Leandro Nascimento Lemos

Reconstrução e análise de genomas de
bactérias de compostagem a partir de
dados metagenômicos

São Paulo
2015

Leandro Nascimento Lemos

Reconstrução e análise de genomas de bactérias de compostagem a partir de dados metagenômicos

Dissertação apresentada ao Instituto de
Matemática e Estatística da Universidade
de São Paulo, para a obtenção de Título
de Mestre em Ciências, na Área de Con-
centração em Bioinformática.

Orientador: Prof. Dr. João Carlos Setubal

São Paulo
2015

Lemos, Leandro N.

Reconstrução e análise de genomas de bactérias de compostagem a partir de dados metagenômicos

59 páginas

Dissertação (Mestrado) - Instituto de Matemática e Estatística da Universidade de São Paulo.

1. Genomas microbianos
2. Bioinformática
3. Metagenômica

I. Universidade de São Paulo. Instituto de Matemática e Estatística.

*Dedico este trabalho ao meu pai Arize da Silva Lemos e à minha mãe Maria Helena
Silva Nascimento.*

"Because in the end, you won't remember the time you spent working in the office or mowing your lawn. Climb that goddamn mountain!"

Jack Kerouac

Agradecimentos

Ao meu pai e à minha mãe, pelo carinho e afeto. E por sempre estarem presentes comigo.

Aos meus avós, irmãos, sobrinhos e tios, pelo constante apoio. Em especial a minha tia Neigmar Lemos, pelo incentivo nos estudos e suporte financeiro nos cinco meses que eu fiquei sem bolsa.

Ao Prof. João Carlos Setubal, pela oportunidade de estudos no Laboratório de Bioinformática, pelos ensinamentos e conselhos durante esses dois anos de mestrado.

Ao Prof. Ronaldo Quaggio (Instituto de Química - Departamento de Bioquímica/USP) pela colaboração em pesquisa, conversas sobre cinema e política nos corredores do IQ e pela construção do consórcio microbiano.

Aos colegas Luciana Antunes Principal e Gianluca Major, pela geração dos dados metagenômicos da compostagem e tratamento inicial dos dados de sequenciamento.

Aos amigos Bruna Bianca Izaguirry, André Agnoletto e Belinha, pela confiança e amizade.

Aos amigos com quem eu dividi moradia durante o tempo em que eu morei na capital de São Paulo. Em especial aos amigos do aptBrover (David Romy II, Douglas Batista, Álvaro Ramos e Brover Romy), pela parceria e apoio.

Aos meus ídolos do Heavy Metal (Rafael Bittencourt, Kai Hansen, Jon Schaffer, Andre Matos, Joey DeMaio, Michael Kiske, Rob Halford e Ozzy Osbourne), pela influência da música na minha vida e pelos shows em que eu pude vê-los ao vivo nesse tempo de moradia em São Paulo. E ao imortal Ronnie James Dio!!!

Aos colegas do Setulab, pela amizade e ensinamentos. E principalmente pelas discussões sobre "a vida, o universo e tudo mais" na hora do café.

Aos colaboradores do Projeto Metazoo, pela oportunidade de colaborações em pesquisa.

À FAPESP pela bolsa concedida (processo 13/05325-5).

A Universidade de São Paulo (USP) e ao Programa de Pós-Graduação em Bioinformática pela oportunidade de realização do Mestrado.

Resumo

Na última década tem sido possível reconstruir o genoma de bactérias e arqueias presentes em comunidades microbianas de ambientes naturais a partir de dados metagenômicos. Isso tem revolucionado nosso entendimento sobre a topologia da árvore da vida e a descoberta de novas capacidades metabólicas, bem como auxiliado na identificação mais acurada de genes de interesse industrial, visto que os dados estão mais completos e menos fragmentados. Com base neste contexto, o objetivo geral deste projeto foi reconstruir o genoma de bactérias ligadas a degradação de biomassa vegetal em comunidades microbianas da compostagem, focando em análises de diversidade de enzimas de Glicosil Hidrolases (GHs), a partir de dados de sequências metagenômicas gerados no projeto temático processo 11/50870-6. Para alcançar os nossos objetivos, foram desenvolvidos pipelines computacionais com softwares já disponíveis na literatura e foram utilizados dois conjuntos principais de dados de sequenciamento massivo (um conjunto de dados seriados que engloba inúmeros estágios do processamento da compostagem e um conjunto de dados do metagenoma de um consórcio microbiano celulolítico e termofílico construído a partir de amostras da compostagem). Foram reconstruídos 13 genomas (sete genomas em amostras dos dados seriados e seis genomas na amostra do consórcio microbiano), sendo identificado no mínimo quatro novas espécies. As análises baseadas em filogenômica indicam a presença de pelo menos uma nova classe dentro do filo Firmicutes, uma nova espécie da família Paenibacillaceae e a reconstrução pela primeira vez do genoma da espécie *Bacillus thermozeamaize*. Também foram identificadas 33 lacunas/ilhas metagenômicas (IMs). Essas regiões apresentaram genes diretamente ligados a biossíntese de polissacarídeos do envelope celular, pseudogenes e proteínas hipotéticas. Algumas dessas proteínas estão diretamente ligadas ao reconhecimento de bacteriófagos durante a fase de infecção viral. A presença de IMs também indica uma

divergência entre as populações microbianas presentes na compostagem com a espécie de referência. Quanto ao potencial de degradação de biomassa vegetal, todos os microrganismos apresentam genes com potencial para degradação de material lignocelulolítico durante o processamento de diferentes estágios da compostagem, indicando a importância do papel funcional dessas bactérias na compostagem.

Palavras-chave: Genomas microbianos, Bioinformática, Metagenômica

Abstract

In the last decade it has been possible to reconstruct Bacteria and Archaea genomes that are in natural microbial communities from metagenomic samples. This has revolutionized our understanding of the topology of the tree of life and the discovery of new metabolic functions, as well as aided in more accurate identification of industrial bioprospecting genes, since the genomic data are more complete and less fragmented. Based on this background, the aim of this project was to reconstruct the bacterial genomes linked to plant biomass degradation in composting communities, focusing on diversity analysis of Glycosyl Hydrolases (GHs) from metagenomic sequence data generated in the Thematic Project (Process 11/50870-6). To achieve our objectives, computational pipelines have been developed (this pipelines were based on software already available in the literature) and we use these pipelines in two massive data sets generated by high-throughput sequencing (one data set of time series compost sample which includes several stages of the composting process and other data set from a cellulolytic and thermophilic microbial consortium). Thirteen genomes were reconstructed (seven genomes from time series metagenomic data and six genomes from microbial consortium). At least four new species have been identified, and the analyzes based on phylogenomic inferences indicate the presence of at least one new class of Firmicutes phylum, and a new Paenibacillaceae family and the reconstruction for the first time the *Bacillus thermozeamaize* genome. They also identified 33 gaps/metagenomic Islands (IMs). These gaps had genes directly linked to polysaccharide biosynthesis of the cell envelope, pseudogenes and hypothetical proteins. Some of these proteins are directly linked to the bacteriophage during the recognition phase of viral infection. The presence of gaps also indicates a divergence between microbial populations present in the compost with the reference genome. All microbial genomes reconstructed in this study

have genes linked to lignocellulolytic potential degradation during the different stages of composting process, indicating the functional role this bacteria in this environment.

Keywords: Microbial genomes, Bioinformatics, Metagenomics

Sumário

1	Introdução	1
1.1	Reconstrução e recrutamento de genomas a partir de metagenomas . . .	1
1.2	Contexto do trabalho	2
1.3	Motivação	3
1.4	Breve descrição dos objetivos alcançados	4
2	Embasamento do trabalho e revisão da literatura	7
2.1	Bactérias abundantes em comunidades microbianas	7
2.2	Métodos e ferramentas computacionais para a reconstrução de genomas microbianos a partir de dados metagenômicos	9
2.2.1	Recrutamento de sequências	9
2.2.2	Binning	10
2.3	Ilhas Metagenômicas	12
2.4	Glicosil hidrolases	12
2.5	Reconstrução de genomas	13
3	Material e Métodos	16
3.1	<i>Pipelines</i> para a reconstrução de genomas a partir de dados metagenômicos	16
3.2	Dados do metagenoma da compostagem (ZC4)	17
3.2.1	Pré-processamento de sequências brutas	19

3.2.2	Recrutamento e montagem dos genomas de bactérias abundantes	19
3.2.3	Identificação de Ilhas Metagenômicas	21
3.2.4	Anotação	21
3.2.5	Caracterização filogenética	22
3.3	Dados do metagenoma do consórcio microbiano da compostagem (ZCTH02)	22
3.3.1	Desenho experimental e sequenciamento	22
3.3.2	Montagem <i>De novo</i> de genomas microbianos	22
3.3.3	Anotação	23
3.3.4	Caracterização filogenética e e filogenômica	23
4	Resultados e Discussão	25
4.1	Genomas recrutados em dados do metagenoma da compostagem (ZC4) .	25
4.1.1	Variações pontuais nos genomas recrutados em ZC4	25
4.1.2	Ilhas Metagenômicas	37
4.2	Consórcio microbiano termofílico e celulolítico (ZCTH02)	39
4.2.1	Posicionamento filogenético dos genomas reconstruídos	39
4.2.2	Glicosil-hidrolases identificadas no consórcio microbiano	42
4.2.3	Anotação Funcional dos genomas reconstruídos do consórcio mi- crobiano	44
5	Conclusões	48
	Referências Bibliográficas	50

Lista de Figuras

2.1	Curva de abundância. Um pequeno número de espécies microbianas são dominantes (abundantes), e o restante são espécies raras. A relação entre espécies abundantes e o ranqueamento é demonstrado na linha azul. Abundância pode mudar com o passar do tempo por inúmeros fatores (linhas pretas). Fonte da imagem: (1)	8
2.2	Etapas de processamento computacional do software MaxBin. O processamento é dividido em duas etapas principais: I. geração de informações de entrada para o software. O nível de cobertura de sequenciamento de cada contigs/ou scaffold, e o perfil de frequências de tetranucleotídeos é calculado pelo usuário. A identificação de genes marcadores de cópia-única é automatizada pelo software MaxBin. A segunda etapa consiste em identificar as populações microbianas (bin) e validar cada população pela presença de genes marcadores de cópia-única. O produto final são populações que representam genomas individuais presentes no metagenoma. Imagem modificada de Wu et al. (2)	11

3.1	Overview das rotinas computacionais utilizadas neste projeto.	
	As análises são divididas inicialmente em recrutamento e montagem <i>De novo</i> de sequências, seguido de identificação de Ilhas Metagenômicas, anotação, identificação de genes de interesse, diversidade das principais famílias de GHs e caracterização filogenômica.	18
4.1	Ilha Metagenômica (IM) de <i>Rhodothermus marinus</i>. O nome de cada produto proteico estão localizados perto das caixas laranjas. . . .	39
4.2	Análise filogenômica dos genomas reconstruídos em amostras de um consórcio termofílico e celulotítico da compostagem e representantes do filo Firmicutes, outros filos de Bacteria e Archaea. Os genomas da compostagem são representados pela sigla ZCTHBIN[1-6]. As cores representam os grupos taxonômicos atuais. A análise filogenômica foi realizada com base no alinhamento de 300 proteínas conservadas e a reconstrução filogenética foi baseada em análises de Máxima Verossimilhança.	45
4.3	Abundância de categories COGs nos genomas reconstruídos do metagenoma de um consórcio termofílico e celulolítico da compostagem.	47

