UNIVERSIDADE DE SÃO PAULO

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

**Análise computacional da diversidade viral presente na
comunidade microbiana do processo de compostagem do
Zoológico de São Paulo**

Dissertação de Mestrado

DEYVID EMANUEL AMGARTEN

ORIENTADOR: JOÃO CARLOS SETUBAL

COORIENTADORA: ALINE MARIA DA SILVA

São Paulo, SP
Outubro de 2016

O estudo da diversidade viral em amostras ambientais tem se tornado cada vez mais importante devido a funções-chave desempenhadas por esses organismos. Estudos recentes têm fornecido evidências de que vírus de bactérias (bacteriófagos) podem ser os principais determinantes em ciclos biogeoquímicos de grandes ecossistemas, além de atuarem no fluxo de genes entre comunidades ambientais e na plasticidade funcional das mesmas frente a estresses ambientais. Neste trabalho, propomos a investigação e caracterização da diversidade viral presente em amostras de compostagem através de abordagens não dependentes e dependentes de cultivo. Na primeira abordagem, coletamos amostras seriadas de uma unidade de compostagem do zoológico de São Paulo para realização de sequenciamento metagenômico. O conjunto de sequências gerado foi extensivamente minerado (*data-mining*) para a produção de resultados de diversidade e abundância de táxons virais ao longo do processo de compostagem. Adicionalmente, procedemos com a montagem e recuperação de sequências virais candidatas a genomas completos e/ou parciais de novos vírus ambientais. Os dois protocolos computacionais utilizados para a mineração de dados encontram-se definidos e automatizados, podendo ser aplicados em quaisquer conjuntos de dados de sequenciamento metagenômico ou metatranscritômico obtidos através da plataforma *Illumina*. A segunda abordagem correspondeu ao isolamento e caracterização de novos fagos de *Pseudomonas* obtidos de amostras de compostagem. Três novos fagos foram identificados e tiveram os seus genomas sequenciados. A caracterização genômica desses fagos revelou genomas com alto grau de novidade, *insights* sobre a evolução de *Caudovirales* e a presença de genes de tRNA, cuja função pode estar relacionada com um mecanismo dos fagos para contornar o viés traducional apresentado pela bactéria hospedeira. A caracterização experimental dos novos fagos isolados demonstrou grande potencial para lise e dissolução de biofilme da cepa *Pseudomonas aeruginosa* PA14, conhecida como agente causador de infecções hospitalares em pacientes imunodeprimidos. Em suma, os dados reunidos nesta dissertação caracterizam a diversidade presente no viroma da compostagem e contribuem para o entendimento dos perfis taxonômico, funcional e ecológico do processo.

Aos três grandes amores da minha vida: Biologia, Computação e Vírus.

ATGCTA01001101ZC01ZC03ZC08010100TAG

AGRADECIMENTOS

Ao Professor João Carlos Setubal, por orientar este trabalho de todas as formas possíveis. Por guiar o foco às vezes disperso do aluno biólogo, por proporcionar um ambiente de trabalho excepcional, por manter uma equipe de alunos com diferentes conhecimentos e formações, pelas inúmeras oportunidades de aperfeiçoamento acadêmico, por orientar em momentos de dúvidas e confusão, enfim, por ter contribuído para a minha formação sendo esse profissional exemplar.

À Professora Aline Maria da Silva, por coorientar este trabalho. Por proporcionar o ambiente necessário para a realização dos trabalhos experimentais, por reunir uma equipe de alunos extremamente prestativos e dispostos a ajudar um biólogo bioinformata, por apontar na direção certa para conseguir a coisa certa, por conceber a ideia inicial que se moldou neste trabalho. Agradeço por compartilhar a sua forma de ver a ciência e o mundo dos microrganismos, por compartilhar dessa paixão que tenho por virologia. Obrigado pelas dicas de livros, de artigos, pelos souvenirs e principalmente, pela empolgação das conversas sempre que o assunto era fago e as possibilidades que o futuro deslumbra sobre eles.

Ao hoje doutorando Luiz Thiberio Rangel, em breve grande professor e orientador. Pelos conhecimentos de bioinformática passados desde a graduação, por contribuir de forma tão significativa para as minhas escolhas acadêmicas. Agradeço pelas inúmeras discussões sobre evolução e pelos *insights* (sim, insights!) que vieram delas. Suas contribuições estão presentes ao longo de todo este trabalho.

Agradeço ao Professor Arthur Gruber, por ter me apresentado à Bioinformática e pela orientação de Iniciação Científica durante a graduação. Os conhecimentos transmitidos foram inúmeros e a contribuição, inestimável.

Agradeço aos Professores Paolo Zanutto, Welligton Luiz de Araújo, Ricardo Vêncio e Armando

Ventura pelas contribuições e sugestões durante os exames de qualificação e defesa desta dissertação. Por apontar os erros e sugerir correções. Pela lista de leituras indicada pelo Prof. Zanotto como pré-requisito de um aluno D3, por compartilhar seu profundo conhecimento sobre virologia e evolução.

À Karen Lombardi e Luciana Principal por realizarem a árdua tarefa de isolar e extrair o DNA dos novos fagos de *Pseudomonas*. Ao Professor Ronaldo Quaggio pela replicação e concentração dos fagos. À Layla Martins pelas incontáveis ajudas com o sequenciamento, cultura dos fagos, experimentos de PCR, experimentos de dissolução de biofilme e testes de infecção. Nosso grupo é privilegiado de contar com uma profissional de tamanha excelência. Obrigado também pela leitura e sugestões no manuscrito.

Ao Gianluca Major, pela criação do Navegador de Metagenomas Caravela a partir de uma ideia inovadora e original. Agradeço por disponibilizar uma versão beta da plataforma antes mesmo da sua conclusão. Como primeiro beta-tester, posso dizer que a ferramenta é útil, me ajudou muito nesta dissertação e irá ajudar ainda mais.

Ao Davi Barbosa, pela ajuda fundamental com o cálculo das correlações das abundâncias de táxons virais e bacterianos e pelo auxílio com o uso da ferramenta Cytoscape.

À Erika Machado, agradeço pelo cuidado e pelo conselho fundamental que tornou este trabalho possível.

Ao meu pai, Gilberto Leonel Amgarten, pelo apoio irrestrito ao longo de todos esses anos. Pelo suporte frente aos problemas e por ser um pai exemplar.

Agradeço à Nadir Dias dos Santos Torezan e Luis Roberto Torezan, pela ajuda nos primeiros anos da graduação. Eu não teria chegado tão longe sem aquele primeiro empurrão e, por isso, sou muito grato. Esse trabalho é fruto do bem que eles fazem para as pessoas.

Agradeço aos meus companheiros de laboratório. Aos membros do Setulab por todo esse tempo junto, pelas inúmeras discussões e ajudas que permeiam toda esta dissertação. Fazer parte de um ambiente saudável e cheio de pessoas incríveis fez toda a diferença.

Agradeço ao Professor Chris Upton por ter me recebido no Viral Bioinformatics Resource Center e por todo o conhecimento transmitido. Agradeço igualmente à equipe do VBRC, em especial ao Chad Smithson e Kathleen McLeod, pela paciência do dia-a-dia e por fazer minha estadia no Canadá memorável.

Por fim, mas não menos importante, agradeço a todos os meus amigos pelos momentos de lazer e descontração que foram essenciais para o desenvolvimento deste trabalho. Em especial, à Telma Melo

pela revisão minuciosa de parte desse texto e por ser alguém que eu posso sempre contar, à Erica de Oliveira por ser farinha do mesmo saco e companheira cativa de moradia e ao Daniel Robles pelas inúmeras viagens e companheirismo de sempre.

Obrigado a todos que contribuíram de forma direta ou indireta para a realização deste trabalho e para a minha formação acadêmica.

Este trabalho foi realizado com o apoio financeiro da Fundação de amparo à pesquisa do estado de São Paulo (FAPESP) através do projeto temático “Estudos da diversidade microbiana no Parque Zoológico do Estado de São Paulo” (Processo 2011/50870-6), de uma bolsa de mestrado (Processo 2014/16450-8), de uma bolsa de estágio no exterior BEPE (Processo 2015/14334-3) e parcialmente por uma bolsa de mestrado da Coordenação de aperfeiçoamento de pessoal de nível superior (CAPES). Agradecemos o apoio imprescindível para a realização deste trabalho.

1	Introdução	2
1.1	Vírus	3
1.1.1	Visão geral	3
1.1.2	Bacteriófagos	3
1.1.3	Diversidade	4
1.1.4	Importância ecológica e aplicações biotecnológicas	5
1.2	Metagenômica	6
1.3	Compostagem	6
1.4	Motivação	7
1.5	Objetivo geral	8
1.6	Relacionamento entre os capítulos	9
	Referências Bibliográficas	10
2	Trabalhos relacionados	13
2.1	Diversidade microbiana na compostagem	14
2.2	Metagenômica e diversidade viral	14
2.3	Isolamento e caracterização de novos fagos cultiváveis	16
	Referências Bibliográficas	18
3	Diversidade no viroma da compostagem	20
3.1	Considerações iniciais	21
3.2	Material e Métodos	21

3.2.1	Conjunto de dados	21
3.2.2	Análises de diversidade e abundância	22
3.2.3	Pipeline de recuperação de contigs virais em metagenomas	25
3.2.4	Estudos de caso	27
3.3	Resultados e Discussão	27
3.3.1	Diversidade e abundância viral na compostagem	27
3.3.2	Recuperação de contigs metagenômicos virais	35
3.3.3	Estudos de caso	41
3.4	Conclusão	48
	Referências Bibliográficas	51

4 Novos fagos isolados da compostagem fornecem informações importantes sobre a evolução e diversidade de fagos com cauda 54

4.1	Considerações iniciais	55
4.2	Authors' contribution and affiliations	55
4.3	Background	56
4.4	Results and Discussion	57
4.4.1	Phage isolation and sequencing	57
4.4.2	Genomic and functional characterization of phage ZC01	57
4.4.3	Genomic and functional characterization of phages ZC03 and ZC08	60
4.4.4	ZC03 and ZC08 specific genes and differences	63
4.4.5	tRNAs and codon bias	64
4.4.6	Homing endonucleases insertion region	66
4.4.7	Phylogenetic analyses	68
4.4.8	Host-range and Phage Morphology	69
4.4.9	Putative depolymerizing enzymes	71
4.4.10	<i>Pseudomonas aeruginosa</i> biofilm degradation	72
4.4.11	Phages in action: metagenomic and metatranscriptomic analyses in the composting process	72
4.5	Conclusion	74
4.6	Methods	75
4.6.1	Bacterial strains and growth conditions	75

4.6.2	Phages isolation and propagation	75
4.6.3	Phage titration	75
4.6.4	Phage DNA extraction and Illumina MiSeq sequencing	76
4.6.5	Genome assembly	76
4.6.6	Verification of ZC01 genome circularity	77
4.6.7	Clusters of phages	77
4.6.8	Genomic and functional characterization	78
4.6.9	Phylogenetic analyses	78
4.6.10	Analyses of phages abundance in composting samples	79
4.6.11	Host-range infection assays	79
4.6.12	Study of bacteriophages effects on biofilm formation	79
4.6.13	Transmission electron microscopy	80
	Referências Bibliográficas	81
5	Conclusão	86
5.1	Contribuições da dissertação	87
5.2	Direções futuras e pesquisas adicionais	88
	Apêndice A Tabelas de abundâncias	89

CAPÍTULO 1

INTRODUÇÃO

1.1 Vírus

1.1.1 Visão geral

Vírus são as entidades biológicas mais numerosas do planeta, ultrapassando todas as outras juntas com uma abundância estimada de 10^{31} partículas virais na biosfera [Bergh et al., 1989, Rohwer and Edwards, 2002]. Eles são encontrados em todas as formas de vida, das bactérias aos cordados, sendo conhecidos como pequenos agentes infecciosos capazes de replicar somente dentro de células vivas. Apresentam grande diversidade quanto à composição e forma de seus genomas, e podem ser compostos de DNA simples-fita, dupla-fita ou DNA com lacunas. Além disso, os vírus são os únicos organismos conhecidos que armazenam informação genética clássica na forma de RNA simples ou dupla-fita. Toda essa diversidade genômica é relevante para os estudos de metagenômica viral, pois cada tipo de material genético exige protocolos diferentes de purificação, extração e análises computacionais que foram considerados ao longo desta dissertação [Thurber et al., 2009].

Os genomas de vírus conhecidos são caracterizados em geral pelo tamanho pequeno (10 - 50 Kpb) e por codificarem poucas proteínas essenciais envolvidas no ciclo de replicação [Steward et al., 2000]. Essa visão, contudo, está sendo repensada devido às descobertas de novos vírus como os Mimiviruses-like e Pandoravirus [Philippe et al., 2013]. Esses vírus têm evidenciado genomas maiores do que o de muitas bactérias parasitas, chegando a atingir 2 Mpb e codificando aproximadamente 2000 proteínas das mais variadas, incluindo enzimas como as aminoacil-tRNA sintetases [Abergel et al., 2007] e genes relacionados à fotossíntese [Lindell et al., 2005].

1.1.2 Bacteriófagos

Os bacteriófagos são vírus que infectam células procarióticas, sendo compostos tipicamente de uma molécula de material genético encapsulada por um componente proteico de morfologia complexa. Tal morfologia é justificada pela presença de uma cabeça icosaédrica ligada a uma cauda, muitas vezes com fibras terminais. A cauda tem função de ligação a receptores na superfície da célula para a injeção do material genético no hospedeiro [Ackermann, 2007]. Todavia, a diversidade de fagos pode ir muito além, sendo que fagos com morfologia filamentosa e organizações alternativas de genoma já foram reportados [Dower and Cwirla, 1995, Semancik et al., 1973].

Fagos apresentam diferentes ciclos de vida, dos quais são conhecidos o ciclo lítico, lisogênico e persistente. Esses ciclos são importantes para estudos de diversidade por influenciar na forma e

apresentação dos genomas [Sime-Ngando, 2014]. No ciclo lítico, o fago se fixa a receptores específicos da parede celular do hospedeiro, injeta seu material genético dentro da célula e utiliza todo o maquinário celular da bactéria para a produção de novas partículas virais que culminam com a lise do hospedeiro e liberação dos virions no ambiente. Já o ciclo lisogênico, característicos de fagos ditos temperados, tende a ser mais brando, ocorrendo a infecção da célula pelo fago e posterior incorporação do material genético no genoma ou no citoplasma do hospedeiro. O fago torna-se um prófago capaz de replicar juntamente com o cromossomo bacteriano e pode permanecer por gerações nesse estado, voltando ao ciclo lítico novamente em casos de estresse ambiental ou por algum outro gatilho específico. A terceira forma de ciclo de vida é a persistente, onde o fago infecta o hospedeiro, utiliza o maquinário celular para replicar-se lentamente e vai sendo liberado aos poucos por brotamentos de membrana.

1.1.3 Diversidade

A diversidade de vírus é classificada de forma análoga à classificação de espécies celulares, todavia com importantes diferenças a serem ressaltadas. Não há consenso na comunidade científica sobre a existência de um ancestral comum único de todos os vírus e, portanto, não há definições de domínios, reinos ou filos. As classificações hierárquicas começam pelo nível de ordem (virales) e seguem pelos níveis de família (viridae), subfamília (virinae), gênero (virus) e espécie. As definições de grupos taxonômicos e espécies são regidas pelo International Committee on Taxonomy of Viruses (ICTV), sendo baseadas historicamente em evidências de morfologia, composição de ácido nucleico (ssDNA, dsDNA, RNA, etc.), e mais recentemente em dados genômicos e moleculares. Segundo o 9º relatório do ICTV [King, 2011], a diversidade viral está contida em 7 ordens, sendo elas: *Caudovirales*, *Herpesvirales*, *Ligamenvirales*, *Mononegavirales*, *Nidovirales*, *Picornavirales* e *Tymovirales*. Além disso, muitos grupos e espécies permanecem sem classificação.

Os bacteriófagos constituem o grupo de vírus com maior número de espécies descritas em banco de dados públicos e são considerados também os mais diversos dentre os vírus, segundo estudos metagenômicos [Breitbart et al., 2003]. Fagos com cauda possuem uma origem comum e constituem a ordem *Caudovirales* [Maniloff and Ackermann, 1998]. Todos os membros da ordem apresentam dsDNA como material genético, e alguns estudos sugerem que eles podem representar entre 10-40% da abundância total de vírus em ambientes aquáticos [Sime-Ngando, 2014]. Segundo dados do 9º relatório do ICTV, dentro de *Caudovirales* há três famílias: *Siphoviridae*, fagos com caudas longas, flexíveis e com ou sem fibras terminais (ex. Fago lambda); *Podoviridae*, fagos com caudas curtas e não contráteis (ex.

Fago T7); e *Myoviridae*, compreendendo fagos com caudas contráteis e fibras terminais de tamanhos variados (ex. Fago T4). Além disso, cerca de 96% das 5500 espécies de bacteriófagos descritas até 2007 são fagos com cauda [Ackermann, 2007], sugerindo portanto, que a ordem *Caudovirales* constitui o grupo taxonômico mais importante em estudos de diversidade de bacteriófagos e vírus em geral.

Vale lembrar que esse número de fagos conhecidos não constitui um retrato fiel da enorme diversidade de bacteriófagos existente, já que o isolamento de novos fagos está condicionado às técnicas de cultivo *in vitro* de seus hospedeiros, que como sabemos, não é uma característica representativa da maioria dos microrganismos [Rappé and Giovannoni, 2003]. É nesse sentido que técnicas de sequenciamento direto do DNA ambiental podem fornecer novos *insights* para o campo de estudo de fagos.

1.1.4 Importância ecológica e aplicações biotecnológicas

O crescente aumento de interesse pelo estudo de fagos tem gerado publicações com evidências de funções estratégicas desempenhadas por estes organismos nas mais variadas comunidades e ecossistemas, além de possíveis aplicações na medicina e biotecnologia. A hipótese *Kill de winner* levantada por Thingstad em 1997 [Thingstad and Lignell, 1997] e recentemente corroborada por diferentes trabalhos [Winter et al., 2010, Fuhrman and Schwalbach, 2003] postula que fagos podem ser responsáveis por direcionar a diversidade microbiana através da predação das espécies com uma explosão populacional, dificultando assim a dominância de uma ou poucas espécies em comunidades microbianas ambientais. Além disso, a lise de células desencadeada pelo ciclo lítico de replicação dos fagos apresenta papel importante no ciclo do carbono e outros nutrientes no oceano, como demonstrado em [Rohwer and Thurber, 2009]. Há também fortes indícios quanto aos bacteriófagos servirem como reservatórios ambientais de genes microbianos, atuando de forma a facilitar a circulação desses genes no ambiente por transferência horizontal ou, muitas vezes, criando a própria variabilidade genética [Mokili et al., 2012, Suttle, 2007].

Além de sua tradicional utilização como ferramentas na tecnologia do DNA recombinante [Haq et al., 2012], bacteriófagos também têm sido propostos como uma alternativa para o controle biológico de bactérias relacionadas a doenças em plantas e animais, uma técnica conhecida como *Phage therapy*. Estes organismos apresentam grande potencial de lise de células bacterianas e de destruição de biofilme em casos de infecções difíceis de tratar com antibióticos [Brüssow, 2005, Zhang and Hu, 2013].

1.2 Metagenômica

A metagenômica pode ser definida como o estudo do material genético recuperado diretamente de amostras ambientais, das quais o DNA ou RNA podem pertencer a espécies de domínios, filos e táxons diferentes, além de constituírem um grande conjunto de dados quando sequenciados. Estudos metagenômicos têm o potencial de fornecer informações valiosas para a caracterização de organismos não cultiváveis por técnicas laboratoriais [Edwards and Rohwer, 2005] e para o entendimento de vias metabólicas complexas como, por exemplo, as presentes no processo de compostagem.

A maioria das sequências obtidas via metagenômica pertence a organismos procarióticos e em menor parcela, a organismos eucarióticos [Dinsdale et al., 2008]. Menor proporção ainda é usualmente atribuída às sequências virais, que a despeito de pertencerem às entidades biológicas mais abundantes da biosfera, são tão pouco representadas. Muitas explicações podem ser levantadas para essa observação, desde o tamanho proporcionalmente menor dos genomas virais (média de 10 - 50 Kpb) até o viés inserido por métodos utilizados para classificar as sequências como pertencentes a um determinado grupo taxonômico. Esses métodos geralmente utilizam de similaridade com sequências conhecidas em banco de dados, que como sabemos, podem apresentar limitações por dependerem de bancos com uma pequena e não representativa parcela da diversidade viral [Edwards and Rohwer, 2005, Rosario and Breitbart, 2011, Sime-Ngando, 2014].

Tendo em vista os desafios impostos pela diluição das sequências virais em estudos metagenômicos tradicionais, tornou-se necessário a adaptação das análises para obtenção de melhores resultados, caso sejam vírus, o foco do estudo. Abordagens bioinformáticas podem ser usadas para filtrar sequências virais e posteriormente caracterizá-las, porém, como já mencionado anteriormente, as peculiaridades das sequências virais e a escassez de homólogos em bancos de dados tornam essa tarefa complexa e às vezes enviesada [Yin and Fischer, 2008]. A mineração de dados virais em estudos metagenômicos totais constitui, portanto, um grande desafio a ser superado.

1.3 Compostagem

A compostagem é uma decomposição bio-oxidativa tipicamente microbiana realizada por uma comunidade complexa, cuja estrutura muda dependendo de fatores como temperatura, pH, aeração, teor de umidade e quantidade de sólidos orgânicos utilizada [Ryckeboer et al., 2003]. O metabolismo microbiano aeróbio impulsiona as alterações de pH e o rápido aumento da temperatura (acima de 50 °C),

seguido de uma fase com temperaturas entre 60-80 °C e terminando com um gradual resfriamento do composto [Kumar, 2011].

O processo converte a matéria orgânica sólida biodegradável em “composto” estável do tipo húmus que pode ser manuseado, armazenado e utilizado como biofertilizante, sendo uma alternativa eficaz e econômica para o tratamento de resíduos orgânicos. No processo de compostagem, geralmente a matéria orgânica é misturada de forma estratificada (por exemplo: estrume e matéria orgânica vegetal alternadamente) observando-se proporções de carbono e nitrogênio de 30:1, aeração adequada e teor de umidade ao redor de 60%.

Os distintos estágios do processo de compostagem estão associados a populações específicas de microrganismos que processam as mais diferentes fontes de carbono ali existentes. Das mais simples (mono e oligossacarídeos) até as mais complexas (lignocelulose, pectina, amido e proteínas). O intenso metabolismo destes microrganismos resulta no aumento de temperatura e diminuição do pH, ocasionado pela degradação de ácidos orgânicos produzidos no processo. Quando a temperatura eleva-se acima de 45°C, microrganismos termófilos substituem os mesófilos. É nesta etapa termofílica que a maioria da matéria orgânica é degradada e, conseqüentemente, mais oxigênio é consumido. A degradação de lignina também acontece nessa fase.

Após a fase termofílica, a atividade microbiana e a temperatura diminuem, propiciando a colonização de microrganismos mesofílicos que lentamente degradam compostos orgânicos mais complexos. Nesse momento, a produção de substâncias húmicas se torna mais expressiva para formação do “composto” maduro [Gajalakshmi and Abbasi, 2008]. Esta dinâmica afeta as populações de microrganismos que conduzem a transformação da matéria orgânica, estabelecendo, assim, um mecanismo de feedback entre os sistemas bióticos e abiótico que impacta na estrutura da comunidade microbiana [Blanc et al., 1999, Gajalakshmi and Abbasi, 2008].

1.4 Motivação

O presente trabalho encontra-se inserido no projeto temático FAPESP 2011/508706 “Estudo da diversidade microbiana no Parque Zoológico do Estado de São Paulo (PZSP)”, que tem como um dos objetivos caracterizar a diversidade microbiana no processo de compostagem, além de entender quais são os principais responsáveis em cada uma das etapas e como eles contribuem para a dinâmica geral do processo. A diversidade microbiana e funcional tem sido acessada tanto pelo sequenciamento direto

do DNA e RNA total, como pelo sequenciamento de *amplicons* do gene do rRNA 16S de amostras coletadas ao longo de todo o processo [Martins et al., 2013, Antunes et al., 2016]. Resultados interessantes sobre a composição e estrutura desta comunidade têm sido gerados e é inquestionável que a definição de perfis taxonômicos bacterianos e a análise de seus genes componentes podem prover informações úteis sobre o papel dos microrganismos neste microambiente. Todavia, restringir os estudos somente a estas análises pode limitar o nosso entendimento do processo como um todo, já que sabemos que cepas muito próximas de uma mesma espécie de bactéria podem apresentar fenótipos diferenciados devido à presença de elementos móveis como transposons e bacteriófagos, por exemplo. Portanto, um estudo necessário e, ao nosso conhecimento, sem precedentes na literatura científica diz respeito à caracterização da diversidade viral presente na comunidade microbiana do processo de compostagem. Dessa forma, poderemos conhecer e entender melhor a importância de todos os atores no processo, assim como investigar a influência de fagos e elementos móveis na eficiência biológica da compostagem e na ecologia da comunidade como um todo.

