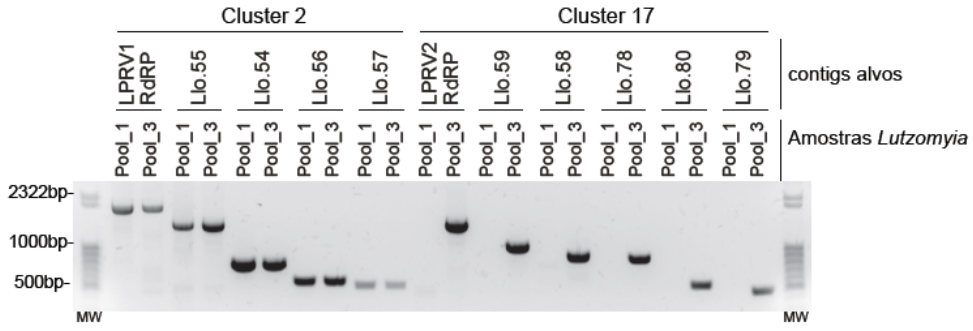

Figura 18. Detecção por RT-PCR dos *contigs* caracterizados através de análise baseada no padrão de pequenos RNAs. Detecção por RT-PCR em duas amostras de flebotomíneos mostraram que os *contigs* no cluster 2 e 17 apresentaram o mesmo padrão de expressão das RdRPs do LPRV1 e LPRV2, respectivamente. Foram utilizadas as mesmas amostras de *Lutzomyia longipalpis* (pool1 e pool3) analisadas na **Figura 14**.

O primeiro domínio é de metallopeptidase dependente de Zinco da superfamília Astacina que é encontrada em enzimas digestivas de animais vertebrados e invertebrados (Wang and Granados 2001, Arolas, Vendrell et al. 2007). O segundo domínio é de Peritrophina-A encontrada em proteínas de ligação de chitina que inclui proteínas de matrix peritrófica das chitinases de insetos que também são encontradas em baculovírus (Lepore, Roelvink et al. 1996). Assim, o *contig* Llo.58 poderia codificar uma proteína que está relacionada na interação entre o LPRV2 e o flebotomíneo visto que vírus notoriamente sequestram e modificam proteínas do hospedeiro para beneficio próprio. Genes envolvidos na interação patógeno-hospedeiro tendem a ser os mais divergente entre os vírus, devido a pressão constante sofrida na interação com o sistema imune do hospedeiro (Marques and Carthew 2007).

É importante ressaltar que o Llo.58 não havia sido detectado por analises de similaridade de sequência contra bancos de dados de referência e não poderia ser classificado como viral somente através da predição de domínio, visto que proteínas com esse domino também podem ser encontradas em proteínas celulares do hospedeiro. Assim, analise baseada do perfil de tamanho dos pequenos RNAs identificaram 23 *contigs* não caracterizados representando adicional segmentos dos genomas do LPRV1 e LPRV2 que não apresentavam similaridade significativa com sequências previamente identificadas depositadas em bancos de dados de referência.

Contigs não caracterizados encontrados nas bibliotecas de Drosophila foram agrupados em 2 clusters separados. O cluster 4 inclui 5 *contigs* não caracterizados que apresentaram alta similaridade com o cluster contendo DUV e DCV (**Figura 17**). O cluster 5 incluiu outros 14 *contigs* não caracterizados que mostravam similaridade com os perfis do DUV e DCV, embora que em menores valores que o cluster 4 (**Figura 17**). Considerando que o genoma completo do DCV é conhecido, estes *contigs* não caracterizados em dois clusters separados provavelmente representam diferente *contigs* do DUV. Reforçando essa hipótese, nós somente identificamos 2 *contigs* do DUV correspondentes a RdRP viral, o que representa somente uma pequena porcentagem do

genoma completo. Adicionalmente, nós somente encontramos esses *contigs* nas bibliotecas de *Drosophila* onde o DUV foi identificado.

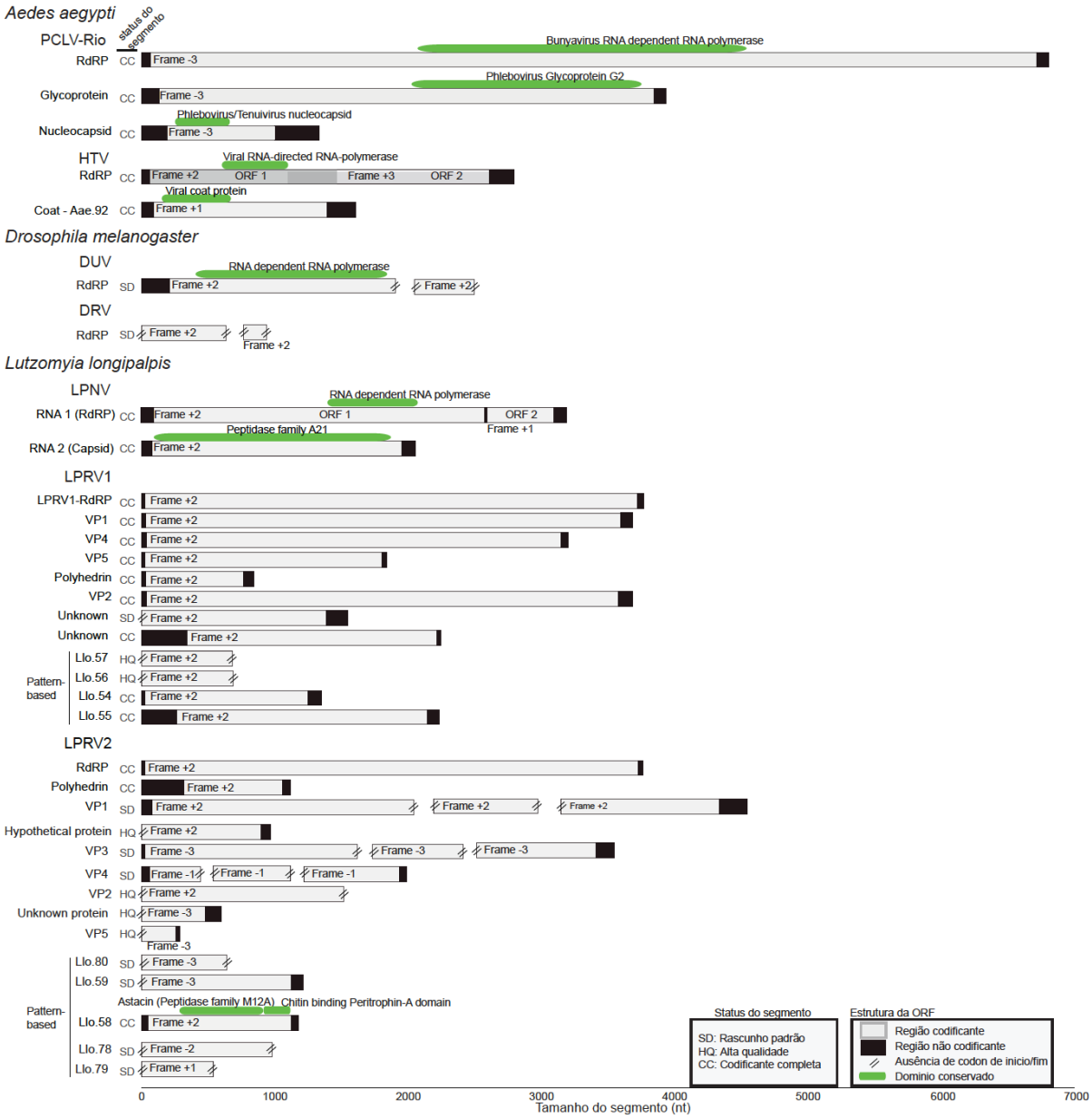

Figura 19. Análise da organização de ORFs e domínios das sequências virais identificadas através da nossa estratégia. Os *contigs* virais são mostrados na mesma escala destacando ORFs preditas e domínios conservados em cinza e verde, respectivamente. O status das sequências virais em relação ao segmento completo esperado e presença de códon de inicio/fim são mostrados. O status da anotação/montagem das novas sequências virais seguiu as orientações previamente descritas por Ladner et al (39).

Nas bibliotecas de *Aedes,* 24 dos 27 *contigs* não caraterizados foram agrupados em 6 clusters diferentes (7,8 ,9,10, 11 e 12) que mostraram alta correlação uns com os

outros e perfil de tamanho de pequenos RNAs consistente com piRNAs de mosquito (pico de 27-28 nt na distribuição de tamanho) (**Figura 17**) (Arensburger, Hice et al. 2011, Morazzani, Wiley et al. 2012). De acordo com o perfil de tamanho, os pequenos RNAs derivados destes *contigs* mostraram enriquecimento para U na posição 1 e A na posição 10, perfil típico de piRNAs de insetos, mas não pico substancial de 21 nt ou cobertura simétrica nas duas fitas. Dessa forma, essas sequências representam possivelmente regiões repetitivas do genoma que geram piRNAs abundantemente, mas não foram montados na versão corrente do genoma do *Aedes aegypti*.

Um resumo das sequências virais identificadas e sua respectiva organização de ORFs e presença de domínios conservados pode ser visualizada na **Figura 19**.

5.2.6. A análise do perfil de pequenos RNAs e sua relação com a biologia do vírus

O perfil de pequenos RNAs gerados pela resposta do hospedeiro depende de características do vírus como a estrutura do genoma, tropismo tecidual e estratégia de replicação. Assim, além de identificação de *contigs* virais, o perfil de tamanho dos pequenos RNAs pode prover informação especifica sobre a biologia de cada vírus. Por exemplo, vírus de RNA tendem a ter perfil de cobertura dos pequenos RNAs homogênea no genoma viral enquanto que vírus de DNA apresentam clara cobertura heterogênea onde somente algumas regiões tem cobertura de pequenos RNAs (Mueller, Gausson et al. 2010, Kemp, Mueller et al. 2013, Marques, Wang et al. 2013). Todos os *contigs* virais descritos neste trabalho foram derivados de vírus de RNA e em sua grande maioria apresentaram cobertura homogênea de pequenos RNAs.

Nós percebemos também que o HTV e PCLV apresentaram perfil de tamanho dos pequenos RNAs distinto apesar de algumas vezes serem encontrados nos mesmos mosquitos (**Figura 15**). O perfil foi HTV apresentou um claro pico de 21 nt consistente com a produção de siRNAs. Em contraste, o perfil do PCLV apresentou 2 picos separados, 21 nt e 26-29 nt, consistente com o padrão de pequenos RNAs gerados por

ambas as vias, siRNA e piRNA respectivamente (**Figura 15A**). Realmente, os pequenos RNAs de 24-29 nt derivados do PCLV mostraram enriquecimento para U na posição 1 e A na posição 10, típico de piRNAs de inseto das fitas senso e antisenso respectivamente (**Figura 20A**).

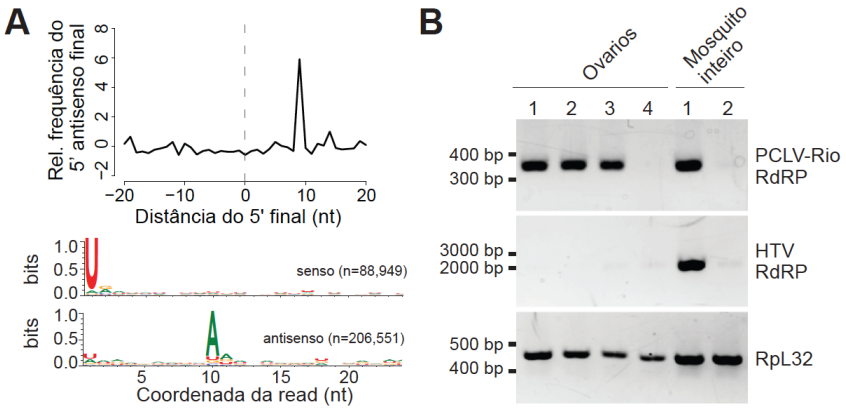

Figura 20. A presença de piRNAs derivados de vírus com assinatura de ping-pong é indicativo de infecção do ovário. (A) Pequenos RNAs de 24-29 nt derivados do PCLV apresentaram sobreposição de 10 nt entre a fita senso e antisenso e enriquecimento de U na posição 1 da fita e enriquecimento de A na posição 10, consistente com piRNAs gerados através do mecanismo de amplificação de ping-pong encontrado em linhagem germinativa de insetos. (B) O PCLV e HTV são detectados em mosquitos inteiros individuais, contudo somente o PCLV é encontrado nos ovários como determinado por RT-PCR. O gene *RpL32* foi utilizado como controle para o RT-PCR.

A via de piRNA de insetos é principalmente ativada na linhagem germinativa onde dois mecanismos de biogênese dos pequenos RNAs pode ocorrer (Malone, Brennecke et al. 2009). piRNAs primários são gerados através processamento endonucleolitico de um transcrito precursor enquanto piRNAs secundários são produzidos por um *loop* de amplificação, referido como mecanismo de *ping-pong*. Nós observamos que os pequenos RNAs de 24-29 nt derivados do PCLV apresentaram sobreposição de 10 nt entre os pequenos RNAs senso e antisenso, consistente com o mecanismo de *ping-pong* de amplificação (**Figura 20A**). Esses resultados sugeriram que o PCLV induz a produção de piRNAs através do mecanismo de *ping-pong* de amplificação quando infectado a linhagem germinativa do inseto. Em acordo com esta hipótese, nós detectamos o PCLV nos ovários de mosquitos individuais (**Figura 20B**). Em contraste, HTV não foi

encontrado em ovários consistente com o fato do vírus não gerar piRNAs (**Figura 20B**). Assim, a presença de uma assinatura clara de piRNAs no perfil de pequenos RNAs poderia ajudar a inferir tropismo tecidual, indicando infecção da linhagem germinativa do inseto.

Além da assinatura gerada pela via de piRNA, a ausência de picos claros no perfil de tamanho dos pequenos RNA pode sugerir inibição das vias de RNAi como os reportados no *Flock House vírus* (FHV). De fato, a proteína B2 codificada pelo FHV é uma potente supressora de silenciamento que bloqueia a via de RNAi (Han, Luo et al. 2011). De forma interessante, ORF 2 no RNA 1 do genoma do LPNV foi predita codificar uma proteína, com similaridade com a B2 do FHV, que poderia agir como uma supressora da via de RNAi (**Figura 19 e Figura 21**). Assim, o perfil amplo de tamanho de pequenos RNA sem picos claros observado no LPNV poderia sugerir inibição das vias de RNAi como uma estratégia de replicação do vírus (**Figura 15A**). Um amplo perfil de tamanho e forte preferência de pequenos RNAs gerados a partir da fita positiva do genoma viral também foi observado para o DCV e DUV em moscas da fruta infectadas (**Figura 15A**). Dessa forma, visto que o gene DCV-1A codifica um potente supressor da via de siRNA (van Rij, Saleh et al. 2006), nosso resultado sugere que o DUV poderia também ser capaz de suprimir a via de RNAi em moscas infectadas.

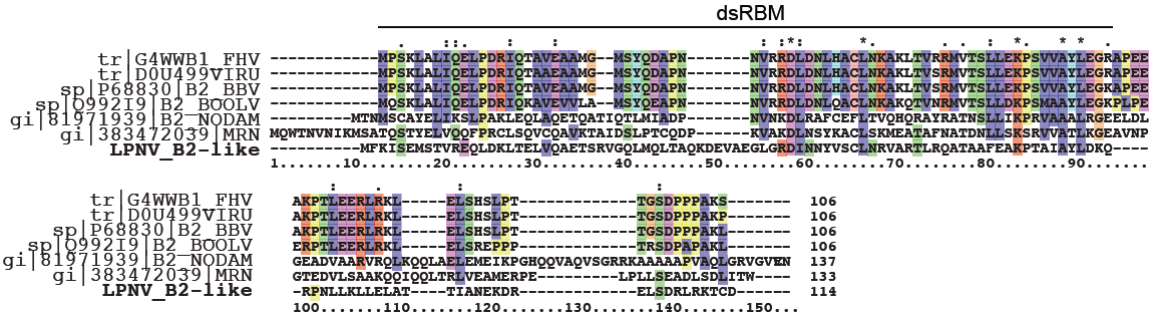

Figura 21. Análise da proteína B2-like do LPNV. Alinhamento de sequências em proteína entre a sequência putativa codificada pelo RNA1 do LPNV e outras proteínas B2 derivadas de Alphanodavírus. Apesar da conservação, o alinhamento em proteína sugeri que faltam diversos resíduos conservados no motivo de ligação do dsRNA (dsRBM) na LPNV B2-like.

5.3. Aplicação do pipeline para análise da dinâmica temporal e espacial do viroma de mosquitos de campo

Um número significativo de vírus, conhecidos como arbovirus são vírus transmitidos a animais vertebrados por artrópodes vetores (Weaver and Barrett 2004). Mais de 500 arbovirus diferentes já foram descritos até o momento (Gubler 2001). Dentre os artrópodes, os mosquitos do gênero *Aedes* são vetores de arbovirus importantes para a saúde pública humana como o DENV, *Yellow Fever vírus* (YFV), Chikungunya vírus (CHK) e *Zika vírus* (ZiV) (Hill, Kafatos et al. 2005, Beaty, Prager et al. 2009, Diallo, Sall et al. 2014, Vega-Rua, Lourenco-de-Oliveira et al. 2015). Existe ainda a ameaça da emergência de novos arbovirus, ainda desconhecidas, que poderiam infectar humanos (Gubler 2001). Neste contexto, a análise metagenômica de mosquitos pode ser utilizada como uma estratégia importante de vigilância para se detectar e identificar vírus circulantes na natureza.

O nosso pipeline identificou com sucesso vírus presentes em moscas de laboratório de *Drosophila* e populações selvagens de *Aedes* e *Lutzomyia* mas que foram mantidas em laboratório por algumas gerações. Assim, nós decidimos testar a aplicação da nossa estratégia em mosquitos capturados diretamente do campo. Este projeto foi uma a colaboração com a Dra. Ana Paula Vilela sobre supervisão da Dra. Erna Kroon do Laboratório de Vírus-UFMG. Nesta colaboração, a Dr. Vilela foi responsável pela coleta de mosquitos de campo na cidade de Caratinga entre os anos de 2010-2011 (Vilela 2013) (**Figura 22**).

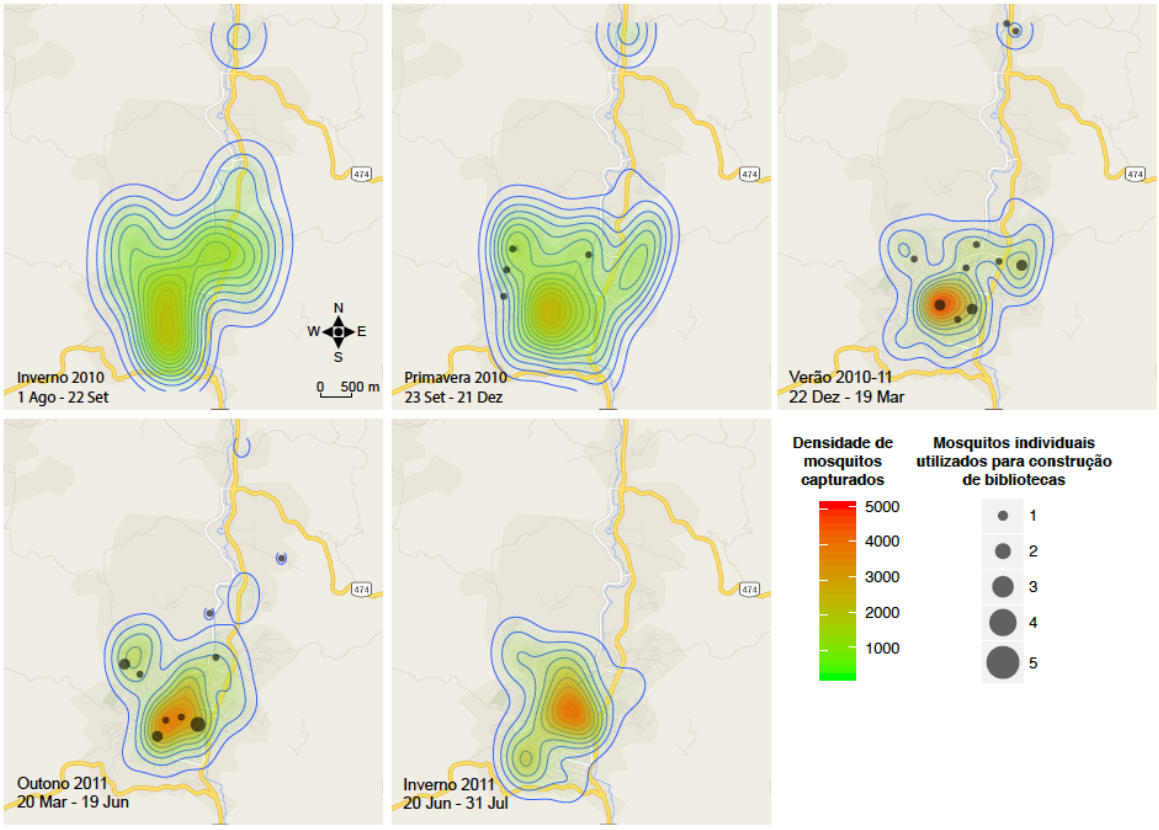

Figura 22. Mapa de captura de mosquitos de campo por estações do ano na cidade de Caratinga. O limite da cidade é delimitado pela área cinza escura. As principais rodovias que cortam a cidade são mostradas em amarelo. A densidade de mosquitos capturados é mostrada de verde para vermelho, onde verde representa as regiões com menor densidade e vermelho com maior densidade de mosquitos capturados. Os círculos em cinza representam os mosquitos individuais utilizados para preparação das bibliotecas de pequenos RNAs.

Os mosquitos foram coletados individualmente e os dados como data da coleta, espécie, sexo e localização anotados para tornar possível realizar correlações, como sazonalidade e geolocalização. A coleta foi realizada através da utilização de MosquiTRAP ™ (Version 3.0, Ecovec Ltda, Brazil) (Álvaro Eduardo Eiras 2009), que são otimizadas para captura de mosquitos de gênero *Aedes*. As armadilhas foram posicionadas durante toda a extensão da cidade, incluindo região urbana e peri-urbana, com uma média de 200m de distância entre elas e checadas semanalmente. No total foram coletados 762 mosquitos das diferentes regiões da cidade durante o período de Agosto de 2010 e Julho de 2011. Destes, 86% eram Aedes albopictus e 14% Aedes aegypti.

Com o intuito de avaliar o viroma circulante nos mosquitos de campo, nós selecionamos 32 mosquitos individuais. Os mosquitos foram escolhidos baseado na qualidade do RNA analisado em Bioanalyzer e na tentativa de utilizar indivíduos coletados em áreas distintas da cidade e em diferentes meses ao longo do ano. O resumo do resultado do sequenciamento pode ser visualizado na **Tabela 8**.

Tabela 8 - Resumo das bibliotecas de pequenos RNAs derivados de mosquitos individuais coletados no campo.

biblioteca	espécie	sexo	# total de sequências	# sequências após trimagem	# sequências mapeadas no hospedeiro	# sequências após pré-processamento	# *contigs* totais montados
RKPM26	*A. aegypti*	F	22.354.271	19.285.763	18.027.991	1.255.969	16
RKPM27	*A. albopictus*	F	20.269.384	16.929.882	10.900.230	6.028.562	5
RKPM28	*A. aegypti*	F	16.991.811	13.646.634	12.423.936	1.221.456	152
RKPM29	*A. aegypti*	F	15.107.645	13.193.646	11.676.526	1.515.952	91
RKPM30	*A. aegypti*	F	18.261.721	15.612.320	13.617.970	1.992.988	52
RKPM31	*A. aegypti*	F	16.645.291	14.070.387	12.402.221	1.666.926	70
RKPM32	*A. aegypti*	F	16.754.889	14.182.775	12.657.732	1.523.777	63
RKPM33	*A. aegypti*	F	18.719.549	16.999.343	16.169.309	828.417	124
RKPM34	*A. aegypti*	F	13.858.564	9.988.175	9.240.894	746.357	127
RKPM35	*A. aegypti*	F	16.967.849	10.789.059	10.363.310	424.713	135
RKPM36	*A. aegypti*	F	20.468.332	14.365.014	13.598.190	765.464	90
RKPM37	*A. aegypti*	F	20.401.791	17.927.318	16.442.857	1.482.817	77
RKPM38	*A. aegypti*	F	17.469.864	14.532.276	13.299.623	1.231.323	72
RKPM39	*A. aegypti*	F	18.055.683	11.179.203	10.794.357	383.767	104
RKPM40	*A. aegypti*	F	19.841.994	17.475.123	15.228.023	2.245.577	29
RKPM41	*A. aegypti*	F	18.570.558	16.149.810	14.021.641	2.126.767	93
RKPM42	*A. aegypti*	F	24.173.561	20.299.307	18.122.015	2.175.480	99
RKPM43	*A. aegypti*	F	29.091.503	24.841.977	23.044.146	1.795.527	92
RKPM44	*A. aegypti*	F	24.732.700	21.595.029	18.036.863	3.556.362	121
RKPM45	*A. aegypti*	F	33.615.150	26.674.362	25.597.925	1.073.877	112
RKPM46	*A. aegypti*	F	22.841.569	20.370.348	18.389.993	1.978.516	72
RKPM47	*A. aegypti*	F	25.127.607	21.257.610	19.689.180	1.566.461	80
RKPM48	*A. aegypti*	M	32.060.499	25.483.917	24.153.840	1.327.662	96
RKPM49	*A. aegypti*	F	26.896.405	21.955.904	20.788.367	1.165.458	0
RKPM50	*A. aegypti*	F	30.925.064	22.078.698	20.895.435	1.181.173	81
RKPM51	*A. aegypti*	F	30.949.497	24.449.452	22.546.104	1.901.093	64
RKPM52	*A. aegypti*	M	28.698.423	25.419.264	22.803.286	2.613.698	128
RKPM53	*A. aegypti*	F	26.157.464	21.488.567	20.032.496	1.454.068	91
RKPM54	*A. aegypti*	F	24.976.855	21.191.895	17.790.872	3.399.244	55
RKPM55	*A. aegypti*	F	42.029.122	36.925.138	32.557.232	4.364.650	37
RKPM56	*A. aegypti*	F	15.352.633	13.594.669	12.099.412	1.494.047	52
RKPM57	*A. aegypti*	F	28.647.112	18.242.393	17.569.563	671.073	94

F: Fêmea; M: Macho

Das 32 bibliotecas de pequenos RNAs de mosquitos individuais, 31 eram derivadas de *Aedes aegypti* e 1 de *Aedes albopictus*. É importante salientar foram capturados cerca de 6 vezes menos *Aedes albopictus* do que *Aedes aegypti*, o que justifica sua subrepresentação nas bibliotecas sequenciadas. O sequenciamento das

bibliotecas de pequenos RNAs de mosquitos individuais gerou em média 23 milhões de sequências. Após as etapas de pré-processamento, em média 1,7 milhões de sequências foram utilizadas para as etapas posteriores de montagem e caracterização de sequência virais. É importante destacar a etapa da nossa estratégia que utiliza o genoma de referência do hospedeiro para retirar sequências não virais. Como o genoma de *Aedes albopictus* ainda não foi sequenciado, para a única biblioteca desta espécie, nós utilizamos o genoma de *A. aegypti*, a espécie mais próxima com genoma disponível, o que resultou em um número menor de sequências filtradas (**Tabela 8**).

