UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Biológicas e da Saúde
Departamento de Genética e Evolução
Laboratório de Biologia Molecular

CÉLIO DIAS SANTOS JÚNIOR

DEGRADAÇÃO DE MATÉRIA ORGÂNICA TERRESTRE POR MICRORGANISMOS DO RIO AMAZONAS – METAGENÔMICA E GENÔMICA POPULACIONAL

SÃO CARLOS - SP
2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Biológicas e da Saúde
Departamento de Genética e Evolução
Laboratório de Biologia Molecular

CÉLIO DIAS SANTOS JÚNIOR

DEGRADAÇÃO DE MATÉRIA ORGÂNICA TERRESTRE POR MICRORGANISMOS DO RIO AMAZONAS – METAGENÔMICA E GENÔMICA POPULACIONAL

Tese de doutorado apresentada ao Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular do Centro de Ciências Biológicas e da Saúde da Universidade Federal de São Carlos para obtenção do título de Doutor em Genética Evolutiva e Biologia Molecular.

Orientador: Prof. Dr. Flávio Henrique Silva
Co-Orientador: Ramiro Logares

SÃO CARLOS - SP
2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Biológicas e da Saúde
Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Célio Dias Santos Júnior, realizada em 14/12/2018:

Prof. Dr. Flavio Henrique da Silva
UFSCar

Prof. Dr. Francis de Morais Franco Nunes
UFSCar

Prof. Dr. Hugo Miguel Preto de Morais Sarmento
UFSCar

Prof. Dr. Marcos Antonio de Oliveira
UNESP

Prof. Dr. Fernando Pellon de Miranda
Petrobras

DEDICATÓRIA

Dedico este trabalho aos meus avós, parte fundamental de quem eu sou e almejo ser. Também dedico aos meus pais, Célio e Sandra, cujo amor incondicional, incentivo e apoio me trouxeram até aqui.

EPÍGRAFE

“O período de maior ganho em conhecimento e experiência é o período mais difícil da vida de alguém. ” - Dalai Lama

SÚMULA CURRICULAR

FORMAÇÃO ACADÊMICA

– Doutorado em Genética Evolutiva e Biologia Molecular (em andamento). Tese: "Degradação de matéria orgânica terrestre por microrganismos do rio Amazonas – metagenômica e genômica populacional". Orientador: Flavio Henrique da Silva. Coorientador: Ramiro Logares. Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular, Universidade Federal de São Carlos, UFSCar, Brasil. Período sanduíche no Institut de Ciencias del Mar, Barcelona, Espanha (Orientador: Ramiro Logares). Bolsista do: Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brasil e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brasil.

– Mestrado em Genética Evolutiva e Biologia Molecular (2016). Dissertação: "Conversão de uma histidina amônia liase em fenilalanina amônia liase por meio de biologia sintética". Orientador: Flavio Henrique da Silva. Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular, Universidade Federal de São Carlos, UFSCar, Brasil. Bolsista do: Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brasil.

– Graduação em Biotecnologia (2013). Título: Diversidade e Potencial Biotecnológico das Bactérias Endofíticas da Mamona (*R. communis*). Orientador: Ana Maria Bonetti. Iniciação Científica no Laboratório de Genética. Graduação em Bacharelado, Universidade Federal de Uberlândia, UFU, Brasil. Período sanduíche na Universidade de Coimbra, Coimbra, Portugal (Orientador: Joana Cardoso Costa). Bolsista do: Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brasil.

PRÊMIOS E TÍTULOS

- Menção Honrosa pelo trabalho “IMPLEMENTAÇÃO DE UM MÓDULO DE IDENTIFICAÇÃO TAXONÔMICA DE MICRORGANISMOS BASEADO NO GENE DO 16S rRNA NO PROGRAMA BEAF E SUA APLICAÇÃO EM DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO”, de autoria de Guilherme Bonotti COPPINI, G.B. ; **SANTOS JUNIOR, C. D.** ; HENRIQUE-SILVA, F.. Instituição: Coordenadoria de Programas de Iniciação Científica (CoPICT), Pró-Reitoria de Pesquisa (ProPq), Universidade Federal de São Carlos (UFSCar) – 2018.

- Medalha de Ouro na *International Genetically Engineered Machine competition* (iGEM) pelo grupo team:UFSCar-Brasil. Instituição: *International Genetically Engineered Machine Foundation* (iGEM Foundation) - 2015.

- Moção de congratulação com relação ao projeto *BugShoo* apresentado no iGEM 2015. Instituição: Câmara de Vereadores de São Carlos-SP – 2015.

- Medalha de Bronze e Prêmio Giant Jamboree na *International Genetically Engineered Machine competition* (iGEM) pelo grupo team:Brasil-SP. Instituição: *International Genetically Engineered Machine Foundation* (iGEM) - 2014.

- Menção Honrosa pelo trabalho "THE FOLATE/PTERIDINE TRANSPORTER OF LEISHMANIA SP.: A POTENTIAL MOLECULAR TARGET WITH AN EVENTFUL EVOLVING HISTORY", de autoria **SANTOS JÚNIOR, C. D.**; ISABEL, T. F. ; TELES, N. ; GUIDO, R. V. C. ; HENRIQUE-SILVA, F.. Instituição: Sociedade Brasileira de Genética - 2014.

- Menção Honrosa pelo pôster "*Metagenomic Analysis of Endophytic Bacterial Biodiversity from Seeds of Ricinus communis*", de autoria **SANTOS JÚNIOR, C. D.**; DIAS, A. C. C. ; BONETTI, A.M. ; Kerr, W.E. ; CAMPOS, T. A.. Instituição: Sociedade Brasileira de Genética - 2011.

CO-ORIENTAÇÕES

1. Guilherme Bonotti Coppini. Desenvolvimento de um módulo de análises taxonômicas do gene 16S rRNA de metagenomas para o programa BEAF (*Referenced Binning Engine for Autonomous Finding*): BEAF-16S. 2018. Trabalho de Conclusão de Curso. (Graduação em Biotecnologia) - Universidade Federal de São Carlos, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Orientador: Flávio Henrique Silva. Co-Orientador: Célio Dias Santos Júnior.

2. Natália Silva da Trindade. *Conservation and Recombinational Contexts in Apoptosis Pathway in Aspergillus fumigatus*. 2015. Trabalho de Conclusão de Curso. (Graduação em Biotecnologia) - Universidade Federal de Uberlândia. Orientadora: Enyara Rezende Morais. Co-Orientador: Célio Dias Santos Júnior.

PUBLICAÇÕES

As publicações em cinza se destacam por possuir relação com os temas abordados na presente tese, assim como, aquelas assinaladas com "***" ao início foram produzidas durante o período de doutorado. O nome do autor desta tese foi mostrado em negrito.

- ***TRINDADE, N. S. ; **SANTOS JÚNIOR, C. D.** ; MACEDO, P. ; VIANA, J. ; ALVES, T. ; FURSTENAU, C. ; GOMES, M. S. ; MORAIS, E. R. . Conservation and Recombinational Contexts in Apoptosis Pathway in Aspergillus fumigatus. Journal of Computer Science & Systems Biology, v. 11, p. 136-153, 2018.

- ***SANTOS-JÚNIOR, C. D.** ; KISHI, L. T. ; TOYAMA, D. ; SOARES-COSTA, A. ; OLIVEIRA, T. C. S. ; DE MIRANDA, F. P. ; HENRIQUE-SILVA, F. . Metagenome Sequencing of Prokaryotic Microbiota Collected from Rivers in the Upper Amazon Basin. Genome Announcements, v. 5, p. e01450-16, 2017.

- ***CARDOSO-JÚNIOR, C. A. M. ; FUJIMURA, P. T. ; **SANTOS-JÚNIOR, C. D.** ; BORGES, N. A. ; UEIRA-VIEIRA, C. ; HARTFELDER, K. ; GOULART, L. R. ; BONETTI, A. M. . Epigenetic modifications and their relation to caste and sex determination and adult division of labor in the stingless bee Melipona scutellaris. Genetics and Molecular Biology (online version), v. 02, p. 1-8, 2017.

- ***NAKAYAMA, D. G. ; **SANTOS JÚNIOR, C. D.** ; KISHI, L. T. ; PEDEZZI, R. ; SANTIAGO, A. C. ; SOARES-COSTA, A. ; HENRIQUE-SILVA, F. . A transcriptomic survey of Migdolus fryanus (sugarcane rhizome borer) larvae. Plos One, v. 12, p. e0173059, 2017.

- ***TOYAMA, D. ; **SANTOS-JÚNIOR, C.D.** ; Kishi, L.T. ; OLIVEIRA, T.C.S. ; GARCIA, J.W. ; SARMENTO, H. ; MIRANDA, F.P. ; HENRIQUE-SILVA, F. . A snapshot on prokaryotic diversity of the Solimões River basin (Amazon, Brazil). GENETICS AND MOLECULAR RESEARCH, v. 16, p. 1-17, 2017.

- ***ISABEL, T. F. ; COSTA, G. N. M. ; PACHECHO, I. B. ; **JÚNIOR, C. D. S.** ; FONSECA, F.P. ; FRANCA, J. B. ; HENRIQUE-SILVA, F. ; YONEYAMA, K. A. ; RODRIGUES, R. S. ; RODRIGUES, V. M. . Expression and Partial Biochemical Characterization of a Recombinant Serine Protease from Bothrops pauloensis Snake Venom. Toxicon (Oxford), v. 115, p. 49-54, 2016.

- ***SANTOS-JÚNIOR, CÉLIO D.**; VERÍSSIMO, A. ; COSTA, J. . The recombination dynamics of Staphylococcus aureus inferred from spA gene. BMC Microbiology (Online), v. 16, p. 143, 2016.

- ***TOYAMA, D. ; KISHI, L. T. ; **SANTOS-JÚNIOR, C. D.** ; SOARES-COSTA, A. ; DE OLIVEIRA, T. C. S. ; DE MIRANDA, F. P. ; HENRIQUE-SILVA, F. . Metagenomics Analysis of Microorganisms in Freshwater Lakes of the Amazon Basin. Genome Announcements, v. 4, p. e01440-16, 2016.

- MENDES, M.G. ; **SANTOS JUNIOR, C.D.** ; DIAS, A.C.C. ; BONETTI, A.M. . Castor bean (Ricinus communis L.) as a potential environmental bioindicator. Genetics and Molecular Research, v. 14, p. 12880-12887, 2015.

- SOUSA, L. B. ; HAMAWAKI, O. T. ; **SANTOS JUNIOR, C.D.** ; OLIVEIRA, V. M. ; NOGUEIRA, A. P. O. ; MUNDIM, F. M. ; HAMAWAKI, R. L. ; HAMAWAKI, C. D. L. . Correlation between yield components in F6 soybean progenies derived from seven biparental crosses. Bioscience Journal (Online), v. 31, p. 1692-1699, 2015.

- LUIZ, D. P. ; **SANTOS JÚNIOR, C. D.** ; BONETTI, A.M. ; BRANDEBURGO, M. M. . Tollip or Not Tollip: What Are the Evolving Questions behind It?. Plos One, v. 9, p. e97219, 2014.

- COSTA, J. ; TEIXEIRA, P. G. ; D'AVÓ, A. F. ; **JÚNIOR, C. S.** ; VERÍSSIMO, A. . Intragenic Recombination Has a Critical Role on the Evolution of Legionella pneumophila Virulence-Related Effector sidJ. Plos One, v. 9, p. e109840, 2014.

- PEDEZZI, R. ; FONSECA, F.P. ; **SANTOS JÚNIOR, C. D.** ; KISHI, L.T. ; TERRA, W.R. ; HENRIQUE-SILVA, F. . A novel β-fructofuranosidase in Coleoptera: Characterization of a β-fructofuranosidase from the sugarcane weevil, Sphenophorus levis. Insect Biochemistry and Molecular Biology, v. 55, p. 31-38, 2014.

- **SANTOS JR., C.D.**; DIAS, A.C.C. ; AMARAL, I.M.R. ; BONETTI, A.M. ; CAMPOS, T.A. . New efficient DNA extraction method to access the microbiome of Ricinus communis seeds. Genetics and Molecular Research, v. 12, p. 1-8, 2013.

- VIEIRA, FCF ; **SANTOS JÚNIOR, C. D.** ; NOGUEIRA, A. P. O. ; DIAS, A. C. C. ; HAMAWAKI, O. T. ; BONETTI, A.M. . ASPECTOS FISIOLÓGICOS E BIOQUÍMICOS DE CULTIVARES D E SOJA SUBMETIDOS A DÉFICIT HÍDRICO INDUZIDO POR PEG 6000. Bioscience Journal (Online), v. 29, p. 543-552, 2013.

- HAMAWAKI, O. T. ; SOUSA, L. B. ; ROMANATO, F. N. ; NOGUEIRA, A. P. O. ; **SANTOS JÚNIOR, C. D.** ; POLIZEL, A. C. . Genetic parameters and

variability in soybean genotypes. Comunicata Scientiae (Online), v. 3, p. 76-83, 2012.