1.5 Objetivo geral

O objetivo geral desta dissertação é realizar um estudo da diversidade viral presente na comunidade microbiana do processo de compostagem do PZSP. Para tal, dividimos os trabalhos realizados em duas abordagens: Na **primeira abordagem**, realizamos a mineração (data mining) do vasto conjunto de dados de sequenciamento metagenômico e metatranscritômico gerado em nosso grupo para amostras tempo-seriadas do processo de compostagem. As análises têm como finalidade identificar a diversidade e eventuais variações nos grupos taxonômicos de vírus e próvirus presentes na comunidade microbiana da compostagem, assim como estabelecer protocolos de reconhecimento e recrutamento de sequências virais longas em dados de sequenciamento de amostras ambientais. Na **segunda abordagem**, focamos a caracterização genômica de bacteriófagos cultiváveis isolados a partir de amostras da compostagem. O isolamento destes bacteriófagos foi realizado através do co-cultivo com bactérias *Pseudomonas aeruginosa* PA14, uma cepa de relevância clínica por causar infecções hospitalares agressivas e difíceis de tratar com abordagens tradicionais.

1.6 Relacionamento entre os capítulos

Os capítulos que se seguem nesta dissertação estão organizados da seguinte forma: Primeiramente, apresentamos no capítulo 2 uma revisão da literatura sobre a diversidade de vírus em microambientes de degradação de biomassa. Também apresentamos o estado-da-arte em recuperação de genomas parciais e/ou completos de vírus a partir de dados de metagenômica e em caracterização genômica de novos fagos isolados. Os capítulos 3 e 4 foram escritos em formato de artigo para publicação e são auto-suficientes. Neles são apresentados os principais dados da literatura, metodologias, e resultados obtidos de forma a atingir o objetivo geral dessa dissertação. De forma resumida, o capítulo 3 versa sobre o estudo da diversidade viral em amostras metagenômicas da compostagem e sobre o desenvolvimento de protocolos computacionais de análises. Já o capítulo 4 apresenta a caracterização genômica de três novos fagos de *Pseudomonas* isolados de amostras da compostagem. O capítulo final, de conclusão, apresenta um panorama dos estudos realizados e contribuições científicas da dissertação, além de perspectivas e ideias que podem dar prosseguimento ao trabalho desenvolvido.

Referências Bibliográficas

- [Abergel et al., 2007] Abergel, C., Rudinger-Thirion, J., Giegé, R., and Claverie, J.-M. (2007). Virus-encoded aminoacyl-trna synthetases: structural and functional characterization of mimivirus tyrrs and metrs. *Journal of virology*, 81(22):12406–12417.
- [Ackermann, 2007] Ackermann, H.-W. (2007). 5500 phages examined in the electron microscope. *Archives of virology*, 152(2):227–243.
- [Antunes et al., 2016] Antunes, L., Martins, L. F., Pereira, R. V., Thomas, A. M., Barbosa, D., Nascimento, L. L., Silva, G. M. M., Moura, L., Epamino, G. W. C., de Oliveira, J. C., Lombardi, K. C., Digiampietri, L. A., Ramos, P. L., Pascon, R. C., Quaggio, R. B., da Cruz, J. B., da Silva, A. M., and Setubal, J. C. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports*.
- [Bergh et al., 1989] Bergh, Ø., Børshheim, K. Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–468.
- [Blanc et al., 1999] Blanc, M., Marilley, L., Beffa, T., and Aragno, M. (1999). Thermophilic bacterial communities in hot composts as revealed by most probable number counts and molecular (16s rDNA) methods. *FEMS Microbiology Ecology*, 28(2):141–149.
- [Breitbart et al., 2003] Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology*, 185(20):6220–6223.
- [Brüssow, 2005] Brüßow, H. (2005). Phage therapy: the escherichia coli experience. *Microbiology*, 151(7):2133–2140.
- [Dinsdale et al., 2008] Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., et al. (2008). Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632.
- [Dower and Cwirla, 1995] Dower, W. J. and Cwirla, S. E. (1995). Transforming cell with filamentous bacteriophage expression vector, culturing for expression and assembly of particle with antibody fab fragment on coat, selecting encoding particle by binding to specific ligand, removing nonbinding particles. US Patent 5,427,908.
- [Edwards and Rohwer, 2005] Edwards, R. A. and Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology*, 3(6):504–510.
- [Fuhrman and Schwalbach, 2003] Fuhrman, J. and Schwalbach, M. (2003). Viral influence on aquatic bacterial communities. *The Biological Bulletin*, 204(2):192–195.
- [Gajalakshmi and Abbasi, 2008] Gajalakshmi, S. and Abbasi, S. (2008). Solid waste management by composting: state of the art. *Critical Reviews in Environmental Science and Technology*, 38(5):311–400.
- [Haq et al., 2012] Haq, I. U., Chaudhry, W. N., Akhtar, M. N., Andleeb, S., and Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: a review. *Virol J*, 9(9).
- [King, 2011] King, A. M. (2011). *Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses*, volume 9. Elsevier.
- [Kumar, 2011] Kumar, S. (2011). Composting of municipal solid waste. *Critical reviews in biotechnology*, 31(2):112–136.

- [Lindell et al., 2005] Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., and Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–89.
- [Maniloff and Ackermann, 1998] Maniloff, J. and Ackermann, H.-W. (1998). Taxonomy of bacterial viruses: establishment of tailed virus genera and the other caudovirales. *Archives of virology*, 143(10):2051–2063.
- [Martins et al., 2013] Martins, L. F., Antunes, L. P., Pascon, R. C., de Oliveira, J. C. F., Digiampietri, L. A., Barbosa, D., Peixoto, B. M., Vallim, M. A., Viana-Niero, C., Ostroski, E. H., et al. (2013). Metagenomic analysis of a tropical composting operation at the sao paulo zoo park reveals diversity of biomass degradation functions and organisms. *PloS one*, 8(4):e61928.
- [Mokili et al., 2012] Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Current opinion in virology*, 2(1):63–77.
- [Philippe et al., 2013] Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., et al. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286.
- [Rappé and Giovannoni, 2003] Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology*, 57(1):369–394.
- [Rohwer and Edwards, 2002] Rohwer, F. and Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16):4529–4535.
- [Rohwer and Thurber, 2009] Rohwer, F. and Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature*, 459(7244):207–212.
- [Rosario and Breitbart, 2011] Rosario, K. and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current opinion in virology*, 1(4):289–297.
- [Ryckeboer et al., 2003] Ryckeboer, J., Mergaert, J., Vaes, K., Klammer, S., De Clercq, D., Coosemans, J., Insam, H., and Swings, J. (2003). A survey of bacteria and fungi occurring during composting and self-heating processes. *Annals of Microbiology*, 53(4):349–410.
- [Semancik et al., 1973] Semancik, J., Vidaver, A., and Van Etten, J. (1973). Characterization of a segmented double-helical rna from bacteriophage ϕ 6. *Journal of molecular biology*, 78(4):617–625.
- [Sime-Ngando, 2014] Sime-Ngando, T. (2014). Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Frontiers in microbiology*, 5.
- [Steward et al., 2000] Steward, G. F., Montiel, J. L., and Azam, F. (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnology and Oceanography*, 45(8):1697–1706.
- [Suttle, 2007] Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812.
- [Thingstad and Lignell, 1997] Thingstad, T. and Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology*, 13(1):19–27.
- [Thurber et al., 2009] Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature protocols*, 4(4):470–483.

- [Winter et al., 2010] Winter, C., Bouvier, T., Weinbauer, M. G., and Thingstad, T. F. (2010). Trade-offs between competition and defense specialists among unicellular planktonic organisms: the “killing the winner” hypothesis revisited. *Microbiology and Molecular Biology Reviews*, 74(1):42–57.
- [Yin and Fischer, 2008] Yin, Y. and Fischer, D. (2008). Identification and investigation of orfans in the viral world. *BMC genomics*, 9(1):24.
- [Zhang and Hu, 2013] Zhang, Y. and Hu, Z. (2013). Combined treatment of pseudomonas aeruginosa biofilms with bacteriophages and chlorine. *Biotechnology and bioengineering*, 110(1):286–295.

CAPÍTULO 2

TRABALHOS RELACIONADOS

2.1 Diversidade microbiana na compostagem

O perfil da comunidade microbiana da compostagem vêm sendo elucidado através de vários trabalhos, cujas abordagens concentram-se na detecção dos grupos de microrganismos mais abundantes (majoritariamente bactérias) e no isolamento de organismos específicos em laboratório. Esses estudos costumam apresentar vieses impostos pela necessidade de clonagem de sequências de rRNA [Bent and Forney, 2008] ou pelo uso de técnicas que não são capazes de reproduzir os fatores necessários para o cultivo da maioria dos organismos em laboratório [Rappé and Giovannoni, 2003]. Contudo, estas limitações têm sido superadas pelos avanços nas tecnologias de sequenciamento direto de amostras ambientais e devido ao desenvolvimento de métodos computacionais mais robustos para a análise das sequências geradas [Thomas et al., 2012]. Em conjunto, estes avanços estabelecem uma visão mais abrangente da composição e da diversidade filogenética da comunidade microbiana presente no processo de compostagem.

Abordagens metagenômicas estão guiando a descoberta de enzimas e microrganismos envolvidos na degradação de biomassa em ambientes análogos à compostagem, tais como o rúmen de bovinos [Brulc et al., 2009, Hess et al., 2011] e uma compostagem experimental para degradação da gramínea *switchgrass* [Allgaier et al., 2010]. Além disso, nosso grupo desenvolveu um estudo metagenômico da comunidade microbiana em amostras coletadas de duas composteiras (ZC1 e ZC2) do PZSP [Martins et al., 2013] e estudos adicionais com outras composteiras (ZC3 e ZC4) incluindo dados de metatranscritômica que foram publicados recentemente [Antunes et al., 2016].

Quanto à diversidade viral em específico, [Cheepudom et al., 2015] reportaram o isolamento de um novo fago termoestável obtido através de amostras da compostagem. O fago é capaz de infectar *Thermobifida fusca*, uma bactéria degradadora de celulose e de grande relevância biotecnológica. Da mesma forma, um trabalho recentemente publicado por [Lima-Junior et al., 2016] isolou oito novos fagos de compostagem através do co-cultivo com *Mycobacterium smegmatis*. Os autores relatam alto grau de novidade genômica para um dos fagos e discutem que ambientes como a compostagem são extremamente ricos quanto à diversidade viral.

2.2 Metagenômica e diversidade viral

Visando a caracterização da diversidade de vírus em amostras ambientais, [Kristensen et al., 2010] propuseram uma re-investigação de dados metagenômicos totais disponíveis publicamente para comu-

nidades marinhas do mar de Sargaço e dados metagenômicos enriquecidos para partículas virais do Ártico, Costa da British Columbia, Golfo do México e Mar de Sargaço. O estudo focou na identificação de sequências virais, reconhecimento de POGs (Phages Orthologous Groups) [Kristensen et al., 2013] e proteínas com alta similaridade às proteínas de fagos. Os resultados obtidos sugerem que os bacteriófagos conhecidos não são representativos de amostras ambientais, constituindo assim, uma pequena parcela da diversidade genética existente. Além disso, as análises realizadas por este grupo indicam que o viroma está povoado de *virus-like-agents* (partículas defectivas ou parcialmente defectivas) com enorme potencial presumido para carregar genes entre as diversas comunidades microbianas marinhas.

De forma similar, a cavidade gastro-intestinal humana também têm sido foco de muitos estudos metagenômicos virais. Em 2003, [Breitbart et al., 2003] publicaram um estudo metagenômico que identificou através do sequenciamento Sanger de partículas virais isoladas uma dominância de fagos no viroma, com estimativas da ordem de 1200 diferentes tipos virais na amostra única analisada. Outro estudo com 252 amostras metagenômicas gastro-intestinais foi realizado por [Waller et al., 2014], no qual genes marcadores táxon-específicos foram usados para identificar, quantificar e monitorar 20 táxons virais, sendo a maioria a nível de gênero. Na média, cinco táxons foram identificados em cada amostra com até três deles sendo altamente abundantes. Mais uma vez os dados indicaram o grande potencial do viroma no fluxo de genes dentro e entre comunidades bacterianas e, nesse mesmo sentido, um estudo muito interessante reportou que a perturbação do microbioma gastro-intestinal por antibióticos é seguida de um aporte de novos genes virais à comunidade bacteriana [Modi et al., 2013]. Esses dados sugerem fortemente que o viroma pode preservar a robustez funcional da comunidade microbiana exposta a estresses ambientais, assim como fornecer mecanismos de adaptação e resistência a antibióticos.

Ainda sobre bacteriófagos da cavidade gastro-intestinal humana, um estudo publicado por [Dutilh et al., 2014] tem causado muita discussão na comunidade científica. Uma abordagem baseada no programa crAss [Dutilh et al., 2012] e na coocorrência de sequências foi utilizada para a montagem de amostras metagenômicas do intestino humano de 12 indivíduos. O resultado foi um genoma circular de aproximadamente 97 Kpb atribuído a um fago nunca antes descrito, chamado pelos pesquisadores de *CrAssphage*. O interessante dessa descoberta diz respeito à abundância das sequências deste fago em conjuntos de dados públicos de amostras metagenômicas fecais humanas, que segundo os autores são seis vezes mais abundantes do que as sequências de todos os outros fagos conhecidos juntos. Além disso, eles estimam que o fago descoberto deve estar presente em 2/3 da população humana. Os

pesquisadores envolvidos na descoberta argumentam que o fago não foi identificado antes devido ao fato de a maioria das proteínas codificadas pelo seu genoma não apresentarem similaridades significativas em bancos de dados públicos. Tal fato evidencia de forma incisiva a necessidade de novas ferramentas e metodologias para a exploração da diversidade de fagos presentes no ambiente.

Por fim, um estudo geral publicado recentemente na revista Nature [Paez-Espino et al., 2016] analisou cerca de 5 TB de dados de sequências metagenômicas dos mais variados ambientes distribuídos pelo planeta através da plataforma IMG [Markowitz et al., 2013]. Os resultados são superlativos em todos os sentidos, sendo que os autores relatam a recuperação de cerca de 125000 novos genomas parciais e/ou completos de vírus e um aumento no número de genes virais conhecidos da ordem de 16x. Além disso, o uso de espaçadores CRISPR e genes de tRNAs permitiu a predição dos prováveis hospedeiros para 9992 desses contigs metagenômicos, alguns dos quais teoricamente capazes de infectar filamentos de bactéria até então sem vírus conhecidos.

2.3 Isolamento e caracterização de novos fagos cultiváveis

A caracterização da diversidade de fagos durante muito tempo esteve atrelada à caracterização da diversidade bacteriana. Ao definir-se uma bactéria de interesse a ser estudada, os pesquisadores podiam proceder com o isolamento de novos fagos capazes de infectar o hospedeiro em questão. Um bom exemplo dessa prática é a bactéria *Mycobacterium smegmatis*, que apresenta grande relevância pela sua similaridade ao agente causador da tuberculose. Milhares de fagos diferentes foram isolados a partir do co-cultivo com *M. smegmatis*, inclusive com a criação de um projeto de ensino em escolas secundaristas nos EUA conhecido como *Phage hunters* [Hatfull et al., 2012]. O projeto contribuiu para a criação do *Actinobacteriophages database* (<http://phagesdb.org/>), coordenado pelo Professor Graham Hatfull da Universidade de Pittsburgh. Nesse banco de dados estão depositados os genomas de cerca de 8000 vírus de bactérias do filo *Actinobacteria* e sua contribuição para a biologia dos fagos e vírus em geral é sem precedentes.

Além dos fagos de *Mycobacterium*, vários estudos têm reportado a caracterização de novos fagos cultiváveis nos mais diferentes hospedeiros. Podemos citar, por exemplo, os trabalhos [Latino et al., 2014, Ceyssens et al., 2008] que isolaram e caracterizaram novos fagos cultiváveis de *Pseudomonas*. Em conjunto, o isolamento e caracterizações de novos fagos cultiváveis fornecem material para o de-

envolvimento de trabalhos como [Kwan et al., 2006] e [Hatfull et al., 2010], que discutem de forma comparativa o genoma de 18 fagos de *Pseudomonas* e 60 fagos de *Mycobacterium*, respectivamente. Ambos os trabalhos demonstram que existe uma enorme diversidade no proteoma desses fagos e discutem, inclusive, a importância desses organismos para o surgimento de cepas resistentes de bactérias em infecções hospitalares.

Referências Bibliográficas

- [Allgaier et al., 2010] Allgaier, M., Reddy, A., Park, J. I., Ivanova, N., D’haeseleer, P., Lowry, S., Sapra, R., Hazen, T. C., Simmons, B. A., VanderGheynst, J. S., et al. (2010). Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. *Plos one*, 5(1):e8812.
- [Antunes et al., 2016] Antunes, L., Martins, L. F., Pereira, R. V., Thomas, A. M., Barbosa, D., Nascimento, L. L., Silva, G. M. M., Moura, L., Epamino, G. W. C., de Oliveira, J. C., Lombardi, K. C., Digiampietri, L. A., Ramos, P. L., Pascon, R. C., Quaggio, R. B., da Cruz, J. B., da Silva, A. M., and Setubal, J. C. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports*.
- [Bent and Forney, 2008] Bent, S. J. and Forney, L. J. (2008). The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *The ISME journal*, 2(7):689–695.
- [Breitbart et al., 2003] Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology*, 185(20):6220–6223.
- [Brulc et al., 2009] Brulc, J. M., Antonopoulos, D. A., Miller, M. E. B., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., Wacklin, P., et al. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences*, 106(6):1948–1953.
- [Ceyssens et al., 2008] Ceyssens, P.-J., Mesyanzhinov, V., Sykilinda, N., Briers, Y., Roucourt, B., Lavigne, R., Robben, J., Domashin, A., Miroshnikov, K., Volckaert, G., et al. (2008). The genome and structural proteome of yua, a new pseudomonas aeruginosa phage resembling m6. *Journal of bacteriology*, 190(4):1429–1435.
- [Cheepudom et al., 2015] Cheepudom, J., Lee, C.-C., Cai, B., and Meng, M. (2015). Isolation, characterization, and complete genome analysis of p1312, a thermostable bacteriophage that infects thermobifida fusca. *Frontiers in microbiology*, 6.
- [Dutilh et al., 2014] Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications*, 5.
- [Dutilh et al., 2012] Dutilh, B. E., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R. A., and Mokili, J. L. (2012). Reference-independent comparative metagenomics using cross-assembly: crass. *Bioinformatics*, 28(24):3225–3231.
- [Hatfull et al., 2012] Hatfull, G. F. et al. (2012). Complete genome sequences of 138 mycobacteriophages. *Journal of virology*, 86(4):2382–2384.
- [Hatfull et al., 2010] Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C.-C., Weber, R. J., Patel, M. C., Germane, K. L., Edgar, R. H., et al. (2010). Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *Journal of molecular biology*, 397(1):119–143.
- [Hess et al., 2011] Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–467.
- [Kristensen et al., 2010] Kristensen, D. M., Mushegian, A. R., Dolja, V. V., and Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends in microbiology*, 18(1):11–19.

- [Kristensen et al., 2013] Kristensen, D. M., Waller, A. S., Yamada, T., Bork, P., Mushegian, A. R., and Koonin, E. V. (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *Journal of bacteriology*, 195(5):941–950.
- [Kwan et al., 2006] Kwan, T., Liu, J., DuBow, M., Gros, P., and Pelletier, J. (2006). Comparative genomic analysis of 18 pseudomonas aeruginosa bacteriophages. *Journal of bacteriology*, 188(3):1184–1187.
- [Latino et al., 2014] Latino, L., Essoh, C., Blouin, Y., Vu Thien, H., and Pourcel, C. (2014). A novel pseudomonas aeruginosa bacteriophage, ab31, a chimera formed from temperate phage paju2 and p. putida lytic phage af: Characteristics and mechanism of bacterial resistance. *PloS one*, 9(4):e93777.
- [Lima-Junior et al., 2016] Lima-Junior, J. D., Viana-Niero, C., Oliveira, D. V. C., Machado, G. E., da Silva Rabello, M. C., Martins-Junior, J., Martins, L. F., Digiampietri, L. A., da Silva, A. M., Setubal, J. C., et al. (2016). Characterization of mycobacteria and mycobacteriophages isolated from compost at the são paulo zoo park foundation in brazil and creation of the new mycobacteriophage cluster u. *BMC microbiology*, 16(1):111.
- [Markowitz et al., 2013] Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I., Tringe, S., and et al. (2013). IMG/m 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 42(D1).
- [Martins et al., 2013] Martins, L. F., Antunes, L. P., Pascon, R. C., de Oliveira, J. C. F., Digiampietri, L. A., Barbosa, D., Peixoto, B. M., Vallim, M. A., Viana-Niero, C., Ostroski, E. H., et al. (2013). Metagenomic analysis of a tropical composting operation at the sao paulo zoo park reveals diversity of biomass degradation functions and organisms. *PloS one*, 8(4):e61928.
- [Modi et al., 2013] Modi, S. R., Lee, H. H., Spina, C. S., and Collins, J. J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, 499(7457):219–222.
- [Paez-Espino et al., 2016] Paez-Espino, D., Eloie-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering earth’s virome. *Nature*, 536(7617):425–430.
- [Rappé and Giovannoni, 2003] Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology*, 57(1):369–394.
- [Thomas et al., 2012] Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. *Microb Inform Exp*, 2(3):1–12.
- [Waller et al., 2014] Waller, A. S., Yamada, T., Kristensen, D. M., Kultima, J. R., Sunagawa, S., Koonin, E. V., and Bork, P. (2014). Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME journal*, 8(7):1391–1402.

CAPÍTULO 3

DIVERSIDADE NO VIROMA DA COMPOSTAGEM

3.1 Considerações iniciais

Os trabalhos desenvolvidos neste capítulo têm por objetivo a caracterização da diversidade viral presente na comunidade microbiana da compostagem através de abordagens independentes de cultivo laboratorial. Para tal, analisamos o conjunto de dados de sequenciamento metagenômico e metatranscritômico de amostras tempo-seriadas obtidas ao longo do processo de compostagem do Parque Zoológico de São Paulo (PZSP).

Dois protocolos computacionais de análises, ou pipelines, foram desenvolvidos para garimpar o conjunto de dados metagenômicos e metatranscritômicos: (1) Pipeline para identificação de espécies virais conhecidas e levantamento das abundâncias relativas e (2) Pipeline para montagem e recuperação de genomas completos e/ou parciais de vírus a partir de dados de sequenciamento de nova geração.

O capítulo está organizado de forma a refletir o embrião de um manuscrito sobre a diversidade viral da compostagem que deverá ser submetido para publicação. Primeiramente, apresentamos a metodologia utilizada nos protocolos computacionais e demais análises de diversidade. Em seguida, descrevemos os resultados obtidos pela implementação desses protocolos nos dados de sequenciamento de amostras da compostagem. Dedicamos uma seção para o detalhamento de estudos de casos que julgamos interessante para ilustrar a aplicação dos pipelines desenvolvidos nessa dissertação. Por fim, concluímos o capítulo destacando os principais resultados obtidos e sua relevância para o estado-da-arte em diversidade viral de amostras ambientais.

3.2 Material e Métodos

3.2.1 Conjunto de dados

O conjunto de dados de sequenciamento utilizado nas análises de diversidade e abundância, na recuperação de contigs longos e na caracterização dos novos fagos de *Pseudomonas* foi obtido a partir de amostras de uma composteira presente no PZSP, denominada ZC4. Esta composteira foi montada principalmente com restos vegetais e fezes de animais do zoológico e vem sendo utilizada em diversos estudos do nosso grupo, como o publicado recentemente na revista Scientific Reports [Antunes et al., 2016]. O processo de compostagem do PZSP dura em média 90 dias e diferentes pontos ao longo deste processo foram amostrados para sequenciamento. Um sumário das amostras tempo-seriadas obtidas para cada tipo de estudo e dados de sequenciamento correspondentes é mostrado na tabela 3.1

e descrito em detalhes no artigo publicado por [Antunes et al., 2016]. Além das amostras para sequenciamento metagenômico e metatranscritômico, amostras do dia 67 foram extraídas e utilizadas em culturas de *Pseudomonas aeruginosa* gentilmente cedidas pela Professora Dra. Regina Baldini (IQ-USP) para isolamento dos novos fagos cultiváveis. O isolamento, extração de DNA e sequenciamento foram realizados conforme protocolos padrão por Karen Lombardi, Luciana Principal e Layla Martins sob orientação da Professora Dra. Aline Maria da Silva (IQ-USP). Um total de 8 isolados foram sequenciados para montagem e caracterização genômica, como será descrito no Capítulo 4.

Tabela 3.1: Distribuição dos reads de sequenciamento Illumina MiSeq na composteira ZC4 e dia do processo amostrado. Os números acompanhados de m representam milhões de pares de sequências. O dia 00 corresponde aos dados obtidos de uma pilha vegetal pré-compostagem.

Dia de coleta	00	01	03	07	15	30	64	67	78	99
Metagenoma ZC4	2m	8m	9m	9m	13m	9m	14m	8m	9m	10m
Metatranscriptoma ZC4	-	15m	15m	14m	12m	9m	17m	-	14m	8m
Isolados	-	-	-	-	-	-	-	15m	-	-

3.2.2 Análises de diversidade e abundância

Pipeline para análises de diversidade e abundância

Em testes iniciais, utilizamos a plataforma online Metavir2 [Roux et al., 2014] para realizar os levantamentos de diversidade e abundância viral na compostagem. Todavia, a ferramenta apresentou problemas técnicos quanto à capacidade de processamento dos servidores online e deixou de aceitar novas submissões. Devido a esta limitação, optamos por desenvolver uma adaptação do protocolo computacional utilizado pela plataforma Metavir2 para implementação local em nossos servidores. Tal abordagem permitiu o desenvolvimento de passos adicionais de aperfeiçoamento da ferramenta, além de adaptações que vão de encontro às necessidades específicas dos nossos dados. O protocolo de análises local foi automatizado através de scripts em linguagem bash e executado em todas as amostras metagenômicas e metatranscritômicas da composteira ZC4 para a obtenção dos resultados de diversidade e abundância. Descrições detalhadas sobre o protocolo computacional desenvolvido são dadas a seguir e ilustradas na figura 3.1.