5.3.1. Identificação do viroma de mosquitos de campo

Para avaliar o viroma presente nos mosquitos sequenciados nós aplicamos o pipeline de detecção e caracterização de vírus descritos nas seções anteriores (**Figura 7**). Nas bibliotecas de mosquitos individuais de campos nós observamos que foi possível a montagem de *contigs* em 31 das 32 amostras sequenciadas. A completa ausência de contigs na biblioteca RKPM49 é ainda um mistério pois o número de sequências e perfil de mapeamento no genoma do mosquito que é similar ao observado para todas as outras. Para as outras bibliotecas, todos os *contigs* montados foram submetidos a análises baseadas em similaridade de sequência em contra bancos de dados de referência para determinação de sua origem. De forma geral, as bibliotecas apresentaram um perfil similar ao visto para os mosquitos analisados anteriormente, onde predominaram *contigs* cuja táxon de origem é desconhecido seguido dos de origem animal e viral (comparar **Figura 8 e Figura 23**). Contudo, apesar do grande número de *contigs* não caracterizados, a proporção de *contigs* derivados de vírus nas bibliotecas de mosquitos individuais de campo foi mais variável que o observado nas populações de *Aedes.* A maior homogeneidade das amostras descritas na **Figura 8** é provavelmente o resultado destas serem provenientes do sequenciamento de grupos de 6 mosquitos que foram criados em laboratório mesmo vindo de ovos coletados em campo. Logo, as

bibliotecas de pequenos RNAs de mosquitos coletados em campo devem refletir melhor a variabilidade observada entre indivíduos.

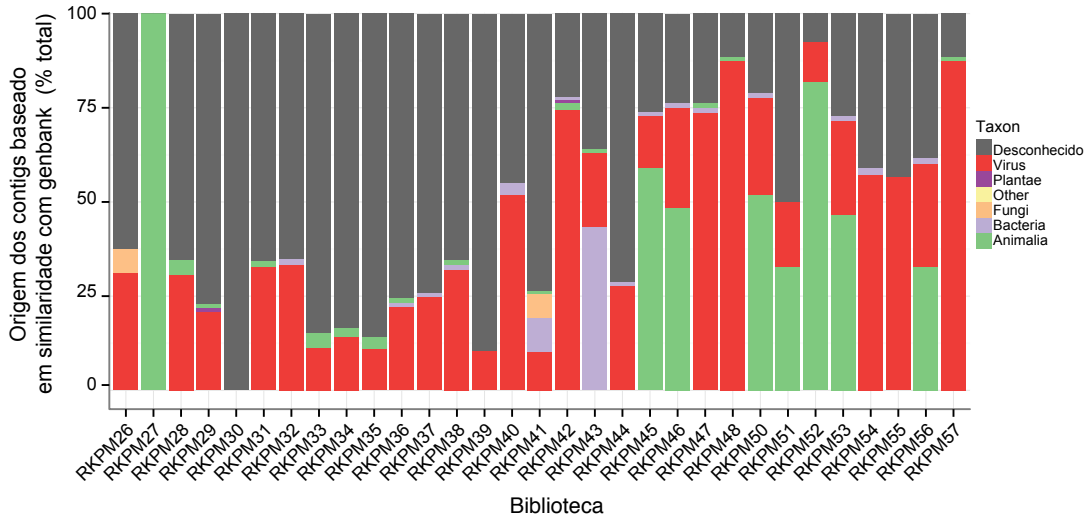

Figura 23. Distribuição de sequências por reino de origem em cada mosquito analisado. *Contigs* montados em cada biblioteca de pequenos RNAs sequenciada que apresentaram similaridade significante com bancos de dados de sequências de referência. A origem dos *contigs* foi classificada por táxon e inclui sequências não caracterizadas designadas ao táxon "Desconhecido".

Das 31 bibliotecas, somente duas não apresentaram *contigs* com similaridade significante com sequências derivadas de vírus incluindo a biblioteca RKPM27 única derivada de *Aedes albopictus* (**Figura 23)**. Nesta biblioteca, a maioria dos *contigs* montados tem origem animal e são provavelmente derivados das sequências do próprio mosquito que não foram filtradas devido a ausência de um genoma de referência. A outra biblioteca sem *contigs* virais foi RPKM30 na qual todos os *contigs* montados apresentaram origem desconhecida. Contudo, apesar de ter montado 52 contigs, o tamanho médio dos contigs foi 63 nt, o que pode ter dificultado as análises baseadas em similaridade de sequência.

Considerando somente os *contigs* de origem viral, nós fizemos a remoção de redundância e consolidação de sequências. Após este procedimento, nós observamos uma grande quantidade de *contigs* apresentando similaridade com vírus das famílias

Bunyaviridae e *Luteoviridae*. Observamos *contigs* de uma única biblioteca, RKPM 44, que apresentaram similaridade com vírus da família *Virgaviridae*.

Tabela 9. Resumos dos vírus encontrados nas bibliotecas de pequenos RNAs de mosquitos individuais de campos.

Família viral	Vírus	Maior *contig* (nt)	Melhor hit segundo BlastP
Bunyaviridae	PCLV	3.927	glycoprotein [Phasi-Charoen like vírus]
	PCLV	6.782	RdRP [Phasi-Charoen like vírus]
	PCLV	1.212	nucleocapsid [Phasi-Charoen like vírus]
Não classificado	HTV	1.597	structural protein precursor [Dosophila A vírus]
	HTV	2.734	putative RdRP [Laehm Sing vírus]
Virgaviridae	MCV	580	polyprotein [Citrus leprosis vírus C]

Através de análises de similaridade de sequências em nucleotídeo, nós observamos que os *contigs* com similaridade com os Bunyavírus e Luteovírus são derivados dos mesmos vírus que identificados previamente nos mosquitos do Rio de Janeiro, PCLV e HTV, apresentando similaridade entre 98-99.3%. Os *contigs* com similaridade com Virgavírus apresentaram baixa similaridade com vírus conhecidos e as análises filogenéticas sugeriram que estas sequências provavelmente representam um novo vírus (**Figura 24**). Esse novo vírus foi nomeado *Mosquito caratinga virgavírus* (MCV), para refletir a localização geográfica e hospedeiro onde ele foi identificado.

Figura 24. Filogenia do novo vírus identificado nas amostras de mosquitos da natureza. Árvore filogenética construída utilizando os melhores hits segundo BlastP sugere que o MCV é um vírus da família *Virgaviridae*.

É importante ressaltar que todos os vírus encontrados nos mosquitos individuais de Caratinga apresentaram perfil de pequenos RNAs compatível com a ativação da via de siRNA (enriquecimento de pequenos RNAs com tamanho de 21 nt e cobertura uniforme do genoma viral em ambas as fitas), o que é indicativo de infecção produtiva (**Figura 25**). Adicionalmente, consistente com o que foi visto em populações do Rio de Janeiro do mosquitos *Aedes aegypti*, o PCLV em mosquitos de Caratinga também apresenta forte ativação da via de piRNAs (**Figura 20 e Figura 25**). O MCV também apresentou algum sinal de ativação da via de piRNAs embora não tão claro como o PCLV (**Figura 25**).

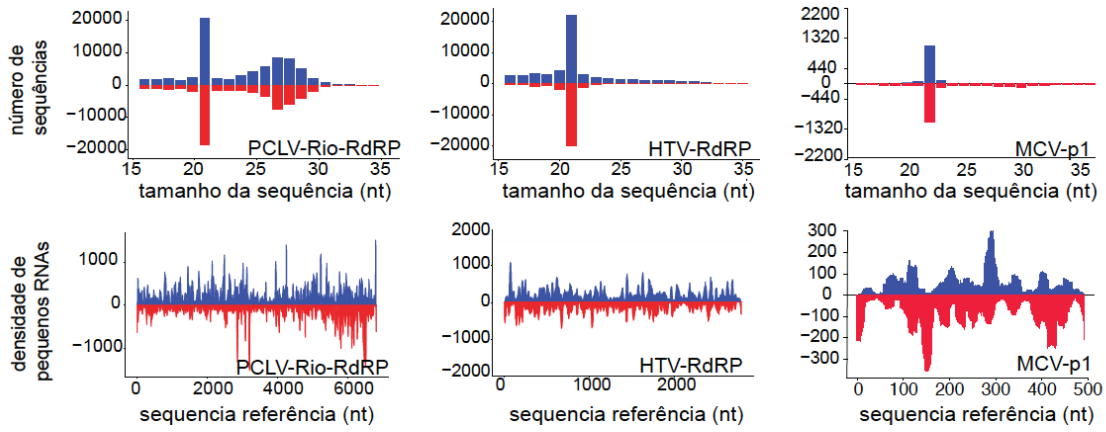

Figura 25. Perfil dos pequenos RNAs derivados dos vírus identificados em mosquitos de campo. Azul e vermelho representam os pequenos RNAs na fita positiva e negativa, respectivamente.

Em seguida, nós aplicamos a nossa estratégia baseada no perfil de tamanho dos pequenos RNAs para tentar identificar *contigs* adicionais que apresentavam perfil similar ao visto nos *contigs* do MCV, mas que não foram caracterizados através de similaridade sequência. Nós utilizamos todos os *contigs* não caracterizados montados na biblioteca onde MCV foi identificado o qual o tamanho do *contig* era maior que 245 nt, N50 dos *contigs* virais montados na biblioteca. No total foram avaliados 28 *contigs* em conjunto

com os *contigs* dos vírus previamente identificados em mosquitos, PCLV, HTV e MCV, além de *contigs* representantes de Animal, Bactéria e Fungo (**Figura 26)**.

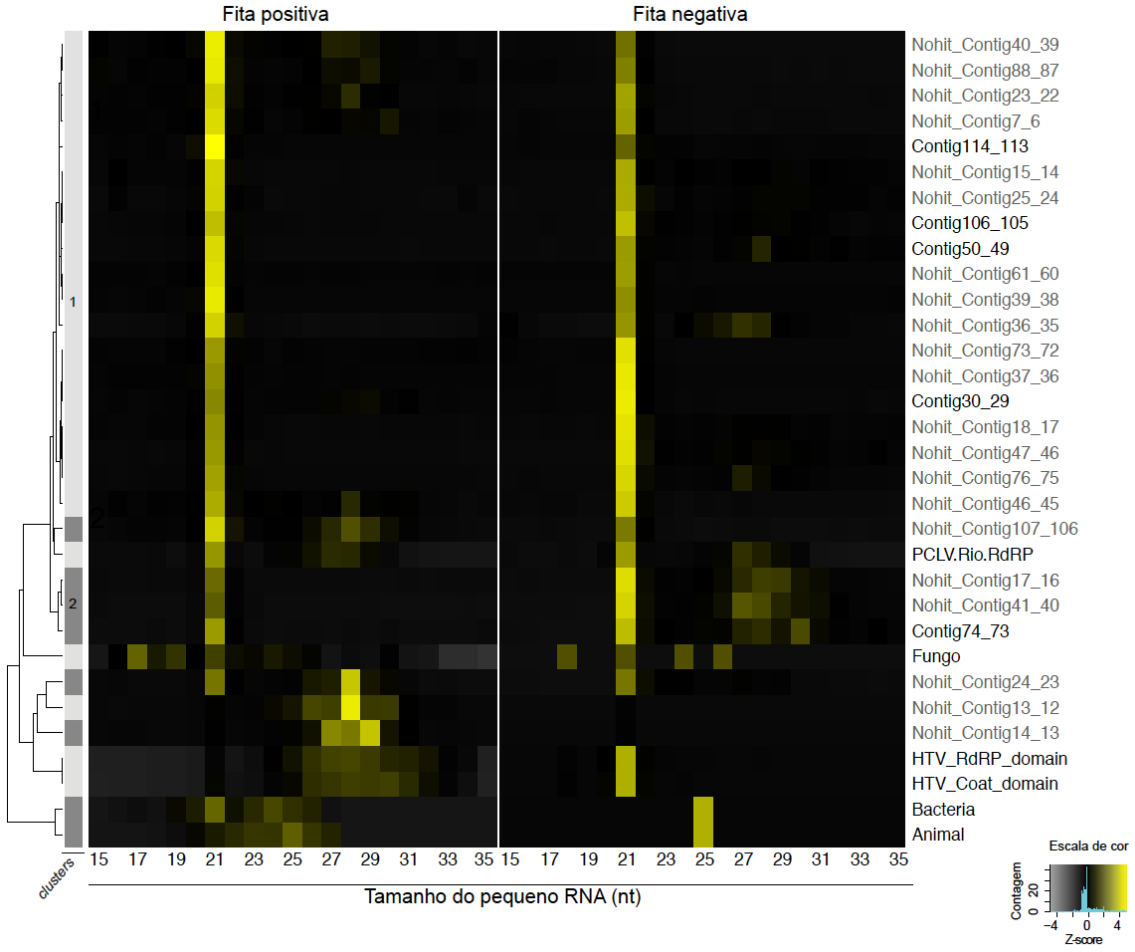

Figura 26. O perfil de tamanho dos pequenos RNAs é capaz de identificar novos *contigs* virais em amostra extraída de mosquito individual de campo. A Clusterização hierárquica foi baseada na correlação de *Pearson* dos perfis de tamanho dos pequenos RNAs dos *contigs* previamente caracterizados (preto) e *contigs* não caracterizados (cinza) oriundos da biblioteca onde o MCV foi identificado, mostrado em forma de mapa de calor "*heatmap*". Os clusters com mais de um *contig* são indicados na barra vertical na esquerda e numerados de acordo com a ordem em que eles aparecem de cima para baixo no *heatmap*.

Através da análise do perfil de pequenos RNAs dos *contigs* não caracterizados nós identificamos 21 novos *contigs* distribuídos em dois clusters (clusters 1 e 2) que poderiam representar partes adicionais do genoma do MCV (**Figura 26**). Estes *contigs* apresentaram correlação de Pearson maior que 0,85 com o perfil de sequências caracterizadas do MCV. Os outros *contigs* não caracterizados não agruparam em nenhum cluster de similaridade e apresentaram perfil de tamanho dos pequenos RNAs

consistentes com piRNAs de mosquitos, o que indica que essas sequências possivelmente representam regiões repetitivas do genoma do mosquito (**Figura 26**). A identificação de *contigs* adicionais derivados do vírus é de extrema importância, pois facilita o desenho de primers e extensão do genoma viral, além de tornar possível outras análises como preferência de base e padrão de uso de códons do vírus.

Para avaliar a prevalência dos vírus encontrados nos mosquitos de Caratinga, cada biblioteca de pequenos RNAs foi mapeada contra as sequências identificadas do PCLV, HTV e MCV. O número de mapeamentos foi então normalizado de acordo com o tamanho da biblioteca, assim tornando possível a comparação entre elas (**Figura 27**).

Com relação a quantidade de sequências identificadas de cada vírus, observou-se um enriquecimento das sequências para sequências de origem do PCLV (**Figura 27**). Isso pode acontecer devido ao fato de o PCLV ser o único que além de ativar a via de siRNA, ativa também a via de piRNAs, responsável por boa parte das sequências com similaridade com o vírus (**Figura 25**).

De forma interessante, mais de 90% das bibliotecas sequenciadas apresentaram pequenos RNAs derivados de mais de um vírus, sugerindo uma alta taxa de co-infecção nos mosquitos individuais de campo. Os vírus previamente identificados, PCLV e HTV apresentaram a maior taxa de prevalência, 92% e 86% respectivamente. Adicionalmente, o PCLV foi encontrado em infecção única em mosquitos de campo, o que não foi visto para o HTV e MCV, encontrados apenas em mosquitos juntamente com o PCLV (**Figura 20**). Assim, a alta taxa de co-infeção de mosquitos com o PCLV e HTV, pode sugerir que há alguma vantagem para ambos os vírus indicando uma possível associação ecológica positiva. Adicionalmente, foi possível encontrar a co-infecção dos três vírus em um mesmo mosquito sugerindo que eles não afetam negativamente uns aos outros (**Figura 27**).

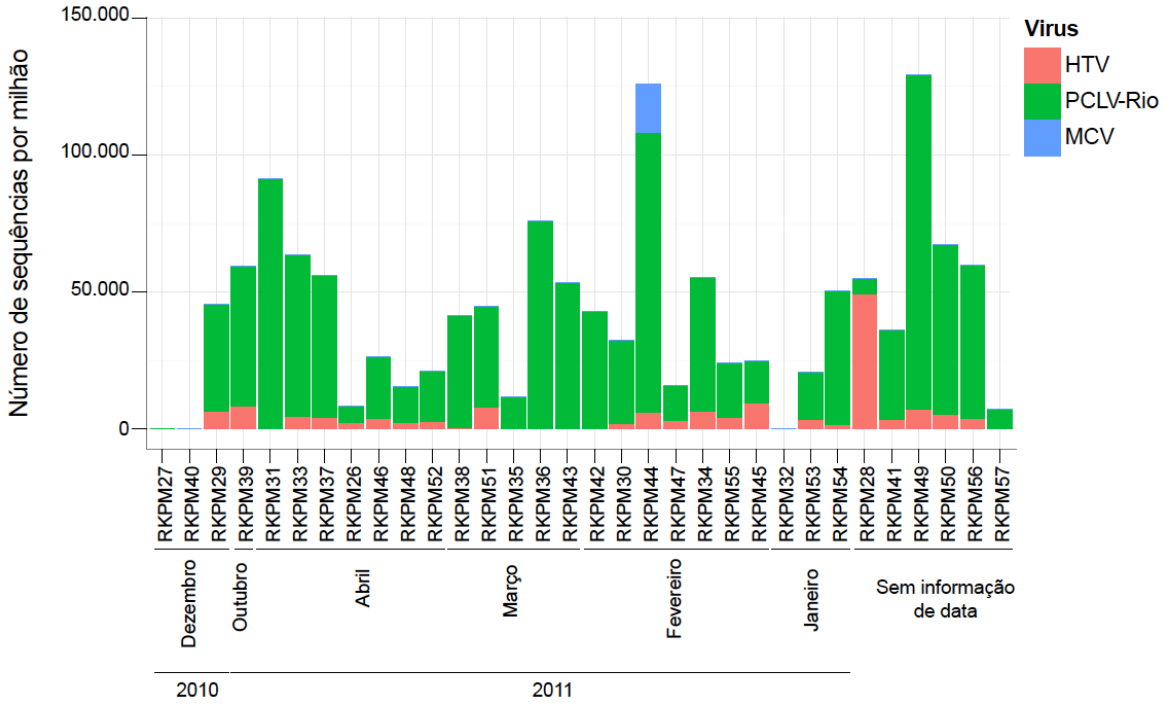

Figura 27. Variação na detecção de RNAs virais nas bibliotecas derivadas de mosquitos individuais de campo. Resumo do número de sequências virais encontradas nas bibliotecas de pequenos RNAs de mosquitos por período de coleta normalizado pelo número de sequências derivadas do hospedeiro.

Nós observamos que o PCLV e HTV foram encontrados em mosquitos coletados em todas as estações do ano e em diferentes regiões da cidade. O fato de os vírus serem encontrados com alta prevalências, nas diversas estações e em diferentes regiões da cidade reforça a ideia de que ambos façam parte do viroma residente dos mosquitos de Caratinga (**Figura 22**). Em contraste, o MCV somente foi encontrado em um único mosquito indicando que ele provavelmente é um componente transiente ou oportunista do viroma dos mosquitos de Caratinga (**Figura 27**). Vale ressaltar que os vírus da família *Virgaviridae* geralmente tem como hospedeiro principal plantas e serem transmitidos através de contato físico ou utilizando fungos como vetores (Adams, Antoniw et al. 2009). Contudo, nós não observamos *contigs* com similaridade significante com sequências derivadas de plantas ou fungos na amostra onde o MCV foi encontrado (ver **Figura 23**, biblioteca RKPM44). Adicionalmente, nós não encontramos na literatura nenhuma evidência de mosquitos infectados por vírus da família *Virgaviridae*, apesar de mosquitos

de ambos os sexos se alimentarem de seiva em algum momento do seu desenvolvimento (Inouye 2010).

Todos os vírus encontrados foram identificados em bibliotecas preparadas a partir de mosquitos *Aedes aegypti* sugerindo que os vírus poderiam ser específicos desta espécie (**Figura 23 e Figura 27**). Nenhum *contig* ou sequência derivada de vírus foram encontradas na biblioteca derivada de *Aedes albopictus*, mas destacamos que realizamos o sequenciamento de uma única amostra derivada desta espécie. Assim, outros experimentos com número maior de indivíduos são necessários para avaliar a capacidade dos vírus identificados de infectar mosquitos do gênero *Aedes albopictus*.

Por fim, é importante observar que o MCV, apesar da baixa prevalência e menor quantidade de sequências geradas, foi passível de identificação e caracterização através da nossa estratégia (**Figura 27**). Dessa forma, nossos resultados sugerem que o pipeline desenvolvido de identificação e caracterização de novos vírus utilizando o sequenciamento de pequenos RNAs pode ser aplicada em amostras complexas como mosquitos individuais capturados no campo. Nesse contexto, nossa estratégia se apresenta como forte ferramenta de vigilância para se caracterizar vírus circulantes tanto em condições de laboratório quanto em condições naturais, como mosquitos individuais derivados diretamente da natureza.

5.4. Detecção de vírus em bibliotecas de pequenos RNAs de insetos, plantas e vertebrados disponíveis em bancos de dados

Nós observamos que o nosso pipeline funcionou eficientemente na identificação de novos vírus em bibliotecas de três insetos de laboratório diferentes construídas por nosso grupo. Adicionalmente, nosso pipeline se mostrou eficaz na caracterização do viroma de mosquitos individuais capturados na natureza. Contudo, através da utilização do nosso pipeline em bibliotecas de insetos nós ainda não havíamos conseguido definir as fronteiras e limitações da nossa estratégia. Assim, nós decidimos avaliar o

comportamento da nossa estratégia em bibliotecas publicadas de insetos de outros laboratórios visando avaliar a eficiência e robustez da estratégia. Adicionalmente, para avaliar a aplicabilidade da nossa estratégia nós ainda aplicamos o nosso pipeline de identificação de vírus em bibliotecas de pequenos RNAs derivadas de plantas e vertebrados. Um resumo das bibliotecas publicadas utilizadas pode ser visto na **Tabela 10**.

Tabela 10. Resumo das bibliotecas de RNA publicas analisadas nesse trabalho.