- **SANTOS JÚNIOR, C. D.**; LUIZ, D. P. ; BONETTI, A.M. . The Effect of 3,5-Dinitrosalicylic Acid on Genomic DNA from Saccharomyces cerevisiae. BAG. Journal of basic and applied genetics, v. 23, p. 219, 2012.

CAPÍTULOS DE LIVROS PUBLICADOS

- **JÚNIOR, C. D. S.**; TELES, N. M. M. ; LUIZ, D. P. ; ISABEL, T. F. . DNA Extraction from Seeds. In: Miodrag Micic. (Org.). Springer Protocols Handbooks. 01ed.New York: Springer New York, 2016, v. 01, p. 265-276.

- **SANTOS JÚNIOR, C. D.**. Aditivos antimicrobiais aplicados à biotecnologia de alimentos. In: BOSCOLLI BARBOSA PEREIRA. (Org.). Aditivos alimentares [livro eletrônico]: conceitos, aplicações e toxicidade. 01ed. Monte Carmelo, MG: FUCAMP, 2013, v. 1, p. 41-76.

PUBLICAÇÕES PENDENTES

- **SANTOS-JÚNIOR, C. D.** ; TOYAMA, D. ; OLIVEIRA, T. C. S.; MIRANDA, F. P.; HENRIQUE-SILVA, F. . Flood season microbiota collected from lakes of Amazon basin revealed through metagenome sequencing. (A ser submetido à revista “Microbiology Resource Announcements”)

- **SANTOS-JÚNIOR, C. D.** ; SARMENTO, H. ; MIRANDA, F. P.; HENRIQUE-SILVA, F. ; LOGARES, R. . Connecting the Amazon river microbial genes and terrestrial organic matter degradation. (A ser submetido à revista “Nature Communications”)

- **SANTOS-JÚNIOR, C. D.** ; TOYAMA, D. ; LOGARES, R. ; MIRANDA, F. P. ; HENRIQUE-SILVA, F. . Population genomes reveal the microbial drivers of the terrestrial organic matter degradation in Amazon river basin. (A ser submetido à revista “Microbiome”)

- **SANTOS-JÚNIOR, C. D.** ; COPPINI, G. B. ; HENRIQUE-SILVA, F. . Referenced Binning Engine for Autonomous Finding - BEAF. (A ser submetido à revista “Bioinformatics”)

AGRADECIMENTOS

À minha família, especialmente meus avós, pais e irmã, pelo apoio, paciência, confiança e incentivo.

Ao Prof. Dr. Flávio Henrique Silva pela orientação e pelos ensinamentos ao longo desses anos.

Ao Prof. Dr. Ramiro Logares e seus alunos do Institut de Ciencies del Mar (Barcelona, Espanha), em especial Lídia Montiel, pela cortesia de me receber em seu laboratório e pelos ensinamentos os quais me auxiliaram a concluir o presente trabalho.

Ao Prof. Dr. Pablo Sanchez do Institut de Ciencies del Mar (Barcelona, Espanha) pela cortesia e ensinamentos essenciais para a conclusão do presente trabalho.

Ao pessoal técnico do servidor MareNostrum sediado em Barcelona/Espanha, pelo auxílio.

Ao Prof. Dr. Hugo Sarmento do Departamento de Hidrobiologia da UFSCar pela sua colaboração na disponibilização de seu tempo e seu servidor para realização do presente trabalho.

Aos Pós-doutorandos Danyelle Toyama e Chakravarthi Mohan, à Doutoranda Priscila Shibao, à mestranda Heloisa Peccin e ao graduando Guilherme Coppini pela valiosa amizade, pelos ensinamentos e pela ajuda prestados.

Aos amigos Pattrick Nguyen, Fatim Belkrane, Anja Cosic, Rejane Monte, Lorena Cruz, Michaella Melo, Thaís Godoy e a todos os outros amigos os quais não foi possível citar, pela animada companhia durante esses anos. Pelo incentivo, ajuda e por tornar meu dia-a-dia mais alegre.

Ao Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular pelo auxílio e suporte, em especial, à Ivanildes que sempre me auxiliou de forma exemplar.

Ao CNPq, CAPES, Petrobrás e CSIC (Consejo Superior de Investigacion Cientifica) pelo suporte financeiro durante a realização do presente trabalho.

RESUMO

Os micróbios da bacia do rio Amazonas representam uma biodiversidade inexplorada, com grande potencial metabólico. Este ecossistema recebe grandes quantidades de matéria orgânica (*organic matter,* OM) terrestre, o que promove o crescimento microbiano heterotrófico. Populações microbianas foram expostas durante milênios a uma OM complexa derivada de plantas, e essas comunidades devem ter desenvolvido vias metabólicas para degradá-la. No presente trabalho, foram analisados 106 metagenomas provenientes de 30 estações de amostragem de rios e lagos da bacia Amazônica, cobrindo as zonas de água doce e costeira. Utilizando-se as técnicas padrão ouro para montagem e predição gênica, foi gerado o primeiro catálogo de genes de microrganismos de água doce, contendo mais de 3,7 milhões de genes não redundantes, com predominância daqueles de origem bacteriana (35,73%). Este estudo compreendeu até então a maior biodiversidade de microrganismos aquáticos amostrados no rio Amazonas. Como um sistema auxiliar, foi criado o programa AGSSY que auxilia na mineração de dados deste catálogo, mostrando-se rápido e eficiente neste processo. A análise de diversidade dos k-mers sugere que esses genes tem origem a partir de processos evolutivos locais. Além disso, há uma estratificação do processamento da OM na coluna de água, possivelmente regulada por um sistema sofisticado de uso alternativo de fontes de carbono, principalmente baseado em tricarboxilatos. A estrutura espacial dos genes de processamento de OM sugere uma substituição dos metabolismos de lignina e hemi-celulose, pelo metabolismo de celulose no oceano. Utilizando-se os métodos de *binning* híbridos, levando em conta composição e abundância dos *contigs* foi possível a reconstrução de 51 genomas populacionais não-redundantes. Por meio da análise destes, foi possível verificar espécies endêmicas abundantes pertencentes a Bactérias e Arquéias, com um predomínio do filo Proteobactéria (39%). Análises dos genomas populacionais, em conjunto com os dados do catálogo de genes, sugerem um sofisticado modelo de *priming* para a degradação de OM terrestre no rio Amazonas. Tal modelo seria baseado na retroinibição sofrida pelos organismos que oxidam a lignina, e numa comunidade de microrganismos utilizadores de fontes alternativas de carbono, que bloqueiam este efeito.

PALAVRAS-CHAVE: Rio Amazonas, Genoma Populacional, Metagenoma, Catálogo de Genes, Efeito *priming*, degradação de matéria orgânica terrestre.

ABSTRACT

Microbes from Amazon river basin represent an unexplored biodiversity, with a huge metabolic potential. This ecosystem receives large amounts of terrestrial organic matter (OM), which promotes heterotrophic microbial growth. Microbial populations have been exposed for millennia to a complex OM derived from plants, and these communities must have developed metabolic pathways to degrade it. In the present work, 106 metagenomes from 30 sampling stations of rivers and lakes of the Amazon river basin were analyzed, covering the freshwater and coastal ocean areas. Using the gold standard techniques for assembly and gene prediction, the first gene catalog of freshwater microorganisms was generated, containing more than 3.7 million non-redundant genes, predominantly those of bacterial origin (35.73%). The present work comprises the hugest biodiversity ever sampled in Amazon river until now. As an auxiliary system, the AGSSY program was created to assist in the data mining of this catalog, proving to be fast and efficient in this process. The analysis of k-mers diversity suggests that these genes originate from local evolutionary processes. In addition, there is a stratification of the OM processing in the water column, possibly regulated by a sophisticated system of alternative carbon sources use, mainly based on tricarboxylates. The spatial structure of the OM processing genes suggests a zonation of lignin, cellulose and hemicellulose degradation. Using hybrid binning methods, taking into account contigs composition and abundance, it was possible to reconstruct 51 non-redundant population genomes. Through the analysis of them, it was possible to verify abundant endemic species belonging to Bacteria and Archaea, with a predominance of the Proteobacterium phylum (39%). Population genome analyzes, together with the gene catalog data, suggest a sophisticated priming model for the degradation of terrestrial OM in the Amazon River. Such a model would be based on the retroinhibition undergone by organisms that oxidize lignin, and in a community of microorganisms using alternative carbon sources that blocks this effect.

KEY WORDS: Amazon river, Population genomes, Metagenomes, Gene catalogue, Priming effect, terrestrial organic matter degradation.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIAÇÕES E SIGLAS

AAI	Identidade média de aminoácidos
AF	Fração alinhada
AMnrGC	Catálogo de Genes Microbianos não-Redundantes da Bacia do Rio Amazonas
ANI	Identidade média de nucleotídeos
BEAF	Dispositivo automático de *binning* por referência
C1	Metabolismo de compostos de 1 carbono, como metano e dióxido de carbono
CAMERA	*Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis*
CBM	*Motif* de ligação à celulose ou carbo-hidratos
COG	*Clusters of Ortholog Genes*
CRAM	Moléculas alicíclicas ricas em carboxila
dbCAN	*Database for automated Carbohydrate-active enzyme ANnotation*
DNA	Ácido desoxirribonucleico
DYP	Peroxidases descolorantes de corantes
eggNOG	*Evolutionary genealogy of genes: Non-supervised Orthologous Groups*
ENA	*European Nucleotide Archive*
GC	Conteúdo de Guanina - Citosina em uma fita de DNA/RNA dado em %
GH	Glicosil hidrolases
Gpb	10^9 pares de base
GTDB	*Genome Taxonomy Database*

HMM	Cadeias ocultas de Markov
KEGG	*Kyoto Enciclopedia for Genes and Genomes*
km	Quilômetros
KO	*KEGG ortholog number*
Kpb	10^3 pares de base
Mpb	10^6 pares de base
NCBI	*National center for Biotechnology Information*
NMDS	Escalonamento multidimensional não-métrico
OF	Fração ortóloga
OM	*Organic matter* ou Matéria Orgânica
OTU	*Operational Taxonomic Unit* / Unidade Taxonômica Operacional
pb	Pares de bases
PFAM	*Protein Families Database*
PG	Genomas populacionais ou MAGs
PHB/A	Poli-hidróxi-butirato e poli-hidróxi-alcanoato
RNA	Ácido ribonucleico
TCA	Ciclo do ácido tricarboxílico
TPM	Transcritos por milhão
tRNA	RNA transportador
TTT	Sistema tripartido de transporte de tricarboxilatos
UFAC	Usuário de fontes alternativas de carbono

SUMÁRIO

Capítulo 1 - Introdução

1.1 Fundamentação e justificativa

A bacia do rio Amazonas é uma das maiores e menos estudadas no mundo, compreendendo cerca de 38% da área continental sul-americana (MIKHAILOV, 2010). Por volta de 60% da matéria orgânica (OM) dissolvida do rio Amazonas é derivada de compostos lignínicos e celulósicos (ERTEL et al., 1986), assim como ácidos húmicos, que são rmetabolizados pelos micróbios (principalmente bactérias) levando à liberação de dióxido de carbono de suas águas (WARD et al., 2013, 2016).

Diversos metagenomas de estações situadas no baixo Amazonas, após a cidade de Manaus até seu estuário, mostraram-se principalmente estruturadas em torno de *Cyanobacteria* e outros táxons diazotróficos (SATINSKY et al., 2015), sugerindo uma substituição da OM terrestre pela algal nas porções finais do rio. Isto pressupõe que os microrganismos ao longo do rio consomem a OM terrestre.

Genes com funções relacionadas à degradação de lignina e celulose, portanto, deveriam ser abundantes. No entanto, estudos anteriores (SATINSKY; CRUMP; et al., 2014; SATINSKY; SMITH; SHARMA; LANDA; et al., 2017; SATINSKY; SMITH; SHARMA; WARD; et al., 2017) mostram que tais funções gênicas são encontradas em baixas quantidades na porção inferior do rio Amazonas, entre Manaus e seu estuário. Estes mesmos autores hipotetizam que esta descoberta poderia ser explicada pela pobre compreensão que temos das enzimas envolvidas na degradação de lignina e celulose em sistemas aquáticos. Entretanto, dentre as razões para baixa detecção poder-se-ia destacar as metodologias analíticas utilizadas, baseadas em recrutamento de *reads* por genes e genomas conhecidos, reduzindo consideravelmente a diversidade amostrada. A determinação do repertório genético presente neste microbioma é um passo chave para a

compreensão dos mecanismos envolvidos na degradação de OM terrestre proveniente da floresta tropical que o circunda.

