UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

Descoberta de novos vírus vegetais e estudo da diversidade viral intra-hospedeiro a partir de dados gerados por sequenciamento em larga escala

JOÃO MARCOS FAGUNDES SILVA

Orientador: Dr. Tatsuya Nagata
Co-orientadora: Dra. Rosana Blawid

Brasília, 2018

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

Descoberta de novos vírus vegetais e estudo da diversidade viral intra-hospedeiro a partir de dados gerados por sequenciamento em larga escala

JOÃO MARCOS FAGUNDES SILVA

Orientador: Dr. Tatsuya Nagata
Co-orientadora: Dra. Rosana Blawid

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Biológicas – Biologia Molecular, do Departamento de Biologia Celular, do Instituto de Ciências Biológicas da Universidade de Brasília como parte dos requisitos para obtenção do título de Mestre em Biologia Molecular.

Brasília, 2018

JOÃO MARCOS FAGUNDES SILVA

Descoberta de novos vírus vegetais e estudo da diversidade viral intra-hospedeiro a partir de dados gerados por sequenciamento em larga escala

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Biológicas – Biologia Molecular, do Departamento de Biologia Celular, do Instituto de Ciências Biológicas da Universidade de Brasília como parte dos requisitos para obtenção do título de Mestre em Biologia Molecular.

Banca Examinadora:

Prof. Dr. Tatsuya Nagata (Orientador) (CEL – UnB)

Prof. Dr. Renato Resende (CEL – UnB)

Dr. Marcio Martinello Sanches (Embrapa CENARGEN)

Suplente:

Prof. Dr. Bergmann Morais Ribeiro (CEL – UnB)

“The most exciting phrase to hear in science, the one that heralds new discoveries, is not
‘Eureka!’ (I've found it!), but ‘That’s funny...’”
- *Isaac Asimov*

Agradecimentos

Agradaço a todos os colegas de laboratório, colaboradores, amigos e familiares que me ajudaram durante esse percurso, direta ou indiretamente. Agradeço ao meu orientador Tatsuya Nagata e co-orientadora Rosana Blawid pelos ensinamentos e pela oportunidade de estudo, aos colaboradores Thor Vinicius Martins Fajardo e Maher Al Rwahnih, a todos os colegas de laboratório e todos aqueles que passaram por lá, mesmo que brevemente, e aos colegas dos laboratórios vizinhos. São muitos nomes para citar, e muias pessoas cujo o suporte e companhia foram essencias para o desenvolvimento do meu trabalho de mestrado, e a vocês sou extremamento grato. Agradeço aos órgãos de fomento à pesquisa CAPES e CNPq, e à Embrapa Uva e Vinho pelo financiamento desse trabalho.

INDÍCE

Resumo viii
Abstract ix
Capítulo 1. Introdução 1
1. Aplicação de HTS para estudos metagenômicos em pepinos e videiras 1
1.1 *Zuchinni lethal chlorosis virus* (ZCLV), um importante patógeno de cucurbitáceas. 1
1.2. Indexação e descoberta de vírus em videiras (*Vitis* spp.) 1
1.3. Visão geral do *pipeline* de uma análise metagenômica a partir de dados gerados por HTS 2
2. Aplicação de HTS para estudos da diversidade viral intra-hospedeiro de vírus de RNA 2
2.1. O modelo de *quasispecies* virais 2
2.2. *Quasispecies* virais e HTS 3
3. Objetivos gerais 3
4. Objetivos específicos 4
5. Referência bibliográfica 5
Capítulo 2. Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline 7
1. Abstract 7
2. Introduction 7
3. Steps for discovering new plant viruses and sequencing their genomes by NGS-enriching viral sequences in nucleic acid preparations 9
4. Analysis of sequence quality 10
5. Trimming NGS data 10
6. *De novo* assembly 12
7. Reference-guided and *de novo* mapping algorithms 14
8. Description of a practical pipeline for the discovery of new viruses – a case study for NGS analysis using symptomatic Cucumis sativus plants 15
8.1. Preparation of virus-enriched samples 16
8.2. Trimming NGS data with Trimmomatic 17
8.3. Benchmark for viral genome assemblies 17
8.4. Extending contigs and *de novo* assembly for discovering new viruses 20
8.5. Evaluating the quality of *de novo* assemblies 21
9. Perspectives 22
10. References 25
Capítulo 3. Discovery and molecular characterization of a novel enamovirus, Grapevine enamovirus-1 29
1. Abstract 29
2. Text 29
3. References 35
Capítulo 4. Descoberta e caracterização inicial de um novo vírus *Virga-like* em videiras 37
1. Resumo 37
2. Introdução 37
3. Metodologia 37
3.1. Material vegetal, preparo das bibliotecas de DNA e sequenciamento 37
3.2. Montagem *de novo* e relação taxonômica 38
3.3. Sequenciamento parcial do genoma 38
3.4. Análise filogenética 39

4. Resultados e discussão 39
5. Referências 43
Capítulo 5. Estudo da diversidade intra-hospedeiro dos vírus associados ao enrolamento da folha da videira 44
1. Resumo 44
2. Introdução 44
3. Metodologia 45
3.1. Material vegetal, preparo das bibliotecas de DNA e sequenciamento 45
3.2. Alinhamento e análise de variantes 45
3.3. Estimativa da diversidade global da poliproteína 46
3.4. Estimativa da pressão seletiva (dN/dS) 46
3.5. Detecção *in silico* de moléculas defectivas de RNA (dRNA) 47
4. Resultados e discussão 47
5. Referências 52
Capítulo 6. Discussão geral 54

Resumo

As tecnologias de sequenciamento em larga escala permitem a caracterização genômica das comunidades virais presentes em tecidos vegetais e animais e em amostras ambientais com alta sensibilidade e acurácia. Devido ao sequenciamento simultâneo de várias sequências genômicas, essa técnica também permite o estudo da alta diversidade genética intra-hospedeiro apresentada pelos vírus de RNA. Nesse trabalho, estudamos e estabelecemos um *pipeline* para a análise de viroma em planta utilizando o modelo de pepino, reportamos a descoberta de dois novos vírus em videiras, Grapevine enamovirus-1 (GEV-1) e Grapevine virga-like virus (GVLV). Após ensaios de amplificação rápida das extremidades do cDNA (*rapid amplification of cDNA ends* – RACE) da extremidade 5' do genoma do GEV-1, foi descrito a sequência genômica quase completa desse vírus (6227 bp), possibilitando a sua classificação como um membro do gênero *Enamovirus* (família *Luteoviridae*) com base na sua organização genômica, estudos filogenéticos e critérios estabelecidos pelo Comitê Internacional de Taxonomia de Vírus (*International Committee on Taxonomy of Viruses* – ICTV). Entretanto, o genoma do GVLV permanece parciamente sequenciado em duas partes: um *contig* de 3348 bp que contém os domínios metiltransferase (Met) e helicase (Hel); e um *contig* de 1272 bp que corresponde à RNA polimerase dependente de RNA (RdRp) parcial. Com base em estudos filogenéticos não foi possível classificar esse vírus, que mostra baixa identidade com ambas as famílias *Virgaviridae* e *Bromoviridae*. Adicionalmente, esse trabalho apresenta um estudo da diversidade genética intra-hospedeiro dos vírus associados ao enrolamento da folha da videira (*Grapevine leafroll-associated virus* – GLRaV), com foco na poliproteína dos GLRaV-2 e -3 (gêneros *Closterovirus e Ampelovirus*, respectivamente), assim como a detecção *in silico* de uma molécula defectiva de RNA do GLRaV-4 (*Ampelovirus*), a partir de dados gerados por HTS. As populações intra-hospedeiro encontradas em dois isolados de GLRaV-2 mostraram apenas 11 polimorfismos de único nucleotídeo (*single nucleotide polymorphisms* – SNPs) em comum (~14% dos SNPs em cada isolado). A diversidade intra-hospedeiro encontrada em dois isolados de GLRaV-3 foi baixa se comparada com os isolados de GLRaV-2.

Abstract

High-throughput sequencing technologies allow for the genomic characterization of viral communities present in plant and animal tissues and environmental samples with high accuracy and sensibility. The simultaneous sequencing of various genomic sequences by this technique also makes it useful for the study of the high intrahost genetic diversity presented by RNA viruses. In this work, we studied and established the conditions of analysis of plant virome using the cucumber model, the discovery of two novel grapevine viruses, Grapevine enamovirus-1 (GEV-1) and Grapevine virga-like virus (GVLV). After rapid amplification of cDNA ends (RACE) assays of the 5' end of GEV-1 genome, we obtained the near full genomic sequence of this virus (6227 bp), enabling its classification as a member of the genus *Enamovirus* (family *Luteoviridae*) based on its genomic properties, phylogenetic studies and criteria stablished by the International Committee on Taxonomy of Viruses (ICTV). However, the genome of GVLV remains only partially sequenced, separated in two parts: a 3348 bp contig containing the methyltranferase (Met) and helicase (Hel) domains; and a 1272 bp contig which corresponds to the partial RNA dependent RNA polimerase (RdRp). Based on phylogenetic studies, were not able to classify this novel virus, which shows low identity with viruses in the families *Virgaviridae* and *Bromoviridae*. Additionally, this works presents a study on the intrahost genetic diversity of Grapevine leafroll-associated viruses (GLRaVs), focusing on the polyprotein of GLRaV-2 and -3 (genera *Closterovirus* and *Ampelovirus*, respectively), as well as an *in silico* detection of a defective RNA molecule of GLRaV-4 (*Ampelovirus*). The intrahost population of two isolates of GLRaV-2 showed only 11 single nucleotide polymorphisms (SNPs) in common (~14 of the SNPs found on each isolate). The intrahost genetic diversity found on two isolates of GLRaV-3 was low compared to GLRaV-2.

Capítulo 1. Introdução

1. Aplicação de HTS para estudos metagenômicos em pepinos e videiras

1.1 *Zuchinni lethal chlorosis virus* (ZCLV), um importante patógeno de cucurbitáceas

Infecções por vírus de RNA são responsáveis por uma grande perda na qualidade e na produção de frutos de cucurbitáceas no Brasil. Dentre os vírus mais importantes que infectam o pepino (*Cucumis Sativus*) destaca-se o *Zucchini lethal chlorosis virus* (ZLCV) (Giampan *et al.*, 2009). Esse vírus pertencente ao gênero *Orthotospovirus* (família *Tospoviridae*) e é transmitido pelo tripes *Frankliniella zucchini* (Giampan *et al.*, 2009). Recentemente, a aplicação de sequenciamento em larga escala (*High-Throughput Sequencing* – HTS), também conhecido como sequenciamento de próxima geração (*Next Generation Sequencing* – NGS), em um *pool* de amostras de pepinos coletados de Planaltina, Distrito Federal, revelou um novo isolado de ZCLV nessas plantas (Lima *et al.*, 2016). Os dados gerados nesse trabalho foram utilizados em um caso de estudo para a avaliação de diferentes *softwares* para análises metagenômicas com dados gerados por HTS (Blawid *et al.*, 2016), devido ao fato de que o resultado gerado por esses *softwares*, em especial aos que são utilizados para a montagem *de novo* dos *contigs*, variam consideravelmente, tendo em vista que cada montador é desenhado para trabalhar com um tipo específico de dado.

1.2. Indexação e descoberta de vírus em videiras (*Vitis* spp.)

A videira (*Vitis* spp.) é a planta lenhosa que abriga o maior número descrito de agentes infecciosos intracelulares, com mais de 60 agentes virais descritos que infectam esse gênero (Martelli, 2014). Devido à propagação vegetativa e intercâmbio de material contaminado, esse gênero sofre de uma grave decadência sanitária ao redor do mundo (Martelli, 2014). Para o diagnóstico de material de intercâmbio de videiras são realizados bioensaios por enxertia em plantas indicadoras, ensaios serológicos por ELISA e ensaios moleculares por RT-PCR ou microarranjo (Maliogka *et al.*, 2015). Entretanto, essas técnicas possuem limitações. Ensaios serológicos e moleculares dependem de um conhecimento prévio dos patógenos para a produção de anticorpos ou *primers* específicos (Al Rwahnih *et al.*, 2015). Bioensaios por enxertia podem ser utilizados para diagnóstico de vírus, enquanto podem demorar meses até o aparecimento de sintomas virais em plantas indicadoras, e muitas vezes não é possível inferir o agente etiológico.

Nesse contexto, as tecnologias de HTS tem sido utilizadas como uma poderosa ferramenta para a descoberta de (novos) vírus em videiras (Martelli, 2014), e são vistas como uma alternativa para a indexação de vírus em material de intercâmbio devido à sua alta sensibilidade e acurácia (Al Rwahnih *et al.*, 2015). O primeiro vírus a ser descoberto por essa técnica em videiras, Grapevine Syrah virus 1, foi descrito em 2009 a partir dados gerados pela plataforma 454 (Al Rwahnih *et al.*, 2009). Atualmente, a plataforma Illumina tem sido a mais utilizada para estudos metagenômicos (Breitwieser *et al.*, 2017).

Essa plataforma fornece, a um preço acessível, uma alta quantidade de dados com uma baixa taxa de erro (0,1% ~ 1%) e alta cobertura (Goodwin *et al.*, 2016; Posada-Cespedes *et al.*, 2016).

1.3. Visão geral do *pipeline* de uma análise metagenômica a partir de dados gerados por HTS

Apesar de fornecer uma detecção sensível e acurada dos vírus presentes em amostras de origem vegetal, a identificação de vírus por HTS pode ser desafiadora devido à necessidade de uma infraestrutura computacional e *softwares* específicos. O método utilizado para o preparo da amostra, a plataforma usada para sequenciamento e limitações computacionais são alguns dos fatores que devem ser considerados em toda análise feita a partir de dados gerados por HTS. A extração de dsRNA tem sido um método bastante utilizado para a extração de ácidos nucleicos de tecido vegetal em análises metagenômicas em videiras (Burger e Maree, 2015; Fajardo *et al.*, 2017; Coetzee *et al.*, 2010; Al Rwahnih *et al.*, 2009; Al Rwahnih *et al.*, 2015). Alternativamente, a extração de *small RNA* (sRNA) e extração de RNA de partículas virais semi-purificadas também são comumente utilizados em ensaios metagenômicos (Kreuze *et al.*, 2009; Pantaleo *et al.*, 2010; Blawid *et al.*, 2017; Kutnjak *et al.*, 2015). A cobertura dos genomas virais, tamanho dos *reads*, abundância de polimorfismos de único nucleotídeo (*single nucleotide polymorphisms* – SNPs) e taxa de erro são algumas das características determinadas tanto pelo método de extração de ácidos nucleicos, quanto pela plataforma usada para o sequenciamento (Kutnjak *et al.*, 2015; Goodwin *et al.*, 2016).

Em uma análise metagênomica feita a partir de dados gerados por HTS, a primeira etapa da análise *in silico* consiste na remoção das extremidades de baixa qualidade dos *reads*, assim como a remoção de sequências de adaptadores. Devido ao fato de não se saber inicialmente os vírus presentes na amostra analisada, os *reads* são montados *de novo*, resultando em *contigs* que são classificados a partir de buscas por similaridade contra um banco de dados de sequencias virais conhecidas. A escolha dos *softwares* e dos parâmetros utilizados para essa análise depende das características do *dataset* gerado. Por exemplo, dados gerados através da extração de sRNA possuem *reads* pequenos, de aproximadamente 25 bp, com uma alta incidência de SNPs devido à amplificação do sinal de RNA interferente (RNAi), comprometendo a montagem *de novo* e resultando em *contigs* menores do que aqueles montados a partir de dados gerados pela extração de RNA de partículas virais semi-purificadas (Kutnjak *et al.*, 2015).

2. Aplicação de HTS para estudos da diversidade viral intra-hospedeiro de vírus de RNA

2.1. O modelo de *quasispecies* virais

Além de terem revolucionado o campo da metagenômica, tornando esse tipo de análise cada vez mais acessível à comunidade científica, o advento das novas tecnologias de sequenciamento permitiu uma visualização detalhada da diversidade viral intra-hospedeiro (revisado por Posada-Cespedes *et al.*, 2017). Devido à alta taxa de erro da RNA polimerase, vírus de RNA formam populações heterogêneas de sequências genômicas distintas porém relacionadas dentro de único hospedeiro, onde o modelo de *quasispecies* moleculares introduzido por Eigen (1977) foi adotado pela virologia para descrever tais populações (Holmes e Moya, 2002). Essa mistura heterogênea de genomas distintos (haplótipos) confere aos vírus de RNA uma alta capacidade de rápida adaptação a novos hospedeiros e tipos celulares, além de permitirem evasão da resposta imune (revisado por Domingo *et al.*, 2012).

No modelo de *quasispecies* virais existe um equilíbrio entre as taxas de mutação e seleção natural, onde os haplótipos dominantes com maior aptidão estariam envolvidos em uma nuvem de haplótipos de baixa frequência. A aptidão de um haplótipo é determinada não apenas pela sua taxa de replicação, mas também pela liberdade desse genoma em sofrer mutações que dão origem a sequencias com a mesma taxa de replicação (Lauring e Andino, 2010), e a frequência de cada haplótipo depende também da probabilidade que esse haplótipo surja espontaneamente. Dessa forma, os haplótipos em uma população intra-hospedeiro estariam interligados mutacionalmente (Holmes e Moya, 2002). Embora o termo *quasispecies* seja muitas vezes usado como sinônimo de diversidade intra-hospedeiro, Holmes e Moya argumentam que o uso desse termo dentro da virologia deveria ser utilizado apenas quando há evidências formais para a adequação desse modelo (Holmes e Moya, 2002).

2.2. *Quasispecies* virais e HTS

A utilização de HTS para o estudo de *quasispecies* virais permite a detecção de variantes com prevalência menor do que as taxas de erro de sequenciamento e de preparo de amostra (< 1%), entretanto, tal análise pode ser desafiadora devido aos erros introduzidos durante o preparo da biblioteca de DNA e sequenciamento (Posada-Cespedes *et al.*, 2017). A minimização desses erros sistemáticos pode ser feita durante o próprio preparo das bibliotecas de DNA, a partir de desenhos experimentais que permitam distinguir polimorfismos reais daqueles introduzidos artificialmente durante as etapas de RT e PCR (Kutnjak *et al.*, 2015; Jabara *et al.*, 2011; Acevedo *et al.*, 2014). A acurácia da detecção de variantes é aprimorada *in silico* através de modelos estatísticos para a identificação e remoção de erros (Posada-Cespedes *et al.*, 2017).

Plataformas que geram *reads* longos, como 454 (Roche) e PacBio (Pacific Biosciences) são capazes de sequenciar grandes porções dos genomas virais, facilitando a montagem *in silico* dos haplótipos presentes nessa população, enquanto os pequenos *reads* gerados pela maioria das plataformas Illumina prejudicam a montagem de haplótipos *in silico* em regiões de baixa complexidade (Giallonardo *et al.*, 2014). A análise de variantes pode ser feita a nível de nucleotídeo, onde há apenas a determinação de SNPs; a nível local, onde os haplótipos locais são montados a partir de SNPs que co-ocorrem em regiões menores que o tamanho dos reads (ou do fragmento sequenciado); ou a nível global, através do faseamento de SNPs. Considerando a dificuldade em usar dados gerados pela plataforma Illumina para a montagem de haplótipos globais pelos métodos de preparo de biblioteca convencionais (Giallonardo *et al*., 2014), a análises de variantes a nível de nucleotídeo ou local é preferível para esse tipo de dado.

3. Objetivos gerais

Para estabelecer a metodologia de análise de dados gerados por HTS, foi feito um estudo de caso com o modelo pepino, onde os principais *softwares* utilizados em análises metagenômicas foram testados. Após a determinação dos *softwares* e parâmetros mais adequados para a análise de dados metagenômicos, o modelo de videira foi utilizado no estudo. Devido à propagação vegetativa, videiras possuem patossistemas complexos. A fim de ampliar o conhecimento sobre os vírus presentes em videiras nas regiões sul, sudeste e nordeste do Brasil, temos como objetivo descrever e estudar as complexas comunidades virais presentes em videiras coletadas dessas regiões pela técnica de HTS, com foco na descoberta de novos vírus.

Temos como segundo objetivo estudar a diversidade genética intra-hospedeiro dos vírus associados ao enrolamento da folha da videira 2 e 3 (*Grapevine leafroll-associated virus* – GLRaV-2 e -3), com foco na poliproteína, a partir do desenvolvimento de um *pipeline* focado em dados gerados pela plataforma Illumina.

4. Objetivos específicos

- Determinar as condições de análise de viroma em pepino, com foco nos *softwares* e parâmetros utilizados para a montagem *de novo*
- Sequenciar o genoma ou a região genômica de dois novos vírus encontrados em videiras, Grapevine enamovirus 1 (GEV-1) e Grapevine virga-like virus (GVLV)
- Analisar os possíveis produtos gênicos e elementos regulatórios presentes nos genomas dos GEV-1 e GVLV através de análises comparativas com sequências virais conhecidas
- Classificar os novos vírus encontrados com base em filogenia molecular, características genômicas e critérios taxonômicos estabelecidos pelo Comitê Internacional de Taxonomia de Vírus (ICTV)
- Estudar a diversidade intra-hospedeiro dos vírus associados ao enrolamento da folha da videira-2 e -3 (GLRaV-2 e -3)
- Determinar se SNPs presentes na poliproteína estão sujeitos à seleção neutra ou positiva entre diferentes isolados de GLRaV-2 e -3

5. Referência bibliográfica

Acevedo A., Brodsky L., Andino R. (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, **505**(7485), 686.

Al Rwahnih M., Daubert S., Golino D., Islas C., & Rowhani A. (2015) Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. *Phytopathology*, **105**(6), 758-763.

Al Rwahnih M., Daubert S., Golino D., Rowhani A. (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology*, **387**(2), 395-401.

Breitwieser F.P., Lu J., Salzberg S.L. (2017) A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

Blawid R., Silva J.M.F., Nagata T. (2017) Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Annals of Applied Biology*, **170**(3), 301-314.

Burger J.T., Maree H.J. (2015) Metagenomic next-generation sequencing of viruses infecting grapevines. In *Plant Pathology: Techniques and Protocols*, pp. 315 – 330. Ed. C. Lacomme. New York, NY, USA: Springer.

Coetzee B., Freeborough M.J., Maree H.J., Celton J.M., Rees D.J.G., Burger J.T. (2010) Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology*, **400**(2), 157-163.

Domingo E., Sheldon J., Perales C. (2012) Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews*, **76**(2), 159-216.

Eigen M., Schuster P. (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*. **64**(11), 541-565.

Fajardo T.V.M., Silva F.N., Eiras M., Nickel O. (2017) High-throughput sequencing applied for the identification of viruses infecting grapevines in Brazil and genetic variability analysis. *Tropical Plant Pathology* doi:10.1007/s40858-017-0142-8

Giallonardo F.D., Töpfer A., Rey M., Prabhakaran S., Duport Y., Leemann C., Schmutz S., Campbell N.K., Joos B., Lecca M.R., Patrignani A. (2014) Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic acids research*, **42**(14), e115-e115.