Biblioteca	ID SRA	Infecção artificial	#total de sequências	# sequências mapeadas hospedeiro	# sequências processadas	# *contigs*	N50 (nt)	Tamanho maior *contig* (nt)	#*contigs* similaridade vírus
Células U4.4	SRR389184	SINV-GFP	27,997,328	20,467,497	6,689,669	1,086	72	1,273	95
Células Aag2	SRR389187	SINV-GFP	35,569,242	21,247,368	9,316,546	763	87	1,874	79
Mosquitos	SRR400496	SINV	4,238,851	2,980,359	1,300,122	226	197	7,361	9
Mosquitoes	SRR400497	SINV-NoVB2	3,522,010	2,874,099	689,541	219	74	201	67
Arabidopsis folhas	SRR1561607	TuMV	15,841,206	6,269,276	6,532,498	75	2,896	7,706	2
Células Garoupa GP	SRR096455	SGIV	7,246,099	4,742,849	936,104	147	64	154	22
Pulmão de camundongo	SRX377856	SARS-CoV	44,436,105	37,200,539	6,835,566	528,894	36	291	0
Pulmão de camundongo	SRR452408	SARS-CoV	22,665,163	16,979,135	2,914,479	924	66	429	140
Pulmão de camundongo	SRR452409	-	24,311,834	17,067,263	3,123,142	757	57	212	0
Células ES de camundongo	SRR640604	EMCV	64,913,697	34,379,621	13,995,769	898	62	255	69
Células ES de camundongo	SRR640602	-	56,784,360	33,088,484	12,742,698	867	61	293	0

5.4.1 Detecção de vírus em amostras de insetos

Para confirmar a eficiência e robustez da nossa estratégia nós decidimos estender a nossa validação analisando 4 bibliotecas públicas de pequenos RNAs derivadas de insetos construídas a partir de mosquitos adultos e linhagens celulares infectadas com SINV (**Tabela 10**) (Myles, Wiley et al. 2008, Vodovar, Bronkhorst et al. 2012). Buscas através de similaridade de sequência mostraram que sequências virais representaram 10.9% dos *contigs* montados nos dados analisados (**Figura 28A**). A diferença entre *contigs* virais e não-virais foi significante na maioria dos dos casos com exceção dos mosquitos infectados com o SINV recombinante (SINV-B2) que codifica a proteína B2 do FHV que quase que completamente bloqueia a via de RNAi (**Figura 28B**) (Adelman, Anderson et al. 2012). Porém, sequência derivadas do SINV foram detectadas em todos conjuntos de *contigs* montados a partir de todas as bibliotecas de inseto, incluindo a biblioteca com SINV-B2 onde a via de RNAi foi inibida. Esse resultado sugeriu que pequenos RNAs derivados de vírus produzidos pela maquinaria de RNAi do hospedeiro são importantes, mas não essenciais para a montagem de *contigs* virais. Nossa estratégia também foi capaz de detectar a presença de diversos *contigs* derivados de vírus que não foram reportadas no momento da primeira publicação. Nós detectamos *contigs* derivados de *Aedes aegypti densovírus 2* (AaDV2), *Mosquito X vírus* (MXV) e *Cell fusion agent vírus* (CFAV) em células Aag2, MXV e *Insect Iridescent vírus- 6* (IIV6) em células U4.4 e *Mosquito nodavírus* (MNV) em mosquitos adultos (**Tabela 11**). Notavelmente, nós identificamos uma sequência de 1.130 nt correspondendo ao MNV que foi originalmente identificado por outro pipeline de identificação de vírus baseado em pequenos RNAs em bibliotecas de mosquito adulto (Wu, Luo et al. 2010). Utilizando o mesmo conjunto de dados, nossa estratégia montou um *contig* de 1.994 nt (AaeS.82) que estendeu em mais de 800 nt a sequência original do MNV previamente publicada (**Figura 28C**). Essa sequência derivada do MNV de 1.994 nt contém a ORF original codificando uma proteína de capsídeo e uma adicional ORF incompleta predita codificar uma proteína

com domínio RdRP_3 (PF00998) (**Figura 28C**). Adicionalmente, nós detectamos um *contig* viral de 1.702 nt (AaeS.81) que mostrou similaridade significante com *Melo necrótica spot vírus*, um membro da família *Tombusviridae* (**Tabela 11**). O *contig* AaeS.81 tem uma ORF complete de 397 aminoacidos e uma segunda ORF incompletas que contém um domínio RdRP_3 (PF00998), o mesmo domínio encontrado no *contig* de 1.994 nt do MNV (**Figura 28C**). O perfil de tamanho dos pequenos RNAs do *contig* AaeS.81 e MNV (AaeS.82) apresentaram alta correlação, > 0.998 (**Figura 28D**). Esses resultados sugeriram que os *contig* do MNV de 1.994 nt e AaeS.81 poderiam representar diferentes fragmentos do mesmo genoma viral (**Figura 28C**). Reforçando a nossa hipótese, o *contig* AaeS.91 e MNV (AaeS.82) foram somente encontrados na mesma biblioteca preparada a partir de mosquitos adultos infectados com SINV.

Tabela 11. Resumo dos vírus identificados em bibliotecas publicas de pequenos RNAs.

ID SRA /Amostra	Família viral	Vírus	Origem da sequência viral	Estratégia	Maior *contig* (nt)	Número de hits	Melhor hit	E-value	Número de acesso
SRR389184/ *Aedes albopictus* células U4.4 + SINV-GFP	*Togaviridae*	*Sindbis vírus*	infecção experimental	blastx	1.128	38	nonstructural polyprotein [Sindbis vírus]	0.0	AAA96975.1
	Iridoviridae	*Insect iridescent vírus-6*	desconhecido	blastx	73	1	IIV6 genome	1,00E-07	AF303741.1
	Birnaviridae	*Mosquitoe X vírus*	desconhecido	blastx	195	54	polyprotein [Mosquitoe x vírus]	4,00E-39	AFU34333.1
	unkonwn	Unknown (*contig* U4.4.84)	desconhecido	padrão	390	1	-	-	-
	unknown	Unknown (*contig* U4.4.85)	desconhecido	padrão	363	1	-	-	-
SRR389187/ *Aedes aegypti* células Aag2 + SINV-GFP	*Flaviviridae*	*Cell fusing agent vírus*	desconhecido	blastx	203	47	polyprotein [Cell fusing agent vírus]	3,00E-11	P33515.1
	Birnaviridae	*Mosquitoe X vírus*	desconhecido	blastx	87	10	putative VP1 [Mosquitoe x vírus]	2,00E-10	AFU34334.1
	Togaviridae	*Sindbis vírus*	infecção experimental	blastx	1.874	18	polyprotein [Sindbis vírus]	0.0	ACU25468.1
	Parvoviridae	*Aedes aegypti densovírus 2*	desconhecido	blastx	52	1	Aedes aegypti densovírus 2	5,00E-08	NC_012636.1
	Togaviridae	*Sindbis vírus*	infecção experimental	blastx	7.361	2	nonstructural polyprotein [Sindbis vírus]	0.0	AAA96974.1
SRR400496/ *Aedes aegypti* + SINV	*Nodaviridae*	*Mosquito nodavírus*	desconhecido	blastx	1,943	2	coat protein [Mosquito nodavírus MNV-1]	1,00E-161	ACY74430.1
		Mosquito nodavírus (*contig* AaeS.82)	desconhecido	blastx	1.729*	2	p89 [Melo necrótica spot vírus]	8,00E-09	AGO36278.1
		Mosquito nodavírus (*contig* AaeS.83)	desconhecido	padrão	709	1	-	-	-
SRR400497/ *Aedes aegypti* + SINV-B2	*Togaviridae*	*Sindbis vírus*	infecção experimental	blastx	169	65	polyprotein [Sindbis vírus]	3,00E-32	BAH70330.1
SRR1561607 / *Arabidopsis* + TuMV	*Potyviridae*	*Turnip mosaic vírus*	infecção experimental	blastn	7.771	3	polyprotein [Reporter vector pCBTuMV-GFP]	0	ABK27329.1
SRR096455 / Garoupa GP +SGIV	*Iridoviridae*	*Singapore grouper iridovírus*	infecção experimental	blastn	112	22	Singapore grouper iridovírus, complete genome	1,00E-17	AY521625.1
SRR452408 / Pulmão de camundongo + SARS-CoV	*Coronaviridae*	*Severe acute respiratory syndrome coronavírus*	infecção experimental	blastn	335	137	SARS coronavírus, complete genome	8,00E-90	JN854286.1
SRR640604 / Células ES camundongo + EMCV	*Picornaviridae*	*Encephalomyocarditis vírus*	infecção experimental	blastn	243	56	Polyprotein [Encephalomyocarditis vírus]	2E-49	ACI47517.1

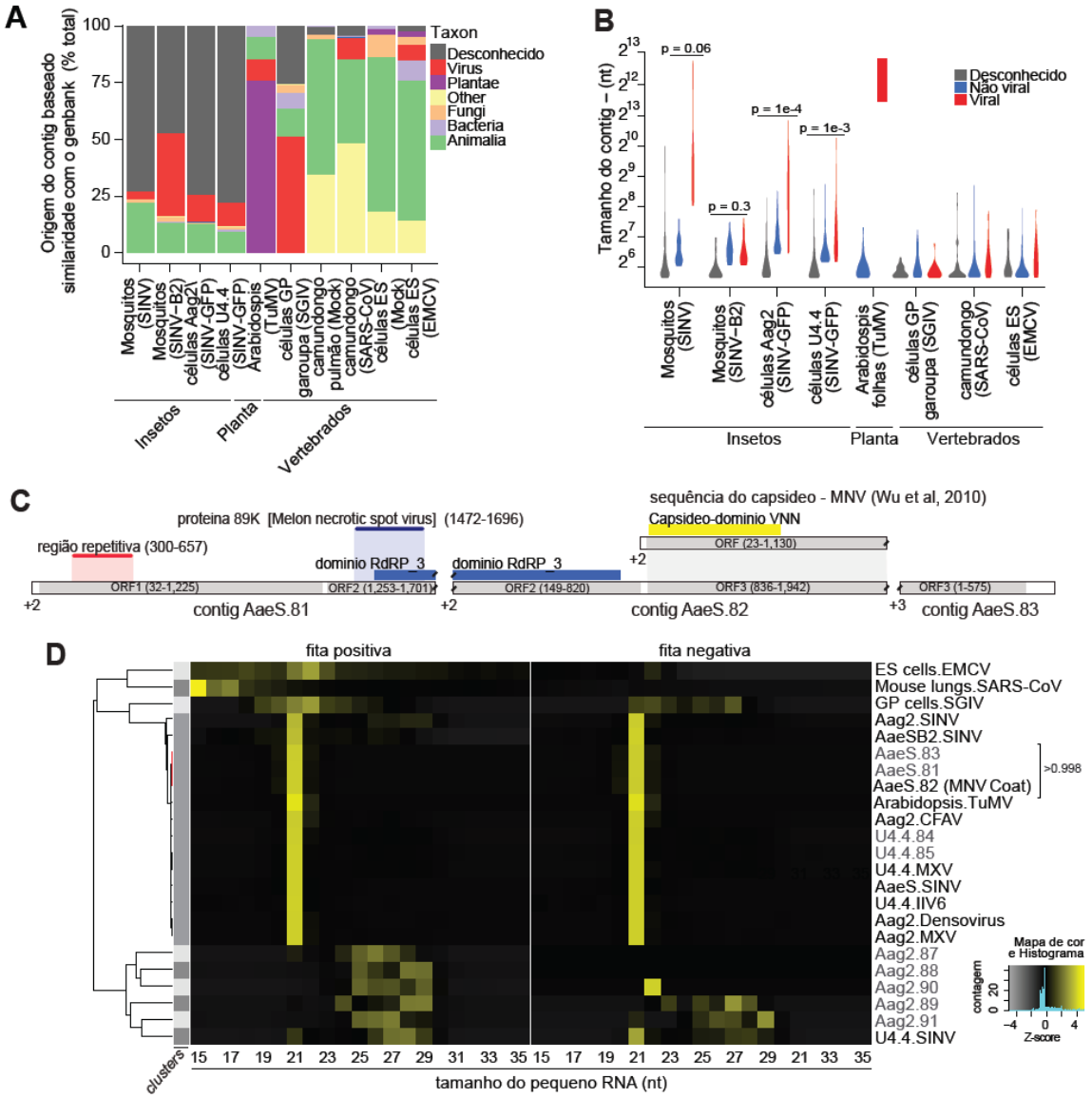

Figura 28. Detecção de vírus baseada no sequenciamento em larga escala de pequenos RNAs é aplicável a animais e plantas. (A) Porcentagem dos *contigs* montados a partir de bibliotecas publicas de pequenos RNAs de insetos, vertebrados e plantas com similaridade significante contra sequências em bancos de dados de referência. A origem dos *contigs* foi classificada por Táxon incluído sequencias "Desconhecidas". (B) Distribuição de tamanho dos *contigs* correspondentes a viral (vermelho), não viral (azul) ou desconhecidas (cinza) em cada biblioteca. P-values das diferenças entre os tamanhos dos *contigs* foram calculados utilizando *test t* de *Student* e indicados na figura. (C) Organização hipotética do genoma do MNV baseada nas análises de ORF e pequenos RNAs dos *contigs* AaeS.81, AaeS.82 d AaeS.83 identificados nesse trabalho. (D) Clusterização hierárquica dos *contigs* virais e não virais montados nas bibliotecas publicas. A clusterização foi baseada em correlação de Pearson do perfil de tamanho dos pequenos RNAs mostrado como forma de mapa de calor '*heatmap*'. Clusters com mais de um elemento foram indicados na barra vertical da esquerda definidos através de correlação de Pearson acima de 0.8. O subcluster destacado em vermelho contém o perfil de pequenos RNAs de *contigs* que mostraram correlação de Pearson acima de 0,998.

Nas bibliotecas publicas de pequenos RNAs derivados de insetos nossa estratégia também montou um total de 1.673 *contigs* não-caracterizados. Os perfis de pequenos RNAs foram analisados para 8 *contigs* com tamanho maior que o N50 observado para *contigs* virais (208 nt). Analisando o perfil de tamanho dos pequenos RNAs, a maioria dos *contigs* virais identificados nos conjuntos de dados publicados foram agrupados em um único cluster apresentando pico de tamanho de 21 nt, consistente com siRNAs canônicos (**Figura 28D**). Diferente do que foi visto para as bibliotecas sequenciadas de *pools Aedes*, *Drosophila* e *Lutzomyia* nesse trabalho, houve uma ausência de diversidade no perfil de pequenos RNAs. Esse resultado pode ser explicado pela alta homogeneidade dos conjuntos de dados públicos de insetos analisados que são principalmente derivados de mosquitos.

Avaliando os 8 *contigs* não caracterizados, um *contigs* de 709 nt (AaeS.83) apresentou perfil de tamanho dos pequenos RNAs similar ao MNV (AaeS.82) e AaeS.81 e foram agrupados no mesmo cluster com alta correlação, > 0.998. Levando em conta todos esses resultados, nós especulamos que o *contig* AaeS.83 pode representar um outro fragmento do genoma do MNV que estenderia o *contig* AaeS.81 (como sugerido na **Figura 28C**). Dois dos *contigs* não caracterizados de 390 e 363 nt derivados de células U4.4 apresentaram perfil de tamanho dos pequenos RNAs similar com diversos vírus os quais foram agrupados juntos (**Figura 28D**).

O *contig* U4.4.84 foi predito codificar duas ORFs incompletas, onde uma delas apresentou similaridade limitada com o *Megavírus terra 1* da superfamília *Megavirales* (**Figura 29A**). Avaliando o perfil de tamanho dos pequenos RNAs, eles apresentam alta correlação, o que sugeri que o U4.4.84 e U4.4.83 podem ter a mesma origem. Nós ainda encontramos pequenos RNAs derivados dos *contigs* U4.4.84 e U4.4.85 na biblioteca de pequenos RNAs preparado a partir de células Aag2 do mesmo laboratório onde foi preparado a biblioteca de células U4.4 (**Figura 29B**). Essas observações

sugeriram esses *contigs* poderiam ser derivados de um vírus que contaminou ambas as culturas células, visto que os pequenos RNAs derivados dos *contigs* U4.4.84 e U4.4.85 não foram observados em células Aag2 do nosso laboratório (**dados não mostrados**). Os outros 5 *contigs* não caracterizados foram montados a partir da biblioteca de células Aag2 mas apresentaram perfil de pequenos RNAs consistente com piRNAs, sugerindo que esses *contigs* provavelmente representam regiões repetitivas não presentes no genoma do mosquito.

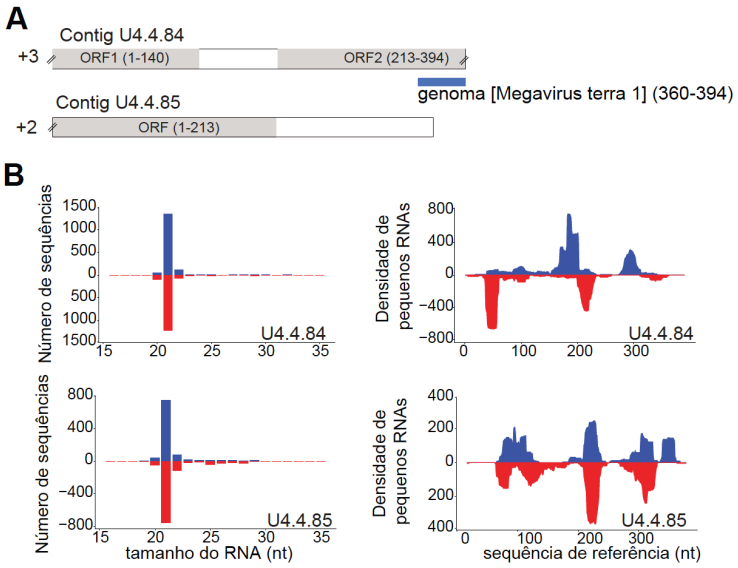

Figura 29. Estrutura de ORFs e perfil de pequenos RNAs das sequências virais não caracterizadas identificadas em bibliotecas de pequenos RNAs de células U4.4. (A) *Contigs* U4.4.84 e U4.4.85 apresentam ORFs incompletas e nenhum domínio conservado. A segunda ORF incompleta do *contig* U4.4.84 apresenta similaridade com *Megavírus terra 1*. (B) Pequenos RNAs com perfil canônico de siRNAs mapeados nos *contigs* U4.4.84 e U4.4.85 são também encontrados na biblioteca derivada de células Aag2 oriundas do mesmo laboratório.

5.4.2. Detecção de vírus em amostras de plantas e animais vertebrados

Com o objetivo de aprofundar os testes da nossa estratégia nós analisamos bibliotecas de pequenos RNAs preparadas a partir de folhas de *Arabidopsis thaliana* infectada com *Turnip mosaic vírus* (TuMV), células GP derivadas do peixe garoupa infectadas com *Singapore grouper iridovírus* (SGIV), pulmão de camundongos infectado com *Severe acute respiratory syndrome coronavírus* (SARS-CoV) e células tronco embrionárias de camundongo infectadas com *Encephalomyocarditis vírus* (EMCV) (Peng, Gralinski et al. 2011, Yan, Cui et al. 2011, Maillard, Ciaudo et al. 2013, Cao, Du et al. 2014) (**Tabela 10)**. As amostras infectadas com vírus conhecidos foram selecionadas para prover prova de conceito de detecção destes vírus em amostras de vertebrados e plantas. Apesar do perfil de pequenos RNAs ser diverso, *contigs* correspondendo a cada vírus foram montados e eficientemente detectados a partir de estratégias baseadas em similaridade de sequência nas respectivas amostras infectadas (**Figura 28D e Figura** 30**A**). Notavelmente, sequências virais montadas a partir da amostra de *Arabidopsis* apresentaram tamanho de *contigs* entre os maiores dentre todos os *contigs* montados nas bibliotecas analisadas neste trabalho (**Figura 28B**). Este fenômeno provavelmente é resultado da eficiente produção de pequenos RNAs derivados de vírus pela maquinaria de RNAi de plantas que favorece a montagem de *contigs* virais longos (**Figura 30A**) (Kreuze, Perez et al. 2009, Pumplin and Voinnet 2013).

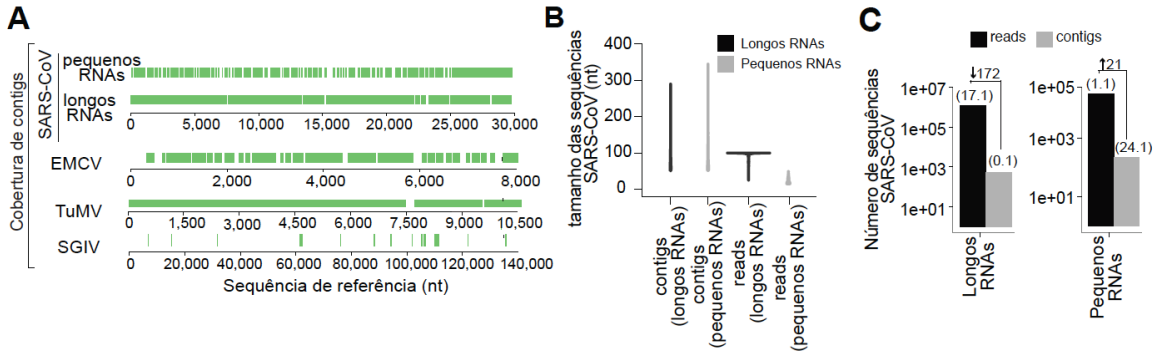

Figura 30. A fração de pequenos RNAs favorece a montagem de *contigs* virais em comparação aos longos RNAs. (A) Cobertura do genoma de SARS-CoV, EMCV, TuMV e SGIV a partir dos *contigs* montados nas bibliotecas de pulmão de camundongo, células ES, folhas de Arabidopsis e células GP de garoupa, respectivamente. (B) Distribuição de tamanho dos *contigs* e sequências brutas derivadas do SARS-CoV nas bibliotecas de RNAs longos (preto) ou pequenos (cinza) RNAs de pulmão de camundongos infectados. (C) Número de sequências brutas e *contigs* derivados do vírus nas bibliotecas de longos e pequenos RNAs preparadas de pulmão de camundongos infectados com SARS-CoV. O número acima das barras indica a porcentagem de sequências virais e *contigs* relativos ao total. Enriquecimento ou depleção de *contigs* virais em comparação a sequências brutas é também indicado.

Em contraste com o que é visto em *Arabidopsis*, *contigs* virais derivados de bibliotecas de camundongo e peixe estão entre os menores *contigs* virais montados (**Figura 28B**). Esse resultado sugeri que *contigs* virais montados a partir de pequenos RNAs em animais vertebrados não é tão eficiente quanto em insetos e plantas. Contudo, *contigs* foram montados e permitiram a identificação de vírus em todas as bibliotecas de pequenos RNAs de peixe e camundongo. Na amostra derivada de células GP, os menores tamanhos vistos pra os *contigs* derivados do SGIV poderiam ser parcialmente explicados pela restrita geração de dsRNA durante o ciclo de replicação de vírus dsDNA (Kemp, Mueller et al. 2013). Adicionalmente, os resultados obtidos com as bibliotecas de pequenos RNAs de camundongo sugerem que a ativação da via de RNAi não é essencial para permitir a montagem de *contigs* virais.

Na biblioteca de células ES, onde a via de RNAi é ativada, foi possível montar e identificar *contigs* derivados do EMCV (Maillard, Ciaudo et al. 2013). Em contraste, *contigs* do SARS-CoV foram montados a partir dos pequenos RNAs derivados das

amostras de pulmão de camundongo, pequenos RNAs estes que provavelmente são gerados a partir da ação de RNAses (**Figura 30A**) (Peng, Gralinski et al. 2011). É importante ressaltar que a RNAse L é um importante fator antiviral que pode degradar RNAs virais em células de mamífero independentemente de RNAi (Girardi, Chane-Woon-Ming et al. 2013). Com relação a distribuição de tamanho dos *contigs* virais e número de pequenos RNAs, o EMCV e SARS-CoV apresentaram métricas similares (**Figura 28B e Figura 30A).** Este resultado sugere que os pequenos RNAs gerados pela via de RNAi ou resultantes de degradação por outras RNAses permitem eficiência similar na montagem de *contigs* virais nas amostras de camundongo**.**

Aproveitando a disponibilidade de dados, nós também decidimos utilizar os dados públicos para tentar reforçar a nossa hipótese que a fração de pequenos RNAs enriquece a montagem de *contigs* para sequências virais em comparação a fração de longos RNAs. Para isso nós avaliamos a identificação de *contigs* virais a partir do sequenciamento de frações diferente de RNAs preparados a partir de amostras de pulmão de camundongos infectados com SARS-CoV (Peng, Gralinski et al. 2011, Josset, Tchitchek et al. 2014). *Contigs* do SARS-CoV montados a partir dos pequenos e longos RNAs apresentam distribuição de tamanho similar, apesar do tamanho e número maior de *sequências* observados na biblioteca preparada a partir de longos RNAs (**Figura 30A e B**). As bibliotecas de pequenos RNAs apresentaram mais de 20 vezes de enriquecimento para sequências virais quando comparado a geração de *contigs* e *sequências* derivadas do vírus (**Figura 30C**). Em contraste, nós observamos uma diminuição de 172 vezes na porcentagem de sequências virais detectadas nos *contigs* montados quando comparado a *sequências* derivadas do vírus na biblioteca de longos RNAS (**Figura 30C**). Assim, a montagem de *contigs* a partir dos pequenos RNAs favorece a montagem de *contigs* do SARS-CoV quando em comparação aos longos RNAs, mesmo quando não é observada uma ativação clara da via de RNAi em

amostras de pulmão de camundongo. Esses resultados preliminares indicam que os pequenos RNAs apresentam enriquecimento para sequências virais e podem ser utilizados para a montagem de *contigs* não só em insetos, mas também em plantas e mamíferos.