Lista de Tabelas

3.1	Número de sequências de cada amostra do metagenoma da compostagem (ZC4) e do consórcio microbiano (ZCTH02).	19
3.2	Características gerais dos genomas usados como referência nas etapas de recrutamento de sequências em ZC4.	20
4.1	Número de sequências candidatas recrutadas com um critério mínimo de 85% de identidade.	26
4.2	Principais características dos genomas de espécies abundantes reconstruídos a partir de análises de recrutamento de sequências em amostras do metagenoma da compostagem (ZC4).	27
4.3	Análise comparativa de genes marcadores de cópia-única presentes nos genomas microbianos das sete espécies mais abundantes da microbiota da compostagem (ZC4).	31
4.4	Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de <i>Pseudoxanthomas suwonensis</i> em amostras de compostagem	32
4.5	Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de <i>Rhodothermus marinus</i> em amostras de compostagem	33
4.6	Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de <i>Thermobispora bispora</i> em amostras de compostagem.	34

4.7	Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de <i>Sphaerobacter thermophilus</i> em amostras de compostagem	35
4.8	Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de <i>Symbiobacterium thermophilum</i> em amostras de compostagem	35
4.9	Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de <i>Thermobifida fusca</i> em amostras de compostagem. . .	36
4.10	Descrição de Ilhas Metagenômicas presentes em espécies abundantes que foram classificadas no metagenoma da compostagem (ZC4).	38
4.11	Genomas microbianos reconstruídos a partir de um consórcio microbiano de compostagem (ZCTH)	41
4.12	Diversidade e abundância de GHs presentes nos genomas reconstruídos em um consórcio microbiano termofílico e celulolítico da compostagem .	46

Capítulo 1

Introdução

1.1 Reconstrução e recrutamento de genomas a partir de metagenomas

Na última década tem sido possível reconstruir o genoma de bactérias e arqueias presentes em comunidades microbianas de ambientes naturais a partir de dados metagenômicos. Isso tem revolucionado o nosso entendimento sobre a topologia da árvore da vida e a descoberta de novas capacidades metabólicas desempenhadas por esses microrganismos, nos permitindo inferir suas funções nos microbiomas onde estão presentes.

Com base neste contexto, a proposta geral desta dissertação foi de aplicar metodologias conhecidas para reconstruir computacionalmente os genomas de bactérias encontradas em um ambiente de compostagem. Nossa expectativa foi de que os microrganismos mais abundantes nesse ambiente (e portanto para os quais temos mais dados genômicos) são relacionadas a degradação de biomassa vegetal, e portanto tivemos como objetivo também fazer uma análise do potencial metabólico para degradação de biomassa dos genomas reconstruídos, analisando em particular as famílias de glicosil hidrolases (GHs).

Para alcançar estes objetivos utilizamos dados gerados no projeto temático "Estudos

da diversidade microbiana no Parque Zoológico do Estado de São Paulo" e softwares já disponíveis na literatura.

1.2 Contexto do trabalho

O projeto temático acima citado é coordenado pelo Prof. Dr. João Carlos Setubal (orientador do aluno) e pela Profa. Dra. Aline Maria da Silva, e tem como objetivo geral coletar, analisar e prospectar dados moleculares de três microbiomas existentes no Parque Zoológico de São Paulo: compostagem vegetal da mata atlântica, lago, e fezes de macacos bugios. Esses três microbiomas contemplam a diversidade de ambientes no parque e a missão de conservação de animais da Fundação Parque Zoológico de São Paulo (FPZSP).

A FPZSP localiza-se em uma área de 900.000 m^2 de Mata Atlântica no município de São Paulo e possui uma Unidade de Compostagem (UPCO) que processa matéria orgânica de várias origens (excremento de aproximadamente 4.000 animais da fauna nativa e exótica, carcaças de animais, etc). Ao final do processo a matéria orgânica é convertida em adubo, que é destinada à fertilização de áreas agrícolas do Zoológico para a produção de alimentos que são consumidos pelos animais do Parque, fechando assim um ciclo de sustentabilidade.

Um processamento típico de compostagem compreende uma série de estágios em que ocorrem aumento de temperatura (até 70°C), manutenção de temperatura elevada por longos dias (vários dias) e resfriamento gradual do material. Outros fatores como gradiente de oxigênio, pH e a disponibilidade de nutrientes também variam durante o processo, que gera como subprodutos ácidos graxos voláteis, amônia, dióxido de carbono e ácidos húmicos. No início da compostagem atuam bactérias mesófilas, actinomicetos, fungos e protozoários, que crescem entre 10 e 45°C e degradam substâncias menos complexas. A ação oxidante destes microrganismos resulta no aumento de temperatura

e queda do pH, causada pela formação de ácidos graxos voláteis, sendo que a degradação de ácidos resulta em subsequente aumento de pH. Quando a temperatura eleva-se acima de 45°C, microrganismos termofílicos substituem os mesofílicos. A maioria da matéria orgânica é degradada nesta etapa e, conseqüentemente, mais oxigênio é consumido. A degradação de lignina também começa nessa fase. Após a fase termofílica, que corresponde ao pico de degradação da matéria orgânica fresca, a atividade microbiana diminui, assim como a temperatura. Durante esta fase, os microrganismos mesofílicos colonizam a compostagem e lentamente degradam compostos orgânicos mais complexos, tais como a lignina. Nesta última fase são produzidas as substâncias húmicas para formar o “composto” maduro (3; 4).

Esta combinação de variáveis afeta as populações microbianas e suas funções na compostagem, estabelecendo, assim, um mecanismo de *feedback* entre os sistemas bióticos e abióticos, o que certamente impacta na estrutura da comunidade que habita o substrato orgânico. Além disso, é plausível supor que as populações e produtos enzimáticos de origem microbiana variem dependendo dos substratos e das condições em que é realizada a compostagem.

1.3 Motivação

Um dos objetivos do projeto temático é estudar o processo de compostagem pela técnica de metagenômica. A compostagem é um processo onde ocorre intensa degradação de biomassa, ao longo de 3 meses (no caso da UPCO). Disso deriva nosso interesse por glicosil hidrolases (GHs). Essas enzimas têm a função de desconstruir a parede celular da biomassa vegetal durante as diferentes fases da compostagem (5). Apesar da grande quantidade de material vegetal disponível em uma base renovável e sustentável, tal como uma composteira, a produção econômica de açúcares fermentáveis a partir da biomassa lignocelulósica ainda é dificultada por vários fatores. Esses fatores incluem as fases de

pré-tratamento, o alto custo de produção de biomassa, e a baixa eficiência catalítica das enzimas que estão disponíveis no mercado. Identificar novas glicosil hidrolases com alta eficiência catalítica, ou variações genéticas nas já existentes, é um passo necessário para a redução dos custos de produção de biocombustíveis (6).

Identificação acurada de quaisquer genes e sua análise detalhada a partir de dados de sequenciamento requerem genomas completos ou quase completos. Isto à primeira vista representa um obstáculo, dado que de um modo geral a grande maioria das sequências gênicas identificadas nos *contigs* de nossas montagens é apenas parcial, como pudemos verificar ao fazer comparação com seus homólogos em bancos de sequências. Essa situação é causada pela alta diversidade de espécies presentes nas amostras (7) e pela cobertura relativamente baixa do sequenciamento, mesmo este gerando milhões de reads. Por outro lado, sabemos que estão presentes na compostagem alguns microrganismos em grande abundância. Isto faz com que seja possível recuperar frações significativas de seus genomas mesmo na situação de cobertura *média* relativamente baixa. Assim sendo, decidimos fazer o levantamento das GHs a partir dos genomas completos ou quase completos que poderemos recuperar dos dados metagenômicos.

Adicionalmente aos dados metagenômicos de comunidades microbianas da compostagem, o uso de um consórcio microbiano termofílico e celulolítico enriquecido com material da compostagem nos ajudou a selecionar microrganismos que apresentam funções específicas de degradação de biomassa vegetal, bem como, reduzir a alta diversidade presente nas comunidades microbianas da compostagem e reconstruir genomas a partir de abordagens *de novo* (abordagens sem o uso de genomas de referência).

1.4 Breve descrição dos objetivos alcançados

- Durante o andamento deste projeto de pesquisa foram desenvolvidos pipelines computacionais para automatizar a reconstrução de genomas microbianos e iden-

tificação de glicosil hidrolases em amostras do metagenoma da compostagem a partir de dados de sequenciadores de nova geração. Este pipeline foi desenvolvido com o uso de softwares já disponíveis na literatura. A automatização das rotinas computacionais foi realizada nas linguagens de programação Perl, Python, R e Shell. Com a aplicação destes pipelines foi possível reconstruir o genoma de 13 espécies bacterianas que apresentam um amplo repertório gênico ligado a funções de degradação de biomassa vegetal. Esses genes apresentaram variações com base em análises comparativas realizadas com as espécies mais próximas, cujo genoma está depositado em banco de dados públicos.

- Alguns genomas reconstruídos neste estudo são de espécies novas, incluindo uma possível classe nova dentro do filo Firmicutes e uma espécie nova de *Thermobacillus*.

Capítulo 2

Embasamento do trabalho e revisão da literatura

2.1 Bactérias abundantes em comunidades microbianas

Quantificar a diversidade taxonômica é um dos primeiros passos para caracterizar uma comunidade microbiana. Isso envolve determinar quais microrganismos estão presentes em uma comunidade (e.g., riqueza de espécies) e suas respectivas abundâncias. A diversidade taxonômica serve como um caminho para analisar o perfil de uma comunidade ecológica e essas análises podem indicar similaridades entre diferentes microbiomas (e.g., comunidade com mais taxa compartilhados são mais similares). (8).

Os primeiros estudos envolvendo análises de diversidade taxonômica a partir do uso de sequenciadores de larga escala identificaram alguns padrões de distribuição de abundância em comunidades microbianas (9). Sogin e colaboradores (9) usaram análises de curvas de abundância (representações gráficas de abundância de taxa presentes em uma comunidade ecológica) e identificaram a presença de um enorme número de taxa presentes em uma porcentagem extremamente pequena em virtualmente quase todas as comunidades analisadas (9; 10). Esses gráficos também indicam a presença de um baixo

número de táxons dominantes, que apresentam uma alta abundância nesses ambientes (Figura 2.1). (9; 1).

Figura 2.1: **Curva de abundância.** Um pequeno número de espécies microbianas são dominantes (abundantes), e o restante são espécies raras. A relação entre espécies abundantes e o ranqueamento é demonstrado na linha azul. Abundância pode mudar com o passar do tempo por inúmeros fatores (linhas pretas). Fonte da imagem: (1)

Bactérias abundantes e raras podem apresentar características e funções biológicas distintas. Por exemplo, uma hipótese plausível é que os microrganismos marinhos raros poderiam aumentar em abundância após uma perturbação ambiental, mantendo o funcionamento contínuo deste ecossistema (10). Em termos ecológicos, as bactérias abundantes contribuem principalmente para a biomassa, fluxo de carbono e ciclagem de nutrientes, enquanto que um grande número de bactérias raras contribuem principalmente para a riqueza de espécies (11).

2.2 Métodos e ferramentas computacionais para a reconstrução de genomas microbianos a partir de dados metagenômicos

Quatro tipos principais de abordagens computacionais são utilizadas para quantificar a diversidade taxonômica. Isso envolve determinar quais microrganismos estão presentes na comunidade microbiana (por exemplo, riqueza de espécies) e determinar qual é a sua respectiva abundância. Essas abordagens são divididas em: (a) análises de genes marcadores, (b) *binning*, (c) classificação e (d) montagem (12). Dentro dessas categorias estabelecidas por Sharpton (12), com exceção das análises diretas de genes marcadores, todas as outras categorias são aplicadas para a reconstrução de genomas microbianos em dados metagenômicos.