Figura 3.1: Descrição do protocolo de análises implementado para a obtenção das matrizes de diversidade e abundância de vírus em dados da compostagem.

Em (1) é feito um controle de qualidade com cada conjunto de dados de sequenciamento, composto de uma etapa simples de filtragem dos reads (sequências obtidas como saída do sequenciador) com tamanho abaixo de 50 pb ou qualidade média abaixo de $Q_p=15$. A filtragem foi realizada pela ferramenta Prinseq [Schmieder and Edwards, 2011]. Somente um dos pares foi utilizado (pair 1) devido à qualidade média melhor das sequências e para evitar resultados conflitantes, já que ambos os pares vêm do mesmo organismo mas poderiam apresentar atribuições taxonômicas diferentes.

Em (2), é realizado uma busca traduzida dos reads nos seis frames de leitura via USEARCH (ublast) [Edgar, 2010] contra a base de dados de sequências proteicas virais RefSeq do National Center for Biotechnology Information (NCBI) [Tatusova et al., 2015]. Arquivos tabulares são gerados no formato blast6out e direcionados como entrada para a próxima etapa.

A etapa (3) corresponde às atribuições taxonômicas e ao cálculo das abundâncias relativas, sendo realizada pela ferramenta GAAS [Angly et al., 2009]. Nessa etapa, o programa faz uma filtragem inicial dos hits para cada sequência-consulta baseado em limiares de e-value $\leq 1 \times 10^{-5}$, identidade $\geq 50\%$ e cobertura do alinhamento $\geq 60\%$. Os hits mantidos após as filtrações (número máximo de 10 por sequência-consulta) são utilizados para a atribuição taxonômica e cálculo de abundâncias. Nesse ponto, optamos pelo método de ponderação dos hits com base no e-value, ao invés dos métodos convencionais de atribuição taxonômica que utilizam o best-hit do blast. Este método intuitivamente consiste do cálculo de um peso para cada hit de uma sequência-consulta metagenômica sendo analisada, de forma que os pesos carreguem o significado estatístico do valor esperado de similaridade relativa a um determinado genoma e que juntos somem 1. Fórmulas e descrições detalhadas dos cálculos podem ser consultadas no artigo correspondente [Angly et al., 2009]. É interessante frisar, todavia, que o método tem o intuito de não descartar informações de similaridade significativas e que, segundo os autores, o método tem como resultado um aumento significativo na acurácia das atribuições taxonômicas e estimativas de abundância se comparados com a técnica usual de best-hit do blast.

A ferramenta GAAS realiza ainda, uma última etapa referente à normalização das abundâncias relativas pelo tamanho do genoma do organismo correspondente aos hits selecionados, de forma a não enviesar os valores pelas diferenças de tamanho nos genomas virais (podem variar da ordem de 1000x). Esta técnica desponta como uma alternativa ao uso dos números brutos de sequências consultas com hits a um determinado genoma para definir sua abundância, sendo utilizada em vários trabalhos metagenômicos, sobretudo estudos de vírus [Lauro et al., 2011, Willner et al., 2011]. A ideia intuitiva da normalização de abundâncias com base em tamanho de genomas é similar ao conceito de *Reads per Kilobase Million* (RPKM) em estudos de RNA-Seq [Mortazavi et al., 2008, Dillies et al., 2013].

Em suma, o fluxo de análises descrito nesta seção utiliza arquivos em formato FASTA contendo os reads de sequenciamento como entrada e fornece um arquivo de saída no formato tabular com informações de abundância relativa de cada espécie frente ao total de vírus identificados (em %) para cada uma das amostras analisadas.

Análises de correlação e redes

Os valores calculados em porcentagem relativa são utilizados como entrada para o programa eLSA [Xia et al., 2013], responsável pela identificação de associações complexas dependentes de séries temporais entre os táxons. Os valores de Local Similarity Analysis (LSA) foram computados com parâmetros default e delay de 1. A visualização e saídas gráficas das associações e dependências em

forma de redes foram feitas através do programa Cytoscape [Shannon et al., 2003]. Coocorrências correlacionadas foram determinadas através de limiares de LS score > 0.975 ou LS score < -0.975 e Q-valor < 0.1 .

3.2.3 Pipeline de recuperação de contigs virais em metagenomas

Para fins de simplicidade, o pipeline de recuperação de sequências virais longas desenvolvido nesta dissertação será tratado pelo nome de MARVEL, em alusão a *Metagenomic Analysis and Retrieve of Viral Long Sequences*. A figura 3.2 traz um esquema geral do fluxo de análises implementado no pipeline.

Figura 3.2: Implementação do pipeline MARVEL. O fluxo descreve cada uma das etapas, que começa com a entrada de reads curtos de sequenciamento metagenômico e termina com dois arquivos FASTA contendo: (1) Genomas parciais e/ou completos de vírus recuperados das amostras e (2) Fragmentos putativos de próvirus. O conjunto final de contigs metagenômicos pode ser usado para a caracterização da comunidade viral e para estudos de caso.

A primeira etapa consiste do controle de qualidade dos dados visando maior eficiência da montagem dos reads de sequenciamento. Sequências com qualidade média abaixo de $Q_p=19$ e/ou tamanho menor do que 50 pb são removidas. A remoção de uma das sequências segundo esses critérios implica na remoção completa do par.

A montagem constitui parte importante do processo para recuperação de contigs longos a partir de reads curtos de sequenciamento metagenômico, por isso realizamos diversos testes e análises de qualidade dos montados antes de incorporar um montador ao pipeline desenvolvido. Vários montadores foram testados e os melhores resultados para os nossos dados da compostagem foram obtidos através do uso dos montadores Soapdenovo2 [Luo et al., 2012] e MIRA 4 [Chevreux, 2007]. Análises de qualidade para todas as montagens via software QUAST [Gurevich et al., 2013] sugerem o montador MIRA com as melhores métricas de N50, total montado e número de genes inteiros preditos (exemplo na tabela 3.2). Sendo assim, as montagens foram realizadas com o software MIRA, modo “genome, accurate” e demais parâmetros default.

Tabela 3.2: Métricas de qualidade das montagens via SOAPdenovo e MIRA calculadas através do software QUAST para a amostra de ZC4 dia 07. N50 é um valor, tal qual 50% da montagem está representada em contigs de pelo menos esse tamanho.

	# Contigs	Total montado (pb)	Maior contig (pb)	GC (%)	N50	# Genes preditos	# Genes completos
SOAPdenovo	68.233	51.494.757	7.492	62,38	720	98.788	16.342
MIRA	55.823	74.945.241	218.691	61,85	1628	110.321	42.922

O conjunto de contigs resultado da montagem é traduzido nos seis possíveis quadros de leitura e submetido a buscas via USEARCH (ublast) contra uma base de dados ampla e customizada de sequências proteicas virais obtidas das bases RefSeq do NCBI [Tatusova et al., 2015], ACLAME de elementos móveis [Leplae et al., 2009] e Actinobacteriophages DB (<http://phagesdb.org>). Limiares de e-value $\leq 1 \times 10^{-7}$, tamanho do alinhamento ≥ 50 aa e identidade mínima de 40% são utilizados para recuperar um conjunto de contigs que, em seguida, é submetido a uma segunda busca via USEARCH (ublast) contra um banco de dados de proteínas-assinatura de vírus. Esse banco é composto de proteínas RefSeq somente vírus obtidas através da busca Entrez para os seguintes termos: “Capsid”, “Large terminase”, “Small terminase”, “RNA polymerase RNA-dependent” and “Tail protein”. Os limiares de corte utilizados nesta segunda busca foram menos estridentes: E-value $\leq 1 \times 10^{-5}$, tamanho do alinhamento ≥ 30 aa e identidade $\geq 20\%$.

Os contigs com hits significativos ao banco de dados de proteínas-assinatura de vírus são submetidos em seguida a dois pontos de checagem para o controle no número de contigs bacterianos (falsos-positivos). O primeiro ponto de checagem consiste na busca por similaridade dos contigs contra um banco de proteínas house-keeping de bactéria. Limiares utilizados foram: Tamanho mínimo do alinhamento de 70 aa e identidade mínima de 50%. O segundo ponto de checagem consiste de uma

busca dos contigs restantes contra o banco completo de proteínas RefSeq bacterianas. Nesse ponto, os contigs são separados em dois grupos: (1) Contigs sem hits significativos contra o RefSeq bacteriano são considerados genomas parciais e/ou completos de vírus e (2) contigs com hits significativos são considerados fragmentos putativos de próvirus.

3.2.4 Estudos de caso

Os contigs metagenômicos candidatos a genomas completos de vírus estudados em detalhes foram checados quanto a estrutura genômica através de alinhamentos BLASTN e dotplots [Altschul, 1997]. Buscas BLASTN contra o banco nr bacteriano/viral do NCBI foram realizadas com parâmetros default. Buscas contra bancos de PSSM e HMM do Conserved Domain Database foram realizadas através da ferramenta CDD-Search do NCBI [Marchler-Bauer et al., 2014] com limiares default utilizando as sequências proteicas preditas pelo software GenMarkS [Besemer et al., 2001] com modelos de genes de fagos.

Os contigs foram submetidos à plataforma IMG-m de análises metagenômicas [Markowitz et al., 2013] e os resultados foram baixados para visualização em versão beta do navegador de metagenomas Caravela, desenvolvido em nosso grupo por Gianluca Major (a ser publicado). Os dados de cobertura e mapeamento dos reads de sequenciamento foram obtidos através do uso da ferramenta Bowtie2 [Langmead and Salzberg, 2012] e foram inseridos no Caravela para visualização. A identificação taxonômica dos reads mapeados no contig foi realizada pela ferramenta Taxonomer [Flygare et al., 2016].

3.3 Resultados e Discussão

3.3.1 Diversidade e abundância viral na compostagem

Diversidade e matrizes de abundâncias

Desenvolvemos nessa dissertação um protocolo computacional de análises de diversidade, que recebe os dados brutos de sequenciamento metagenômico ou metatranscritômico e devolve um conjunto de matrizes com os táxons virais identificados e suas respectivas abundâncias relativas. As abundâncias são dadas em porcentagem e passam previamente por um processo de normalização por tamanho do genoma para remoção do viés devido à grande variação de tamanho de genomas virais. Detalhes da metodologia empregada são descritos na seção de métodos 3.2.2.

O protocolo de análises de diversidade foi implementado através de scripts em linguagem bash e aplicado para os dados de metagenômica e metatranscritômica de todas as amostras seriadas da composteira ZC4. Ao final das análises, obtivemos arquivos tabulares em formato texto, onde são encontrados as porcentagens de abundância relativa de cada espécie dentre todos os vírus identificados em determinada amostra. Os arquivos foram concatenados de modo a apresentar uma matriz com a abundância relativa para cada espécie ao longo de todos os dias amostrados (tabelas com os 100 mais abundantes no metagenoma e metatranscritoma foram disponibilizadas no apêndice A). A tabela 3.3 mostra as dez espécies de vírus conhecidos mais abundantes no metagenoma da compostagem ao longo dos dez dias amostrados.

Tabela 3.3: Dez espécies de vírus de DNA mais abundantes ao longo dos dias no metagenoma da compostagem. Estes dados representam as abundâncias relativas (%) entre todos os vírus identificados pelo protocolo de análises de diversidade.

Espécie	D01	D03	D07	D15	D30	D64	D67	D78	D99
<i>Streptococcus</i> phage 20617	2.1	2.18	2.33	2.57	2.02	2.18	1.9	1.56	1.67
<i>Geobacillus</i> virus E2	2.09	2.94	2.59	2.11	1.49	1.69	2.8	2.52	1.16
Torque teno virus 2	1.52	1.98	1.16	1.49	2.15	2.23	1.39	1.89	3.13
<i>Clostridium</i> phage phiMMP04	0.43	1.46	1.84	0.73	3.2	1.13	2.11	2.85	1.35
<i>Enterobacteria</i> phage fiAA91-ss	0.92	1.28	2.48	2.01	1.26	1.75	1.63	1.69	1.47
<i>Microbacterium</i> phage Min1	1.15	1.27	1.15	1.47	1.5	1.07	1.0	0.96	1.07
<i>Geobacillus</i> phage GBSV1	1.05	1.19	1.68	1.37	2.11	0.95	1.17	1.2	0.6
<i>Bacillus</i> virus 1	0.74	1.05	1.58	1.47	2.06	0.82	1.21	1.4	0.65
<i>Lactococcus</i> prophage bIL312	1.07	1.1	0.79	1.26	1.1	1.11	0.96	0.77	0.88
<i>Burkholderia</i> phage phiE125	0.94	0.94	0.72	1.18	1.13	1.17	0.87	1.21	1.33

Streptococcus phage 20617 foi caracterizado como uma fago incorporado ao genoma de *Streptococcus thermophilus* DSM 20617 (prófago), uma cepa inicialmente isolada em lente fermentado. Se olharmos para as proteínas putativas codificadas no genoma deste fago, podemos encontrar uma constituição não usual de genes. Vários produtos proteicos envolvidos no metabolismo de aminoácidos e carboidratos podem ser observados como enolases, glicerato quinase e 3-dehydroquinatase I. Usando o genoma RefSeq do prófago, pudemos recuperar alguns fragmentos entre os contigs montados

pelo MIRA, entre eles uma pequena sequência de 1597 pb com um domínio enolase completo presente. Essas evidências podem sugerir um papel importante para o fago no processo de degradação de biomassa e, portanto, uma ideia para explicar sua alta abundância ao longo dos dias da compostagem.

Além disso, a presença de um Torque teno virus (TTV) é de alguma forma intrigante. O primeiro vírus do tipo foi isolado em um paciente japonês com hepatite diferente dos tipos A-E e, desde então, eles vêm sendo encontrados e isolados em diferentes pacientes e outros animais pelo mundo [Nishizawa et al., 1997, Springfield et al., 2000]. Tirando algumas observações de alta carga viral em pacientes com miopatias inflamatórias severas, câncer e lupus, a explicação para a alta diversidade e abundância de TTV em diferentes ambientes permanece uma incógnita. TTVs geralmente apresentam genomas pequenos (cerca de 3,8 Kpb) com duas ORFs majoritárias e diversas outras menores. Ao nosso conhecimento, não há relatos de TTVs infectando espécies de bactérias e, desde que o processo de compostagem sendo estudado por este projeto é predominante bacteriano, tal dado soa contraditório. Se, de fato, um TTV está presente e se reproduzindo na compostagem, algum organismo é responsável pela sua replicação. Por ora, não sabemos de qual se trata e mais estudos devem ser realizados com o intuito de esclarecer esses resultados.

Olhando para a diversidade acessada através do metatranscritoma das mesmas amostras (tabela 3.4), observamos que o TTV não encontra-se entre os mais abundantes. Todavia, transcritos do vírus foram identificados ao longo dos dias do processo de compostagem e a presença de moléculas de RNA mesmo após 90 dias do início do processo de compostagem fornece indícios de que os fragmentos identificados não são majoritariamente devido ao aporte inicial de materiais. O RNA é uma molécula instável e não deve resistir no ambiente após os vários dias do processo. Como dito anteriormente, os dados levantados neste trabalho sugerem que o TTV desenvolve seu ciclo de vida ativamente no processo de compostagem, contudo sem um hospedeiro de nosso conhecimento.

Tabela 3.4: Espécies de vírus de RNA ou transcritos dos vírus mais abundantes ao longo dos dias amostrados no metatranscritoma da compostagem. Estes dados representam tanto as abundâncias relativas (%) de vírus com genomas de RNA, quanto porcentagem de mRNA sendo transcritos por vírus de DNA ou RNA. Asteriscos verdes representam espécies identificadas entre os mais abundantes tanto no metagenoma quanto no metatranscritoma de ZC4. Sublinhado foi usado para ressaltar os vírus de genoma de RNA.

Espécie	D01	D03	D07	D15	D30	D64	D78	D99
* <i>Lactococcus</i> prophage bIL312	7.33	16.72	11.33	10.27	5.59	7.58	18.71	2.45
<u><i>Pleurotus ostreatus</i> virus 1 RNA-2</u>	1.45	2.46	2.04	3.57	14.48	11.27	17.35	21.24
<u><i>Enterobacteria</i> phage Hgal1</u>	13.62	7.48	0.89	0.47	5.25	0.53	0.19	0.38
<u><i>Enterobacteria</i> phage FI</u>	1.38	0.84	2.26	1.15	20.42	0.75	0.63	1.17
* <i>Streptococcus</i> prophage 20617	2.49	4.91	3.43	5.38	0.63	5.24	2.73	0.73
<u><i>Penicillium stoloniferum</i> virus F</u>	0.25	0.18	0.26	0.15	1.78	2.23	1.11	18.88
<i>Cellulophaga</i> phage phi14:2	1.59	2.76	2.12	2.80	0.52	3.89	1.22	0.42
<u><i>Rosellinia necatrix</i> partitivirus 1-W8</u>	0.71	0.53	1.36	1.25	0.90	1.28	2.38	5.53
<u><i>Enterobacteria</i> phage Qbeta</u>	2.36	1.18	1.04	0.48	6.11	0.55	0.35	0.86
<i>Cellulophaga</i> phage phi38:1	1.41	2.15	1.73	2.05	0.35	3.59	0.78	0.33

É interessante notar a presença de três vírus de RNA que infectam fungos entre os dez mais abundantes no metatranscritoma (tabela 3.4). *Pleurotus ostreatus* virus 1 apresenta genoma organizado em dois fragmentos dupla-fita de RNA, sendo que o fragmento 1 codifica uma RNA polimerase RNA-dependente e o fragmento 2 uma proteína de capsídio. O fragmento dois apresenta maior abundância, chegando a atingir o pico de 21% entre todos as sequências identificadas como vírus no dia 99. O mesmo acontece com os vírus *Penicillium stoloniferum* virus F e *Rosellinia necatrix* partitivirus 1-W8, que atingem o pico de 18.88% e 5.53% no último dia da compostagem, respectivamente. Embora a nossa compostagem não apresente quantidades significativas de fungos identificados [Antunes et al., 2016], esses dados sugerem uma maior abundância relativa desses organismos nos últimos dias do processo.

Além da saída em modo texto, os resultados de diversidade e abundância podem ser visualizados de forma gráfica e interativa através dos recursos do pacote KRONA Tools [Ondov et al., 2011]. A figura 3.3 traz a representação gráfica da composição taxonômica e abundância relativa para o metagenoma do dia 01 na composteira ZC4. Os resultados de diversidade e abundância para todas as amostras

foram disponibilizados publicamente para consulta interativa em HTML através do endereço: <http://projetos.lbi.iq.usp.br/metazoo/deyvid/dissertation/ZC4-viral-diversity-charts.html>

Figura 3.3: Diversidade de vírus presente na composteira ZC4 dia 1. A abundância relativa para cada um dos táxons é mostrada na figura.

Análises de correlação e redes

Os dados de abundância para os táxons virais foram utilizados em conjunto com dados de diversidade e abundância dos demais microrganismos identificados nas amostras [Antunes et al., 2016] para a aplicação da técnica de LSA. Esta técnica estatística permite a identificação de associações tempo-dependente entre táxons microbianos, sendo o seu uso adequado para levantamentos de diversidade e abundância em amostras tempo-seriadas [Steele et al., 2011, Chow et al., 2014], como é o caso do nosso estudo.

Os valores de correlação (LS score) foram filtrados de forma a apresentar somente as associações com alta significância (LS score > 0.975 ou LS score < -0.975 e Q-valor < 0.1). Optamos por valores extremamente estridentes com o intuito de diminuir o ruído que pode ser inserido em análises de correlação com dados de abundância relativa. As associações são mostradas através de redes (figuras 3.4, 3.5, 3.6 e 3.7), onde os nós azuis em formato de círculos representam táxons bacterianos e os nós vermelhos em formato de triângulo representam táxons virais. As arestas ligando os nós representam co-ocorrências correlacionadas (associações) identificadas através dos limiares de LS score, sendo que

linhas cheias foram usadas para correlações positivas e linhas tracejadas para correlações negativas.

A figura 3.4 apresenta um panorama geral das interações bactéria/vírus para as amostras do metagenoma da compostagem. Na figura, é possível identificar três grandes agrupamentos de interações não conectados. Oito táxons virais foram identificados com seis ou mais interações bacterianas e compreendem o que chamamos de *vírus-hub*. Estes vírus podem estar influenciando significativamente a diversidade microbiana no processo de compostagem e, por isso, foram investigados em mais detalhes.

Figura 3.4: Panorama geral da rede de associações vírus-bactérias identificadas através da técnica de LSA para as amostras tempo-seriadas da composteira ZC4. Nós azuis e vermelhos representam táxons bacterianos e virais, respectivamente. Arestas em linhas cheias representam interações positivas entre os nós, enquanto linhas tracejadas representam interações negativas. Setas indicam os vírus-hub com seis ou mais interações bacterianas que serão discutidos em detalhes posteriormente. **Textos na imagem podem não estar legíveis devido ao tamanho da fonte, todavia, cada uma das redes de associações será tratada a seguir com detalhes e figuras adicionais.**

O primeiro grupo de interações (figura 3.5) apresenta 4 fagos-hub: 3 fagos de *Mycobacterium* (PMC, SG4 e Frutiloop) e um fago de *Klebsiella* (F19). Os fagos de *Mycobacterium* pertencem a um mesmo grupo dentro da família *Siphoviridae*, os *Che8-like viruses*, e seus genomas apresentam similaridade média a nível de nucleotídeos em torno de 70%. A maior interconectividade entre os nós mantidos pelos fagos-hub de *Mycobacterium* é coerente com a suposição de que estes fagos devem ser capazes de infectar um mesmo hospedeiro ou conjunto de hospedeiros devido à similaridade entre eles. Todavia, nenhuma bactéria do gênero *Mycobacterium* demonstra o alto nível de correlação dos demais táxons mostrados na figura. É interessante notar cinco bactérias do filo *Deinococcus-Thermus*, que são bactérias conhecidas como altamente resistentes a uma série de danos, como ultra-violeta

e calor [Ferreira et al., 1997]. Além disso, as correlações que estes fagos-hub estabelecem com as bactérias em questão são majoritariamente positivas, ou seja, o aumento em abundância do fago implica em aumento das abundâncias das bactérias da interação. Este tipo de co-ocorrência correlacionada pode ser fruto de dinâmicas presa-predador, parasita-hospedeiro ou de competição como discutido em [Lenski, 1988, Chibani-Chennoufi et al., 2004]. Todavia, estudos mais detalhados e direcionados precisariam ser feitos para hipotetizar qualquer tipo de interação entre os táxons em questão. Última nota deve ser dado à presença acentuada de microrganismos extremófilos nesse conjunto de interações, que incluem *Desulfobacterales*, *Acidobacteriales*, *Halobacteriales*, entre outras.

Figura 3.5: Grupo I de interações vírus-bactéria identificadas através da técnica LSA para amostras temporizadas da composteira ZC4.

O segundo grupo de interações (figura 3.6) apresenta três vírus como hub: *Tetraselmis viridis* S20, *Salmonella phage RE-2010* e *Ralstonia phage phiRSA1*. As associações entre o vírus de *Tetraselmis* S20 e vários táxons de bactéria é de certa forma inesperada, tendo em vista que este vírus infecta microalgas eucarióticas. A maioria dessas interações são positivas e direcionadas a bactérias do filo *Proteobacteria* e *Acidobacteria*. O mesmo tipo de padrão de associações positivas foi observado no estudo [Chow et al., 2014], onde os autores estimaram que as associações positivas correspondem a cerca de 80% das

do ambiente estudado (compostagem termofílica).

Figura 3.7: Grupo III de interações vírus-bactéria identificadas através da técnica LSA para amostras tempo-seriadas da composteira ZC4.

3.3.2 Recuperação de contigs metagenômicos virais

As análises de diversidade reportadas na seção anterior caracterizam a diversidade e abundância de **espécies conhecidas de vírus**, assim como começam a trazer luz para algumas das interações que esses organismos estabelecem na comunidade microbiana da qual fazem parte. Contudo, há uma outra parte da diversidade que não pode ser acessada pelas técnicas e protocolos descritos anteriormente, e que acreditamos compreender uma grande e significativa parcela da diversidade viral. Essa diversidade compreende desde vírus com média similaridade genômica a espécies conhecidas até **vírus totalmente desconhecidos** e sem paralelo em bancos de dados públicos. Para lidar com esse desafio, desenvolvemos o pipeline MARVEL para recuperação de sequências virais longas a partir de dados de sequenciamento metagenômico e metatranscritômico.