Contudo, apesar de conseguir montar *contigs* derivados de vírus, nós montamos poucos *contigs* que não foram caracterizados nas bibliotecas de camundongo, peixe e plantas (**Figura 28 A e B**).

Assim, nossa estratégia de utilização do perfil de tamanho dos pequenos RNAs para caracterização de *contigs* não caracterizados não pode ser propriamente testada nessas amostras. É importante ressaltar que mais testes são necessários para avaliar a aplicabilidade da nossa estratégia baseada em padrão em amostras derivadas de plantas e animais vertebrados. Contudo, os nossos resultados sugerem que o sequenciamento de bibliotecas de pequenos RNAs também pode ser aplicado para o monitoramento de vírus circulantes não só em insetos, mas também em plantas e vertebrados.

6. Conclusões

1 - O sequenciamento em larga escala de pequenos RNAs otimiza a detecção de *contigs* derivados de vírus quando comparado com o sequenciamento de longos RNAs;

2 - O padrão de pequenos RNAs permite a identificação de sequências virais não caracterizadas através de similaridade de sequências;

3 - O perfil e padrão molecular dos pequenos RNAs permite inferir aspectos da biologia do vírus, tais como estrutura do genoma, tropismo tecidual e estratégia de replicação;

4 - O pipeline desenvolvido de caracterização de sequências virais identificou 6 novos vírus em populações de laboratório de *Drosophila*, *Aedes* e *Lutzomyia*;

5 - Análise temporal/espacial do viroma de mosquitos Aedes na cidade de Caratinga identificou 3 vírus como componentes do viroma;

6 - O PCLV e HTV são vírus com alta prevalência em populações de mosquitos no Brasil;

7 - A estratégia desenvolvida de identificação e caracterização de novos vírus a partir do sequenciamento de pequenos RNAs pode ser aplicado para plantas e animais vertebrados;

7. Perspectivas

1 - Avaliar a aplicação do nosso pipeline de identificação e caracterização de sequências virais em outros organismos vertebrados e invertebrados;

2 - Isolar e caracterizar biologicamente os novos vírus identificados e verificar uma possível interferência na transmissão de outros patógenos pelos insetos vetores;

3 - Estudar a importância da via de piRNAs em mosquitos para o combate a infecção pelo PCLV;

4 - Avaliar a prevalência dos vírus identificados nas bibliotecas de pequenos RNAs de mosquitos individuais de campo em todos os mosquitos individuais coletados na cidade de Caratinga;

5 - Correlacionar a presença dos vírus encontrados em mosquitos individuais de campo com dados de sazonalidade e presença de Dengue;

8. Referências

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Siden-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin and J. C. Venter (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-2195.

Adams, M. J., J. F. Antoniw and J. Kreuze (2009). "Virgaviridae: a new family of rod-shaped plant viruses." Arch Virol **154**(12): 1967-1972.

Adelman, Z. N., M. A. Anderson, M. Liu, L. Zhang and K. M. Myles (2012). "Sindbis virus induces the production of a novel class of endogenous siRNAs in Aedes aegypti mosquitoes." Insect Mol Biol **21**(3): 357-368.

Aguiar, E. R., R. P. Olmo, S. Paro, F. V. Ferreira, I. J. de Faria, Y. M. Todjro, F. P. Lobo, E. G. Kroon, C. Meignin, D. Gatherer, J. L. Imler and J. T. Marques (2015). "Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host." Nucleic Acids Res.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Álvaro Eduardo Eiras, M. C. R. (2009). "Preliminary evaluation of the “Dengue-MI” technology for Aedes aegypti monitoring and control." Caderno de Saúde Pública Rio de Janeiro **25**(1): 14.

Ambrose, R. L., G. C. Lander, W. S. Maaty, B. Bothner, J. E. Johnson and K. N. Johnson (2009). "Drosophila A virus is an unusual RNA virus with a T=3 icosahedral core and permuted RNA-dependent RNA polymerase." J Gen Virol **90**(Pt 9): 2191-2200.

Arensburger, P., R. H. Hice, J. A. Wright, N. L. Craig and P. W. Atkinson (2011). "The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs." BMC Genomics **12**: 606.

Arensburger, P., K. Megy, R. M. Waterhouse, J. Abrudan, P. Amedeo, B. Antelo, L. Bartholomay, S. Bidwell, E. Caler, F. Camara, C. L. Campbell, K. S. Campbell, C. Casola, M. T. Castro, I. Chandramouliswaran, S. B. Chapman, S. Christley, J. Costas, E. Eisenstadt, C. Feschotte, C. Fraser-Liggett, R. Guigo, B. Haas, M. Hammond, B. S. Hansson, J. Hemingway, S. R. Hill, C. Howarth, R. Ignell, R. C. Kennedy, C. D. Kodira, N. F. Lobo, C. Mao, G. Mayhew, K. Michel, A. Mori, N. Liu, H. Naveira, V. Nene, N. Nguyen, M. D. Pearson, E. J. Pritham, D. Puiu, Y. Qi, H. Ranson, J. M. Ribeiro, H. M. Roberston, D. W. Severson, M. Shumway, M. Stanke, R. L. Strausberg, C. Sun, G. Sutton, Z. J. Tu, J. M. Tubio, M. F. Unger, D. L. Vanlandingham, A. J. Vilella, O. White, J. R. White, C. S. Wondji, J. Wortman, E. M. Zdobnov, B. Birren, B. M. Christensen, F. H. Collins, A. Cornel, G. Dimopoulos, L. I. Hannick, S. Higgs, G. C. Lanzaro, D. Lawson, N. H. Lee, M. A. Muskavitch, A. S. Raikhel and P. W. Atkinson (2010). "Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics." Science **330**(6000): 86-88.

Arolas, J. L., J. Vendrell, F. X. Aviles and L. D. Fricker (2007). "Metallocarboxypeptidases: emerging drug targets in biomedicine." Curr Pharm Des **13**(4): 349-366.

Bahder, B. W., S. Poojari, O. J. Alabi, R. A. Naidu and D. B. Walsh (2013). "Pseudococcus maritimus (Hemiptera: Pseudococcidae) and Parthenolecanium corni (Hemiptera: Coccidae) are capable of transmitting grapevine leafroll-associated virus 3 between Vitis x labruscana and Vitis vinifera." Environ Entomol **42**(6): 1292-1298.

Baltimore, D. (1971). "Expression of animal virus genomes." Bacteriol Rev **35**(3): 235-241.

Beaty, B. J., D. J. Prager, A. A. James, M. Jacobs-Lorena, L. H. Miller, J. H. Law, F. H. Collins and F. C. Kafatos (2009). "From Tucson to genomics and transgenics: the vector biology network and the emergence of modern vector biology." PLoS Negl Trop Dis **3**(3): e343.

Bishop-Lilly, K. A., M. J. Turell, K. M. Willner, A. Butani, N. M. Nolan, S. M. Lentz, A. Akmal, A. Mateczun, T. N. Brahmbhatt, S. Sozhamannan, C. A. Whitehouse and T. D. Read (2010). "Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing." PLoS Negl Trop Dis **4**(11): e878.

Bonami, J. R. and J. Sri Widada (2011). "Viral diseases of the giant fresh water prawn Macrobrachium rosenbergii: a review." J Invertebr Pathol **106**(1): 131-142.

Bonami, J. R. and S. Zhang (2011). "Viral diseases in commercially exploited crabs: a review." J Invertebr Pathol **106**(1): 6-17.

Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam and F. Rohwer (2002). "Genomic analysis of uncultured marine viral communities." Proc Natl Acad Sci U S A **99**(22): 14250-14255.

Cao, M., P. Du, X. Wang, Y. Q. Yu, Y. H. Qiu, W. Li, A. Gal-On, C. Zhou, Y. Li and S. W. Ding (2014). "Virus infection triggers widespread silencing of host genes by a distinct

class of endogenous siRNAs in Arabidopsis." Proc Natl Acad Sci U S A **111**(40): 14613-14618.

Cargnelutti, J. F., R. G. Olinda, L. A. Maia, G. M. de Aguiar, E. G. Neto, S. V. Simoes, T. G. de Lima, A. F. Dantas, R. Weiblen, E. F. Flores and F. Riet-Correa (2014). "Outbreaks of Vesicular stomatitis Alagoas virus in horses and cattle in northeastern Brazil." J Vet Diagn Invest **26**(6): 788-794.

Carrington, L. B. and C. P. Simmons (2014). "Human to mosquito transmission of dengue viruses." Front Immunol **5**: 290.

Carter, J. B. and V. A. Saunders (2007). Virology : principles and applications. Chichester, England ; Hoboken, NJ, John Wiley & Sons.

Chakrabarti, A., S. Banerjee, L. Franchi, Y. M. Loo, M. Gale, Jr., G. Nunez and R. H. Silverman (2015). "RNase L activates the NLRP3 inflammasome during viral infections." Cell Host Microbe **17**(4): 466-477.

Chakrabarti, A., B. K. Jha and R. H. Silverman (2011). "New insights into the role of RNase L in innate immunity." J Interferon Cytokine Res **31**(1): 49-57.

Chao, J. A., J. H. Lee, B. R. Chapados, E. W. Debler, A. Schneemann and J. R. Williamson (2005). "Dual modes of RNA-silencing suppression by Flock House virus protein B2." Nat Struct Mol Biol **12**(11): 952-957.

Cheadle, C., M. P. Vawter, W. J. Freed and K. G. Becker (2003). "Analysis of microarray data using Z score transformation." J Mol Diagn **5**(2): 73-81.

Coffey, L. L., B. L. Page, A. L. Greninger, B. L. Herring, R. C. Russell, S. L. Doggett, J. Haniotis, C. Wang, X. Deng and E. L. Delwart (2014). "Enhanced arbovirus surveillance with deep sequencing: Identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes." Virology **448**: 146-158.

Diallo, D., A. A. Sall, C. T. Diagne, O. Faye, O. Faye, Y. Ba, K. A. Hanley, M. Buenemann, S. C. Weaver and M. Diallo (2014). "Zika virus emergence in mosquitoes in southeastern Senegal, 2011." PLoS One **9**(10): e109442.

Diez-Domingo, J., E. G. Perez-Yarza, J. A. Melero, M. Sanchez-Luna, M. D. Aguilar, A. J. Blasco, N. Alfaro and P. Lazaro (2014). "Social, economic, and health impact of the respiratory syncytial virus: a systematic search." BMC Infect Dis **14**: 544.

Ding, S.-W. (2010). "RNA-based antiviral immunity." Nature reviews. Immunology **10**: 632-644.

Djikeng, A., R. Kuzmickas, N. G. Anderson and D. J. Spiro (2009). "Metagenomic analysis of RNA viruses in a fresh water lake." PLoS One **4**(9): e7264.

Domingo, E., C. Escarmis, N. Sevilla, A. Moya, S. F. Elena, J. Quer, I. S. Novella and J. J. Holland (1996). "Basic concepts in RNA virus evolution." FASEB J **10**(8): 859-864.

Eddy, S. R. (2009). "A new generation of homology search tools based on probabilistic inference." Genome Inform **23**(1): 205-211.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Edwards, R. A. and F. Rohwer (2005). "Viral metagenomics." Nat Rev Microbiol **3**(6): 504-510.

Fajardo, T. V. M., M. Eiras, P. G. Schenato, O. Nickel and G. B. Kuhn (2005). "Avaliação da variabilidade do Grapevine leafroll-associated virus 1 e 3 por análise de seqüências de nucleotídeos e polimorfismo conformacional de fita simples." Fitopatologia Brasileira **30**: 177-182.

Fajardo, T. V. M., G. B. Kuhn, M. Eiras and O. Nickel (2002). "Detecção de Closterovirus em videira e caracterização parcial de um isolado do Grapevine leafroll-associated virus 3." Fitopatologia Brasileira **27**: 58-64.

Figueiredo, L. T. (2007). "Emergent arboviruses in Brazil." Rev Soc Bras Med Trop **40**(2): 224-229.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate and M. Punta (2014). "Pfam: the protein families database." Nucleic Acids Res **42**(Database issue): D222-230.

Flenniken, M. L. (2014). "Honey bee-infecting plant virus with implications on honey bee colony health." MBio **5**(2): e00877-00814.

Galiana-Arnoux, D., C. Dostert, A. Schneemann, J. A. Hoffmann and J. L. Imler (2006). "Essential function in vivo for Dicer-2 in host defense against RNA viruses in drosophila." Nat Immunol **7**(6): 590-597.

Girardi, E., B. Chane-Woon-Ming, M. Messmer, P. Kaukinen and S. Pfeffer (2013). "Identification of RNase L-dependent, 3'-end-modified, viral small RNAs in Sindbis virus-infected mammalian cells." MBio **4**(6): e00698-00613.

Gonzalez, M. W. and W. R. Pearson (2010). "Homologous over-extension: a challenge for iterative similarity searches." Nucleic Acids Res **38**(7): 2177-2189.

Gubler, D. J. (2001). "Human arbovirus infections worldwide." Ann N Y Acad Sci **951**: 13-24.

Han, B. W., W. Wang, C. Li, Z. Weng and P. D. Zamore (2015). "Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production." Science **348**(6236): 817-821.

Han, Y.-H., Y.-J. Luo, Q. Wu, J. Jovel, X.-H. Wang, R. Aliyari, C. Han, W.-X. Li and S.-W. Ding (2011). "RNA-based immunity terminates viral infection in adult Drosophila in the absence of viral suppression of RNA interference: characterization of viral small interfering RNA populations in wild-type and mutant flies." J Virol **85**: 13153-13163.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy and R. M. Goodman (1998). "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products." Chem Biol **5**(10): R245-249.

Hill, C. A., F. C. Kafatos, S. K. Stansfield and F. H. Collins (2005). "Arthropod-borne diseases: vector control in the genomics era." Nat Rev Microbiol **3**(3): 262-268.

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-877.

Hugenholtz, P., B. M. Goebel and N. R. Pace (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." J Bacteriol **180**(18): 4765-4774.

Inouye, D. W. (2010). "Mosquitoes: more likely nectar thieves than pollinators." Nature **467**(7311): 27.

Josset, L., N. Tchitchek, L. E. Gralinski, M. T. Ferris, A. J. Eisfeld, R. R. Green, M. J. Thomas, J. Tisoncik-Go, G. P. Schroth, Y. Kawaoka, F. P. Manuel de Villena, R. S. Baric, M. T. Heise, X. Peng and M. G. Katze (2014). "Annotation of long non-coding RNAs expressed in collaborative cross founder mice in response to respiratory virus infection reveals a new class of interferon-stimulated transcripts." RNA Biol **11**(7): 875-890.

Kemp, C., S. Mueller, A. Goto, V. Barbier, S. Paro, F. Bonnay, C. Dostert, L. Troxler, C. Hetru, C. Meignin, S. Pfeffer, J. A. Hoffmann and J.-L. Imler (2013). Broad RNA interference-mediated antiviral immunity and virus-specific inducible responses in Drosophila. J. Immunol. **190:** 650-658.

Koenraadt, C. J., T. Balenghien, S. Carpenter, E. Ducheyne, A. R. Elbers, M. Fife, C. Garros, A. Ibanez-Justicia, H. Kampen, R. J. Kormelink, B. Losson, W. H. van der Poel, N. De Regge, P. A. van Rijn, C. Sanders, F. Schaffner, M. M. Sloet van Oldruitenborgh-Oosterbaan, W. Takken, D. Werner and F. Seelig (2014). "Bluetongue, Schmallenberg - what is next? Culicoides-borne viral diseases in the 21st Century." BMC Vet Res **10**: 77.

Kreuze, J. F., A. Perez, M. Untiveros, D. Quispe, S. Fuentes, I. Barker and R. Simon (2009). "Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses." Virology **388**(1): 1-7.

Ladner, J. T., B. Beitzel, P. S. Chain, M. G. Davenport, E. F. Donaldson, M. Frieman, J. R. Kugelman, J. H. Kuhn, J. O'Rear, P. C. Sabeti, D. E. Wentworth, M. R. Wiley, G. Y. Yu, C. Threat Characterization, S. Sozhamannan, C. Bradburne and G. Palacios (2014). "Standards for sequencing viral genomes in the era of high-throughput sequencing." MBio **5**(3): e01360-01314.

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin and N. R. Pace (1985). "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses." Proc Natl Acad Sci U S A **82**(20): 6955-6959.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Lepore, L. S., P. R. Roelvink and R. R. Granados (1996). "Enhancin, the granulosis virus protein that facilitates nucleopolyhedrovirus (NPV) infections, is a metalloprotease." J Invertebr Pathol **68**(2): 131-140.

Li, C. X., M. Shi, J. H. Tian, X. D. Lin, Y. J. Kang, L. J. Chen, X. C. Qin, J. Xu, E. C. Holmes and Y. Z. Zhang (2015). "Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses." Elife **4**(eLife.05979): eLife.05979.

Li, Y., J. Lu, Y. Han, X. Fan and S. W. Ding (2013). "RNA interference functions as an antiviral immunity mechanism in mammals." Science **342**(6155): 231-234.

Lobo, F. P., B. E. Mota, S. D. Pena, V. Azevedo, A. M. Macedo, A. Tauch, C. R. Machado and G. R. Franco (2009). "Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts." PloS one **4**(7): e6282.

Longdon, B., M. A. Brockhurst, C. A. Russell, J. J. Welch and F. M. Jiggins (2014). "The evolution and genetics of virus host shifts." PLoS Pathog **10**(11): e1004395.

Maillard, P. V., C. Ciaudo, A. Marchais, Y. Li, F. Jay, S. W. Ding and O. Voinnet (2013). "Antiviral RNA interference in mammalian cells." Science **342**(6155): 235-238.

Malone, C. D., J. Brennecke, M. Dus, A. Stark, W. R. McCombie, R. Sachidanandam and G. J. Hannon (2009). "Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary." Cell **137**(3): 522-535.

Marques, J. T. and R. W. Carthew (2007). "A call to arms: coevolution of animal viruses and host innate immune responses." Trends Genet **23**(7): 359-364.

Marques, J. T., K. Kim, P. H. Wu, T. M. Alleyne, N. Jafari and R. W. Carthew (2010). "Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila." Nat Struct Mol Biol **17**(1): 24-30.

Marques, J. T., J. P. Wang, X. Wang, K. P. de Oliveira, C. Gao, E. R. Aguiar, N. Jafari and R. W. Carthew (2013). "Functional specialization of the small interfering RNA pathway in response to virus infection." PLoS Pathog **9**(8): e1003579.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011, next generation sequencing; small RNA; microRNA; adapter removal: Cutadapt removes adapter sequences from high-throughput sequencing reads. **17**.

Merkling, S. H. and R. P. van Rij (2013). "Beyond RNAi: antiviral defense strategies in Drosophila and mosquito." J Insect Physiol **59**(2): 159-170.

Miller, J. R., S. Koren and G. Sutton (2010). "Assembly algorithms for next-generation sequencing data." Genomics **95**(6): 315-327.

Morazzani, E. M., M. R. Wiley, M. G. Murreddu, Z. N. Adelman and K. M. Myles (2012). "Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma." PLoS Pathog **8**: e1002470.

Mueller, S., V. Gausson, N. Vodovar, S. Deddouche, L. Troxler, J. Perot, S. Pfeffer, J. A. Hoffmann, M. C. Saleh and J. L. Imler (2010). "RNAi-mediated immunity provides strong protection against the negative-strand RNA vesicular stomatitis virus in Drosophila." Proc Natl Acad Sci U S A **107**(45): 19390-19395.

Mulder, N. and R. Apweiler (2007). "InterPro and InterProScan: tools for protein sequence classification and comparison." Methods Mol Biol **396**: 59-70.

Myles, K. M., M. R. Wiley, E. M. Morazzani and Z. N. Adelman (2008). "Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes." Proc Natl Acad Sci U S A **105**(50): 19938-19943.

Naccache, S. N., A. L. Greninger, D. Lee, L. L. Coffey, T. Phan, A. Rein-Weston, A. Aronsohn, J. Hackett, Jr., E. L. Delwart and C. Y. Chiu (2013). "The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns." J Virol **87**(22): 11966-11977.

Nan, Y., G. Nan and Y. J. Zhang (2014). "Interferon induction by RNA viruses and antagonism by viral pathogens." Viruses **6**(12): 4999-5027.

Nene, V., J. R. Wortman, D. Lawson, B. Haas, C. Kodira, Z. J. Tu, B. Loftus, Z. Xi, K. Megy, M. Grabherr, Q. Ren, E. M. Zdobnov, N. F. Lobo, K. S. Campbell, S. E. Brown, M. F. Bonaldo, J. Zhu, S. P. Sinkins, D. G. Hogenkamp, P. Amedeo, P. Arensburger, P. W.

Atkinson, S. Bidwell, J. Biedler, E. Birney, R. V. Bruggner, J. Costas, M. R. Coy, J. Crabtree, M. Crawford, B. Debruyn, D. Decaprio, K. Eiglmeier, E. Eisenstadt, H. El-Dorry, W. M. Gelbart, S. L. Gomes, M. Hammond, L. I. Hannick, J. R. Hogan, M. H. Holmes, D. Jaffe, J. S. Johnston, R. C. Kennedy, H. Koo, S. Kravitz, E. V. Kriventseva, D. Kulp, K. Labutti, E. Lee, S. Li, D. D. Lovin, C. Mao, E. Mauceli, C. F. Menck, J. R. Miller, P. Montgomery, A. Mori, A. L. Nascimento, H. F. Naveira, C. Nusbaum, S. O'Leary, J. Orvis, M. Pertea, H. Quesneville, K. R. Reidenbach, Y. H. Rogers, C. W. Roth, J. R. Schneider, M. Schatz, M. Shumway, M. Stanke, E. O. Stinson, J. M. Tubio, J. P. Vanzee, S. Verjovski-Almeida, D. Werner, O. White, S. Wyder, Q. Zeng, Q. Zhao, Y. Zhao, C. A. Hill, A. S. Raikhel, M. B. Soares, D. L. Knudson, N. H. Lee, J. Galagan, S. L. Salzberg, I. T. Paulsen, G. Dimopoulos, F. H. Collins, B. Birren, C. M. Fraser-Liggett and D. W. Severson (2007). "Genome sequence of Aedes aegypti, a major arbovirus vector." Science **316**(5832): 1718-1723.

Newman, C. M., F. Cerutti, T. K. Anderson, G. L. Hamer, E. D. Walker, U. D. Kitron, M. O. Ruiz, J. D. Brawn and T. L. Goldberg (2011). "Culex flavivirus and West Nile virus mosquito coinfection and positive ecological association in Chicago, United States." Vector Borne Zoonotic Dis **11**(8): 1099-1105.

Nicaise, V. (2014). "Crop immunity against viruses: outcomes and future challenges." Front Plant Sci **5**: 660.

O'Neil, S. T. and S. J. Emrich (2013). "Assessing De Novo transcriptome assembly metrics for consistency and utility." BMC Genomics **14**: 465.

Oh, J., A. L. Byrd, C. Deming, S. Conlan, N. C. S. Program, H. H. Kong and J. A. Segre (2014). "Biogeography and individuality shape function in the human skin metagenome." Nature **514**(7520): 59-64.

Oude Munnink, B. B., S. M. Jazaeri Farsani, M. Deijs, J. Jonkers, J. T. Verhoeven, M. Ieven, H. Goossens, M. D. de Jong, B. Berkhout, K. Loens, P. Kellam, M. Bakker, M. Canuti, M. Cotten and L. van der Hoek (2013). "Autologous antibody capture to enrich immunogenic viruses for viral discovery." PLoS One **8**(11): e78454.

Parrish, C. R., E. C. Holmes, D. M. Morens, E. C. Park, D. S. Burke, C. H. Calisher, C. A. Laughlin, L. J. Saif and P. Daszak (2008). "Cross-species virus transmission and the emergence of new epidemic diseases." Microbiol Mol Biol Rev **72**(3): 457-470.

Peng, X., L. Gralinski, M. T. Ferris, M. B. Frieman, M. J. Thomas, S. Proll, M. J. Korth, J. R. Tisoncik, M. Heise, S. Luo, G. P. Schroth, T. M. Tumpey, C. Li, Y. Kawaoka, R. S. Baric and M. G. Katze (2011). "Integrative deep sequencing of the mouse lung transcriptome reveals differential expression of diverse classes of small RNAs in response to respiratory virus infection." MBio **2**(6): e00198-00111.

Pfeffer, S., M. Zavolan, F. A. Grasser, M. Chien, J. J. Russo, J. Ju, B. John, A. J. Enright, D. Marks, C. Sander and T. Tuschl (2004). "Identification of virus-encoded microRNAs." Science **304**(5671): 734-736.

Pumplin, N. and O. Voinnet (2013). "RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence." Nat Rev Microbiol **11**(11): 745-760.

Quan, P. L., T. Briese, G. Palacios and W. Ian Lipkin (2008). "Rapid sequence-based diagnosis of viral infection." Antiviral Res **79**(1): 1-5.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Redinbaugh, M. G. and J. L. Zambrano (2014). "Control of virus diseases in maize." Adv Virus Res **90**: 391-429.

Roossinck, M. J., D. P. Martin and P. Roumagnac (2015). "Plant Virus Metagenomics: Advances in Virus Discovery." Phytopathology **105**(6): 716-727.

Sanjuan, R. (2012). "From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses." PLoS Pathog **8**(5): e1002685.