Giampan J.S., Rezende J.A.M., Piedade S.M.D.S. (2009) Yield loss caused by Zucchini lethal chlorosis virus (ZLCV) on zucchini squash'Caserta'. *Summa Phytopathologica*, **35**(3), 223-225.

Goodwin S., McPherson J.D., McCombie W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6), 333.

Holmes E.C., Moya A. (2002) Is the quasispecies concept relevant to RNA viruses? *Journal of Virology,* **76**(1), 460-465.

Jabara C.B., Jones C.D., Roach J., Anderson J.A., Swanstrom R. (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences*, 108(50), 20166-20171.

Kreuze J.F., Perez A., Untiveros M., Quispe D., Fuentes S., Barker I., 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), 1-7.

Kutnjak D., Rupar M., Gutierrez-Aguirre I., Curk T., Kreuze J.F., Ravnikar M. (2015) Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscapes of a plant virus population. *Journal of Virology*, **89**(9), 4760-4769.

Lauring A.S., Andino R. (2010) Quasispecies theory and the behavior of RNA viruses. *PLoS pathogens*, 6(7), e1001005.

Lima R.N., De Oliveira A.S., Leastro M.O., Blawid R., Nagata T., Resende R.O., Melo F.L. (2016) The complete genome of the tospovirus Zucchini lethal chlorosis virus. *Virology journal*, **13**(1), 123.

Maliogka V.I., Martelli G.P., Fuchs M., Katis N.I. (2015) Control of viruses infecting grapevine. Adv *Virus Research*, **91**, 175-227.

Martelli G.P. (2014) Directory of virus and virus-like diseases of the grapevine and their agents. *Journal of Plant Pathology*, **96**(Suppl. 1), 1–136.

Pantaleo V, Saldarelli P, Miozzi L, Giampetruzzi A, Gisel A, Moxon S, Dalmay T, Bisztray G, Burgyan J. (2010) Deep sequencing analysis of viral short RNAs from an infected Pinot Noir grapevine. *Virology*. **408**(1), 49-56.

Posada-Cespedes S., Seifert D., Beerenwinkel N. (2017) Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Research*, **239**, 17-32.

Capítulo 2. Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline

1. Abstract

Small-scale sequencing has improved substantially in recent decades, culminating in the development of next-generation sequencing (NGS) technologies. Modern NGS methods have helped the discovery of many new plant viruses. Nevertheless, there is still a need to establish solid assembly pipelines targeting small genomes characterised by low identities to known viral sequences. Here, we describe and discuss the fundamental steps required for discovering and sequencing new plant viral genomes by NGS. A practical pipeline and standard alternative tools used in NGS analysis are presented.

Keywords

Assembly pipeline, next-generation sequencing, plant viral genome discovery, plant virology.

Este capítulo foi publicado na revista Annals of Applied Biology.
Blawid R., Silva J.M.F., Nagata T. (2017) Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Annals of Applied Biology*, **170**, 301–314

2. Introduction

The first approach to DNA sequencing was developed roughly 40 years ago. Sequences from 84 kbp to 1 Gbp per run were possible by the end of the 1980s and the beginning of the 1990s, as in the Human Genome Projects (by NIH and Celera). The advancement of sequencing methods that produced massively short reads began the era of the 'next generation' in genomic sequencing. The amount of data in a single run is currently in the order of terabases, which has prompted a new way to resolve sequence assembly.

Early in the development of next-generation sequencing (NGS; Fig. 1), entire new viral genomes were determined using 454 sequencing of nucleic acids extracted from diseased plants (see e.g. Roossinck *et al*., 2010). In 2005, the former Solexa (today Illumina) developed a technology based on sequencing by synthesis using reversible dye-terminator chemistry. Khalifa *et al*. (2016) used dsRNA extracts to compare the results from Illumina sequencing with those from Sanger-sequenced clones, finding a high level of identity of 99.3–100%. Many research groups are currently choosing the Illumina platform for discovering new plant viral sequences. For example, Candresse *etal*. (2014) used both Illumina and 454 GS FLX Titanium sequencing technology to effectively detect plant viruses that had escaped routine quarantine viral detection; virus-derived small interfering RNAs (siRNAs) and virion-associated nucleic acids were useful for characterising and identifying two quarantined mastreviruses (Candresse *et al*., 2014). The Illumina sequencing system has also been applied to discover new RNA viruses. Ndunguru *et al*. (2015) used RNA extracts from potyvirus-infected cassava plants to prepare cDNA libraries compatible with the Illumina MiSeq system; *de novo* assembly and reference-guided assembly (Geneious programme, Biomatters, Auckland, New Zealand) were used to generate two genome-wide consensus sequences that matched ipomoviruses. Lastly, sequencing by oligo ligation detection (SOLiDTM) and two-base sequencing (Applied Biosystems, Foster City, CA, USA) have also been used to discover new plant viruses. For example, Sela *et al*. (2013) used the SOLiDTM technology to detect novel viruses; a new partitivirus was detected coinfecting with the *Melon necrotic spot virus* in watermelon plants (Sela *et al*., 2013). Today, the Illumina platform is leading the market of NGS technologies mainly because of the resonable price of their service that can be afforded by small plant, bacterial, viral and fungi diagnostic laboratories.

In this report, we describe the sample preparation and analytical steps required for routine sequencing and present the pipeline that we are currently using, which is cost-effective and useful for the discovery of plant viral sequences in metagenomics NGS data. The same methodology can be used for other types of samples, such as environmental water and human clinical samples, with slight modifications in sample preparation. Although here, the described pipeline is somehow similar to other established pipelines, we provide a case study for benchmarking different assembler programmes for discovering new plant virus genomes. This article has the objective of motivating researchers to explore existing bioinformatics tools and guiding newcomers to NGS through a new field of knowledge.

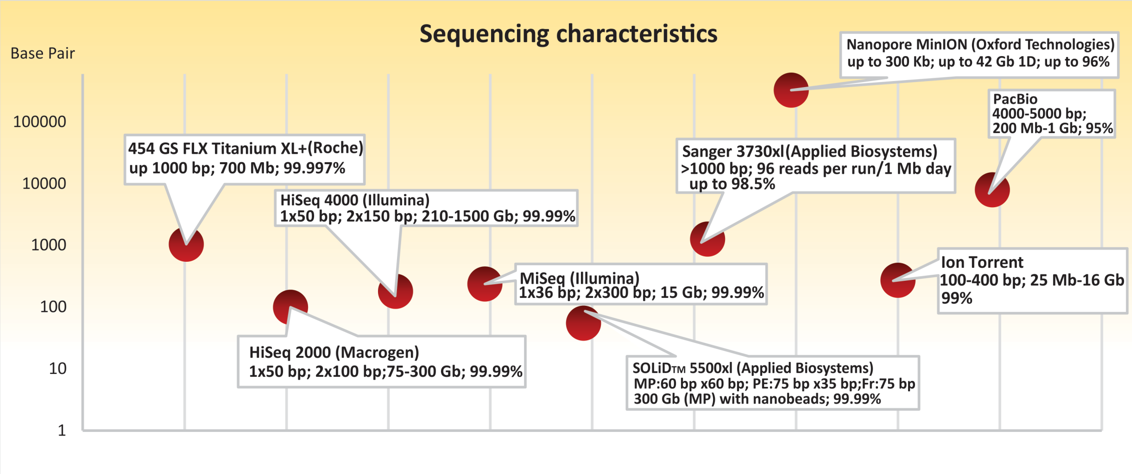


Figure 1 Examples of currently used sequencing platforms for next-generation sequencing and their characteristics. The information inside the boxes is in the following order: read length, typical throughput per run, system accuracy. MP, mate pair; PE, pair end; FR, fragment.

3. Steps for discovering new plant viruses and sequencing their genomes by NGS-enriching viral sequences in nucleic acid preparations

Viral sequences are often the only target for virologists in metagenomics analyses. Enriching viral sequences in preparations of extracted nucleic acids is thus critical. Plant DNA viruses have circular (Geminiviridae and Nanoviridae) or pseudocircular (Caulimoviridae) genomes, and rolling-circle amplification (RCA) (Inoue-Nagata *et al.*, 2004) is often a good choice for sample preparation (Rosario *et al.*, 2013; Idris *et al.*, 2014). The caulimoviral genome contains discontinuities in both strands but can also be amplified by RCA because replicative forms are closed circles and can serve as templates for exponential amplification reactions. RCA reactions can be also performed with satellite DNAs because all reported satellite DNAs are circular. This technique may nevertheless impede the discovery of novel plant DNA viruses with linear genomes.

Several strategies can enrich RNA viral sequences in samples of viral sequences and reduce off-target RNAs. A common approach is the extraction of dsRNA (Al Rwahnih *et al.*, 2009; Adams *et al.*, 2013; Candresse *et al.*, 2013; Blouin *et al.*, 2016), which is a replicative intermediary. The extraction of dsRNA by Valverde *et al.* (1986) or as modified by Gentit *et al.* (2001) used CF-11 cellulose, but the product by Whatman (now a part of GE Healthcare Life Sciences) has been discontinued. An alternative cellulose powder, Sigmacell cellulose (Sigma-Aldrich, S6790), still needs to be evaluated for dsRNA extraction and NGS library construction. DNase I treatment is usually recommended because of dsDNA contamination in dsRNA extractions. Another alternative to enrich dsRNA from plant extracts is to use anti-dsRNA monoclonal antibodies in pull-down experiments. Blouin *et al.* (2016) developed a novel protocol using the anti-dsRNA mAb 2G4 (O'Brien *et al.*, 2015) in pull-down assays in order to identify and discover new viruses from the infected cultivated crops.

siRNAs are abundantly produced by the RNAi machinery of plants infected by RNA viruses. Both RNA and DNA viral infections are associated with the production of siRNAs (reviewed by Wu *et al.*, 2015). In contrast to dsRNA preparation, a number of commercial kits are available for the extraction of small RNAs. These strategies are useful for the study of plant viromes because all viral and subviral agents undergo siRNA biogenesis. For instance, Kutnjak *et al.* (2014) demonstrated the benefit of the sequencing approach with small RNAs to distinguish strains of Potato virus X isolates in mixed infections. Later, Kutnjak *et al.* (2015) compared two viral sequence enrichment methods (enriching virus-derived small RNAs and viral particles) to assess the efficiency of virus consensus genome reconstruction. The authors concluded that the small RNA sequencing is more time-efficient and a more generic approach. On the other hand, they stressed that sequencing RNA extracted from virus particle preparation not only enhanced the efficiency of sequence reconstruction but also allowed the detection of non-homologous recombination hot spots in viral genomes.

Several virus-like particle (VLP) purification procedures have been used not only for viromic studies of living cells but also for the preparation of environmental samples such as surface waters and soils (Rosario *et al.*, 2009; Kleiner *et al.*, 2015; Reavy *et al.*, 2015). We are currently using an unselective procedure of semi-purification of plant viral and VLPs by sequential centrifugation without a prior filtration step combined with rRNA depletion for metagenomics studies (unpublished data). The use of a tandem tangential flow filtration system to separate bacteria from viral-sized particles followed by an ultracentrifugation step to enrich samples with viral particles has been demonstrated by Djikeng *et al*. (2009). Although some authors state that non-encapsidated agents and unstable particles cannot be isolated by the partial purification of VLPs (Kreuze *et al*., 2009), we have been able to identify endornaviruses (non-encapsidated) from RNA extracts using the plant VLP semi-purification protocol described in this study.

4. Analysis of sequence quality

The quality control of sequence data provides a quick impression on the possible problems that might have occurred during sequencing procedures. The quality assessment output helps in choosing the correct preprocessing parameters because low-base call quality scores can, for instance, negatively impact assembling and mapping. The Phred quality score (*Q*) was developed for Sanger sequencing, but the same idea of quality control is used by NGS approaches. *Q* scores in NGS reflect the base-calling error probabilities. Illumina sequencers typically produce reads with poor quality at the end of the reads. Using sequenced reads with low *Q* scores (lower than 20) for downstream approaches may decrease assembly quality and waste time. The FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) tool provides quality analysis for NGS data. FASTQC runs on the Java platform and is therefore suitable for all major operating systems. This tool is distributed by the Bio-Linux platform (http://environmentalomics.org/bio-linux/) and is also available on Galaxy web interfaces (https://usegalaxy.org/). FASTQ analysis provides important information for subsequent trimming. For example, the ‘per base N content’ result of FASTQC contains significant proportions of N bases (reads without a confident base-call quality), indicating a biased sequence composition. The region indicated by this result might be eliminated from reads in a later trimming step. The ‘per base sequence quality’ result of FASTQC indicates statistically lower base quality, which is indicative of the threshold of the reading for quality-based trimming in a later step. FASTQC provides insight to the sequence quality of a library and helps the choice of better trimming parameters for avoiding the loss of sequence information.

5. Trimming NGS data

Several tools are currently available for trimming single- and/or paired-end read data (Table 1). Jiang *et al*. (2014) compared several trimming tools on both simulated and real sRNA data, paired-end reads and Nextera LMP sequencing data. Jiang *et al*. (2014) also defined three parameters for the performance of 15 trimming methods: *sensitivity* (Sen, the ratio of correctly trimmed reads to the number of contaminant reads), *specificity* (Spec, the ratio of the number of untrimmed non-contaminant reads to the number of non-contaminant reads) and *positive predictive value* (PPV, the ratio of the number of correctly trimmed reads to the number of trimmed reads). *SeqTrim*, *TagCleaner*, *SeqPrep* and *AdapterRemoval* have the slowest processing speeds (Mbp/s) (Jiang *et al*., 2014); *TagCleaner*, *Trimmomatic* and *Skewer* have the highest PPV and Spec, and *Cutadapt* has a high Sen. Some of these tools were developed specifically either for single- or paired-end data and others for both.

Most of the currently available trimming methods include adapter and/or quality trimming. Some can separate multiplexed reads based on barcodes and include a package with different Illumina adaptors, as with *Trimmomatic*. The main innovation of the *Trimmomatic* algorithm is its ability to remove known adapter sequences originating from Illumina sequencing and to perform quality trimming. *Trimmomatic* removes adapters and low-quality N bases and offers two approaches for quality filtering: a sliding window and a maximum information quality filtering. The trimming thus starts by scanning from the 5′ end of the reads and clipping when the average quality per base drops below a given quality. Over-trimming often causes the loss of information, whereas retaining contaminants (e.g. primers, adapters or barcode indices) and low-quality base reads can interfere with downstream sequencing analyses, for example, mapping and the *de novo* assembly of sequences. Preprocessed paired-end data should not impair downstream analysis because some assemblers, for example, SPAdes (St. Petersburg genome assembler, Bankevich *et al*., 2012), use the positional relationship between pairs for further data processing. Quality-based trimming may increase the mappability of reads but can also reduce the absolute number of aligned reads and may consequently lead to the loss of information. If low-score parameters allow too many base-calling errors that might interfere with downstream processing, trimming with high scores may waste sequencing information. We suggest that trimming should be used with caution to avoid, for example, false expression estimates for sequencing data (Williams *et al*., 2016). Taken together, trimming is an important process in order to reconstruct whole viral genomic consensus sequences but should be used carefully to avoid the loss of genetic information. After trimming procedures, it is always appropriate to verify sequence quality using, for example, the above described FASTQC tool.

Table 1 Currently available software for trimming sequences from NGS data

Trimmer	Repository (2016)	Requirements	Characteristics
Skewer	https://github.com/relipmoc/skewer	Linux	
FastX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/	Linux, OpenSolaris, FreeBSD, MacOS X	Several tools available (e.g. Fastq information, collapser, renamer)
SeqTrim	https://github.com/dariogf/SeqtrimNext	Ruby 1.9.2 or greater CD-HIT 4.5.3 or greater Blast plus 2.24 or greater	Especially suited for Illumina and 454 data
TagCleaner	http://tagcleaner.sourceforge.net/manual.html	Linux/Unix, MacOS Perl 5 (higher)	Also available as web based

SeqPrep	https://github.com/jstjohn/SeqPrep	Linux/Unix	Merge paired-end Illumina data
Flexbar	https://github.com/seqan/flexbar	Linux, MacOS TBB library 4.0 or later SeqAn library v.2.1.1	
Scythe	https://github.com/vsbuffalo/scythe	GCC or Clang compiler, Zlib, Fastq	Uses Bayesian approach to classify contaminants
TrimGalore	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/	Cutadapt, FastQC (optional), Perl	Remove biased methylation position for Reduced Representation Bisufite-Seq(RRBS) type libraries
Cutadapt	https://github.com/marcelm/cutadapt/	Python, PYPI, C compiler	Interesting filtering outputs
AdapterRemoval	https://github.com/MikkelSchubert/adapterremoval	Zlib, bzlib2	
AlienTrimmer	ftp://ftp.pasteur.fr/pub/gensoft/projects/AlienTrimmer		
NextClip	https://github.com/richardmleggett/nextclip	GCC compiler	

6. *De novo* assembly

The *de novo* assembly of contigs (assembling sequence reads as contigs without previously known genomic sequences) is the key process in NGS workflows for discovering new viral sequences. Selected assemblers and parameter settings can influence contig characteristics, which can interfere with *de novo* (also in guided reference mapping) genome assembly. Algorithms for *de novo* assembly can be divided into two main categories: overlap layout consensus (OLC) and algorithms based on *de Bruijn* graphs (Vázques-Castellanos *et al*., 2014). Both classes are based on a graph, a structure containing nodes connected to each other by edges, representing the reads and their overlaps. Miller *et al*. (2010), however, revised NGS assembler algorithms and divided them into three categories: OLC, *de Bruijn* and Greedy graph algorithms. The Greedy graph algorithms can use either OLC or *de Bruijn* graphs, although contigs in the Greedy approach are built by Greedy extension in which only high-scored edges are considered.

Algorithms based on *de Bruijn* graphs do not rely on pair-wise alignments to detect overlaps. *De Bruijn* graph algorithms are therefore considered to be more computationally efficient, a major advantage for solving assembly problems. These algorithms hash the reads into substrings of a fixed length *k*, designated *k*-mers (see Compeau *et al*., 2011), that represent the nodes in the graph. Two nodes are linked with an edge if they share a *k*-1-mer. Overlaps between the reads are then represented as shared *k*-mers. The algorithms based on *de Bruijn* graphs include, amongst others, the assemblers (Table 2) Velvet, SOAPdenovo2 (Luo *et al*., 2012), ABySS (Simpson *et al*., 2009), SPAdes, IDBA-UD (Peng *et al*., 2012) and MEGAHIT (Li *et al*., 2015), and the OLC assemblers include programmes such as the Celera Assembler revised pipeline CABOG (Miller *et al*., 2008), SSAKE (Warren *et al*., 2006) and VCAKE (Jeck *et al*., 2007).

Several assemblers currently used for metagenomics short-read data can be compiled for web-interface platforms or be installed on different operating systems. Lai *et al*. (2015) compared five assemblers [ABySS, IDBA-UD, CABOG, MetaVelvet (Namiki *et al*., 2012) and SOAPdenovo (Li *et al*., 2009)] to evaluate their performance on metagenomic data sets. Several parameters were measured, including cover rate, assembly errors and contig/genome length, each versus coverage. SOAPdenovo and MetaVelvet generated fewer errors in all ranges of coverage studied but produced shorter contigs (Li *et al*., 2009). In addition to coverage, GC content may affect assembly performance across the sequence, so Lin *et al*. (2011) generated a table of recommendations for selecting appropriate *de novo* tools depending on the read property of the GC content, single- or paired-end reads and read length for small and large genomes.

Several pipelines for bioinformatics assembly for the discovery of viral genomes have been developed to facilitate data processing (Lorenzi, 2013; Wang *et al*., 2013; Ho & Tzanetakis, 2014; Jayasundara *et al*., 2014; Burger & Maree, 2015; Wan *et al*., 2015; Nakamura *et al*., 2016; Yamashita *et al*., 2016). Wan *et al*. (2015) developed a pipeline named VirAmp that combines different tools in a user-friendly Galaxy web interface for viral genome assembly. Two *de Bruijn* graphic algorithms are available in VirAmp (www.viramp.com), Velvet (Zerbino & Birney, 2008) and SPAdes (St. Petersburg genome assembler, Bankevich *et al*., 2012), and an OLC algorithm-based *de novo* method, VICUNA (Yang *et al*., 2012), is also available. VirAmp tools also allow the use of a reference genome for guidance and have their own tool for further scaffolding and contig extension. VirusTap, another user-friendly assembly pipeline for viral genomes, was released in 2016 (Yamashita *et al*., 2016, https://gph.niid.go.jp/cgi-bin/virustap/index.cgi). An additional trimming filter for host and bacterial contamination were offered by VirusTap for read subtraction to facilitate effective *de novo* assemblies, which are especially important when analysing samples highly contaminated by phage and rRNA, such as clinical stool samples. An advantage of using such web interfaces is the availability of a combination of assembly methods that follow very clear protocols. Not all configuration sets, however, are offered in web interfaces, and in many cases, users might be interested in additional optional sets. Such web interfaces are nevertheless important to those not familiar with command lines and without access to a computer facility.

Table 2 The most common *de Bruijn* assemblers used for discovering plant viral genomes

Assembler	Characteristics	Website	Availability (2016)
ABySS	Improved scalability for assemblies of long sequencing reads	http://www.bcgsc.ca/platform/bioinfo/software/abyss	https://github.com/bcgsc/abyss
IDBA-UD	Designed for uneven-coverage data; iterative assembly with multiple *k*-mer values; removes	http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/	https://github.com/loneknightpy/idba

	short low-depth nodes based on neighbouring nodes		
MEGAHIT	Achieves low-memory assembly; very fast; suitable for assembly of metagenomic data; Iterative assembly with multiple values of *k*	Not available	https://github.com/voutcn/megahit
MetaVelvet	Uses length-weighted coverage for multiple species assembly	http://metavelvet.dna.bio.keio.ac.jp/	http://metavelvet.dna.bio.keio.ac.jp/MV.html#download
SOAPdenovo2	Supports max *k*-mer values of 63/127; focused on single-genome assembly	http://soap.genomics.org.cn/soapdenovo.html	https://sourceforge.net/projects/soapdenovo2/files/SOAPdenovo2/
SPAdes	Designed for uneven-coverage data; also supports Illumina long paired-end reads (2x150 and 2x250) and long reads from PacBio (hybrid assemblies); includes an error correction; iterative assembly with multiple values of *k*	http://bioinf.spbau.ru/spades	http://bioinf.spbau.ru/spades
StriDe	Identify and decompose reads into sub-reads only in error prone prone regions; extend paired-end reads to create artificial long reads; creates an assembly graph from the corrected sub-reads, extended artificial reads and original reads	Not available	https://github.com/ythuang0522/StriDe
Velvet	Developed specifically for short reads; removes nodes below a coverage/length threshold	https://www.ebi.ac.uk/~zerbino/velvet/	https://github.com/dzerbino/velvet/

7. Reference-guided and *de novo* mapping algorithms

In addition to *de novo* assembly, reference-guided assembly can be used when the objective is to detect genomic variants of known viral genomes from a library. Generated short contigs and merged/interleaved reads can thus be mapped to a reference genomic sequence.