Em 2013, verificou-se que somente cerca de 5% da lignina que entrava no rio Amazonas chegava aos oceanos (WARD et al., 2013), de modo que esse material recalcitrante poderia estar sendo consumido pelos microrganismos do rio. Recentemente, a degradação de lignina no rio Amazonas foi verificada como sendo potencializada pela degradação de compostos mais lábeis (WARD et al., 2016), o chamado *priming effect.* Pouco se sabe sobre este efeito, no que tange seus mecanismos moleculares ou agentes microbianos, assim a resolução dessas questões poderia trazer maior conhecimento acerca do funcionamento de sistemas fluviais tropicais e auxiliar na compreensão da ciclagem de carbono global.

A geração de catálogos de genes fornece uma fonte única de informação baseada nos genes e funções das proteínas preditas. Diferentemente do recrutamento de *reads,* em que curtos fragmentos de DNA são associados à determinadas funções gênicas, os catálogos de genes são uma coleção de genes preditos obtidos de *reads* montadas em *contigs*. As inferências obtidas por meio de catálogos de genes têm maior confiabilidade do que aquelas tomadas a partir da análise funcional de *reads*, pois utilizam um conjunto de *motifs* e permitem analisar as abundâncias dos mesmos por metagenoma de modo mais acurado, eliminando o efeito de *singletons*. Dessa forma, a análise de dados metagenômicos por essa estratégia poderia revelar efeitos antes não observados no conjunto de dados da microbiota do rio Amazonas.

Outra questão interessante é avaliar os genes em termos genômicos, uma vez que esses dados sem contexto podem revelar relações ecológicas pouco explícitas, que

contextualizadas, fornecem um panorama completo. Por isso, genomas populacionais, obtidos a partir da classificação de *contigs* amazônicos, são outra importante ferramenta para elucidar esses aparatos bioquímicos de degradação de OM terrestre.

1.2 Objetivos e escopo

Esta tese descreve os aparatos bioquímicos presentes na microbiota fluvial amazônica, por meio da geração de um catálogo de genes e recuperação de genomas populacionais à partir de *contigs* obtidos dos metagenomas. Apesar do panorama geral da composição microbiana em termos taxonômicos estar razoavelmente estabelecida por meio de estudos anteriores, em termos de funções metabólicas, pouco se sabe sobre a degração de OM terrestre no sistema fluvial amazônico. O principal objetivo foi elucidar de modo geral os principais mecanismos moleculares envolvidos na degradação de OM terrestre pelos microrganismos da bacia Amazônica. O problema de pesquisa foi dividido nas seguintes questões:

Q1. A diversidade gênica presente nos microrganismos do rio Amazonas reflete processos evolutivos locais ou é similar a outros sistemas fluviais?

Q2. Há uma estrutura espacial na ocorrência de genes ao longo do curso do rio Amazonas que poderia indicar especialização metabólica?

Q3. Quais as principais famílias de proteínas associadas à degradação de OM terrestre?

Q4. Há mecanismos alternativos promovendo o uso de fontes de carbono secundárias ou o armazenamento de carbono na microbiota da bacia Amazônica?

Q5. Há evidências de um *priming effect* e quais os mecanismos para tal?

Q6. As inferências obtidas por meio do uso do catálogo de genes são mantidas quando avaliado o contexto genômico?

Q7. Pode-se produzir um *software* que realize as análises básicas de metagenomas sem a necessidade de conhecimentos amplos na área de bioinformática?

Nesta tese as questões 1 até 6 foram abordadas dentro dos capítulos 3 e 4. Nos capítulos 5 e 6 são apresentadas ferramentas bioinformáticas que foram desenvolvidas pelo autor e que podem ser utilizadas no desenvolvimento de trabalhos similares, representando um recurso alternativo às ferramentas atuais ou *pipelines* caseiras.

Capítulo 2 - Revisão bibliográfica

2.1 O rio Amazonas

O rio Amazonas nasce nos Andes (Peru) e troca sete vezes de nome ao longo do seu curso. No Brasil, é chamado de rio Solimões até a sua confluência com o Rio Negro, a partir de onde passa a ser o rio Amazonas. O rio Amazonas e seus afluentes compreendem uma área equivalente a quase 38% da área continental da América do Sul (MIKHAILOV, 2010) e sua descarga representa 17 bilhões de toneladas por dia, o que representa aproximadamente 18% de toda a água fluvial do planeta que acaba no oceano (SUBRAMANIAM et al., 2008). Além disso, a floresta amazônica contribui com aproximadamente 10% da produção primária global, gerando cerca de 8.5 Pg de carbono por ano (FIELD et al., 1998; MALHI et al., 2008). O rio Amazonas é um sistema túrbido, rico em material orgânico particulado (SIOLI, 1984). Em torno de 60% da matéria orgânica (OM) presente nas águas do rio Amazonas são compostos derivados de celulose e lignina (ERTEL et al., 1986), bem como ácidos húmicos, que são respirados pelos microrganismos (principalmente micróbios heterotróficos) levando à liberação de CO_2 para a atmosfera (WARD et al., 2013, 2016). Isso representa uma liberação de quase 1.4 Tg de carbono por ano dos rios para a atmosfera (SAWAKUCHI et al., 2017).

As águas do rio Amazonas podem ser classificadas em três tipos, de acordo com Sioli (SIOLI, 1984): as águas brancas - ricas em nutrientes e material em suspensão, tendo pH próximo de neutro; águas pretas – pobres em nutrientes e menos túrbidas, tendo pH mais baixo devido a uma quantidade superior de ácidos húmicos e águas claras - pobres em nutrientes e partículas, tendo pH próximo a 7. Na Figura 2.1 é

possível observar na carta hidrográfica da bacia do rio Amazonas diferentes partes assinaladas com seus tipos de água.

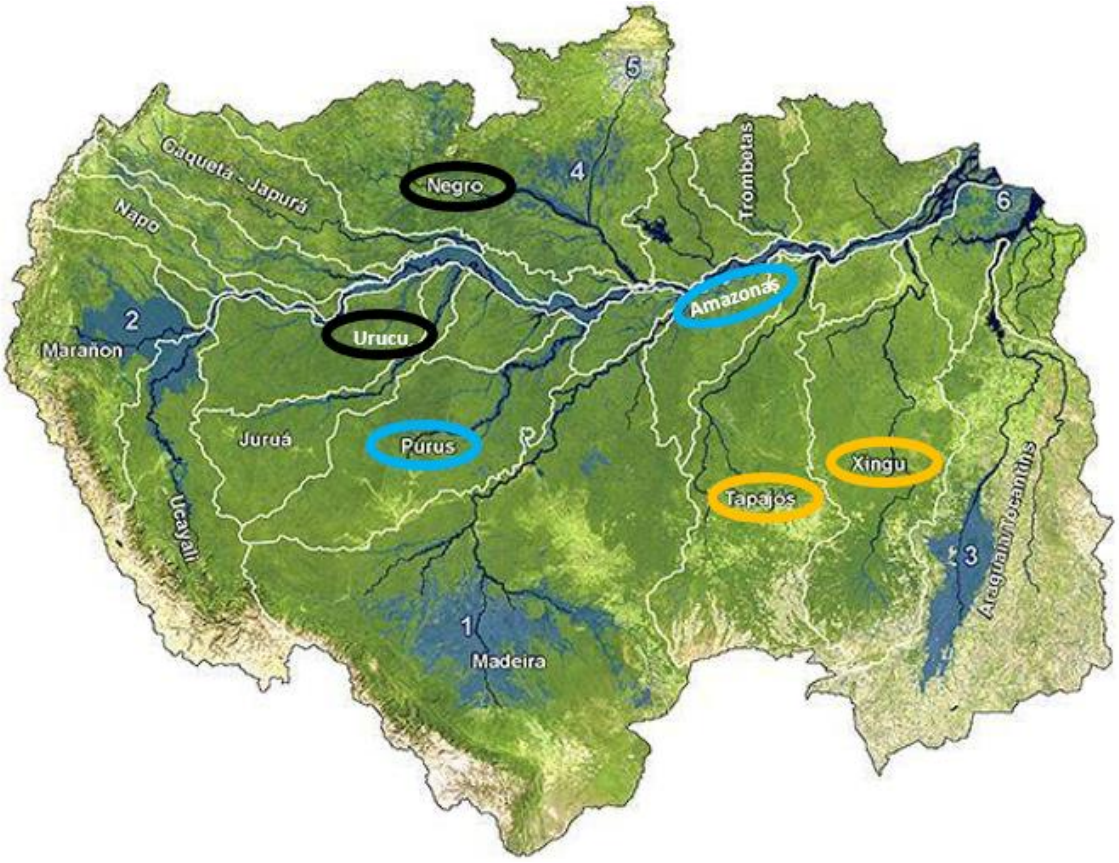

Figura 2.1. Carta hidrográfica da bacia do rio Amazonas evidenciando exemplos de afluentes com diferentes tipos de água. Os rios Urucu e Negro são exemplos de rios de água preta (elipses em preto), enquanto os rios Purus e Amazonas são exemplos de rios de água branca (elipses em azul claro). Por sua vez, os rios Tapajós e Xingu são exemplos de rios de água clara (elipses em amarelo). Fonte: Modificado de Castello et al. (CASTELLO et al., 2012).

Outro papel importante do rio Amazonas é a circulação de água dentro do próprio continente sul-americano. A teoria conhecida como "bomba biótica" postula que a condensação da água, devido à evapotranspiração da floresta amazônica que é mantida pelo rio Amazonas, diminui a pressão atmosférica que altera as correntes de ar úmido do mar para a terra (MAKARIEVA et al., 2013; MAKARIEVA; GORSHKOV, 2014). Esse processo gera uma espécie de "rio vertical" que alimenta as nuvens como uma bomba que suga a umidade do oceano Atlântico e do solo e a obriga a circular pelo continente da América do Sul. Este processo é capaz de gerar precipitações em regiões mais meridionais da América do Sul. A porção brasileira da floresta amazônica, até 2013, perdeu 763.000 Km^2 de sua área original. O desmatamento da floresta altera os padrões de pressão e pode causar o declínio dos ventos carregados de umidade que vêm do oceano para o continente (NOBRE, 2014).

Assim, a compreensão do ecossistema amazônico, altamente dependente do rio Amazonas, é essencial para a preservação de todo o continente sul-americano. A dinâmica hídrica do rio Amazonas é diferente daquela apresentada por rios de clima temperado, em que se acredita haver um *continuum* ecológico. No rio Amazonas, há o regime de pulsos de inundação. Isto significa que a partir da reunião das águas pluviais de toda a bacia de drenagem amazônica com as águas oriundas do degelo anual andino há um alagamento sazonal do rio Solimões, o que causa uma elevação do nível das águas todos os anos, formando uma planície de inundação nos meses de junho-julho (AYRES, 1995; JUNK, 1993; JUNK; BAYLEY; SPARKS, 1989; PIEDADE; JUNK; PAROLIN, 2000). Essa flutuação teria, segundo Junk et al. (JUNK; BAYLEY; SPARKS, 1989), um profundo efeito nas características limnológicas, ecológicas e biológicas desses corpos hídricos amazônicos. De modo que não é possível, portanto,

estabelecer claramente uma estrutura de comunidades ecológicas isoladas neste contexto (TOYAMA et al., 2017).

Recentemente, demonstrou-se que a variação nas quantidades de matéria orgânica (OM) durante os pulsos de inundação é o fator sazonal mais forte responsável pela variação do metabolismo bacteriano heterotrófico entre os corpos hídricos principais da Amazônia e os lagos de várzea, impulsionados principalmente pela qualidade do OM disponível (VIDAL et al., 2015). Assim, a influência do transporte lateral e da dinâmica temporal, principalmente motivados pelos pulsos de inundação, afetam grandemente as estruturas das comunidades bacterianas de lagos de várzea da Amazônia (MELO et al., 2018). Isto, reflete-se também de modo global, nos ciclos biogeoquímicos realizados por estas comunidades.

2.2 O fluxo de carbono e efeito *priming*

As águas continentais desempenham um papel biogeoquímico importante, ligando os ecossistemas terrestres ao oceano. Os sistemas fluviais são fundamentais neste processo, pois são altamente dinâmicos e biologicamente ativos, transformando carbono orgânico de origem terrestre e transportando-o para o mar (COLE et al., 2007).

Aproximadamente 1.9 Pg de carbono por ano chegam às águas continentais vindos do solo e apenas 50% desse carbono (0,9 Pg de carbono por ano) é entregue aos oceanos (COLE et al., 2007; MAYORGA et al., 2005). Isso sugere que a OM terrestre deve ser consumida ativamente. Outro fato que corrobora com a ideia de que haja altas taxas de degradação da OM terrestre é que apenas 30% do carbono orgânico terrestre

está armazenado nos sedimentos marinhos (BURDIGE, 2005). O tempo médio de residência do carbono orgânico dissolvido de origem terrestre em oceano aberto é menor que 100 anos (HERNES; BENNER, 2003), isso mostra que a remoção de OM terrestre acontece em questão de décadas.

A respiração e produção líquida dos ecossistemas tende a aumentar das cabeceiras dos rios até os seus estuários, assim como as taxas de respiração de carbono orgânico tendem a declinar, devido a uma maior produção primária (BATTIN et al., 2008). A degradação seletiva ao longo do tempo de residência nos sistemas estuarinos e fluviais pode deixar a OM mais resistente, o que acaba envelhecendo este material previamente antes de atingir os oceanos (RAYMOND; BAUER, 2001).