Schmidt, T. M., E. F. DeLong and N. R. Pace (1991). "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing." J Bacteriol **173**(14): 4371-4378.

Shepard, D. S., L. Coudeville, Y. A. Halasa, B. Zambrano and G. H. Dayan (2011). "Economic impact of dengue illness in the Americas." Am J Trop Med Hyg **84**(2): 200-207.

Shepard, D. S., E. A. Undurraga and Y. A. Halasa (2013). "Economic and disease burden of dengue in Southeast Asia." PLoS Negl Trop Dis **7**(2): e2055.

Tamura, K. (1992). "Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases." Mol Biol Evol **9**(4): 678-687.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar (2013). "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0." Mol Biol Evol **30**(12): 2725-2729.

Thompson, M. R., J. J. Kaminski, E. A. Kurt-Jones and K. A. Fitzgerald (2011). "Pattern recognition receptors and the innate immune response to viral infection." Viruses **3**(6): 920-940.

Tokarz, R., S. H. Williams, S. Sameroff, M. Sanchez Leon, K. Jain and W. I. Lipkin (2014). "Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses." J Virol **88**(19): 11480-11492.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz and E. M. Rubin (2005). "Comparative metagenomics of microbial communities." Science **308**(5721): 554-557.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar and J. F. Banfield (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature **428**(6978): 37-43.

van Cleef, K. W., J. T. van Mierlo, P. Miesen, G. J. Overheul, J. J. Fros, S. Schuster, M. Marklewitz, G. P. Pijlman, S. Junglen and R. P. van Rij (2014). "Mosquito and Drosophila entomobirnaviruses suppress dsRNA- and siRNA-induced RNAi." Nucleic Acids Res **42**(13): 8732-8744.

van Rij, R. P., M. C. Saleh, B. Berry, C. Foo, A. Houk, C. Antoniewski and R. Andino (2006). "The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster." Genes Dev **20**(21): 2985-2995.

Vega-Rua, A., R. Lourenco-de-Oliveira, L. Mousson, M. Vazeille, S. Fuchs, A. Yebakima, J. Gustave, R. Girod, I. Dusfour, I. Leparc-Goffart, D. L. Vanlandingham, Y. J. Huang, L. P. Lounibos, S. Mohamed Ali, A. Nougairede, X. de Lamballerie and A. B. Failloux (2015). "Chikungunya virus transmission potential by local Aedes mosquitoes in the Americas and Europe." PLoS Negl Trop Dis **9**(5): e0003780.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers and H. O. Smith (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

Vijayakumar, K., B. George, T. S. Anish, R. S. Rajasi, M. J. Teena and C. M. Sujina (2013). "Economic impact of chikungunya epidemic: out-of-pocket health expenditures during the 2007 outbreak in Kerala, India." Southeast Asian J Trop Med Public Health **44**(1): 54-61.

Vilela, A. P. P. (2013). Detecção de Dengue virus em fêmeas de Aedes aegypti coletadas em Caratinga (MG) e no campus Pampulha, UFMG. Doutorado, Universidade Federal de Minas Gerais.

Vivancos, A. P., M. Guell, J. C. Dohm, L. Serrano and H. Himmelbauer (2010). "Strand-specific deep sequencing of the transcriptome." Genome Res **20**(7): 989-999.

Vodovar, N., A. W. Bronkhorst, K. W. van Cleef, P. Miesen, H. Blanc, R. P. van Rij and M. C. Saleh (2012). "Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells." PLoS One **7**(1): e30861.

Walboomers, J. M., M. V. Jacobs, M. M. Manos, F. X. Bosch, J. A. Kummer, K. V. Shah, P. J. Snijders, J. Peto, C. J. Meijer and N. Munoz (1999). "Human papillomavirus is a necessary cause of invasive cervical cancer worldwide." J Pathol **189**(1): 12-19.

Wang, P. and R. R. Granados (2001). "Molecular structure of the peritrophic membrane (PM): identification of potential PM target sites for insect control." Arch Insect Biochem Physiol **47**(2): 110-118.

Wang, X. B., Q. Wu, T. Ito, F. Cillo, W. X. Li, X. Chen, J. L. Yu and S. W. Ding (2010). "RNAi-mediated viral immunity requires amplification of virus-derived siRNAs in Arabidopsis thaliana." Proc Natl Acad Sci U S A **107**(1): 484-489.

Weaver, S. C. and A. D. Barrett (2004). "Transmission cycles, host range, evolution and emergence of arboviral disease." Nat Rev Microbiol **2**(10): 789-801.

Webster, C. L., F. M. Waldron, S. Robertson, D. Crowson, G. Ferrari, J. F. Quintana, J. M. Brouqui, E. H. Bayne, B. Longdon, A. H. Buck, B. P. Lazzaro, J. Akorli, P. R. Haddrill and D. J. Obbard (2015). "The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster." PLoS Biol **13**(7): e1002210.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." Mol Biol Evol **18**(5): 691-699.

Willner, D., M. Furlan, M. Haynes, R. Schmieder, F. E. Angly, J. Silva, S. Tammadoni, B. Nosrat, D. Conrad and F. Rohwer (2009). "Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals." PLoS One **4**(10): e7370.

Wu, D., M. Wu, A. Halpern, D. B. Rusch, S. Yooseph, M. Frazier, J. C. Venter and J. A. Eisen (2011). "Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees." PLoS ONE **6**(3): e18011.

Wu, Q., Y. Luo, R. Lu, N. Lau, E. C. Lai, W.-X. Li and S.-W. Ding (2010). "Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs."

Proceedings of the National Academy of Sciences of the United States of America **107**: 1606-1611.

Yamao, T., Y. Eshita, Y. Kihara, T. Satho, M. Kuroda, T. Sekizuka, M. Nishimura, K. Sakai, S. Watanabe, H. Akashi, Y. Rongsriyam, N. Komalamisra, R. Srisawat, T. Miyata, A. Sakata, M. Hosokawa, M. Nakashima, N. Kashige, F. Miake, S. Fukushi, M. Nakauchi, M. Saijo, I. Kurane, S. Morikawa and T. Mizutani (2009). "Novel virus discovery in field-collected mosquito larvae using an improved system for rapid determination of viral RNA sequences (RDV ver4.0)." Arch Virol **154**(1): 153-158.

Yan, Y., H. Cui, S. Jiang, Y. Huang, X. Huang, S. Wei, W. Xu and Q. Qin (2011). "Identification of a novel marine fish virus, Singapore grouper iridovirus-encoded microRNAs expressed in grouper cells by Solexa sequencing." PLoS One **6**(4): e19148.

Zerbino, D. R. and E. Birney (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome research. **18:** 821-829.

Zhou, A., J. Paranjape, T. L. Brown, H. Nie, S. Naik, B. Dong, A. Chang, B. Trapp, R. Fairchild, C. Colmenares and R. H. Silverman (1997). "Interferon action and apoptosis are defective in mice devoid of 2',5'-oligoadenylate-dependent RNase L." EMBO J **16**(21): 6355-6363.

Zhuang, L., Z. Zhang, X. An, H. Fan, M. Ma, B. D. Anderson, J. Jiang, W. Liu, W. Cao and Y. Tong (2014). "An efficient strategy of screening for pathogens in wild-caught ticks and mosquitoes by reusing small RNA deep sequencing data." PLoS One **9**(3): e90831.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." J Theor Biol **8**(2): 357-366.

9. Anexos

9.1. Trabalhos completos publicados durante o período do doutorado

Aguiar, E. R. G. R. ; Olmo, R. P. ; Paro, S. ; Ferreira, F. V. ; De Faria, I. J. D. S. ; Todjro, Y. M. H. ; Lobo, F. P. ; Kroon, E. G. ; Meignin, C. ; Gatherer, D. ; Imler, J.-L. ; Marques, J. T. . "Sequence-Independent Characterization Of Víruses Based On The Pattern Of Viral Small Rnas Produced By The Host". Nucleic Acids Research, Vol. 43, N. 13., P. 6191-6206, 2015.

Campos, Rafael K ; Boratto, Paulo V ; Assis, Felipe L ; **Aguiar, Eric RGR** ; Silva, Lorena CF ; Albarnaz, Jonas D ; Dornas, Fabio P ; Trindade, Giliane S ; Ferreira, Paulo P ; Marques, João T ; Robert, Catherine ; Raoult, Didier ; Kroon, Erna G ; La Scola, Bernard ; Abrahão, Jônatas S . "Samba Vírus: A Novel Mimivírus From A Giant Rain Forest, The Brazilian Amazon". Virology Journal, V. 11, P. 95, 2014.

Marques, Joao Trindade ; Wang, Ji-Ping ; Wang, Xiaohong ; De Oliveira, Karla Pollyanna Vieira ; Gao, Catherine ; **Aguiar, Eric Roberto Guimaraes Rocha** ; Jafari, Nadereh ; Carthew, Richard W. . "Functional Specialization Of The Small Interfering Rna Pathway In Response To Vírus Infection". Plos Pathogens (Online), V. 9, P. E1003579, 2013.

9.2. Outros trabalhos desenvolvidos em durante o período do doutorado

1- Descoberta de novos vírus através do sequenciamento dos pequenos RNAs de mosquitos individuais de campo – Trabalho em andamento;

2- Comparação dos pequenos RNAs e genes das vias de RNAi em insetos vetores *Lutzomyia longipalpis* e *Aedes aegypti* e o organismo modelo *Drosophila melanogaster* – trabalho em preparação;

3- Entendendo a resposta imune inata a vírus de DNA utilizando a *Drosophila melanogaster* como modelo – Trabalho em andamento;

4- Análise da resposta do vetor Aedes aegypti à infecção por Dengue vírus: foco na barreira do intestino – trabalho em preparação;

5- Monitoramento da circulação do DENV entre mosquitos de campo e pacientes humanos – trabalho em preparação;

6- Identificação em picorna-like vírus de uma estrutura de RNA capaz de ativar resposta antiviral de RNAi – Trabalho em preparação;

9.3. Revisões da literatura escritas durante o doutorado:

1- Uso de pequenos RNAs como biomarcadores;

P.P. Vilela, A., E. R.G.R. Aguiar, F. V. Ferreira, L. S. Ribeiro, **R. P. Olmo** and J. T. Marques (2012). "Small Non-Coding RNAs as Biomarkers." Recent Patents on Biomarkers 2: 119-130.

Published online 3 June 2015

Nucleic Acids Research, 2015, Vol. 43, No. 13 ***6191–6206***
doi: 10.1093/nar/gkv587

Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host

Eric Roberto Guimarães Rocha Aguiar[1,2], Roenick Proveti Olmo[1,2], Simona Paro[2], Flavia Viana Ferreira[3], Isaque João da Silva de Faria[1], Yaovi Mathias Honore Todjro[1], Francisco Pereira Lobo[4], Erna Geessien Kroon[3], Carine Meignin[2,5], Derek Gatherer[6], Jean-Luc Imler[2,5,7] and João Trindade Marques[1,*]

[1]Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, CEP 30270-901, Brazil, [2]CNRS-UPR9022, Institut de Biologie Moléculaire et Cellulaire, 67084 Strasbourg Cedex, France, [3]Department of Microbiology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, CEP 30270-901, Brazil, [4]Laboratório Multiusuário de Bioinformática, Embrapa Informática Agropecuária, Campinas, São Paulo, CEP 13083-886, Brazil, [5]Faculté des Sciences de la Vie, Université de Strasbourg, 67083 Strasbourg Cedex, France, [6]Division of Biomedical and Life Sciences, Faculty of Health and Medicine, Lancaster University, Lancaster, Lancashire, LA1 4YQ, UK and [7]Institut d'Etudes Avancées de l'Université de Strasbourg (USIAS), 67084 Strasbourg Cedex, France

Received March 18, 2015; Revised May 5, 2015; Accepted May 24, 2015

ABSTRACT

Virus surveillance in vector insects is potentially of great benefit to public health. Large-scale sequencing of small and long RNAs has previously been used to detect viruses, but without any formal comparison of different strategies. Furthermore, the identification of viral sequences largely depends on similarity searches against reference databases. Here, we developed a sequence-independent strategy based on virus-derived small RNAs produced by the host response, such as the RNA interference pathway. In insects, we compared sequences of small and long RNAs, demonstrating that viral sequences are enriched in the small RNA fraction. We also noted that the small RNA size profile is a unique signature for each virus and can be used to identify novel viral sequences without known relatives in reference databases. Using this strategy, we characterized six novel viruses in the viromes of laboratory fruit flies and wild populations of two insect vectors: mosquitoes and sandflies. We also show that the small RNA profile could be used to infer viral tropism for ovaries among other aspects of virus biology. Additionally, our results suggest that virus detection utilizing small RNAs can also be applied to vertebrates, although not as efficiently as to plants and insects.

INTRODUCTION

Viruses are highly abundant in most biological systems and are characterized by an extraordinary diversity (1). Large-scale sequencing of RNA and DNA has been commonly used in metagenomic studies to assess the genetic diversity of viruses in a biological sample, referred to as the virome (1–5). In some cases, sample manipulation prior to sequencing, such as centrifugation and column filtration, are applied in order to enrich for viral sequences although such techniques can sometimes lead to contamination (6–8). Thus, direct nucleic acid extraction with few or no sample manipulation steps is the preferred strategy to minimize external contamination. However, the lack of viral enrichment may sometimes result in a majority of non-viral sequences in the library (9). Whether or not enrichment is employed, virus identification by metagenomics is inherently limited since it mostly relies on sequence similarity comparisons against reference databases. New strategies need to be developed to improve virus detection and help characterize novel unknown sequences commonly found in large-scale sequencing studies that are sometimes referred to as the 'dark matter' of metagenomics (10,11).

The characterization of insect viromes has particular public health significance since mosquitoes and other insect species can transmit human viral pathogens, such as

*To whom all correspondence should be addressed. Tel: +55 31 3409 2623; Fax: +55 31 3409 2613; Email: jtm@ufmg.br

Dengue virus and *Chikungunya virus* (12,13). Sequencing of small or long RNAs has been used to identify viruses in insects, although the potential advantages and disadvantages of each strategy are unclear (7,9,14–17). Notably, while long RNAs are direct products of viral replication and transcription, the biogenesis of small RNAs involves further processing of viral RNA products by host antiviral pathways such as RNA interference (RNAi). In insects and most animals, there are at least three different RNAi pathways that involve the production of distinct types of small RNAs, namely microRNAs (miRNAs), piwi-interacting RNAs (piRNAs) and small interfering RNAs (siRNAs). Each type of small RNA has a unique size distribution and nucleotide preference related to the RNAi pathway to which it belongs. In insects, RNA byproducts of viral replication can trigger the production of virus-derived small RNAs of length 21 and 24–29 nt, suggesting the activation of siRNA and piRNA pathways, respectively (15,18–21). Virus-derived siRNAs originate uniformly from both strands of genomes by processing of viral double-stranded RNA (dsRNA) generated in infected cells (18,22). The siRNA pathway is a major antiviral response against viruses containing DNA or RNA genomes since dsRNA seems to be a common by-product of viral replication (20,22–26). In contrast, virus-derived piRNAs that normally show a less uniform genome coverage can be generated from single-stranded RNA precursors but their function in controlling infection is less clear (20,21). Virus-derived siRNAs and piRNAs will associate with argonaute proteins to form the RNA-induced silencing complex that degrades complementary viral RNAs (18,22,23). In contrast to siRNAs and piRNAs, miRNAs are mostly derived from specific non-coding loci in the host genome and have no direct role in silencing of viral transcripts. Viruses can sometimes produce their own miRNAs but this seems to be mostly restricted to DNA viruses that infect vertebrate animals (27). In addition to the siRNA and piRNA pathways, other unrelated mechanisms can also generate virus-derived small RNAs from the degradation of viral RNAs, such as RNase L in mammals (28).

Here, we took advantage of the production of virus-derived small RNAs by the host response, to identify viruses within laboratory stocks of *Drosophila melanogaster* and wild populations of *Aedes aegypti* and *Lutzomyia longipalpis*. We show that small RNAs are relatively enriched and favour the detection of viral sequences compared to long RNAs in the same sample. This suggests that the production of virus-derived small RNAs by host antiviral pathways causes an enrichment of viral sequences in small RNAs compared to long RNAs. Moreover, we show that the size profile of small RNAs produced by host pathways is unique to each virus and can be used as a signature to classify and identify viral contigs independent of sequence similarity comparisons to known references. This pattern-based strategy overcomes a severe limitation of metagenomic approaches, allowing identification of novel viral contigs, which otherwise would have escaped detection by sequence-based methods. In addition, we noted that the small RNA profile could reflect aspects of virus biology since the activation of RNAi pathways is affected by viral genome structure and tissue tropism. We show that the profile of virus-derived small RNAs consistent with activation of the piRNA pathway in the germline, was successfully used to infer viral infection of mosquito ovaries. Using this small RNA based approach, we identified novel viruses from the *Bunyaviridae, Reoviridae* and *Nodaviridae* families that compose the virome of wild and laboratory insect populations. Using published small RNA datasets, we show that this strategy can also be broadly employed for the detection of viruses in animals and plants, although in vertebrates this application requires further validation.

MATERIALS AND METHODS

Sample processing and nucleic acid extraction

Aedes aegypti mosquitoes used on the experiments were obtained from laboratory colonies established from eggs collected in three neighbourhoods of Rio de Janeiro (Humaita, Tubiacanga and Belford Roxo), in southeastern Brazil. Laboratory colonies of *L. longipalpis* sandflies were derived from wild-caught animals captured in the city of Teresina, in northeastern Brazil. *Drosophila* libraries were prepared from wild-type laboratory stocks that were infected with *Vesicular stomatitis virus*, *Drosophila C virus* or *Sindbis virus* as described previously (29). Individual or pooled insects were anesthetized with carbon monoxide and directly ground in Trizol using glass beads. Ovaries were dissected from female mosquitoes and directly homogenized in Trizol reagent using a pipette. Total RNA or DNA was extracted using Trizol according to the manufacturer's protocol (Invitrogen).

RNA library construction

Total RNA extracted from three separate pools of mosquitoes, sandflies and fruit flies were used to construct independent small RNA libraries. In the case of mosquitoes, the same total RNA was used to also construct three independent long RNA libraries. Small RNAs were selected by size (~18–30 nt) on a denaturing PAGE before being used for construction of libraries as previously described (30). Long RNA libraries were constructed from total RNA that was poly(A) enriched and depleted for ribosomal RNA (rRNA) using a TruSeq Stranded Total RNA kit according to the manufacturer's protocol (Illumina). Sequencing was performed by the IGBMC Microarray and Sequencing platform, a member of the 'France Génomique' consortium (ANR-10-INBS-0009). Sequence strategy was 1 × 50 base pairs (bp) for small RNAs libraries and 2 × 100 bp (forward and reverse sequencing) resulting in an average read length of 190 nt total.

Pre-processing of RNA libraries

Raw sequenced reads from small RNA libraries were submitted to quality filtering and adaptor trimming using fastx_quality_filter and fastq_clipper respectively, both part of the fastx-toolkit package (version 0.0.14) (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Small RNA reads with phred quality below 20, shorter than 15 nt after trimming of adaptors or containing ambiguous bases, were discarded. In the case of long RNA libraries, raw sequenced reads were submitted to quality filtering using

fastx_quality_filter. Reads with phred quality below 20 or containing ambiguous bases were discarded. Remaining sequences from small or long RNA libraries were mapped to reference sequences from transposable elements, bacterial genomes (2739 complete genomes deposited in Genbank) and host genomes (*L. longipalpis*, *A. aegypti* and *D. melanogaster*) using Bowtie (version 1.1.1) for small RNA libraries or Bowtie2 (version 2.2.4) for long RNA libraries (one mismatch allowed) (31,32). The *Drosophila* genome (version v5.44) was downloaded from flybase.org. The latest versions of mosquito (Liverpool strain L3) and sandfly (Jacobina strain J1) genomes were downloaded from VectorBase (https://www.vectorbase.org/). Sequences of transposable elements were obtained from TEfam (http://tefam.biochem.vt.edu/tefam/). Remaining sequenced reads that did not map to transposable elements, host or bacterial genomes, referred to as processed reads, were used for contig assembly and subsequent analysis.

Contig assembly strategy

Processed reads were utilized for contig assembly using Velvet (version 1.0.13) (33). Assembly was performed in parallel using different strategies for each library. For small RNA libraries, we performed contig assembly using different size ranges of small RNA reads (20–23, 24–30 and 20–30 nt). In each case, parallel assembly strategies were performed using a fixed k-mer value (k-mer 15) with default parameters or a k-mer value between 15 and 31 defined automatically by VelvetOptimser (version 2.2.5) (http://bioinformatics.net.au/software.velvetoptimiser.shtml). For long RNA libraries, contig assembly was performed using a fixed k-mer value (k-mer 31) or an automatically defined k-mer value from 15 to 91. For each library, results from parallel contig assembly strategies were merged using CAP3 (version date 12-21-07) with max gap length in overlap of 2 and overlap length cutoff of 20 (34). Results from assembly strategies utilizing different size ranges of small RNAs were also combined using CAP3. Removal of redundant contig sequences was performed using BLASTClust program within the standalone BLAST package (version 4.0d) (35), requiring 50% of length with at least 50% of identity between contigs. Non-redundant contigs larger than 50 nt received specific IDs to indicate their origin and were further characterized.

Sequence-based characterization of contigs

Assembled contigs were characterized by sequence similarity (nucleotide and protein) to known sequences, with analysis of conserved domains if detected, and also examined for the presence of ORFs. We used BLAST for sequence similarity searches against non-redundant NCBI databases (nucleotides and protein). InterproScan (version 5.3–46.0) (36) and HMMer (version 3.0) (37) were used to verify the presence of open reading frames (ORFs) and conserved domains and the Pfam database (version 27.0) (38) to analyse protein domains. Hits with an *E*-value smaller than $1e^{-5}$ for nucleotide comparison or $1e^{-3}$ for protein comparison were considered significant. Viral genomic segments were classified as described (39).

Analysis of small RNA profiles

For pattern-based analysis, processed small RNA reads were mapped against contig or reference sequences using Bowtie allowing one mismatch. Small RNA size profile was calculated as the frequency of each small RNA read size from 15–35 nt mapped on the reference genome or contig sequence considering each polarity separately. We used a Z-score to normalize the small RNA size profile and to plot heatmaps for each contig or reference sequence using R (version 3.0.3) with gplots package (version 2.16.0). To evaluate the relationship between small RNA profiles from different contig or reference sequences, we computed the Pearson correlation (confidence interval >95%) of the Z-score values. Similarities between small RNA profiles were defined using hierarchical clustering with UPGMA as the linkage criterion. Groups of sequences with more than one element with at least 0.8 of Pearson correlation between each other were assigned to clusters. The density of small RNA coverage was calculated as the number of times that small RNA reads covered each nucleotide on the reference genome or contig sequence. Small RNA size profile and density of coverage were calculated using in-house Perl (version 5.12.4) scripts using BioPerl (version 1.6.923) and plotted using R with ggplot2 (version 1.0.1).

RT-PCR and Sanger sequencing

About 200 ng of total RNA extracted from insects was reverse transcribed into cDNA using MMLV reverse transcriptase. cDNA or DNA were subjected to polymerase chain reaction (PCR) reactions using specific primers. Oligonucleotide primers are listed in Supplementary Table S1 and were designed according to contig sequences obtained from our assembly pipeline. PCR products were subjected to direct Sanger sequencing.