In addition to using commercially available programmes, such as Geneious, DNASTAR's Lasergene Genomics Suite and CLC Workbench, alternatives are available for obtaining complete genomes by mapping reads against a target reference genome or to a constructed database. Bowtie (Langmead *etal.*, 2009), Bowtie2 (Langmead & Salzberg, 2012) and BWA (Li & Durbin, 2009) are examples of programmes used in reference-guided approaches. Bowtie2 can map reads against a reference genome supporting gapped, ambiguous characters and local and paired-end alignments, in contrast to Bowtie that can only align reads end to end, without insertions or deletions.

BWA also maps low-divergent sequences against a reference genome using the FM (Ferragina and Manzini) index for indexing and a backtracking method for inexact matches. Some studies have compared the alignment performances of Bowtie and BWA (Ruffalo *et al.*, 2011; Medina-Medina *et al.*, 2012; Yu *et al.*, 2012). A general concern is that Bowtie is usually faster than BWA, but BWA consistently aligns more reads than Bowtie without affecting read alignment quality, even when more errors are allowed. Therefore, it is generally recommended that if sensitivity is required, BWA should be used, especially because it offers a greater number of parameters including, for instance, maximum number of gaps or deletions.

The coverage quality of reference-guided assembled data may be measured mainly by the number of gaps, contig sizes and depths. Gaps might be filled by short insert sequences at higher depths and/or larger inserts that provide the ability to solve simple sequence repeats, which are relatively frequent in the genomes of plant viroids (Qin *et al.*, 2014). A mixture of short and long inserts in the library may, in these cases, therefore produce a final longer consensus sequence with higher accuracy.

Mapping can also be used to purge unwanted host sequences from a library. Several commercial software platforms also have their own mapping algorithms that can subtract reads. Subtracting host and/or bacterial sequences by mapping the reads against a reference genome can computationally simplify assembly. Filtering host/contaminated sequences may produce longer contigs, depending on the assembler used. Host-read subtraction is commonly used to detect human pathogens by metagenomics (Naccache *et al.*, 2014; Greninger *et al.*, 2015) and is also part of the VirusTAP pipeline (Yamashita *et al.*, 2016). As previously mentioned, this subtraction step must be performed with caution to avoid jeopardising genuine viral reads. Lastly, the mapping of reads to a piled-up genome (the final consensus sequence) as a reference allows the estimation of sequence diversity within a population. Such analyses help to shed new light on many questions of viral evolution and phylogeography (Beerenwinkel *et al.*, 2012; Kortenhoeven *et al.*, 2015).

8. Description of a practical pipeline for the discovery of new viruses – a case study for NGS analysis using symptomatic Cucumis sativus plants

We are using the following practical pipeline quite successfully to discover new viral genomic sequences using the Illumina platform and paired-end libraries from samples enriched for viral sequences (Fig. 2). Although the pipeline presented here is similar to most currently established approaches, we additionally encourage researchers to use different *de novo* assemblers in order to explore their NGS data to the maximum. Here, we present a case study for NGS analysis using *Cucumis sativus* plants with virus-like symptoms collected in commercial fields in Brasília, Brazil.

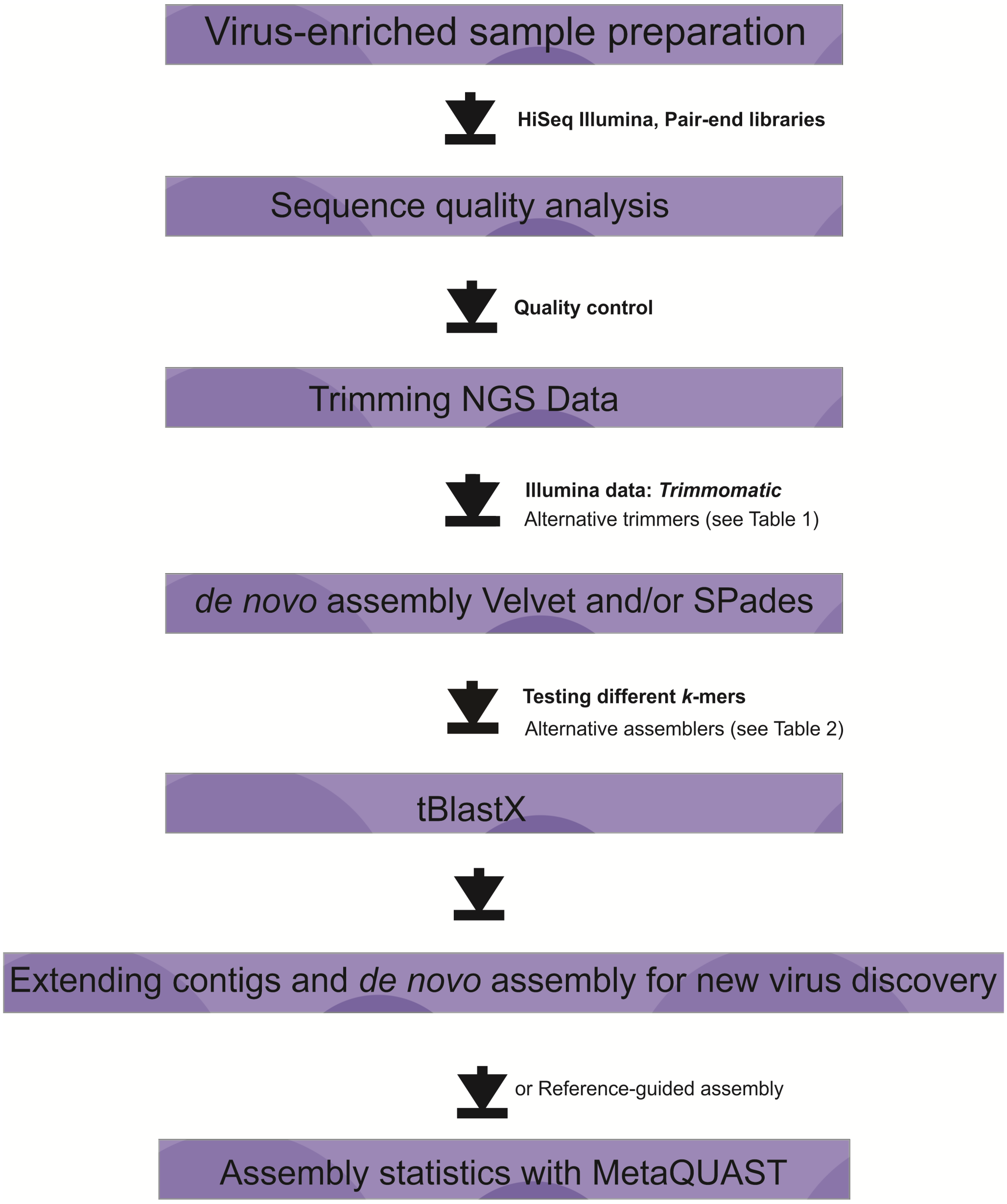

Virus-enriched sample preparation
HiSeq Illumina, Pair-end libraries
Sequence quality analysis
Quality control
Trimming NGS Data
Illumina data: *Trimmomatic*
Alternative trimmers (see Table 1)
de novo assembly Velvet and/or SPades
Testing different *k*-mers
Alternative assemblers (see Table 2)
tBlastX
Extending contigs and *de novo* assembly for new virus discovery
or Reference-guided assembly
Assembly statistics with MetaQUAST

Figure 2 Pipeline used for discovering the genomes of new plant viruses.

8.1. Preparation of virus-enriched samples

The following protocol was used for preparing virus-enriched samples. A total of 20 individual plant samples showing viral-like symptoms were homogenised, percolated through a fine mesh, mixed with Triton X-100 (final concentration of 1%) and centrifuged at low speed (5000 *g*) for 20 min. Then, the supernatant was ultracentrifuged at 140 000 *g* for 1 h. The RNA pool was extracted from the pellet with the ZR Plant RNA MiniPrep kit (Zymo Research, Irvine, CA, USA). We have used this procedure quite successfully for detecting RNA and DNA viral sequences (the kit is not totally selective for RNA unless the sample is treated with DNase), transposons and subviral agents. An rRNA depletion step using the Plant Leaf RiboZero kit (Illumina) was performed to prevent masking the viral sequences prior to library construction using the TruSeq kit for RNA library preparation (Illumina). The library was processed by Illumina HiSeq 2000 machines generating 2×100bp paired-end data (4–5G scale sets).

8.2. Trimming NGS data with Trimmomatic

Amongst the options shown in Table 1, we chose *Trimmomatic* (Bolger *et al.*, 2014) (http://usadellab.org/cms/) for trimming the paired-end Illumina HiSeq2000 data (100 bp paired-end). *Trimmomatic* runs on a Java platform, so it can be used on Linux, Mac or Windows operating systems. This tool can also be found on Galaxy web interfaces (https://usegalaxy.org/). We have achieved the best results with both adapter and quality trimming. FASTQC tools help to detect inherent errors and artefacts in sequencing outputs and show the average *Q* scores across the sequence. We have often used a sliding window of 4 : 20 (window size:required quality) for trimming our NGS data. Quality trimming was performed when the average quality within the window fell below the Phred score of 20. We used the following parameters for our case study analysis: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:20 and MINLEN:36. After performing the specified processing steps, two 'paired' output files were obtained, each containing 10 323 599 reads. The total 'unpaired' reads were 2 223 323.

8.3. Benchmark for viral genome assemblies

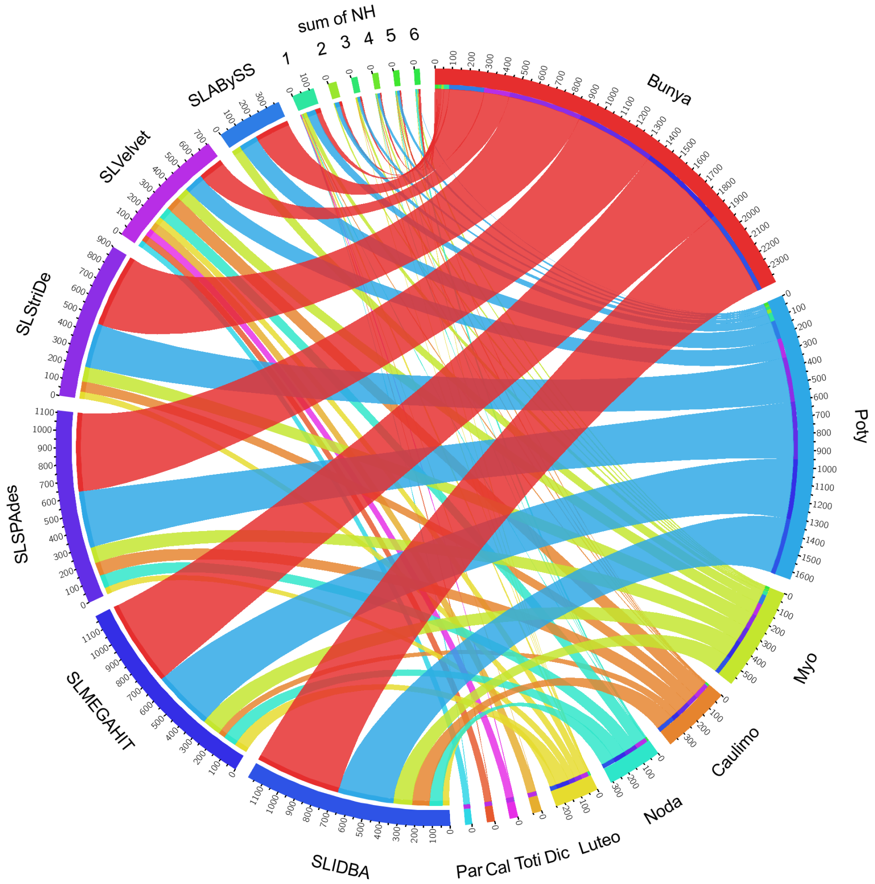
We used six *de Bruijn* assemblers, ABySS, IDBA-UD, MEGAHIT, SPAdes, StriDe and Velvet, for genome assembly. Velvet runs on both Windows (with 36 GB RAM) in the Linux-like emulator, Cygwin and Bio-linux (http://environmentalomics.org/bio-linux/) (with 16 GB RAM) under Ubuntu OS. SPAdes runs on a 64-bit Linux system or Mac OS and requires the installation of Python (http://pypi.python.org/pypi/SPADE). SPAdes and Velvet assemblers are also part of web-interface pipelines, such as the VirAmp pipeline (http://www.viramp.com/). All other assemblers were compiled according to recommendations for the Linux platform. Many of these assemblers can be run in web-interface pipelines, but we found, for instance, that manually running Velvet without expected coverage and coverage *cut-off* parameters produced scaffolds containing more low-coverage sequences. We were therefore able to discover new viral genomic segments that were present in our samples with extremely low coverage. We have used *k*-mers ranging from 21 to 99 for benchmarking several assemblers such as ABySS, IDBA-UD, MEGAHIT, SPAdes, StriDe and Velvet (Table 3). Velvet generated the most diverse scaffolds, while SPAdes generated the longest contig, StriDe the highest overall number of contigs with a list of 253 415 and IDBA-UD the lowest number with only 663 contigs.

Table 3 Benchmarking six different *de novo* assemblers for discovering novel virus-like sequences

Assembler	Max contig length (nt)	Min contig length (nt)	Number of Contigs	Total length (nt)	N50	L50
ABySS	3727	55	47 428	4.256488	109	18 620
IDBA-UD	4902	215	663	375 192	591	175
MEGAHIT	5205	202	690	406 694	623	188
SPAdes	5385	78	5018	1.116126	255	1074
StriDe	4782	84	253 415	27.191145	101	118 197
Velvet 21	2843	41	54 525	6.033835	130	18 390
Velvet 33	1571	65	41 831	5.218958	132	16 006
Velvet 55	3747	109	2987	557 907	191	924
Velvet 77	2345	153	1305	247 259	153	497

Employed *k*-mers: ABySS (55); IDBA-UD (21,41,61,77); MEGAHIT (21,41,61,81,99); SPAdes (21,33,55,77); StriDe (55); Velvet (21,33,55,77). Max, maximum; Min, minimum; Total length, the total number of bases in the assembly; N50, the length for which the collection of all contigs of that length or longer covers at least half the assembly; L50, the number of contigs equal to or longer than N50.

Once the scaffolds/contigs were generated, we used the commercial programme Geneious R8 (Biomatters) for tBlastX searches. The generated contigs were imported into Geneious, and tBlastX searches (E-value 1e-10, Matrix BLOSUM62, Word Size 3) were performed using the last updated virus RefSeq records from GenBank (http://www.ncbi.nlm.nih.gov/genome/viruses/). To discover false-positive sequences, Blastn searches were performed with those contigs showing low amino acid identity (less than 55%) to viral subject sequences. Fig. 3 shows the cumulative sequence length mean (length mean of the input query that matches viral subject sequences in amino acid) for the different viral families per assembler (ABySS, IDBA-UD, MEGAHIT, SPAdes, StriDe and Velvet) after tBlastX searches. For instance, MEGAHIT gave the highest cumulative sequence mean length (SL) and ABySS the least cumulative SL, indicating that the MEGAHIT assembler generated longer single contigs for each viral sequence belonging to a particular family. It is our experience that often, MEGAHIT, IDBA-UD and SPAdes tools give the longest sequence lengths. From Fig. 3, it is also possible to observe the cumulative SL per viral family. For instance, the cumulative SL value found for the *Bunyaviridae* family was generated mainly from the sum of the SL produced from the IDBA-UD, MEGAHIT, SPAdes and StriDe assemblers. Although the Velvet assembler does not produce long sequence lengths, a higher number of viral tBlastx hits were found for each viral family and across a range of diverse families (see Fig. 3). The average amino acid identity of query sequences that matches sequences from a certain viral family is shown in Table 4.



Families	NHABySS	NHIDBA	NHMEGAHIT	NHSPAdes	NHStriDe	NHVelvet	SLABySS	SLIDBA	SLMEGAHIT	SLSPAdes	SLStriDe	SLVelvet	Sum of SL per Family
Bunyaviridae	17	8	8	12	9	32	208	516	517	474	441	152	2308
Caliciviridae	0	0	0	0	0	1	0	0	0	0	0	42	42
Caulimoviridae	0	3	4	1	3	4	0	106	43	83	63	89	384
Dicistroviridae	0	0	0	0	0	1	0	0	0	0	0	50	50
Luteoviridae	0	1	1	2	1	8	0	40	78	40	40	41	239
Myoviridae	6	6	6	6	8	22	57	113	112	94	90	72	538
Nodaviridae	0	1	1	8	0	2	0	77	77	77	0	69	300
Parvoviridae	0	0	0	0	0	1	0	0	0	0	0	36	36
Potyviridae	25	14	13	1	14	51	111	335	358	342	266	114	1526
Totiviridae	0	0	0	0	0	3	0	0	0	0	0	44	44
Sum of SL and NH	48	33	33	30	35	125	376	1187	1185	1110	900	709	

Figure 3 Circular plot of the represented table. The outermost rings on the left side represent the cumulative sequence mean length (SL) in numbers per assembler (ABySS, IDBA-UD, MEGAHIT, SPAdes, StriDe and Velvet), that is the cumulative sequence mean length of the input query (contig) that aligns to the viral subject sequence. The outermost rings on the right side represent the sum of the SL per viral family plus the sum of the number of tBlastX hits (NH). Sum of number of tBlastX hits (NH) with: 1 = Velvet; 2 = ABySS; 3 = StriDe; 4 = MEGAHIT; 5 = IDBA-UD; 6 = SPAdes; Segment colours represents viral families: Bunya, Bunyaviridae (red); Cal, Caliciviridae (dark brown); Caulimo, Caulimoviridae (light brown); Dic, Dicistroviridae (light brown); Luteo, Luteoviridae (yellow); Myo, Myoviridae (light green); Noda, Nodaviridae (turquoise); Par, Parvoviridae (light-blue); Poty, Potyviridae (blue); Toti, Totiviridae (pink).

Table 4 Number of contigs hits to viral sequences and the average of amino acid identities using tBlastX

Family	ABySS	IDBA-UD	MEGAHIT	SPAdes	StriDe	Velvet
	Number of Hits/ Average of Amino Acid Identities (%)					
Bunyaviridae	19/89	8/83	8/83	12/83	9/83	32/84
Caliciviridae	0	0	0	0	0	1/95
Caulimoviridae	0	3/43	4/43	1/45	3/50	4/47
Dicistroviridae	0	0	0	0	0	1/58
Luteoviridae	0	1/77	1/77	2/77	1/77	8/87
Myoviridae	6/76	6/72	6/72	6/79	8/76	22/75
Nodaviridae	0	1/39	1/39	8/39	0	2/45
Parvoviridae	0	0	0	0	0	1/75
Potyviridae	25/93	14/96	13/95	1/90	14/95	51/94
Totiviridae	0	0	0	0	0	3/68

Employed *k*-mers: ABySS (55); IDBA-UD (21,41,61,77); MEGAHIT (21,41,61,81,99); SPAdes (21,33,55,77); StriDe (55); Velvet (21,33,55,77).

Based on our experience, we recommend that users explore their data using different assemblers (especially Velvet and SPAdes) and assembler parameters, especially when other information, for example, parsimony and number of error-free bases, is of interest in addition to obtaining new viral genomic sequences.

Commercial packages such as DNASTAR's Lasergene Genomics Suite and CLC Workbench contain automated programmes for contig assembly. The use of commercial software is an option if you are not sure about the choice of *de novo* contig – assembly software and parameters, although in our opinion, the data can be better explored by manually assembling the parameters. In addition, the automated *de novo* assembler of Geneious requires a lot of memory and so is not a practical choice for the *de novo* assembly of NGS data sets.

8.4. Extending contigs and *de novo* assembly for discovering new viruses

We used the following procedure with Geneious R8 to obtain contigs representing full (or longer) viral genomic sequences. The contigs were analysed by tBlastX searches as described in step 3 and were then organised by their identity values. The contigs with identities similar to the same virus (or a related virus in the same genus) were clustered together in the same folder. Each contig was extended by mapping (Geneious 'Map to reference' command with medium–low/fast sensitivity) the trimmed merged reads (also imported into Geneious after trimming) to the chosen contig as a reference. Unassembled reads were mapped to the reference, extending the contig length until all reads of the library were mapped (you may repeat this step until the contig length stops increasing). Finally, the extended sequences can be reassembled into a unique contig consensus using the '*de novo* assembly' command of Geneious (medium–low/fast sensitivity, fine tuning of iteration up to five times). We have therefore successfully obtained full viral genomes in many cases (if not longer contigs) from short reads generated by Illumina using the mapping and *de novo* assembler tools of Geneious. Note that the Geneious read mapper requires a lot of memory when the parameters are set for high sensitivity and that the high sensitivity may lead to artificial chimeric contigs. Geneious must be carefully used to prevent the extension of artificial chimeric sequences by properly stopping the extension.

Using this method, the complete sequence of the S, M and L segments of the Zucchini lethal chlorosis virus (ZLCV, accession numbers KU681010-KU681012), a tospovirus, was determined and recently published by Lima *et al.* (2016).

8.5. Evaluating the quality of *de novo* assemblies

With a variety of pipelines, tools and even parameters of tools, many assembly outcomes are possible, and the improvement of the analysis is of great interest. NGS metrics might be useful for assessing the assembly quality, although it is not a straightforward task. The QUAST (Gurevich *et al.*, 2013) and MetaQUAST (Mikheenko *et al.*, 2015) programmes are options to assess and compare the quality of our assemblies. Contig size is one of the metrics that can be evaluated for *de novo* assembly with or without a reference genome. As already stated, SPAdes usually generated longer contigs than Velvet (Table 3), although the longest contigs are not always viral sequences. The N50 and L50 statistics can be calculated directly from the contig list output, without the aid of a reference sequence. N50 measures the length of a set of sequences, in which the sum of the lengths of the longest sequences equals at least 50% of the total assembly size. L50 is the number of contigs that are longer than or equal to the N50 value, reflecting the minimum number of contigs that covers at least half of the total assembly size. Assemblies always have false positives, so N50 and L50 alone do not necessarily assess the assembly quality and can sometimes even be misleading. Table 3 shows the N50 and L50 values obtained for ABySS, IDBA-UD, MEGAHIT, SPAdes, StriDe and Velvet using *k*-mers varying from 21 to 99. We also used the S, M and L sequences of ZLCV to compare assembly metrics based on a reference genome.

Another useful metric in genome assembly is the genome fraction (GA%), which is the total number of bases aligned to the reference sequence divided by the genome size. In other words, it calculates the percentage of a reference genome that is covered by contigs. Both Velvet and SPAdes assemblers generated high GA percentages (Table 5), even for nodes with middle to low coverages, meaning that they could assemble almost entire genomes. This information alone, however, does not give any information regarding the contiguity of the contigs. The NG50 can be calculated when a reference genome is provided; it measures the minimum contig length from which all contigs of that length or longer cover half of the reference genome. MetaQUAST also generates other valuable metrics as the NA50 and NGA50. These metrics are obtained by breaking the contigs mapped to a reference into aligned blocks and calculating the N50 and NG50 using only these blocks. SPAdes also often gives the highest value of NG50 and NGA50, as shown for our case study with all segments of ZLCV (Table 5).