O efeito *priming* foi primeiramente definido a partir de experimentos com degradação de compostos em solo (BINGEMANN; VARNER; MARTIN, 1953), e foi definido como a mudança nos tempos de renovação da OM a partir de tratamentos moderados desse material (KUZYAKOV; FRIEDEL; STAHR, 2000). O cometabolismo, uma componente do efeito *priming*, foi definido, por sua vez, como a degradação facilitada dos compostos recalcitrantes a partir da adição de pequenas quantidades de OM lábil, resultando na liberação de compostos ricos em carbono e nitrogênio, quando comparados à condição sem a adição (HORVATH, 1972).

Um exemplo de efeito *priming* (Figura 2.2) é o caso em que a degradação da OM lábil por uma comunidade de decompositores, traz energia para os decompositores da OM recalcitrante e essa energia permite que essa outra comunidade de decompositores produza enzimas que degradam a OM recalcitrante. Isso, por sua vez,

libera nutrientes para os decompositores de OM recalcitrante e lábil, criando uma relação mutualística entre essas duas comunidades microbianas.

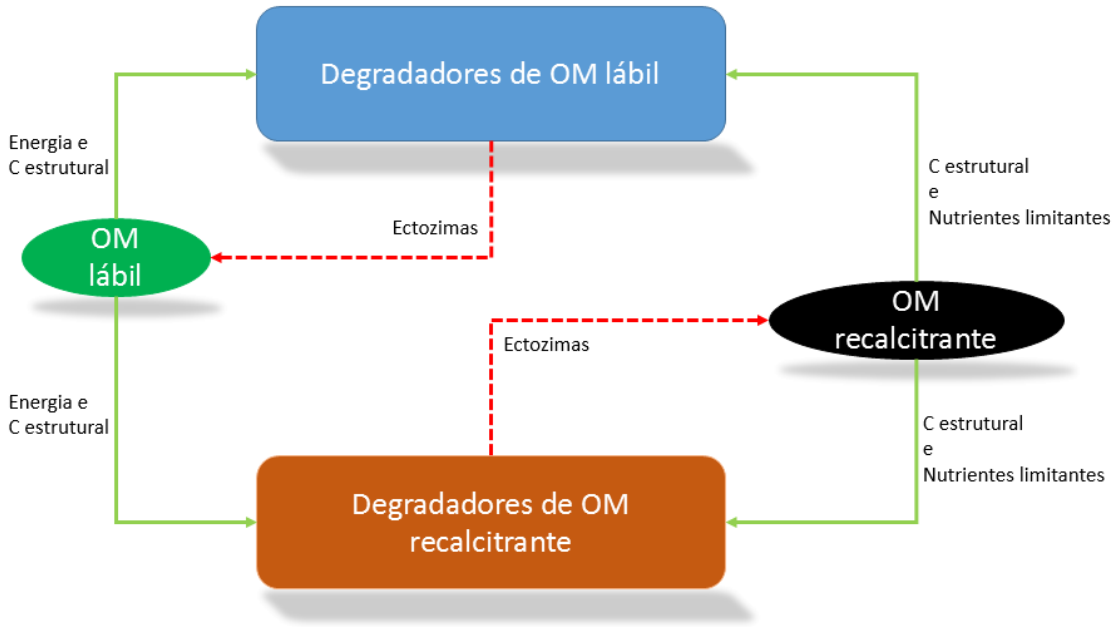

Figura 2.2. Exemplo de efeito *priming*. Neste modelo há 2 comunidades especializadas em diferentes tipos de OM, que interagem entre si por meio dos benefícios que seus produtos levam uma à outra, estabelecendo uma espécie de mutualismo. Fonte: Adaptado de Bianchi et al. (BIANCHI, 2011).

Apesar do mecanismo geral do efeito *priming* não ser compreendido, acredita-se que ele seja devido à combinação de fatores, tais como: (1) aumento da biomassa microbiana devido a consumo de substratos reativos, (2) promoção da produção microbiana de enzimas extracelulares capazes de quebrar OM menos reativa dirigida pela energia da quebra de OM mais reativa, (3) desequilíbrio estequiométrico entre

substratos e microrganismos, de modo que haja uma seleção de nutrientes e, por fim, (4) o consumo preferencial de substratos mais reativos (GUENET et al., 2010).

Em sistemas aquáticos, um exemplo disso é o ácido benzóico que tem um efeito de *priming* para a assimilação de ácidos fúlvicos pela *Arthrobacter* spp. em lagos (DE HAAN, 1977). Especula-se que o componente mais antigo e mais recalcitrante do *pool* de carbono orgânico dissolvido no oceano Pacífico Norte foi disponibilizado para assimilação e utilização bacteriana via cometabolismo e/ou fotoxidação (CHERRIER et al., 1999).

O sistema do rio Amazonas é basicamente heterotrófico e baseado na supersaturação com CO_2, em que a maioria deste carbono vem da respiração microbiana *in situ* de carbono orgânico (MAYORGA et al., 2005). O rio Amazonas tem taxas muito baixas de produção algal (WISSMAR et al., 1981), enquanto mais de 50% do carbono orgânico particulado no alto rio Mississipi é derivado de algas (KENDALL; STEVEN; KELLY, 2001). Esse fato mostra que a OM terrestre é abundante no ecossistema amazônico.

A degradação da OM terrestre no rio Amazonas é dificultada. Devido à grande turbidez das águas do rio Amazonas cerca de 0.05% das taxas de *outgassing* (a liberação de CO_2 das águas para a atmosfera) são derivadas de oxidação por ultravioleta *in situ* (REMINGTON; KRUSCHE; RICHEY, 2011). Cerca de 60% da lignina produzida na floresta é direcionada ao rio, onde ela é continuamente decomposta em monômeros, que são remineralizados diretamente para CO_2, ou ainda reduzidos a intermediários de baixo peso molecular (LMW). Menos de 5% da lignina que a floresta produz fica armazenada dentro da bacia amazônica ou é entregue ao oceano (WARD et

al., 2013). Esta taxa de degradação (~92%) é muito alta para um dos principais componentes da OM terrestre ainda considerado recalcitrante e mostra que a idade do substrato não interfere na sua biodisponibilidade (MAYORGA et al., 2005). De fato, a quebra da lignina mantém de 30 a 50% das taxas de respiração microbiana brutas no rio Amazonas (WARD et al., 2013).

Ward et al. (WARD et al., 2016) realizaram experimentos de incubação *in situ* e *ex situ* com matéria orgânica algal e monômeros de vanilina marcados com ^{13}C verificando as taxas de produção de CO_2, consumo de O_2 e a degradação de substratos vegetais (lignificada e do tipo macrofítico). Seus resultados indicaram que nas zonas de confluência entre rios do ecossistema amazônico há um aumento da concentração de CO_2, o que indica um potencial de aumento da degradação biológica da OM ao longo dessa zona. As descobertas destes autores reforçam a ideia de que zonas de confluência sejam *hot spots* da produção bacteriana e de que os níveis de CO_2 nessas regiões sejam mais elevados, como também visto por outros autores (FARJALLA, 2014). O efeito *priming* no rio Amazonas foi descrito como rápido, levando de minutos a horas, sendo mais comumente verificado em frações de microrganismos associados a partículas, onde há um maior efeito de enzimas extracelulares (WARD et al., 2016).

2.3 Degradação de matéria orgânica terrestre

A OM terrestre é composta principalmente por celulose e lignina, a primeira lábil e a última, considerada recalcitrante. A lignina é o segundo composto orgânico natural mais abundante da terra, sendo o primeiro a celulose (SCHLESINGER, 1977). Ela consiste em um polímero de fenilpropanóides que tem a função de conferir

integridade estrutural (VOELKER et al., 2011) e resistência a patógenos e ao colapso celular sob tensão associada ao transporte de água nas plantas (MIEDES et al., 2014). Poucas espécies de algas foram descritas conter essa substância (DELWICHE; GRAHAM; THOMSON, 1989; MARTONE et al., 2009). Dessa forma, a fonte principal de lignina nas frações de carbono orgânico particulado e dissolvido nos ecossistemas é de origem terrestre, e por sua dinâmica de degradação lenta, a lignina é considerada um material recalcitrante (BIANCHI, 2011). Assim, compreender quais os genes são responsáveis pelo decaimento da lignina é um conhecimento chave para a compreensão da ciclagem de carbono no contexto global.

Alguns efeitos ambientais podem reduzir a velocidade de degradação da OM, como a sorção de partículas ou a adsorção a partículas e de compostos, por exemplo, substâncias húmicas (ARNOSTI, 2011; MAYER et al., 2004). Por outro lado, os microrganismos podem organizar-se em agregados, o que facilita a degradação de OM, uma vez que se tornam *hot spots* de atividade enzimática (GROSSART et al., 2007; MOLDREUP et al., 2001).

Bactérias celulolíticas utilizam um arsenal de enzimas com atividades sinérgicas e complementares, tais como glicosil-hidrolases (GHs) para clivagens de ligações glicosídicas, polissacarídeo-esterases para suportar a ação de GHs sobre hemicelulose, e polissacarídeos liases para despolimerização de biopolímeros (PAYNE et al., 2015; VAN DEN BRINK; DE VRIES, 2011). Curiosamente, a dinâmica do ecossistema pode interferir na degradação da celulose bacteriana. Foi demonstrado que a predação pode resultar em agregação bacteriana, aumentando os níveis de metabolismo de quitina e celulose em água doce (CORNO et al., 2015). Alguns filos de água doce podem ser associados à degradação de celulose e hemicelulose: beta-proteobactérias,

principalmente o grupo das *Burkholderia* (HUTALLE-SCHMELZER et al., 2010; KONG et al., 2001; LIANG et al., 2014); Fibrobacteres (NEWTON et al., 2011) e Verrucomicrobia da família Opitutae (CABELLO-YEVES; GHAI; et al., 2017).

A lignina é um polímero heterogêneo de diversos fenilpropanóides, que é rico em ligações do tipo β-aril éter (β-O-4), podendo ser conjugado a diversos compostos, como o monolignol ferulato (KARLEN et al., 2016). A degradação de lignina tende a ocorrer em duas etapas: a despolimerização da lignina, realizada por enzimas extracelulares, e a mineralização de seus monômeros que ocorre intracelularmente (KAMIMURA et al., 2017). Os principais decompositores de lignina em ecossistemas terrestres são os fungos, por meio da secreção de oxidorredutases. Entretanto, sua atuação em ecossistemas aquáticos é bastante limitada. Bactérias tendem a contribuir com a despolimerização de lignina na natureza (LLADÓ; LÓPEZ-MONDÉJAR; BALDRIAN, 2017), principalmente por meio de peroxidases descolorantes ou *dye-decolorizing peroxidases* – DYPs (GONZALO et al., 2016); mas tendem numa maior extensão promover a mineralização dos compostos aromáticos heterogêneos de baixo peso molecular derivados de lignina (MASAI; KATAYAMA; FUKUDA, 2007; VICUÑA, 1988).

As bactérias oxidadoras de lignina foram identificadas principalmente como actinobactérias Gram-positivas, pertencentes aos gêneros *Rhodococcus* (AHMAD et al., 2011) e *Amycolatopsis* (BROWN; BARROS; CHANG, 2012); e também como bactérias Gram-negativas do grupo das γ-proteobactérias do gênero *Pseudomonas* (RAHMANPOUR; BUGG, 2015) que demonstrou utilizar-se de peroxidases dependentes de Mn(II). Diversos mecanismos de oxidação da lignina, além dos já mencionados, também puderam ser verificados em bactérias, como o uso de lacases

pelo gênero *Streptomyces* (MAJUMDAR et al., 2014) e β-eterases dependentes de glutationa no gênero *Sphingobium* (GALL et al., 2014).

Ensaios com lignina e celulose marcadas com ^{14}C revelaram que a principal degradação bacteriana da lignina ocorre em ambientes anóxicos na ausência de fungos (BENNER; MORAN; HODSON, 1986), o que sugere os sedimentos como a principal região de degradação. Entretanto, mais recentemente, verificou-se que a degradação da lignina tende a ocorrer em zonas oxigenadas, uma vez que utiliza mecanismos de oxidação que são dependentes de oxigênio e potencializados pelo mesmo (SANCHEZ, 2009).

A produção de peróxido de hidrogênio extracelular, um composto altamente reativo, é um primeiro passo para a oxidação de lignina mediada por enzimas como a lignina peroxidase e lacases dependentes de cobre (CRAGG et al., 2015). Elas produzem juntas um *pool* de compostos aromáticos heterogêneos, que forma uma fração húmica de carbono dissolvido, previamente detectada por outros estudos realizados no curso do rio Amazonas (ERTEL et al., 1986; SEIDEL et al., 2016). A lignina ainda tem um destino diferente de outros biopolímeros durante a sua decomposição na lâmina d'água, uma vez que sua degradação é controlada pela disponibilidade de outras fontes de carbono facilmente decomponíveis (KLOTZBÜCHER et al., 2011).