No contexto genômico, os processos computacionais de *binning* e classificação são utilizados em análises de genomas completos, ou parcialmente completos, em dois tipos principais: análises com genoma de referência (classificação) e análises sem genoma de referência (*binning*).

2.2.1 Recrutamento de sequências

Os métodos de classificação nos permitem recrutar sequências específicas de conjuntos de dados metagenômicos, sendo possível identificar variações nessas sequências com base em alinhamentos contra os genomas de referências disponíveis em bancos de dados públicos (13; 14).

As ferramentas computacionais Bowtie2 (15) e Frhit (16) têm sido utilizadas para o recrutamento de sequências e posterior montagem de genomas microbianos. Por exemplo, a ferramenta Frhit utiliza a construção de tabelas *hash* de *k*-mers para a compactação do banco de referências. A aplicação desse tipo de estratégia reduz a complexidade computacional e o tempo de processamento de dados, sendo possível

alinhar milhares de sequências em um tempo reduzido (16).

Ao final do recrutamento de sequências os genomas são montados com o uso de softwares de montagem *de novo* de sequências. Quanto a questão de montagem, inúmeros softwares estão disponíveis na literatura (17; 18; 19; 20).

2.2.2 Binning

A estratégia conhecida como *binning* é utilizada para agrupar sequências pelas suas características composicionais comuns. O *Binning* tem sido uma estratégia importante para a análise de genomas completos, ou quase completos, em dados metagenômicos (21; 22). Essa abordagem permite analisar o genoma de microrganismos não-cultiváveis (23). Atualmente, o *binning* é dividido em três abordagens principais: a. Mapas Auto-Organizados (do inglês, ESOM (*emerging self-organizing map*)); b. Cobertura diferencial (22); e c. Algoritmos de *binning* automático (2).

Os mapas auto-organizados são construídos com base na frequência de tetranucleotídeos e são aplicados em amostras individuais de metagenomas (24). Porém, ESOMs de dados seriados requerem múltiplos conjuntos de dados para um *binning* acurado. Uma abordagem relacionada ao *binning* de dados seriados de ESOM são as abordagens de *binning* com cobertura diferencial, que usam gráficos de duas dimensões (eixos x e y) para plotar a cobertura diferencial de reads usadas na montagem de contigs para separar as populações microbianas individuais (2).

No entanto, ambas as abordagens citadas acima têm como resultado final um gráfico, onde os genomas, ou bins individuais, são escolhidos manualmente (2). Recentemente, Wu e colaboradores (2) desenvolveram a ferramenta computacional MaxBin (2) para automatizar a identificação e validação de bins. O MaxBin automatiza a busca por bins baseado no algoritmo de Maximização de expectativas (do inglês, *expectation-maximization algorithm*). Inicialmente são calculadas as frequências de tetranucleotídeos e o nível de cobertura diferencial de cada *contig*, e essas informações são combina-

das para a identificação do bins. Os bins são validados com base na presença de genes marcadores de cópia-única (Figura 2.2). (2).

Uma das principais limitações dessas técnicas é que as aplicações são restritas a ambientes com baixa diversidade (e.g., consórcios microbianos). No caso da compostagem da FPZSP, um estudo anterior indicou que esse ambiente apresenta uma alta diversidade, tanto taxonômica, quanto funcional (7), sendo de difícil aplicação. Essa classe de abordagem computacional ligada a reconstrução de genomas microbianos foi aplicada nos dados metagenômicos do consórcio microbiano da compostagem.

Figura 2.2: **Etapas de processamento computacional do software MaxBin.** O processamento é dividido em duas etapas principais: I. geração de informações de entrada para o software. O nível de cobertura de sequenciamento de cada contigs/ou scaffold, e o perfil de frequências de tetranucleotídeos é calculado pelo usuário. A identificação de genes marcadores de cópia-única é automatizada pelo software MaxBin. A segunda etapa consiste em identificar as populações microbianas (bin) e validar cada população pela presença de genes marcadores de cópia-única. O produto final são populações que representam genomas individuais presentes no metagenoma. Imagem modificada de Wu et al. (2)

2.3 Ilhas Metagenômicas

Outro conceito importante em nosso projeto é o de *Ilhas Metagenômicas*. Comparações entre sequências de genomas extraídos de metagenomas contra genomas de isolados de referência têm demonstrado regiões com baixa representatividade e regiões não mapeadas, e portanto preditas como regiões únicas das espécies de referência¹. Em geral essas ilhas apresentam um conteúdo GC atípico, baixa densidade gênica, alto número de pseudogenes e pequenas proteínas hipotéticas (25), sendo provável que sua presença seja devida a transferência lateral.

No contexto deste trabalho, a identificação das IMs foi aplicada como um indicativo de diferenciação genômica entre diferentes cepas e espécies. Mesmo sendo lacunas, as IMs podem contribuir para a análise de genomas reconstruídos em metagenomas.

2.4 Glicosil hidrolases

O banco de dados público CAZy (*Carbohydrate Active enZymes* database - CAZy (26); <http://www.cazy.org>) subdivide as enzimas relacionadas ao processamento de carboidratos e glicoconjugados em quatro subgrupos: esterases (CE), polissacarídeo liases (PL), glicosil hidrolases (GH) e glicosil transferase (GT).

Glicosil hidrolases (GHs) são enzimas que catalisam a hidrólise de ligações glicosídicas entre dois ou mais carboidratos ou entre um carboidrato e uma outra biomolécula, produzidas por organismos dos domínios Archaea, Eucarya e Bacteria. As GHs são divididas em 133 famílias protéicas (CAZy Database - <http://www.cazy.org/Glycoside-Hydrolases.html>) – Data de Acesso: 01/04/2015). Cada família contém enzimas que estão relacionadas por similaridade e estrutura tridimensional.

Devido à sua importância em aplicações biotecnológicas e biomédicas, as GHs constituem até agora o grupo melhor caracterizado bioquimicamente no banco de dados

¹Notar que o termo *ilha* neste contexto é enganador, pois se tratam de regiões *ausentes* dos genomas recuperados dos metagenomas. Um termo melhor seria *lacunas metagenômicas*.

CAZy (27). Exemplos de famílias de glicosil hidrolases em processos bioquímicos são: GH1, GH3, GH5, GH7 e GH16 (e.g., degradação de celulose), GH10, GH11 e GH43, e GH2 e GH38 (e.g., polímeros de xilose e manose), assim como GH13, GH31 e GH57 (e.g., amilases) e GH18 e GH85 (e.g., quitinases) (28). As famílias GH5, GH7 e GH9 são as mais diversas, e apresentam um grande interesse e potencial para possíveis aplicações industriais, devido a sua atividade em despolimerizar componentes das paredes celulares de plantas (29). A família GH5, historicamente conhecida como "família de celulasas A", foi a primeira família de celulasas descritas (30).

Associado ao sistema de classificação do banco de dados CAZy, Pope e colaboradores (31) relacionaram quatro grupos funcionais de enzimas ligadas aos processos de desconstrução da parede vegetal de plantas: celulasas (GH5, GH6, GH7, GH9, GH44, GH45, GH48), endohemicelulasas (GH8, GH10, GH11, GH12, GH26, GH28, GH53), enzimas desramificantes (GH51, GH54, GH62, GH67, GH78) e enzimas degradadoras de oligossacarídeos (GH1, GH2, GH3, GH29, GH35, GH38, GH39, GH42, GH43, GH52).

2.5 Reconstrução de genomas

Com o advento do sequenciamento metagenômico foi possível descobrir novos genes (32; 33) e reconstruir genomas quase completos ou completos com qualidade suficiente para inferências filogenéticas e reconstruções metabólicas acuradas (22; 21). Esses estudos têm revolucionado as descobertas nos campos da Ecologia e Evolução Microbiana. No entanto, esses estudos tiveram como foco principal as análises via sequenciamento de genes conservados, tais como o gene que codifica a subunidade 16S rRNA (34; 35).

No ano de 2004, Tyson e colaboradores (36) publicaram o primeiro estudo sobre reconstrução de genomas a partir de dados metagenômicos. Esses autores utilizaram um método baseado em separação de populações microbianas via conteúdo GC, conseguindo diferenciar dois genomas completos e mais três genomas quase completos de

comunidades microbianas formadas em resíduos de mineração (do inglês, *acid mine drainage*). A reconstrução desses genomas foi viável porque este ambiente apresenta uma baixa diversidade e as populações microbianas dos dois genomas reconstruídos são as mais abundantes (*Leptospirillum* group II e *Ferroplasma* type II). Os autores ainda discutem que a aplicação dessa abordagem em outros ambientes será limitada caso o microbioma apresente uma alta riqueza de espécies e uma heterogeneidade na abundância dos membros da comunidade microbiana, bem como pela presença de extensivos rearranjos genômicos (36; 23).

Com o desenvolvimento de novas ferramentas computacionais de montagem de genomas que são específicas para dados metagenômicos (19; 37; 18; 38) e de atribuição operacional taxonômica baseada em métodos composicionais de fragmentos de sequências em organismos específicos (*binning*) foi possível reconstruir o genoma de microrganismos de populações raras (abundância menor que 1%) em ambientes de baixa diversidade (21; 22). Wrighton e colaboradores (21) reconstruíram o genoma de 49 membros de filos candidatos (e.g, BD1-5, OP11, e OD1) que são amplamente encontrados em estudos de diversidade microbiana via 16S rRNA e analisaram estratégias metabólicas até então desconhecidas no domínio Bacteria (21). Martín e colaboradores (39) também usaram uma estratégia computacional similar a descrita anteriormente para a reconstrução do genoma da espécie bactéria Candidatus *Accumulibacter phosphatis*.

Em 2011, Morowitz et al. (40) a partir de dados de sequenciamento *shotgun* usando a tecnologia de pirosequenciamento 454 reconstruíram o genoma de duas espécies dominantes (*Serratia* e *Citrobacter*) em amostras do microbioma intestinal. Este estudo revelou características genotípicas em nível de cepas que diferenciam duas cepas de *Citrobacter* e que pode explicar as flutuações nos níveis de abundância durante a sucessão microbiana. Esse estudo foi limitado em caracterizar o genoma de espécies abundantes, não sendo possível reconstruir o genoma de espécies com baixa abundância, ainda que, esses organismos apresentem uma maior contribuição em termos de funções gerais den-

tro da comunidade microbiana (40). Além de genomas bacterianos, também foi possível reconstruir o genoma de Archaea dominantes em microbiomas de oceanos (41).

Com o advento da tecnologia de sequenciamento Illumina, novas oportunidades para a compreensão do estudo de comunidades microbianas e suas características ecológicas foram possíveis. Dois estudos publicados em 2013 demonstraram o potencial para a reconstrução de genomas microbianos de populações raras em ambientes de baixa diversidade, Albertsen et al. (22) desenvolveu um *pipeline* computacional para a reconstrução de genomas a partir do uso de características composicionais e cobertura diferencial em amostras similares de um mesmo microbioma. Com essa abordagem foi possível reconstruir 31 genomas bacterianos presentes em biorreatores de tratamento de esgoto doméstico, incluindo o genoma do filo candidato TM7 (22). Paralelamente, no mesmo ano, Shaton e colaboradores (42) desenvolveram uma estratégia computacional similar de Albertsen et al., (22) usando amostras do microbioma intestinal coletadas em diferentes estágios do desenvolvimento de recém-nascidos prematuros. Neste estudo foi possível reconstruir o genoma de populações bacterianos que apresentam um abundância de 0.05%.

Em 2015, Brown et al. (43) reconstruíram aproximadamente 800 genomas, que representam pelo menos 35 novos filos microbianos, a partir de amostras de metagenomas. Até a presente data, este foi o estudo com o maior número de genomas reconstruídos.