RESULTS

Optimizing contig assembly from small RNA sequences

Large scale sequencing of small RNAs has been used for virus identification in insects and plants (15,16). Thus, we constructed small RNA libraries from laboratory stocks of *D. melanogaster* and wild populations of *A. aegypti* mosquitoes and *L. longipalpis* sandflies, two important vectors for human pathogens (Supplementary Table S2). *Drosophila* libraries were prepared from laboratory strains infected with three distinct viruses, *Drosophila C virus* (DCV), *Sindbis virus* (SINV) and *Vesicular stomatitis virus* (VSV) in order to help optimize our virus detection pipeline for small RNA sequences. Small RNA libraries were prepared from whole insects with no sample manipulation prior to RNA extraction to minimize risks of sample contamination in the laboratory, which is essential when processing field samples. After sequencing of small RNA libraries, data were processed to enrich for potential viral sequences by removing sequences derived from host and bacterial genomes. Host sequences corresponded to the vast majority of small RNAs (73–92%) in our libraries but a substantial percentage of sequences (3.4–14% of libraries) remained after these processing steps (Figure 1). However,

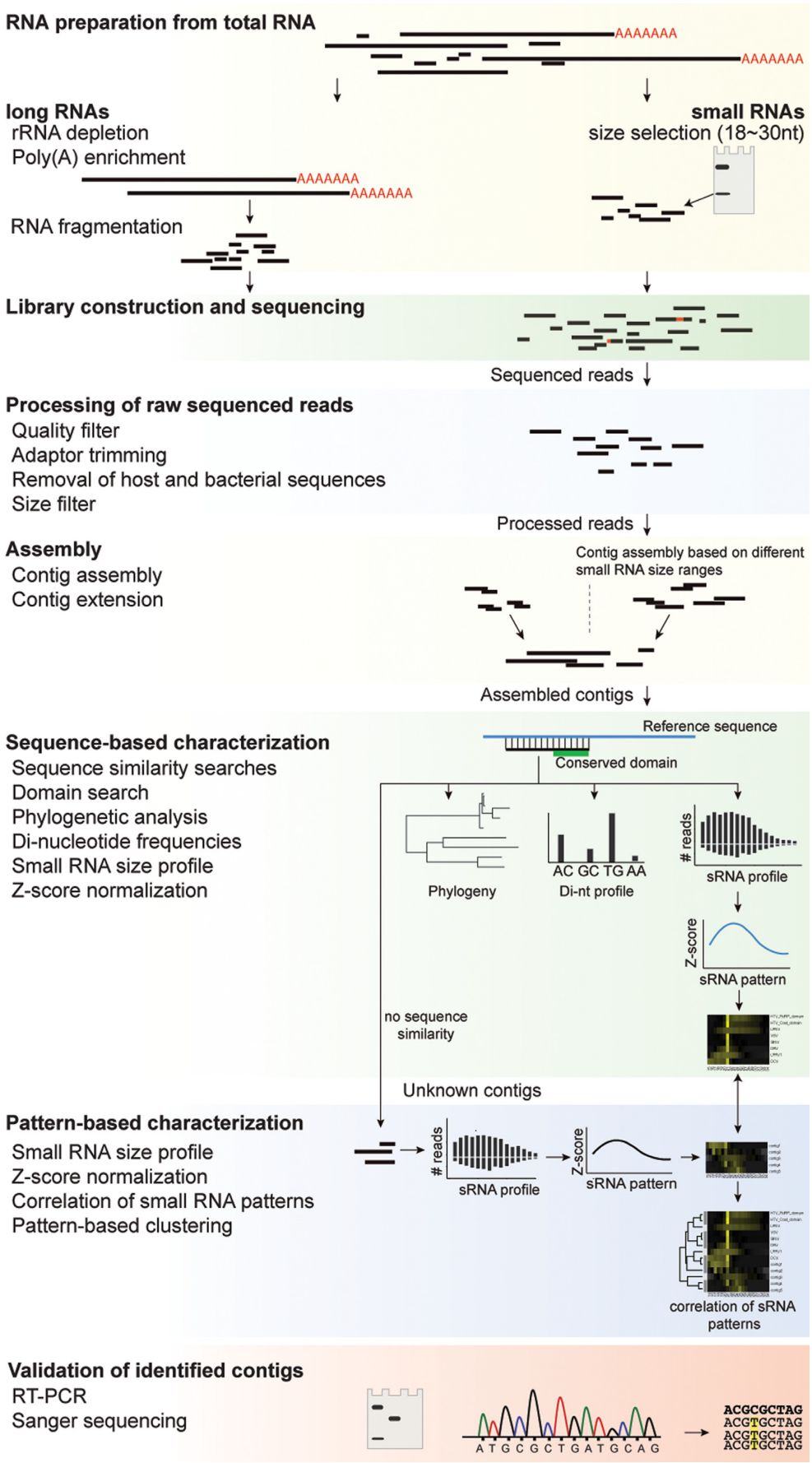

Figure 1. Overview of the pipeline for virus detection based on long and small RNAs. Different RNA fractions were utilized for the construction of small and long RNA libraries. Sequenced reads were processed to enrich for potential virus sequences. Processed reads were then utilized for contig assembly and extension. Contigs were characterized using both sequence-based and pattern-based strategies. Viral contigs were further validated by RT-PCR and Sanger sequencing. See text for details.

these are short sequences that need to be assembled into longer contiguous sequences (contigs) before being used for sequence-similarity searches against reference databases. Several studies have shown that virus-derived small RNAs are mostly 21 nt-long siRNAs (18,22,23). However, we reasoned that focusing on 21-nt long small RNAs could be shortsighted. Indeed, the piRNA pathway or degradation by other exonucleases may also generate virus-derived small RNAs in insect hosts (15,20,21). Importantly, these small RNAs of different origins could cooperate to allow the assembly of longer contigs. Therefore, we tested the use of different size ranges of small RNAs for contig assembly (Figure 2A). The best number and largest size of contigs were obtained when 20–23 and 24–30 nt small RNAs were utilized to assemble contigs separately and results combined afterwards (Figure 2A). Contig assembly utilizing other small RNA size ranges including 20–23, 24–30 or 20–30 nt small RNAs resulted in variable metrics depending on the library. Thus, the combination of contig assembly results utilizing 20–23 and 24–30 nt small RNAs separately seems to be more broadly applicable without prior knowledge of the small RNA profile.

Libraries from infected *Drosophila* were used to directly assess our virus detection strategy. In these libraries, we detected 42, 40 and 1 contigs that showed significant similarity against VSV, SINV and DCV, respectively (Supplementary Figure S1A). Thus, our approach could detect viruses known to be present in flies, although detection was limited by the number of viral small RNAs. We observed 1572 small RNAs derived from DCV that only allowed assembly of one contig covering 0.8% of the viral genome (Supplementary Figure S1). In contrast, 53 620 and 9588 small RNAs derived from VSV and SINV, respectively, allowed assembly of multiple independent contigs that covered 81.1 and 23.4% of the respective genomes (Supplementary Figure S1A). Thus, high coverage of viral genomes is important to allow contig assembly from overlapping small RNAs.

Next, all unique contigs assembled from *D. melanogaster, A. aegypti* and *L. longipalpis* small RNA libraries were utilized for sequence similarity searches against the NCBI non-redundant databases (nucleotide and protein). The vast majority of contigs assembled in all nine small RNA libraries (10 577 out of 11 806) did not show any significant sequence similarity and are hereafter referred to as unknown (Figure 2B). The large majority of these contigs (92%) are shorter than 100 nt thus hampering more detailed analyses. Nevertheless, clustering of our libraries based on the similarity of unknown sequences separates *Drosophila*, *Aedes* and *Lutzomyia* samples, suggesting that these contigs are host-specific (Supplementary Figure S2). The remaining 1229 non-redundant contigs were classified according to the taxon assigned to their most significant BLAST hit (Figure 2B). Several contigs showed similarity to animal sequences especially in the case of mosquito and sandfly libraries (Figure 2B). These likely belong to the insect genome but were not successfully removed in the pre-processing step. This may reflect the fact that the genomes of *A. aegypti* and *L. longipalpis* are not as well curated as the *D. melanogaster* genome (40,41). Several of the remaining contigs are derived from bacteria and fungi and could be part of the insect microbiome.

Sequence-based detection of viruses in contigs assembled from small RNAs

Out of 1229 non-redundant contigs, 223 (∼18%) showed significant similarity to viral sequences in reference databases (Figure 2B and Supplementary Table S2). The mean size of viral contigs was significantly longer than all the rest and included all the largest assembled sequences (Figure 2C). These results suggest that our small RNA based strategy favours assembly of long viral contigs compared to sequences of other origin. We removed 83 contigs derived from DCV, SINV or VSV that were among the 223 viral contigs. The remaining 140 viral contigs were filtered to eliminate similar sequences detected in more than one library from the same insect species. We also used overlap between contigs to further extend viral sequences. These steps allowed us to generate merged results from the three independent small RNA libraries from each insect population, *Drosophila*, *Aedes* and *Lutzomyia*. We were able to reduce 140 total viral contigs to 34 non-redundant sequences that could be assigned to at least seven viruses based on the most significant BLAST hit in reference databases (Table 1). Phylogenetic analysis suggests that six out of the seven viruses represent completely new species. Regarding the virome of each insect species, two viruses were detected in mosquitoes, three in sandflies and two in fruit flies.

In *Aedes* mosquitoes we detected contigs that belong to a novel strain of *Phasi Charoen Like-virus* (PCLV), a bunyavirus previously identified in mosquitoes from Thailand (Supplementary Figure S3A)(42). In addition, we also identified contigs from a novel virus related to *Laem Singh virus* (LSV) and two other recently described tick viruses, *Ixodes scapularis associated virus* 1 and 2 (Supplementary Figure S3B) (7), all of which are taxonomically unclassified. This new virus was named *Humaita-Tubiacanga virus* (HTV) to reflect the origin of its host.

In sandflies, we observed several non-redundant contigs showing similarity to reoviruses and nodaviruses (Table 1). Specifically, 23 non-redundant contigs showed sequence similarity to viruses of the genus *Cypovirus* from the family *Reoviridae* (Table 1). This number of unique viral contigs is high even considering the fact that reoviruses can have up to 12 genomic segments. Based on the phylogenetic analysis of genomic segments encoding viral RNA-dependent RNA polymerases (RdRPs), we were able to identify two distinct reoviral sequences that belong to the genus *Cypovirus* (Supplementary Figure S3C). These viruses were named *Lutzomyia Piaui reovirus* 1 (LPRV1) and *Lutzomyia Piaui reovirus 2* (LPRV2) to reflect their host and geographical location. Analyses of the other viral contigs assembled from sandfly libraries that showed similarity to nodaviruses suggest they belong to a novel virus related to *Nodamura virus* and a member of the genus *Alphanodavirus* (Supplementary Figure S3D). This novel virus was named *Lutzomyia Piaui nodavirus* (LPNV).

In fruit flies, we detected contigs that showed similarity to two viral families unrelated to the viruses used for experimental infections. That suggested the *Drosophila* laboratory stocks we used already carried unrecognized viral infections (Table 1). One set of contigs showed similarity to

Table 1. Summary of viruses identified in *Drosophila melanogaster*, *Aedes aegypti* and *Lutzomyia longipalpis*

Host	Virus family	Virus	Largest contig (nt)	Segment status[1]	# contig (sum of libraries)	ID strategy	Best hit	*E*-value	Accession number (size of reference in nt)
A. aegypti	*Bunyaviridae*	PCLV	3936	CC	4	blastx	glycoprotein precursor [Phasi Charoen-like virus]	0E + 00	AIF71031.1 (3852)
		PCLV	6807	CC	23	blastx	RdRP [Phasi Charoen-like virus]	0E + 00	AIF71030.1 (6783)
		PCLV	1332	CC	3	blastx	nucleocapsid [Phasi Charoen-like virus]	2E − 72	AIF71032.1 (1398)
	Unassigned	HTV	1609	CC	8	blastx	structural protein precursor [Drosophila A virus]	2E − 65	YP_003038596.1 (1326)
		HTV	2793	CC	13	blastx	putative RdRP [Laem Singh virus]	8E − 34	AAZ95951.1 (507)
L. longipalpis	*Reoviridae*	LPRV1	3762	CC	11	blastx	RdRP [Choristoneura occidentalis cypovirus 16]	3E − 173	ACA53380.1 (3675)
		LPRV1	3687	CC	5	blastx	VP3 [Inachis io cypovirus 2]	1E − 81	YP_009002593.1 (3450)
		LPRV1	3200	CC	2	blastx	VP4 [Inachis io cypovirus 2]	4E − 63	YP_009002588.1 (3201)
		LPRV1	1842	CC	2	blastx	VP5 [Inachis io cypovirus 2]	2E − 16	YP_009002589.1 (1899)
		LPRV1	841	CC	1	blastx	polyhedrin [Simulium ubiquitum cypovirus]	6E − 69	ABH85367.1 (836)
		LPRV1	3685	CC	1	blastx	VP2 [Inachis io cypovirus 2]	5E − 24	YP_009002587.1 (3649)
		LPRV1	1547	HQ	2	phmmer	unknown [Choristoneura occidentalis cypovirus 16]	900E − 03	ABW87641.1 (1946)
		LPRV1	2237	CC	3	blastx	unknown [Choristoneura occidentalis cypovirus 16]	200E − 01	ABW87640.1 (2214)
		LPRV1	2231	CC	1	pattern-based			
		LPRV1	1345	CC	1	pattern-based			
		LPRV1	688	HQ	1	pattern-based			
		LPRV1	680	HQ	1	pattern-based			
		LPRV2	3680	CC	1	blastx	RdRP [Bombyx mori cypovirus 1]	0E+00	AAK20302.1 (3854)
		LPRV2	1116	CC	1	blastx	polyhedrin [Heliothis armigera cypovirus 14]	4E-11	AAY34355.1 (956)
		LPRV2	2043 + 779 + 1392	SD	3	blastx	VP1 protein [Dendrolimus punctatus cypovirus 1]	4E-70	AAN84544.1 (4164)
		LPRV2	964	HQ	1	blastx	hypothetical protein LdcV14s9gp1 [Cypovirus 14]	2E-09	NP_149143.1 (1141)
		LPRV2	678 +1035 + 1617	SD	3	blastx	VP3 [Bombyx mori cypovirus 1]	5E-14	ADB95943.1 (3262)
		LPRV2	443 + 579 + 769	SD	3	blastx	viral structural protein 4 [Bombyx mori cypovirus 1]	2E-10	ACT78457.1 (1796)
		LPRV2	1516	HQ	1	blastx	VP2 protein [Dendrolimus punctatus cypovirus 1]	8E-53	AAN86620.1 (3846)
		LPRV2	599	HQ	4	blastx	unknown [Operophtera brumata cypovirus 18]	4E-10	ABB17215.1 (2883)
		LPRV2	286	HQ	2	blastx	putative VP5 [Dendrolimus punctatus cypovirus 1]	3E-02	AAO61786.1 (1501)
		LPRV2	641	SD	1	pattern-based			
		LPRV2	1212	SD	1	pattern-based			
		LPRV2	1174	CC	1	pattern-based			
		LPRV2	976	SD	1	pattern-based			
		LPRV2	535	SD	1	pattern-based			
	Nodaviridae	LPNV	2054	CC	5	blastx	capsid protein [Nudaurelia capensis beta virus]	1E − 42	NP_048060.1 (1836)
		LPNV	3189	CC	23	blastx	RdRP [Nodamura virus]	9E − 82	NP_077730.1 (3129)
D. melanogaster	*Unassigned*	DUV	1905 + 452	SD	2	blastx	protein P1 (RdRP) [Acyrthosiphon pisum virus]	2E −63	NP_620557.1 (10 035)
	Reoviridae	DRV	635 + 175	SD	2	blastx	RdRP [Fiji disease virus]	8E − 05	YP_249762.1 (4532)

[1]Segment status defined as described by Ladner *et al.* (39): SD: Standard Draft, HQ: High quality, CC: Coding complete, C: Complete, F: Finished.

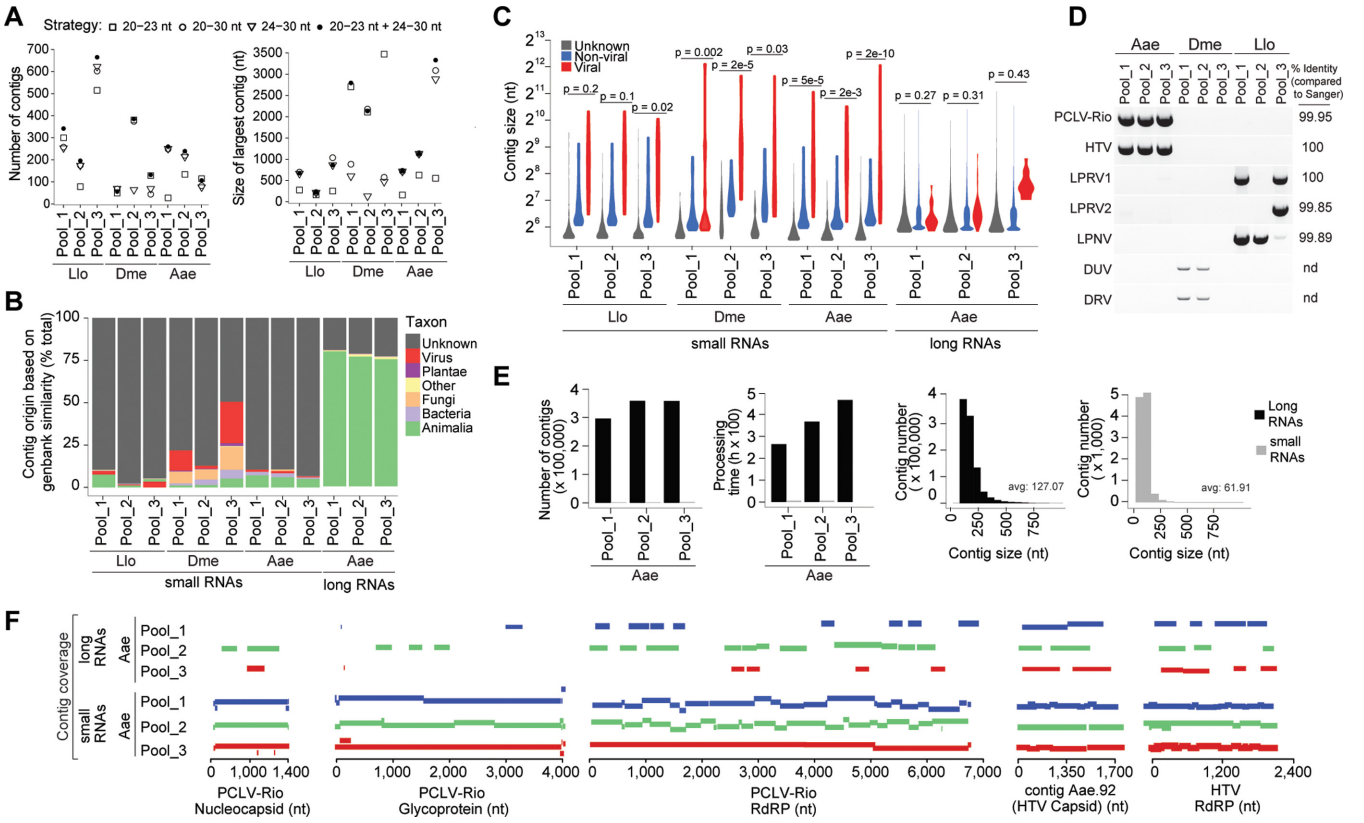

Figure 2. Small RNA sequencing identifies viral sequences more efficiently than long RNAs. (**A**) Comparison of number of contigs and size of largest contig in each small RNA library using different size ranges of small RNAs in the assembly step. (**B**) Proportion of contigs assembled in each library with significant similarity to reference sequences. The origin of contigs is classified by taxon and includes unknown sequences. (**C**) Size distribution of viral (red), non-viral (blue) and unknown contigs (grey) for each library. *P*-values for the difference between viral and non-viral contig sizes are indicated (Student *t*-test). (**D**) Viral RNA sequences were detected by RT-PCR from total RNA extracted from three separate pools of *Drosophila*, *Aedes* and *Lutzomyia* populations. Sanger sequencing of PCR products showed high identity to the sequence determined by our metagenomics approach as shown in the right column (not done; nd). (**E**) Comparison of processing time, number of contigs and frequency distribution of contig sizes for small and long RNA libraries shown in grey and black, respectively. (**F**) Coverage of PCLV and HTV genome segments by contigs assembled in each small and long RNA libraries from mosquitoes. Biological replicate samples are shown in blue, green and red.

reoviruses. Phylogeny suggests these belong to a virus of the genus *Fijivirus* of the family *Reoviridae* (Supplementary Figure S3C). This virus was named *Drosophila reovirus* (DRV). Another viral contig showed similarity to *Acyrthosiphon pisum virus* but could not be assigned to any known viral families by phylogeny (Supplementary Figure S3E). This virus was consequently named *Drosophila uncharacterized virus* (DUV).

Sequences corresponding to all seven potential new viruses were successfully amplified by PCR from reverse transcribed RNA but not from DNA (Figure 2D and data not shown). This indicates they are present in an RNA form, which is consistent with the observation that they presumably belong to viral families with RNA genomes (Figure 2D and Table 1). Sanger sequencing of PCR products showed 99–100% sequence identity to the contigs assembled using our strategy (Figure 2D). Importantly, all single nucleotide differences were also present in small RNA sequences from the individual libraries prior to contig assembly suggesting natural variations in virus populations (data not shown). Notably, the presence of these seven viruses was only detected in the corresponding insect populations utilized for the construction of small RNA libraries where they were first identified, *Drosophila, Aedes* or *Lutzomyia*. These results indicate that our strategy is not prone to generate artifacts.

Small RNAs are naturally enriched for viral sequences compared to long RNAs

We successfully detected seven viruses using small RNA libraries but had no basis to compare how this strategy would fare against other alternatives for detection of viral sequences. Other strategies utilize some type of sample manipulation in order to enrich for viral sequences prior to nucleic acid extraction although this may result in contamination (7,9). Direct sequencing of long RNAs is also utilized but can be limited by the abundance of host rRNA molecules that represent the vast majority of sequences in long RNA libraries (43). As an alternative, we constructed long RNA libraries after rRNA depletion and poly(A) enrichment from the same total RNA of *A. aegypti* populations used to prepare small RNA libraries (Supplementary Table S2). This allowed us to directly compare results from large scale sequencing of small and long RNAs from the same samples without manipulation prior to RNA extrac-

tion. The number and length of sequences obtained with long RNA libraries resulted in 10.4-fold more data compared to small RNA libraries. As a result, long RNA libraries generated a total of 1 011 347 contigs with N50 of ∼136 nt compared to 6066 contigs with N50 of ∼48 nt for small RNA libraries (Figure 2E and Supplementary Table S2). The larger number of contigs resulted in 43- to 72-fold longer processing times for similarity searches against databases comparing long and small RNA libraries (Figure 2E). Most contigs assembled from long RNA libraries (>60%) showed similarity to animal sequences and are likely to be unassembled parts of the *A. aegypti* genome (Figure 2B). Long RNA libraries also contained a large number of unknown sequences but they did not represent the majority of contigs as observed for small RNA libraries (Figure 2B). These results would suggest that long RNA libraries are more indicated for virus detection since they had significantly more contigs. However, the total number of viral contigs was very similar in small and long RNA libraries (Supplementary Table S2). Furthermore, the average size of viral contigs was longer in small RNA libraries, which resulted in larger coverage of viral genomes in all three independent samples (Figure 2C and F). Even though both strategies allowed detection of the same viruses, PCLV and HTV, these results indicate that small RNA libraries were enriched and naturally favoured assembly of viral sequences compared to long RNAs. The mechanism of small RNA biogenesis by host pathways appears to favour the generation of overlapping sequences that are likely to be important in allowing significant contig extension compared to sequencing of long RNAs. It is possible that viral RNAs could be further enriched in long RNAs had we not limited our sequencing to polyadenylated RNA molecules. Nevertheless, rRNA depletion alone could still bias sequencing results and small RNAs show natural enrichment for viral sequences without extensive processing steps prior to library construction.

Classifying viral sequences using small RNA pattern analysis

Our results indicate that small RNAs libraries favour the detection of viruses compared to long RNAs. However, the majority of contigs assembled from small RNAs were not identified by sequence similarity searches against reference databases. Sequence independent strategies are necessary to identify highly divergent viruses that have no known relatives. The size profile of virus derived small RNAs produced by the host pathways was unique for each virus analysed in this study including PCLV, SINV, VSV, DCV, HTV, LPNV, LPRV1, LPRV2, DRV and DUV (Figure 3A and Supplementary Figure S1B). Additionally, small RNA size profiles observed for contigs derived from other organisms such as Fungi and Bacteria were also very distinct (Supplementary Figure S4). In the case of segmented viral genomes, the small RNA size profile was remarkably similar for different segments of the same virus such as PCLV and LPNV (Figure 3A). These small RNA size profiles were consistent with diverse origins of small RNAs including production of siRNAs (peaks at 21 nt), piRNAs (peak from 27–28 nt) or degradation of viral RNAs (no size enrichment, strong bias for small RNAs corresponding to the genomic strand) but are hard to classify visually (Figure 3A and Supplementary Figure S1B). Thus, we used a Z-score to normalize the small RNA size profile and generate heatmaps for each contig that could be subjected to hierarchical clustering based on pairwise correlations to evaluate their relationship (Figure 3A). Using this strategy, small RNA size profiles of different viruses usually showed low correlation. By contrast, the small, medium and large segments of PCLV were grouped in a common cluster of similarity (cluster 7) as well as the RdRP and capsid segments of LPNV (cluster 5) (Figure 3B). Since contigs representing genomic segments of the same virus were grouped together, we tested whether the correlation of small RNA size profiles could help classify additional contigs that showed similarity to viral sequences but could not be further characterized.

In sandflies, based on the analysis of sequences encoding viral RdRPs, we were able to identify two separate reoviruses, namely LPRV1 and LPRV2. However, we observed another 21 non-redundant contigs showing similarity to reoviruses that could not be assigned to LPRV1 or LPRV2 solely based on sequence similarity. Since the small RNA size profile of LPRV1 and LPRV2 RdRP segments were clearly distinct (Figure 3A), we hypothesized it could be used to classify the origin of the remaining 21 reovirus contigs. Using this strategy, we observed that seven reovirus contigs were grouped together with the RdRP of LPRV1 (cluster 3) while 8 formed a cluster with LPRV2 RdRP (cluster 8) based on the similarity of the small RNA size profile (Table 1 and Figure 3B). Thus we analysed the expression of contigs in clusters 3 and 8 compared to the RdRPs of LPRV1 and LPRV2. Consistent with the small RNA profile similarity, contigs in cluster 3 are detected in the same libraries as the LPRV1 RdRP while contigs in cluster 8 follow the expression of the RdRP of LPRV2 (data not shown).