Os compostos aromáticos de baixo peso molecular, gerados a partir da oxidação da lignina, tendem a apresentar moderada a elevada toxicidade para os microrganismos que não são capazes de cataboliza-los. Dentre os efeitos de inibição, temos: inibição enzimática, que atua sobre determinadas classes de enzimas, como as celulases (QIN, Lei et al., 2016); inibição fermentativa, que atua sobre a formação de produtos da

fermentação e da dinâmica de crescimento celular anaeróbio (MONLAU et al., 2014; XUE et al., 2018); e a inibição de crescimento de culturas aeróbias puras (ASTON et al., 2016). O catabolismo dos compostos gerados pela oxidação de lignina ocorre por meio de uma via complexa, que pode apresentar-se em sua totalidade ou apenas parcialmente em algumas espécies (Figura 2.3).

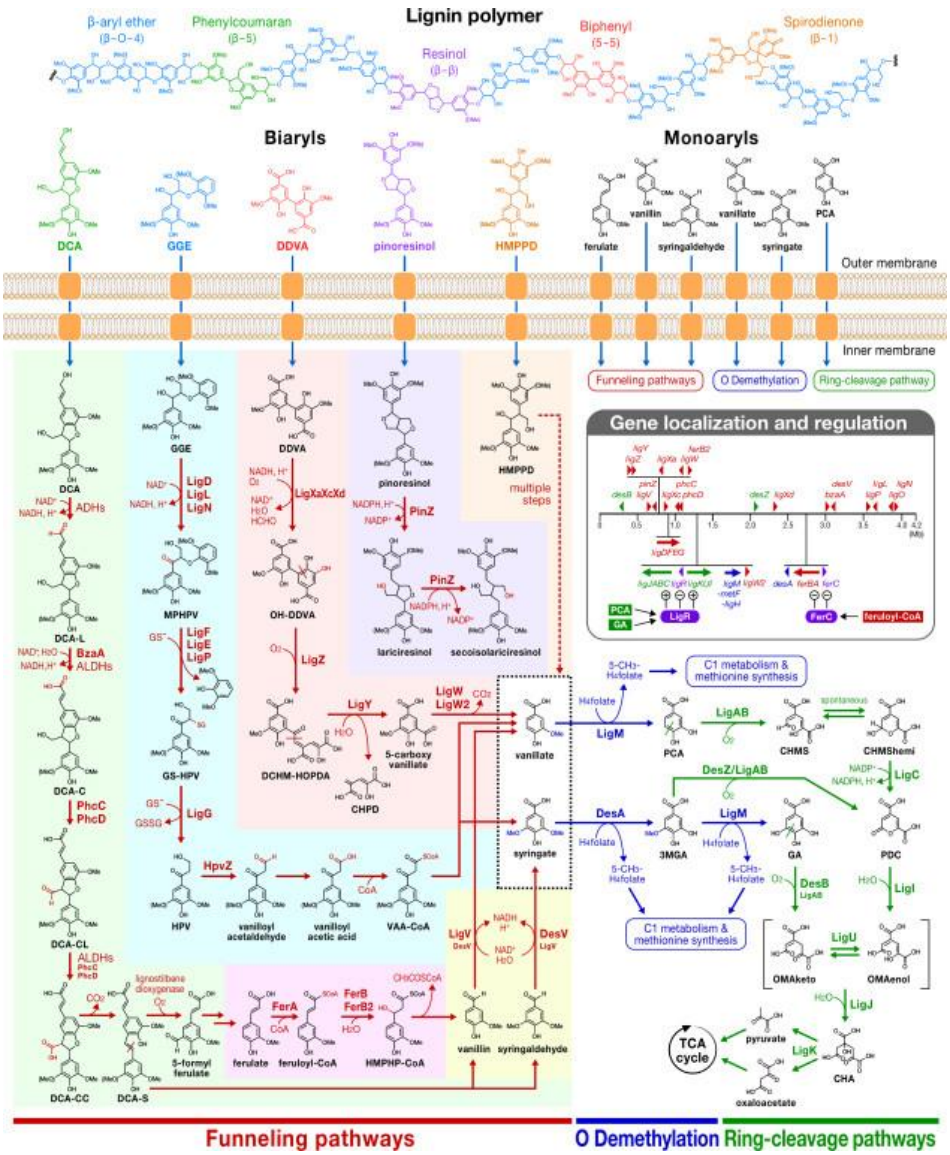

Figura 2.3. Vias do catabolismo de compostos aromáticos derivados de lignina. As vias do microrganismo modelo *Sphingobium* SYK-6 estão mostradas. Diferentes

compostos di- e mono-arílicos, derivados de lignina, são canalizados para as vias do vanilato e siringato. O vanilato e siringato são O-desmetilados produzindo, assim, compostos aromáticos (catecóis, protocatecuato e galato), que são degradados por meio das vias de clivagem de anéis, terminando no ciclo do ácido tricarboxílico. Fonte: Extraído de Kamimura et al. (KAMIMURA et al., 2017).

Resumidamente, a captação dos substratos através da membrana interna bacteriana acontece via transportadores ativos, que podem ser de quatro tipos diferentes: transportadores do cassete de ligação de ATP (ABC), transportadores da superfamília facilitadora (MFS), sistema tripartido periplasmático independente de ATP (TRAP) e um membro da superfamília dos transportadores iônicos (IT). Uma vez dentro da célula, o catabolismo se dá a partir de três passos principais: (1) os compostos arílicos gerados pela oxidação da lignina e transportados para o interior da célula são canalizados a uma das vias de degração de bi- ou monoaris para a formação de vanilato ou siringato; (2) o vanilato e siringato são O-desmetilados; (3) os compostos já desmetilados, contendo anéis catecólicos, tem suas estruturas de ressonância rompidas até a formação de compostos mais simples, como piruvato e oxaloacetato, que são direcionados para o ciclo do TCA.

2.4 Metagenômica da Amazônia

Uma vez que cerca de 99% dos microrganismos presentes em ambientes naturais não podem ser cultivados em laboratório, comunidades bacterianas apenas

podem ser analisadas, em sua totalidade, via técnicas independentes de cultivo, como microscopia ou a metagenômica (GARZA; DUTILH, 2015; STREIT; SCHMIDTZ, 2004).

A metagenômica é uma técnica que permite o estudo de uma população microbiana de uma única vez, por meio do sequenciamento maciço de DNA ambiental, obtendo não apenas informação de caráter taxonômico, mas também, funcional (SHARPTON, 2014). Essa abordagem é ideal para obter sequências de milhares de genomas de uma única vez.

A composição microbiológica de lagos provenientes de quatro bacias hídricas (rio Amazonas, rio Araguaia, rio Paraná e Pantanal), obtida de bibliotecas de *amplicons* do gene 16S rRNA com sequenciamento 454, mostrou-se similar a de outros sistemas de água doce ao redor do mundo, se considerados níveis taxonômicos mais altos. Entretanto, os sistemas brasileiros apresentaram divergências específicas quanto a sua composição ao nível de família, que foi correlacionada com fatores abióticos, como o pH (TESSLER et al., 2017).

O primeiro estudo da microbiota aquática amazônica foi realizado em 2011, pelo grupo de pesquisa coordenado pelo Prof. Dr. Flávio Henrique-Silva. Este estudo analisou um ponto do Rio Solimões a 400 km à montante de Manaus, gerando os primeiros dados de metagenômica dos micro-organismos da coluna d'água deste rio (GHAI; RODRIGUEZ-VALERA; et al., 2011).

Recentemente, vários estudos investigaram zonas específicas da bacia amazônica por meio de tecnologias ômicas e agora vamos discutir um pouco dos achados mais recentes utilizando-se esta técnica.

A parte superior da bacia Amazônica revelou-se dominada por Proteobactérias (principalmente Betaproteobactérias), Cianobactérias, Actinobactérias, Planctomicetos, Bacteroidetes e Firmicutes (SANTOS-JÚNIOR et al., 2017; TOYAMA, 2016; TOYAMA et al., 2016). Vários metagenomas de estações de amostragem da bacia do baixo Amazonas apresentaram maiores contribuições de cianobactérias e outros táxons diazotróficos (SATINSKY et al., 2015). Na pluma do rio Amazonas sobre o Oceano Atlântico, há evidências indicando a substituição de espécies de água doce por espécies correspondentes de água salobra (DOHERTY et al., 2017; SATINSKY et al., 2015; SATINSKY; ZIELINSKI; et al., 2014). Poucos táxons de arqueia foram descritos para o rio Amazonas, no entanto, as Thaumarchaeota foram observadas na bacia do alto rio Amazonas (SANTOS-JÚNIOR et al., 2017), enquanto o grupo das Euryarchaeota parece dominar a pluma (SATINSKY et al., 2015).

Uma vez que a maior parte da matéria orgânica presente no rio é material derivado de plantas (ERTEL et al., 1986; GAGNE-MAYNARD et al., 2017; SEIDEL et al., 2016), como a lignina e a celulose, espera-se que as funções gênicas associadas à sua degradação sejam abundantes no rio. Em vez disso, estudos metagenômicos anteriores (SATINSKY; CRUMP; et al., 2014; SATINSKY; SMITH; SHARMA; LANDA; et al., 2017; SATINSKY; SMITH; SHARMA; WARD; et al., 2017) detectaram tais funções em quantidades muito baixas no baixo rio Amazonas, levantando a hipótese de que isso ocorreu devido à má compreensão das enzimas envolvidas na degradação de lignina e celulose em sistemas aquáticos, bem como a natureza heterogênea dessas moléculas.

Estudos investigando conexões entre estilos de vida microbiano na coluna de água e degradação de matéria orgânica derivada de plantas revelaram que micróbios

associados a partículas apresentaram mais funções relacionadas à degradação de carbono do que micróbios de vida livre (SATINSKY; CRUMP; et al., 2014; SATINSKY et al., 2015). Até o momento, o catabolismo dos compostos derivados de lignina, bem como seu transporte não foram ainda avaliados no ecossistema aquático amazônico.

2.5 Catálogos de genes

Inicialmente, a utilização da anotação de *reads* dominava estudos metagenômicos. Uma vez que as funções gênicas atribuídas daquela forma podem não refletir de modo satisfatório o conteúdo gênico da amostra, criou-se os catálogos gênicos. Estes bancos de dados são obtidos através da montagem das *reads* e predição de genes nos *contigs* resultantes, com posterior eliminação da redundância desses genes obtidos. Assim, os bancos de dados, também conhecidos como catálogos gênicos, podem ser utilizados para anotação com diferentes abordagens e referências, gerando um perfil metabólico mais próximo da realidade da amostra.

Os catálogos de genes fornecem uma fonte única de informações baseadas nos genes e nas funções de proteínas preditas. Ao contrário dos metagenomas, os catálogos de genes são coleções de genes preditos obtidos a partir de *reads* montadas. As inferências obtidas dos catálogos de genes têm muito mais confiabilidade do que aquelas retiradas da atribuição funcional de *reads*.

Eles têm a vantagem de amostrar uma abundância funcional mais realista por metagenoma, além de possibilitar a bioprospecção e auxiliar no desenvolvimento de teorias utilizadas em testes experimentais de características ambientais funcionais.

Alguns catálogos de genes já foram publicados anteriormente, como por exemplo dos oceanos (CARRADEC et al., 2018; MENDE et al., 2017; SUNAGAWA et al., 2015), de solo (BAHRAM et al., 2018) e intestinos de animais (PAN et al., 2018; QIN, Junjie et al., 2010). Estes catálogos tem auxiliado os cientistas a descobrir diversos aspectos importantes das comunidades microbianas que vivem naqueles ambientes.

Um exemplo importante de inferência ecológica foi aquela obtida do catálogo de genes do solo global (BAHRAM et al., 2018), em que foi possível inferir que a diversidade genética bacteriana, mas não a fúngica, é mais alta em habitats temperados e que a composição gênica microbiana varia mais fortemente de acordo com as variáveis ambientais (precipitação e pH do solo) do que com a distância geográfica. Este catálogo, também forneceu evidências de um provável antagonismo bacteriano-fúngico em habitats de solo e oceanos.

Apesar de toda a cobertura metagenômica do rio Amazonas, da extensão amostrada e da forma como os experimentos se sobrepuseram, estes dados ainda permanecem pouco trabalhados. As *reads* obtidas dos projetos de sequenciamento já realizados não foram sequer montadas e analisadas em nível de *contig*. Isto, aliado ao fato de não haver atualmente um catálogo de genes de ambientes de água doce, reforça a necessidade de um catálogo do rio Amazonas.