Capítulo 3

Material e Métodos

Neste projeto foram utilizados dois conjuntos de dados. Um primeiro conjunto de dados faz parte de uma série temporal que foi coletada durante algumas etapas do processamento da compostagem (ZC4) (seção 3.2). E o segundo conjunto de dados faz parte do metagenoma de um consórcio microbiano de bactérias termofílicas e celulolíticas (ZCTH02) enriquecido com amostras de compostagem do PZSP (seção 3.3).

3.1 *Pipelines* para a reconstrução de genomas a partir de dados metagenômicos

Um dos objetivos específicos deste projeto foi desenvolver *pipelines* para automatizar as rotinas computacionais de reconstrução de genomas microbianos e análises de diversidade e abundância de GHs. Para alcançar este objetivo foram escritos *scripts* nas linguagens de programação Perl, Python, R e bash, e usou softwares de terceiros que já estavam disponíveis na literatura (Figura 3.1). Os *pipelines* são divididos em três etapas principais: 1. Recrutamento ou Binning; 2. Anotação; e 3. Caracterização filogenômica/filogenética. O recrutamento de genomas (Etapa 1a) é aplicado em ambientes de alta diversidade e com o uso de genomas de referência. Neste caso, esse método

foi aplicado no conjunto de dados ZC4 (seção 3.2.2). Enquanto que o método de *binning* (Etapa 1b) é aplicado em ambientes com uma microdiversidade. Por exemplo, no consórcio microbiano da compostagem (ZCTH02) (seção 3.3) foi aplicado a técnica de *binning*, que é o agrupamento de sequências com base em suas características composicionais (e.g., conteúdo GC, frequência de tetranucleotídeos, etc). A etapa de anotação (Etapa 2) é aplicada em genomas individuais oriundos de ambas as etapas anteriores de reconstrução de genomas, e a caracterização filogenética é aplicada tanto em dados provenientes da etapa de recrutamento, quanto na etapa de *Binning*. Enquanto que a caracterização filogenômica foi aplicada apenas nos dados do consórcio microbiano.

3.2 Dados do metagenoma da compostagem (ZC4)

Desenho experimental e sequenciamento

O procedimento de coleta das amostras da compostagem foi realizado a partir dos mesmos métodos descritos por Martins e colaboradores (7). Cinco pontos da composteira foram combinados para compor uma única amostra que foi congelada -80°C. Esse mesmo material foi utilizado para a extração de DNA da análise metagenômica. A extração de DNA foi realizada com base no protocolo e kit MoBio PowerMax Soil DNA (MoBio Laboratories). As bibliotecas de DNA foram preparadas para o sequenciamento na plataforma MiSeq (Illumina) usando o Nextera DNA SamplePrep kit (Illumina).

O desenho experimental, coletas, extração de DNA e preparação de bibliotecas para o sequenciamento na plataforma Illumina MiSeq foi realizado pela aluna de doutorado Luciana Antunes Principal, integrante do projeto temático.

Informações adicionais sobre os dados metagenômicos da compostagem, quanto ao número de sequências e datas de coletas estão disponíveis na **Tabela 3.1**.

Figura 3.1: *Overview* das rotinas computacionais utilizadas neste projeto. As análises são divididas inicialmente em recrutamento e montagem *De novo* de sequências, seguido de identificação de Ilhas Metagenômicas, anotação, identificação de genes de interesse, diversidade das principais famílias de GHs e caracterização filogenômica.

Tabela 3.1: Número de sequências de cada amostra do metagenoma da compostagem (ZC4) e do consórcio microbiano (ZCTH02).

Amostras	Data de coleta	Total (dados brutos)	Total (dados limpos)
ZC4_01	06/08/2013	4.106.932	3.992.763
ZC4_03	08/08/2013	4.703.839	4.506.766
ZC4_07	12/08/2013	4.571.336	4.303.900
ZC4_15	20/08/2013	7.231.432	6.744.827
ZC4_30	04/09/2013	3.651.669	3.032.359
ZC4_64	08/10/2013	5.055.450	4.943.709
ZC4_67	11/10/2013	4.203.006	4.014.859
ZC4_78	22/10/2013	5.285.343	4.744.857
ZC4_99	12/11/2013	5.899.518	4.673.060
ZCTH02	30 ^o dia	3.449.215	3.110.717

3.2.1 Pré-processamento de sequências brutas

O pré-processamento das sequências geradas no projeto temático foi realizado por Gianluca Machado da Silva (também aluno de mestrado do orientador). Para a remoção de sequências de baixa qualidade foi utilizado o software Sickle (44). Foram removidas do conjunto de dados sequências menores que 50 bp e com Phred score menor que 20.

3.2.2 Recrutamento e montagem dos genomas de bactérias abundantes

A identificação e quantificação das espécies mais abundantes presentes no metagenoma da compostagem foi realizada no software MyTaxa (45). As investigações sobre a estrutura de comunidades microbianas, incluindo o estudo da diversidade taxonômica, faz parte da tese de doutorado da aluna Luciana Antunes Principal.

Durante a etapa de recrutamento de sequências candidatas de cada genoma bacteriano foi realizado um alinhamento das sequências do metagenoma da compostagem contra os genomas de referências das espécies abundantes. Foram utilizados sete genomas de referência já depositados no NCBI, os quais estão completos e anotados (**Tabela 3.2**).

Tabela 3.2: Características gerais dos genomas usados como referência nas etapas de recrutamento de sequências em ZC4.

Espécie de Referência	Ordem	Número de Acesso (Referência)	Tamanho do genoma (bp)	Número de CDS	Plasmídeo	Tamanho do plasmídeo (bp)
<i>Pseudoxanthomonas suwonensis</i> 11-1	Xanthomonadales	NC_014924.1	3.419.049	3.045	Ausente	Ausente
<i>Thermoaerobacter marianensis</i> DSM	Clostridiales	NC_014831.1	2.844.696	2.308	Ausente	Ausente
<i>Thermobispora bispora</i> DSM	Rhodothermaceae	NC_014165.1	4.189.976	3.476	Ausente	Ausente
<i>Sphaerobacter thermophilus</i> DSM	Sphaerobacterales	NC_013523.1	2.741.033	3.439	Ausente	Ausente
<i>Rhodothermus marinus</i> DSM	Clostridiales	NC_013501.1	3.261.604	2.841	Presente	125.133
<i>Thermobifida fusca</i> YX	Actinomycetales	NC_007333.1	3.642.249	3.024	Ausente	Ausente
<i>Symbiobacterium thermophilum</i> IAM	Actinomycetales	NC_006177.1	3.566.135	3.180	Ausente	Ausente

A etapa de recrutamento é realizada a partir de um alinhamento local no software FR-HIT (46). O software FR-HIT utiliza a construção de tabelas hash de k-mers para a compactação do banco de dados de referências. Isso reduz a complexidade computacional do processo, acelerando o alinhamento de milhares de sequências.

Para reduzir o problema de sequências conservadas presentes nos genomas microbianos e possível perda de informação em termos de descartes de sequências que alinham em mais de um genoma, sequências que alinham em mais de um genoma foram selecionadas com base no *best-hit* e valor de e-value. Esta estratégia de recrutamento de sequências foi similar a estratégia utilizada por Kay e colaboradores (47).

O critério geral de filtro foi de 85% de identidade e *e-value* menor que 0.00005 e o processo de montagem foi realizado no software Newbler (48).

3.2.3 Identificação de Ilhas Metagenômicas

As Ilhas Metagenômicas (IMs) foram identificadas no visualizador de alinhamentos de metagenomas MGA-Viewer (49). Os genes presentes nessas regiões foram identificados via anotação disponível no arquivo Genbank do NCBI.

3.2.4 Anotação

O processo de anotação (Etapa 2) foi realizado no software PROKKA (50). O PROKKA utiliza uma série de softwares para inferir características específicas de genomas microbianos. A identificação de regiões codificantes (CDS) foi realizada no software Prodigal (51) e os genes ribossomais foram identificados no software RNAmmer (52).

As glicosil hidrolases foram identificadas a partir de Modelos Ocultos de Markov (hmm) disponíveis no banco de dados dbCan (53). O banco de dados público CAZy disponibiliza informações sobre a nomenclatura e classificação de GHs de acordo com suas sequências de aminoácidos.

Para as análise exploratória de dados foi utilizado a plataforma R - R Development

Core Team, 2011.

3.2.5 Caracterização filogenética

A caracterização filogenética foi realizada a partir da comparação das sequências de genes marcadores de cópia-única presentes nos genomas microbianos reconstruídos usando a ferramenta BLASTn (54).

3.3 Dados do metagenoma do consórcio microbiano da compostagem (ZCTH02)

3.3.1 Desenho experimental e sequenciamento

O consórcio microbiano utilizado neste projeto foi construído pelo Prof. Dr. Ronaldo Quaggio (Instituto de Química da USP). Inicialmente uma amostra de compostagem foi enriquecida em um meio de cultura termofílico em fonte de carbono 1% de carboximetilcelulose em uma temperatura de 65°C. A extração de DNA e sequenciamento do DNA metagenômico foi realizada da mesma forma que em ZC4. Informações gerais da amostra estão disponíveis na tabela 3.1.

A composteira escolhida foi a que estava com mais alta temperatura no dia de realização da coleta deste material (composteira de número 30 com uma temperatura de aproximadamente 60-70°C). O tempo de processamento foi no dia 30. Ela estava no trigésimo dia de processamento da compostagem.

3.3.2 Montagem *De novo* de genomas microbianos

O processo de montagem das sequências do consórcio microbiano (Etapa 1b) foi realizado no software Spades (38) com parâmetros específicos para sequências *paired-end*. A reconstrução de genomas individuais com base em métodos de composição (*binning*) foi realizada no software MaxBin (2).

3.3.3 Anotação

O processo de anotação para identificar CDSs e glicosil hidrolases foi o mesmo utilizado nos dados de ZC4, conforme descrito na seção anterior.

Uma das vantagens da aplicação de métodos de binning é a recuperação de genomas de microrganismos ainda não descritos na literatura. Com isso, é pertinente adicionar uma etapa de identificação de categorias funcionais. A identificação dessas categorias foram identificadas a partir do uso da ferramenta Psiblast (54), comparando as sequências do genoma contra o banco de dados do COG (*Clusters of Orthologous Groups*).

3.3.4 Caracterização filogenética e filogenômica

As sequências do 16S rRNA, quando presentes, foram comparadas com o banco de dados RDP (55). Buscas por genes marcadores de cópia-única complementaram essa etapa de identificação de quais bactérias estão mais próximas filogeneticamente das presentes na compostagem.

As análises filogenéticas foram realizadas a partir de um alinhamento múltiplo no software MUSCLE (56) e a inferência filogenética realizada no software FastTree2 (57)

Para complementar a caracterização filogenética foi realizada uma análise filogenômica dos genomas microbianos reconstruídos em ZCTH02 (Etapa 3) no software PhyloPhlAn (58).

Capítulo 4

Resultados e Discussão

4.1 Genomas recrutados em dados do metagenoma da compostagem (ZC4)

Nas amostras do conjunto de dados ZC4 foram mapeadas 2.570.483 sequências candidatas em sete genomas de referência com um critério de 85% de similaridade (Tabela 4.1). Sendo possível obter genomas com até 80% do genoma de referência (Tabela 4.2).