O pipeline foi implementado e automatizado através de scripts bash (descrito em detalhes na seção de metodologias 3.2.3) e aplicado para as amostras tempo-seriadas da composteira ZC4. Primeiramente, os dados de sequenciamento brutos foram submetidos a protocolos de filtragem por qualidade e em seguida submetidos ao montador. O conjunto total de contigs foi então buscado contra um amplo

banco de dados de proteínas virais, que inclui os bancos RefSeq Viral, Aclame de fagos, prófagos e elementos móveis e Actinobacteriophages DB. Limiares moderadamente estridentes foram utilizados nesta etapa, sendo esperado que muitos dos contigs que passam pelo filtro ainda sejam oriundos de genomas bacterianos. Os contigs com hits significantes contra o amplo banco de dados viral é então buscado contra um segundo banco de dados mais restrito, composto apenas de proteínas-assinatura de vírus como por exemplo proteínas de capsídio e RNA-polimerase-RNA-dependente. Os limiares de corte utilizados nesta etapa foram menos estridentes, tendo em vista que o conjunto buscado de contigs era menor e estamos em busca de sequências com baixa similaridade (homologia distante) a proteínas conhecidas do banco de dados. Como ainda observamos a presença de muitos contigs bacterianos mesmo após essa etapa, o pipeline procede com dois pontos de checagem. O primeiro, realizando uma busca contra proteínas house-keeping de bactérias e o segundo, realizando uma busca contra o banco RefSeq de proteínas bacterianas. De forma similar às buscas contra as proteínas virais, as duas etapas de buscas contra proteínas bacterianas apresentam limiares menos estridentes contra as proteínas house-keeping e mais estridentes contra o RefSeq bacteriano. A tabela 3.5 traz uma síntese do número de contigs selecionados após cada etapa do pipeline para os dados da composteira ZC4.

Tabela 3.5: Números de contigs obtidos após cada uma das etapas do pipeline MARVEL para recuperação de grandes sequências virais. O conjunto inicial de contigs (total) passou por sucessivas filtrações para obtenção do conjunto final de fragmentos genômicos de vírus. Dados correspondem às amostras tempo-seriadas da composteira ZC4. Dia00 corresponde a uma pilha vegetal pré compostagem.

	Inclusão		Inclusão	Exclusão	Checagem
	Iniciais	Amplo-vir	Proteínas-assinatura vir	House-keeping bac	RefSeq-bac
Dia 00	11491	2693	147	146	94
Dia 01	51823	12865	582	577	271
Dia 03	49080	13477	1054	1036	623
Dia 07	49935	14111	1120	1110	718
Dia 15	76792	17485	1014	1006	681
Dia 30	57852	13350	911	903	684
Dia 64	101173	25058	1443	1430	992
Dia 78	115661	31615	3304	3280	2593
Dia 99	109935	26757	2163	2154	1819

Tanto as saídas blast quanto os contigs em formato FASTA são armazenados em arquivos intermediários que facilitam a inspeção do conjunto de sequências produzido após cada etapa do pipeline. Os contigs que não apresentam hits significativos na etapa final de busca contra o banco RefSeq bacteriano são organizados em um arquivo FASTA final, sendo considerados **genomas parciais e/ou completos de vírus** presentes na comunidade microbiana das amostras analisadas. Por outro lado, os contigs que apresentam hits significativo na busca final são organizados em outro arquivo FASTA e são considerados **fragmentos putativos de próvirus**.

As figuras 3.8 e 3.9 mostram plots cumulativo de pares de base por número de contigs amostrado dos dois conjuntos recuperados pela aplicação do pipeline em dados da composteira ZC4: Genomas completos e/ou parciais e fragmentos putativos de próvirus. Podemos verificar que as amostras dos dias finais da compostagem são as mais informativas quanto a tamanho total montado para os dois conjuntos de contigs.

Figura 3.8: Gráfico cumulativo de pares de bases por contig amostrado. Os contigs foram obtidos a partir das amostras tempo-seriadas da composteira ZC4 e são referentes ao conjunto final de genomas parciais e/ou completos de vírus recuperados pelo pipeline MARVEL desenvolvido nesta dissertação.

Figura 3.9: Gráfico cumulativo de pares de bases por contigs amostrados. Os contigs foram obtidos a partir das amostras tempo-seriadas da composteira ZC4 e são referentes ao conjunto final fragmentos putativos de provírus recuperados pelo pipeline MARVEL desenvolvido nesta dissertação.

As tabelas 3.6 e 3.7 exibem outras métricas de qualidade para os contigs recuperados, entre elas o tamanho do maior contig e número de genes preditos para o conjunto de sequências de cada amostra. Esse conjunto de contigs e genes constitui o que chamamos de **viroma da compostagem**, permitindo um olhar mais de perto e detalhado sobre a diversidade de genes e diversidade funcional presente na compostagem. A seguir, selecionamos alguns desses contigs para a realização de estudos de caso.

Tabela 3.6: Métricas de qualidade para o conjunto de genomas parciais e/ou completos recuperados pelo pipeline MARVEL desenvolvido nesta dissertação. O número de genes completos preditos para o conjunto de contigs constitui uma métrica importante para análises adicionais de caracterização funcional.

Vírus recuperados de ZC4					
	# Contigs	# Contigs (>10 kpb)	Maior contig	N50	# Genes preditos
Dia00	94	1	35118	2222	201
Dia01	271	5	20797	1924	623
Dia03	623	9	23497	1861	1435
Dia07	718	9	23418	2453	1751
Dia15	681	10	35692	2234	1524
Dia30	684	14	29339	1986	1577
Dia64	992	25	56411	3306	2684
Dia67	834	8	36157	1965	1892
Dia78	2593	52	37251	2926	7059
Dia99	1819	48	48544	3357	5067

Tabela 3.7: Métricas de qualidade para o conjunto de próvirus putativos recuperados pelo pipeline MARVEL desenvolvido nesta dissertação. O número de genes completos preditos para o conjunto de contigs constitui uma métrica importante para análises adicionais de caracterização funcional.

Próvirus recuperados de ZC4					
	# Contigs	# Contigs (>10 kbp)	Maior contig	N50	# Genes preditos
Dia00	52	0	9.192	2.871	131
Dia01	306	33	86.836	20.660	1.646
Dia03	413	87	170.855	32.903	3.826
Dia07	392	47	129.333	31.104	2.738
Dia15	325	65	179.473	50.715	3.246
Dia30	219	31	98.142	30.722	1.455
Dia64	438	49	70.307	16.787	2.159
Dia67	224	19	157.831	46.456	1.300
Dia78	687	109	280.998	40.621	5.566
Dia99	335	37	87.387	14.461	1.748

3.3.3 Estudos de caso

Nesta seção, reportamos como estudo de caso a investigação dos maiores contigs candidatos a genomas completos e/ou parciais recuperados pelo MARVEL. Os contigs foram analisados quanto à presença de genes virais, domínios conservados e características genômicas típicas de genomas de vírus. A tabela 3.8 exibe um sumário dos contigs estudados:

Tabela 3.8: Sumário dos três maiores contigs candidatos a genomas completos de fagos para estudo de caso.

Genomas candidatos de vírus metagenômicos			
	Dia64-MIRA-c101	Dia99-MIRA-c351	Dia99-MIRA-c408
Tamanho (pb)	56.411	48.544	45.345
Cobertura (reads/pos)	17,1	66,5	19,3
Repetições	Terminais invertidas de 119 pb	Terminais diretas de 4.532 pb	Terminais diretas de 333 pb
BLASTN	Nenhum hit >500 pb	Nenhum hit >500 pb	Phage BcepC6B (cov 2% ID 73%)
CDD	Capsid, tail, tail tube, phage recombinase	Capsid, tail, prohead, terminase	Capsid, tail, Cas4 CRISPR

Contig metagenômico Dia64-MIRA-c101

O contig Dia64-MIRA-c101 foi inspecionado no navegador de metagenomas Caravela (desenvolvido por Gianluca Major a ser publicado) quanto a inconsistências fruto de artefatos de montagem e coerência taxonômica dos reads de sequenciamento mapeados no contig. A figura 3.10 mostra o mapeamento dos reads no contig, onde reads de cor verde foram identificados como bactérias do filo *Firmicutes* e reads de cor roxa foram identificados como bactérias do filo *Actinobacteria*. Todavia, é interessante notar a massiva maioria de reads em cor cinza, representando reads sem nenhuma atribuição taxonômica segundo o software Taxonomer. Essa grande proporção de reads não classificados sugere que o contig apresenta alta coerência taxonômica mesmo quando investigado ao nível das suas sequências curtas componentes. Ainda na figura 3.10, podemos observar que o contig apresenta cobertura média de 17 reads/posição com distribuição uniforme e maior que 5 reads/posição em toda a sua extensão.

Figura 3.10: Visualização gráfica do contig Dia64-MIRA-c101 no navegador de metagenomas Caravela. Os hits contra bancos de dados obtido via IMG-M podem ser observados na parte superior da figura, enquanto o mapeamento dos reads componentes do contig e suas respectivas classificações taxonômicas via Taxonomer podem ser observados na parte inferior da figura. Reads de cor verde foram identificados como bactérias do filo *Firmicutes*, reads de cor roxa como bactérias do filo *Actinobacteria* e reads em cor cinza não foram atribuídos taxonomicamente.

Na parte superior da figura, encontram-se uma série de *features* (CDS, tRNAs, miRNAs, etc.) preditas para o contig de acordo com dados de análises realizadas na plataforma IMG-M do Joint Genome Institute. O IMG realiza buscas por similaridade das sequências de amino ácidos dos genes preditos no contig contra bancos que incluem: KEGG Orthology, PFAM, COG e Enzyme Commission (EC).

Adicionalmente, o contig foi buscado contra o banco nr bacteriano/viral via BLASTN e nenhum alinhamento maior do que 500 pb foi detectado, evidenciando o alto nível de novidade do novo genoma recuperado. O total de 85 genes preditos pelo GenMarkS foram submetidas a buscas via PSSM e HMM através da ferramenta CDD-Search e os resultados indicam proteínas típicas de fagos, como mostra a tabela 3.9. Além disso, é interessante ressaltar a presença de um aparato proteico para metilação de bases do DNA presente no contig, sugerindo que esse pode ser um mecanismo importante para o ciclo de vida do organismo em questão.

Tabela 3.9: Anotação dos genes presentes no contig metagenômico Dia64-MIRA-c101 e hits significativos ao conjunto de bancos de dados do Conserved Domain Database (CDD). Genes sem hits significativos não foram incluídos na tabela.

Anotação do contig metagenômico Dia64-MIRA-c101				
	Tamanho (aa)	Anotação	Referência	Evidência
Gene 1	310	Phage tyrosine recombinase	TIGR02225	1.00E-39
Gene 25	146	Single-stranded DNA-binding protein	COG0629	2.00E-25
Gene 29	302	Phage Mu protein F like protein	pfam04233	2.00E-06
Gene 30	234	Phage serine protease XkdF	pfam14550	7.00E-30
Gene 33	308	Phage major capsid protein E	pfam03864	5.00E-04
Gene 43	1419	Phage-related minor tail protein	TIGR01760	4.00E-35
Gene 48	541	Sulfatase maturation enzyme AslB	COG0641	3.00E-28
Gene 50	274	Glycosyltransferase involved in cell wall biosynthesis	COG0463	4.00E-06
Gene 53	165	Glycosyl hydrolase 108 + PG binding domain	pfam05838, pfam09374	3E-029, 5E-05
Gene 66	287	CRISPR/Cas system-associated protein Cas4	pfam12684	2.00E-13
Gene 67	292	RecT protein	COG3723	2.00E-77
Gene 71	435	DNA methylase	pfam01555	9.00E-36
Gene 72	286	Site-specific DNA-cytosine methylase	COG0270	1.00E-22
Gene 73	227	S-adenosylmethionine-dependent methyltransferases	pfam08241	3.00E-06
Gene 74	276	phage conserved hypothetical protein	TIGR02220	7.00E-37
Gene 75	577	DNA primase	TIGR01391	2.00E-14
Gene 76	229	FAD-dependent thymidylate synthase	PRK00847	1.00E-49

Por fim, analisamos a estrutura genômica do contig e constatamos a presença de repetições terminais do tipo invertidas (fita +/fita-) de 119 pb. Esse tipo de repetições é típica de elementos transponíveis (transposons) e do genoma de alguns fagos, como os fagos de *Escherichia* PRD1 e L17 [Savilahti and Bamford, 1986]. Trata-se de uma forte evidência indicando que o contig metagenômico Dia64-MIRA-c101 é o genoma completo de um novo fago.

Contig metagenômico Dia99-MIRA-c351

O contig Dia99-MIRA-c351 foi inspecionado quanto a inconsistências fruto de artefatos de montagem e coerência taxonômica no navegador de metagenomas Caravela e, de forma similar ao contig c101, apresentou cobertura alta e uniforme ao longo de sua extensão (66 reads/posição), com todas as

posições sendo cobertas por ao menos 11 reads de sequenciamento (figura 3.11). Além disso, o contig apresenta 98,2% de seus reads componentes não classificados em nenhum nível taxonômico, indicando o grau de novidade da sequência e a coerência taxonômica da montagem.

Figura 3.11: Visualização gráfica do contig Dia99-MIRA-c351 no navegador de metagenomas Caravela. Os hits contra bancos de dados obtido via IMG-M podem ser observados na parte superior da figura, enquanto o mapeamento dos reads componentes do contig e suas respectivas classificações taxonômicas via Taxonomer podem ser observados na parte inferior da figura. Reads em roxo foram classificados como bactérias do filo *Actinobacteria*, em azul como bacteriófagos da ordem *Caudovirales* e reads em cinza não foram identificados taxonomicamente. A presença de reads atribuídos a bactérias em regiões pontuais do contig pode ser explicada pela presença de proteínas com domínios conservados entre vírus e bactérias ou mesmo por se tratarem de regiões com eventos de transferência horizontal.

65 genes foram preditos pelo GenMarkS, entre eles várias proteínas-assinatura de fagos, como proteínas da cauda e terminases (tabela 3.10). Vale citar a presença de uma proteína associada ao fator de virulência E em bactérias e um cassete de proteínas repressoras metal-dependente.

Tabela 3.10: Anotação dos genes presentes no contig metagenômico Dia99-MIRA-c351 com hits significativos contra o conjunto de banco de dados do CDD. Genes sem hits significativos não foram incluídos na tabela.

Anotação do contig metagenômico Dia99-MIRA-c351				
	Tamanho (aa)	Anotação	Referência	Evidência
Gene 4	149	Restriction endonuclease	cd00523	2.00E-05
Gene 7	621	DNA polymerase A	TIGR000593	6.00E-31
Gene 8	730	Virulence-associated protein E	TIGR01613	2.00E-33
Gene 9	69	HipB Transcriptional regulator	COG1396	1.00E-09
Gene 10	536	DEAD-like helicase	smart00487	2.00E-22
Gene 14	157	Transcription factor WhiB	pfam02467	1.00E-16
Gene 17	327	Peptidoglycan recognition protein	cd06583	2.00E-10
Gene 21	1037	Phage minor-tail protein + cell wall hydrolase	COG5280, COG0791	7E-028, 3E-11
Gene 24	135	Lambda phage tail tube	pfam06199	5.00E-03
Gene 32	250	Caudovirus prohead serine protease	TIGR01543	7.00E-07
Gene 33	465	Phage portal protein	pfam04860	3.00E-68
Gene 35	513	Phage Terminase large subunit	pfam03354	4.00E-15
Gene 36	162	Phage terminase small subunit	TIGR01558	1.00E-09
Gene 40	205	Arsenical Resistance Operon Repressor like	cd00090	6.00E-05
Gene 42	620	Arsenical Resistance Operon Repressor	smart00418	4.00E-06
Gene 45	62	OmpR DNA-binding response regulator	COG0745	7.00E-07
Gene 47	221	LexA-binding, inner membrane-associated hydrolase	pfam04307	1.00E-13

A estrutura do contig evidenciou repetições terminais do tipo direta de 4532 pb, como mostra o dotplot na figura 3.12. Repetições diretas são comuns em genomas de fagos como o T5, que possui genoma de 122 Kpb com repetições diretas de 10 Kpb [Wang et al., 2005]. Mais uma vez, evidência contundente da completude do genoma recuperado através da aplicação do pipeline MARVEL em dados de metagenômica.

Figura 3.12: Dotplot do contig metagenômico alinhado contra si mesmo, onde as linhas indicam regiões alinhadas. As repetições terminais diretas podem ser observadas nas extremidades do gráfico.

Contig metagenômico Dia99-MIRA-c408

As análises via Caravela mostram que o contig Dia99-MIRA-c408 apresenta poucas variações na cobertura de reads mapeados ao longo de sua extensão (figura 3.13), com uma cobertura média de 19 reads/posição e nenhuma posição coberta por menos que 9 reads curtos de sequenciamento. Novamente, a grande maioria dos reads não foi atribuída taxonomicamente para nenhum grupo, sendo um indicativo de que não há problemas quanto a inconsistências de montagem.

Figura 3.13: Visualização gráfica do contig Dia99-MIRA-c401 no navegador de metagenomas Caravela. Os hits contra bancos de dados obtido via IMG-M podem ser observados na parte superior da figura, enquanto o mapeamento dos reads componentes do contig e suas respectivas classificações taxonômicas via Taxonomer podem ser observados na parte inferior da figura. Reads em verde claro foram classificados como bactérias do filo *Actinobacteria*, em azul escuro como bactérias do filo *Proteobacteria*, em amarelo como bacteriófagos da ordem *Caudovirales* e reads em cinza não foram identificados taxonomicamente.

O contig de 45345 pb teve 53 genes preditos, cuja anotação preliminar indica a presença de algumas proteínas características de fagos, como uma major-capsid-protein e uma head-to-tail connecting protein. Esse contig, todavia, apresenta cerca de 70% de suas proteínas como hipotéticas e sem similaridades significativas com proteínas de função conhecida em banco de dados públicos. Como podemos ver na parte superior da figura 3.13, várias proteínas longas (>500 aa) estão sendo codificadas pelo contig metagenômico sem apresentar, contudo, pistas de suas funções.

A busca do contig via BLASTN contra o banco nr bacteriano/viral apresentou hits pequenos e pouco significativos a nível genômico, sendo o fago de *Burkholderia* BcepC6B o melhor hit da busca (cobertura 2% e identidade 73%). A estrutura genômica do contig evidenciou repetições terminais do tipo direta de 333 pb, sugerindo se tratar de um genoma completo do que parece ser mais um novo fago recuperado de amostras metagenômicas da compostagem.

Tabela 3.11: Anotação dos genes presentes no contig metagenômico Dia99-MIRA-c408 com hits significativos contra o conjunto de banco de dados do CDD. Genes sem hits significativos não foram incluídos na tabela.

Anotação do contig metagenômico Dia99-MIRA-c408				
	Tamanho (aa)	Anotação	Referência	Evidência
Gene 1	511	DNA polymerase A	cd08642	3.00E-85
Gene 3	459	DEAD-like helicase	smart00487	2.00E-15
Gene 4	190	Metal dependent phosphohydrolase	COG1896	1.00E-04
Gene 19	334	Major capsid protein	PHA00665	0
Gene 20	207	Putative protease	PHA00666	1.00E-18
Gene 25	113	D-alanyl-D-alanine carboxypeptidase	pfam13539	2.00E-03
Gene 28	558	Bacteriophage head to tail connecting protein	PHA00670	0
Gene 31	151	N-Acyltransferase	PHA00673	4.00E-23
Gene 34	137	Thermolysin-like proteinase	cd09607	2.00E-03
Gene 37	86	Competence damage-inducible protein A	PRK00549	2.00E-03
Gene 39	661	Bifunctional DNA primase/polymerase	pfam09250, pfam08707	5.00E-15
Gene 48	105	HNH endonuclease	pfam13392	1.00E-10
Gene 51	442	CRISPR/Cas system-associated protein Cas4	pfam10926	2.00E-79
Gene 52	200	Phage protein of unknown function	DUF2815, pfam10991	5.00E-39
Gene 53	285	DnaQ-like exonuclease	pfam01612	1.00E-05

3.4 Conclusão

Neste capítulo, descrevemos dois conjuntos de métodos para a exploração da diversidade viral através de dados de sequenciamento metagenômico. O primeiro protocolo computacional visou realizar um levantamento da diversidade de grupos de vírus conhecidos nas amostras, assim como calcular as abundâncias relativas dos táxons virais identificados. O protocolo foi aplicado para o estudo da diversidade viral em amostras metagenômicas e metatranscritômicas da compostagem, produzindo resultados tabulares, gráficos e redes de interações entre os táxons virais e bacterianos presentes nas amostras da compostagem.

Os resultados mostram que a comunidade viral é majoritariamente composta de fagos, como es-

tudos similares em diversos ambientes têm reportado [Breitbart et al., 2003, Waller et al., 2014, Sime-
Ngando, 2014]. Vale lembrar que os dados de diversidade obtidos nesta etapa dizem respeito somente
à parcela conhecida e cultivável de vírus. Uma imensa maioria do que chamamos “Materia escura
viral” ainda permanece para ser descoberta e caracterizada.

Os dados de abundância fornecem evidências de que alguns fagos e vírus de RNA são muito
abundantes na microbiota da compostagem, todavia, não podemos descartar a possibilidade de esses
dados apresentarem um viés para organismos super-representados em banco de dados públicos, como
é o caso dos Mycobacteriófagos por exemplo. Além disso, o que definimos como espécies virais pode
estar longe de representar fielmente os dados que observamos em estudos de comunidades microbianas,
podendo ser mais adequado, por exemplo, a identificação de gêneros ou quasiespécies (grupo de vírus
relacionados com alta similaridade genômica competindo em um ambiente altamente mutagênico).
Sendo assim, as espécies identificadas em nossas análises na verdade poderiam ser agrupadas de forma
a refletir entidades únicas que englobariam espécies de vírus com um conjunto mínimo de similaridade
genômica.

As análises de coocorrência via LSA fornecem evidências iniciais para a elucidação de algumas
interações estabelecidas no ambiente entre vírus e bactérias. Sobretudo, a presença dos vírus-hub deve
ser mais estudada quanto à influência desses organismos nas dinâmicas de diversidade da comunidade
microbiana. Não dispomos de evidências suficientes para afirmar que qualquer das coocorrências
correlacionadas implicam em interações do tipo presa-predador ou parasita-hospedeiro, por exemplo,
e nesse sentido, somente estudos adicionais podem ajudar a trazer luz a essas dinâmicas.

O segundo protocolo computacional, o pipeline de análise e recuperação de sequências virais longas
MARVEL, foi implementado através de scripts bash e aplicado para as amostras metagenômicas da
composteira ZC4. O pipeline recebeu como entrada um conjunto de reads de sequenciamento de cada
amostra e produziu como saída: (1) Arquivo FASTA contendo candidatos a genomas virais completos
e/ou parciais e (2) arquivo FASTA com candidatos a fragmentos de próvirus. Um estudo de caso foi
realizado para os três maiores contigs recuperados pelo pipeline, de forma a demonstrar sua aplicação e
potencial para resultados. As evidências preliminares fornecidas pelos estudos de caso sugerem que os
três contigs metagenômicos são de fato genomas completos de novos fagos ambientais com excepcional
grau de novidade genômica.

Os dados coletados por este trabalho, aliado aos dados de trabalhos realizados por outros grupos
[Rappé and Giovannoni, 2003, Mokili et al., 2012, Sime-
Ngando, 2014, Paez-Espino et al., 2016] sugerem

que a diversidade viral conhecida atualmente é apenas uma minúscula parcela da diversidade total de vírus presentes no ambiente. O mais empolgante talvez seja pensar que esta enorme “Matéria escura viral” começou a ser desvendada e os resultados a partir deste ponto são imprevisíveis.