2.6 Genomas populacionais ou *Metagenome Assembled Genomes*

A estruturação dos dados de sequenciamento de alto rendimento em genomas pode representar um passo crucial para contextualizar o potencial metabólico observado com outras abordagens. Assim, um catálogo de genes seria importante para inferir evidências de relações ecológicas e metabólicas, no entanto, há a necessidade do contexto genômico para se avaliar o potencial metabólico dos organismos de uma determinada população oriundos de ambientes específicos, bem como seu potencial regulador e outras possíveis características. Entretanto, os genomas obtidos representarão as populações microbianas mais abundantes no ambiente, uma vez que contribuem em maior extensão para o DNA ambiental, consequentemente gerando um maior grau de redundância nos conjuntos de dados finais. Esses genomas populacionais (PGs) representam uma população de organismos, mas não seus indivíduos.

Os PGs podem nos ajudar a desvendar as principais adaptações metabólicas dessas populações microbianas a diferentes ambientes (GEORGES et al., 2014; MULLER et al., 2014). Assim, também auxiliam na elucidação de seus arranjos genéticos e previsão dos papéis ecológicos desempenhados por microrganismos não cultiváveis ou, ainda, na verificação de adaptações de organismos conhecidos nesses ambientes.

Também conhecidos como *metagenome-assembled genomes* (MAGs), os PGs tem sido recuperados de diversos ambientes (BAKER et al., 2015; CABELLO-YEVES et al., 2018; DICK et al., 2009; GHAI; PAŠIĆ; et al., 2011; HUGERTH et al., 2015; PARKS et al., 2017; TULLY et al., 2017). Dentre as técnicas para sua recuperação destacam-se os sistemas híbridos de categorização baseados em frequência de k-mers e

abundância diferencial de *contigs*, assim como aquele empregado no *software* Metabat (KANG et al., 2015). Os *contigs* são agrupados de acordo com as características acima descritas em grupos de sequências denominados *bins*, que são posteriormente refinados eliminando-se quimeras e *contigs* de origem exógena ao PG em questão.

Descobertas importantes foram obtidas a partir desta estratégia em diferentes ambientes, como novos representantes de clados bacterianos (PARKS et al., 2017) e mecanismos de micróbios de água doce na captação de luz e performance dos ciclos biogeoquímicos (CABELLO-YEVES et al., 2018; CABELLO-YEVES; GHAI; et al., 2017; CABELLO-YEVES; HARO-MORENO; et al., 2017). No entanto, pouco foi feito, nesse sentido, em ecossistemas tropicais de água doce, principalmente na bacia do rio Amazonas. Uma análise do virioplâncton de água doce da Amazônia foi recentemente realizada (SILVA et al., 2017), isolando-se PGs de bacteriófagos onipresentes no *continuum* do rio sobre o oceano Atlântico. O impacto destes vírus no balanço de carbono mostrou-se importante para facilitar a liberação de matéria orgânica vegetal (por exemplo, lignina e celulose) e diminuir a quantidade total de produção primária por hospedeiros fotossintéticos.

Capítulo 3 - Evidências moleculares de *priming effect* no rio Amazonas

3.1 Introdução e objetivos

Apesar da relevância amplamente reconhecida do processamento heterotrófico microbiano de carbono terrestre na bacia do rio Amazonas (BATTIN et al., 2008; BENNER et al., 1995), pouco se sabe sobre o conjunto de genes e processos bioquímicos utilizados por eles para realizar tal tarefa. Em 2013, verificou-se que somente cerca de 5% da lignina que entrava no rio Amazonas chegava aos oceanos (WARD et al., 2013), entretanto, genes com funções relacionadas à degradação de lignina e celulose são encontradas em baixas quantidades na porção inferior do rio Amazonas, entre Manaus e seu estuário (SATINSKY; CRUMP; et al., 2014; SATINSKY; SMITH; SHARMA; LANDA; et al., 2017; SATINSKY; SMITH; SHARMA; WARD; et al., 2017). Recentemente, a degradação de lignina no rio Amazonas foi verificada como sendo potencializada pela degradação de compostos mais lábeis (WARD et al., 2016), o chamado efeito *priming.* Pouco se sabe sobre este efeito, no que tange seus mecanismos moleculares ou agentes microbianos, tornando-se uma questão chave para a compreensão da ciclagem de carbono global.

Neste capítulo, apresentamos o primeiro catálogo de genes do maior rio do mundo, por meio da análise de 106 metagenomas (mais de 500 Gpb), originados de 30 estações, cobrindo um total de aproximadamente 2106 km, do alto rio Solimões até a pluma do rio Amazonas no oceano Atlântico. Este catálogo foi utilizado para a avaliação dos genes responsáveis pela degradação de OM terrestre na bacia do rio Amazonas, permitindo a geração de um modelo do efeito *priming*, previsto por outros autores (WARD et al., 2016). Portanto, abordamos aqui as seguintes questões principais:

- A diversidade genética presente na bacia do rio Amazonas reflete processos evolutivos locais ou é semelhante a outros ecossistemas de água doce?

- Existe uma estrutura espacial na ocorrência de genes ao longo do curso do rio que sugira um zoneamento das populações quanto ao seu metabolismo?

- Quais são as principais funções bioquímicas associadas à degradação da OM terrestre?

- Existem funções bioquímicas específicas que promovem o uso de fontes alternativas de carbono na bacia do rio Amazonas?

Assim, hipotetizamos que a grande quantidade de OM complexa presente no rio Amazonas possa ter sido crucial na seleção de linhagens com novos genes capazes de degradá-la. Além disso, esperamos que as funções associadas ao ciclo de carbono e nitrogênio sejam predominantes em zonas de confluência, como a dos rios Negro e Solimões, onde águas negras e brancas se misturam (FARJALLA, 2014; LARAQUE; GUYOT; FILIZOLA, 2009), e a região da pluma do rio Amazonas no oceano Atlântico. Nessas regiões também foi hipotetizado por outros autores (BIANCHI, 2011), que provavelmente devido à transição de comunidades microbianas haveria um maior potencial para que o efeito *priming* ocorra. Nessas zonas, grandes quantidades de nutrientes e luz criam condições ideais para que haja uma alta produção primária, promovendo o efeito *priming* (HILTON et al., 2015; SATINSKY et al., 2015; SATINSKY; SMITH; SHARMA; WARD; et al., 2017). Também levantamos a hipótese de que exista uma extensa semelhança funcional entre os microbiomas do rio Amazonas e oceano, dado que múltiplas funções metabólicas centrais (ou seja, aquelas relacionadas ao metabolismo microbiano basal) seriam necessárias tanto para os

micróbios de água doce quanto de água salgada. Por outro lado, não esperamos que a similaridade genética entre os microbiomas do rio e do oceano seja alta, já que os táxons dominantes nesses ambientes mudam (LOGARES et al., 2009). Além disso, os processos evolutivos experimentados pela microbiota do rio Amazonas, ao longo de milênios sendo exposta a material advindo da floresta tropical, provavelmente moldaram localmente a diversidade genética. Essa diversidade gênica deveria diferenciar o microbioma do rio Amazonas de outros ambientes (MEYER et al., 2017; SATINSKY et al., 2015; STALEY et al., 2014a; TOYAMA et al., 2017; VAN ROSSUM et al., 2015).

3.2 Material e métodos

Foram analisados 106 metagenomas gerados em diferentes trabalhos (SANTOS-JÚNIOR et al., 2017; SATINSKY; ZIELINSKI; et al., 2014; TOYAMA, 2016; TOYAMA et al., 2016) provenientes de 30 estações distribuídas ao longo do curso da bacia Amazônica, com cobertura média de 5.0 Gpb por metagenoma (desvio padrão de 7.29 Gpb). As estações do rio Solimões e lagos no curso do rio Amazonas à montante de Manaus até a pluma do rio Amazonas no oceano Atlântico cobriram ao todo mais de 2100 km, e foram divididas em 5 seções (Figura 3.1 e Apêndice 1).

Estas seções foram:

1) Montante - Seção a montante, incluindo as porções da bacia amazônica à montante da cidade de Manaus;
2) Jusante - Trecho a jusante, situado entre a cidade de Manaus até o início do estuário do rio Amazonas, compreendendo a mistura das águas brancas ricas em partículas do rio Solimões e das águas ricas em compostos húmicos do rio Negro (FARJALLA, 2014; LARAQUE; GUYOT; FILIZOLA, 2009);
3) Estuário - Seção de estuário, incluindo a porção do rio que termina no oceano Atlântico. Nesta seção há uma redução na velocidade das águas, predomínio de forças deposicionais e uma substituição das fontes de carbono por produção primária (SATINSKY et al., 2015);
4) Pluma – A pluma que cobre uma enorme área onde o oceano é influenciado pela água do rio Amazonas e possui salinidade alterada (SATINSKY; ZIELINSKI; et al., 2014);
5) Oceano - Estações na borda da pluma, onde há um teor salino similar às águas do Atlântico.

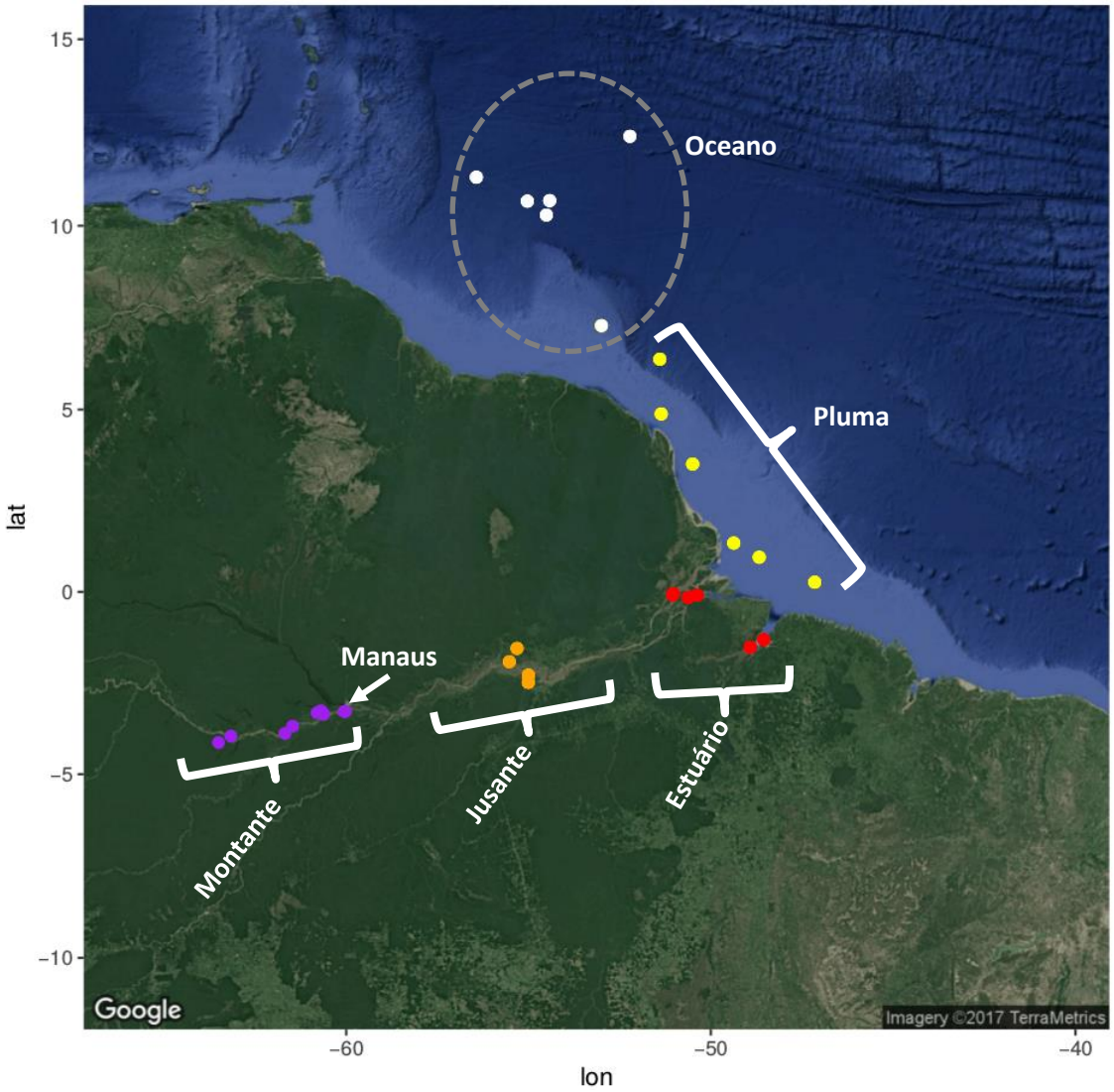

Figura 3.1. Estações de amostragem no rio Amazonas e suas seções. Os 106 metagenomas utilizados neste trabalho (estações representadas como pontos amarelos, alaranjados, brancos e vermelhos são aquelas pertencentes aos estudos de Satinski et al. (2014, 2015, 2017), enquanto que as estações representadas como pontos roxos são aquelas pertencentes aos estudos de Santos-Júnior (2017), Toyama (2016) e Toyama et al. (2016)) e sua distribuição ao longo dos cinco trechos do rio Amazonas são mostrados. (Fonte: Próprio autor)

O DNA ambiental foi extraído dos filtros utilizados para separar as populações microbianas de vida livre e associadas a partículas de acordo com a malha de filtragem. Dependendo do estudo original, as bactérias associadas a partículas foram definidas como aquelas células retidas no intervalo de filtragem de 297 µm a 2 ou 5 µm, e micróbios de vida livre foram definidos como aquelas células retidas no intervalo de 5 ou 2 µm a 0,2 µm (Figura 3.2). Os metagenomas foram obtidos a partir de bibliotecas preparadas com tecnologia Illumina, utilizando os kits Nextera XT ou TruSeq. Foram utilizadas diferentes plataformas de sequenciamento: Illumina Genome Analyzer IIx, Illumina HiSeq 2500 ou Illumina MiSeq. Mais informações sobre as condições específicas de cada metagenoma utilizado neste estudo estão disponíveis no Apêndice 1.