4.1.1 Variações pontuais nos genomas recrutados em ZC4

As bactérias identificadas são conhecidas degradadoras de biomassa vegetal em compostagem (e.g., *Thermobispora bispora* e *Thermobifida fusca*), ao passo que para outros que identificamos esta seria uma observação nova (e.g., *T. marianensis*). Neste último caso estamos provavelmente diante de cepas novas em relação às conhecidas, e possivelmente espécies novas (ou seja, uma classificação de espécie incorreta feita pelo software de identificação taxonômica).

Tabela 4.1: Número de sequências candidatas recrutadas com um critério mínimo de 85% de identidade.

Espécie	ZC4_01	ZC4_03	ZC4_07	ZC4_15	ZC4_30	ZC4_64	ZC4_67	ZC4_78	ZC4_99
<i>P. suwonensis</i>	102.345	104.246	39.419	36.267	11.487	27.219	25.901	16.166	13.329
<i>T. marianensis</i>	8.442	14.242	34.729	52.588	13.382	17.811	35.430	20.009	7.590
<i>T. bispora</i>	120.529	247.572	101.361	93.197	65.921	39.674	32.658	58.210	36.842
<i>S. thermophilus</i>	8.128	21.735	15.807	25.449	13.690	17.485	18.221	38.219	24.228
<i>R. marinus</i>	4.665	9.123	14.583	190.710	90.357	22.715	29.153	108.264	19.149
<i>T. fusca</i>	10.667	15.411	7.042	9.656	40.799	44.162	16.285	32.277	24.179
<i>S. thermophilum</i>	40.377	83.597	64.758	36.732	7.820	41.468	29.166	11.108	6.732

Tabela 4.2: Principais características dos genomas de espécies abundantes reconstruídos a partir de análises de recrutamento de sequências em amostras do metagenoma da compostagem (ZC4).

Espécie de Referência	Número de contigs candidatos	Mapeamento no genoma de referência (%)	Número de genes candidatos
<i>P. suwonensis</i>	2.581	67,4	3.297
<i>T. marianensis</i>	2.576	80,6	3.306
<i>T. bispora</i>	623	99,9	3.936
<i>S. thermophilus</i>	394	95,9	2.580
<i>R. marinus</i>	905	91,8	2.721
<i>T. fusca</i>	490	97,9	3.255
<i>S. thermophilum</i>	3.221	96,9	4.781

A identificação das espécies foi confirmada usando sequências de genes marcadores de cópia-única. A média de identidade obtida para as espécies *Thermobispora bispora*, *Sphaerobacter thermophilus* e *Thermobifida fusca* foi de 98% (Tabela 4.3). Enquanto que para *Rhodothermus marinus* a média foi de 96.1%, mas as análises baseadas em um segundo gene marcador mostraram que as sequências de *R. marinus* da compostagem do FPZSP tem um nível de identidade de 99.3% com as sequências de uma cepa de *R. marinus* disponível no IMG (Integrated Microbial Genomes). Para as espécies *Symbiobacterium thermophilum* (95.2%), *Pseudoxanthomonas suwonensis* (93%) e *Thermaerobacter marianensis* (91.5%) foi observado uma média baixa no nível de identidade, sugerindo que os microrganismos encontrados na compostagem podem ser espécies diferentes de um mesmo gênero. De fato, no caso de *Thermaerobacter*, nós identificamos, via análises do gene 16S rRNA, que essa espécie está mais relacionada com a espécie *Thermaerobacter* sp SS (Número de acesso GenBank AB444428).

A seguir mostraremos observações sobre cada um dos genomas das espécies identificadas.

***Pseudoxanthomonas suwonensis*.** Comparações realizadas com as sequências de genes marcadores de cópia-única recrutadas e reconstruídas a partir de dados do metagenoma da compostagem indicaram uma variação de 91 até 94% no nível de identidade nas sequências de nucleotídeos com base naquelas presentes no genoma de referência. Quanto a capacidade de degradação de açúcares complexos em açúcares menores foi identificado sete genes (Tabela 4.4), sendo possível identificar variações pontuais nessas sequências. Em termos de repertório gênico relacionado a degradação de biomassa vegetal, a presença de uma Mannan endo-1,4-beta-mannosidase indica uma potencial capacidade e associação desta espécie com funções de degradação de biomassa vegetal durante o processamento da compostagem. Bem como a presença de proteínas das famílias GH43, GH3 e GH11 que são potenciais endohemicelulases termoestáveis. A presença de enzimas acessórias para a degradação de biomassa vegetal, tal como uma representante da família CE12, reforça ainda mais o papel funcional desta bactéria na compostagem.

***Rhodothermus marinus*.** Com base em análises comparativas de genes marcadores de cópia-única foi possível observar uma variação de 94 a 100% de identidade nas sequências de nucleotídeos. Foram identificados onze genes que codificam enzimas potenciais para a degradação de biomassa vegetal (Tabela 4.5). Similar as análises comparativas realizadas com *P. suwonensis*, no genoma de *R. marinus* foram identificadas inúmeras variações pontuais nas sequências de nucleotídeos, conforme observado pelas variações nos níveis de identidade nos alinhamentos. Um exemplo de variação está no gene que codifica uma Intracellular exo-alpha-(1->5)-L-arabinofuranosidase 1 (94% de identidade em nível de nucleotídeos). Esta é uma enzima-chave para a degradação de componentes de parede vegetal de plantas (59). Neste genoma também estão presentes

genes que codificam enzimas das famílias GH3. Enzimas desta família são responsáveis pela conversão de celobiose em glicose (60).

Thermobispora bispora. Bactéria aeróbica, gram-positiva e termofílica (8). Apresenta um único genoma depositado no NCBI. Essa cepa (R51) foi isolada de amostras de estrume maduro (61). As sequências de genes marcadores de cópia-única, com base em análises comparativas com o genoma de referência, apresentaram variações de 92 até 100%. Com exceção do gene que codifica a DNA-directed RNA polymerase subunit alpha (92%), todos os outros genes apresentaram um alto nível de identidade. O baixo nível de identidade da DNA-directed RNA polymerase subunit alpha pode estar relacionado com o baixo nível de cobertura do alinhamento (80%). Quanto ao repertório gênico associado a degradação de biomassa vegetal, não foi possível observar alterações pontuais nas sequências, no qual todas as sequências apresentaram quase 100% de identidade ao genoma de referência (Tabela 4.6).

Thermaerobacter sp. Os genes marcadores de cópia-única apresentaram uma variação de 90 até 93% nas sequências de nucleotídeos. Devido a fragmentação das sequências e com base nos critérios mínimos de identificação recomendados pelo dbCan database não foi possível identificar nenhuma enzima associada a degradação de biomassa vegetal neste genoma. No entanto, este genoma deve possuir enzimas degradadoras, visto que o registro do GenBank para *Thermaerobacter sp.* SS (número de acesso AB444428) indica que ela foi isolada de lama residual de um sistema de saneamento municipal no Japão.

Sphaerobacter thermophilus. Bactéria aeróbia obrigatória isolada de um municipal sludge (62). A *Sphaerobacter* identificada nas amostras do metagenoma do zoológico apresentou um alto nível de identidade (86-100%) com a cepa de referência depositada no NCBI. Os genes associados a degradação de biomassa vegetal também apresentaram

poucas variações (Tabela 4.7).

Tabela 4.3: Análise comparativa de genes marcadores de cópia-única presentes nos genomas microbianos das sete espécies mais abundantes da microbiota da compostagem (ZC4).

Marker gene	P. <i>suwonensis</i>	T. <i>marianensis</i>	T. <i>bispora</i>	S. <i>thermophilus</i>	R. <i>marinus</i>	T. <i>fusca</i>	S. <i>thermophilum</i>
50S							
ribosomal protein L6	92	93	100	99	99	100	90
50S							
ribosomal protein L21	91	91	100	99	100	99	99
DNA-directed							
RNA polymerase subunit beta	91	90	100	99	98	100	100
Ribosomal							
RNA small subunit methyltransferase E	98	Not detected	99	99	95	100	Missing in the reference
DNA							
primase	92	93	100	99	94	100	86
DNA-directed							
RNA polymerase subunit alpha	93	92	92	100	98	100	100
DNA							
translocase SpoIIIE							
[cell	94	90	99	99	95	100	96
division protein FtsK/SpoIIIE]							

Tabela 4.4: Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de *Pseudoxanthomas suwonensis* em amostras de compostagem

Sequência	Cobertura (%)	Identidade (%)	E-value	Família do CAZy	Atividades conhecidas
PROKKA_00450	100	92.71	0.0	GH43	xylan 1,4-beta-xylosidase
PROKKA_00615	100	100.00	0.0	CE12	esterase
PROKKA_00791	97	92.17	0.0	CE6	hypothetical protein
PROKKA_01179	99	90.97	0.0	GH5	Mannan endo-1,4-beta-mannosidase
PROKKA_01199	99	92.45	0.0	GH3	glycoside hydrolase
PROKKA_01981	100	92.73	0.0	GH11	glycoside hydrolase family protein
PROKKA_02178	100	90.71	0.0	GH43	glycoside hydrolase family protein

Tabela 4.5: Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de *Rhodothermus marinus* em amostras de compostagem

Sequence	Coverage (%)	Identity (%)	E-value	CAZy family	Atividades conhecidas
PROKKA_00191	100	97.48	0.0	GH26	Mannan endo-1,4-beta-mannosidase
PROKKA_00782	99	93.08	0.0	CE3	GDSL-like Lipase/Acylhydrolase
PROKKA_00857	100	96.88	0.0	CE1	enterobactin/ferric enterobactin esterase
PROKKA_00907	100	97.02	0.0	GH3	Beta-glucosidase
PROKKA_01043	100	98.13	0.0	GH10	Endo-1,4-beta-xylanase A precursor
PROKKA_01168	100	94.49	0.0	GH51	Intracellular exo-alpha-L-arabinofuranosidase 2
PROKKA_01405	100	99.56	0.0	GH26	Mannan endo-1,4-beta-mannosidase precursor
PROKKA_01501	100	99.78	0.0	CE1	2-hydroxymuconate semialdehyde hydrolase
PROKKA_01786	99	94.27	0.0	CE1	Carboxymethylenebutenolidase
PROKKA_01816	100	96.32	0.0	GH43	Xylan 1,3-beta-xylosidase
PROKKA_01894	100	99.95	0.0	GH67	Extracellular xylan exo-alpha-(1->2)-glucuronosidase precursor

Tabela 4.6: Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de *Thermobispora bispora* em amostras de compostagem.