Referências Bibliográficas

- [Altschul, 1997] Altschul, S. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.
- [Angly et al., 2009] Angly, F. E., Willner, D., Prieto-Davó, A., Edwards, R. A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D. A., Barott, K., Cottrell, M. T., Desnues, C., et al. (2009). The gaas metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol*, 5(12):e1000593.
- [Antunes et al., 2016] Antunes, L., Martins, L. F., Pereira, R. V., Thomas, A. M., Barbosa, D., Nascimento, L. L., Silva, G. M. M., Moura, L., Epamino, G. W. C., de Oliveira, J. C., Lombardi, K. C., Digiampietri, L. A., Ramos, P. L., Pascon, R. C., Quaggio, R. B., da Cruz, J. B., da Silva, A. M., and Setubal, J. C. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports*.
- [Besemer et al., 2001] Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, 29(12):2607–2618.
- [Breitbart et al., 2003] Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology*, 185(20):6220–6223.
- [Chevreux, 2007] Chevreux, B. (2007). Mira: an automated genome and est assembler.
- [Chibani-Chennoufi et al., 2004] Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.-L., and Brüssow, H. (2004). Phage-host interaction: an ecological perspective. *Journal of bacteriology*, 186(12):3677–3686.
- [Chow et al., 2014] Chow, C.-E. T., Kim, D. Y., Sachdeva, R., Caron, D. A., and Fuhrman, J. A. (2014). Top-down controls on bacterial community structure: microbial network analysis of bacteria, t4-like viruses and protists. *The ISME journal*, 8(4):816–829.
- [Dillies et al., 2013] Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683.
- [Edgar, 2010] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461.
- [Ferreira et al., 1997] Ferreira, A. C., Nobre, M. F., Rainey, F. A., Silva, M. T., Wait, R., Burghardt, J., Chung, A. P., and Da Costa, M. S. (1997). *Deinococcus geothermalis* sp. nov. and *deinococcus murrayi* sp. nov., two extremely radiation-resistant and slightly thermophilic species from hot springs. *International Journal of Systematic and Evolutionary Microbiology*, 47(4):939–947.
- [Flygare et al., 2016] Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., Tardif, K. D., Kapusta, A., Ryneerson, S., et al. (2016). Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mrna expression profiling. *Genome biology*, 17(1):1.
- [Gurevich et al., 2013] Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.
- [Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

- [Lauro et al., 2011] Lauro, F. M., DeMaere, M. Z., Yau, S., Brown, M. V., Ng, C., Wilkins, D., Raftery, M. J., Gibson, J. A., Andrews-Pfannkoch, C., Lewis, M., et al. (2011). An integrative study of a meromictic lake ecosystem in antarctica. *The ISME journal*, 5(5):879–895.
- [Lenski, 1988] Lenski, R. E. (1988). Dynamics of interactions between bacteria and virulent bacteriophage. In *Advances in microbial ecology*, pages 1–44. Springer.
- [Leplae et al., 2009] Leplae, R., Lima-Mendez, G., and Toussaint, A. (2009). Aclame: a classification of mobile genetic elements, update 2010. *Nucleic acids research*, page gkp938.
- [Luo et al., 2012] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18.
- [Marchler-Bauer et al., 2014] Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., et al. (2014). Cdd: Ncbi’s conserved domain database. *Nucleic acids research*, page gku1221.
- [Markowitz et al., 2013] Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Pagani, I., Tringe, S., and et al. (2013). IMG/m 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 42(D1).
- [Mokili et al., 2012] Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Current opinion in virology*, 2(1):63–77.
- [Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.
- [Nishizawa et al., 1997] Nishizawa, T., Okamoto, H., Konishi, K., Yoshizawa, H., Miyakawa, Y., and Mayumi, M. (1997). A novel dna virus (ttv) associated with elevated transaminase levels in post-transfusion hepatitis of unknown etiology. *Biochemical and biophysical research communications*, 241(1):92–97.
- [Ondov et al., 2011] Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12(1):385.
- [Paez-Espino et al., 2016] Paez-Espino, D., Eloie-Fadrosch, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering earth’s virome. *Nature*, 536(7617):425–430.
- [Rappé and Giovannoni, 2003] Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Reviews in Microbiology*, 57(1):369–394.
- [Roux et al., 2014] Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC bioinformatics*, 15(1):76.
- [Savilahti and Bamford, 1986] Savilahti, H. and Bamford, D. H. (1986). Linear dna replication: inverted terminal repeats of five closely related escherichia coli bacteriophages. *Gene*, 49(2):199–205.
- [Schmieder and Edwards, 2011] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.
- [Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

- [Sime-Ngando, 2014] Sime-Ngando, T. (2014). Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Frontiers in microbiology*, 5.
- [Slesarev et al., 2002] Slesarev, A. I., Mezhevaya, K. V., Makarova, K. S., Polushin, N. N., Shcherbina, O. V., Shakhova, V. V., Belova, G. I., Aravind, L., Natale, D. A., Rogozin, I. B., et al. (2002). The complete genome of hyperthermophile methanopyrus kandleri av19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences*, 99(7):4644–4649.
- [Springfeld et al., 2000] Springfield, C., Bugert, J. J., Schnitzler, P., Tobiasch, E., Kehm, R., and Darai, G. (2000). Tt virus as a human pathogen: significance and problems. *Virus Genes*, 20(1):35–45.
- [Steele et al., 2011] Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C.-E. T., Sachdeva, R., Jones, A. C., Schwalbach, M. S., et al. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME journal*, 5(9):1414–1425.
- [Tatusova et al., 2015] Tatusova, T., Ciufu, S., Fedorov, B., O’Neill, K., and Tolstoy, I. (2015). Refseq microbial genomes database: new representation and annotation strategy. *Nucleic acids research*, 43(7):3872.
- [Waller et al., 2014] Waller, A. S., Yamada, T., Kristensen, D. M., Kultima, J. R., Sunagawa, S., Koonin, E. V., and Bork, P. (2014). Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME journal*, 8(7):1391–1402.
- [Wang et al., 2005] Wang, J., Jiang, Y., Vincent, M., Sun, Y., Yu, H., Wang, J., Bao, Q., Kong, H., and Hu, S. (2005). Complete genome sequence of bacteriophage t5. *Virology*, 332(1):45–65.
- [Willner et al., 2011] Willner, D., Furlan, M., Schmieder, R., Grasis, J. A., Pride, D. T., Relman, D. A., Angly, F. E., McDole, T., Mariella, R. P., Rohwer, F., et al. (2011). Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4547–4553.
- [Xia et al., 2013] Xia, L. C., Ai, D., Cram, J., Fuhrman, J. A., and Sun, F. (2013). Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*, 29(2):230–237.

CAPÍTULO 4

NOVOS FAGOS ISOLADOS DA COMPOSTAGEM FORNECEM
INFORMAÇÕES IMPORTANTES SOBRE A EVOLUÇÃO E
DIVERSIDADE DE FAGOS COM CAUDA

4.1 Considerações iniciais

Três novos bacteriófagos foram isolados através do co-cultivo de amostras da composteira ZC4 com *Pseudomonas aeruginosa* PA14. Os novos fagos tiveram os seus genomas sequenciados e caracterizados ao longo do desenvolvimento desta dissertação e os principais resultados foram compilados em um manuscrito para submissão à revista *BMC Genomics*. Apresentamos a versão final do manuscrito a seguir.

4.2 Authors' contribution and affiliations

The Following authors contributed to this manuscript:

Deyvid Amgarten^{1,2}, Layla Farage Martins¹, Karen Cristina Lombardi¹, Luciana Principal Antunes¹, Ana Paula Silva de Souza¹, Gianluca Gonçalves Nicastro¹, Elliott Watanabe Kitajima³, Ronaldo Bento Quaggio¹, Chris Upton⁴, João Carlos Setubal^{1,2,5*} and Aline Maria da Silva^{1,2*}

1 Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil

2 Programa de Pós-Graduação Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, Brazil

3 Departamento de Fitopatologia e Nematologia, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, Brazil

4 Biochemistry and Microbiology Department, University of Victoria, BC, Canada

5 Biocomplexity Institute of Virginia Tech, Blacksburg, VA, USA

* These authors shared senior authorship

Correspondence and requests for materials should be addressed to: J.C.S. (setubal@iq.usp.br) or A.M.D.S. (almsilva@iq.usp.br)

AMDS, DA and JCS conceived the study and designed the experiments. DA performed genomic characterization and bioinformatics analyses. KCL and LPA isolated the new phages from composting. KCL, LPA and LFM performed phage DNA isolation and sequencing. LFM and RBQ performed phage infection, biofilm assays and PCR validations of genome assemblies. GGN contributed to the design of experiments and acquisition of data of *P. aeruginosa* PA14 phage infection assays. EWK and APSS generated electron microscopy images. CU guided the work with the VOCs database. AMDS, DA and JCS wrote the manuscript.

4.3 Background

Viruses present remarkable diversity regarding morphology, genomes, and proteins [Breitbart and Rohwer, 2005]. Among viruses, bacteriophages (or simply phages) are a group of special interest, given their interactions with bacteria that are important to biotechnology and human health. As new phage genomes are characterized, unusual features are found, including new genes and novel genome architectures [Rohwer, 2003, Hendrix, 2003, Paez-Espino et al., 2016]. Thus, the study of phage diversity can contribute to the understanding of their evolution and their influence on any microbial community.

Composting is a diverse microbial environment in which complex organic molecules such as lignocellulose are converted into humus-like substances suitable for use as a soil amendment [Ryckeboer et al., 2003]. The study of composting microbial communities is important for elucidating the pathways of biomass degradation, and has contributed to the discovery of novel microorganisms and enzymes valuable for biotechnological applications [Jurado et al., 2014]. Aside from the great diversity of bacteria and fungi species in this environment [Partanen et al., 2010, Martins et al., 2013, Neher et al., 2013, López-González et al., 2015], phages have also been identified in composting material [Marks and Hamilton, 2014, Lima-Junior et al., 2016, Cheepudom et al., 2015]. Recent studies have reported novel phage genomes in composting [Lima-Junior et al., 2016], and interesting features such as phage thermostable enzymes [Cheepudom et al., 2015].

In this work, we describe three new phages isolated from a composting operation at the Sao Paulo Zoo Park (Brazil) using *Pseudomonas aeruginosa* PA14 as host. This reference strain is a clinical and highly virulent isolate that represents the most common clonal group worldwide [Mosquera-Rendón et al., 2016]. Along with the novelty of these phages, this work also presents results concerning transcribed phage genes and phage abundance variation from a three-month time-series sampling of the composting process. To our knowledge, this is the first study to present such results concerning the complex microbial context in which these phages live.

4.4 Results and Discussion

4.4.1 Phage isolation and sequencing

In order to access a slice of the cultivable phage diversity present in the composting microbiota, we screened composting samples from the São Paulo Zoo Park (São Paulo, Brazil) for phages infecting *Pseudomonas aeruginosa* PA14. Three new phages were isolated, which we named *Pseudomonas* phage ZC01, ZC03 and ZC08. Their genomes were fully sequenced, assembled, and annotated. Overall characteristics of these phage genomes are summarized in Table 4.1. For ZC01, a large contig with terminal direct repeats of 214 bp was obtained, indicating a circular genome of 57,061 bp. Genome circularity for ZC01 was confirmed by PCR assay (data not shown). Assemblies for isolates ZC03 and ZC08 resulted in two slightly different sequences with 69,844 bp and 70,774 bp, respectively. For all three phage genomes, coverage was above 8,000x and uniform through the entire contig.

Tabela 4.1: Genomic features of the three new *Pseudomonas* phages described in this work.

	Phage ZC01	Phage ZC03	Phage ZC08
Accession Number	KU356689	KU356690	KU356691
Genome size (bp)	57061	69844	70774
GC content (%)	63	42	42
Genomic structure	Circular	Linear	Linear
Genes predicted	78	85	83
tRNA genes	None	10	9 + 1 Pseudogene

4.4.2 Genomic and functional characterization of phage ZC01

The majority of the ZC01 genome consists of coding sequences, with the exception of three main non-coding regions: 370 bp at the 5' end, 249 bp around the 8 Kbp position and 933 bp around the 31 Kbp position. A common characteristic for these non-coding regions is their lower GC content (41%) relative to the average GC content for the entire ZC01 genome (63%, Table 4.1). This variation is due to an increase of T nucleotides from mean value of 16% in the genome to up to 37% in these

non-coding regions.

We have searched the National Center for Biotechnology Information (NCBI) virus-only nucleotide database [Altschul, 1997] for genomes similar to phage ZC01. The most similar genomes include phage Ab18 (98% coverage and 96% identity), phage PaMx11 (80% coverage and 72% identity) and phage YuA (31% coverage and 69% identity). Phage YuA is part of a genus in the *Siphoviridae* family, the *YuA-like viruses* [King, 2011]. Based on this information, we have created a new cluster of viral genomes in the Viral Orthologous Clusters database (VOCs) [Ehlers et al., 2002], which was named *Siphoviridae* YuA-like and is publicly available through the VOCs Java client <http://virology.uvic.ca>. Genomes in the *YuA-like* cluster were selected using BLASTN [Altschul, 1997] results and the Jaccard index of similarity. This cluster contains a total of 14 different genomes (listed in Table 4.2), 932 genes and 401 ortholog groups (OGs).

Tabela 4.2: Phage genomes assigned to *Siphoviridae* *YuA-like* and *Podoviridae* *N4-like* VOCs clusters. *Data is available in the VOCs database of large DNA viruses. **New phages described in this study.

Siphoviridae cluster*		Podoviridae Cluster*	
Phage species	Accession number	Phage species	Accession number
<i>Burkholderia</i> phage BcepGomr	NC_009447	<i>Enterobacter</i> phage EcP1	NC_019485
Phage phiJL001	NC_006938	Enterobacteria phage N4	NC_008720
<i>Pseudomonas</i> phage 73	NC_007806	<i>Erwinia</i> phage vB_EamP-S6	NC_019514
<i>Pseudomonas</i> phage Ab18	LN610577	<i>Escherichia</i> phage vB_EcoP_G7C	NC_15933
<i>Pseudomonas</i> phage B3	NC_006548	<i>Pseudomonas</i> phage LIT1	NC_013692
<i>Pseudomonas</i> phage D3112	NC_005178	<i>Pseudomonas</i> phage LUZ7	NC_013691
<i>Pseudomonas</i> phage DMS3	NC_008717	<i>Pseudomonas</i> phage ZC03**	KU356690
<i>Pseudomonas</i> phage M6	NC_007809	<i>Pseudomonas</i> phage ZC08**	KU356691
<i>Pseudomonas</i> phage MP22	NC_009818	Roseophage DSS3P2	NC_012697
<i>Pseudomonas</i> phage MP29	NC_011611	Roseophage EE36P1	NC_012696
<i>Pseudomonas</i> phage MP38	NC_011611		
<i>Pseudomonas</i> phage PaMx11	NC_0028770		
<i>Pseudomonas</i> phage YuA	NC_010116		
<i>Pseudomonas</i> phage ZC01**	KU356689		

All ten most conserved genes in the *YuA-like* cluster phages have orthologous representatives in phage ZC01 (Table 4.3). Likewise, conserved DNA helicase, RecD-like protein, deoxyuridylylate hydroxymethyltransferase and DNA polymerase A were annotated and assigned to ortholog groups with 6 or more orthologous genes. Due to the high similarity among phages ZC01, Ab18 and PaMx11, there are 36 ortholog groups shared by them only. These include well-known proteins such as a holin and a DNA ligase, but also several hypothetical proteins. The large number of unique proteins for these

three phages (approximately 44% of their genes) suggests they might constitute a new genus inside the *Siphoviridae* family, instead of belonging to the *YuA-like* genus. Moreover, phages Ab18 and ZC01 share two specific ortholog groups (VOCs ID: 18968, 18971), which were annotated as hypothetical and have no similarity with anything else in the NCBI nr database. Their function remains to be discovered. In addition, we found in the ZC01 genome a gene coding for a protein that is the only member in the VOCs ortholog group 18954. The predicted protein is the Rz1 smaller lipoprotein. It was manually annotated through the inspection of the Rz larger lipoprotein ORF (Rz1 is nested within Rz). The annotation was based on similar findings for the lambda phage genome, where Rz1 was experimentally isolated and characterized [Zhang and Young, 1999]. We verified that ZC01, Ab18 and PaMx11 share orthologous proteins for the Rz larger lipoprotein (VOCs ID: 22052); therefore, we conclude that the Rz1 smaller nested lipoprotein is also present in Ab18 and PaMx11 genomes, but it was not annotated.

Tabela 4.3: List of 20 most conserved ortholog groups in the *Siphoviridae* *YuA-like* cluster. *Data is available in the VOCs database of large DNA viruses.

VOCs ID*	Ortholog Group	Number of genes	Number of genomes
18328	Putative tail assembly protein (PSP-YuA-073)	12	12
18331	Putative tail protein (PSP-YuA-076)	12	12
18325	Tail fiber protein (PSP-YuA-070)	12	11
18326	structural phage protein (PSP-YuA-071)	11	11
18327	Tail assembly protein (PSP-YuA-072)	11	11
18330	Conserved tail assembly protein (PSP-YuA-075)	11	11
18306	Terminase large subunit (PSP-YuA-051)	8	8
18323	Virion structural protein (PSP-YuA-068)	8	8
18278	Putative deoxycytidylate deaminase (PSP-YuA-023)	7	7
18318	Structural phage protein (PSP-YuA-063)	7	7
18257	DNA helicase (PSP-YuA-002)	6	6
18258	Hypothetical protein (PSP-YuA-003)	6	6
18259	Hypothetical protein (PSP-YuA-004)	6	6
18260	Hypothetical protein (PSP-YuA-005)	6	6
18261	RecD-like DNA helicase (PSP-YuA-006)	6	6
18269	Hypothetical protein (PSP-YuA-014)	6	6
18272	Deoxyuridylate hydroxymethyltransferase (PSP-YuA-017)	6	6
18273	Hypothetical protein (PSP-YuA-018)	6	6
18275	Bacteriophage conserved protein (PSP-YuA-020)	6	6
18276	DNA Polymerase I (PSP-YuA-021)	6	6

Functional annotation of the predicted proteins based on VOCs clusters context and additional tools showed that 41% of ZC01 proteins have unknown function. “DNA metabolism and replication”

and “structural proteins” are the functional categories with most genes. A few protein products were annotated as involved in host interaction pathways, e.g., membrane lysis and metabolism regulation. ZC01 genome annotation is summarized in Fig. 4.1. Detailed annotation of the ZC01 genome and genes is available through the VOCs Java client (<http://virology.uvic.ca>).

Figure 4.1: Circular representation of the *Pseudomonas* phage ZC01 genome. The outer circle represents genes (all genes are on the plus strand). Functional blocks were defined according to annotation and are represented by colors: DNA metabolism and replication proteins (red), host metabolism interaction and regulation proteins (pink), host membrane interaction proteins (yellow) and Structural proteins (blue). Gaps in the functional block circle represent proteins with unknown function. The central graph (in purple) shows genomic GC content variation computed in 100 bp windows.

4.4.3 Genomic and functional characterization of phages ZC03 and ZC08

Phages ZC03 and ZC08 present very similar genomes, with 95% of their nucleotide sequences aligned averaging 98% identity. Differences are mainly due to two indels at the 47 Kbp and 51 Kbp positions. The first indel consists of a unique sequence (≈ 800 bp) present in phage ZC03, while the

second indel consists of a unique sequence (≈ 1000 bp) present in phage ZC08.

ZC03 and ZC08 non-coding regions also present variation on the GC content. However, in contrast to ZC01, the ZC03 and ZC08 non-coding regions display an increase in the GC content ($\approx 60\%$) with respect to the average CG content of the genome (42%, Table 1). The average GC content from phages ZC03 and ZC08 significantly diverges from the GC content of the assumed host genome (66%), suggesting that *P. aeruginosa* may not be the optimal host for these two phages [Bahir et al., 2009].

BLASTN searches of ZC03 and ZC08 genomes against the nt database returned hits with low genomic coverage and identity. The best hits included Enterobacteria phage N4 (10% coverage and 70% identity), *Erwinia* phage Ea9-2 (8% coverage and 70% identity) and Enterobacteria phage IME11 (7% coverage and 71% identity). These results strongly indicate that phages ZC03 and ZC08 are rather different from known phage species and that they probably belong to a new genus in the *Podoviridae* family. Creating a cluster of similar genomes in this case was challenging due to the shortage of similar genomes. Phage N4 is the most similar known genome and also the only officially representative of a genus recognized by the International Committee on Taxonomy of Viruses (ICTV), the *N4-like-virus* [King, 2011]. Thus, considering BLASTN results and the Jaccard index of similarity to select phage genomes, we created a new cluster of viruses in the VOCs database, which we named *Podoviridae N4-like* cluster. A total of 10 different genomes were selected to be part of this cluster (Table 4.2), comprising a total of 876 genes and 491 ortholog groups. Data for this cluster is also publicly available through the VOCs Java client at the Viral Bioinformatics Resource Center (VBRC) web platform (<http://virology.uvic.ca>).

The ZC03 and ZC08 genomes harbor representatives from all the 15 core genes found for the *N4-like* cluster, or 14 core genes according to the literature [Chan et al., 2014]. These genes include RNA polymerases 1 and 2, DNA helicase, DNA polymerase, primase, exonuclease, terminase small and large subunits and coat proteins (Table 4.4).

Tabela 4.4: List of the 15 most conserved ortholog groups corresponding to the core-genes in the *Podoviridae* *N4-like* cluster. *Data is available in the VOCs database of large DNA viruses.

VOCs ID*	Ortholog Group	Number of genes	Number of genomes
18472	RNAP1 (EBP-N4-015)	10	10
18473	RNAP2 (EBP-N4-016)	10	10
18481	AAA ATPase containing protein (EBP-N4-024)	12	10
18494	DNA helicase (EBP-N4-037)	10	10
18496	DNAP (EBP-N4-039)	10	10
18499	Putative exonuclease (EBP-N4-042)	10	10
18500	Putative primase (EBP-N4-043)	10	10
18501	gp44 (EBP-N4-044)	10	10
18502	Single-stranded DNA-binding protein (EBP-N4-045)	10	10
18512	gp55 (EBP-N4-055)	10	10
18513	Major coat protein (EBP-N4-056)	10	10
18514	gp57 (EBP-N4-057)	10	10
18516	Putative portal protein (EBP-N4-059)	10	10
18525	Terminase large subunit (EBP-N4-068)	10	10
18526	gp69 (EBP-N4-069)	10	10

A total of 44 genes were assigned to specific ortholog groups from phages ZC03 and ZC08, corresponding to more than half of the total set of genes in each genome. ZC03 and ZC08 genes are very different from gene sequences available in public databases, indicating the high degree of novelty of these genomes. Among the 44 specific genes, only two products could be annotated: a putative peptidoglycan hydrolase (VOCs ID: 18902) and a putative homing endonuclease (VOCs ID: 18919). The ZC03 genome contains 6 unique genes and an additional tRNA gene that is not present in the ZC08 genome (probably a pseudogene, as we discuss later). On the other hand, ZC08 presents 5 unique genes that are not present in the ZC03 genome. Given the evidence, we hypothesize that these two phage species have diverged in the recent evolutionary past.

Functional annotation of the predicted proteins for phages ZC03 and ZC08 based on VOCs clusters context and other tools showed that approximately 50% of the proteins have unknown function. ZC03 genome annotation and features are summarized in Fig. 4.2. Additional information about genes and

annotation is available through the VOCs Java client (<http://virology.uvic.ca>).

Figure 4.2: Linear representation of the *Pseudomonas* phage ZC03 genome. The two central bands represent genes being codified by the plus strand (green) or minus strand (orange). Functional blocks were defined according to annotation and are represented by colors in the top strand: DNA metabolism and replication proteins (red), host metabolism interaction and regulation proteins (pink), host membrane interaction proteins (yellow) and structural proteins (blue). Gaps in the functional blocks band represent proteins with unknown function. The bottom purple graph shows GC content variation computed in 100 bp windows. Hairpin symbol shows genome region where tRNA genes were predicted. Ins/del regions are shown comparing ZC03 and ZC08 phages; ZC03 is the reference for indels definition.

4.4.4 ZC03 and ZC08 specific genes and differences

Given their overall similarity, the differences between the ZC03 and ZC08 genomes can help in the understanding of viral evolution. Major differences include one gene region (*ZC03_002* and *ZC08_002*) with lower nucleotide identity than its neighborhood (48% to 94% identity, respectively) and three indel regions responsible for encoding several genes (Table 4.5).

Tabela 4.5: Phages ZC03 and ZC08 specific genes. *Hits were considered weak hits if the alignment presented coverage ≥ 0.3 and identity $\geq 20\%$ through HMM or PSSM searches.

Gene	Protein Size (aa)	Annotation	BLAST and CDD weak hits*
ZC03_002	331	Hypothetical protein	No significant hits
ZC03_047	143	Hypothetical protein	Conserved Hypothetical protein (CDD:DUF2461), Cellulase-like domain (pfam12876)
ZC03_051	60	Hypothetical protein	SH3 Proline recognition superfamily (CDD:c117036)
ZC03_052	61	Hypothetical protein	Metallophosphatase superfamily (CDD:c113995), Conserved hypothetical protein (CDD:DUF1501)
ZC03_053	166	DrpA-like recombination mediator (InterPro:IPR003488)	—
ZC03_054	32	Hypothetical protein	Flagellar basal body-associated protein FlhL (COG1580)
ZC03_055	37	Hypothetical protein	No significant hits found
ZC03_073	65	Hypothetical protein	Metallochaperone hypA superfamily (pfam01155)
—	—	—	—
ZC08_002	337	Hypothetical protein	Tail assembly protein Xyllela phage Salvo (AHB12240.1)
ZC08_048	82	Hypothetical protein	Borrelia lipoprotein (pfam00820)
ZC08_055	209	Putative HNH homing endonuclease	—
ZC08_073	47	Hypothetical protein	No significant hits found
ZC08_074	132	Hypothetical protein	Phage head-tail joining protein (COG5614)
ZC08_075	129	Hypothetical protein	Metal-responsive transcriptional regulator (pfam15611)

The genes *ZC03_002* and *ZC08_002* were annotated as hypothetical proteins, although BLASTP results show similarity (coverage 41%, identity 29%) with a tail assembly protein of *Xyella* phage Salvo (AHB12240.1). Genes *ZC03_002* and *ZC08_002* were assigned to separate individual VOCs ortholog groups, likely because their encoded amino acid sequences present only 29% identical residues in a full alignment. Multiple alignment of these genomes indicate a syntenic relationship between these genes, providing additional evidence for the hypothesis that they are distant orthologs. To explain the different nucleotide identity between the genes and their neighborhoods we hypothesize that the genes *ZC03_002* and *ZC08_002* may be involved with a host infection mechanism and, therefore, be under different selective pressure, in case their respective ideal hosts are two different strains of *Pseudomonas aeruginosa*.

ZC03 presents a cassette of genes that might have been originated from a horizontal gene transfer event. Five specific genes are encoded by this region, and most of the genes were annotated as hypothetical. Weak hits suggest functions for the genes as shown in Table 4.5, but such evidence was not considered enough for a robust annotation. The only annotated gene in the cassette is a DrpA-like DNA recombination mediator. Other two indels events consist of unique sequences in ZC08 that encode one putative HNH homing endonuclease and three hypothetical proteins, respectively. Details of the homing endonuclease insertion region will be discussed later in this work.