MetaQUAST supports more than one reference sequence and calculates the reference-based statistics separately for each sequence. QUAST can also use the GAGE

(Salzberg *et al*., 2012) method for genome assessment in the GAGE mode to calculate, amongst other metrics, the E size, which measures the assembly contiguity based on a reference genome. None of these metrics for assessing assembly quality is perfect, and errors in a newly assembled genome are often very hard to detect.

Table 5 Evaluation of assemblers employing MetaQuast using the L, M and S segments of Zucchini lethal chlorosis virus (ZLCV) as reference genomes

Reference genomes	**ABySS**	**IDBA-UD**	**MEGAHIT**	**SPAdes**	**StriDe**	**Velvet 21**	**Velvet 33**	**Velvet 55**	**Velvet 77**
S GA%	99,6	99,8	99,8	100	99,6	85,5	89,9	97,6	98
S NG50	2,200	3,502	3,499	3,507	3,494	282	884	472	622
S NGA75	1,314	3,502	3,499	3,507	3,494	137	162	421	620
M GA%	99,6	100	100	100	99,6	98,5	99,5	100	95,3
M NG50	3,727	4,798	4,798	4,801	4,782	1,067	1,083	3,747	1,221
M NGA75	3,727	4,798	4,798	4,798	4,782	319	761	3,747	1,008
L GA%	55,7	92,9	93,2	97,8	87,4	96,9	97,4	93	26,5
L NG50	195	1,648	1,648	1,667	1,338	838	756	507	none
L NGA75	none	1,274	1,274	1,277	605	458	430	344	none
RL GA%	76,9	96,3	96,4	98,9	93,3	95	96,4	95,9	60,3

Accession numbers for S, M and L segments (KU681010-KU681012) of ZLCV; GA%, genome fraction covered by contigs; NG50, the length for which the collection of all contigs of that length or longer covers at least half the reference genome; NGA75, the length for which the collection of all aligned contigs of that length or longer covers at least 75% the reference genome. Reference length (RL) is the nucleotide sum of all segments S, M and L from ZLCV. Employed *k*-mers: ABySS (55); IDBA-UD (21,41,61,77); MEGAHIT (21,41,61,81,99); SPAdes (21,33,55,77); StriDe (55); Velvet (21,33,55,77).

9. Perspectives

Applications in the areas of new pathogen discovery and phytosanitation are expected to increase as NGS technologies become more accessible. In contrast to other detection techniques, such as ELISA and RT-PCR, massive parallel sequencing can assay the viral content of a sample with high resolution, without the aid of specific oligonucleotides or antibodies. Bioassays provide a sensitive method for indexing but nevertheless have important drawbacks, such as long incubation periods until symptoms appear in indicator plants. High-throughput sequencing is more sensitive and faster than bioassays and other indexing approaches for detecting and identifying viruses (Al Rwahnih *et al*., 2015). These techniques, however, are complementary rather than exclusive, and bioassays should be performed to confirm results from NGS experiments that help to understand the aetiology, symptomatology and epidemiology of complex plant viral diseases. Moreover, high-throughput sequencing might detect contaminant viral sequences or viruses that might not be actually replicating at the plant tissue where it was found.

Oxford Nanopore has invested in the concept of 'strand sequencing' by developing portable devices that can generate long reads. The concept uses a protein-linked nanopore as a biosensor that is essentially a nano-scale hole in an electrically resistant synthetic membrane. An ionic current is passed through the nanopore, and as analytes pass through the membrane, the system allows the identification of nucleotide bases, RNA and microRNA and even proteins via aptamer-protein coupling. Long read sequences are important for spanning repetitive elements and resolving repetitive regions, especially for complex genomes. The reduction of sequencing bias by directly sequencing RNA without having to convert it to cDNA is one of the newest developments. This technology, for example, has been shown by Bolisetty *et al.* (2015) to be a powerful method to characterise exon connectivity. Another company, PacBio (Pacific Biosciences), has also invested in technologies that produce long read data. The advent of single-molecule real-time sequencing has, for example, made possible the discovery of circular RNAs in rice that have specific tissue and/or developmental phase expression known to play a role in transcriptional and post-transcriptional gene regulation (Lu *et al.*, 2015). The demand is increasing for developing and improving assemblers capable of handling high error rates derived from long sequencing reads and assemblers that can process hybrid data (mixtures of short and long reads derived from different technologies). The self-correction of reads was developed for PacBio data but often still requires high coverage of the longest reads. Despite the relatively high error rates, emergent technologies have begun a new revolution in sequencing. Powerful new sequencing techniques using nanoelectronics are clearly on the horizon. The possibility today to directly sequence molecules in real time using either biological or synthetic (e.g. silicon nitride, silicon oxide or graphene) pores and therefore the possibility to generate bias-free data provide new avenues in genomic sequencing.

De Bruijn graph algorithms have become very useful in the last years, and in fact, their usage has increased since the introduction of short read sequencing. The choice of the *k*-mer length is an essential step for this class of assemblers. As the overlap between reads are represented as shared *k*-mers of a fixed length of *k*, for smaller values of *k*, the number of detectable overlaps increases, especially for low coverage sequences, but the chance of finding spurious false-positive overlaps is also increased. Genomes that exhibit repeat structures that are larger than the value of *k* make the graph collapse, and the algorithm is unable to resolve the correct path with high precision. Therefore, currently, many research groups are developing assemblers that accept hybrid data; for instance, SPAdes now affords hybrid assemblies in addition to the non-hybrid ones. Viral metagenomics graph collapsing is not a major problem in viral genomes as repeats are rare. It is important to mention that there is always a *trade-off* between the correctness of assembly for low/high coverage regions, and users should always try different values of *k*.

Last but not least, the search for highly divergent new viruses depends mainly on the identification of distant homologs that share significant sequence similarity than would be expected by chance. Homology inference tools fail when too far diverged sequences exist or when, by chance, an unexpected similarity between sequences occurs, for example, the occurrence by chance of identical *k*-mers that originated from different parts of the genome. Therefore, future challenges are expected for the development of *de novo* assemblies, especially for the development of new tools that detect remote homologies and for tools based on non-homology searches.

10. References

Adams I.P., Miano D.W., Kinyua Z.M., Wangai A., Kimani E., Phiri N., Reeder R., Harju V., Glover R., Hany U., Souza-Richards R., Deb Nath P., Nixon T., Fox A., Barnes A., Smith J., Skelton A., Thwaites R., Mumford R., Boonham N. (2013) Use of next generation sequencing for the identification and characterization of *Maize chlorotic mottle virus* and *Sugarcane mosaic virus* causing lethal necrosis in Kenya. *Plant Pathology*, **62**, 741–749.

Al Rwahnih M., Daubert S., Golino D., Rowhani A. (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology*, **387**, 395–401.

Al Rwahnih M., Daubert S., Golino D., Islas C., Rowhani A. (2015) Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in Grapevine. *Phytopathology*, **105**, 758–763.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**, 455 – 477.

Beerenwinkel N., Günthard H.F., Roth V., Metzner K.J. (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, **3**, 329.

Blouin A.G., Ross H.A., Hobson-Peters J., O'Brien C.A., Warren B., MacDiarmid R. (2016) A new virus discovered by immunocapture of double-stranded RNA, a rapid method for virus enrichment in metagenomics studies. *Molecular Ecology Resources*, **16**, 1255–1263.

Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bolisetty M.T., Rajadinakaran G., Graveley B.R. (2015) Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology*, **16**, 204.

Burger J.T., Maree H.J. (2015) Metagenomic next-generation sequencing of viruses infecting grapevines. In *Plant Pathology: Techniques and Protocols*, pp. 315 – 330. Ed. C. Lacomme. New York, NY, USA: Springer.

Candresse T., Marais A., Faure C., Gentit P. (2013) Association of *Little cherry virus 1* (LChV1) with the Shirofugen stunt disease and characterization of the genome of a divergent LChV1 isolate. *Phytopathology*, **103**, 293–298.

Candresse T., Filloux D., Muhire B., Julian C., Galzi S., Fort G., Bernardo P., Daugrois J.H., Fernandez E., Martin D.P., Varsani A., Roumagnac P. (2014) Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS One*, **9**, e102945.

Compeau P.E., Pevzner P.A., Tesler G. (2011) How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, **29**, 987–991.

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

Gentit P., Foissac X., Svanella-Dumas L., Peypelut M., Candresse T. (2001) Characterization of two different apricot latent virus variants associated with peach asteroid spot and peach sooty ringspot diseases. *Archives of Virology*, **146**, 1453 – 1464.

Greninger A.L., Naccache S.N., Federman S., Yu G., Mbala P., Bres V., Stryke D., Bouquet J., Somasekar S., Linnen J.M., Dodd R., Mulembakani P., Schneider B.S., Muyembe-Tamfum J.-J., Stramer S.L., Chiu C.Y. (2015) Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, **7**, 99.

Gurevich A., Saveliev V., Vyahhi N., Tesler G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Ho T., Tzanetakis I.E. (2014) Development of a virus detection and discovery pipeline using next generation sequencing. *Virology*, **471–473**, 54–60.

Idris A., Al-Saleh M., Piatek M.J., Al-Shahwan I., Ali S., Brown J.K. (2014) Viral metagenomics: analysis of begomoviruses by Illumina high-throughput sequencing. *Viruses*, **6**, 1219–1236.

Inoue-Nagata A.K., Albuquerque L.C., Rocha W.B., Nagata T. (2004) A simple method for cloning the complete begomovirus genome using the bacteriophage phi29 DNA polymerase. *Journal of Virological Methods*, **116**, 209–211.

Jayasundara D., Saeed I., Maheswararajah S., Chang B.C., Tang S.L., Halgamuge S.K. (2014) ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics*, **31**, 886 – 896.

Jeck W.R., Reinhardt J.A., Baltrus D.A., Hickenbotham M.T., Magrini V., Mardis E.R., Dangl J.L., Jones C.D. (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics*, **23**, 2942–2944.

Jiang H., Lei R., Ding S.W., Zhu S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, **15**, 182.

Khalifa M.E., Varsani A., Ganley A.R.D., Pearson M.N. (2016) Comparison of Illumina de novo assembled and Sanger sequenced viral genomes: a case study for RNA viruses recovered from the plant pathogenic fungus *Sclerotinia sclerotiorum*. *Virus Research*, **219**, 51–57.

Kleiner M., Hooper L.V., Duerkop B.A. (2015) Evaluation of methods to purify virus-like particles for metagenomics sequencing of intestinal viromes. *BMC Genomics*, **16**, 7.

Kortenhoeven C., Joubert F., Bastos A.D.S., Abolnik C. (2015) Virus genome dynamics under different propagation pressures: reconstruction of whole genome haplotypes of West Nile viruses from NGS data. *BMC Genomics*, **16**, 118.

Kreuze J.F., Perez A., Untiveros M., Quispe D., Fuentes S., Barker I., 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.

Kutnjak D., Silvestre R., Cuellar W., Perez W., Müller G., Ravnikar M., Kreuze J. (2014) Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach. *Virus Research*, **191**, 45–50.

Kutnjak D., Rupar M., Gutierrez-Aguirre I., Curk T., Kreuze J.F., Ravnikar M. (2015) Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscapes of a plant virus population. *Journal of Virology*, **89**, 4760–4769.

Lai B., Wang F., Wang X., Duan L. (2015) InteMAP: integrated metagenomic assembly pipeline for NGS short reads. *BMC Bioinformatics*, **16**, 244.

Langmead B., Salzberg S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

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

Li H., Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754 – 1760.

Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K., Li S., Yang H., Wang J., Wang J. (2009) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research*, **20**, 265–272.

Li D., Liu C.M., Luo R., Sadakane K., Lam T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.

Lima R.N., De Oliveira A.S., Leastro M.O., Blawid R., Nagata T., Resende R.O., Melo F.L. (2016) The complete genome of the tospovirus *Zucchini lethal chlorosis virus*. *Virology Journal*, **13**, 123.

Lin Y., Li J., Shen H., Zhang L., Papasian C.J., Deng H.-W. (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics*, **27**, 2031–2037.

Lorenzi H. (2013) Viral metagenome annotation pipeline. In *Encyclopedia of Metagenomics*, pp. 1–12. Ed. K.E. Nelson. New York, NY, USA: Springer.

Lu T., Cui L., Zhou Y., Zhu C., Fan D., Gong H., Zhao Q., Zhou C., Zhao Y., Lu D., Luo J., Wang Y., Tian Q., Feng Q., Huang T., Han B. (2015) Transcriptome-wide investigation of circular RNAs in rice. *RNA*, **21**, 2076–2087.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.-M., Peng S., Zhu X., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.-W., Wang J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**, 18.

Medina-Medina N., Broka A., Lacey S., Lin H., Klings E.S., Baldwin C.T., Steinberg M.H., Sebastiani P. (2012) Comparing Bowtie and BWA to align short reads from a RNA-Seq experiment. In *6th International Conference on Practical Applications of Computational Biology & Bioinformatics*, **154**, pp. 197–207. Eds M.P. Rocha, N. Luscombe, F. Fdez-Riverola, J.M. Corchado Rodríguez. Berlin, Germany: Springer-Verlag.

Mikheenko A., Saveliev V., Gurevich A. (2015) MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, **32**, 1088.

Miller J.R., Delcher A.L., Koren S., Venter E., Walenz B.P., Brownley A., Johnson J., Li K., Mobarry C., Sutton G. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**, 2818–2824.

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

Naccache S.N., Federman S., Veeraraghavan N., Zaharia M., Lee D., Samayoa E., Bouquet J., Greninger A.L., Luk K.C., Enge B., Wadford D.A., Messenger S.L., Genrich G.L., Pellegrino K., Grard G., Leroy E., Schneider B.S., Fair J.N., Martínez M.A., Isa P., Crump J.A., DeRisi J.L., Sittler T., Hackett J., Jr., Miller S., Chiu C.Y. (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, **24**, 1180–1192.

Nakamura Y., Yasuike M., Nishiki I., Iwasaki Y. (2016) V-GAP: viral genome assembly pipeline. *Gene*, **576**, 676 – 680.

Namiki T., Hachiya T., Tanaka H., Sakakibara Y. (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, **40**, e155.

Ndunguru J., Sseruwagi P., Tairo F., Stomeo F., Maina S., Djikeng A., Djinkeng A., Kehoe M., Boykin L.M. (2015) Analyses of twelve new whole genome sequences of Cassava Brown Streak Viruses and Ugandan Cassava Brown Streak Viruses from East Africa: diversity, supercomputing and evidence for further speciation. *PLoS One*, **10**, e0139321.

O'Brien C.A., Hobson-Peters J., Yam A.W., Colmant A.M., McLean B.J., Prow N.A., Watterson D., Hall-Mendelin S., Warrilow D., Nq M.L., Khromykh A.A., Hall R.A. (2015) Viral RNA intermediates as targets for detection and discovery of novel and emerging mosquito-borne viruses. *PLoS Neglected Tropical Diseases*, **23**, e0003629.

Peng Y., Leung H.C.M., Yiu S.M., Chin F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Qin L., Zhang Z., Zhao X., Wu X., Chen Y., Tan Z., Li S. (2014) Survey and analysis of simple sequence repeats (SSRs) present in the genomes of plant viroids. *FEBS Open Biology*, **4**, 185–189.

Reavy B., Swanson M.M., Cock P.J., Dawson L., Freitag T.E., Singh B.K., Torrance L., Mushegian A.R., Taliansky M. (2015) Distinct circular single-stranded DNA viruses exist in different soil types. *Applied and Environmental Microbiology*, **81**, 3934–3945.

Roossinck M.J., Saha P., Wiley G.B., Quan J., White J.D., Lai H., Chavarría F., Shen G., Roe B.A. (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology*, **19** Suppl. 1, 81–88.

Rosario K., Nilsson C., Lim Y.W., Ruan Y., Breitbart M. (2009) Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology*, **11**, 2806–2820.

Rosario K., Padilla-Rodriguez M., Kraberger S., Stainton D., Martin D.P., Breitbart M., Varsani A. (2013) Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, **171**, 231–237. DOI:10.1016/j.virusres.2012.10.01.

Ruffalo M., LaFramboise T., Koyutürk M. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, **27**, 2790–2796.

Salzberg S.L., Phillippy A.M., Zimin A., Puiu D., Magoc T., Koren S., Treangen T.J., Schatz M.C., Delcher A.L., Roberts M., Marçais G., Pop M., Yorke J.A. (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, **22**, 557–567.

Sela N., Lachman O., Reingold V., Dombrovsky A. (2013) A new cryptic virus belonging to the family Partitiviridae was found in watermelon co-infected with Melon necrotic spot virus. *Virus Genes*, **47**, 382–384.

Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Valverde R.A., Dodds J.A., Heick J.A. (1986) Double-stranded ribonucleic acid from plants infected with viruses having elongated particles and undivided genomes. *Phytopathology*, **76**, 459–465.

Vázques-Castellanos J.F., García-Lopéz R., Pérez-Brocal V., Pignatelli M., Moya A. (2014) Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics*, **15**, 37.

Wan Y., Renner D.W., Albert I., Szpara M.L. (2015) VirAmp: a galaxy-based viral genome assembly pipeline. *GigaScience*, **4**, 19.

Wang Q., Jia P., Zhao Z. (2013) VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One*, **8**, e64465.

Warren R.L., Sutton G.G., Jones S.J.M., Holt R.A. (2006) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500–501.

Williams C.R., Baccarella A., Parrish J.Z., Kim C.C. (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, **17**, 103.

Wu Q.F., Ding S.W., Zhang Y.J., Zhu S.F. (2015) Identification of viruses and viroids by next-generation sequencing and homology-dependent and homology-independent algorithms. *Annual Review of Phytopathology*, **53**, 425–444.

Yamashita A., Sekizuka T., Kuroda M. (2016) VirusTAP: viral genome-targeted assembly pipeline. *Frontiers in Microbiology*, **7**, 32.

Yang X., Charlebois P., Gnerre S., Coole M.G., Lennon N.J., Levin J.Z., Qu J., Ryan E.M., Zody M.C., Henn M.R. (2012) De novo assembly of highly diverse viral populations. *BMC Genomics*, **13**, 475.

Yu X., Guda K., Willis J., Veigl M., Wang Z., Markowitz S., Adams M.D., Sun S. (2012) How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining*, **5**, 6.

Zerbino D.R., Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Capítulo 3. Discovery and molecular characterization of a novel enamovirus, Grapevine enamovirus-1

1. Abstract

In this study, we describe a novel putative *Enamovirus* member, Grapevine enamovirus-1 (GEV-1), discovered by high-throughput sequencing (HTS). A limited survey using HTS of 17 grapevines (*Vitis* spp.) from the south, southeast, and northeast regions of Brazil led to the detection of GEV-1 exclusively on southern plants, infecting four grapevine cultivars (Cabernet Sauvignon, Semillon, CG 90450, and Cabernet franc) with a remarkable identity of around 99% at the nucleotide level. This novel virus was only detected in multiple-virus infected plants exhibiting viral-like symptoms. GEV-1 was also detected on a cv. Malvasia Longa by RT-PCR. We performed graft-transmissibility assays on GEV-1. The organization, products, and cis-acting regulatory elements of GEV-1 genome are also discussed here. The near complete genome sequence of GEV-1 was obtained during the course of this study, lacking only part of the 3' untranslated terminal region. This is the first report of a virus in the family *Luteoviridae* infecting grapevines. Based on its genomic properties and phylogenetic analyses, GEV-1 should be classified as a new member of the genus *Enamovirus*.

Keywords

High-throughput sequencing, Virus Discovery, *Luteoviridae*, Grapevine enamovirus-1, GEV-1

Este capítulo foi publicado na revista Virus Genes.
Silva J.M.F., Al Rwahnih M., Blawid R., Nagata T., Fajardo T.V.M. (2017) Discovery and molecular characterization of a novel enamovirus, Grapevine enamovirus-1. *Virus Genes*, **53**(4), 667-671.

2. Text

A limited survey was performed on 17 grapevine samples subjected to high-throughput sequencing (HTS). These plants were collected from three different grapevine collections from the south (11), northeast (2), and southeast (4) regions of Brazil, and the symptoms in the *V. vinifera* hosts were downward rolling of leaves and reddening or yellowing, whereas other genotypes were asymptomatic (Fajardo *et al.*, 2017; Fajardo *et al.*, 2016). Following a typical metagenomic pipeline using HTS, we were able to identify a new putative *Enamovirus* member, tentatively named Grapevine enamovirus-1 (GEV-1), infecting distinct grapevine cultivars in Brazil. The family *Luteoviridae* comprises three genera, *Luteovirus*, *Polerovirus*, and *Enamovirus*. These viruses have a positive-sense RNA genome of around 5.2–6.3 kb (Domier, 2012). The genus *Enamovirus* has only one recognized viral species, *Pea enation mosaic virus-1* (PEMV-1), and two putative members, Citrus vein enation virus (CVEV) and Alfalfa enamovirus-1 (AEV-1). The systemic movement of PEMV-1, type species of the genus Enamovirus, is provided by an umbravirus (Peter *et al*., 2009), although this has not been reported for the remaining putative enamoviruses. Viruses in the family *Luteoviridae* are transmitted by aphids in a circulative non-persistent manner (Domier, 2012).

To characterize the viromes of these plants, double-stranded RNA (dsRNA) extracts were subjected to HTS on the Illumina HiSeq 2000 platform. Briefly, reads were trimmed, de novo assembled (CLC bio, Qiagen, USA), and subjected to a BLASTX search against the NCBI viral RefSeq database. This led to the identification of a novel luteovirid in four grapevine samples of different cultivars: Cabernet Sauvignon (S1M-CS), CG 90450 (S12-CG), Semillon (S16-SE), and Cabernet franc (S19-CF). All vines exhibiting symptoms of viral infection were collected from the Rio Grande do Sul (RS) state, south region of Brazil. Seven sets of primers (Table 1) were designed to confirm the infection of GEV-1 on the S1M-CS vine, and amplicons corresponding to sets 1, 6, and I were sequenced, which verified an identity of 99% with the contigs built by de novo assembly. Contigs with Luteoviridae hits in the samples S1M-CS (1), S12-CG (5), S16-SE (1), and S19-CF (5) covered, respectively, 99, 93, 99, and 95% of the near complete GEV-1 sequence, obtained after rapid amplification of the cDNA ends (RACE) in the 5' extremity by the Terminal deoxynucleotidyl transferase (TdT) method. After scaffolding of the fragmented S12-CG and S19-CF contigs, an identity of around 99% at the nucleotide level was verified for GEV-1 among these four samples. Mixed infections were present in all GEV-1 positive samples (Fajardo *et al*., 2017; Fajardo *et al*., 2016) and the virus communities in these samples included 11 pathogens: Grapevine Cabernet Sauvignon reovirus (GCSV), *Grapevine vein clearing virus* (GVCV), *Grapevine Red Globe virus* (GRGV), *Grapevine leafroll-associated virus 2 and 3* (GLRaV-2, -3), *Grapevine rupestris stem pitting-associated virus* (GRSPaV), *Grapevine virus A* (GVA), *Grapevine virus B* (GVB), *Grapevine fleck virus* (GFkV), *Grapevine rupestris vein feathering virus* (GRVFV), and *Grapevine yellow speckle viroid 1* (GYSVd-1). Additionally, GEV-1 was detected on a Malvasia Longa vine by RT-PCR during the virus indexing of this cultivar. The near complete genome of GEV-1 isolate CS-BR (6227 bp), lacking only part of the 3' untranslated sequence, was deposited in GenBank under accession KX645875. The isolate SE-BR near complete sequence (6176 bp), obtained only by de novo assembly, was also deposited in GenBank under accession KY820716.