Sequência	Cobertura (%)	Identidade (%)	E-value	Família do Cazy	Atividades Conhecidas
PROKKA_00047	99.95	100	0.0	GH6	Exoglucanase_A_precursor
PROKKA_00069	99.93	100	0.0	GH10	Endo-1,4-beta-xylanase_A_precursor
PROKKA_00119	100.00	100	0.0	GH5	Mannan_endo-1,4-beta-mannosidase_precursor
PROKKA_00405	99.92	100	0.0	CE1	Exoglucanase/xylanase_precursor
PROKKA_00458	99.87	100	0.0	CE1	Alpha/beta_hydrolase_family_protein
PROKKA_00460	99.92	100	0.0	GH1	Beta-glucosidase_A
PROKKA_00613	99.96	100	0.0	GH3	Thermostable_beta-glucosidase_B
PROKKA_00619	99.93	100	0.0	GH3	Thermostable_beta-glucosidase_B
PROKKA_00935	100.00	100	0.0	CE1	3-oxoadipate_enol-lactonase_2
PROKKA_01040	100.00	100	0.0	GH48	Exoglucanase_B_precursor
PROKKA_01041	100.00	100	0.0	GH10	Endo-1,4-beta-xylanase_A_precursor
PROKKA_01096	100.00	100	0.0	GH5	Endoglucanase_C307_precursor
PROKKA_01097	100.00	100	0.0	GH43	arabinofuranosidase
PROKKA_01110	99.93	100	0.0	GH1	Thermostable_beta-glucosidase_B
PROKKA_01302	99.65	100	0.0	GH1	Beta-glucosidase_A
PROKKA_01304	99.92	100	0.0	GH1	Beta-glucosidase_A
PROKKA_01505	99.88	100	0.0	CE4	Peptidoglycan-N-acetylglucosamine_deacetylase
PROKKA_01541	99.62	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_01602	99.62	100	0.0	CE4	Polysaccharide_deacetylase
PROKKA_01606	99.92	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_01705	100.00	100	0.0	CE1	Exoglucanase/xylanase_precursor
PROKKA_01936	99.79	100	0.0	GH1	Beta-glucosidase_B
PROKKA_02333	99.90	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_02350	99.93	100	0.0	CE7	Alpha/beta_hydrolase
PROKKA_02410	99.88	100	0.0	CE4	Polysaccharide_deacetylase
PROKKA_02758	99.87	100	0.0	CE4	Peptidoglycan-N-acetylmuramic_acid_deacetylase_PdaC
PROKKA_03162	100.00	100	0.0	GH62	arabinofuranosidase_precursor
PROKKA_03306	100.00	100	0.0	CE1	Trioxoandrosta
PROKKA_03426	100.00	100	0.0	GH11	Endo-1,4-beta-xylanase_B_precursor

Tabela 4.7: Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de *Sphaerobacter thermophilus* em amostras de compostagem

Sequence	Coverage (%)	Identity (%)	E-value	CAZy family	Atividades conhecidas
PROKKA_00143	100	99.50	0	GH5	Cellulase
PROKKA_00552	100	99.50	0	GH39	Poly-beta-1,6-N-acetyl-D-glucosamine
PROKKA_01255	100	99.50	0	GH5	hypothetical_protein
PROKKA_01336	100	99.78	0	GH57	Alpha amylase
PROKKA_01731	100	99	0	GH1	1,4-beta-D-glucan_glucohydrolase
PROKKA_02139	100	98	0	GH39	Cellulase

***Symbiobacterium thermophilum*.** Bactéria termofílica gram-negativa. Apresenta um crescimento dependente de co-culturas com associações a bactérias do gênero *Bacillus* (63). Um único genoma de *S. thermophilum* está depositado no NCBI (64). As análises comparativas de genes marcadores de cópia-única apresentam variações de 86 até 100% com base nas sequências de nucleotídeos. Foi possível identificar quatro carboxylesterases relacionadas a deconstrução de biomassa vegetal (Tabela 4.8).

***Thermobifida fusca*.** Tal como em *T. bispora*, não foi possível identificar variações nas sequências de nucleotídeos dos genes que codificam GHs ligadas a degradação de biomassa vegetal (Tabela 4.9).

Tabela 4.8: Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de *Symbiobacterium thermophilum* em amostras de compostagem

Sequence	Coverage (%)	Identity (%)	E-value	CAZy family	Atividades conhecidas
PROKKA_00099	100	100	0	CE4	Bifunctional_xylanase
PROKKA_01646	100	99	0	CE4	Hypothetical protein
PROKKA_02021	100	100	0	CE1	Alpha/beta_hydrolase_family_protein
PROKKA_02397	100	99	0	CE1	Carboxylesterase
PROKKA_04771	100	94	0	CE4	Bifunctional_xylanase

Tabela 4.9: Potenciais enzimas relacionadas a degradação de carboidratos identificadas no genoma de *Thermobifida fusca* em amostras de compostagem.

Sequência	Cobertura (%)	Identidade (%)	E-value	Família do CAZy	Atividades conhecidas
PROKKA_00220	99.79	100	0.0	GH57	hypothetical_protein
PROKKA_00322	100.00	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_00323	99.74	100	0.0	CE12	GDSL-like_Lipase/Acylhydrolase
PROKKA_00417	99.96	100	0.0	GH9	Endoglucanase_E-4_precursor
PROKKA_00630	99.90	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_00688	99.86	100	0.0	GH1	Beta-glucosidase_A
PROKKA_00796	99.55	100	0.0	CE3	Endo-1,4-beta-xylanase_A_precursor
PROKKA_00809	99.92	100	0.0	GH2	Exo-beta-D-glucosaminidase_precursor
PROKKA_00824	99.93	100	0.0	GH5	Endoglucanase_E-5_precursor
PROKKA_00825	99.93	100	0.0	GH5	Mannan_endo-1,4-beta-mannosidase_precursor
PROKKA_00831	99.80	100	0.0	GH11	Endo-1,4-beta-xylanase_B_precursor
PROKKA_01122	99.25	100	0.0	CE4	hypothetical_protein
PROKKA_01232	99.91	100	0.0	CE1	Homoserine_O-acetyltransferase
PROKKA_01269	100.00	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_01428	99.77	100	0.0	CE1	trioxoandrosta
PROKKA_01632	100.00	100	0.0	GH4	putative_6-phospho-beta-glucosidase
PROKKA_01915	99.82	100	0.0	GH43	Xylan_1,3-beta-xylosidase
PROKKA_01920	99.73	100	0.0	CE4	Peptidoglycan-N-acetylglucosamine_deacetylase
PROKKA_01972	99.84	100	0.0	GH6	Exoglucanase_A_precursor
PROKKA_02327	99.95	100	0.0	GH9	Cellulase_1_precursor
PROKKA_02483	99.83	100	0.0	GH48	Exoglucanase_B_precursor
PROKKA_02503	99.52	100	0.0	CE1	Esterase_PHB_depolymerase
PROKKA_02506	100.00	100	0.0	CE1	Tropinesterase
PROKKA_02614	100.00	100	0.0	CE3	GDSL-like_Lipase/Acylhydrolase
PROKKA_03082	100.00	100	0.0	GH1	Bifunctional_beta-D-glucosidase/beta-D-fucosidase
PROKKA_03122	99.91	100	0.0	CE2	Lipase_1_precursor

4.1.2 Ilhas Metagenômicas

Nas análises de recrutamento foram identificadas 33 IMs (Tabela 4.10). Essas regiões apresentam várias características típicas de regiões genômicas de alta instabilidade, conforme já registrado na literatura para outras espécies (65; 66), bem como a presença de pseudogenes, proteínas hipotéticas, proteínas de membrana e conteúdo GC atípico (25).

Dentre as principais características pontuais observadas nas IMs estão a presença de proteínas que fazem a biossíntese de componentes de membrana celular, receptores de membrana (e.g., receptor TonB em *R. marinus*) e proteínas hipotéticas (presentes em todas IMs identificadas nesta análise de recrutamento de sequências).

Um exemplo de identificação de receptores de membranas em IMs foi a presença de receptores dependentes de TonB e cupin 2 em *R. marinus*. Os receptores dependentes de TonB são complexos de sinalização celular da parte externa do envelope celular de bactérias (e.g., *Escherichia coli*) (67), sendo um sítio de ligação de bacteriófagos (68). Essa região sendo ausente na *R. marinus* da compostagem do PZSP é um possível mecanismo de defesa contra a predação de fagos.

Também foram identificadas excisionase e integrases (Figura 4.1) nas IMs. Essas duas enzimas participam dos processos de integração e excisão de bacteriófagos em cromossomos microbianos (69). As integrases também são conhecidas por indicarem uma possível presença de elementos profagos em genomas bacterianos (70; 71; 72).

Quanto a presença de proteínas ligadas a estrutura de membranas, identificamos aminoptidases e as undecaprenil-fosfato. Em análises dos genomas de cepas da espécie *Vibrio cholerae* foi encontrado um genoma de bacteriófago que estava integrado em regiões que codificavam aminopeptidases (73).

Tabela 4.10: Descrição de Ilhas Metagenômicas presentes em espécies abundantes que foram classificadas no metagenoma da compostagem (ZC4).

Espécie	Nº de Ilhas Metagenômicas	Tamanho da maior Ilha Metagenômica (kbp)	Principais características das Ilhas Metagenômicas
<i>P. suuonensis</i>	9	20	Proteínas hipotéticas; Proteínas de membrana (YhhN family protein); Transportador ABC; proteínas de biossíntese de polissacarídeos.
<i>T. marianensis</i>	6	50	Proteínas hipotéticas; Lipoproteína A; proteína de montagem de fimbria.
<i>T. bispora</i>	2	5	Proteínas hipotéticas.
<i>S. thermophilus</i>	2	10	Proteínas hipotéticas; resolvase.
<i>R. marinus</i>	8	12	Proteínas hipotéticas; excisionase; integrase; receptor dependente de TonB; cupin 2.
<i>T. fusca</i>	0		
<i>S. thermophilum</i>	6	15	Proteínas hipotéticas; ligase.

Figura 4.1: Ilha Metagenômica (IM) de *Rhodothermus marinus*. O nome de cada produto proteico estão localizados perto das caixas laranjas.

4.2 Consórcio microbiano termofílico e celulolítico (ZCTH02)

4.2.1 Posicionamento filogenético dos genomas reconstruídos

Foram identificados seis genomas individuais (bins) no consórcio microbiano termofílico e celulolítico da compostagem, todos eles com mais de 90% de uma coleção de genes marcadores de cópia-única (Tabela 4.11).

Identificamos em Bin1 uma sequência de 16S rRNA que tem 98% de identidade com um microrganismo não-cultivável isolado de uma comunidade microbiana de compostagem (número de acesso do NCBI: FN667161) (74) e um nível de identidade de 89% com base nas comparações entre as sequências do gene que codifica a DNA primase da espécie *Thermobacillus composti*. O baixo nível de identidade obtido nas análises comparativas da DNA primase é consistente com o resultado obtido nas análises de 16S rRNA e sugere tratar-se de uma espécie nova do gênero *Thermobacillus*.

A sequência de 16S rRNA presente no genoma Bin2 apresentou um nível de identidade de 99% com o gene do 16S rRNA da espécie *Bacillus thermozeamaize*. A espécie

B. thermozeamaize foi isolada de amostras de instalações termofílicas e hipertermofílicas de de processamento de alimentos (75). Com o resultado de comparações do gene 16S rRNA de Bin2 indicando uma alta similaridade com a espécie *B. thermozeamaize* e uma baixa similaridade com outros genomas (maior nível de identidade foi de 47% em sequências do gene que codifica a DNA primase) é plausível supor que se trata do primeiro registro do genoma da espécie *B. thermozeamaize*.

As sequências dos genes marcadores de cópia-única dos genomas Bin4 e Bin5 apresentaram um alto nível de similaridade (maior ou igual a 99%) com as espécies *Geobacillus thermoglucosidasius* e *Caldibacillus debilis*, respectivamente. Esse resultado é um indicativo de que esses dois genomas são representantes das espécies *G. thermoglucosidasius* e *Caldibacillus debilis*.