4.4.5 tRNAs and codon bias

Analysis of the ZC03 and ZC08 tRNA genes revealed anti-codons for seven different amino acids, with proline and leucine appearing twice. Prediction software could not accurately assign the anti-codon for one ZC03 tRNA gene, which also happens to be the one missing in the ZC08 genome (Fig. 4.3). It seems that the equivalent region to this tRNA gene in ZC03 and ZC08 genomes may have accumulated enough substitutions to produce a pseudo tRNA gene (Fig. 4.3).

Figura 4.3: Genomic region with the tRNA genes in ZC03 and ZC08. Region corresponding to the sixth tRNA gene is highlighted due to the differences between the genomes; it may be a pseudogene.

We analyzed codon usage for phage ZC03 and the host *P. aeruginosa* PA14 proteins. Table 4.6 shows predicted codons concerning ZC03 tRNAs and usage bias for each codon among all the codons for the same amino acid. The results show that tRNAs carried by the phage correspond to codons rarely used by the host. There is even an extreme case for the codon TTA (Leu), whose tRNA is not encoded in the *P. aeruginosa* PA14 genome. This data corroborates reports in the literature, suggesting that a selective recruitment of tRNAs compensates for the compositional differences between phage and host genomes [Bailly-Bechet et al., 2007]. The only exceptions to this pattern seem to be the tRNAs for asparagine and methionine, which do not present any detectable bias.

Tabela 4.6: Codon usage for phage ZC03 and *P. aeruginosa* PA14. *The anti-codon for one tRNA could not accurately be predicted, and, therefore, this might be a tRNA(Val) pseudogene.

Codon	UUA	CUA	AUG	AAC	CCA	AGA	UCA	ACA	GUA
Amino Acid	Leu	Leu	Met	Asn	Pro	Arg	Ser	Thr	Val
tRNAs encoded by phage genome	1	1	1	1	2	1	1	1	1*
tRNAs encoded by host genome	0	1	4	2	1	1	1	1	2
Random usage	0.166	0.166	1	0.5	0.25	0.166	0.166	0.25	0.25
Phage ZC03 usage	0.132	0.15	1	0.532	0.3	0.242	0.154	0.275	0.3
<i>P. aeruginosa</i> PA14 usage	0.003	0.013	1	0.85	0.047	0.007	0.014	0.025	0.059

We have investigated proteins with high frequency of proline/leucine (which tRNAs are presented twice) and proteins with high usage of the nine codons corresponding to tRNA carried by the phages. These were an Rz lipoprotein (*ZC03_025*), a putative class II holin (*ZC03_027*), two homing endonucleases (*ZC03_052* and *ZC08_055*), one hypothetical protein containing a cellulase-like domain

(*ZC03_047*), plus some hypothetical proteins. Some studies have shown that proteins highly transcribed and important to the organism are less susceptible to codon negative bias and wobble base-pairing, since the translation on those cases could be less efficient [Bulmer et al., 1987, Stadler and Fire, 2011]. Therefore, we formulate the hypothesis that the genes listed are especially important for the phages' activity within the host.

4.4.6 Homing endonucleases insertion region

Homing Endonucleases (HEs) are site-specific DNA endonucleases encoded by genes inside mobile elements such as self-splicing introns and inteins. These mobile elements can insert themselves within conserved genes and still preserve their function due to posterior self-splicing activity at the protein level [Stoddard, 2005]. They undergo a life cycle that starts with the invasion of a population, continues with the spreading through individuals, and ends when the element is fixed and is no longer under selection. At this point, the HE sequence degenerates and loses its function through random processes [Gogarten and Hilario, 2006].

HEs are commonly found in phage genomes, with reports indicating up to 15 elements in phage T4 [Edgell et al., 2010]. Phage T4 is thought to be an outlier, however, since many *T4-like viruses* have been studied and not as many HEs have been found in them. Moreover, it remains a challenge to understand the influence of HEs in producing phenotypes and in the mosaic evolution of phage genomes [Hatfull and Hendrix, 2011].

At least two different HEs were identified within phages ZC03 and ZC08 genomes, both resembling endonucleases from the HNH family [Stoddard, 2005]. Fig. 4.4 shows a genome plot of this region, where one can observe a common element equally inserted within an ATPase-domain-containing protein of 350 aa in ZC03 and ZC08 genomes. Because of the insertion, this protein was predicted as two separate pieces, and multiple alignment of orthologous proteins in the *N4-like* cluster (VOCs ID: 18481) indicates that ZC03 and ZC08 are the only genomes to present an HE within this ORF. Searches of the fusion protein against the Reference Proteome HMMER database of HMMs [Finn et al., 2011] suggest that the protein is a ribonucleotide reductase. A similar case of HE disrupting a conserved ribonucleotide reductase gene was reported for phage Aeh1 [Friedrich et al., 2007].

Figure 4.4: Comparison of the HNH endonuclease insertion region in ZC03 and ZC08 genomes. Blue lines in the top strand are substitutions, red lines are deletions in ZC03 genome and blank spaces are identical aligned regions. ADCP: ATPase Domain Containing Protein.

The second HE is inserted within the RNA polymerase (RNAP) subunit 2 gene, in the ZC08 genome only. ZC08 HE was assigned to the ortholog group VOCs ID:18862, which also contains two other proteins from phages G7C (YP_004782150.1) and EE36P1 (YP_002898939.1). However, the HE element is inserted in a different location inside phages G7C and EE36P1 genomes, more specifically between the two RNA polymerases subunit genes RNAP1 and RNAP2 (YP_004782141.1 and YP_004782143.1 in phage G7C, respectively).

A closer investigation about the RNAP genes in phages from the *Podoviridae* family shows three organization types: (I) One single-unit protein of about 880 aa (T7-type), (II) two adjacent genes encoding for a two-subunits RNAP (N4-type), and (III) two genes spaced by one or more non-related genes encoding a two-subunit RNAP (G7C-type). Multiple alignment of the RNAP protein sequences from the T7-type, N4-type and G7C-type strongly suggests that the two-subunit proteins are actually non-overlapping pieces from the larger RNAP (T7-type). Thus, HE insertion within the RNA polymerase (RNAP) subunit 2 gene in ZC08 genome may present a distinct organization that might have implications to understand the evolutionary history behind the RNAP in phages.

Altogether these findings suggest that single-unit and two-subunit RNAP in *Podoviridae* may be linked by a common evolutionary pathway. Our hypothesis is that a single-unit RNAP was present in the common ancestor of *T7-like* and *N4-like* phages. After these two lineages diverged, this single-unit protein was probably disrupted by the insertion of an HE element in the lineage that originated the *N4-like* phages. Since then, random changes have led to two-subunit genes that continue to present affinity in creating the complex required for transcription of the phage genes, as was experimentally demonstrated in [Willis et al., 2002]. This hypothesis is also supported by the comparative analysis of phage genomes from the *T7-like* and *N4-like* groups, where detailed inspection shows different derivations in the spacing region between RNAP1 and RNAP2 genes. For example, phage G7C presents only one putative HNH endonuclease (YP_004782142.1) between the RNAP genes, while phage EE36P1 presents one putative HNH endonuclease (YP_002898939.1) plus seven predicted hypothetical proteins. There are examples from zero up to eight spacing genes between RNAP1 and RNAP2.

4.4.7 Phylogenetic analyses

We performed phylogenetic analyses based on the Terminase Large Subunit gene (*terL*) of each cluster of phages. The *YuA-like* cluster presents three different ortholog groups for this gene (VOCs ID:18306, 19671, 19804) which may indicate non-orthologous or distant orthologous proteins. A maximum likelihood phylogenetic tree was generated (Fig. 4.5). The presence of three distinct groups, which correspond exactly to the VOCs ortholog groups assignment, suggests that these sequences have long undergone separate evolutionary pathways. Phages Lambda and N15 belong to an external group, since they are phages from the *Siphoviridae* family but from a different and closely related group (*Lambda-like* phages).

Figura 4.5: Maximum likelihood phylogenetic tree based on the *terL* gene for phages in the *YuA-like* VOCs cluster. The tree was rooted by two external groups represented here by Enterobacteria phage Lambda (EBP-Lambda) and Enterobacteria phage N15 (EBP-N15). Bootstrap values are shown close to the nodes in percentages.

All *terL* genes from the N4-like cluster were assigned to the same ortholog group (VOCs ID: 18525), which was also part of the N4-like core-genome. Fig. 4.6 shows an unrooted maximum likelihood tree for phages from the *N4-like* group. Internal nodes display weak bootstrap support, indicating that the relationships among phages inside the *N4-like* cluster are unresolved. However, the bootstrap values strongly support the existence of two clades, one grouping ZC03 and ZC08 and another grouping *N4-like* and related phages. This data supports the claim that phages ZC03 and ZC08 constitute a

new genus inside the *Podoviridae* family.

Figure 4.6: Maximum likelihood phylogenetic tree based on the *terL* gene for phages in the *N4-like* VOCs cluster. Bootstrap values are shown close to the nodes in percentages.

Building phylogenetic trees for phages remains a real challenge due to the non-existence of marker genes present in all species. The *terL* gene represents a candidate for this purpose in the order *Caudovirales*, but it is clearly limited by its lack of identifiable homology among families inside the order.

4.4.8 Host-range and Phage Morphology

As previously mentioned, the three phages under study were isolated from composting samples using *P. aeruginosa* PA14 as host. In infection assays of *P. aeruginosa* PA14, phage ZC01 exhibits large burst size, with well-defined borders and diameter of 2.0 - 2.5 mm. Burst sizes for ZC03 and ZC08 are similar to each other but much smaller than ZC01 plaques (0.5-1 mm) (data not shown). Host-range for these phages was evaluated using *P. aeruginosa* PAO1 and *Stenotrophomonas maltophilia*. Lysis plaques were not observed for neither of these two bacterial strains, when the three phages were dropped into their respective bacterial lawns (data not shown). These results indicate that the new phages are specific to *P. aeruginosa* PA14, but other undiscovered potential hosts cannot be ruled out.

Morphology features for the three new phages were assessed by transmission electron microscopy (Fig. 4.7. Phage ZC01 has the typical morphology from phages of the *Siphoviridae* family (Fig. 4.7A). Structural description for this group states that phages present icosahedral and head-tail geometries (morphotype B1) or a prolate capsid (morphotype B2) and T=7 symmetry. Head diameter is usually around 60 nm, being more or less elongated depending on the phage genus. Members of the family are also characterized by their filamentous, cross-banded, non-contractile tails, usually with short terminal or sub-terminal fibers [Ceyssens et al., 2008].

Figure 4.7: Transmission electron micrographs of negatively stained *Pseudomonas* phages virions found in composting: (A) *Pseudomonas* phage ZC01 with typical morphology of members of the *Siphoviridae* family; (B) (C) *Pseudomonas* phage ZC08, empty shelled and full virion, respectively, with typical morphology of members of the *Podoviridae* family. Note the short tail. Phage ZC03 image is not shown due to its similarity to phage ZC08.

Phages ZC03 and ZC08 exhibit morphology from phages of the *Podoviridae* family. Fig. 4.7 shows phage ZC08 in two different positions (panels B and C), where the icosahedral head and a short tail with terminal fibers can be seen. Image for phage ZC03 is not shown due to its similarity to ZC08.

4.4.9 Putative depolymerizing enzymes

We have screened phage genomes for genes possibly involved in pathways of biomass degradation and depolymerization, since the lysis and turnover of bacterial cell components constitute important steps in this process [Weinbauer, 2004]. Most genes involved with biomass degradation in phages are responsible for breaking cell wall peptidoglycan components and lysing the lipid membrane, an important step in infection [Moak and Molineux, 2004]. A total of nine proteins were found, including peptidoglycan hydrolases, N-acetylmuramidases, Rz lipoproteins, holins and an endolysin (Table 4.7).

Tabela 4.7: Putative depolymerization enzymes encoded in ZC01, ZC03 and ZC08 phages. Additional information include significant predictions of binding sites, transmembrane regions, domain regions and significant hits to similar proteins.

Phage	Gene name	Annotation	Size (aa)	Additional information
Phage ZC01	ZC01_075	Putative endolysin	168	TIGR02594 family protein
Phage ZC01	ZC01_076	Putative holin	70	Two transmembrane domains found with TMHMM {REF}
Phage ZC01	ZC01_078	Rz/Rz1 lipoprotein	182	Bacteriophage Rz lysis protein (pfam03245)
Phage ZC03	ZC03_016	Peptidoglycan hydrolase	831	N-acetyl-D-glucosamine binding site; lysozyme-like domain (pfam01464)
Phage ZC03	ZC03_025	Putative RZ/Rz1 lipoprotein	164	Similar to Rz/RzI spanin protein in phage EC1-UPM (AGC31575.1)
Phage ZC03	ZC03_026	N-acetylmuramidase	194	Glycosyl hydrolase 108 (pfam05838); Pepitidoglycan binding domain
Phage ZC08	ZC08_016	Peptidoglycan hydrolase	831	N-acetyl-D-glucosamine binding site; lysozyme-like domain (pfam01464)
Phage ZC08	ZC08_025	Putative Rz/Rz1 lipoprotein	164	Similar to Rz/RzI spanin protein in phage EC1-UPM (AGC31575.1)
Phage ZC08	ZC08_026	N-acetylmuramidase	194	Glycosyl hydrolase 108 (pfam05838); Pepitidoglycan binding domain

A peptidoglycan hydrolase (831 aa) is present both in ZC03 and ZC08 genomes. These proteins were assigned to the ortholog group VOCs ID: 18902, which contains only these two genes and no other homologs in the N4-like cluster. We identified a central lysozyme-like domain (pfam1464) and N-acetyl-D-glucosamine binding sites in the protein, with reports indicating a similar architecture for phage transglycosylases [Koonin and Rudd, 1994]. Accordingly, the infection tests in *S. maltophilia* provided evidence suggesting that this peptidoglycan hydrolase appears to be quite effective in breaking cell-wall: a halo of cell debris formed around the spot test when phage titers were moderately high (1×10^6 PFU ml⁻¹, data not shown). It is worthy to mention that *S. maltophilia* is not susceptible to infection by the phages.

4.4.10 *Pseudomonas aeruginosa* biofilm degradation

To investigate the ability of phages to mediate biofilm degradation we challenged *P. aeruginosa* PA14 biofilms during 24 hours with the three different phages isolated in this study. Exposure to the three phages strongly reduced biofilm cell densities, mainly for phage ZC01 (Fig. 4.8). These results indicate a promising degradation potential for these three novel phages against *P. aeruginosa* PA14 biofilms. This strain is highly virulent in susceptible animal hosts and known to form a biofilm resistant to currently available antibiotics [Pawar et al., 2015]. In these assays we used lower phage titers (5×10^5 PFU ml⁻¹) than the titers normally used in biofilm degradation assays (1×10^6 PFU ml⁻¹ to 1×10^{10} PFU ml⁻¹) [Cerca et al., 2007, Doolittle et al., 1995, Doolittle et al., 1996] highlighting the antibiofilm effectiveness of these phages.

Figure 4.8: *Pseudomonas aeruginosa* PA14 biofilms of 24h and 48h were exposed to phages ZC01, ZC03, and ZC08 for 24h. Image shows the results after exposure for each of the phages compared to the control, which was exposed to a buffer solution only. Images are representative of n=4 experiments.

4.4.11 Phages in action: metagenomic and metatranscriptomic analyses in the composting process

This work is part of a project that aims to understand the composting process at the microbial and molecular levels [Martins et al., 2013, Antunes et al., 2016]. In this project, several time-series samples of a composting unit were obtained and corresponding DNA and mRNA sequence datasets were generated. As the samples from this composting unit were used for both the metagenomic and phage isolation work, it was feasible to verify the presence of the phages described here as well as the

host *P. aeruginosa* in the DNA and mRNA sequence datasets for each sampling day. We indeed found sequences that correspond to these genomes in all datasets and the relative abundance was inferred for phage-host populations (Fig. 4.9). Phages ZC03 and ZC08 relative abundance variation parallels that of *P. aeruginosa* but with an apparent delay, which is consistent with mathematical models that have been proposed for phage-host variation [Gourley and Kuang, 2004, Krysiak-Baltyn et al., 2015]. We calculated a correlation score applying the Local Similarity Analysis (LSA) technique for time-series samples [Xia et al., 2011], and a positive LS score of 0.71 was obtained for *P. aeruginosa* and phages ZC03/ZC08 abundances (p-value < 0.02). This data suggests that *P. aeruginosa* and phages ZC03/ZC08 present a mutualistic relationship that is characteristic of temperate phages, as discussed in [Obeng et al., 2016]. In support of this finding, we note that both ZC03 and ZC08 encode genes for a holliday junction resolvase (*ZC03_013*), a putative ligase (*ZC03_0078*) and a DNA recombination protein A (*ZC03_053*); these proteins are involved with the ability of a phage to perform the lysogenic cycle [Christiansen et al., 1996].

Figure 4.9: Relative abundance of metagenomic reads through the composting process. Raw read count for phages and host was divided by the total number of reads in each sample and normalized by the genome size of the organism (given in percentage).

We investigated active phage-related functions in the composting process through the identification of metatranscriptomic reads mapped to phage genes. Fig. 4.10 shows the proportion of mRNA reads identified for each day in the respective function. “Structural” and “DNA metabolism and replication” proteins are the predominant functions observed through the days of the composting process. We

identified mRNA for host lysis in the sample of day 7, more specifically mRNA reads for a class II phage holin (*ZC03_027*). It is interesting to note that a spike in phage abundance is also observed on day 7, as well as a marked decrease in host abundance (Fig. 4.9). This observation suggests a cause-effect relationship, but additional studies are necessary to gather additional evidence for this hypothesis.

Figure 4.10: Phage-related protein functions along the composting process assessed by the identification of mRNA sequences from phages and the host *P. aeruginosa* in time-series metatranscriptomic samples. Values are shown in percentage of the total of proteins with known function.

4.5 Conclusion

In this work, three new *Pseudomonas* phages have been characterized in terms of genomic structure, genes, and the putative proteins encoded by their genomes. Two of the three phages present remarkable novelty at both the genomic and protein levels and may be members of a new genus in the *Podoviridae* family. Comparative analyses of the new phages in a context with phages from *YuA-like* and *N4-like* clusters provided important insights about the evolution and diversity of tailed phages. Moreover, infectivity and biofilm degradation experiments showed a narrow host-range and promising application as anti-microbial agents against *P. aeruginosa* infections. Finally, metagenomic and metatranscriptomic analyses provided data to situate phages ZC01, ZC03, and ZC08 in the microbial community to which they belong, yielding interesting clues about phage population and gene expression dynamics in this complex environment.

4.6 Methods

4.6.1 Bacterial strains and growth conditions

Pseudomonas aeruginosa PA14, *Pseudomonas aeruginosa* PAO1 and *Stenotrophomonas maltophilia* type strain 810-2 (ATCC 13637) were grown at 37 °C in LB-medium. Solid LB medium contained 1.5% w/v of Bacto agar (Difco) and the soft agar top-layer contained 0.7% of Bacto agar. All strains were subcultured once and glycerol stocks were done and stored frozen at −80 °C until further use.

4.6.2 Phages isolation and propagation

Composting sample for phage isolation was collected from the composting facility in the São Paulo Zoo Park, São Paulo, Brazil following the procedure previously described [Martins et al., 2013] upon 67 days after completion of the composting pile. The procedure for phage isolation was adapted from [Yoshida et al., 2013]. The compost sample (≈ 75 g) was suspended in 300 mL of SM buffer (10 mmol MgSO_4 ; 50 mmol Tris-HCl, pH 7.5) containing 3% NaCl (w/v), dispensed into 50 mL centrifuge tubes and incubated for 60 min at 4 °C. Suspensions were homogenized for 5 min at maximum speed using the TissueLyser II (Qiagen) and centrifuged at 3000 xg for 10 min. The supernatants were filtered through a 0.2 μm membrane and immediately used for infection by *Pseudomonas aeruginosa* PA14 using the soft-agar overlay method [Adams et al., 1959]. After overnight incubation at 37 °C several individual lytic plaques were collected, suspended in 100 μL of SM buffer and used for a new round of infection to warrant phage purification. The genomes of seven phage isolates were fully sequenced and out of them, three were found to be distinct (ZC01, ZC03 and ZC08). The other five isolates were identical to one of these three selected phages.

Phages were propagated using the soft-agar overlay method using *P. aeruginosa* PA14 as the host strain [Adams et al., 1959]. Briefly, 10 μL of isolated phage lysate were mixed with overnight bacterial culture and 3-5 mL of top-agar LB, and then added onto a LB Petri dish. After incubation the lysate from a clear Petri dish was eluted with SM buffer and stored at 4 °C for further use.

4.6.3 Phage titration

Bacteriophage titer was determined as described by [Adams et al., 1959]. Briefly, 100 μL of diluted phage solution, 100 μL of a bacterial overnight culture, and 5 mL of LB top agar were mixed in a tube

and poured into a LB agar containing Petri dish. Plates were incubated for 18 h at 37 °C after which plaque forming units (PFU) were counted.

4.6.4 Phage DNA extraction and Illumina MiSeq sequencing

For DNA extraction, phages were propagated in *P. aeruginosa* PA14 strain and collected after complete bacterial lysis, using 10 mL of SM buffer per Petri dish. The phage suspension was filtered through a 0.2 µm membrane and viral particles were precipitated with 10% polyethylene glycol (PEG) 8000 (w/v) and 1 mol NaCl overnight at 4 °C. Viral particles were collected by centrifugation at 3,000 ×g for 5 min. The pellet was suspended in 1 mL of SM buffer and treated with DNase (TURBO DNA-free, Life Technologies) as a way attempt to reduce contamination with *P. aeruginosa* DNA. The intact viral particles suspension were treated with phenol:chloroform:isoamylalcohol and phage DNA was extracted using MoBio PowerMax Soil DNA kit (MoBio Laboratories). Purified phage DNA was subjected to a final clean-up step using QIAamp mini spin columns (Qiagen, USA) and stored at −80 °C.

DNA purity and concentration were evaluated on a ND-1000 spectrophotometer (Nano Drop Technologies, USA) at 260 nm, 280 nm and 230 nm. Further quantification was performed with Quant-iT Picogreen dsDNA assay kit (Life Technologies, USA). DNA integrity was examined with DNA 7500 chip using 2100 Bioanalyzer and were mostly enriched in fragments higher than 10 Kbp. Shotgun genomic libraries were prepared using an Illumina Nextera DNA library preparation kit (Illumina, Inc., USA) with total DNA input of 20-35 ng. The resulting DNA fragment libraries were cleaned up with Agencourt AMPure XP beads (Beckman Coulter, Inc., USA) and fragment size within the range of 400-700 bp was verified by running in the 2100 Bioanalyzer using Agilent High Sensitivity DNA chip. Quantification of Illumina sequencing libraries with KAPA Library Quantification Kit, normalization, and pooling were performed following standard protocols for sequencing in the Illumina MiSeq platform. Pooled libraries were subjected to one run using the MiSeq Reagent kit v2 (500-cycle format, paired-end (PE) reads). On average, Illumina PE read1 and read2 presented, respectively, >80% and >75% of bases with quality score at least 30 (Q30).

4.6.5 Genome assembly

Raw reads were subject to host DNA contamination removal with Deconseq [Schmieder and Edwards, 2011] followed by a three-way protocol for digital normalization of high coverage libra-

ries using KHMER Perl scripts [Brown et al., 2012]. Resulting reads were assembled with MIRA 4 (mode:genome, accurate, other parameters default) [Chevreux et al., 2004] and final genomes were assessed by manual inspection of coverage and mapping on IGV [Thorvaldsdóttir et al., 2012].

4.6.6 Verification of ZC01 genome circularity

In order to verify genome circularity of the phage ZC01, a pair of primers was designed near to the repeated regions as described: Primer-Foward-ZC01 (5'AGCGCCTCTAGGGGTGCTAG) in the region 56975 bp to 56994 bp and primer-Reverse-ZC01 (5'TCGGAAATGATTGCCTCCGGG) in the region 306 to 326. The PCR mixture (25 μ L final volume) included 15 ng of genomic DNA, 1x KAPA Hifi HotStart Ready Mix (Bio-Rad) and 0.2 μ mol of each primer. PCR amplification was performed using the following program: 95 °C for 3 min, 25 cycles of 95 °C for 30 sec, 57 °C for 30 sec and 72 °C for 30 sec; final extension of 72 °C for 5 min and hold at 4 °C. PCR product of 495 pb, as expected for circular genome, was analyzed on a Bioanalyzer DNA 1000 chip (Agilent).

4.6.7 Clusters of phages

In order to define clusters of similar genomes to be implemented in the VOCs database, we counted the number of shared genes between two phage genomes according to Phage Ortholog Groups (POGs) data available by FTP [Kristensen et al., 2013]. Then, numbers of shared genes were used to calculate the Jaccard index (or Jaccard coefficient of similarity) for each pair of genomes according with the following expression:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (4.1)$$

$$0 \leq J(A, B) \leq 1$$

Where A and B are the group of genes from species A and B , respectively.

We consider that this index reflects similarity with more reliability than only using an absolute number of shared genes, since the J-index also consider the similarity between the total number of genes in the two phages being assessed. Lastly, we selected one reference genome for each one of the clusters and grouped genomes with a pairwise $J_{(ref, X)} \geq 0.1$. Phages YuA and N4 were chosen as

reference genomes due to their close phylogenetic relationship with ZC01 and ZC03/ZC08, respectively. The J-index cutoff was defined based on exploratory analyses of our data.