To assess the graft transmissibility of the novel virus, cv. Cabernet Sauvignon (S1M-CS) was grafted onto 16 healthy cv. 1103P plants *(V. berlandieri* x *V. rupestris*). Graft-transmissibility of GEV-1 was confirmed in 13 out of 16 positive samples by performing RT-PCR 5 months after grafting.

Luteovirids are known to harbor five to ten open reading frames (ORFs) usually displayed as two gene blocks separated by a non-coding intergenic region (Ashoub *et al*., 1998; Smirnova *et al*., 2015; Jaag *et al*., 2003). The 5'-proximal block contains the two partially overlapping ORFs 1 and 2, plus an additional ORF encoding a silencing suppressor protein (ORF 0) in the genera *Enamovirus* and *Polerovirus*. The 3'-proximal gene block contains the ORFs corresponding to the coat protein (ORF 3), an extension of the coat protein translated by an in-frame stop codon readthrough (ORF 5) and a movement protein (ORF 4) located within ORF 3 in the genera *Luteovirus* and *Polerovirus* that is absent in the genus *Enamovirus*. ORFs 3, 4, and 5 are translated from a subgenomic RNA (sgRNA). GEV-1 genomic features (Fig. 1a) are discussed below.

Table 1 Primers used for detection of Grapevine enamovirus-1 (GEV-1)

Name	Sequence	5' nt position	Length (bp)
Set1F	CACACTTGCTTCTCTTCTCG	50	749
Set1R	CCAACGTAAGCGAATAGTCG	798	
Set2F	GTTGGAGAGAGGAAAGAATCGG	1323	684
Set2R	GGGTTTGTCTGTGACCTCATAGTC	2006	
Set3F	AGGCCAAGAGGGGCAAGAAATTGT	2434	530
Set3R	CGGCAGATTTTGATCTAGCAGCTC	2963	
Set4F	ACAAGCAGGAGTTGAGGATG	3423	600
Set4R	CGACGAGCATTTTACCCACA	4022	
Set5F	GGACAGAGGTTGCATTGCGTAT	4622	505
Set5R	TTGAAACCGAGCCAAGTGAGTGTC	5126	
Set6F	TTCCCTTGGGAGACTCGGTTCTAT	5263	735
Set6R	AAACATGACCACCCGTCTCATAGC	5997	
SetIF	AAAGTGGTGTGTCGCTATGG	3850	638
SetIR	GGCAAACGAATTTACCAAGAACG	4487	

ORF 0 (nt 335–1279) overlaps ORF 1 and potentially encodes a 34.9 kDa protein (P0), presumably a suppressor of host RNAi machinery. The F-box-like domain (LPxxI/L(x10–13)P) found in the P0 of polero- and enamoviruses is necessary for its silencing suppressor activity (Fusaro *et al.*, 2012, Pazhouhandeh *et al.*, 2006). Interestingly, it was verified that only the first leucine is conserved on GEV-1. ORF 1 (nt 471–2912) potentially encodes an 89.8 kDa protein (P1), it contains a conserved 3C-like serine peptidase followed by the genome-linked viral protein (VPg). The conserved domain H(x25)D(x70–80)GxSG of the S39 serine protease is positioned between nt 1350 and 1841 (Li *et al.*, 2000). Alignments with PEMV-1 suggest that the first VPg cleavage site (E/S) at the GEV-1 genome is positioned at nt 1938 (Gorbalenya and Koonin, 1993; Wobus *et al.*, 1999), but we were unable to deduce the second proteolysis site. The W(A, G)D motif followed by a DE-rich region (Mäkinen *et al.*, 1995) is located between nt 2079 and 2111. ORF 2 (nt 2168–4165) is translated by a -1 frameshift from ORF 1, originating a fusion protein (P1–P2) containing the RNA-dependent RNA polymerase (RdRp), of a predicted molecular mass of 36.55 kDa. The highly conserved GDD box motif (Koonin and Dolja, 1993) is located between nt 3869 and 3877. ORF 3 (nt 4281–4871) potentially encodes a 21.4 kDa protein, which corresponds to the coat protein (CP). ORF 5 (nt 4872–5915) encodes the readthrough domain (RTD) of the CP-RTD fusion protein, predicted to have a total molecular mass of 59.7 kDa. This protein is needed for efficient aphid transmission (Liu *et al.*, 2009). GEV-1 lacks the C-terminal portion of the CP-RTD protein that is responsible for limiting the infection of luteo- and poleroviruses to the phloem (Peter *et al*., 2009). The amino acid identity between GEV-1 and others enamoviruses is below 44% for all ORFs.

Viruses in the family *Luteoviridae* employ a wide range of translational mechanisms which are regulated by cis-acting RNA elements (CRE) embedded in the virus genome (Smirnova *et al*., 2015; Newburn and White, 2015). GEV-1 ORF 0 possesses a leaky start codon UAU**AUG**U, allowing the translation of ORF 1 (Kosak, 1987; Firth and Bierley, 2012). Two signals are required for the -1 ribosomal frameshift at ORF 1, the heptanucleotide sequence XXXYYYZ and a downstream pseudoknot or very stable RNA secondary structure located six to eight nucleotides from the frame-shift site (Giedroc and Cornish, 2009). We found the TTTAAAC sequence located at nt 2168 and a pseudoknot seven nt downstream of this site, predicted with the RNAPKplex program (Lorenz *et al*., 2011). The CCNNNN tandem repeat motif associated with ORF 3 stop codon readthrough (Brown *et al*., 1996) is located between nt 4887 and 4935, 15 nucleotides downstream from the termination site. Remarkably, GEV-1 readthrough site at ORF 3 (UUG**UGA**UAU) is not similar to any previously reported *Luteoviridae* (Firth and Bierley, 2012).

Maximum likelihood trees for the family *Luteoviridae* were estimated based on the P2- and CP-translated sequences (Fig. 1b). The ORF 2 is separated from ORF 3 by an intergenic region which is a probable hot spot for recombination among luteovirids (Moonan *et al*., 2000; Pagán and Holmes, 2010) so incongruences in the trees when considering these two distinct regions are expected. In both trees, GEV-1 clusters together with PEMV-1 and AEV-1, indicating that GEV-1 is more closely related to the genus *Enamovirus*.

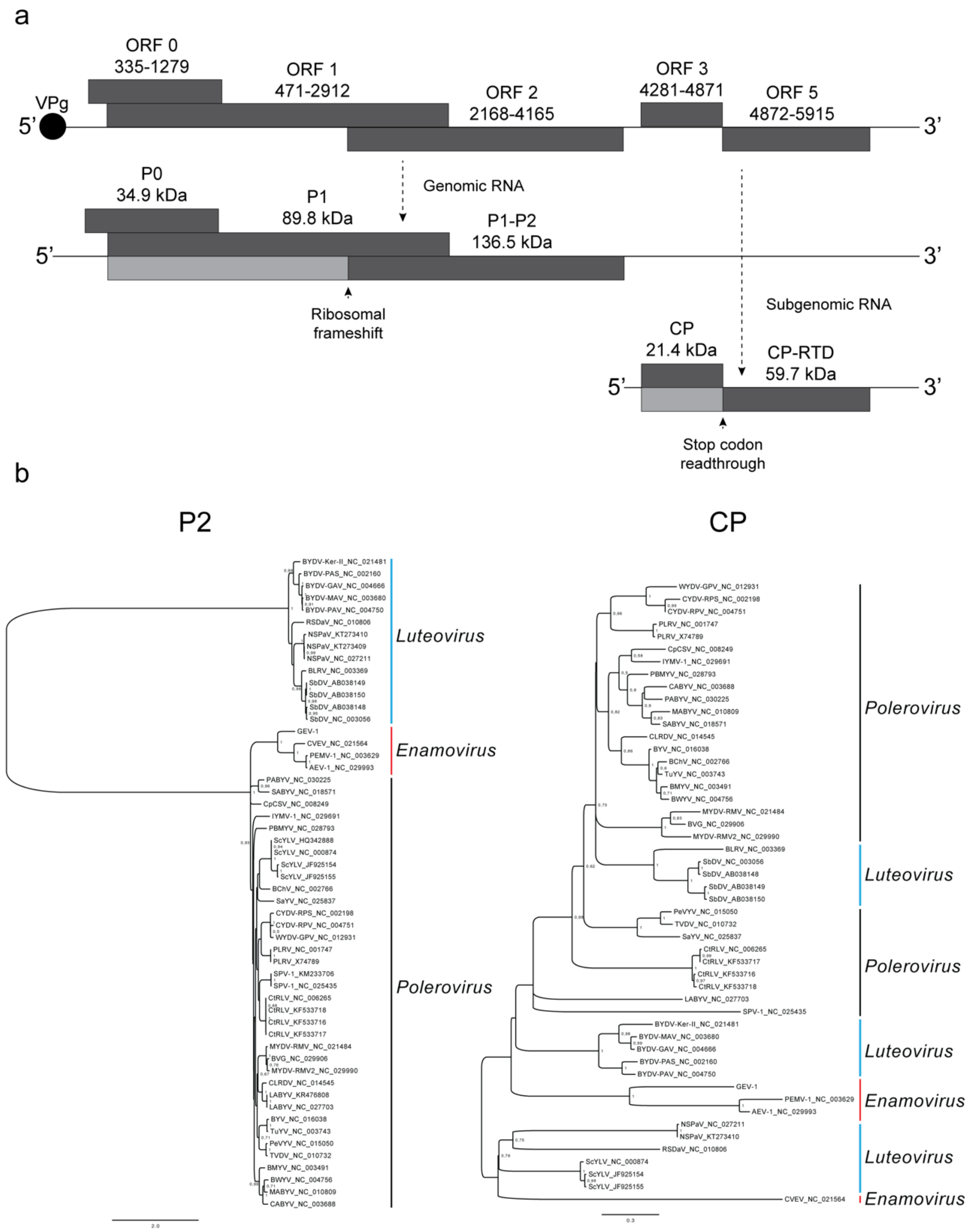


Fig. 1 **a** GEV-1 genome organization and **b** Maximum likelihood trees (JTT + G(4) + I; Bootstrap = 1000 replications) for the family *Luteoviridae* using the P2 and CP amino acid sequences. Trees were inferred with MEGA 7 (Kumar *et al*., 2016). Alignments were performed with MUSCLE. Trees were midpoint rooted

Using HTS of dsRNA extracts from 17 samples, some of them exhibiting viral-like symptoms, we identified in four samples, within a virus community, a new putative *Enamovirus* member, provisionally named Grapevine enamovirus-1 (GEV-1), infecting distinct grapevine cultivars from the south region of Brazil. The distinguishing feature of the genus Enamovirus is the lack of a movement protein (Domier, 2012). No ORF corresponding to this protein could be identified on GEV-1. Bioassays confirmed infection and graft-transmissibility of GEV-1. Due to the lack of a movement protein, it is possible that GEV-1 needs co-infection with another virus for its cell-to-cell movement and graft-transmissibility. Phylogenetic analyses revealed that GEV-1 is more closely related to the genus *Enamovirus*. Based on these data, GEV-1 should be classified as a new member of the genus *Enamovirus*. Due to its sensitivity, HTS have been proposed as a diagnostic tool for biosecurity and quarantine surveillance (Massart *et al.*, 2014; Bag *et al.*, 2015; Al Rwahnih *et al.*, 2015). Despite being a valuable tool for discovering novel viruses in metagenomic samples, information regarding the biological significance of a newly discovered virus such as pathogenicity, transmission, host range, and epidemiology often cannot be obtained by these means (Massart *et al.*, 2017). In addition, grapevines often present complex pathosystems, and further studies are needed to understand the interaction between these pathogens and their effect on the vines health, development, and quality of the grapes.

3. References

Al Rwahnih M., Daubert S., Golino D., Islas C., Rowhani A. (2015) Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. *Phytopathology*, **105**(6), 758-763.

Ashou, A., Rohde W., Prüfer D. (1998) *In planta* transcription of a second subgenomic RNA increases the complexity of the subgroup 2 luteovirus genome. *Nucleic Acids Research*, **26**(2), 420-426.

Bag S., Al Rwahnih M., Li A., Gonzalez A., Rowhani A., Uyemoto J.K., Sudarshana M.R. (2015) Detection of a new luteovirus in imported nectarine trees: a case study to propose adoption of metagenomics in post-entry quarantine. *Phytopathology*, **105**(6), 840-846.

Brown C.M., Dinesh-Kumar S.P., Miller W.A. (1996) Local and distant sequences are required for efficient readthrough of the barley yellow dwarf virus PAV coat protein gene stop codon. *Journal of Virology*, **70**(9), 5884-5892.

Domier L.L. (2012) Family *Luteoviridae*. In *Virus taxonomy. Ninth report of the International Committee on Taxonomy of Viruses*, pp. 1045–1053. Eds King A.M.Q., Adams M.J., Carstens E.B., Lefkowitz E.J. San Diego, USA: Elsevier Academic

Fajardo T.V.M., Eiras M., Nickel O. (2016) Detection and molecular characterization of Grapevine yellow speckle viroid 1 isolates infecting grapevines in Brazil. *Tropical Plant Pathology*, **41**(4), 246-253.

Fajardo T.V.M., Silva F.N., Eiras M., Nickel O. (2017) High-throughput sequencing applied for the identification of viruses infecting grapevines in Brazil and genetic variability analysis. *Tropical Plant Pathology* doi:10.1007/s40858-017-0142-8

Firth A.E., Brierley I. (2012) Non-canonical translation in RNA viruses. *Journal of General Virology*, **93**(7), 1385-1409.

Fusaro A.F., Correa R.L., Nakasugi K., Jackson C., Kawchuk L., Vaslin M.F., Waterhouse P. M. (2012) The *Enamovirus* P0 protein is a silencing suppressor which inhibits local and systemic RNA silencing through AGO1 degradation. *Virology*, **426**(2), 178-187.

Giedroc D.P., Cornish P.V. (2009) Frameshifting RNA pseudoknots: structure and mechanism. *Virus Research*, **139**(2), 193-208.

Gorbalenya A.E., Koonin E.V. (1993) Comparative analysis of amino-acid sequences of key enzymes of replication and expression of positive-strand RNA viruses: validity of approach and functional and evolutionary implications. *Sov Sci Rev D Physicochem Biol*, **11**, 1-84.

Jaag H.M., Kawchuk L., Rohde W., Fischer R., Emans N., Prüfer D. (2003) An unusual internal ribosomal entry site of inverted symmetry directs expression of a potato leafroll polerovirus replication-associated protein. *Proceedings of the National Academy of Sciences*, **100**(15), 8939-8944.

Koonin E.V., Dolja V.V., Morris T.J. (1993) Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Critical Reviews in Biochemistry and Molecular Biology*, **28**(5), 375-430.

Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, **15**(20), 8125-8148.

Kumar S., Stecher G., Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**(7), 1870-1874.

Li X., Ryan M.D., Lamb J.W. (2000) Potato leafroll virus protein P1 contains a serine proteinase domain. *Journal of General Virology*, **81**(7), 1857-1864.

Liu S., Sivakumar S., Wang Z., Bonning B.C., Miller W.A. (2009) The readthrough domain of pea enation mosaic virus coat protein is not essential for virus stability in the hemolymph of the pea aphid. *Archives of Virology*, **154**(3), 469-479.

Lorenz R., Bernhart S.H., Zu Siederdissen C.H., Tafer H., Flamm C., Stadler P.F., Hofacker I.L. (2011) ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, **6**(1), 26.

Mäkinen K., Tamm T., Næss V., Truve E., Puurand Ü., Munthe T., Saarma M. (1995) Characterization of cocksfoot mottle sobemovirus genomic RNA and sequence comparison with related viruses. *Journal of General Virology*, **76**(11), 2817-2825.

Massart S., Candresse T., Gil J., Lacomme C., Predajna L., Ravnikar M., Reynard J.S., Rumbou A., Saldarelli P., Škorić D., Vainio E.J. (2017) A Framework for the Evaluation of Biosecurity, Commercial, Regulatory, and Scientific Impacts of Plant Viruses and Viroids Identified by NGS Technologies. *Frontiers in Microbiology*, **8**.

Massart S., Olmos A., Jijakli H., Candresse T. (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research*, **188**, 90-96.

Moonan F., Molina J., Mirkov T.E. (2000) Sugarcane yellow leaf virus: an emerging virus that has evolved by recombination between luteoviral and poleroviral ancestors. *Virology*, **269**(1), 156-171.

Newburn L.R., White K.A. (2015) *Cis*-acting RNA elements in positive-strand RNA plant virus genomes. *Virology*, **479**, 434-443.

Pagán I., Holmes E.C. (2010) Long-term evolution of the Luteoviridae: time scale and mode of virus speciation. *Journal of Virology*, **84**(12), 6177-6187.

Pazhouhandeh M., Dieterle M., Marrocco K., Lechner E., Berry B., Brault V., Hemmer O., Kretsch T., Richards K.E., Genschik P., Ziegler-Graff V. (2006) F-box-like domain in the polerovirus protein P0 is required for silencing suppressor function. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(6), 1994-1999.

Peter K.A., Gildow F., Palukaitis P., Gray S. M. (2009) The C terminus of the polerovirus p5 readthrough domain limits virus infection to the phloem. *Journal of Virology*, **83**(11), 5419-5429.

Smirnova E., Firth A.E., Miller W.A., Scheidecker D., Brault V., Reinbold C., Rakotondrafara A.M., Chung B.Y., Ziegler-Graff V. (2015) Discovery of a small non-AUG-initiated ORF in poleroviruses and luteoviruses that is required for long-distance movement. *PLoS Pathogens*, **11**(5), e1004868

Wobus C.E., Skaf J.S., Schultz M.H., de Zoeten G.A. (1998) Sequencing, genomic localization and initial characterization of the VPg of pea enation mosaic enamovirus. *Journal of General Virology*, **79**(8), 2023-2025.

Capítulo 4. Descoberta e caracterização inicial de um novo vírus *Virga-like* em videiras

1. Resumo

Nesse trabalho reportamos o sequenciamento parcial do genoma de um novo vírus, chamado provisoriamente de Grapevine virga-like virus (GVLV), descoberto através de uma reanálise de dados gerados por HTS usando um banco de dados de referência de sequencias virais atualizado (RefSeq - Virus, maio de 2017). Esse vírus possui um complexo de replicação *alpha-like* e mostra baixa identidade com vírus pertencentes às famílias *Virgaviridae* e *Bromoviridae*, apresentando maior identidade com o vírus não classificado Citrus virga-like virus (CVLV), que também se encontra apenas parcialmente sequenciado. Após a montagem *de novo* dos *contigs* e sequenciamento pelo método Sanger de três *amplicons*, obtivemos 4620 bp do genoma do GVLV. Análise filogenéticas do domínio metiltransferase (Met) posicionaram os GVLV e CVLV como um grupo externo às famílias *Virgaviridae* e *Bromoviridae*, enquanto as análises do domínio parcial da helicase (Hel) os posicionaram mais próximos à família *Virgaviridae*. Uma grande porção da replicase dos GVLV e CVLV não possui similaridade com nenhuma sequência viral conhecida, indicando que esses vírus podem pertencer a um novo grupo ainda não descrito de vírus. Uma visualização das partículas virais, assim como a sequência genômica completa do GVLV é necessária para a sua classificação e caracterização completa.

2. Introdução

As famílias *Virgaviridae* e *Bromoviridae* são compostas por vírus de ssRNA(+) que possuem um complexo de replicação *alpha-like*, que contém os domínios metiltransferase (Met) e helicase (Hel), e uma estrutura *tRNA-like* localizada na extremidade 3' do genoma. A principal diferença entre esses dois grupos é a morfologia de suas partículas virais: enquanto os vírus pertencentes à família *Virgaviridae* formam partículas alongadas, os vírus pertencentes à família *Bromoviridae* formam partículas isosaédricas ou em forma de bacilo (Adams *et al.*, 2009; Bujarski *et al.*, 2012). A distinção entre esses dois grupos é suportada por estudos filogenéticos (Adams *et al.*, 2009), e enquanto os genomas dos vírus pertencentes à família *Bromoviridae* são trisegmentados, os genomas dos que pertencem à família *Virgaviridae* possuem de 1 a 3 segmentos genômicos. Através da técnica de HTS, identificamos um novo vírus em videiras, provisoriamente chamado de Grapevine virga-like virus (GVLV). Esse vírus foi identificado em 3 cultivares distintos de videira: *V. flexuosa* (amostra 2M-VF; 12 *reads*), *V. vinifera* cv. Semillon (amostra S16-S; 26 *reads*) e *V. vinifera* cv. Cabernet Franc (amostra S19-CF; 2 *reads*), e possui maior identidade com o vírus ainda não classificado Citrus virga-like virus (CVLV) (Matsumura *et al.*, 2017).

3. Metodologia

3.1. Material vegetal, preparo das bibliotecas de DNA e sequenciamento

17 videiras, coletadas de vinhedos nas regiões Sul (11 plantas), Sudeste (4) e Nordeste (2) do Brasil foram sujeitas à sequenciamento na plataforma Illumina HiSeq 2000 (100 bp paired-end). A extração de ácidos nucleicos foi feita a partir da extração de RNA fita-dupla (dsRNA) de plantas clonais com celulose CF-11 seguindo o protocolo de Valverde (1986). As bibliotecas de cDNA (TruSeq RNA) foram montadas pela Macrogen (Seul, Coréia do Sul) ou Eurofins Genomics (Huntsville, EUA).

3.2. Montagem *de novo* e relação taxonômica

O *trimming* e remoção de sequências de adaptadores dos *reads* foram feitos com o programa Trimmomatic v0.36 (Bolger *et al.*, 2014). Os *reads* derivados do hospedeiro foram removidos com o BWA v0.7.15 (Li e Durbin, 2010) e SAMtools v1.3.1 (Li *et al.*, 2009). A montagem *de novo* foi feita com o SPAdes v3.5 (Bankevich *et al.*, 2012), e a relação taxonômicas dos *contigs* feita com o tBlastx (Altschul et al, 1990). Adicionalmente, a relação taxonômica dos *reads* foi feita pelo Kaiju webserver (Menzel *et al.*, 2016), e os *reads* que alinharam contra o CVLV foram extraidos e montados *de novo* com o Velvet v0.1.19 (Zerbino e Birney, 2008). Na análise feita pelo Kaiju, o GVLV foi identificado em três amostras distintas: *V. flexuosa* (amostra 2M-VF; 12 *reads*), *V. vinifera* cv. Semillon (amostra S16-S; 26 *reads*) e *V. vinifera* cv. Cabernet Franc (amostra S19-CF; 2 *reads*). Devido à maior cobertura desse vírus na amostra S16-S, as análises seguintes, incluindo o sequenciamento pelo método Sanger, foram feitas utilizando apenas para essa amostra. Ao total, cinco *contigs* foram montados para a amostra S16-S (Fig. 1a). Para verificar que esses *contigs* são de origem viral, buscas por similaridade foram feitas usando o programa Blastn contra o banco de dados *nr* pelo portal do NCBI (blast.ncbi.nlm.nih.gov).