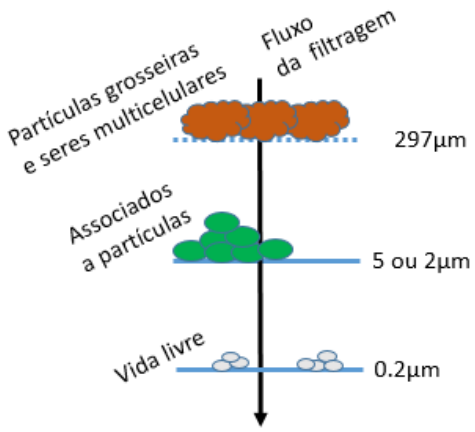

Figura 3.2. Esquema de filtragem utilizado na preparação dos metagenomas utilizados neste estudo. A água amostrada passou por um filtro maior, eliminando assim partículas e organismos maiores. Logo em seguida, passou por um filtro intermediário que coleta os microrganismos associados a partículas e por fim, passa por um filtro menor que coleta as células livres. (Fonte: Próprio autor)

3.2.1 Processamento dos metagenomas

Os metagenomas foram obtidos do banco de dados *Sequence Reads Archive* (SRA) do *National Center for Biotechnology Information* (NCBI). As sequências dos adaptadores Illumina e as bases de baixa qualidade foram removidos dos metagenomas usando o programa Cutadapt (MARTIN, 2011). Apenas *reads* maiores que 80 pb e contendo uma qualidade média $Q \geq 24$ foram mantidas ao fim do processo. A qualidade das *reads* foi verificada com o programa FASTQC *tool* (ANDREWS, 2017). As *reads* dos metagenomas pertencentes a uma mesma estação de amostragem foram montadas conjuntamente utilizando-se o programa MEGAHIT v1.0 (LI, Dinghua et al., 2016) com os parâmetros estabelecidos para metagenomas grandes e complexos (meta-large) e tamanho mínimo dos *contigs* após montagem de 1 Kpb, como recomendado por Vollmers et al. (VOLLMERS; WIEGAND; KASTER, 2017). A qualidade das montagens foi avaliada com auxílio do programa QUAST (GUREVICH et al., 2013) e MetaQUAST (MIKHEENKO; SAVELIEV; GUREVICH, 2016).

3.2.2 Análise da diversidade de k-mers em diferentes ambientes

Os k-mers são *motifs* com comprimento “k” observados mais de uma vez em uma sequência genômica. Assim o cálculo de distâncias ecológicas padrão comparativas de diversos conjuntos de dados, substituindo-se as contagens de espécies por contagens de k-mers, pode superar as limitações de técnicas baseadas na atribuição taxonômica/funcional (dependentes de subconjuntos de sequências anotadas) e em métodos *de novo* (que comparam os conjuntos inteiros de sequências). Assim, realizou-

se um teste de diversidade k-mers envolvendo os metagenomas provenientes da bacia do rio Amazonas e metagenomas provenientes de rios temperados e do solo da floresta amazônica (Apêndice 2) com o fim de se confirmar se a diversidade gênica inferida a partir de amostras do rio Amazonas reflete um processo de evolução específico local.

Os metagenomas da bacia do rio Amazonas (106) foram comparados com 37 metagenomas do rio Mississippi (STALEY et al., 2014a) e 91 metagenomas provenientes de três bacias hidrográficas distantes até 130 km no sudoeste da Colúmbia Britânica - Canadá (VAN ROSSUM et al., 2015), ambos representantes de rios de clima temperado, e 11 metagenomas oriundos do solo da floresta Amazônica (MEYER et al., 2017). Todos os metagenomas foram processados como descrito no item 3.2.1, quanto à eliminação de sequências adaptadoras e *reads* de baixa qualidade e comprimento menor que 80 pb. Após seu processamento, os metagenomas foram submetidos ao programa SIMKA versão 1.4 (BENOIT et al., 2016) com k = 21, tamanho mínimo de *read* de 80 pb, eliminando-se *reads* de baixa complexidade e k-mers com índice de Shannon < 1.5 (isso reduz o efeito de ruído devido à sequências repetitivas), seguido de normalização pelo tamanho da amostra. A distância de Jaccard obtida a partir da matriz de presença-ausência de k-mers foi usada para gerar um escalonamento multidimensional não métrico (NMDS) com análise posterior da variância da beta-distribuição e análise de similaridade entre grupos implementada no pacote Vegan (DIXON, 2003) em linguagem R.

3.2.3 Catálogo de genes microbianos não-redundantes da bacia amazônica (AMnrGC – Amazon river basin microbial non-redundant genes catalogue*)*

Os genes foram preditos a partir dos *contigs* obtidos pela montagem descrita no item 3.2.1 por meio do uso do programa Prodigal versão 2.6.3 (HYATT et al., 2010). Apenas as fases aberta de leitura (ORFs) preditas como completas, aceitando-se o uso de códons de iniciação alternativos (GUG e UUG), e maiores que 150 pb foram mantidas. Neste processo obteve-se mais de 6 milhões de genes redundantes que foram agrupados em *clusters* nos quais apenas um gene representativo foi mantido. As sequências foram agrupadas utilizando-se o programa CD-HIT-EST versão 4.6 (FU et al., 2012; LI, W.; GODZIK, 2006) com uma identidade de nucleotídeos igual ou superior a 95% e no mínimo 90% de sobreposição com o menor gene, assim como previamente descrito por outros autores (MENDE et al., 2017). Todas as análises posteriores utilizaram as sequências representativas. O conteúdo de GC por gene foi inferido utilizando o *software* Infoseq, implementado no pacote EMBOSS versão 6.6.0.0 (RICE; LONGDEN; BLEASBY, 2000).

3.2.4 Estimativas de abundância gênicas

As *reads* com qualidade mínima foram mapeadas contra o catálogo de genes não redundantes utilizando-se os programas BWA versão 0.7.12-r1039 (LI, H.; DURBIN, 2009), SamTools versão 1.3.1 (LI, H. et al., 2009) e BEDTools versão 2.26.0 (QUINLAN; HALL, 2010). As abundâncias gênicas foram estimadas por meio do programa eXpress versão 1.5.1 (ROBERTS; PACHTER, 2012) desconsiderando-se a

correção de *bias*, e foram dadas como o equivalente a transcritos por milhão de *reads* (TPM). Os valores de TPM são calculados similarmente aos índices RPKM (*Reads Per Kilobase of gene per Million reads*) ou FPKM (*Fragments Per Kilobase of gene per Million reads*), entretanto, a normalização quanto ao comprimento do gene acontece primeiro no cálculo do TPM e somente então normaliza-se pela profundidade de sequenciamento. Isto implica que os valores de TPM, quando somados para uma mesma amostra sempre tem o mesmo valor. Isso facilita a comparação da proporção de *reads* mapeadas para um gene em cada amostra.

Utilizou-se valores de TPM ≥ 1.00 para que um gene fosse reconhecido como presente em uma amostra, e uma abundância média maior que zero ($\mu_{TPM} > 0.0$) para que fosse reconhecido como presente em uma seção do rio ou no tipo de água (água doce, água salina e zona de mistura, também dita como pluma).

3.2.5 Anotação funcional

Os genes representativos foram anotados por meio de busca de ortólogos nos seguintes bancos de dados de referência: Enciclopédia de genes e genomas de Kyoto - KEGG versão 2015-10-12 (KANEHISA et al., 2012); Conglomerados de grupos de proteínas ortólogas - COG versão 2014 (TATUSOV et al., 2003); CAMERA *Prokaryotic Proteins Database* versão 2014 (SUN et al., 2011); UniProtKB versão 2016-08 (UNIPROT CONSORTIUM, 2015). A busca de ortólogos se deu por meio do algoritmo Blastp implementado no programa Diamond v.0.9.22 (BUCHFINK; XIE; HUSON, 2014), utilizando-se uma cobertura ≥ 50% , identidade de aminoácidos ≥ 45% e e-*value* ≤ 1e-5, com um *Score* ≥ 50. O mapeamento das vias bioquímicas foi realizado

por meio dos códigos KO (implementados como identificadores do banco de dados KEGG) foi feita por meio da ferramenta KEGG *mapper* (KANEHISA et al., 2017).

Algumas vezes a anotação dependente de ortólogos falha no sentido da necessidade direta de identidade entre as sequências envolvidas na busca. Uma busca direcionada por probabilidade de estados estatísticos foi desenvolvida para suprir esse problema e superar o desafio do tempo envolvido no processo, que era demasiado longo. Utilizando-se perfis de cadeias ocultas de Markov, obtidos a partir de alinhamentos de proteínas pertencentes à mesma família, pode-se criar um método eficiente de anotação de domínios sem a necessidade de associação a um ortólogo específico. A anotação dos genes utilizando-se perfis de cadeias ocultas de Markov de famílias protéicas (HMM) foi feito com o programa HMMSearch versão 3.1b1 (EDDY, 2011) contra os bancos de dados: dbCAN versão 5 (YIN et al., 2012), PFAM versão 30 (FINN et al., 2016) e eggNOG versão 4.5 (HUERTA-CEPAS et al., 2016). Os resultados foram filtrados utilizando-se e-*value* $\leq$ 1e-5 e probabilidade posterior ao alinhamento dos resíduos ≥ 0.9, desconsiderando-se a sobreposição de domínios.

Por fim, produziu-se duas curvas de acumulação, uma contabilizando os genes obtidos por amostra e outra contabilizando o número de famílias de proteínas preditas anotadas com o banco de dados PFAM por amostra. Ambas as curvas foram obtidas por meio de uma análise acumulativa com progressão randômica de comparação pareada utilizando-se 100 pseudo-replicatas.

3.2.6 Atribuição de taxonomia aos genes

A taxonomia dos genes foi anotada utilizando-se os melhores resultados em termos de: *Score*, e-*value* e identidade. Para isso as anotações obtidas com as bases de dados referência KEGG versão 2015-10-12 (KANEHISA et al., 2012), UniProtKB versão 2016-08 (UNIPROT CONSORTIUM, 2015) e CAMERA *Prokaryotic Proteins Database* versão 2014 (SUN et al., 2011) foram cruzadas e o melhor ortólogo foi selecionado. O último ancestral comum (LCA) foi obtido a partir dos números de identificação taxonômica (TaxID) do NCBI associados às entradas disponíveis nos bancos de dados de referência UniRef100 e KEGG. Informações da base de dados CAMERA também foram usadas para recuperar a taxonomia (TaxID). As proteínas foram anotadas como "sem assinatura" se os registros continham representantes de vários domínios da vida ou se tivessem apenas a função atribuída sem informações taxonômicas. Sequências de referência previamente detectadas em outros estudos metagenômicos, permanecendo pouco caracterizadas e sem espécies associadas, foram utilizadas para anotar o grupo aqui denominado "Metagenômico".

3.2.7 Maquinaria bioquímica para a degradação de matéria orgânica (OM) terrestre

Para investigar a degradação da OM, agrupamos as amostras por seção do rio. Uma relação completa de sequências de referência e famílias de proteínas usadas nesta busca está disponível no Apêndice 3.

A degradação da lignina inicia-se com a oxidação do polímero carreada extracelularmente, seguida pela internalização e metabolismo de monômeros ou

dímeros heterogêneos produzidos por este processo. As famílias de proteínas relacionadas à oxidação da lignina (PF05870, PF07250, PF11895, PF04261 e PF02578) foram buscadas na anotação dos genes do catálogo com o banco de dados PFAM. Os genes relacionados ao metabolismo de compostos aromáticos derivados de lignina foram anotados via algoritmo de busca Blastp implementado no programa Diamond versão 0.9.22 (BUCHFINK; XIE; HUSON, 2014), com cobertura ≥ 50%, identidade de aminácidos ≥ 40% e e-*value* ≤ 1e-5 como recomendado por outros autores (KAMIMURA et al., 2017), usando seus bancos de dados como referência.