Ortholog assignment methodology and VOCs implementation were made as previously described in [Ehlers et al., 2002]. All VOCs information and genomic information about phage genomes in the two clusters used in this work are public available in the VBRC web platform <http://virology.uvic.ca>.

4.6.8 Genomic and functional characterization

Genes were predicted by GenMarkS [Besemer and Borodovsky, 2005] and Prodigal [Hyatt et al., 2010] using models for phage genes and t-RNA predictions were performed by Aragorn [Laslett and Canback, 2004]. Proteins were automatically annotated by ProKKA [Seemann, 2014] followed by additional manual characterization using CDD-Search [Marchler-Bauer et al., 2014], HMMER-Search [Finn et al., 2011] and BLAST searches [Altschul, 1997] (Jan 2016). Hits were considered robust and significant for annotation when above the following alignment thresholds: E-value: 1×10^{-5} , alignment coverage: 60% and identity: 50%. VOCs ortholog groups and embedded tools were used for transitive annotation and comparative analyses [Upton et al., 2000, Ehlers et al., 2002, Tcherepanov et al., 2006, Hillary et al., 2011].

4.6.9 Phylogenetic analyses

We identified ortholog groups for the Terminase Large Subunit gene (*terL*) by similarity in each of the clusters and performed multiple alignments through the VOCs GUI interface using MAFFT 7 (L-INS-i iterative algorithm and others parameters default) [Kato and Standley, 2013]. Guidance [Penn et al., 2010] was used to test the robustness of the multiple alignments and columns with confidence score below 0.4 were removed. The evolutionary history was inferred by using the maximum likelihood method based on the Whelan-Goldman model [Whelan and Goldman, 2001] and Le-Gascuel model [Le and Gascuel, 2008] for the Siphoviridae cluster and Podoviridae cluster, respectively. Discrete Gamma distributions with 5 categories were used to model evolutionary rate differences among sites. Robustness of branches were tested by 1000 interactions of bootstrap [Felsenstein, 1985]. Best fitting models and evolutionary analyses were conducted in RAxML version 8 [Stamatakis, 2014].

4.6.10 Analyses of phages abundance in composting samples

Metagenomics and metatranscriptomics datasets from composting time-series used in the analyses were generated from a composting unit at the Sao Paulo Zoo Park (Brazil) [Antunes et al., 2016]. Sequences were subject to mapping using Bowtie2 [Langmead and Salzberg, 2012] (default parameters) against the isolated phage genomes and *P. aeruginosa* PA14. Reads mapping were considered as belonging to the new phages or the host and counted. Relative abundances were calculated dividing the number of reads by the total number of sequencing reads generated for the sample being analyzed. We applied genome size normalization in order to compare relative abundance of phages and the host *P. aeruginosa*.

4.6.11 Host-range infection assays

The isolated phages were investigated for host range specificity and lysis efficiency in screening tests against different *Pseudomonas aeruginosa* PA14 and PA01 strains, and *Stenotrophomonas maltophilia* Type strain 810-2 (ATCC 13637) [Davenport et al., 2014]. Bacterial lawns of the different strains were propagated on LB agar plates and 5 μ L droplets of phages (1×10^4 up to 1×10^7 PFU mL^{-1}) were put on the lawns. The plates were incubated 18 h and checked for presence of plaque.

4.6.12 Study of bacteriophages effects on biofilm formation

Biofilms were allowed to form on 8-well chamber stainless slides for 24h or 48h as described in [Jurcisek et al., 2011]. Briefly, bacterial culture (200 μ L) with an OD600 of 0.05 - 0.1, which corresponds to approximately 1.5×10^7 cells was added to each well. The slide was incubated at 37 °C for 24h or 48h without shaking. The medium was replaced once a day during the whole experiment. Afterwards, the slides were washed twice with LB medium and the biofilms were challenged with 100 μ L of LB and 100 μ L of phage solution with a concentration of 5×10^5 PFU mL^{-1} during 24 hours at 37 °C. Control experiments were performed at the same conditions with the slides incubated with 100 μ L of LB and 100 μ L of SM buffer. Biofilms attached to slides before and after phage infection were stained with 1% of crystal violet solution in ethanol 96% and analyzed in the microscope.

4.6.13 Transmission electron microscopy

For transmission electron microscopy, copper grids covered with carbon-coated Formvar films were floated, membrane side down, on a drop of phage suspension for about 10 min. After eliminating excess liquid and washing with distilled water, grids were floated on a drop of 1% (w/v) uranyl acetate for 5 min. After eliminating excess liquid, dried grids were examined in a JEOL JEM 1011 or Philips EM 300 transmission electron microscope and the images registered digitally.

Declarations

List of abbreviations

NCBI: National Center for Biotechnology Information; VOCs: Viral Orthologous Clusters; OG: Ortholog group; ICTV: International Committee for Taxonomy of Viruses; HE: Homing Endonuclease; ADCP: ATPase domain containing Protein; RNAP: RNA Polymerase; *terL*: Terminase Large Subunit; POGs: Phage Ortholog Groups.

Availability of data and material

All the new phage genomes here described have been deposited in GenBank. Accession numbers are: KU356689 (ZC01), KU356690 (ZC03), and KU356691 (ZC08). The VOCs database of viral genomes and orthologs groups (clusters *YuA-like*, *N4-like*) used for comparative analyses are publicly available in the Viral Bioinformatics Resource Center online platform: <http://virology.uvic.ca/>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grant 2011/50870-6 from the São Paulo State Research Foundation (FAPESP). DA was supported by fellowships from FAPESP (2014/16450-8 and 2015/14334-3) and from the Coordination for the Improvement of Higher Education Personnel (CAPES). AMDS, JCS and KCL received Research Fellowship Awards from National Council for Scientific and Technological Development (CNPq). The funders had no role in study design, data collection, analysis, decision to publish or preparation of the manuscript.

Acknowledgements

The authors would like to thank Dr. João Batista da Cruz and Dr. Paulo Bressan from the São Paulo

Zoo Park for continued support for this project and for making the composting operation available for sampling. We express our gratitude to Dr. Regina Lucia Baldini for providing the *Pseudomonas aeruginosa* strains and for suggestions in phage infection assays. We also thank Carlos Morais for help in computational analyses, Luiz Thiberio Rangel for fruitful discussions, and the Viral Bioinformatics Resource Center team for computational tool support.

Referências Bibliográficas

- [Adams et al., 1959] Adams, M. H. et al. (1959). Bacteriophages. *Bacteriophages*.
- [Altschul, 1997] Altschul, S. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.
- [Antunes et al., 2016] Antunes, L., Martins, L. F., Pereira, R. V., Thomas, A. M., Barbosa, D., Nascimento, L. L., Silva, G. M. M., Moura, L., Epamino, G. W. C., de Oliveira, J. C., Lombardi, K. C., Digiampietri, L. A., Ramos, P. L., Pascon, R. C., Quaggio, R. B., da Cruz, J. B., da Silva, A. M., and Setubal, J. C. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports*.
- [Bahir et al., 2009] Bahir, I., Fromer, M., Prat, Y., and Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Molecular systems biology*, 5(1):311.
- [Bailly-Bechet et al., 2007] Bailly-Bechet, M., Vergassola, M., and Rocha, E. (2007). Causes for the intriguing presence of trnas in phages. *Genome research*, 17(10):1486–1495.
- [Besemer and Borodovsky, 2005] Besemer, J. and Borodovsky, M. (2005). Genemark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research*, 33(suppl 2):W451–W454.
- [Breitbart and Rohwer, 2005] Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, 13(6):278–284.
- [Brown et al., 2012] Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv preprint arXiv:1203.4802*.
- [Bulmer et al., 1987] Bulmer, M. et al. (1987). Coevolution of codon usage and transfer rna abundance. *Nature*, 325(6106):728–730.
- [Cerca et al., 2007] Cerca, N., Oliveira, R., and Azeredo, J. (2007). Susceptibility of staphylococcus epidermidis planktonic cells and biofilms to the lytic action of staphylococcus bacteriophage k. *Letters in Applied Microbiology*, 45(3):313–317.
- [Ceyssens et al., 2008] Ceyssens, P.-J., Mesyanzhinov, V., Sykilinda, N., Briers, Y., Roucourt, B., Lavigne, R., Robben, J., Domashin, A., Miroshnikov, K., Volckaert, G., et al. (2008). The genome and structural proteome of yua, a new pseudomonas aeruginosa phage resembling m6. *Journal of bacteriology*, 190(4):1429–1435.

- [Chan et al., 2014] Chan, J. Z.-M., Millard, A. D., Mann, N. H., and Schäfer, H. (2014). Comparative genomics defines the core genome of the growing n4-like phage genus and identifies n4-like roseophage specific genes. *Frontiers in microbiology*, 5:506.
- [Cheepudom et al., 2015] Cheepudom, J., Lee, C.-C., Cai, B., and Meng, M. (2015). Isolation, characterization, and complete genome analysis of p1312, a thermostable bacteriophage that infects *thermobifida fusca*. *Frontiers in microbiology*, 6.
- [Chevreux et al., 2004] Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E., Wetter, T., and Suhai, S. (2004). Using the miraest assembler for reliable and automated mrna transcript assembly and snp detection in sequenced ests. *Genome research*, 14(6):1147–1159.
- [Christiansen et al., 1996] Christiansen, B., Brøndsted, L., Vogensen, F. K., and Hammer, K. (1996). A resolvase-like protein is required for the site-specific integration of the temperate lactococcal bacteriophage tp901-1. *Journal of bacteriology*, 178(17):5164–5173.
- [Davenport et al., 2014] Davenport, K., Daligault, H., Minogue, T., Broomall, S., Bruce, D., Chain, P., Coyne, S., Gibbons, H., Jaissle, J., Li, P.-E., et al. (2014). Complete genome sequence of *stentrophomonas maltophilia* type strain 810-2 (atcc 13637). *Genome announcements*, 2(5):e00974–14.
- [Doolittle et al., 1995] Doolittle, M., Cooney, J., and Caldwell, D. (1995). Lytic infection of *escherichia coli* biofilms by bacteriophage t4. *Canadian Journal of Microbiology*, 41(1):12–18.
- [Doolittle et al., 1996] Doolittle, M., Cooney, J., and Caldwell, D. (1996). Tracing the interaction of bacteriophage with bacterial biofilms using fluorescent and chromogenic probes. *Journal of industrial microbiology*, 16(6):331–341.
- [Edgell et al., 2010] Edgell, D. R., Gibb, E. A., and Belfort, M. (2010). Mobile dna elements in t4 and related phages. *Virology journal*, 7(1):1.
- [Ehlers et al., 2002] Ehlers, A., Osborne, J., Slack, S., Roper, R. L., and Upton, C. (2002). Poxvirus orthologous clusters (pocs). *Bioinformatics*, 18(11):1544–1545.
- [Felsenstein, 1985] Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791.
- [Finn et al., 2011] Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, page gkr367.
- [Friedrich et al., 2007] Friedrich, N. C., Torrents, E., Gibb, E. A., Sahlin, M., Sjöberg, B.-M., and Edgell, D. R. (2007). Insertion of a homing endonuclease creates a genes-in-pieces ribonucleotide reductase that retains function. *Proceedings of the National Academy of Sciences*, 104(15):6176–6181.
- [Gogarten and Hilario, 2006] Gogarten, J. P. and Hilario, E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC evolutionary biology*, 6(1):94.
- [Gourley and Kuang, 2004] Gourley, S. A. and Kuang, Y. (2004). A delay reaction-diffusion model of the spread of bacteriophage infection. *SIAM Journal on Applied Mathematics*, 65(2):550–566.
- [Hatfull and Hendrix, 2011] Hatfull, G. F. and Hendrix, R. W. (2011). Bacteriophages and their genomes. *Current opinion in virology*, 1(4):298–303.
- [Hendrix, 2003] Hendrix, R. W. (2003). Bacteriophage genomics. *Current opinion in microbiology*, 6(5):506–511.

- [Hillary et al., 2011] Hillary, W., Lin, S.-H., and Upton, C. (2011). Base-by-base version 2: single nucleotide-level analysis of whole viral genome alignments. *Microbial informatics and experimentation*, 1(1):1.
- [Hyatt et al., 2010] Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1.
- [Jurado et al., 2014] Jurado, M., López, M. J., Suárez-Estrella, F., Vargas-García, M. C., López-González, J. A., and Moreno, J. (2014). Exploiting composting biodiversity: study of the persistent and biotechnologically relevant microorganisms from lignocellulose-based composting. *Bioresource technology*, 162:283–293.
- [Jurcisek et al., 2011] Jurcisek, J. A., Dickson, A. C., Bruggeman, M. E., and Bakaletz, L. O. (2011). In vitro biofilm formation in an 8-well chamber slide. *JoVE (Journal of Visualized Experiments)*, (47):e2481–e2481.
- [Katoh and Standley, 2013] Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780.
- [King, 2011] King, A. M. (2011). *Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses*, volume 9. Elsevier.
- [Koonin and Rudd, 1994] Koonin, E. V. and Rudd, K. E. (1994). A conserved domain in putative bacterial and bacteriophage transglycosylases. *Trends in biochemical sciences*, 19(3):106–107.
- [Kristensen et al., 2013] Kristensen, D. M., Waller, A. S., Yamada, T., Bork, P., Mushegian, A. R., and Koonin, E. V. (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *Journal of bacteriology*, 195(5):941–950.
- [Krysiak-Baltyn et al., 2015] Krysiak-Baltyn, K., Martin, G. J., Stickland, A. D., Scales, P. J., and Gras, S. L. (2015). Computational models of populations of bacteria and lytic phage. *Critical reviews in microbiology*, pages 1–27.
- [Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.
- [Laslett and Canback, 2004] Laslett, D. and Canback, B. (2004). Aragorn, a program to detect trna genes and tmrna genes in nucleotide sequences. *Nucleic Acids Research*, 32(1):11–16.
- [Le and Gascuel, 2008] Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320.
- [Lima-Junior et al., 2016] Lima-Junior, J. D., Viana-Niero, C., Oliveira, D. V. C., Machado, G. E., da Silva Rabello, M. C., Martins-Junior, J., Martins, L. F., Digiampietri, L. A., da Silva, A. M., Setubal, J. C., et al. (2016). Characterization of mycobacteria and mycobacteriophages isolated from compost at the são paulo zoo park foundation in brazil and creation of the new mycobacteriophage cluster u. *BMC microbiology*, 16(1):111.
- [López-González et al., 2015] López-González, J., Suárez-Estrella, F., Vargas-García, M., López, M., Jurado, M., and Moreno, J. (2015). Dynamics of bacterial microbiota during lignocellulosic waste composting: studies upon its structure, functionality and biodiversity. *Bioresource technology*, 175:406–416.

- [Marchler-Bauer et al., 2014] Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., et al. (2014). Cdd: Ncbi’s conserved domain database. *Nucleic acids research*, page gku1221.
- [Marks and Hamilton, 2014] Marks, T. J. and Hamilton, P. T. (2014). Characterization of a thermophilic bacteriophage of *geobacillus kaustophilus*. *Archives of virology*, 159(10):2771–2775.
- [Martins et al., 2013] Martins, L. F., Antunes, L. P., Pascon, R. C., de Oliveira, J. C. F., Digiampietri, L. A., Barbosa, D., Peixoto, B. M., Vallim, M. A., Viana-Niero, C., Ostroski, E. H., et al. (2013). Metagenomic analysis of a tropical composting operation at the sao paulo zoo park reveals diversity of biomass degradation functions and organisms. *PloS one*, 8(4):e61928.
- [Moak and Molineux, 2004] Moak, M. and Molineux, I. J. (2004). Peptidoglycan hydrolytic activities associated with bacteriophage virions. *Molecular microbiology*, 51(4):1169–1183.
- [Mosquera-Rendón et al., 2016] Mosquera-Rendón, J., Rada-Bravo, A. M., Cárdenas-Brito, S., Corredor, M., Restrepo-Pineda, E., and Benítez-Páez, A. (2016). Pangenome-wide and molecular evolution analyses of the *pseudomonas aeruginosa* species. *BMC genomics*, 17(1):1.
- [Neher et al., 2013] Neher, D. A., Weicht, T. R., Bates, S. T., Leff, J. W., and Fierer, N. (2013). Changes in bacterial and fungal communities across compost recipes, preparation methods, and composting times. *PLoS One*, 8(11):e79512.
- [Obeng et al., 2016] Obeng, N., Pratama, A. A., and van Elsas, J. D. (2016). The significance of mutualistic phages for bacterial ecology and evolution. *Trends in microbiology*, 24(6):440–449.
- [Paez-Espino et al., 2016] Paez-Espino, D., Eloie-Fadrosch, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering earth’s virome. *Nature*, 536(7617):425–430.
- [Partanen et al., 2010] Partanen, P., Hultman, J., Paulin, L., Auvinen, P., and Romantschuk, M. (2010). Bacterial diversity at different stages of the composting process. *BMC microbiology*, 10(1):1.
- [Pawar et al., 2015] Pawar, V., Komor, U., Kasnitz, N., Bielecki, P., Pils, M. C., Gocht, B., Moter, A., Rohde, M., Weiss, S., and Häussler, S. (2015). In vivo efficacy of antimicrobials against biofilm-producing *pseudomonas aeruginosa*. *Antimicrobial agents and chemotherapy*, 59(8):4974–4981.
- [Penn et al., 2010] Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., and Pupko, T. (2010). Guidance: a web server for assessing alignment confidence scores. *Nucleic acids research*, 38(suppl 2):W23–W28.
- [Rohwer, 2003] Rohwer, F. (2003). Global phage diversity. *Cell*, 113(2):141.
- [Ryckeboer et al., 2003] Ryckeboer, J., Mergaert, J., Vaes, K., Klammer, S., De Clercq, D., Coosemans, J., Insam, H., and Swings, J. (2003). A survey of bacteria and fungi occurring during composting and self-heating processes. *Annals of Microbiology*, 53(4):349–410.
- [Schmieder and Edwards, 2011] Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6(3):e17288.
- [Seemann, 2014] Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, page btu153.
- [Stadler and Fire, 2011] Stadler, M. and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. *Rna*, 17(12):2063–2073.
- [Stamatakis, 2014] Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

- [Stoddard, 2005] Stoddard, B. L. (2005). Homing endonuclease structure and function. *Quarterly reviews of biophysics*, 38(01):49–95.
- [Tcherepanov et al., 2006] Tcherepanov, V., Ehlers, A., and Upton, C. (2006). Genome annotation transfer utility (gatu): rapid annotation of viral genomes using a closely related reference genome. *BMC genomics*, 7(1):1.
- [Thorvaldsdóttir et al., 2012] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2012). Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, page bbs017.
- [Upton et al., 2000] Upton, C., Hogg, D., Perrin, D., Boone, M., and Harris, N. L. (2000). Viral genome organizer: a system for analyzing complete viral genomes. *Virus research*, 70(1):55–64.
- [Weinbauer, 2004] Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS microbiology reviews*, 28(2):127–181.
- [Whelan and Goldman, 2001] Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699.
- [Willis et al., 2002] Willis, S., Kazmierczak, K., Carter, R., and Rothman-Denes, L. (2002). N4 rna polymerase ii, a heterodimeric rna polymerase with homology to the single-subunit family of rna polymerases. *Journal of bacteriology*, 184(18):4952–4961.
- [Xia et al., 2011] Xia, L. C., Steele, J. A., Cram, J. A., Cardon, Z. G., Simmons, S. L., Vallino, J. J., Fuhrman, J. A., and Sun, F. (2011). Extended local similarity analysis (elsa) of microbial community and other time series data with replicates. *BMC systems biology*, 5(2):1.
- [Yoshida et al., 2013] Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T., and Takai, K. (2013). Metagenomic analysis of viral communities in (hado) pelagic sediments. *PLoS One*, 8(2):e57271.
- [Zhang and Young, 1999] Zhang, N. and Young, R. (1999). Complementation and characterization of the nested rz and rz1 reading frames in the genome of bacteriophage λ . *Molecular and General Genetics MGG*, 262(4-5):659–667.

CAPÍTULO 5

CONCLUSÃO

5.1 Contribuições da dissertação

Esta dissertação teve por objetivo principal o estudo e caracterização da diversidade de vírus em amostras de compostagem. Para tal, foram desenvolvidos dois protocolos computacionais de análises que são detalhadamente descritos na seção de métodos do capítulo 3. É nossa intenção disponibilizar os pipelines de forma *user-friendly* em breve.

O primeiro pipeline visa identificar espécies de vírus conhecidos a partir de dados de sequenciamento metagenômico e metatranscritômico de amostras ambientais. Após a identificação, o pipeline procede com o cálculo das abundâncias relativas e normalizadas, gerando saídas em formato tabular para análises estatísticas e de redes.

O segundo pipeline, intitulado MARVEL, tem como objetivo a recuperação de sequências virais longas a partir de dados de sequenciamento metagenômico e metatranscritômico de reads curtos. O pipeline MARVEL procede com a montagem das sequências curtas e identificação de fragmentos genômicos virais, fornecendo arquivos FASTA com as sequências recuperadas como saída. Como demonstrado pelos estudos de caso, algumas das sequências recuperadas pelo pipeline MARVEL são candidatas a genomas completos e/ou parciais de vírus ou provírus e podem ser posteriormente caracterizadas em detalhes de acordo com o objetivo do estudo em questão.

Os pipelines foram aplicados em dados de amostras seriadas de uma composteira do Zoológico de São Paulo e os resultados são apresentados nesta dissertação. Os dados de diversidade e abundância fornecem pistas importantes para o esclarecimento das dinâmicas ecológicas de microrganismos em suas comunidades ambientais, sobretudo os dados de co-ocorrências correlacionadas e análises de redes. De forma complementar, os fragmentos genômicos virais recuperados da compostagem permitem o acesso direto aos genes e proteínas sendo produzidas pelo viroma da comunidade ambiental, fornecendo evidências para o entendimento da influência destes organismos no perfil funcional do processo de compostagem. Juntos, perfil taxonômico e perfil funcional caracterizam de forma abrangente a diversidade no viroma da compostagem.

Na abordagem dependente de cultivo, isolamos três novos fagos de *Pseudomonas* provenientes de amostras da compostagem. Esses fagos foram caracterizados detalhadamente quanto aos genomas e anotação das proteínas putativas. Além disso, realizamos uma série de experimentos para comprovar resultados das análises computacionais e para auxiliar na caracterização geral dos fagos. Mostramos que os três fagos apresentam atividade lítica e dissolvem eficientemente biofilme de *Pseudomonas aeruginosa* PA14, com destaque especial para o potencial do fago ZC01. Essas informações podem

servir de base para estudos posteriores visando testar o potencial e aplicação clínica dos fagos como agentes antimicrobianos.

5.2 Direções futuras e pesquisas adicionais

Os pipelines desenvolvidos nesta dissertação têm o potencial para serem utilizados em qualquer conjunto de dados de sequenciamento metagenômico ou metatranscritômico via *Illumina short reads*. Ademais, outros conjuntos de dados também podem ser analisados com a implementação de adaptações simples na ferramenta. Para complementar um artigo futuro a ser publicado sobre a diversidade no viroma da compostagem, pretendemos:

- Aplicar os pipelines adicionalmente aos dados de sequenciamento metagenômico das composteiras ZC1, ZC2 e ZC3;
- Comparar os resultados obtidos para as diferentes composteiras no sentido de corroborar a consistência dos padrões observados;
- Aperfeiçoar as análises de correlação e redes para inferência de interações vírus/hospedeiro, vírus/vírus e vírus/fatores abióticos.

Além disso, os novos fagos isolados demonstraram grande potencial como agentes anti-microbianos e para dissolução de biofilmes de *Pseudomonas aeruginosa* PA14. Estudos adicionais precisam ser realizados para comprovar sua eficácia no tratamento de infecções hospitalares e combate a cepas resistentes.