In mosquitoes, we identified one viral contig (Aae.92) of 1609 nt predicted to encode a protein with a coat domain (PF00729). This contig showed similarity with the capsid protein of *Drosophila A virus* (DAV) but phylogenetic analysis suggests the two viruses are considerably distinct (Supplementary Figure S5A). Phylogenetic analysis would seem to suggest that contig Aae.92 belongs to a viral family distinct from the two viruses that were also found in mosquitoes, PCLV and HTV. However, we note that DAV is an unusual virus, whose RdRP and capsid proteins show similarity to different viral families (Supplementary Figure S5) (44). Notably, the small RNA size profile for the HTV RdRP and contig Aae.92 were remarkably similar and clustered together based on correlation of the small RNA size profile (Figure 3B). Hence, we hypothesized that contig Aae.92 encodes the capsid protein of HTV as we only characterized a segment corresponding to the RdRP of this virus. In agreement with this hypothesis, we observed 100% correlation between the detection by RT-PCR of contig Aae.92 and the RdRP segment of HTV in individual mosquitoes (Figure 3C and data not shown).

A pattern-based strategy that identifies viral contigs in a sequence-independent manner

Viral contigs show unique small RNA size profiles that can be used to assign sequences to specific viruses in our

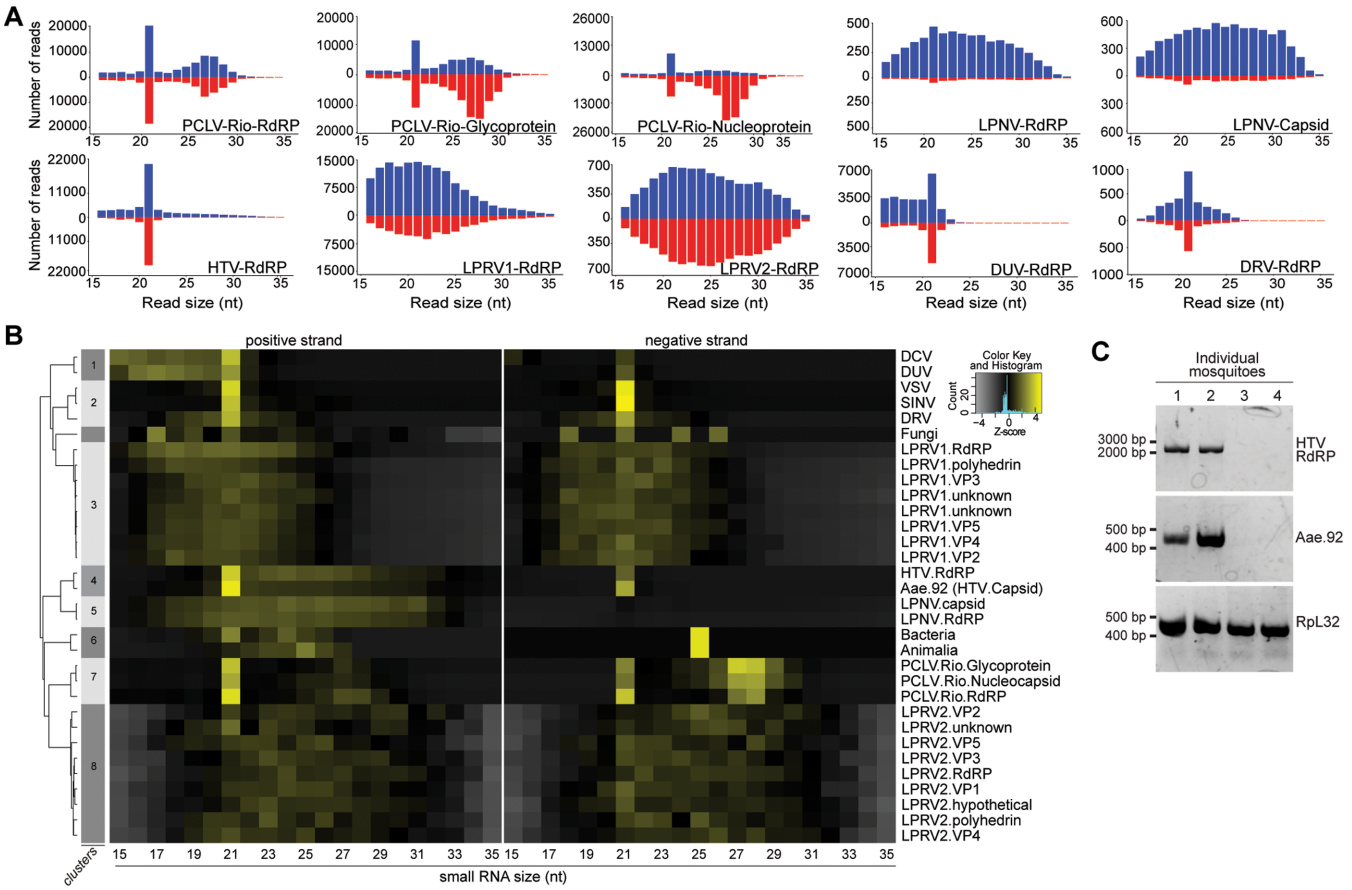

Figure 3. Small RNA size profile can classify uncharacterized viral contigs. (**A**) Small RNA size profile of previously characterized virus segments identified by sequence similarity searches. Blue and red represent small RNAs in the positive and negative strands, respectively. (**B**) Hierarchical clustering of viral contig sequences assembled in fruit fly, mosquito and sandfly libraries. Clustering was based on Pearson correlation of small RNA size profile shown as a heatmap. Clusters with more than one contig are indicated on the left vertical bar and numbered according to the order in which they appear from top to bottom. Clusters were defined by Pearson correlation above 0.8. (**C**) Contig Aae.92 and the segment corresponding to the HTV RdRP that grouped together by similarity of the small RNA size profile in panel (B) show perfect correlation of expression in individual mosquitoes as determined by RT-PCR. Results are representative of 46 individual mosquitoes that were analysed. The endogenous gene *Rpl32* was used as control for the RT-PCR.

samples. Possibly, this pattern analysis strategy could also help identify unknown contigs independently of sequence-similarity searches. In order to select prospective unknown contigs to be analysed, we noted that viral contigs were the largest assembled in our small RNA libraries (N50 of 208 nt compared to 63 nt for non-viral contigs) (Figure 2C). Thus, we used N50 as a proxy to filter 10 577 contigs representing unknown sequences and select 106 candidates longer than 208 nt. We eliminated sequence redundancy among these candidates, which resulted in 79 unique unknown contigs that were labelled according to their library of origin, *Lutzomyia* (Llo), *Drosophila* (Dme) or *Aedes* (Aae). Small RNA size profile was determined for all 79 unknown contigs and compared to previously characterize viral contigs using hierarchical clustering. This analysis generated 17 clusters containing more than one element, which were numbered sequentially according to the position in the heatmap. We observed that 72 out of the 79 unique unknown contigs were grouped in 11 different clusters of similarity (Figure 4A). Interestingly, clusters were composed of contigs assembled exclusively in libraries from the same insect, *Lutzomyia*, *Drosophila* or *Aedes*.

Unknown contigs found in *Lutzomyia* libraries were grouped in three separate clusters that showed clearly distinct small RNA patterns (Clusters 2, 6 and 17 in Figure 4A). Cluster 6 contained contigs with small RNA size profiles consistent with insect piRNAs (peak size between 27–28 nt) suggesting they could be derived from transposable elements (Figure 4A). Cluster 2 contained four unknown contigs that were grouped together and showed high correlation to previously identified LPRV1 segments (highlighted in red). Cluster 17 contained 19 unknown contigs that showed good correlation to LPRV2 segments (Figure 4A). Notably, in cluster 17, we observed that 5 of the 19 contigs formed a subgroup with correlation higher than 0.93 to LPRV2 RdRP segment (highlighted in red). These results suggest that some of the unknown contigs in *Lutzomyia* libraries could actually represent additional segments of LPRV1 and LRPV2. Indeed, based on the multi-segmented nature of reovirus genomes, we expected to find more segments for both LPRVs than were detected by sequence similarity searches (Table 1). In order to investigate this possibility, we analysed the expression of selected unknown contigs highlighted in cluster 2 and cluster 17 that presented

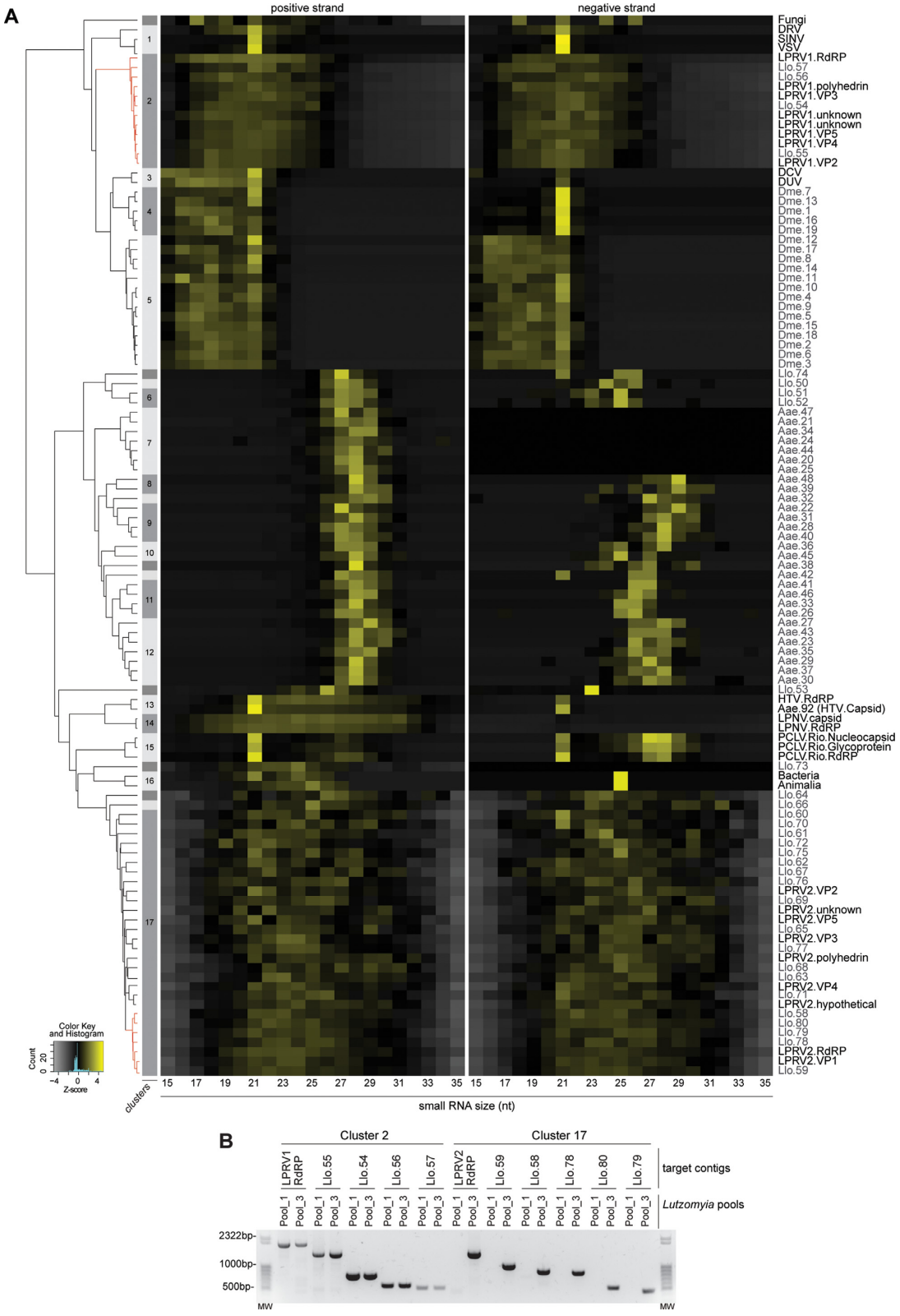

Figure 4. Small RNA pattern-based analysis identifies viral contigs without known relatives in reference databases. (**A**) Hierarchical clustering of viral and unknown contig sequences assembled in fruit fly, mosquito and sandfly libraries. Clustering was based on Pearson correlation of the small RNA size profile shown as a heatmap. Clusters with more than one contig are indicated on the left vertical bar and numbered according to the order in which they appear from top to bottom. Clusters were defined by Pearson correlation above 0.8. (**B**) Detection by RT-PCR in two separate pools of sandflies shows that contig sequences in Clusters 2 and 17 mimic the expression of RdRP segments of LPRV1 or LPRV2, respectively. The same pools of *Lutzomyia longipalpis* (pool1 and pool3) analysed in Figure 2D were used.

the highest correlation to the small RNA size profile of RdRP segments from LPRV1 or LPRV2, respectively. All four unknown contigs in cluster 2 perfectly mimicked the expression profile of the RdRP segment from LPRV1 while all five contigs from cluster 17 copied the expression of LPRV2 (Figure 4B). None of these nine new LPRV contigs showed significant similarity to reovirus sequences in reference databases, suggesting they are less conserved. Only one of these nine unknown contigs, contig Llo.58, assigned to LPRV2, had a complete ORF that was predicted to encode a 361 amino acid protein containing two putative domains (Supplementary Figure S6). The first domain is a Zn-dependent metallopeptidase from the Astacin superfamily found in digestive enzymes in both invertebrates and vertebrates (45,46). The second domain is a Peritrophin-A found in chitin-binding proteins that includes peritrophic matrix proteins of insect chitinases also found in baculoviruses (47). Thus, contig Llo.58 could encode a protein involved in the interaction between LPRV2 and sandflies since viruses commonly hijack and repurpose cellular proteins to their own advantage. Genes involved in host-pathogen interactions tend to be more divergent among viruses. Importantly, Llo.58 was not detected by similarity searches against reference databases and would not have been classified as viral based solely on domain prediction since these could also be found in cellular proteins. Thus, analysis of the small RNA size profile identified 23 unknown contigs representing additional segments of LPRV1 and LPRV2 genomes that have no similarity to known sequences in reference databases.

Unknown contigs found in *Drosophila* libraries were grouped in two separate clusters. Cluster 4 included five unknown contigs that showed high similarity to the cluster containing both DUV and DCV (Figure 4A). Cluster 5 contained another 14 unknown contigs that showed similarity to the profile of DUV and DCV albeit at lower correlation than cluster 4 (Figure 4A). Since the full genome sequence of DCV is known, these unknown contigs in the two separate clusters most likely represent different contigs from DUV. Indeed, we only identified two DUV contigs corresponding to the viral RdRP, which represents a small percentage of the full genome. In agreement with this hypothesis, these contigs were only found in the *Drosophila* library where DUV was identified (data not shown).

In *Aedes* libraries, 24 out of 27 unknown contigs were grouped in six different clusters (7, 8, 9, 10, 11 and 12) that showed high correlation to each other and a small RNA size profile consistent with mosquito piRNAs (27–28 nt peak in the size profile) (Figure 4A) (20,48). Accordingly, small RNAs derived from these contigs showed enrichment for U at position 1 and A at position 10, typical of insect piRNAs but no substantial 21 nt size peak nor symmetric coverage of both strands (data not shown). Thus, these sequences are most likely derived from repetitive regions that generate abundant piRNAs but are still absent from the current version of the *A. aegypti* genome.

The small RNA profile can provide information about virus biology

The pattern of small RNAs generated by the host response depends on virus characteristics such as genome structure, tissue tropism or strategy of replication. Thus, besides identifying viral contigs, the small RNA size profile may also provide specific information on the biology of each virus. For example, RNA viruses tend to have a very homogenous small RNA coverage of the viral genome while DNA viruses show clear hotspots of small RNAs (18,22,23). All viral contigs described here were derived from RNA viruses and mostly had homogeneous small RNA coverage.

We also noticed that HTV and PCLV showed very distinct small RNA size profiles despite being sometimes found in the same mosquitoes (Figures 2D and 3A). The profile of HTV showed a clear 21-nt peak size consistent with production of siRNAs. In contrast, the profile of PCLV showed two separate peaks of 21 and 24–29 nt, consistent with small RNAs generated by both siRNA and piRNA pathways. Indeed, 24–29 nt small RNAs derived from PCLV showed enrichment for U at position 1 and A at position 10, typical of sense and antisense insect piRNAs, respectively (Figure 5A). The insect piRNA pathway is mostly active in the germline where two mechanisms of small RNA biogenesis may occur (49). Primary sense piRNAs are generated by endonucleolytic processing of precursor transcripts while secondary piRNAs are produced by an amplification loop referred to as the ping-pong mechanism. We observed that 24–29 nt small RNAs derived from PCLV showed 10-nt overlap between sense and antisense RNAs consistent with the ping-pong amplification mechanism (Figure 5A) (50,51). These results suggest that PCLV induces the production of piRNAs by this mechanism when infecting the insect germline. In agreement with this hypothesis, we observed 75% prevalence for PCLV in ovaries of individual mosquitoes (Figure 5B). In contrast, HTV was not found in ovaries consistent with the fact it does not generate piRNAs (Figure 5B). Thus, the presence of a clear piRNA signature in the small RNA profile could help infer tissue tropism for the insect germline.

The lack of clear peaks in the size distribution of small RNAs may suggest inhibition of RNAi pathways such as reported for *Flock House virus* (FHV). Indeed, the B2 protein encoded by FHV is a powerful suppressor of silencing that blocks the RNAi pathway (52). Interestingly, ORF 2 in RNA 1 from the LPNV genome is predicted to encode a protein with similarity to the FHV B2 protein that could act as a suppressor of silencing (Supplementary Figures S6 and S7). Thus, the broad small RNA size profile with no clear peaks observed for LPNV could suggest inhibition of RNAi pathways as a strategy of replication (Figure 3A). A broad size profile and strong preference for small RNAs generated from the positive strand of the viral genome was also observed for DCV and DUV in infected fruit flies (Figure 3 and Supplementary Figure S1). Since the DCV-1A gene encodes a potent suppressor of the siRNA pathway (53), this suggests that DUV may also be capable of suppressing RNAi in infected flies.

Virus detection in published insect small RNA libraries

We decided to validate our strategy by analysing four published insect small RNA libraries constructed from adult mosquitoes and cell lines infected with SINV (Supplementary Table S2) (21,25). Sequence similarity searches showed

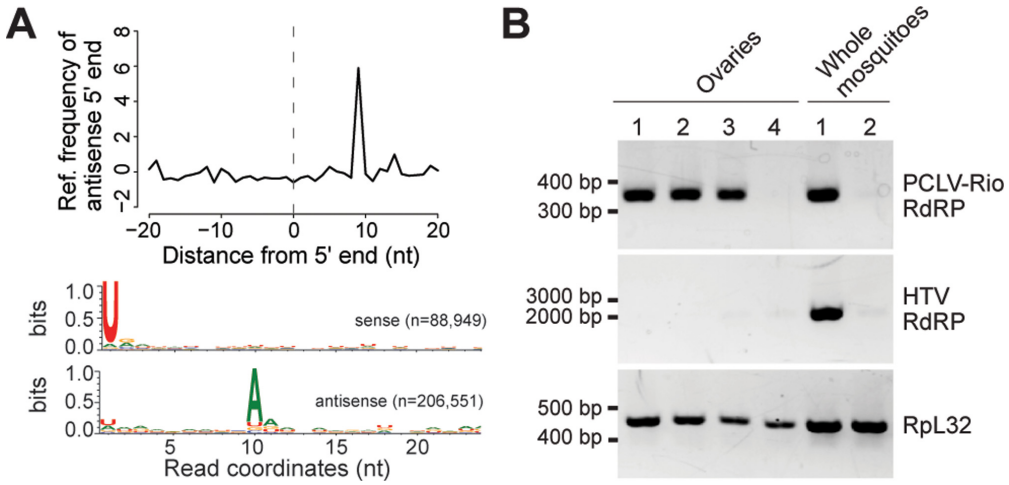

Figure 5. The presence of virus-derived piRNAs with a ping-pong signature is indicative of ovary infection. (**A**) About 24–29 nt small RNAs derived from PCLV show a 10 nt overlap between sense and antisense strands and U enrichment at position 1 and A enrichment at position 10 consistent with piRNAs generated by the ping-pong amplification mechanism found in the insect germline. (**B**) Both PCLV and HTV are detected in individual mosquitoes but only PCLV is present in ovaries as determined by RT-PCR. Results are representative of eight ovaries of individual mosquitoes that were analysed. The endogenous gene *Rpl32* was used as control for the RT-PCR.

that viral sequences represented 10.9% of contigs assembled from these datasets (Figure 6A). Size difference between viral and non-viral contigs was significant in most cases with the exception of mosquitoes infected with a recombinant SINV encoding the FHV B2 protein that almost completely blocks the RNAi pathway (Figure 6B) (54). Nevertheless, SINV sequences were detected among viral contigs in all libraries including the one where the RNAi was inhibited. This suggests that virus-derived small RNAs produced by host RNAi pathways are important but not essential for the assembly of viral contigs. Our approach also detected the presence of several contigs derived from viruses that were not reported at the time of first publication. We detected contigs derived from *A. aegypti densovirus 2* (AaDV2), *Mosquitoe X virus* (MXV) and *Cell fusion agent virus* (CFAV) in Aag2 cells, MXV and *Insect Iridescent virus- 6* (IIV6) in U4.4 cells and *Mosquito nodavirus* (MNV) in adult mosquitoes (Supplementary Table S3). Notably, a 1130 nt sequence corresponding to MNV was originally identified by another small RNA-based analysis pipeline in the library from adult mosquitoes (15). Using the same dataset, our strategy assembled a contig of 1994 nt (AaeS.82) that extended the original published MNV sequence of 1130 nt (Figure 6C). This 1994 nt MNV sequence contains the original ORF encoding the capsid protein and an additional incomplete ORF predicted to encode a protein with an RdRP_3 domain (PF00998) (Figure 6C). In addition, we detected a viral contig of 1702 nt (AaeS.81) that showed significant similarity to *Melon necrotic spot virus*, a member of *Tombusviridae* family (Supplementary Table S3). Contig AaeS.81 has one complete ORF of 397 aminoacids and a second incomplete ORF that contains a RdRP_3 conserved domain (PF00998), the same domain found in the MNV contig (Figure 6C). The small RNA size profile of contig AaeS.81 and MNV (AaeS.82) are very similar and showed correlation above 0.998 (Figure 6D). These results suggest that the 1994 nt MNV sequence and contig AaeS.81 could represent different fragments of the same viral genome (Figure 6C). In agreement with this hypothesis, contig AaeS.81 and MNV (AaeS.82) were only found in the same library prepared from adult mosquitoes infected with SINV.

In these published libraries, our pipeline also assembled a total of 1673 unknown contigs. Small RNA profiles were analysed for eight unknown contigs longer than the N50 observed for viral contigs (208 nt). Regarding the small RNA size profile, most viral contigs identified in published insect datasets were grouped in a single large cluster showing a 21 nt peak size consistent with typical siRNAs (Figure 6D). The lack of diversity in the small RNA profile can be explained by the higher homogeneity of these samples that are mostly derived from mosquitoes. Nevertheless, one unknown contig of 709 nt, contig AaeS.83, showed small RNA size profile similar to MNV (AaeS.82) and AaeS.81 and were grouped in the same cluster with correlation above 0.998 (Figure 6D). It is tempting to speculate that contig AaeS.83 might represent another missing piece of the MNV genome together with AaeS.81 (as suggested in Figure 6C).