3.3. Sequenciamento parcial do genoma

Quatro pares de *primers* foram desenhados para a confirmar a infecção e unir os *contigs* montados na etapa anterior (Tabela 1). Apenas a amostra S16-S foi utilizada nessa etapa. As reações de PCR foram feitas com a enzima LongAmp (New England Biolabs) e os *amplicons* foram sequenciados pelo método Sanger pela Macrogen, resultando em dois *contigs*, designados GVLV-Met-Hel e GVLV-RdRp. Esses *contigs* foram extendidos e os caractéres ambíguos removidos através de uma remontagem com *contigs* previamente montados pelo CLC assember (CLC Bio, Qiagen, EUA).

Tabela 1 *Primers* utilizados para detecção e sequenciamento do genoma do GVLV

Nome	Sequência	Tamanho do *amplicon*
1F	GGACGAAGTCACAACCAACACAGTTT	478 bp

1R	CGCGAGTAGGTCTGACAACTTTCATTAT	
2F	GGTACGTACATCGCAGACGGAAT	450 bp
2R	TTGTCGGAGCGACTTGGCATATCTAT	
3F	AGGGCAGCCAGGTTTGCTTT	1278 bp
3R	CACCCGGTTTAGCAGTGTGAGTAATA	
4F	CCGACATAAGACCGCTGCGAAGTATT	311 bp
4R	CGGCTCCACCCAAAAATATCTCAAAA	

3.4. Análise filogenética

Alinhamentos das sequências traduzidas dos domínios metiltransferase e helicase parcial foram feitos com o MUSCLE (Edgar, 2004) a partir de sequências virais obtidas através do banco de dados RefSeq -Virus. Os alinhamentos incluem vírus das famílias *Virgaviridae* e *Bromoviridae* e do gênero *Idaevirus*. As árvores filogenéticas foram montadas com o MEGA 7 (Kumar *et al*., 2016) pelo método de máxima verossimilhança (LG + G(5) + I; 1000 replicatas de *bootstrap*).

4. Resultados e discussão

Seguindo um *pipeline* típico de análises metagenômicas, nosso grupo identificou previamente um novo vírus em videiras, Grapevine enamovirus-1 (GEV-1) (Silva *et al*., 2017). Através de uma reanálise desses dados utilizando um banco de dados de sequências virais mais atualizado (RefSeq - Virus, maio de 2017), descobrimos mais um novo vírus nessas amostras, Grapevine virga-like virus (GVLV), mostrando a importância em se usar banco de dados atualizados. O genoma do GVLV foi parcialmente sequenciado, resultando em dois *contigs*: GVLV-Met-Hel (3348 bp), que contém os domínios metiltransferase e helicase parcial; e GVLV-RdRp (1272 bp), que corresponde à RNA polimerase dependente de RNA (*RNA dependent RNA polymerase* – RdRp) parcial (Fig. 1a). Não foi possível, até o momento, a amplificação da região supostamente localizada entre esses *contigs*, sugerindo que eles podem estar localizados em seguimentos genômicos distintos. Com base em alinhementos feitos pelo Blastx, o *contig* GVLV-Met-Hel mostra 26% de identidade com o Bacopa chlorosis virus (*Bromoviridae*; *query cover* = 48%), enquanto o *contig* GVLV-RdRp mostra 32% de identidade com *Rehmania mosaic virus* (*Virgaviridae*; *query cover* = 40%) (Fig. 1b).

Árvores filogenéticas dos domínios metiltransferase e helicase parcial foram montadas por máxima verossimilhança (Figs. 2 e 3). A árvore do domínio metiltransferase posicionou os GVLV e CVLV como um grupo externo às famílias *Virgaviridae* e *Bromoviridae* (Fig. 2), enquanto a árvore montada para o domínio helicase parcial os posicionou dentro da família *Virgaviridae* (Fig. 3), porém com baixo suporte dos valores de *bootstrap*. Incongruências nas árvores filogenéticas, principalmente quanto à posição do gênero *Idaeovirus*, do micovírus não classificado Macrophomina phaseolina tobamo-like virus (YP_009109561) e dos GVLV e CVLV, sugerem possíveis eventos de recombinação entre ancestrais das famílias *Virgaviridae* e *Bromoviridae*. Uma grande porção do *contig* GVLV-RdRp não possui similaridade com nenhum outro vírus conhecido além do CVLV (Fig. 1b), indicando que esses vírus podem pertencer à um novo grupo não descrito de vírus. Para uma caracterização completa do GVLV, as partículas virais serão visualizadas por microscopia eletrônica de transmissão (MET), as extremidades dos *contigs* GVLV-Met-Hel e GVLV-RdRp serão amplificadas por ensaios de amplificação rápida das extremidades do cDNA (*rapid amplification of cDNA ends* – RACE) e RNA extraído de partículas virais semi-purificadas serão enviadas para sequenciamento na plataforma Illumina.

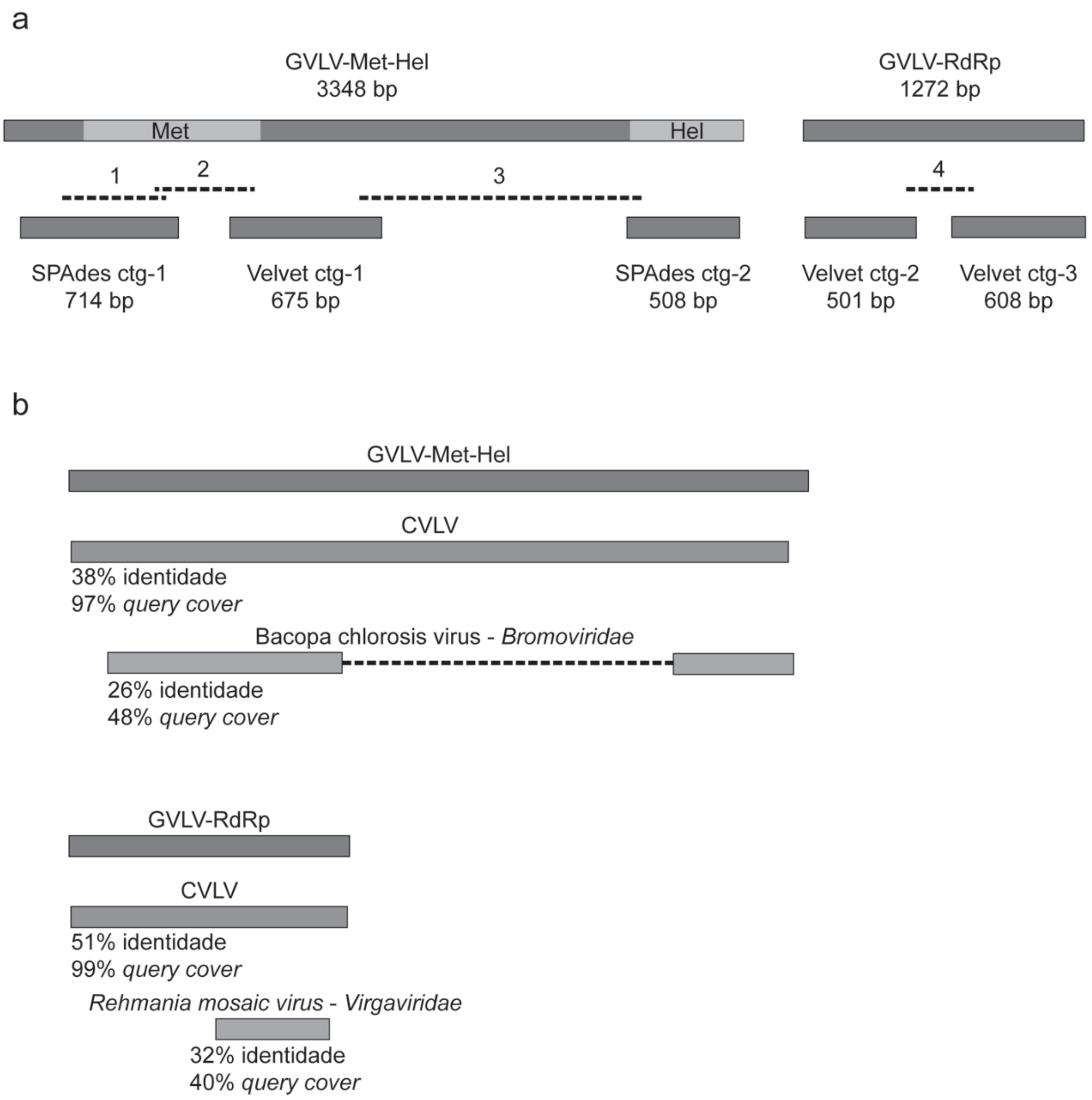


Figura 1 a *Contigs* finais (GVLV-Met-Hel e GVLV-RdRp) obtidos após montagem *de novo* e sequenciamento pelo método Sanger (*amplicons* representados por linhas rachuradas, indicando o par de *primers* utilizado [tabela 1]), destacando os domínios metiltransferase (Met) e helicase parcial (Hel) e *contigs* inicialmente montados pelos programas SPAdes e Velvet; **b** Representação dos alinhamentos dos *contigs* GVLV-Met-Hel e GVLV-RdRp obtidos pelo Blastx

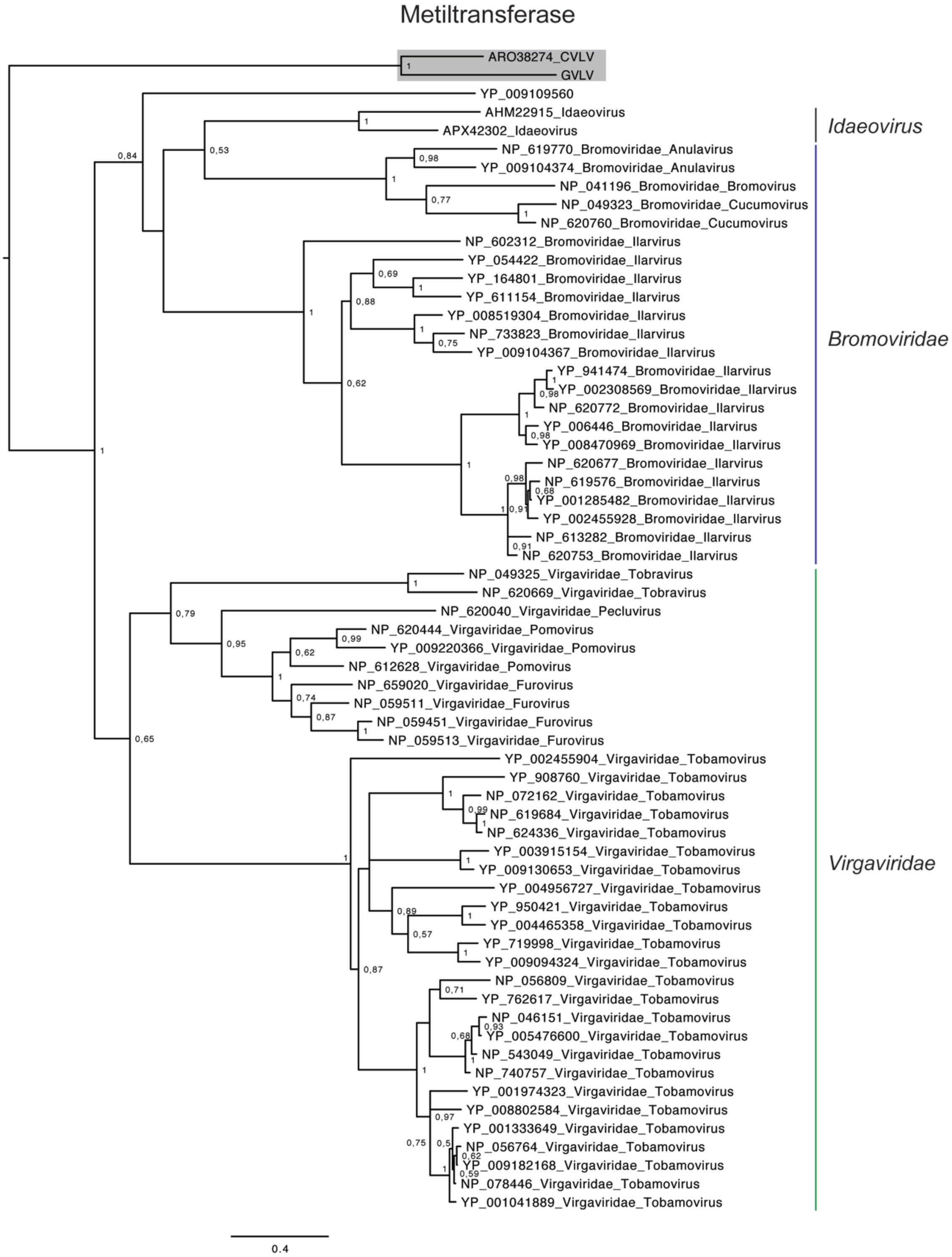

Metiltransferase
ARO38274_CVLV
1
GVLV
YP_009109560
AHM22915_Idaeovirus
1
APX42302_Idaeovirus
Idaeovirus
0,84
0,53
NP_619770_Bromoviridae_Anulavirus
0,98
YP_009104374_Bromoviridae_Anulavirus
1
NP_041196_Bromoviridae_Bromovirus
0,77
NP_049323_Bromoviridae_Cucumovirus
1
NP_620760_Bromoviridae_Cucumovirus
NP_602312_Bromoviridae_Ilarvirus
YP_054422_Bromoviridae_Ilarvirus
0,69
YP_164801_Bromoviridae_Ilarvirus
1
0,88
YP_611154_Bromoviridae_Ilarvirus
1
YP_008519304_Bromoviridae_Ilarvirus
1
NP_733823_Bromoviridae_Ilarvirus
0,75
YP_009104367_Bromoviridae_Ilarvirus
Bromoviridae
YP_941474_Bromoviridae_Ilarvirus
1
0,98
YP_002308569_Bromoviridae_Ilarvirus
0,62
NP_620772_Bromoviridae_Ilarvirus
1
YP_006446_Bromoviridae_Ilarvirus
0,98
YP_008470969_Bromoviridae_Ilarvirus
1
NP_620677_Bromoviridae_Ilarvirus
1
0,98
NP_619576_Bromoviridae_Ilarvirus
0,68
YP_001285482_Bromoviridae_Ilarvirus
1
0,91
YP_002455928_Bromoviridae_Ilarvirus
NP_613282_Bromoviridae_Ilarvirus
0,91
NP_620753_Bromoviridae_Ilarvirus
NP_049325_Virgaviridae_Tobravirus
1
NP_620669_Virgaviridae_Tobravirus
0,79
NP_620040_Virgaviridae_Pecluvirus
NP_620444_Virgaviridae_Pomovirus
0,99
0,95
YP_009220366_Virgaviridae_Pomovirus
0,62
NP_612628_Virgaviridae_Pomovirus
1
NP_659020_Virgaviridae_Furovirus
0,74
NP_059511_Virgaviridae_Furovirus
0,87
NP_059451_Virgaviridae_Furovirus
1
0,65
NP_059513_Virgaviridae_Furovirus
YP_002455904_Virgaviridae_Tobamovirus
YP_908760_Virgaviridae_Tobamovirus
1
NP_072162_Virgaviridae_Tobamovirus
0,99
NP_619684_Virgaviridae_Tobamovirus
1
NP_624336_Virgaviridae_Tobamovirus
YP_003915154_Virgaviridae_Tobamovirus
1
1
YP_009130653_Virgaviridae_Tobamovirus
Virgaviridae
YP_004956727_Virgaviridae_Tobamovirus
0,89
YP_950421_Virgaviridae_Tobamovirus
1
YP_004465358_Virgaviridae_Tobamovirus
0,57
YP_719998_Virgaviridae_Tobamovirus
1
YP_009094324_Virgaviridae_Tobamovirus
0,87
NP_056809_Virgaviridae_Tobamovirus
0,71
YP_762617_Virgaviridae_Tobamovirus
NP_046151_Virgaviridae_Tobamovirus
0,93
0,68
YP_005476600_Virgaviridae_Tobamovirus
NP_543049_Virgaviridae_Tobamovirus
1
1
NP_740757_Virgaviridae_Tobamovirus
YP_001974323_Virgaviridae_Tobamovirus
YP_008802584_Virgaviridae_Tobamovirus
0,97
YP_001333649_Virgaviridae_Tobamovirus
0,75
0,5
NP_056764_Virgaviridae_Tobamovirus
0,62
YP_009182168_Virgaviridae_Tobamovirus
1
0,59
NP_078446_Virgaviridae_Tobamovirus
YP_001041889_Virgaviridae_Tobamovirus
0.4

Figura 2 Árvore filogenética montada por máxima verossimilhança do domínio metiltransferase, destacando em cinza os GVLV e CVLV. Valores de *bootstrap* abaixo de 50% foram omitidos

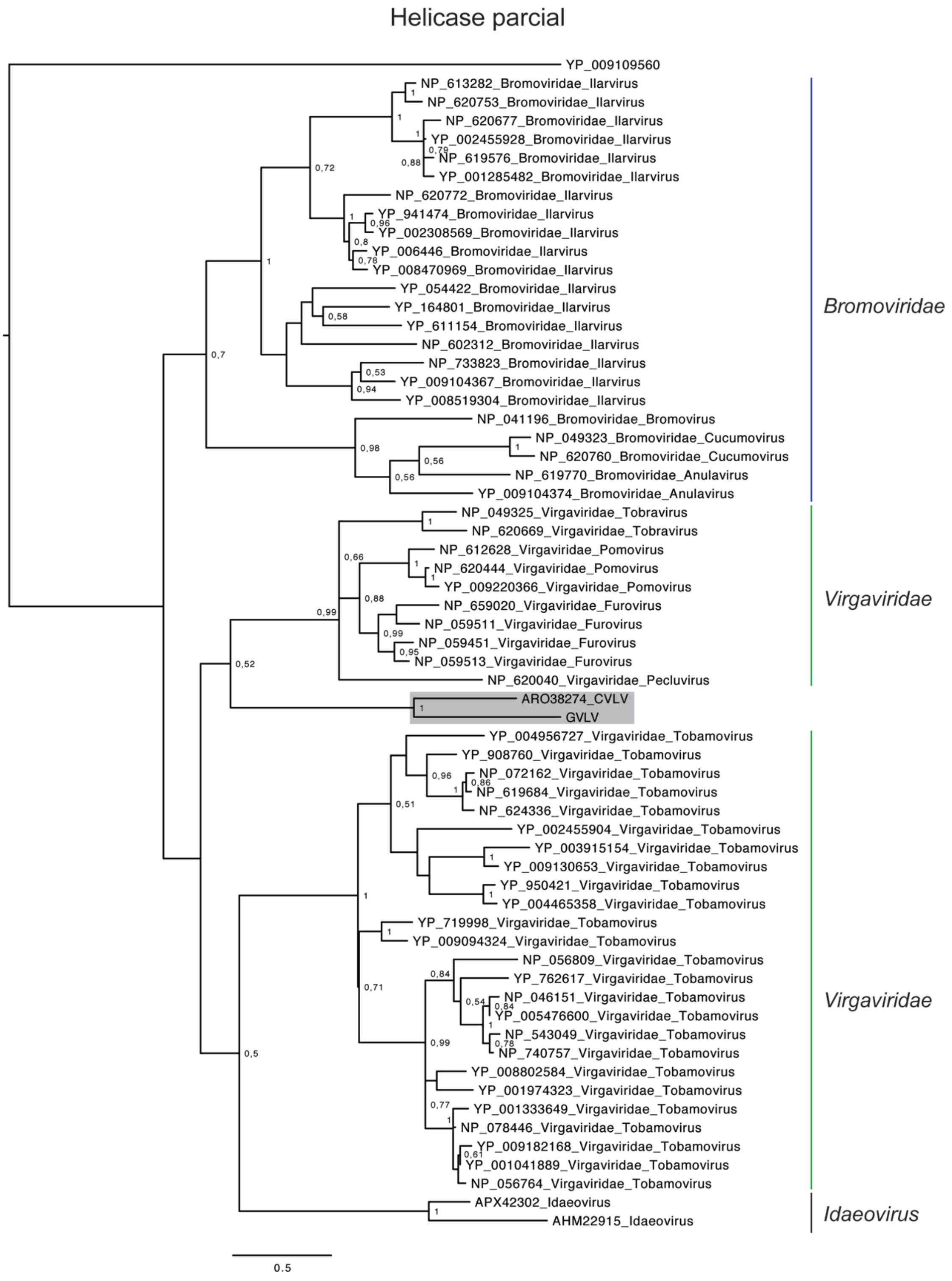

Helicase parcial
YP_009109560
NP_613282_Bromoviridae_Ilarvirus
NP_620753_Bromoviridae_Ilarvirus
NP_620677_Bromoviridae_Ilarvirus
YP_002455928_Bromoviridae_Ilarvirus
NP_619576_Bromoviridae_Ilarvirus
YP_001285482_Bromoviridae_Ilarvirus
NP_620772_Bromoviridae_Ilarvirus
YP_941474_Bromoviridae_Ilarvirus
YP_002308569_Bromoviridae_Ilarvirus
YP_006446_Bromoviridae_Ilarvirus
YP_008470969_Bromoviridae_Ilarvirus
YP_054422_Bromoviridae_Ilarvirus
YP_164801_Bromoviridae_Ilarvirus
YP_611154_Bromoviridae_Ilarvirus
NP_602312_Bromoviridae_Ilarvirus
NP_733823_Bromoviridae_Ilarvirus
YP_009104367_Bromoviridae_Ilarvirus
YP_008519304_Bromoviridae_Ilarvirus
NP_041196_Bromoviridae_Bromovirus
NP_049323_Bromoviridae_Cucumovirus
NP_620760_Bromoviridae_Cucumovirus
NP_619770_Bromoviridae_Anulavirus
YP_009104374_Bromoviridae_Anulavirus
Bromoviridae
NP_049325_Virgaviridae_Tobravirus
NP_620669_Virgaviridae_Tobravirus
NP_612628_Virgaviridae_Pomovirus
NP_620444_Virgaviridae_Pomovirus
YP_009220366_Virgaviridae_Pomovirus
NP_659020_Virgaviridae_Furovirus
NP_059511_Virgaviridae_Furovirus
NP_059451_Virgaviridae_Furovirus
NP_059513_Virgaviridae_Furovirus
NP_620040_Virgaviridae_Pecluvirus
Virgaviridae
ARO38274_CVLV
GVLV
YP_004956727_Virgaviridae_Tobamovirus
YP_908760_Virgaviridae_Tobamovirus
NP_072162_Virgaviridae_Tobamovirus
NP_619684_Virgaviridae_Tobamovirus
NP_624336_Virgaviridae_Tobamovirus
YP_002455904_Virgaviridae_Tobamovirus
YP_003915154_Virgaviridae_Tobamovirus
YP_009130653_Virgaviridae_Tobamovirus
YP_950421_Virgaviridae_Tobamovirus
YP_004465358_Virgaviridae_Tobamovirus
YP_719998_Virgaviridae_Tobamovirus
YP_009094324_Virgaviridae_Tobamovirus
NP_056809_Virgaviridae_Tobamovirus
YP_762617_Virgaviridae_Tobamovirus
NP_046151_Virgaviridae_Tobamovirus
YP_005476600_Virgaviridae_Tobamovirus
NP_543049_Virgaviridae_Tobamovirus
NP_740757_Virgaviridae_Tobamovirus
YP_008802584_Virgaviridae_Tobamovirus
YP_001974323_Virgaviridae_Tobamovirus
YP_001333649_Virgaviridae_Tobamovirus
NP_078446_Virgaviridae_Tobamovirus
YP_009182168_Virgaviridae_Tobamovirus
YP_001041889_Virgaviridae_Tobamovirus
NP_056764_Virgaviridae_Tobamovirus
Virgaviridae
APX42302_Idaeovirus
AHM22915_Idaeovirus
Idaeovirus
0.5

Figura 3 Árvore filogenética montada por máxima verossimilhança do domínio helicase parcial, destacando em cinza os GVLV e CVLV. Valores de *bootstrap* abaixo de 50% foram omitidos

5. Referências

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

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*. **215**, 403-410.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, **19**, 455 – 477.

Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bujarski, J., Figlerowicz, M., Gallitelli, D., Roossinck, M.J., Scott, S.W. (2012) Family *Bromoviridae*. In *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses*, pp. 965-976. Eds King A.M.Q., Adams M.J., Carstens E.B., Lefkowitz E.J. San Diego, USA: Elsevier Academic

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

Kumar S., Stecher G., Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**(7), 1870-1874.

Li H., Durbin R. (2010) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754 – 1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*. **25**(16), 2078-2079.

Matsumura, E.E., Coletta-Filho, H.D., Nouri, S., Falk, B.W., Nerva, L., Oliveira, T.S., Dorta, S.O., Machado, M.A (2017) Deep sequencing analysis of RNAs from citrus plants grown in a citrus sudden death-affected area reveals diverse known and putative novel viruses. *Viruses*, **9**(92). doi: 10.3390/v9040092.

Menzel P., Ng K. L., Krogh A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, **7**.

Silva J.M.F., Al Rwahnih M., Blawid R., Nagata T., Fajardo T.V.M. (2017) Discovery and molecular characterization of a novel enamovirus, Grapevine enamovirus-1. *Virus Genes*, 1-5.

Valverde R.A., Dodds J.A., Heick J.A. (1986) Double-stranded ribonucleic acid from plants infected with viruses having elongated particles and undivided genomes. *Phytopathology*, **76**, 459–465.**Zerbino D.R., Birney E. (2008**) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Capítulo 5. Estudo da diversidade intra-hospedeiro dos vírus associados ao enrolamento da folha da videira

1. Resumo

O enrolamento das folhas da videira é causada por vários vírus pertencentes à família *Closteroviridae*, designados vírus associados ao enrolamento da folha da videira (*Grapevine leafroll-associated viruses* – GLRaVs), e é reconhecida como uma das doenças mais destrutivas que acomete esse gênero. Nesse estudo, utilizamos dados gerados por HTS para estudar a diversidade genética intra-hospedeiro da poliproteína (ORF 1a) dos GLRaV-2 e -3 (gêneros *Closterovirus* e *Ampelovirus*, respectivamente) e detectar uma molécula defectiva de RNA (dRNA) de um GLRaV-4 *strain* 5. Dois isolados de GLRaV-2 (S1-ISA e S19-CF) apresentaram uma alta diversidade intra-hospedeiro se comparados com dois isolados de GLRaV-3 (S1-ISA e S17-TC). O consenso da poliproteína dos dois isolados de GLRaV-2 difere em apenas 14 posições, e apresentam 99,7% de identidade entre si a nível de amino ácido. Entretanto, mais de 70 polimorfismos de único nucleotídeo (*single nucleotide polymorphisms* – SNPs) foram identificados em cada isolado, que compartilham 9 SNPs em comum. Apenas 21 e 12 SNPs foram identificados em cada isolado de GLRaV-3 (S1-ISA e S17-TC, respectivamente). A porcentagem de SNPs não-sinônimos (≤ 50%) encontrados na poliproteína dos 4 isolados avaliados (GLRaV-2 S1-ISA e S19-CF e GLRaV-3 S1-ISA e S19-CF) sugere uma seleção negativa agindo sobre essas populações intra-hospedeiro, considerando que na ausência de pressão evolutiva (seleção neutra), a maioria das mutações em qualquer códon do código genético resultam em mutações não-sinônimas. O dRNA do GLRaV-4 *strain* 5 encontrado é composto pelo complexo da replicase e uma porção da extremidade 3' que contém a capa proteíca divergida (*diverged coat protein* – CPd) truncada.

2. Introdução

O enrolamento das folhas da videira (GLRD) é reconhecida como uma das doenças mais destrutivas de videiras, e é provavelmente a doença mais difundida entre esse gênero (Martelli e Boudon-Padieu, 2006). Os vírus associados ao enrolamento da folha da videira (*Grapevine leafroll-associated viruses* – GLRaVs) pertencem à família *Closteroviridae*, sendo os GLRaVs-1, -3 e -4 pertencentes ao gênero *Ampelovirus* e os GLRaVs -2 e -7 pertencentes aos gêneros *Closterovirus* e *Velarivirus*, respectivamente (Maliogka *et al*., 2015). A família *Closteroviridae* é composta por vírus de ssRNA(+) e possuem genomas de aproximadamente 15 a 20 kb. Os GLRaVs pertencentes ao gênero *Ampelovirus* são transmitidos principalmente por cochonilhas, enquanto o transmissor dos GLRaVs-2 e -7 não são conhecidos (Maliogka *et al*., 2015). O complexo da replicase desses vírus é formado por uma poliproteína, que contém os domínios protease (Pro), metiltranferase (Met) e helicase (Hel), fusionada a um domínio de RNA polimerase dependente de RNA (RdRp) expresso através de uma mudança no quadro de leitura (*frameshift* +1) da poliproteína (Rubio *et al*., 2013).

Devido à alta taxa de erro de replicação do RNA viral e ao tamanho do genoma dos vírus pertencentes à família *Closteroviridae*, estima-se que pelo menos duas mudanças de nucleotídeo sejam introduzidas a cada replicação do genoma (Bar-Joseph e Mawassi, 2013; Lauring e Andino, 2010). Eventos de recombinação aumentam a complexidade da diversidade populacional intra-hospedeiro, e são responsáveis pela formação de novos genótipos e de moléculas defectivas de RNA (dRNA), além de terem permitido a aquisição de novos genes através da recombinação com mRNAs ou outros genomas virais (Rubio *et al*., 2013). A diversidade genética intra-hospedeiro (Ayllón *et al*., 1999; Lozano *et al*., 2009), a estabilidade genética de um único consenso ao longo do tempo (Albiach-Martí *et al*., 2000) e a baixa diversidade genética entre isolados de um mesmo genogrupo (Rubio *et al*., 2013) sugerem tanto a adequação do modelo de *quasispecies* dentro da família *Closteroviridae* quanto uma forte seleção negativa agindo sobre esses vírus (Holmes e Moya, 2002). Estudos sobre a diversidade intra-hospedeiro dos GLRaVs são limitados, onde até então a diversidade intra-hospedeiro desses importantes patógenos de videiras foi estudada apenas por métodos de baixa acurácia como polimorfismo de conformação de fita simples (SSCP), ensaios de mobilidade de DNA heteroduplex (HMA) ou clonagem molecular seguida por sequenciamento de um baixo número de clones (Turturo *et al*., 2005; Jarungula *et al*., 2010; Bertazzon *et al*., 2010; Esteves *et al*., 2012). Nesse trabalho, a diversdade genética intra-hospedeiro da poliproteína de dois isolados de GLRaV-2 (S1-ISA e S19-CF) e dois isolados de GLRaV-3 (S1-ISA e S17-TC) foi analisada a partir de dados gerados pela plataforma Illumina. Adicionalmente, reportamos a detecção *in silico* de uma molécula defectiva de RNA do GLRaV-4 *strain* 5.

3. Metodologia

3.1. Material vegetal, preparo das bibliotecas de DNA e sequenciamento

RNA fita-dupla (dsRNA) foi extraído de 17 plantas coletadas de diferentes coleções das regiões sul, sudeste e nordeste do Brasil. As plantas formas propagadas vegetativamente em casa de vegetação antes da extração de dsRNA seguindo o protocolo de Valverde (1986). As bibliotecas de cDNA (TruSeq RNA) foram montadas pela Macrogen (Seul, Coréia do Sul) ou Eurofins Genomics (Huntsville, EUA), e o sequenciamento foi feito pela plataforma Illumina HiSeq 2000 (100 bp paied-end).

3.2. Alinhamento e análise de variantes

Os GLRaVs presentes nas amostras analisadas foram previamente descritos (Fajardo *et al*., 2017). Os *reads* foram alinhados contra as sequências de referência com o programa BWA (Li e Durbin, 2010). *Reads* duplicados ou não alinhandos foram removidos com o SAMtools (Li *et al*., 2009). A partir dessa análise, os *datasets* com as maiores coberturas ao longo dos genomas dos GLRaVs foram selecionados para a análise de variantes, que foi feita com o programa LoFreq (Wilm *et al.*, 2012). Esse programa utiliza a qualidade Phred de cada variante para modelar uma distribuição binomial da taxa de erro de sequenciamento e distinguir variantes reais de artefatos. Entretanto, esse programa não consegue distinguir variantes reais de artefatos introduzidos durante as etapas de RT e PCR da montagem da biblioteca de cDNA, portanto, para minimizar o número de falsos positivos, apenas SNPs presentes em regiões com cobertura acima de 35 x e com frequência acima de 5% foram analisados. Em cada posição, a diversidade expressa pela entropia de Shannon (Schneider *et al.*, 1986) foi calculada a partir da equação: $H(i) = -\sum_{i=a}^{t} f(b,i) \log_2 f(b,i)$; onde b é uma das bases (a, g, c ou t), *H(i)* é a entropia de Shannon (ou incerteza) na posição i e *f(b,i)* é a frequência da base b na posição i. A entropia em cada posição varia de 0 *bits*, quando não há variantes, a 2 *bits*, quando as quatro bases são equiprováveis (onde a frequência de cada variante é de 25%). Para cada SNP, foi determinado o tipo de mutação (sinônima ou não-sinômina) causada, onde dois SNPs em um mesmo códon foram considerados independentemente.

3.3. Estimativa da diversidade global da poliproteína

Dois métodos foram utilizados para estimar a diversidade global da poliproteína: *Pn*, calculada através do número total de sítios polimórficos divido pelo tamanho da poliproteína (Cabot *et al.*, 1999); e *Sn*, calculada pela soma da entropia de Shannon de cada SNP normalizada pela entropia máxima da sequência da poliproteína, obtida pela equação *Sn = St/2p*, onde *St* é a soma da entropia de Shannon de cada posição e *p* corresponde ao tamanho da poliproteína, excetuando o códon de terminação. Em ambos os casos, a diversidade global varia de 0 (sem diversidade) a 1 (máxima diversidade). Entretanto, a frequência das bases alternativas é considerada apenas por *Sn*.

3.4. Estimativa da pressão seletiva (dN/dS)

A pressão seletiva agindo a poliproteína dos GLRaV-2 e -3 foi estimada com o programa codeml do pacote paml v4.8 (Yang, 2007). A pressão seletiva foi avaliada atráves do número de substituições não-sinônimas por sítios não-sinônimos pelo número de substituições sinônimas por sítios sinônimos (dN/dS). Alinhamentos da poliproteína foram realizados com sequências obtidas pelo GenBank, onde 7 sequências completas foram obtidas para o GLRaV-2 e 15 foram obtidas para o GLRaV-3. Para determinar se os dados utilizados são sensíveis às suposições feitas sobre os modelos, e para calacular os valores globais de dN/dS, corridas preliminares foram feitas usando o modelo mais básico do codeml (model = 0, NSsites = 0; modelo M0), testando diferentes frequências de códon e fixando ou estimando *kappa* (taxa de transição por transversão). Após os testes preliminares, o modelo de seleção por sítio (model = 0, NSsites = 7 8) (Nielsen e Yang 1998) foi utilizado, e testes da razão da verossimilhança foram realizados entre os modelos M7 (evolução neutra) e M8 (seleção positiva) para identificar sítios possivelmente selecionados positivamente (Tabela 1). Os valores da média da distribuição posterior de *omega* (dN/dS) para cada sítio da poliproteína dos GLRaV-2 e -3 foram obtidos pela abordagem *Bayes empirical Bayes* (BEB) implementada no modelo M8 (Yang *et al.*, 2005)

Tabela 1 Testes da razão da verossimilhança entre os modelos M7 (evolução neutra) e M8 (seleção positiva)

Vírus	Modelo	Log-verossimilhança (lnL)	Nível de significância (M7-M8)
GLRaV-2	M7	-29307,685974	0,0146383
	M8	-29303,461859	
GLRaV-3	M7	-25374,309180	0,0039268
	M8	-25368,769237	

3.5. Detecção *in silico* de moléculas defectivas de RNA (dRNA)

Para a detecção de dRNAs a partir dos dados Illumina foi usado o programa ViReMa (Routh e Johnson, 2013). Esse programa utiliza o Bowtie (Langmead *et al.*, 2009) para identificar *reads* recombinantes que se alinham em mais de uma região do genoma de referência. Devido a um número alto de falsos positivos causados pela atividade de troca de fita molde das enzimas usadas nas reações de RT, resultando em micro deleções e inserções (microindel) (Görzer *et al.*, 2010), uma segunda etapa *in silico* foi feita para confirmar a presença de dRNAs. Nessa etapa os *reads* foram alinhados contra o genoma de referência com o BWA e a cobertura dos *reads* ao longo do genoma foi avaliada. Caso dRNAs estejam presentes, a cobertura de *reads* da região deletada será inferior se comparada com o resto do genoma. Essa análise revelou a presença de um dRNA de um GLRaV-4 *strain* 5 em uma videira *V. vinifera* cv. Trajadura (S18-TRAJ) exibindo infecções virais múltiplas.

4. Resultados e discussão

Estudos sobre a evolução dos vírus pertencentes à família *Closteroviridae* sugerem que a estrura genética e diversificação desses vírus está relacionada à ação de uma forte seleção negativa, eventos de recombinação e a efeitos gargalo causados pela transmissão para novos hospedeiros ou pela transmissão por vetores (Rubio *et al.*, 2013). A estabilidade de um único consenso ao longo do tempo e a baixa diversidade genética encontrada dentro dos genogrupos da família *Closteroviridae* sugerem, além de uma forte seleção negativa, a adequação do modelo de *quasispecies* para explicar a dinâmica populacional intra-hospedeiro desses vírus (Albiach-Martí *et al.*, 2000; Rubio *et al.*, 2013; Holmes e Moya, 2002). Nossos resultados corroboram com essas observações. O número total de SNPs, de SNPs sinônimos e não-sinônimos, a diversidade global (*Pn* e *Sn*) e os valores de dN/dS obtidos para a poliproteína dos GLRaV-2 e -3 estão disponíveis na Tabela 2. Os valores globais de dN/dS obtidos para a poliproteína dos GLRaV-2 e -3

(0,12679 e 0,09458, respectivamente) se aproximam dos valores obtidos para a capa proteíca (CP) por Rubio *et al.* (0,161 e 0,073 para os GLRaV-2 e -3, respectivamente; 2013), com o GLRaV-3 estando sujeito à uma maior pressão seletiva negativa. Adicionalmente, a porcentagem de SNPs não-sinônimos (≤ 50%) encontrada em ambos os isolados de GLRaV-2 e -3 sugere uma forte seleção negativa agindo sobre essas populações intra-hospedeiro.

Tabela 2 Número total de SNPs, de SNPs sinônimos e não-sinônimos, diversidade global (*Pn* e *Sn*) e valores de dN/dS da poliproteína dos GLRaV-2 e -3.

Vírus	Isolado	SNPs (total)	SNPs sinônimos	SNPs não-sinônimos	*Pn*	*Sn*	dN/dS (M0)
GLRaV-2	**S1-ISA**	78	41	37	0,00885	0,00335	0,12679
	S19-CF	76	38	38	0,00862	0,00353	
GLRaV-3	**S1-ISA**	21	11	11	0,00327	0,00089	0,09458
	S17-TC	12	7	5	0,00178	0,00047	

A diversidade genética dos dois isolados de GLRaV-2 analisados (S1-ISA e S19-CF) foi maior do que a dos dois isolados de GLRaV-3 (S1-ISA e S17-TC) (Tabela 2 e Fig. 1). Os isolados GLRaV-2 S1-ISA e S19-CF apresentaram populações intra-hospedeiro complexas e distintas, com mais de 70 SNPs identificadas em cada isolado, que compartilharam apenas 9 SNPs (~11% em cada isolado). Apesar da porcentagem de sítios polimórficos (*Pn*) da poliproteína ser maior no isolado GLRaV-2 S1-ISA do que no GLRaV-2 S19-CF (0,00885 e 0,00862, respectivamente), a diversidade global da poliproteína expressa por *Sn* foi maior no isolado S19-CF (0,00335 e 0,00353, para os isolados S1-ISA e S19-CF, respectivamente), devido a esse método considerar a frequência de cada variante. O isolado GLRaV-2 S19-CF apresentou um grande número de SNPs com aproximadamente 1 *bit* de entropia de Shannon (Fig. 1), sugerindo a presença de dois haplótipos dominantes nessa população, ambos com uma frequência de ~50%. Infecções mistas com mais de um GLRaV e infecções compostas por dois ou mais variantes divergentes são comuns entre esses vírus (Fajardo *et al.*, 2017; Rubio *et al.*, 2013; Bertazzon *et al.*, 2010), e foram observadas para outros membros da família *Closteroviridae* (Rubio *et al.*, 2013). A co-infecção com diferentes vírus ou isolados de um mesmo vírus podem ter efeitos sinestérgicos ou antagonistas, e podem alterar a distribuição de haplótipos de certos vírus (Rubio *et al.*, 2013).

SNPs não-sinônimos dos isolados GLRaV-2 S1-ISA e S19-CF e GLRaV-3 S1-ISA e S17-TC tiveram, em geral, valores médios da distribuição posterior de *omega* maiores do que SNPs sinônimos (Fig. 2), indicando que devido à seleção negativa, SNPs não-sinônimos ocorrem com maior frequência em sítios neutralmente ou, possivelmente, positivamente selecionados. Através de testes da razão da verossimilhança entre os modelos M7 (evolução neutra) e M8 (seleção positivos) não foi possível determinar, com uma significancia alta (p ≤ 0,001), sítios positivamente selecionados (Tabela 1). Entretanto, esses dados corroboram com a hipótese de que a emergência de alguns novos isolados esteja relacionada a efeitos gargalos causados pela transmissão para novos hospedeiros ou pela transmissão por vetores, considerando que SNPs não-sinônimos na população intra-hospedeiro ocorreram com maior frequência em sítios que estão sujeitos à seleção neutra ou positiva entre diferentes isolados desses vírus. No modelo *Citrus tristeza virus* (CTV; gênero *Closterovirus*), a transmissão do isolado T317 de cidra (*Citrus medica* L.) para laranjeira (*Citrus sinensis* L.) originou, em duas ocasiões distintas, os novos isolados T318 e T317D, onde foi observado uma mudança na estrutura populacional intra-hospedeiro desses isolados (Rubio *et al.*, 2000). A alta diversidade genética intra-hospedeiro encontrada nos dois isolados de GLRaV-2 pode facilitar a transmissão para novos hospedeiros, e sugere a presença de dois ou mais variantes divergentes, enquanto as populações intra-hospedeiros dos dois isolados de GLRaV-3 analisados são compostas por um haplótipo dominante e variantes de menor frequência (Fig. 1).

Eventos de recombinação aumentam a complexidade das populações virais intra-hospedeiro, dando origem a novos haplótipos e isolados, e são responsáveis pela formação de dRNAs (Rubio *et al.*, 2013). Essas formas deletérias do RNA genômico retém os elementos regulatórios necessários para a sua replicação, mas necessitam da sequência parental para prover funções essenciais para a infecção e replicação do dRNA, como encapsidação, replicação e movimento sistêmico (Rubio *et al.*, 2013; Bar-Joseph e Mawassi, 2013). Ao utilizar dados gerados por HTS, nós identificamos um dRNA de um GLRaV-4 *strain* 5 em uma videira *V. vinifera* cv. Trajadura co-infectada com GLRaV-3, *Grapevine Red Globe virus* (GRGV), *Grapevine Syrah virus 1* (GSyV-1), *Grapevine rupestris stem pitting-associated virus* (GRSPaV), *Grapevine virus A* (GVA), *Grapevine fleck virus* (GFkV) and *Grapevine rupestris vein feathering virus* (GRVFV) (Fajardo *et al.*, 2017), que podem estar facilitando a infecção do dRNA de GLRaV-4 *strain* 5 através da supressão da resposta de defesa do hospedeiro. Esse dRNA é composto pelo complexo de replicase e uma porção da extremidade 3' que contém a capa proteica divergida (*diverged coat protein* – CPd) truncada (Fig. 3). O sítio de junção desse dRNA está localizado nas posições 8785 e 13313 do RNA genômico (número de acesso do GenBank: KX828702). dRNAs similares, compostos pelo complexo replicase e uma porção da extremidade 3' contendo uma proteína truncada, foram descritos para o closterovírus CTV. Esses dRNAs são eficientemente transmitidos mecanicamente para citros e protoplastos de *Nicotiana benthamiana*, e são provavelmente capazes de auto-replicação (Che *et al.*, 2002). Um alto números de SNPs foi anotado ao longo do genoma do GLRaV-4 *strain* 5 (Fig. 3), sugerindo uma alta diversidade intra-hospedeiro, porém, não é possível determinar se esses variantes estão presentes no dRNA ou RNA genômico do GLRaV-4 *strain* 5.