Tabela 4.11: Genomas microbianos reconstruídos a partir de um consórcio microbiano de compostagem (ZCTH)

Identificador do genoma	Bin1	Bin2	Bin3	Bin4	Bin5	Bin6
Melhor hit (16S rRNA)	Thermobacillus sp Uncultured bacteria	Bacillus thermoazeamaize	NA	Geobacillus thermoglucosidasius	Caldibacillus debilis	Uncultured compost bacterium
Cobertura/Identidade (16S rRNA) (%)	98/99	99/99	NA	14/100	8/100	95/97
Melhor hit (DNA primase)	<i>Thermobacillus</i> <i>composti</i>	<i>Calditalkalibacillus</i> <i>thermarum</i>	<i>Cohnella</i> <i>laeviribosi</i>	<i>Geobacillus</i> <i>thermoglucosidasius</i>	<i>Caldibacillus</i> <i>debilis</i>	<i>Clostridiales</i> <i>bacterium</i>
Cobertura/Identidade (DNA primase) (%)	100/89	98/47	99/66	100/99	100/100	94/41
Completeness (%)	100	92	98.1	97.2	98.1	95.3
GC (%)	64	54	63	43	52	66
Número de contigs	46	143	378	245	244	374
Tamanho do genoma (bp)	3.371.584	3.439.117	3.487.502	4.385.880	2.863.490	2.906.748
Tamanho do maior contig (bp)	509.962	208.013	120.895	122.932	49.485	62.495
N50	168.514	69.814	27.823	33.740	18.062	12.495
Número de CDS	3.058	3.319	3.288	4.273	2.647	2.742
Número de RNAs	60	67	46	69	56	53

Tal como a sequência que codifica o gene 16S rRNA presente no genoma do Bin1, a sequência de 16S rRNA do Bin6 também apresentou um alto nível de identidade (97%) com uma bactéria não-cultivável. A sequência mais próxima de Bin6 foi isolada de uma comunidade microbiana presente em um digestor aeróbio termofílico e autotermal de tratamento de esgotos (número de acesso do NCBI: FN687454) (76). O microrganismo mais próximo filogeneticamente do Bin6, com base na comparação da sequência da DNA primase é uma "Clostridiales bacterium" (nível de identidade de 66%). As análises filogenômicas reforçaram os resultados obtidos por comparações individuais de genes marcadores de cópia-única e pelo 16S rRNA do genoma Bin6, e de uma forma geral para todos os outros cinco genomas (Figura 4.2). Neste caso, genoma da espécie Bin6 está relacionado em um grupo divergente das classes Bacilli e Clostridia. Esse resultado sugere que este genoma provavelmente é de um organismo pertencente a uma possível nova classe dentro do filo Firmicutes. Este resultado corrobora as recentes análises filogenômicas baseadas em proteínas conservadas no filo Firmicutes (77).

Não foi possível identificar padrões de altas similaridades nas sequências presentes no genoma Bin3, visto que a sequência de DNA primase divergiu muito do microrganismos mais conhecido (66% de identidade com a espécie *Cohnella laeviribosi*). Para esta espécie também não foi possível recuperar uma sequência completa ou quase completa do gene 16S rRNA. As análises baseadas em filogenômica indicam que esta espécie é representante da família Paenibacillaceae, sendo um indicativo de um gênero novo identificado em amostras da compostagem.

4.2.2 Glicosil-hidrolases identificadas no consórcio microbiano

Foram identificadas 60 Glicosil-hidrolases associadas a degradação de biomassa vegetal nos seis genomas reconstruídos em amostras do consórcio microbiano da compostagem (Tabela 4.12). A maior parte dos genes apresentou um alto nível de cobertura, e por isso, foi possível realizar análises comparativas mais detalhadas em termos de alinhamentos

com sequências depositadas em bancos de dados públicos.

A presença desses genes indica um potencial para degradação de biomassa vegetal, bem como a mineração de enzimas termoestáveis de interesse industrial. Por exemplo, no genoma Bin1 as enzimas mais abundantes foram as pertencentes a família GH43. Dentro deste grupo estão presentes várias enzimas com particular interesse na produção de biocombustíveis, tais como as beta-glucosidases, que são responsáveis pela conversão de celobiose em glicose (78).

Bin1 se destaca por ser o mais rico em diversidade e abundância de GHs, tendo o maior potencial para degradação de biomassa vegetal na compostagem. Os demais microrganismos seriam "acessórios" ou complementares (visto que possuem, coletivamente, 6 famílias de GHs não presentes em Bin1). Neste caso, Bin1 foi a espécie dominante no consórcio microbiano. E ela pode depender de um ou maior microrganismos para realizar determinadas funções de degradação de biomassa.

Cinco famílias (GH39, GH26, GH1, GH5, GH52 e GH57) ocorreram unicamente em um desses cinco genomas e não em Bin1. Por exemplo, a família GH57 foi identificada apenas no genoma Bin6. Representantes desta família apresentam funções de alfa-amilases e estão associadas as rotas metabólicas de sacarificação enzimáticas, seguido de fermentação (79). O maior nível de identidade encontrado na sequência de aminoácido desta GH57 deste genoma comparado com o banco de dados do nr foi de 63% (*Candidatus Brocadia sinica* JPN1 - Número de acesso: GAN32319). Enquanto que as famílias GH26 e GH39 ocorreram apenas em Bin3.

Uma das enzimas da família GH9 presente no genoma Bin1 apresentou um nível de identidade de 79% com uma endoglucanase depositada no NCBI (código de acesso: WP_015253529.1). As enzimas da família GH9 apresentam funções ligadas a quebra de celulose (80). Outros grupos presentes neste mesmo genoma, tais como as GH10 e GH11, estão relacionadas com a degradação de xilano. Microrganismos com a capacidade de degradação de xilano são mais raros (78), e no consórcio microbiano, todos

os microrganismos com o genoma reconstruído apresentam um potencial gênico para a degradação deste polissacarídeos.

4.2.3 Anotação Funcional dos genomas reconstruídos do consórcio microbiano

A categoria COG associada ao metabolismo e transporte de carboidratos (G) foi a mais abundantes no genoma Bin1, enquanto que no genoma Bin6 a categoria mais abundante foi a de metabolismo e transporte de aminoácidos (E) (Figura 4.3). Os outros quatro genomas tiveram como categoria mais abundante a categoria R (Funções gerais apenas preditas).

A presença de um grande número de genes associados ao metabolismo e transporte de carboidratos no genoma Bin1 pode está relacionado com a sua capacidade em degradação açúcares maiores (polissacarídeos) em açúcares menores (monossacarídeos). Essa funções é de extrema importância no processamento de material de origem vegetal presente na compostagem.

Uma maior abundância da categoria E no genoma Bin6 está associada com a presença de uma diversidade de proteases e peptidases neste genoma, visto que foram identificadas inúmeras peptidase sinal lipoprotéicas, peptidases de processamento de fatores de esporulação e outras peptidases.

As diferenças encontradas nas proporções das categories COG nos seis genomas microbianos, principalmente na categoria G (Metabolismo e Transporte de Carboidratos), pode está relacionado com as diferentes estratégias de degradação de açúcares.

Tabela 4.12: Diversidade e abundância de GHs presentes nos genomas reconstruídos em um consórcio microbiano termofílico e celulolítico da compostagem

Família do CAZy	Atividades conhecidas	Bin1	Bin2	Bin3	Bin4	Bin5	Bin6
GH43	Alpha-L-arabinofuranosidase; beta-D-xylosidase	9	0	0	0	1	0
GH51	Endoglucanase; alpha-L-arabinofuranosidase	4	1	3	0	0	0
GH4	Alpha-Glucosidase; alpha-galactosidase; alpha-glucuronidase	3	2	0	1	3	0
GH2	eta-Galactosidase; Beta-mannosidase	3	1	0	1	0	0
GH10	Xylanase; Beta-1,3-endoxylanase	3	0	0	2	0	0
GH9	endoglucanase	2	0	0	0	0	0
GH3	Beta-1,4-Glucosidase; Beta-1,4-xylosidase; Beta-1,3-glucosidase; alpha-l-arabinofuranosidase	2	1	0	0	1	0
GH8	Cellulase; Beta-1,3-glucosidase; Beta-1,4-endoxylanase; Beta-1,4-endomannanase	1	0	1	0	0	0
GH67	Alpha-Glucuronidase	1	0	0	0	0	0
GH30	Beta-1,6-glucanase; Beta-xylosidase	1	1	0	0	0	0
GH11	endo-beta-1,4-xylanase; endo-beta-1,3-xylanase	1	0	0	0	0	0
GH113	eta-Xylosidase; Alpha-L-iduronidase	1	0	0	0	0	0
GH39	Beta-mannanase	0	0	1	0	0	0
GH26	Beta-Glucosidase; Beta-galactosidase; Beta-mannosidase	0	0	1	0	0	0
GH1	Cellulase; Beta-1,4-endoglucanase; Beta-1,3-glucosidase; Beta-1,4-endoxylanase; Beta-1,4-endomannanase	0	0	1	0	1	1
GH5	Beta-xylosidase Alpha-amylase	0	0	0	0	1	0
GH52		0	0	0	0	1	0
GH57		0	0	0	0	0	1

Figura 4.3: Abundância de categorias COGs nos genomas reconstruídos do metagenoma de um consórcio termofílico e celulolítico da compostagem.

Capítulo 5

Conclusões

Um dos objetivos específicos deste projeto foi desenvolver *pipelines* computacionais que automatizasse a reconstrução de genomas e identificação de GHs a partir de dados metagenômicos da compostagem. Inicialmente estava previsto a reconstrução de genomas apenas nas amostras de compostagem (ZC4), mas um novo experimento de construção de consórcio microbiano (ZCTH02) foi desenvolvido durante o andamento do temático. Parte dos scripts e softwares utilizados nas análises de ZC4 também foram aplicadas em ZCTH02.

Os outros três objetivos deste projeto estavam ligados a identificação e estudo de diversidade de genes que codificam glicosil hidrolases nos genomas reconstruídos. As sete espécies mais abundantes em comunidades microbianas da compostagem apresentam um amplo repertório gênico para a degradação de biomassa vegetal e também apresentam variações pontuais nas sequências de nucleotídeos de genes marcadores de cópia-única, sendo um indicativo de novas cepas, espécies ou até mesmo gêneros.

Nas análises do consórcio foi possível reconstruir seis genomas microbianos, todos eles associados com a degradação de biomassa vegetal. Além do mais, análises filogenéticas e filogenômicas indicam a presença de pelo menos uma nova classe dentro do filo Firmicutes, bem como a identificação de uma nova espécie da família Paenibacillaceae e

uma espécie do gênero *Thermobacillus*. Também conseguimos reconstruir pela primeira vez o genoma da espécie *Bacillus thermozeamaize*.

Referências Bibliográficas

- [1] Jay T Lennon and Stuart E Jones. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology*, 9(2):119–130, 2011.
- [2] Y. W. Wu, Y. H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2:26, 2014.
- [3] S Gajalakshmi and SA Abbasi. Solid waste management by composting: state of the art. *Critical Reviews in Environmental Science and Technology*, 38(5):311–400, 2008.
- [4] George E Fitzpatrick, Eva C Worden, and Wagner A Vendrame. Historical development of composting technology during the 20th century. *HortTechnology*, 15(1):48–51, 2005.
- [5] Michael J Dougherty, Patrik D’haeseleer, Terry C Hazen, et al. Glycoside hydrolases from a targeted compost metagenome, activity-screening and functional characterization. *BMC biotechnology*, 12(1):38, 2012.
- [6] Fei Wen, Nikhil U Nair, and Huimin Zhao. Protein engineering in designing tailored enzymes and microorganisms for biofuels production. *Current opinion in biotechnology*, 20(4):412–419, 2009.

- [7] Layla Farage Martins, Luciana Principal Antunes, Renata C Pascon, et al. Metagenomic analysis of a tropical composting operation at the são paulo zoo park reveals diversity of biomass degradation functions and organisms. *PloS one*, 8(4):e61928, 2013.
- [8] L. Wang and I. Ganly. The oral microbiome and oral cancer. *Clin. Lab. Med.*, 34(4):711–719, Dec 2014.
- [9] M. L. Sogin, H. G. Morrison, J. A. Huber, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.*, 103(32):12115–12120, Aug 2006.
- [10] David A Caron and Peter D Countway. Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquatic Microbial Ecology*, 57(3):227, 2009.
- [11] Carlos Pedrós-Alió. The rare bacterial biosphere. *Annual review of marine science*, 4:449–466, 2012.
- [12] Thomas Jefferson Sharpton. An introduction to the analysis of shotgun metagenomic data. *Plant Genetics and Genomics*, 5:209, 2014.
- [13] S. Schloissnig, M. Arumugam, S. Sunagawa, et al. Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50, Jan 2013.
- [14] D. B. Rusch, A. L. Halpern, G. Sutton, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, 5(3):e77, Mar 2007.
- [15] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Apr 2012.