APÊNDICE A

TABELAS DE ABUNDÂNCIAS

**Abundância relativa (%) dos 100 vírus mais abundantes no metagenoma da
composteira ZC4**

Espécie	D00	D01	D03	D07	D15	D30	D64	D67	D78	D99	Média/dia
Streptococcus phage 20617	1.978	2.098	2.180	2.329	2.568	2.021	2.177	1.899	1.564	1.674	2.049
Geobacillus virus E2	0.631	2.090	2.940	2.594	2.106	1.489	1.686	2.799	2.518	1.162	2.001
Torque teno virus 2	2.651	1.522	1.981	1.164	1.491	2.148	2.232	1.388	1.888	3.134	1.960
Clostridium phage phiMMP04	0.445	0.429	1.464	1.837	0.725	3.202	1.126	2.111	2.850	1.349	1.554
Enterobacteria phage fiAA91-ss	0.545	0.920	1.276	2.483	2.011	1.259	1.755	1.626	1.694	1.471	1.504
Microbacterium phage Min1	2.158	1.146	1.271	1.151	1.470	1.499	1.067	0.996	0.958	1.073	1.279
Geobacillus phage GBSV1	0.659	1.048	1.186	1.678	1.375	2.110	0.949	1.166	1.202	0.597	1.197
Bacillus virus 1	0.521	0.743	1.050	1.579	1.471	2.062	0.823	1.210	1.396	0.651	1.150
Lactococcus prophage bIL312	1.680	1.069	1.099	0.789	1.261	1.102	1.113	0.958	0.767	0.879	1.072
Burkholderia phage phiE125	1.027	0.940	0.940	0.721	1.177	1.126	1.172	0.871	1.210	1.331	1.052
Paenibacillus phage Davies	0.717	0.710	1.600	2.428	1.066	0.564	0.781	1.248	0.851	0.402	1.037
Enterobacteria phage If1	1.761	0.891	1.191	0.807	0.906	0.898	0.972	0.823	0.950	0.674	0.987
Thermoanaerobacterium phage THSA	0.215	0.343	0.832	1.437	1.036	1.524	0.651	1.292	1.664	0.836	0.983
Streptomyces phage SV1	1.102	0.639	1.277	0.709	0.733	0.724	0.446	1.102	1.099	1.131	0.896
Paenibacillus phage PG1	0.909	0.616	0.933	0.907	0.596	0.938	0.789	1.107	1.449	0.674	0.892
Geobacillus phage GBK2	0.253	0.547	0.772	1.833	1.821	0.824	0.694	1.090	0.581	0.312	0.873
Emiliana huxleyi virus 86	1.352	0.484	1.185	0.626	1.166	1.120	0.558	0.508	0.719	1.002	0.872
Phage phi OH2 DNA	0.316	0.572	1.264	1.503	1.135	0.480	0.740	0.954	0.973	0.517	0.845
Deep-sea thermophilic phage D6E	0.441	0.755	1.024	1.110	1.026	0.823	0.689	1.015	0.845	0.514	0.824
Streptomyces phage Sujidade	1.926	0.675	0.771	0.715	0.748	0.705	0.662	0.611	0.546	0.785	0.815
Streptomyces phage phiSASD1	1.296	0.591	0.431	0.275	0.365	0.709	0.728	0.604	1.253	1.734	0.799
Bacillus phage G	0.733	0.851	0.750	0.932	0.968	0.741	0.922	0.734	0.533	0.549	0.771
Streptomyces phage Zemlya	1.715	0.649	0.731	0.674	0.744	0.636	0.642	0.592	0.560	0.674	0.762
Streptococcus phage Dp-1	0.597	0.739	0.777	0.927	0.909	0.674	0.832	0.746	0.542	0.572	0.731
Lactococcus phage P087	0.562	0.892	0.789	0.884	0.823	0.604	0.859	0.713	0.493	0.505	0.712
Streptomyces phage Lika	1.455	0.594	0.675	0.640	0.697	0.634	0.622	0.567	0.529	0.704	0.712
Torque teno virus 3	0.585	0.488	0.429	0.282	0.479	0.849	1.061	0.539	0.856	1.316	0.688
Planktothrix phage PaV-LD	0.718	0.725	0.584	0.816	0.758	0.661	0.719	0.629	0.487	0.573	0.667
Synechococcus phage S-CBS3	1.030	0.434	0.624	0.656	0.721	0.862	0.601	0.508	0.374	0.584	0.639
Human herpesvirus 4	0.904	0.350	0.916	0.525	0.826	0.747	0.435	0.421	0.453	0.585	0.616

Streptococcus prophage 315.1	0.342	0.699	0.638	0.752	0.613	0.762	0.669	0.475	0.382	0.410	0.574
Shigella phage SfIV	0.248	0.840	0.666	0.500	0.653	0.707	0.617	0.543	0.480	0.456	0.571
Paenibacillus phage phiIBB_P123	0.361	0.302	0.615	1.240	0.524	0.611	0.464	0.555	0.437	0.264	0.537
Ralstonia phage phiRSA1	1.352	0.775	0.738	0.348	0.327	0.196	0.311	0.253	0.251	0.183	0.473
Streptomyce phage TG1	1.141	1.113	0.524	0.183	0.195	0.248	0.399	0.292	0.260	0.294	0.465
Lactococcus phage Q54	0.377	0.586	0.443	0.724	0.538	0.396	0.525	0.462	0.236	0.291	0.458
Bacillus phage IEBH	0.263	0.267	0.632	0.394	0.241	0.917	0.725	0.352	0.455	0.286	0.453
Stx2-converting phage 1717	0.341	0.486	0.329	0.238	0.492	0.539	0.503	0.352	0.337	0.705	0.432
Bacillus phage PBC1	0.221	0.355	0.474	0.520	0.611	0.313	0.439	0.526	0.495	0.200	0.415
Arthrobacter phage vB_ArS-ArV2	1.255	0.362	0.343	0.294	0.276	0.312	0.290	0.330	0.287	0.386	0.413
Staphylococcus phage StauST398-5	0.485	0.412	0.438	0.544	0.474	0.322	0.434	0.391	0.286	0.277	0.406
Bacillus phage phi105	0.177	0.176	0.311	0.547	0.410	0.610	0.353	0.556	0.452	0.268	0.386
Streptomyces phage VWB	0.668	0.455	0.307	0.183	0.209	0.383	0.209	0.276	0.365	0.792	0.385
Enterobacteria phage mEp235	0.063	0.238	0.206	0.297	0.374	0.458	0.260	0.599	0.732	0.444	0.367
Pseudomonas phage PAJU2	0.331	0.423	0.464	0.235	0.312	0.383	0.295	0.339	0.371	0.513	0.367
Cardioderma polyomavirus KY336	0.634	0.199	0.463	0.419	0.470	0.313	0.238	0.256	0.266	0.362	0.362
Vibrio phage X29	0.123	0.270	0.402	0.354	0.570	0.214	0.332	0.499	0.452	0.275	0.349
Rhodococcus phage REQ3	0.600	0.143	0.323	0.258	0.251	0.375	0.239	0.275	0.561	0.390	0.342
Human herpesvirus 4	0.518	0.175	0.519	0.277	0.462	0.388	0.251	0.217	0.244	0.304	0.336
Rhodococcus phage REQ1	0.696	0.120	0.341	0.165	0.203	0.254	0.165	0.155	0.333	0.895	0.333
Ralstonia phage RSY1 DNA	0.211	0.664	0.725	0.221	0.231	0.195	0.422	0.252	0.196	0.207	0.332
Cellulophaga phage phi14:2	0.411	0.276	0.356	0.348	0.402	0.293	0.332	0.337	0.240	0.295	0.329
Mycobacterium phage Bernal13	0.647	0.147	0.328	0.169	0.291	0.385	0.181	0.191	0.454	0.468	0.326
Pseudomonas phage Pf3	0.143	1.042	0.771	0.211	0.271	0.079	0.145	0.409	0.069	0.103	0.324
Streptomyces phage mu1/6	0.449	0.222	0.499	0.286	0.258	0.521	0.139	0.122	0.130	0.587	0.321
Bathycoccus sp. RCC1105 virus BpV1	0.403	0.304	0.324	0.311	0.379	0.333	0.328	0.283	0.233	0.287	0.319
Bacillus phage 0305phi8-36	0.190	0.400	0.360	0.439	0.342	0.252	0.371	0.319	0.209	0.199	0.308
Enterobacteria phage BP-4795	0.294	0.283	0.285	0.196	0.299	0.359	0.347	0.283	0.360	0.364	0.307
Clostridium phage phi3626	0.116	0.185	0.306	0.469	0.460	0.259	0.322	0.411	0.303	0.207	0.304
Bacillus phage Eoghan	0.015	0.169	0.221	0.605	0.363	0.118	0.220	0.890	0.330	0.083	0.301
Streptomyces phage phiBT1	0.707	0.568	0.308	0.143	0.100	0.147	0.305	0.245	0.238	0.230	0.299
Bacillus prophage phBC6A52	0.133	0.289	0.299	0.296	0.327	0.272	0.542	0.280	0.261	0.269	0.297
Actinoplanes phage phiAsp2	0.137	0.055	0.134	0.169	0.234	0.298	0.287	0.248	0.159	1.218	0.294
Bacillus phage SPBc2	0.165	0.289	0.327	0.382	0.336	0.270	0.375	0.369	0.233	0.173	0.292

Burkholderia phage KS9	0.285	0.311	0.273	0.225	0.272	0.287	0.304	0.303	0.302	0.289	0.285
Pseudomonas phage phi297	0.113	0.271	0.427	0.222	0.237	0.213	0.293	0.319	0.366	0.389	0.285
Torque teno virus 1	0.215	0.213	0.170	0.234	0.178	0.335	0.294	0.344	0.340	0.414	0.274
Bacillus phage Fah	0.148	0.157	0.287	0.425	0.316	0.151	0.253	0.431	0.367	0.180	0.271
Streptomyces phage R4	0.855	0.191	0.198	0.109	0.193	0.170	0.153	0.161	0.293	0.383	0.271
Bacillus phage BCJA1c	0.297	0.214	0.315	0.255	0.250	0.321	0.270	0.279	0.313	0.176	0.269
Streptomyces phage phiC31	0.672	0.552	0.281	0.098	0.100	0.115	0.219	0.179	0.189	0.157	0.256
Enterobacteria phage phi92	0.308	0.412	0.246	0.196	0.266	0.233	0.283	0.209	0.158	0.211	0.252
Xylella phage Xfas53	0.432	0.225	0.158	0.109	0.224	0.246	0.124	0.222	0.267	0.499	0.251
Grapevine fleck virus	0.213	0.181	0.280	0.176	0.319	0.398	0.193	0.181	0.228	0.317	0.249
Enterococcus phage phiFL4A	0.082	0.350	0.331	0.125	0.122	0.214	0.403	0.279	0.337	0.207	0.245
Clostridium phage vB_CpeS-CP51	0.236	0.292	0.244	0.318	0.200	0.098	0.369	0.261	0.263	0.166	0.245
Bacillus phage WBeta	0.127	0.143	0.249	0.364	0.276	0.134	0.225	0.374	0.318	0.156	0.237
Flavobacterium phage 6H	0.214	0.179	0.152	0.233	0.326	0.209	0.260	0.256	0.224	0.304	0.236
Bacillus prophage phBC6A51	0.124	0.230	0.253	0.364	0.297	0.156	0.309	0.320	0.196	0.108	0.236
Synechococcus phage S-CBS1	0.357	0.187	0.225	0.185	0.269	0.273	0.216	0.214	0.137	0.282	0.234
Clostridium phage phiCTP1	0.114	0.318	0.213	0.406	0.264	0.159	0.302	0.245	0.158	0.142	0.232
Cronobacter phage vB_CsaM_GAP32	0.284	0.287	0.212	0.222	0.254	0.236	0.252	0.191	0.157	0.200	0.230
Clostridium phage phiMMP02	0.023	0.126	0.250	0.682	0.380	0.083	0.152	0.351	0.178	0.067	0.229
Loktanella phage pCB2051-A	0.239	0.176	0.199	0.199	0.270	0.298	0.229	0.253	0.185	0.246	0.229
Salicola phage CGphi29	0.048	0.084	0.142	0.272	0.224	0.265	0.301	0.283	0.383	0.279	0.228
Helicobacter phage phiHP33	0.284	0.177	0.208	0.218	0.306	0.327	0.206	0.212	0.151	0.178	0.227
Myxococcus phage Mx8	0.162	0.105	0.234	0.156	0.283	0.217	0.157	0.283	0.427	0.212	0.224
Thermus phage P23-77	0.032	0.153	0.230	0.493	0.267	0.134	0.205	0.240	0.264	0.207	0.222
Micromonas pusilla virus 12T	0.243	0.200	0.208	0.217	0.268	0.238	0.242	0.212	0.167	0.222	0.222
Stenotrophomonas phage S1	0.120	0.303	0.312	0.224	0.191	0.161	0.196	0.208	0.268	0.225	0.221
Prochlorococcus phage P-SSM2	0.227	0.211	0.224	0.275	0.276	0.211	0.240	0.197	0.147	0.182	0.219
Vibrio phage pYD38-A	0.060	1.455	0.297	0.073	0.062	0.085	0.034	0.021	0.022	0.038	0.215
Bacillus phage Blastoid	0.018	0.113	0.153	0.419	0.261	0.072	0.160	0.678	0.210	0.049	0.213
Bacillus phage Cherry	0.101	0.118	0.218	0.297	0.251	0.118	0.151	0.381	0.342	0.154	0.213
Torque teno virus 10	0.322	0.090	0.096	0.142	0.175	0.320	0.245	0.166	0.276	0.288	0.212
Torque teno virus 12	0.345	0.193	0.099	0.114	0.172	0.268	0.366	0.155	0.141	0.254	0.211
Bacillus phage Gamma	0.100	0.119	0.216	0.297	0.246	0.116	0.152	0.377	0.331	0.150	0.210
Bacillus phage Finn	0.019	0.131	0.130	0.441	0.261	0.047	0.132	0.683	0.204	0.042	0.209

Bacillus phage SPP1	0.137	0.203	0.142	0.275	0.245	0.283	0.244	0.209	0.193	0.152	0.208
---------------------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Abundância relativa (%) dos 100 transcritos e genomas de RNA no metatranscritoma da composteira ZC4

Espécie	D01	D03	D07	D15	D30	D64	D78	D99	Média
Lactococcus prophage bIL312	7.332	16.730	11.339	10.280	5.594	7.589	18.717	2.457	10.005
Pleurotus ostreatus virus 1 RNA-2	1.458	2.461	2.048	3.570	14.487	11.271	17.353	21.248	9.237
Enterobacteria phage Hgal1	13.628	7.483	0.897	0.470	5.254	0.537	0.192	0.386	3.606
Enterobacteria phage FI	1.387	0.847	2.267	1.150	20.424	0.752	0.635	1.173	3.579
Streptococcus phage 20617	2.493	4.920	3.431	5.390	0.637	5.245	2.738	0.733	3.198
Penicillium stoloniferum virus F segment 1	0.256	0.187	0.266	0.150	1.790	2.231	1.118	18.886	3.111
Cellulophaga phage phi14:2	1.591	2.764	2.129	2.809	0.526	3.898	1.225	0.424	1.921
Rosellinia necatrix partitivirus 1-W8 segment 1	0.718	0.538	1.369	1.254	0.909	1.289	2.387	5.531	1.749
Enterobacteria phage Qbeta	2.365	1.189	1.042	0.486	6.119	0.550	0.360	0.862	1.622
Cellulophaga phage phi38:1	1.414	2.159	1.736	2.059	0.355	3.590	0.783	0.332	1.554
Sclerotinia sclerotiorum partitivirus S segment 1	0.127	0.091	0.037	4.449	2.882	1.377	1.459	0.966	1.423
Staphylococcus phage StauST398-5	0.443	1.453	2.836	2.913	0.318	1.377	1.562	0.070	1.372
Tetraselmis viridis virus S1	4.492	1.430	0.064	0.031	0.038	3.831	0.198	0.625	1.339
Fusarium poae virus 1 RNA 1	0.481	0.409	0.866	1.250	0.858	1.209	2.380	2.751	1.275
Enterobacterio phage MS2	1.790	0.559	0.621	0.466	5.480	0.510	0.013	0.103	1.193
Zucchini yellow mosaic virus	8.914	0.093	0.018	0.000	0.006	0.037	0.000	0.000	1.133
Pseudomonas phage PRR1	1.300	0.594	0.823	0.799	4.373	0.510	0.021	0.549	1.121
Geobacillus virus E2	0.997	1.602	1.875	2.095	0.204	1.146	0.731	0.112	1.095
Bacillus phage G	0.505	1.434	1.366	1.696	0.327	1.450	1.318	0.381	1.060
Synechococcus phage S-CBS3	0.415	1.238	1.130	1.268	0.425	0.889	2.237	0.466	1.009
Bacillus prophage phBC6A52	3.181	1.078	0.451	0.230	0.070	1.896	0.285	0.717	0.988
Enterobacteria phage GA	1.962	0.399	0.467	0.412	4.125	0.350	0.033	0.135	0.985
Bathycoccus sp. RCC1105 virus BpV1	0.559	1.262	0.919	0.939	0.262	1.947	0.969	0.367	0.903
Lactococcus phage P087	0.537	1.036	0.997	2.225	0.358	0.782	0.554	0.466	0.869
Bacillus phage SPBc2	0.525	1.425	1.189	1.202	0.242	0.647	0.513	0.110	0.732

Enterobacteria phage C-1 INW-2012	0.931	0.511	0.669	0.225	2.158	0.372	0.304	0.475	0.705
Micromonas pusilla virus 12T	0.463	1.027	0.735	0.647	0.208	1.331	0.744	0.200	0.669
Squash mosaic virus RNA 1	4.368	0.690	0.000	0.014	0.014	0.000	0.000	0.000	0.636
Geobacillus phage GBK2	0.258	0.660	1.512	1.432	0.090	0.695	0.333	0.044	0.628
Deep-sea thermophilic phage D6E	0.736	0.593	0.907	1.133	0.138	0.440	0.596	0.051	0.574
Bacillus prophage phBC6A51	0.367	0.823	0.944	0.867	0.162	0.643	0.721	0.041	0.571
Paenibacillus phage Davies	0.296	0.792	1.011	1.491	0.099	0.604	0.247	0.022	0.570
Caulobacter phage phiCb5	0.240	0.177	0.311	0.609	0.859	0.274	0.069	1.977	0.564
Enterobacteria phage M	0.682	0.444	0.599	0.282	0.706	0.442	0.178	0.991	0.540
Enterobacteria phage fiAA91-ss	0.139	0.534	1.495	0.827	0.213	0.193	0.742	0.062	0.526
Cryphonectria parasitica mitovirus 1-NB631	0.241	0.185	0.556	0.001	2.631	0.322	0.090	0.106	0.516
Diaporthe ambigua RNA virus 1	0.108	0.359	0.647	0.072	0.511	1.522	0.447	0.249	0.489
Mycobacterium phage Myrna	0.295	0.616	0.480	0.438	0.184	0.482	1.127	0.167	0.474
Cronobacter phage vB_CsaM_LGAP32	0.283	0.575	0.477	0.490	0.152	0.798	0.499	0.264	0.442
Brochothrix phage NF5	0.401	0.942	0.573	0.380	0.050	0.839	0.086	0.049	0.415
Streptococcus phage Dp-1	0.188	0.522	0.628	0.683	0.170	0.378	0.467	0.096	0.391
Rhizoctonia solani virus 717 RNA1	0.156	0.095	0.220	0.210	0.223	0.157	0.945	1.118	0.390
Microbacterium phage Min1	0.217	0.437	0.641	0.414	0.196	0.254	0.691	0.235	0.386
Cotesia congregata bracovirus segment Circle4	0.093	0.585	0.340	0.245	0.196	0.212	1.146	0.063	0.360
Synechococcus phage S-CBS1	0.136	0.402	0.449	0.492	0.142	0.288	0.830	0.138	0.360
Geobacillus phage GBSV1	0.626	0.420	0.657	0.744	0.034	0.203	0.106	0.030	0.353
Bacillus phage 0305phi8-36	0.316	0.616	0.537	0.501	0.114	0.374	0.303	0.050	0.352
Helicobacter phage phiHP33	0.126	0.599	0.632	0.734	0.058	0.363	0.210	0.050	0.347
Atkinsonella hypoxylon partitivirus RNA 1	0.000	0.004	0.014	0.045	0.163	0.303	0.490	1.643	0.333
Human picobirnavirus RNA segment 1	2.009	0.139	0.193	0.042	0.000	0.022	0.000	0.000	0.301
Bacillus virus 1	0.470	0.246	0.506	0.747	0.037	0.161	0.147	0.032	0.293
Thermoanaerobacterium phage THSA-485A	0.040	0.148	0.590	1.248	0.038	0.121	0.132	0.029	0.293
Hop trefoil cryptic virus 2 IPP_GelbSK RNA 1	0.052	0.047	0.110	0.191	0.177	0.185	0.508	0.985	0.282
Pseudomonas phage PP7	0.233	0.024	0.052	0.032	0.067	0.025	0.000	1.816	0.281
Cyanophage KBS-S-2A	0.114	0.337	0.313	0.335	0.074	0.193	0.808	0.067	0.280
Vibrio phage henriette 12B8	0.291	0.490	0.246	0.279	0.098	0.428	0.278	0.103	0.277
Bacteriophage Aaphi23	0.233	0.376	0.276	0.325	0.068	0.488	0.366	0.077	0.276
Red clover cryptic virus 2 IPP_Nemaro RNA 1	0.105	0.084	0.133	0.179	0.195	0.256	0.517	0.731	0.275
Phage phi OH2 DNA	0.193	0.322	0.511	0.760	0.052	0.192	0.115	0.025	0.271

Lactococcus phage Q54	0.148	0.417	0.406	0.299	0.060	0.355	0.241	0.101	0.253
Planktothrix phage PaV-LD	0.104	0.254	0.367	0.407	0.124	0.268	0.359	0.131	0.252
Flavobacterium phage 6H	0.156	0.316	0.292	0.389	0.074	0.270	0.342	0.055	0.237
Streptococcus prophage 315.1	0.197	0.344	0.458	0.399	0.048	0.290	0.127	0.025	0.236
Bacillus phage Fah	0.148	0.510	0.348	0.311	0.041	0.185	0.247	0.024	0.227
Botryotinia fuckeliana partitivirus 1 RNA3	0.658	0.209	0.043	0.000	0.021	0.098	0.009	0.615	0.207
Bacillus phage WBeta	0.134	0.441	0.302	0.269	0.036	0.168	0.213	0.021	0.198
Mycobacterium phage DrDrey	0.094	0.237	0.265	0.318	0.060	0.250	0.181	0.164	0.196
Primula malacoides virus China/Mar2007 RNA 1	0.072	0.063	0.110	0.233	0.139	0.159	0.325	0.458	0.195
Acinetobacter phage AP205	0.716	0.390	0.356	0.023	0.000	0.052	0.000	0.000	0.192
Enterobacteria phage vB_KleM-RaK2	0.133	0.233	0.175	0.174	0.087	0.368	0.191	0.155	0.190
Bacillus phage PM1 DNA	0.195	0.460	0.146	0.224	0.015	0.379	0.066	0.028	0.189
Listeria phage LP-030-3	0.172	0.393	0.226	0.234	0.037	0.322	0.093	0.017	0.187
Bacillus phage BCD7	0.146	0.220	0.308	0.306	0.080	0.306	0.092	0.017	0.185
White clover cryptic virus 2 IPP_Lirepa RNA 1	0.105	0.027	0.138	0.195	0.112	0.134	0.248	0.502	0.183
Cafeteria roenbergensis virus BV-PW1	0.109	0.254	0.176	0.174	0.065	0.318	0.195	0.157	0.181
Mycobacterium phage Kostya	0.082	0.208	0.239	0.291	0.055	0.225	0.157	0.160	0.177
Burkholderia phage phiE255	0.129	0.206	0.233	0.207	0.081	0.224	0.139	0.126	0.168
Shigella phage SfIV	0.087	0.132	0.178	0.133	0.123	0.135	0.363	0.185	0.167
Gremmeniella abietina mitochondrial RNA virus S2	0.242	0.201	0.502	0.000	0.180	0.037	0.171	0.000	0.167
Burkholderia phage phiE125	0.054	0.132	0.267	0.116	0.102	0.169	0.291	0.143	0.159
Cardioderma polyomavirus isolate KY336	0.038	0.123	0.317	0.344	0.085	0.095	0.232	0.036	0.159
Botrytis cinerea debilitation-related virus	0.001	0.147	0.370	0.014	0.052	0.084	0.549	0.041	0.157
Bacillus phage Slash	0.142	0.262	0.235	0.199	0.046	0.215	0.138	0.015	0.157
Gremmeniella abietina RNA virus L1	0.000	0.000	0.000	0.000	0.002	0.020	0.000	1.222	0.156
Streptomyces phage Sujidade	0.080	0.142	0.302	0.224	0.058	0.121	0.242	0.057	0.153
Burkholderia phage BcepMu	0.121	0.193	0.216	0.202	0.066	0.191	0.135	0.094	0.152
Clostridium phage phi3626	0.045	0.228	0.278	0.248	0.051	0.123	0.206	0.022	0.150
Abalone shriveling syndrome-associated virus	0.046	0.213	0.247	0.312	0.074	0.096	0.177	0.019	0.148
Streptomyces phage Zemlya	0.074	0.142	0.317	0.229	0.044	0.116	0.194	0.050	0.146
Shrimp white spot syndrome virus	0.092	0.135	0.083	0.051	0.058	0.055	0.068	0.618	0.145
Ophiostoma mitovirus 4	0.159	0.168	0.438	0.041	0.175	0.053	0.111	0.012	0.145
Ophiostoma mitovirus 5	0.230	0.146	0.183	0.000	0.155	0.025	0.402	0.000	0.143
Streptomyces phage Lika	0.068	0.127	0.299	0.209	0.047	0.113	0.205	0.051	0.140

Staphylococcus prophage phiPV83	0.266	0.318	0.091	0.047	0.015	0.326	0.048	0.007	0.140
Saccharomyces 20S RNA narnavirus	0.113	0.257	0.180	0.000	0.226	0.048	0.000	0.275	0.138
Listeria phage A006	0.153	0.266	0.173	0.150	0.014	0.271	0.041	0.024	0.137
Phaeocystis globosa virus strain 16T	0.057	0.164	0.150	0.199	0.052	0.171	0.164	0.105	0.133
Streptomyces phage SV1	0.084	0.072	0.473	0.111	0.040	0.052	0.178	0.043	0.132
Clostridium phage vB_CpeS-CP51	0.088	0.188	0.220	0.199	0.074	0.088	0.160	0.028	0.131
Ceratocystis polonica partitivirus segment 1	0.000	0.005	0.001	0.037	0.086	0.104	0.195	0.606	0.129