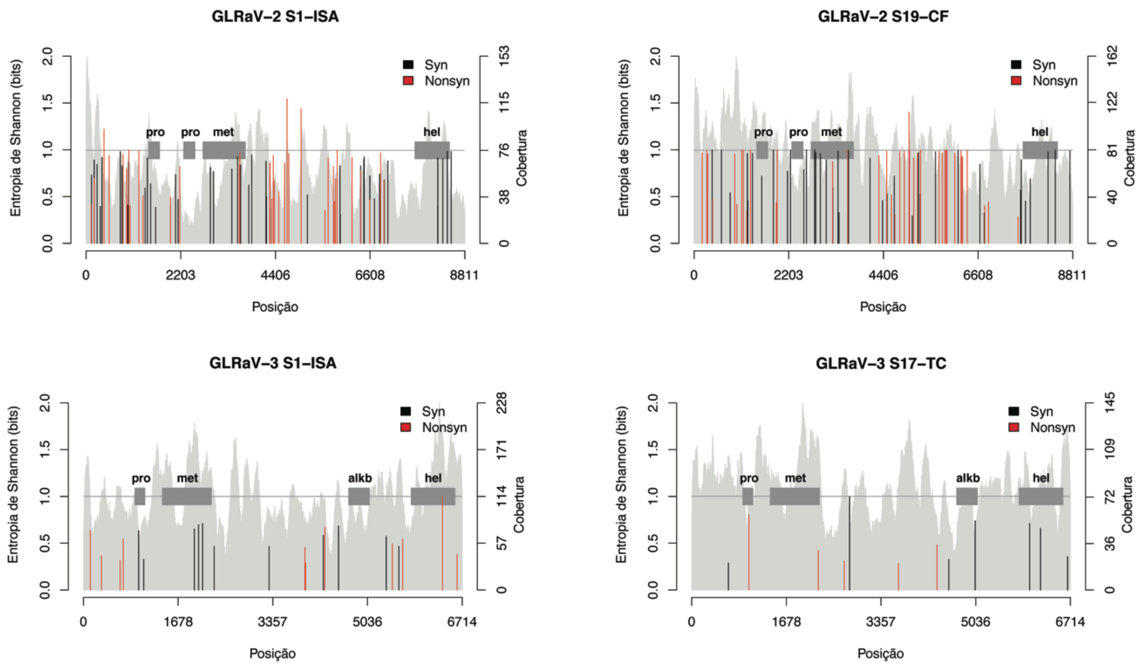


Figura 1 Gráficos da diversidade (entropia de Shannon) de cada SNP e da cobertura ao longo da poliproteína (cinza ao fundo) dos dois isolados de GLRaV-2 (S1-ISA e S19-CF) e GLRaV-3 (S1-ISA e S17-TC), destacando em vermelho os SNPs não-sinônimos e em preto SNPs sinônimos; e posição dos domínios protease (pro), metiltransferase (met), helicase (hel) e AlkB (alkb)

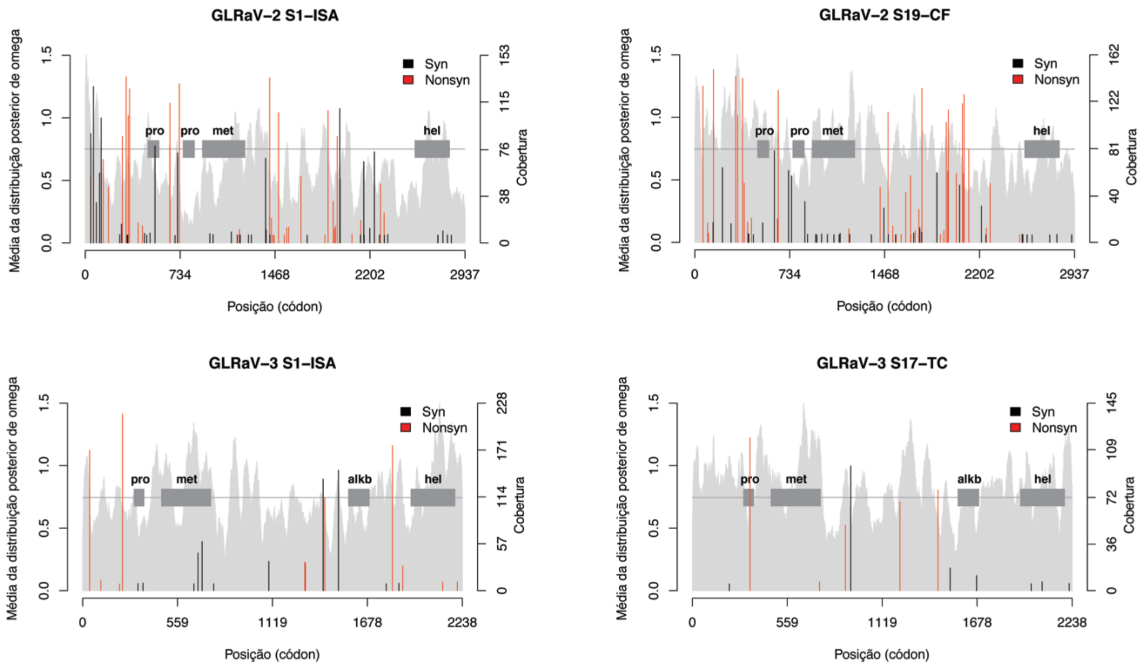


Figura 2 Gráficos da média da distribuição posterior de *omega* (dN/dS) de cada SNP e da cobertura ao longo da poliproteína (cinza ao fundo) dos dois isolados de GLRaV-2 (S1-ISA e S19-CF) e GLRaV-3 (S1-ISA e S17-TC), destacando em vermelho os SNPs

não-sinônimos e em preto SNPs sinônimos; e posição dos domínios protease (pro), metiltransferase (met), helicase (hel) e AlkB (alkb)

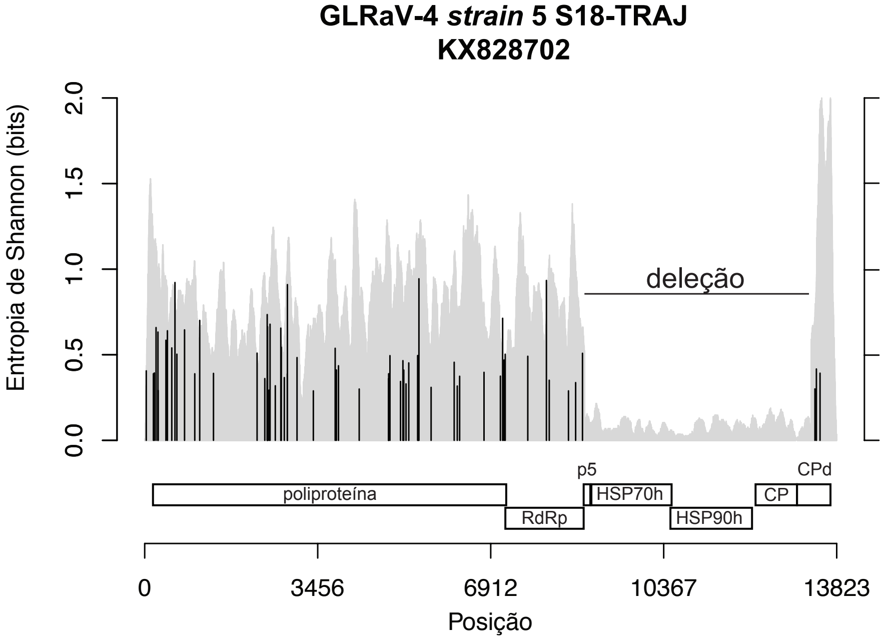


Figura 3 Gráfico da diversidade (entropia de Shannon) em cada SNP, cobertura ao longo do genoma (cinza ao fundo) e organização genômica do GLRaV-4 *strain* 5 (KX828702), mostrando a poliproteína, RNA polimerase dependente de RNA (RdRp), p5, HSP70h, HSP90h, capa proteíca (CP) e capa proteíca divergida (CPd). A região deletada no dRNA está destacada

5. Referências

Ayllón M.A., Rubio L., Moya A., Guerri J., Moreno P. (1999) The haplotype distribution of two genes of citrus tristeza virus is altered after host change or aphid transmission. *Virology*, **255**(1), 32-39.

Albiach-Martí M.R., Mawassi M., Gowda S., Satyanarayana T., Hilf M.E., Shanker S., Almira E.C., Vives M.C., López C., Guerri J., Flores R. (2000) Sequences of citrus tristeza virusseparated in time and space are essentially identical. *Journal of Virology*, **74**(15), 6856-6865.

Bar-Joseph M., Mawassi M. (2013) The defective RNAs of Closteroviridae. *Frontiers in Microbiology*, **4**, 132.

Bertazzon N., Borgo M., Vanin S., Angelini E. (2010) Genetic variability and pathological properties of Grapevine leafroll-associated virus 2 isolates. *European Journal of Plant Pathology*, **127**(2), 185-197.

Cabot B., Martell M., Esteban J.I., Sauleda S., Otero T., Esteban R., Guàrdia J., Gómez J. (2000) Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. *Journal of Virology*, **74**(2), 805-811.

Che X., Mawassi M., Bar-Joseph M. (2002) A novel class of large and infectious defective RNAs of Citrus tristeza virus. *Virology*, **298**(1), 133-145.

Esteves F., Santos M.T., Eiras-Dias J.E., Fonseca F. (2012) Occurrence of grapevine leafroll-associated virus 5 in Portugal: genetic variability and population structure in field-grown grapevines. *Archives of Virology*, **157**(9), 1747-1765.

Fajardo T.V.M., Silva F.N., Eiras M., Nickel O. (2017) High-throughput sequencing applied for the identification of viruses infecting grapevines in Brazil and genetic variability analysis. *Tropical Plant Pathology* doi:10.1007/s40858-017-0142-8

Görzer I., Guelly C., Trajanoski S., Puchhammer-Stöckl E. (2010) The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. *Journal of Virological Methods*, **169**(1), 248-252.

Holmes E.C., Moya A. (2002) Is the quasispecies concept relevant to RNA viruses? *Journal of Virology,* **76**(1), 460-465.

Jarugula S., Alabi O.J., Martin R.R., Naidu R.A. (2010) Genetic variability of natural populations of Grapevine leafroll-associated virus 2 in Pacific Northwest vineyards. *Phytopathology*, **100**(7), 698-707.

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

Lauring A.S., Andino R. (2010) Quasispecies theory and the behavior of RNA viruses. *PLoS pathogens*, 6(7), e1001005.

Li H., Durbin R. (2010) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754 – 1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078-2079.

Lozano G., Grande-Pérez A., Navas-Castillo, J. (2009) Populations of genomic RNAs devoted to the replication or spread of a bipartite plant virus differ in genetic structure. *Journal of Virology*, **83**(24), 12973-12983.

Martelli G.P., Boudon-Padieu E. (2006) Directory of infectious diseases of grapevines, pp. 9–194. Eds. Martelli G.P., Boudon-Padieu E. Options Méditerranéennes Série B: Vol. 55. Bari, Italy: CIHEAM-IAMB.

Maliogka V.I., Martelli G.P., Fuchs M., Katis N.I. (2015) Control of viruses infecting grapevine. Adv *Virus Research*, **91**, 175-227.

Nielsen R., Yang Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**(3), 929-936.

Routh A., Johnson J.E. (2013) Discovery of functional genomic motifs in viruses with ViReMa–a Virus Recombination Mapper–for analysis of next-generation sequencing data. *Nucleic Acids Research*, **42**(2), e11-e11.

Rubio L., Guerri J., Moreno P. (2000) Characterization of Citrus tristeza virus isolates by single-strand conformation polymorphism analysis of DNA complementary to their RNA population. In *Proceedings of* *the 14th Conference of the International Organization of Citrus Virologists, Vol. 14,* pp 12-17. Eds da Graça J.V., Lee R.F., Yokomi R.K. Riverside, CA: IOCV

Rubio L., Guerri J., Moreno P. (2013) Genetic variability and evolutionary dynamics of viruses of the family *Closteroviridae*. *Frontiers in Microbiology*, **4**, 151.

Schneider T.D., Stormo G.D., Gold L., Ehrenfeucht A. (1986) Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, **188**(3), 415-431.

Turturo C., Saldarelli P., Yafeng D., Digiaro M., Minafra A., Savino V., Martelli G.P. (2005) Genetic variability and population structure of Grapevine leafroll-associated virus 3 isolates. Journal of General *Virology*, **86**(1), 217-224.

Wilm A., Aw P.P., Bertrand D., Yeo G.H., Ong S.H., Wong C.H., Khor C.C., Petric R., Hibberd M.L., Nagarajan N. (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, **40**(22): 11189-11201.

Yang Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8), 1586-1591.

Yang Z., Wong W.S., Nielsen R. (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, **22**(4), 1107-1118.

Capítulo 6. Discussão geral

Devido ao barateamento das tecnologias de HTS, essa técnica tem sido extensamente utilizada em estudos metagenômicos ao longo da última década. No entanto, a alta cobertura e pequeno tamanho dos *reads* proporcionados pelas plataformas de HTS mais utilizadas impõem desafios durante o processamento e análise dos dados. No capítulo 2 revisamos o uso de HTS para a descoberta de novos vírus em plantas, e descrevemos um *pipeline* básico para a análise de comunidades virais presentes em amostras de tecido vegetal a partir de dados gerados por HTS, usando o modelo *Cucumis sativus* para apresentar um estudo de caso. Cobrimos nesse trabalho: os principais métodos de enriquecimento de sequências virais, com foco na extração de RNA a partir de partículas virais semi-purificadas; as plataformas mais utilizadas para estudos metagenômicos, com foco na plataforma Illumina, que tem sido a mais utilizada (Blawid *et al.*, 2017; Breitwieser *et al.*, 2017); os principais *softwares* usados para a análise de qualidade e pre-processamento dos dados (*trimming* e remoção de sequências de adaptadores); principais *softwares* de código aberto para a montagem *de novo* dos *contigs*; e relação taxonômica dos *contigs* através de buscas por similaridades. Apresentamos nesse trabalho um estudo de caso utilizando dados gerados pela plataforma Illumina HiSeq 2000 a partir da extração de RNA de partículas virais semi-purificadas, testando vários montadores e parâmetros, e mostramos que os resultados obtidos variam consideravelmente.

As principais características de *datasets* gerados para estudos metagenômicos são a cobertura desigual entre os diferentes vírus presentes e o alto grau de polimorfismo. Montadores baseados em grafos *de Bruijn* que utilizam apenas um valor de *k* como o Velvet (Zerbino e Birney, 2008) e ABySS (Simpson *et al.*, 2009) foram desenhados para a montagem de genomas eucarióticos, e se baseam na cobertura de *k*-mers para identificar e remover erros, pois eles assumem uma cobertura uniforme ao longo do genoma. Sem os parâmetros optimizados para análises metagenômicas, esses montadores acabam podem descartar sequências virais de baixa cobertura, e mostramos que ao executar o Velvet sem os paramêtros *cov_cutoff* e *exp_cov* gerou um maior número de *contigs* de baixa cobertura. Entretando, essa montagem resultou em um grande número de *contigs* fragmentados, complicando a análise dos resultados devido ao alto número de *hits*. Os montadores baseados em grafos *de Bruijn* iterativos, que usam mais de um valor de *k* para a montagem, como o SPAdes (Bankevich *et al.*, 2012), IDBA-UD (Peng *et al.*, 2012) e MEGAHIT (Li *et al.*, 2015) são os mais recomendados para análises metagenômicas. Em suma, existe uma relação custo-benefício entre a sensibilidade e a acurácia de detecção de sequências virais quanto aos *softwares* e parâmetros utilizados nas análises bioinformáticas. Os montadores *de novo* que geraram os maiores *contigs* virais (MEGAHIT, SPAdes e IDBA-UD) acabaram por descartar sequências virais de baixa cobertura, enquanto o montador que teve o maior número de famílias virais detectadas (Velvet) produziu apenas *contigs* pequenos. O tamanho do *contig*, além de poder facilitar a montagem do genôma viral completo, influencia o tamanho do alinhamento nas buscas por similaridade, onde *contigs* pequenos podem dificultar a identificação de sequências virais genuínas.

Seguindo adaptações do *pipeline* descrito pelo trabalho acima, identificamos dois novos vírus em videiras (Capítulos 3 e 4): Grapevine enamovirus-1 (GEV-1), encontrado em 4 cultivares distintos de videira (Cabernet Sauvignon, CG 90450, Semillon e Cabernet franc); e Grapevine virga-like virus (GVLV), encontrado em baixa cobertura em 3 cultivares distintos (*Vitis flexuosa*, Semillon e Cabernet franc). Videiras comumente exibem patossistemas complexos, onde infecções mistas foram observadas em todos os casos acima. O genoma quase completo do GEV-1 foi obtido (6227 bp), possibilitando a sua classificação como um membro do gênero *Enamovirus* (família *Luteoviridae*). Interessantemente, o domínio *F-box-like* responsável pela supressão da maquinaria de RNAi do hospedeiro encontrado na proteíno P0 dos vírus pertencentes aos gêneros *Enamo-* e *Polerovirus* (Fusaro *et al.*, 2012, Pazhouhandeh *et al.*, 2006) não é conservado na P0 do GEV-1. Isso sugere uma adaptação à videira ou, caso esse domínio não seja funcional, que o GEV-1 é dependente da co-infecção com outros vírus que vão prover a função de supressão de silenciamento do hospedeiro ou que essa função é fornecida por um outro domínio ou proteína do próprio GEV-1. O genoma do GVLV foi parcialmente sequenciado (4620 bp). Esse vírus mostra baixa similaridade com vírus pertencente às famílias *Virgaviridae* e *Bromoviridae*. Essas famílias são similares quanto à organização genômica, porém, os vírus pertencentes à família *Virgaviridae* formam partículas virais alongadas, e os que pertencem à família *Bromoviridae* formam partículas icosaédricas ou em forma de bacilo (Adams *et al.*, 2009; Bujarski *et al.*, 2012). Análises filogenéticas dos domínios metiltranferase e helicase posicionaram o GLVL, respectivamente, como grupo externo às famílias *Virgaviridae* e *Bromoviridae* ou dentro da família *Virgaviridae*, porém com baixo suporte estatístico. Para uma caracterização completa do GVLV pretendemos visualizar as partículas virais por microscopia eletrônica de transmissão (MET) e sequenciar o seu genoma completo com o apoio de dados gerados por HTS a partir da extração de RNA de partículas virais semi-purificadas.

No Capítulo 5 utilizamos dados Illumina para estudar a diversidade intra-hospedeiro dos vírus associados ao enrolamento da folha da videira (*Grapevine leafroll-associated virus* – GLRaV). Nesse estudo, dois isolados de GLRaV-2 (S1-ISA e S19-CF; gênero *Closterovirus*) apresentaram uma diversidade intra-hospedeiro substancialmente maior do os que dois isolados de GLRaV-3 (S1-ISA e S17-TC) analisados, com mais de 70 polimorfismos de único nucleotídeo (*single nucleotide polymorphisms* - SNPs) identificados em cada isolado de GLRaV-2. Esses dados sugerem que devido à propagação vegetativa, os dois isolados de GLRaV-2 estudados acumularam mais mutações ao longo do tempo, enquanto efeitos gargalos causados pela transmissão pelo inseto vetor dos dois isolados de GLRaV-3 fizaram que esses isolados apresentassem uma menor diversidade genética intra-hospedeiro. O consenso da poliproteína dos GLRaV-2 S1-ISA e S19-CF difere em apenas 11 posições (99,7% de identidade a nível de amino ácido), entretanto, esses vírus apresentaram populações intra-hospedeiro complexas e distintas, com apenas 9 SNPs em comum (~11% em cada isolado). Em todos os 4 isolados analisados (GLRaV-2 S1-ISA e S19-CF e GLRaV-3 S1-ISA e S17-TC), SNPs não-sinônimos obtiveram, em geral, valores médios da distribuição posterior de dN/dS maior do que SNPs sinônimos, indicando que devido à seleção negativa, SNPs não-sinônimos ocorrem com maior frequência em sítios sujeitos à seleção neutra ou positiva. Adicionalmente, dectectamos uma molécula defectiva de RNA (dRNA) de um GLRaV-4 *strain* 5 em uma videira multi-infectada. Esse dRNA é composto pelo complexo replicase mais uma porção da extremidade 3' que contém a capa proteica divergida (*diverged coat protein* - CPd) truncada, similar a dRNAs de classe III do closterovírus *Citrus tristeza virus* (CTV) (Bar-Joseph e Mawassi, 2013; Che *et al.*, 2002).

As tecnologias de HTS impulsionaram a descoberta de novos vírus na última década, além de permitirem estudos aprofundados sobre as interações patógeno-hospedeiro e a diversidade genética intra-hospedeiro apresentada pelos vírus de RNA (Blaswid *et al.*, 2017; Breitwieser *et al.*, 2017; Parameswaran *et al.*, 2017; Posada-Cespedes *et al.*, 2017). Ao descrever novos vírus atráves de HTS, fatores importantes como a patogenicidade, modo de trasmissão, prevalência no campo e amplitude de hospedeiros muitas vezes não podem ser avaliados, realçando que a técnica de HTS, apesar de poderosa, é complementar aos ensaios tradicionalmente empregados para a caracterização de novos patógenos.

Referências

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

Bar-Joseph M., Mawassi M. (2013) The defective RNAs of Closteroviridae. *Frontiers in Microbiology*, **4**, 132.

Blawid R., Silva J.M.F., Nagata T. (2017) Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Annals of Applied Biology*, **170**(3), 301-314.

Breitwieser F.P., Lu J., Salzberg S.L. (2017) A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

Bujarski, J., Figlerowicz, M., Gallitelli, D., Roossinck, M.J., Scott, S.W. (2012) Family *Bromoviridae*. In *Virus Taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses*, pp. 965-976. Eds King A.M.Q., Adams M.J., Carstens E.B., Lefkowitz E.J. San Diego, USA: Elsevier Academic

Che X., Mawassi M., Bar-Joseph M. (2002) A novel class of large and infectious defective RNAs of Citrus tristeza virus. *Virology*, **298**(1), 133-145.

Fusaro A.F., Correa R.L., Nakasugi K., Jackson C., Kawchuk L., Vaslin M.F., Waterhouse P. M. (2012) The *Enamovirus* P0 protein is a silencing suppressor which inhibits local and systemic RNA silencing through AGO1 degradation. *Virology*, **426**(2), 178-187.

Li D., Liu C.M., Luo R., Sadakane K., Lam T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.

Parameswaran P., Sklan E., Wilkins C., Burgon T., Samuel M.A., Lu R., Ansel K.M., Heissmeyer V., Einav S., Jackson W., Doukas T. (2010) Six RNA viruses and forty-one hosts: viral small RNAs and modulation of small RNA repertoires in vertebrate and invertebrate systems. *PLoS pathogens*, **6**(2), p.e1000764.

Pazhouhandeh M., Dieterle M., Marrocco K., Lechner E., Berry B., Brault V., Hemmer O., Kretsch T., Richards K.E., Genschik P., Ziegler-Graff V. (2006) F-box-like domain in the polerovirus protein P0 is required for silencing suppressor function. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(6), 1994-1999.

Peng Y., Leung H.C.M., Yiu S.M., Chin F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Posada-Cespedes S., Seifert D., Beerenwinkel N. (2017) Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Research*, **239**, 17-32.

Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Zerbino D.R., Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.