- [16] B. Niu, Z. Zhu, L. Fu, S. Wu, and W. Li. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*, 27(12):1704–1705, Jun 2011.
- [17] Jonathan Laserson, Vladimir Jojic, and Daphne Koller. Genovo: de novo assembly for metagenomes. *Journal of Computational Biology*, 18(3):429–443, 2011.
- [18] Yu Peng, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–i101, 2011.
- [19] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155–e155, 2012.
- [20] Binbin Lai, Ruogu Ding, Yang Li, Liping Duan, and Huaiqiu Zhu. A de novo metagenomic assembly program for shotgun dna reads. *Bioinformatics*, 28(11):1455–1462, 2012.
- [21] K. C. Wrighton, B. C. Thomas, I. Sharon, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, 337(6102):1661–1665, Sep 2012.
- [22] M. Albertsen, P. Hugenholtz, A. Skarshewski, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, 31(6):533–538, Jun 2013.
- [23] I. Sharon, M. J. Morowitz, B. C. Thomas, et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, 23(1):111–120, Jan 2013.
- [24] G. J. Dick, A. F. Andersson, B. J. Baker, et al. Community-wide analysis of microbial genome sequence signatures. *Genome Biol.*, 10(8):R85, 2009.

- [25] F. Rodriguez-Valera, A. B. Martin-Cuadrado, B. Rodriguez-Brito, et al. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.*, 7(11):828–836, Nov 2009.
- [26] Vincent Lombard, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M Coutinho, and Bernard Henrissat. The carbohydrate-active enzymes database (cazy) in 2013. *Nucleic acids research*, 42(D1):D490–D495, 2014.
- [27] Brandi L Cantarel, Pedro M Coutinho, Corinne Rancurel, et al. The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics. *Nucleic acids research*, 37(suppl 1):D233–D238, 2009.
- [28] JPLF Cairo, Flávia C Leonardo, Thabata M Alvarez, et al. Functional characterization and target discovery of glycoside hydrolases from the digestome of the lower termite *Coptotermes gestroi*. *Biotechnol. Biofuels*, 4:50, 2011.
- [29] Michael E Himmel, Shi-You Ding, David K Johnson, et al. Biomass recalcitrance: engineering plants and enzymes for biofuels production. *science*, 315(5813):804–807, 2007.
- [30] B Henrissat, M Claeyssens, P Tomme, L Lemesle, and J-P Mornon. Cellulase families revealed by hydrophobic cluster analysis. *Gene*, 81(1):83–95, 1989.
- [31] PB Pope, SE Denman, M Jones, et al. Adaptation to herbivory by the tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proceedings of the National Academy of Sciences*, 107(33):14793–14798, 2010.
- [32] M. Hess, A. Sczyrba, R. Egan, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–467, Jan 2011.

- [33] D. Wu, M. Wu, A. Halpern, et al. Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLoS ONE*, 6(3):e18011, 2011.
- [34] L. F. Roesch, R. R. Fulthorpe, A. Riva, et al. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*, 1(4):283–290, Aug 2007.
- [35] J. G. Caporaso, C. L. Lauber, W. A. Walters, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.*, 108 Suppl 1:4516–4522, Mar 2011.
- [36] G. W. Tyson, J. Chapman, P. Hugenholtz, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, Mar 2004.
- [37] Ruibang Luo, Binghang Liu, Yinlong Xie, et al. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18, 2012.
- [38] Sergey Nurk, Anton Bankevich, Dmitry Antipov, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In *Research in Computational Molecular Biology*, pages 158–170. Springer, 2013.
- [39] H. Garcia Martin, N. Ivanova, V. Kunin, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, 24(10):1263–1269, Oct 2006.
- [40] M. J. Morowitz, V. J. Denef, E. K. Costello, et al. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc. Natl. Acad. Sci. U.S.A.*, 108(3):1128–1133, Jan 2011.

- [41] V. Iverson, R. M. Morris, C. D. Frazar, et al. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, 335(6068):587–590, Feb 2012.
- [42] I. Sharon, M. J. Morowitz, B. C. Thomas, et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, 23(1):111–120, Jan 2013.
- [43] C. T. Brown, L. A. Hug, B. C. Thomas, et al. Unusual biology across a group comprising more than 15Bacteria. *Nature*, 523(7559):208–211, Jul 2015.
- [44] N.A. Joshi and J.N. Fass. Sickel: A sliding-window, adaptive, quality-based trimming tool for fastq files. *Github*, 1(1):1, 2011.
- [45] C. Luo, L. M. Rodriguez-R, and K. T. Konstantinidis. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.*, 42(8):e73, Apr 2014.
- [46] Beifang Niu, Zhengwei Zhu, Limin Fu, Sitao Wu, and Weizhong Li. Fr-hit, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*, 27(12):1704–1705, 2011.
- [47] G. L. Kay, M. J. Sergeant, V. Giuffra, et al. Recovery of a Medieval *Brucella melitensis* Genome Using Shotgun Metagenomics. *MBio*, 5(4), 2014.
- [48] M. Margulies, M. Egholm, W. E. Altman, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.
- [49] Zhengwei Zhu, Beifang Niu, Jing Chen, et al. Mgaviewer: A desktop visualization tool for analysis of metagenomics alignment data. *Bioinformatics*, 29(1):122–123, 2013.

- [50] T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, Jul 2014.
- [51] D. Hyatt, G. L. Chen, P. F. Locascio, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, 2010.
- [52] K. Lagesen, P. Hallin, E. A. Rodland, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, 35(9):3100–3108, 2007.
- [53] Yanbin Yin, Xizeng Mao, Jincai Yang, et al. dbcan a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research*, 40(W1):W445–W451, 2012.
- [54] S. F. Altschul, T. L. Madden, A. A. Schaffer, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997.
- [55] J. R. Cole, Q. Wang, J. A. Fish, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, 42(Database issue):D633–642, Jan 2014.
- [56] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, Aug 2004.
- [57] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, 2010.
- [58] N. Segata, D. Bornigen, X. C. Morgan, and C. Huttenhower. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun*, 4:2304, 2013.
- [59] I. Sa-Nogueira and L. J. Mota. Negative regulation of L-arabinose metabolism

- in *Bacillus subtilis*: characterization of the *araR* (*araC*) gene. *J. Bacteriol.*, 179(5):1598–1608, Mar 1997.
- [60] K. B. Krogh, P. V. Harris, C. L. Olsen, et al. Characterization and kinetic analysis of a thermostable GH3 beta-glucosidase from *Penicillium brasilianum*. *Appl. Microbiol. Biotechnol.*, 86(1):143–154, Mar 2010.
- [61] Konstantinos Liolios, Johannes Sikorski, Marlen Jando, et al. Complete genome sequence of *thermobispora bispora* type strain (r51t). *Standards in genomic sciences*, 2(3):318, 2010.
- [62] A. Pati, K. Labutti, R. Pukall, et al. Complete genome sequence of *Sphaerobacter thermophilus* type strain (S 6022). *Stand Genomic Sci*, 2(1):49–56, 2010.
- [63] K. Ueda, M. Ohno, K. Yamamoto, et al. Distribution and diversity of symbiotic thermophiles, *Symbiobacterium thermophilum* and related bacteria, in natural environments. *Appl. Environ. Microbiol.*, 67(9):3779–3784, Sep 2001.
- [64] Kenji Ueda, Michiyo Ohno, Kaori Yamamoto, et al. Distribution and diversity of symbiotic thermophiles, *symbiobacterium thermophilum* and related bacteria, in natural environments. *Applied and environmental microbiology*, 67(9):3779–3784, 2001.
- [65] L. Pasic, B. guez Mueller, A. B. Martin-Cuadrado, et al. Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics*, 10:570, 2009.
- [66] Lejla Pašić, Beltran Rodriguez-Mueller, Ana-Belen Martin-Cuadrado, et al. Metagenomic islands of hyperhalophiles: the case of *salinibacter ruber*. *Bmc Genomics*, 10(1):570, 2009.
- [67] Ralf Koebnik. Tonb-dependent trans-envelope signalling: the exception or the rule? *Trends in microbiology*, 13(8):343–347, 2005.

- [68] Wolfgang Rabsch, Li Ma, Graham Wiley, et al. Fepa-and tonb-dependent bacteriophage h8: receptor binding and genomic sequence. *Journal of bacteriology*, 189(15):5658–5674, 2007.
- [69] Eun Hee Cho, Richard I Gumport, and Jeffrey F Gardner. Interactions between integrase and excisionase in the phage lambda excisive nucleoprotein complex. *Journal of bacteriology*, 184(18):5200–5203, 2002.
- [70] Marco Ventura, Aldert Zomer, Carlos Canchaya, et al. Comparative analyses of prophage-like elements present in two lactococcus lactis strains. *Applied and environmental microbiology*, 73(23):7771–7780, 2007.
- [71] Marco Ventura, Francesca Turrone, Gipsi Lima-Mendez, et al. Comparative analyses of prophage-like elements present in bifidobacterial genomes. *Applied and environmental microbiology*, 75(21):6929–6936, 2009.
- [72] Marco Ventura, Carlos Canchaya, David Pridmore, Bernard Berger, and Harald Brüßow. Integration and distribution of lactobacillus johnsonii prophages. *Journal of bacteriology*, 185(15):4603–4608, 2003.
- [73] Daniel Rios Garza, Cristiane C Thompson, Edvaldo Carlos Brito Loureiro, et al. Genome-wide study of the defective sucrose fermenter strain of vibrio cholerae from the latin american cholera epidemic. *PloS one*, 7(5):e37283, 2012.
- [74] Pasi Partanen, Jenni Hultman, Lars Paulin, Petri Auvinen, and Martin Romantschuk. Bacterial diversity at different stages of the composting process. *BMC microbiology*, 10(1):94, 2010.
- [75] P.S. Mak. *Isolation and Characterization of Thermophilic and Hyperthermophilic Microorganisms from Food Processing Facilities*. Iowa State University, 2003.

- [76] D. Hayes, L. Izzard, and R. Seviour. Microbial ecology of autothermal thermophilic aerobic digester (ATAD) systems for treating waste activated sludge. *Syst. Appl. Microbiol.*, 34(2):127–138, Apr 2011.
- [77] W. Zhang and Z. Lu. Phylogenomic evaluation of members above the species level within the phylum Firmicutes based on conserved proteins. *Environ Microbiol Rep*, 7(2):273–281, Apr 2015.
- [78] D. B. Wilson. Three microbial strategies for plant cell wall degradation. *Ann. N. Y. Acad. Sci.*, 1125:289–297, Mar 2008.
- [79] C. Li, M. Du, B. Cheng, et al. Close relationship of a novel Flavobacteriaceae α -amylase with archaeal α -amylases and good potentials for industrial applications. *Biotechnol Biofuels*, 7(1):18, 2014.
- [80] Y. Honda, N. Shimaya, K. Ishisaki, M. Ebihara, and H. Taniguchi. Elucidation of exo-beta-D-glucosaminidase activity of a family 9 glycoside hydrolase (PB-PRA0520) from *Photobacterium profundum* SS9. *Glycobiology*, 21(4):503–511, Apr 2011.