We also identified two unknown contigs of 390 and 363 nt in U4.4 cells, U4.4.84 and U4.4.85, that showed a size profile similar to several viruses grouped together (Figure 6D). Contig U4.4.84 is predicted to encode two incomplete ORFs one of which shows limited similarity to *Megavirus terra 1* (Supplementary Figure S8A). High correlation of the small RNA size profile suggests U4.4.84 and U4.4.85 have the same origin. We also found small RNAs derived from contigs U4.4.84 and U4.4.85 in the small library prepared from Aag2 cells in the same laboratory as U4.4 cells (Supplementary Figure S8B). These observations could suggest an infectious virus that contaminated both cell cultures since small RNAs derived from contigs U4.4.84 and U4.4.85 were not observed in Aag2 cells from our own laboratory (data not shown). Another five unknown contigs were assembled in the library from Aag2 cells but

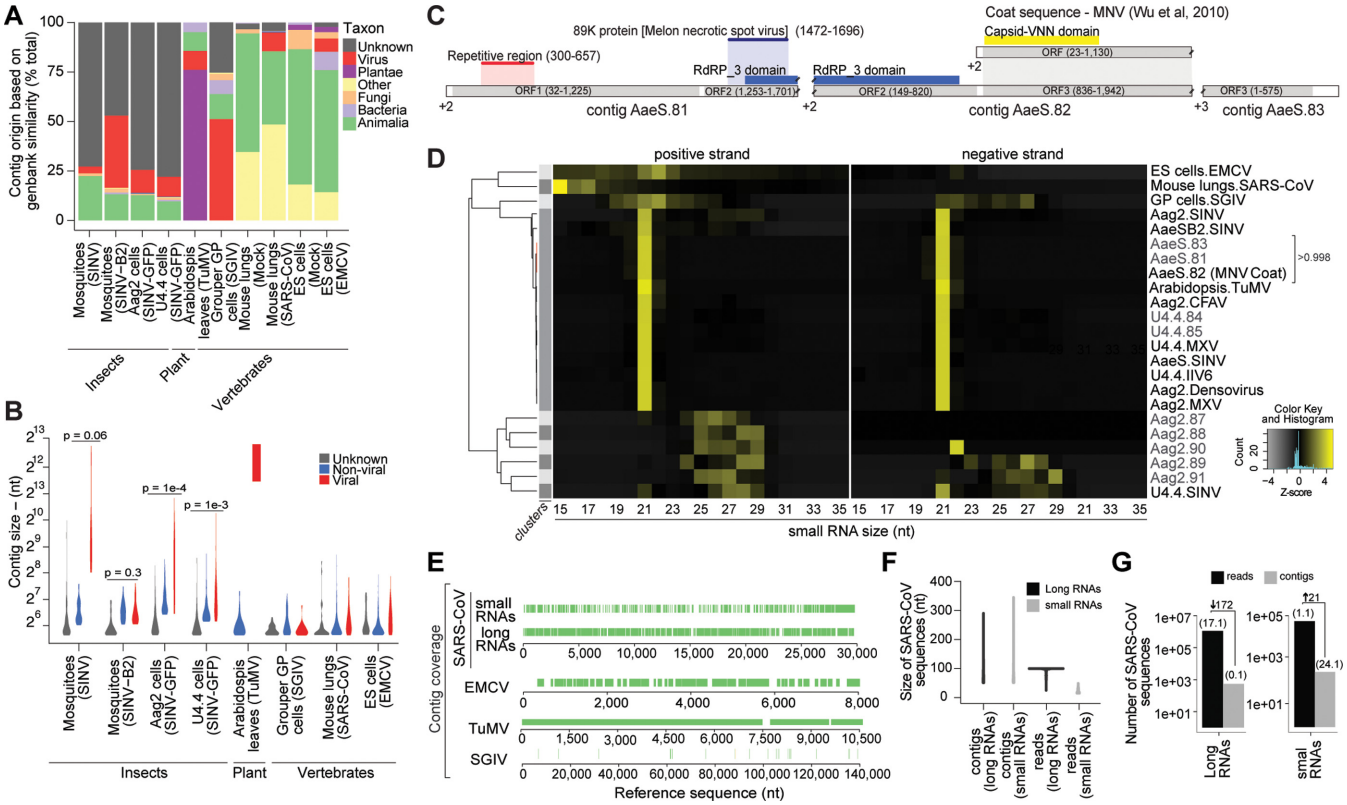

Figure 6. Virus detection based on large-scale sequencing of small RNAs is applicable to animals and plants. (**A**) Percentage of contigs assembled from published small RNA libraries from insects, plants and vertebrate animals with significant similarity against reference sequences. The origin of contigs is classified by taxon and includes unknown sequences. (**B**) Size distribution of contigs corresponding to viral (red), non-viral (blue) or unknown sequences (grey) in each library. *P*-values for the difference between contig sizes are indicated (Student *t*-test). (**C**) Hypothetical genome organization of MNV based on ORF and small RNA analysis of contigs AaeS.81, AaeS.82 and AaeS.83 identified in this study. (**D**) Hierarchical clustering of viral and unknown contig sequences assembled in published libraries. Clustering was based on Pearson correlation of the small RNA size profile shown as a heatmap. A single cluster with more than one contig is indicated on the left vertical bar as defined by correlation above 0.8. A sub-cluster highlighted in red contains small RNA profiles of three contigs that show Pearson correlation above 0.998. (**E**) Coverage of SARS-CoV, EMCV, TuMV and SGIV genomes by contigs assembled in RNA libraries from mouse lungs, ES cells, *Arabidopsis* and fish GP cells, respectively. (**F**) Size distribution of contigs and raw sequenced reads derived from SARS-CoV in long (black) or small (grey) RNA libraries from infected mouse lungs. (**G**) Number of raw reads and contigs sequences derived from viruses in long and small RNA libraries prepared from SARS-CoV infected mouse lungs. The number above bars indicates the percentage of viral reads and contigs sequences relative to the total. Fold enrichment or depletion of virus sequences comparing contigs to raw reads is shown.

showed small RNA profiles consistent with piRNAs suggesting these might represent repetitive regions absent from the mosquito genome.

Small RNAs allow efficient virus detection in plants and vertebrate animals

Virus detection utilizing small RNAs has been applied to insects and plants but not to vertebrates to the best of our knowledge (14–17). In order to further test our strategy, we analysed small RNA libraries prepared from *Arabidopsis thaliana* leaves infected with *Turnip mosaic virus* (TuMV), grouper fish GP cells infected with *Singapore grouper iridovirus* (SGIV), mouse lungs infected with *Severe acute respiratory syndrome coronavirus* (SARS-CoV) and mouse embryonic stem cells infected with *Encephalomyocarditis virus* (EMCV) (55–58). Samples infected with known viruses were chosen to provide proof-of-concept for detection. Although the small RNA profile was diverse, contigs corresponding to each virus were efficiently and specifically detected by sequence-based comparisons in the respective infected samples (Figure 6D and E). Notably, viral sequences assembled in *Arabidopsis* were among the largest contigs assembled in all libraries we analysed in this study (Figure 6B). This is most likely the result of highly efficient production of virus-derived small RNAs by the plant RNAi pathway that favours assembly of long viral contigs (Figure 6E) (16,59).

In contrast to *Arabidopsis*, viral contigs assembled in mouse and fish libraries were among the shortest (Figure 6B). This suggests that viral contig assembly from small RNAs in vertebrate animals is not as efficient as in insects and plants. Nevertheless, contigs were assembled and allowed identification of viruses in all fish and mouse small RNA libraries. In fish GP cells, the smaller size of SGIV contigs could be partially explained by restricted generation of dsRNA during the replication cycle of dsDNA viruses (23). Results obtained with mouse small RNA libraries suggest that activation of RNAi is not essential to allow assembly of viral contigs. EMCV contigs were detected in ES cells where the RNAi pathway is activated (56). In contrast, SARS-CoV contigs were assembled in mouse lungs

from small RNAs that were most likely generated by other RNAses (Figure 6F) (57). Notably, RNase L is an important antiviral factor that can degrade viral RNAs in mammalian cells independently of RNAi (28). The size distribution of viral contigs and number of virus-derived small RNAs was similar for EMCV and SARS-CoV (Figure 6B and E). Thus, small RNAs generated by the RNAi pathway or resulting from degradation by other RNases both allowed similar assembly of viral contigs in mouse samples.

We also directly compared virus identification from different RNA fractions prepared from mouse lungs infected with SARS-CoV (57,60). SARS-CoV contigs assembled from long and small RNAs had a similar size distribution despite the larger read size and numbers observed in the library prepared from long RNAs (Figure 6E and F). Small RNA libraries showed more than 20-fold enrichment of viral sequences among contigs when compared to raw reads (Figure 6G). In contrast, there is a 172-fold decrease in the percentage of viral sequences detected in assembled contigs compared to raw reads from long RNA libraries (Figure 6G). Thus, contig assembly from small RNAs favours assembly of SARS-CoV sequences compared to long RNAs even though no clear RNAi response is observed in mouse lungs. These preliminary results suggest that small RNAs show enrichment for viral sequences and can be used to assemble contigs not only in insects and plants but also mammals.

We also observed that very few contigs assembled by our pipeline in small RNA libraries from mice, fish and plants were unknown sequences (Figure 6A and B). Thus, our strategy to use the small RNA size profile to characterize unknown contigs could not be properly tested in these samples. Further testing is required to evaluate the application of our pattern-based approach to vertebrates and also plants.

DISCUSSION

In this study we describe a powerful approach based on small RNAs that allows for successful identification of viruses without any prior information about their presence. The majority of the viruses we identified potentially represent new species, illustrating the power of our strategy. Importantly, our results strongly indicate that virus identification from small RNAs provides four notable advantages compared to other metagenomic strategies. Firstly, preparation of small RNA libraries requires little sample manipulation and no column filtration steps before RNA extraction. This minimizes the chance of sample contamination or bias that can affect virus discovery by metagenomic studies (8). Secondly, we demonstrate that large scale sequencing of small RNAs optimizes the detection of viruses since these are naturally enriched for viral sequences and favour assembly of longer contigs compared to long RNAs. This is likely a result of the mechanism of small RNA biogenesis by host antiviral pathways that seem to efficiently generate large amounts of overlapping virus-derived small RNAs. Thirdly, we show that the small RNA size profile can help identify and characterize potential novel viral sequences for which we would otherwise have no other information. Indeed, large-scale sequencing projects currently face limitations due to the amount of sequences without known relatives in reference databases (10). We observed that small RNA size profiles are quite specific, and show that pattern similarities can be used to identify novel viral sequences. Using this approach we characterized novel viral segments of three viruses described in this study, HTV, LPRV1 and LPRV2. Fourthly, we show that the small RNA profile could help infer specific features of virus biology such as genome structure, tissue tropism and replication strategies. Indeed, based on the presence of a signature observed for activation of the piRNA pathway in the insect germline, we demonstrated that PCLV but not HTV is found in mosquito ovaries.

A large part of our strategy was based on the diversity of virus-derived small RNA profiles observed in infected insects. Although virus-derived small RNAs profiles can be very heterogeneous in infected insects, only the production of 21 nt long virus-derived siRNAs has classically been considered a hallmark of antiviral immunity (15,17,18,20–23,61). Our high-throughput analysis of three insect species infected with 10 different viruses shows a diversity of virus-derived small RNAs profiles that do not reflect technical differences in sample preparation, processing or analysis. Rather, these distinct profiles of virus-derived small RNA profiles seem to reflect divergent strategies of viral replication and host-specific antiviral responses.

Using our small RNA-based approach we characterized the virome of laboratory stocks of fruit flies and wild populations of two vector insects, mosquitoes and sandflies. These included six novel viruses and a strain of PCLV previously described in mosquitoes from Thailand. Of particular significance, we identified viruses belonging to viral families (e.g. *Bunyaviridae* and *Reoviridae*) that include several mammalian pathogens. Future studies should evaluate the presence of these viruses in wild mosquito and sandfly populations in Brazil as a potential threat for humans and livestock. In addition, these viruses could affect the ability of vector insects to carry other human pathogens such as *Dengue virus* and *Leishmania*, naturally transmitted by mosquitoes and sandflies, respectively. Together our results indicate that sequencing of small RNAs is a powerful virus surveillance strategy in research laboratories as well as natural settings.

Our small RNA based strategy was also successful in characterizing viruses in published small RNA datasets from plants, fish and mammals in addition to insects. In the case of mouse samples, enrichment of viral sequences in the small RNA fraction was observed even in the absence of activation of the RNAi pathway. Thus, efficient production of virus-derived small RNAs might be a broad phenomenon that can be further explored for virus detection. Indeed, multiple mammalian antiviral pathways, including RNAi and RNase L, can generate small RNAs during viral infection (28,56). However, viral contigs assembled from small RNAs were all identified by sequence similarity searches against reference databases. Thus, more extensive analyses are still required in order to evaluate whether our small RNA profile based approach can have broad applications to plants and animals.

ACCESSION NUMBERS

Datasets were deposited on the Small Read Archive of the National Center for Biotechnology Information under accession numbers described in Supplementary Table S2. Viral sequences described in Table 1 were deposited in Genbank under accession numbers KR003784–KR003824.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank L. Moreira (Fiocruz-MG) for providing the mosquito colony; N. Gontijo, M. Sant'Anna and C. Nonato (Universidade Federal de Minas Gerais) for the sandfly colony; R. Carthew for invaluable suggestions on the manuscript; J.A. Hoffmann, B. Drummond and members of the Marques and Imler laboratories for discussion; B. Claydon, E. Santiago, A. Courtin, K. Pansanato and R. Bianchini for technical help. We thank the IGBMC core facility in Strasbourg for sequencing.

FUNDING

Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) [478986/2010-6; 590069/2011-0; 400648/2013-0; 473092/2013-1]; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) [CBB - APQ-01228-11; CBB-APQ-00239-14] (to J.T.M. and E.G.K.); Agence Nationale de la Recherche [ANR-11-ASV3-002]; Investissement d'Avenir Programs [ANR-10-LABX-36; ANR-11-EQPX-0022]; National Institute of Health [PO1 AI070167] (to J.L.I.). E.R.G.R.A, R.P.O., F.V.F., I.J.S.F. and Y.M.H.T. received fellowships from CAPES and CNPq. Funding for open access charge: CAPES and CNPq.
Conflict of interest statement. None declared.

REFERENCES

1. Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat. Rev. Microbiol.*, **3**, 504–510.
2. Djikeng,A., Kuzmickas,R., Anderson,N.G. and Spiro,D.J. (2009) Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One*, **4**, e7264.
3. Riesenfeld,C.S., Schloss,P.D. and Handelsman,J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, **38**, 525–552.
4. Victoria,J.G., Kapoor,A., Dupuis,K., Schnurr,D.P. and Delwart,E.L. (2008) Rapid identification of known and new RNA viruses from animal tissues. *PLoS Pathog.*, **4**, e1000163.
5. Willner,D., Furlan,M., Haynes,M., Schmieder,R., Angly,F.E., Silva,J., Tammadoni,S., Nosrat,B., Conrad,D. and Rohwer,F. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*, **4**, e7370.
6. Oude Munnink,B.B., Jazaeri Farsani,S.M., Deijs,M., Jonkers,J., Verhoeven,J.T., Ieven,M., Goossens,H., de Jong,M.D., Berkhout,B., Loens,K. *et al.* (2013) Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS One*, **8**, e78454.
7. Tokarz,R., Williams,S.H., Sameroff,S., Sanchez Leon,M., Jain,K. and Lipkin,W.I. (2014) Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. *J. Virol.*, **88**, 11480–11492.
8. Naccache,S.N., Greninger,A.L., Lee,D., Coffey,L.L., Phan,T., Rein-Weston,A., Aronsohn,A., Hackett,J. Jr, Delwart,E.L. and Chiu,C.Y. (2013) The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J. Virol.*, **87**, 11966–11977.
9. Li,C.X., Shi,M., Tian,J.H., Lin,X.D., Kang,Y.J., Chen,L.J., Qin,X.C., Xu,J., Holmes,E.C. and Zhang,Y.Z. (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife*, **4**, eLife.05979.
10. Oh,J., Byrd,A.L., Deming,C., Conlan,S., Program,N.C.S., Kong,H.H. and Segre,J.A. (2014) Biogeography and individuality shape function in the human skin metagenome. *Nature*, **514**, 59–64.
11. Wu,D., Wu,M., Halpern,A., Rusch,D.B., Yooseph,S., Frazier,M., Venter,J.C. and Eisen,J.A. (2011) Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLoS One*, **6**, e18011.
12. Shepard,D.S., Undurraga,E.A. and Halasa,Y.A. (2013) Economic and disease burden of dengue in Southeast Asia. *PLoS Negl. Trop. Dis.*, **7**, e2055.
13. Vijayakumar,K., George,B., Anish,T.S., Rajasi,R.S., Teena,M.J. and Sujina,C.M. (2013) Economic impact of chikungunya epidemic: out-of-pocket health expenditures during the 2007 outbreak in Kerala, India. *Southeast Asian J. Trop. Med. Public Health*, **44**, 54–61.
14. Cook,S., Chung,B.Y., Bass,D., Moureau,G., Tang,S., McAlister,E., Culverwell,C.L., Glucksman,E., Wang,H., Brown,T.D. *et al.* (2013) Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. *PLoS One*, **8**, e80720.
15. Wu,Q., Luo,Y., Lu,R., Lau,N., Lai,E.C., Li,W.X. and Ding,S.W. (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 1606–1611.
16. Kreuze,J.F., Perez,A., Untiveros,M., Quispe,D., Fuentes,S., Barker,I. and Simon,R. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology*, **388**, 1–7.
17. Ma,M., Huang,Y., Gong,Z., Zhuang,L., Li,C., Yang,H., Tong,Y., Liu,W. and Cao,W. (2011) Discovery of DNA viruses in wild-caught mosquitoes using small RNA high throughput sequencing. *PLoS One*, **6**, e24758.
18. Mueller,S., Gausson,V., Vodovar,N., Deddouche,S., Troxler,L., Perot,J., Pfeffer,S., Hoffmann,J.A., Saleh,M.C. and Imler,J.L. (2010) RNAi-mediated immunity provides strong protection against the negative-strand RNA vesicular stomatitis virus in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 19390–19395.
19. Wang,X.H., Aliyari,R., Li,W.X., Li,H.W., Kim,K., Carthew,R., Atkinson,P. and Ding,S.W. (2006) RNA interference directs innate immunity against viruses in adult Drosophila. *Science*, **312**, 452–454.
20. Morazzani,E.M., Wiley,M.R., Murreddu,M.G., Adelman,Z.N. and Myles,K.M. (2012) Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. *PLoS Pathog.*, **8**, e1002470.
21. Vodovar,N., Bronkhorst,A.W., van Cleef,K.W., Miesen,P., Blanc,H., van Rij,R.P. and Saleh,M.C. (2012) Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PLoS One*, **7**, e30861.
22. Marques,J.T., Wang,J.P., Wang,X., de Oliveira,K.P., Gao,C., Aguiar,E.R., Jafari,N. and Carthew,R.W. (2013) Functional specialization of the small interfering RNA pathway in response to virus infection. *PLoS Pathog.*, **9**, e1003579.
23. Kemp C1,M.S., Goto,A., Barbier,V., Paro,S., Bonnay,F., Dostert,C., Troxler,L., Hetru,C., Meignin,C., Pfeffer,S. *et al.* (2013) Broad RNA interference-mediated antiviral immunity and virus-specific inducible responses in Drosophila. *J. Immunol.*, **190**, 650–658.
24. Aliyari,R., Wu,Q., Li,H.W., Wang,X.H., Li,F., Green,L.D., Han,C.S., Li,W.X. and Ding,S.W. (2008) Mechanism of induction and suppression of antiviral immunity directed by virus-derived small RNAs in Drosophila. *Cell Host Microbe*, **4**, 387–397.
25. Myles,K.M., Wiley,M.R., Morazzani,E.M. and Adelman,Z.N. (2008) Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 19938–19943.
26. Weber,F., Wagner,V., Rasmussen,S.B., Hartmann,R. and Paludan,S.R. (2006) Double-stranded RNA is produced by

positive-strand RNA viruses and DNA viruses but not in detectable amounts by negative-strand RNA viruses. *J. Virol.*, **80**, 5059–5064.
27. Umbach,J.L. and Cullen,B.R. (2009) The role of RNAi and microRNAs in animal virus replication and antiviral immunity. *Genes Dev.*, **23**, 1151–1164.
28. Girardi,E., Chane-Woon-Ming,B., Messmer,M., Kaukinen,P. and Pfeffer,S. (2013) Identification of RNase L-dependent, 3'-end-modified, viral small RNAs in Sindbis virus-infected mammalian cells. *MBio*, **4**, e00698–00613.
29. Galiana-Arnoux,D., Dostert,C., Schneemann,A., Hoffmann,J.A. and Imler,J.L. (2006) Essential function in vivo for Dicer-2 in host defense against RNA viruses in drosophila. *Nat. Immunol.*, **7**, 590–597.
30. Pfeffer,S., Zavolan,M., Grasser,F.A., Chien,M., Russo,J.J., Ju,J., John,B., Enright,A.J., Marks,D., Sander,C. *et al.* (2004) Identification of virus-encoded microRNAs. *Science*, **304**, 734–736.
31. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
32. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
33. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
34. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
35. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
36. Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.*, **396**, 59–70.
37. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
38. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
39. Ladner,J.T., Beitzel,B., Chain,P.S., Davenport,M.G., Donaldson,E.F., Frieman,M., Kugelman,J.R., Kuhn,J.H., O'Rear,J., Sabeti,P.C. *et al.* (2014) Standards for sequencing viral genomes in the era of high-throughput sequencing. *MBio*, **5**, e01360–01314.
40. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.
41. Nene,V., Wortman,J.R., Lawson,D., Haas,B., Kodira,C., Tu,Z.J., Loftus,B., Xi,Z., Megy,K., Grabherr,M. *et al.* (2007) Genome sequence of Aedes aegypti, a major arbovirus vector. *Science*, **316**, 1718–1723.
42. Yamao,T., Eshita,Y., Kihara,Y., Satho,T., Kuroda,M., Sekizuka,T., Nishimura,M., Sakai,K., Watanabe,S., Akashi,H. *et al.* (2009) Novel virus discovery in field-collected mosquito larvae using an improved system for rapid determination of viral RNA sequences (RDV ver4.0). *Arch. Virol.*, **154**, 153–158.
43. Vivancos,A.P., Guell,M., Dohm,J.C., Serrano,L. and Himmelbauer,H. (2010) Strand-specific deep sequencing of the transcriptome. *Genome Res.*, **20**, 989–999.
44. Ambrose,R.L., Lander,G.C., Maaty,W.S., Bothner,B., Johnson,J.E. and Johnson,K.N. (2009) Drosophila A virus is an unusual RNA virus with a T = 3 icosahedral core and permuted RNA-dependent RNA polymerase. *J. Gen. Virol.*, **90**, 2191–2200.
45. Wang,P. and Granados,R.R. (2001) Molecular structure of the peritrophic membrane (PM): identification of potential PM target sites for insect control. *Arch. Insect Biochem. Physiol.*, **47**, 110–118.
46. Arolas,J.L., Vendrell,J., Aviles,F.X. and Fricker,L.D. (2007) Metallocarboxypeptidases: emerging drug targets in biomedicine. *Curr. Pharm. Des.*, **13**, 349–366.
47. Lepore,L.S., Roelvink,P.R. and Granados,R.R. (1996) Enhancin, the granulosis virus protein that facilitates nucleopolyhedrovirus (NPV) infections, is a metalloprotease. *J. Invertebr. Pathol.*, **68**, 131–140.
48. Arensburger,P., Hice,R.H., Wright,J.A., Craig,N.L. and Atkinson,P.W. (2011) The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics*, **12**, 606.
49. Malone,C.D., Brennecke,J., Dus,M., Stark,A., McCombie,W.R., Sachidanandam,R. and Hannon,G.J. (2009) Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell*, **137**, 522–535.
50. Gunawardane,L.S., Saito,K., Nishida,K.M., Miyoshi,K., Kawamura,Y., Nagami,T., Siomi,H. and Siomi,M.C. (2007) A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science*, **315**, 1587–1590.
51. Brennecke,J., Aravin,A.A., Stark,A., Dus,M., Kellis,M., Sachidanandam,R. and Hannon,G.J. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, **128**, 1089–1103.
52. Han,Y.-H., Luo,Y.-J., Wu,Q., Jovel,J., Wang,X.-H., Aliyari,R., Han,C., Li,W.-X. and Ding,S.-W. (2011) RNA-based immunity terminates viral infection in adult Drosophila in the absence of viral suppression of RNA interference: characterization of viral small interfering RNA populations in wild-type and mutant flies. *J. Virol.*, **85**, 13153–13163.
53. van Rij,R.P., Saleh,M.C., Berry,B., Foo,C., Houk,A., Antoniewski,C. and Andino,R. (2006) The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster. *Genes Dev.*, **20**, 2985–2995.
54. Adelman,Z.N., Anderson,M.A., Liu,M., Zhang,L. and Myles,K.M. (2012) Sindbis virus induces the production of a novel class of endogenous siRNAs in Aedes aegypti mosquitoes. *Insect Mol. Biol.*, **21**, 357–368.
55. Cao,M., Du,P., Wang,X., Yu,Y.Q., Qiu,Y.H., Li,W., Gal-On,A., Zhou,C., Li,Y. and Ding,S.W. (2014) Virus infection triggers widespread silencing of host genes by a distinct class of endogenous siRNAs in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 14613–14618.
56. Maillard,P.V., Ciaudo,C., Marchais,A., Li,Y., Jay,F., Ding,S.W. and Voinnet,O. (2013) Antiviral RNA interference in mammalian cells. *Science*, **342**, 235–238.
57. Peng,X., Gralinski,L., Ferris,M.T., Frieman,M.B., Thomas,M.J., Proll,S., Korth,M.J., Tisoncik,J.R., Heise,M., Luo,S. *et al.* (2011) Integrative deep sequencing of the mouse lung transcriptome reveals differential expression of diverse classes of small RNAs in response to respiratory virus infection. *MBio*, **2**, e00198–e00111.
58. Yan,Y., Cui,H., Jiang,S., Huang,Y., Huang,X., Wei,S., Xu,W. and Qin,Q. (2011) Identification of a novel marine fish virus, Singapore grouper iridovirus-encoded microRNAs expressed in grouper cells by Solexa sequencing. *PLoS One*, **6**, e19148.
59. Pumplin,N. and Voinnet,O. (2013) RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence. *Nat. Rev. Microbiol.*, **11**, 745–760.
60. Josset,L., Tchitchek,N., Gralinski,L.E., Ferris,M.T., Eisfeld,A.J., Green,R.R., Thomas,M.J., Tisoncik-Go,J., Schroth,G.P., Kawaoka,Y. *et al.* (2014) Annotation of long non-coding RNAs expressed in collaborative cross founder mice in response to respiratory virus infection reveals a new class of interferon-stimulated transcripts. *RNA Biol.*, **11**, 875–890.
61. Marques,J.T. and Carthew,R.W. (2007) A call to arms: coevolution of animal viruses and host innate immune responses. *Trends Genet.*, **23**, 359–364.