A degradação de celulose e hemicelulose envolve o ataque daqueles polímeros por glicosil hidrolases (GH). As famílias de proteínas mais comuns associadas a atividades celulolíticas - GH1, GH3, GH5, GH6, GH8, GH9, GH12, GH45, GH48, GH51 e GH74 (BRUMM, 2013), e *motifs* de ligação à celulose - CBM1, CBM2, CBM3, CBM6, CBM8, CBM30 e CBM44 (BRUMM, 2013; LÓPEZ-MONDÉJAR et al., 2016) foram buscados nas anotações dos genes do catálogo obtidas com o banco de dados PFAM. As famílias protéicas com atividade hemocelulolítica - GH2, GH10, GH11, GH16, GH26, GH30, GH31, GH39, GH42, GH43 e GH53 (LÓPEZ-MONDÉJAR et al., 2016) também foram buscadas dentro das anotações obtidas com o banco de dados PFAM, assim como as monoxigenases líticas de polissacarídeos - LPMO (LÓPEZ-MONDÉJAR et al., 2016), que realizam simultaneamente a degradação de hemicelulose e celulose.

Os microorganismos tendem a liberar ectozimas que ao atuar degradando os materiais refratários e lábeis disponíveis no ambiente, geram uma mistura complexa de compostos de baixo peso molecular. O consumo dessa mistura é mediado por uma vasta diversidade de sistemas de transporte (PORETSKY et al., 2010). Os transportadores

comumente associados à degradação da lignina (transportador MFS, família AAHS, transportadores ABC, família MHS, superfamília ITS e transportador TRAP) foram buscados via Blastp com o programa Diamond versão 0.9.22 (BUCHFINK; XIE; HUSON, 2014), utilizando os mesmos parâmetros supracitados contra o banco de dados referência estabelecido por Kamimura et al. (KAMIMURA et al., 2017).

A degradação da lignina termina na produção de 4-carbóxi-4-hidróxi-2-oxoadipato, que é convertido em piruvato ou oxaloacetato, que são substratos para o ciclo do ácido tricarboxílico (TCA) (KAMIMURA et al., 2017). Da mesma forma, os produtos de degradação da celulose/hemicelulose geram glicose e dímeros de glicose. Recentemente, foram encontradas diversas proteínas de ligação ao substrato do sistema tripartido de transporte de tricarboxilatos (TTT) relacionadas a subprodutos da degradação do carbono orgânico terrestre, como adipato (ROSA et al., 2017) e tereftalato (HOSAKA et al., 2013). Para investigar o metabolismo dos compostos lábeis e a possível ligação entre o sistema TTT e a degradação de lignina, celulose e hemicelulose, as famílias de proteínas TctA (PF01970), TctB (PF07331) e TctC (PF03401) foram pesquisadas diretamente nas anotações obtidas com o banco de dados PFAM.

Os genes encontrados por meio destas estratégias foram submetidos ao programa PSORT v.3.0 (YU et al., 2010), para determinação de sua localização subcelular. Utilizou-se uma previsão nos três táxons possíveis (Gram negativo, Gram positivo e Arqueia), e o melhor *score* foi utilizado para determinar a localização da proteína. Os genes atribuídos com uma localização "desconhecida" foram eliminados, assim como os genes com atribuições errôneas, como, por exemplo, proteínas conhecidamente extracelulares atribuídas à membrana citoplasmática.

A quantidade total de genes encontrados por função (oxidação de lignina, transporte, degradação de celulose e hemicelulose e metabolismo de compostos aromáticos derivados de lignina) por seção de rio normalizada pelas contagens máximas por local e metagenomas foi usada para inferir as diferentes populações microbianas sobre o rio Amazonas. Correlogramas foram feitos usando o número de genes encontrados por metagenoma e a distância linear de cada amostra da nascente do rio Amazonas (rio Mantaro, 10° 43′ 55″ S / 76° 38′ 52″ W), calculada usando os pacotes *Fields* (NYCHKA et al., 2017), Corrplot (WEI; SIMKO, 2017) e RColorBrewer (NEUWIRTH, 2014) implementados em R. Apenas correlações significativas ($p < 0,01$) foram mantidas nas análises finais. O número de genes relacionados com as principais funções foi duplamente plotado contra as populações de bactérias livres e associadas a partículas, a fim de verificar a associação de funções com o estilo de vida.

3.2.8 Disponibilidade dos dados

Os metagenomas utilizados na produção do catálogo de genes do rio Amazonas (AMnrGC) são apresentados no Apêndice 1, e são referentes aos projetos de sequenciamento do banco de dados NCBI-SRA: SRP044326, PRJEB25171 e SRP039390. O AMnrGC, bem como os arquivos de anotação estão disponíveis em: 10.5281/zenodo.1484504. Os metagenomas utilizados na comparação de diversidade genética por meio de k-mers foram detalhados no Apêndice 2 e são provenientes dos projetos de sequenciamento da Proposta JGI 685/300791 (Metagenomas da Floresta Amazônica) e projetos NCBI-SRA: SRP018728 (Rio Mississippi) e PRJNA287840 (bacias hidrográficas do Canadá).

3.3 Resultados

3.3.1 O Catálogo de Genes Microbianos não-Redundantes da Bacia do Rio Amazonas (AMnrGC)

Nosso conjunto de dados original conteve 106 metagenomas oriundos de 30 estações (Figura 3.1) cobrindo cerca de 2100 km do rio Amazonas e sua pluma sobre o oceano Atlântico. A montagem dos metagenomas produziu um total de 2.747.383 *contigs* ≥ 1000 pb, com uma densidade de codificação de proteínas estimada em torno de 70%, totalizando um comprimento total de ~ 5.5 Gpb (Apêndice 4). Identificamos 6.074.767 genes com mais de 150 pb, permitindo códons de iniciação alternativos (GUG e UUG). Após eliminação da redundância por meio do agrupamento de genes com identidade mínima de 95% e sobreposição de 90% do gene mais curto, o AMnrGC incluiu 3.748.772 genes não redundantes com metade deles com comprimento superior a 867 pares de bases.

As diversidades gênica e funcional recuperadas foram representativas da diversidade total presente nos locais de amostragem, conforme indicado pelas curvas de acumulação, que tenderam à saturação (Figura 3.3).

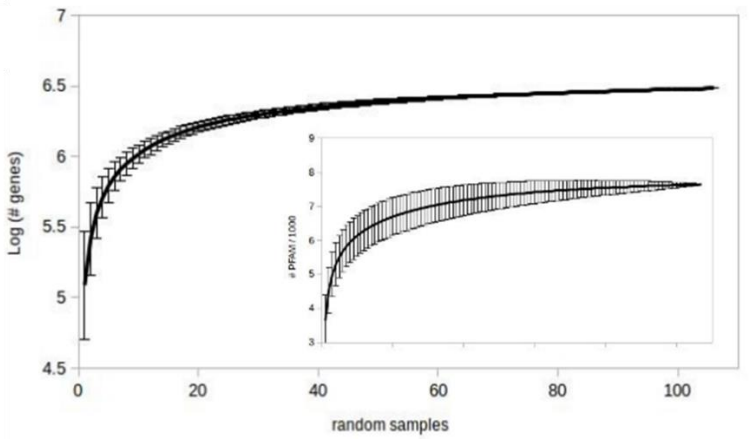

Figura 3.3. Curvas de rarefação em termos de diversidade funcional e gênica. O número de genes por metagenoma foi utilizado para a construção da curva de rarefação superior, enquanto que o número de famílias protéicas encontradas por metagenoma foi utilizado na construção da curva inferior. A randomização na construção de ambas as curvas foi obtida com 100 pseudo-replicatas. (Fonte: Próprio autor)

Enquanto 52% dos genes que compõe o AMnrGC foram anotados com pelo menos um banco de dados de referência (Figura 3.4), por volta de 86% dos genes anotados foram anotados por dois ou mais banco de dados.

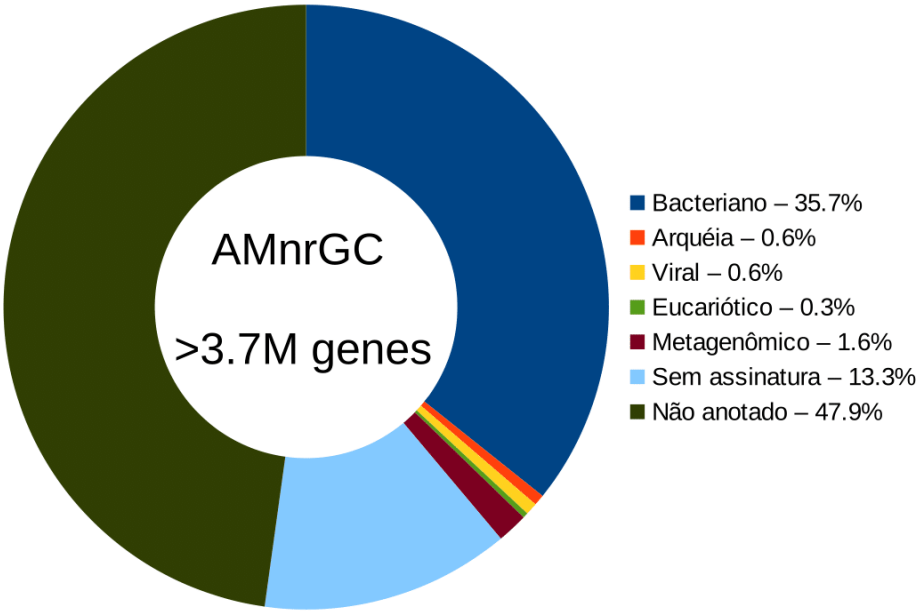

Figura 3.4. Classificação taxonômica dos mais que 3,7 milhões de genes no AMnrGC. Os genes apenas anotados através de métodos probabilísticos foram classificados como "Sem assinatura" e aqueles que apresentavam ortólogos para proteínas mal caracterizadas previamente encontradas em outros metagenomas foram atribuídos apenas como "Metagenômico". Os genes que não apresentaram ortólogos ou quaisquer outras anotações foram identificados como "Não anotado". (Fonte: Próprio autor)

3.3.2 Efeito do filtro ambiental

A legitimidade da divisão do rio Amazonas em 5 seções (montante, jusante, estuário, pluma e oceano) foi investigada por meio da diversidade de k-mers entre os metagenomas usados para gerar o AMnrGC. Essa categorização foi validada por uma β-dispersão significativa (p = 1e-3) e um teste de PERMANOVA significativo (F = 1,52,

$p < 9{,}9x10^{-5}$), como pode ser observado na Figura 3.5. Isto significa que os grupos têm uma grande variabilidade interna, assim como se diferenciam uns dos outros pela sua variabilidade conjunta.

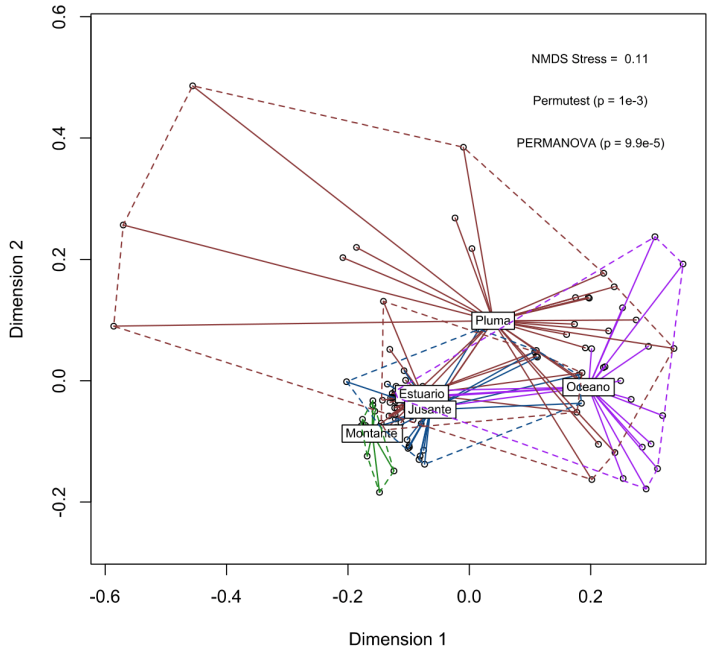

Figura 3.5. Efeito da divisão do rio Amazonas em 5 seções. A diversidade de k-mers dos metagenomas que compõe as diferentes seções do rio Amazonas foi plotada em um NMDS juntamente com a estatística PERMANOVA e o teste de β-dispersão (PERMUTEST). (Fonte: Próprio autor)

Para quantificar a diversidade genética contida no AMnrGC, ou seja, se esta diversidade é advinda de processos locais de evolução, realizou-se um teste de diversidade de k-mers envolvendo amostras de outros ecossistemas. A diversidade de k-mers (Figura 3.6) mostrou uma separação clara entre o rio Amazonas e as amostras de

solo da floresta amazônica, bem como outros rios de água doce (bacias hidrográficas do Canadá e do rio Mississippi). A β-dispersão foi significativa (PERMUTEST, F = 25,7, p < 0,001), sugerindo diferenças na variância intragrupos (distância média diferente dos centroides). No entanto, o teste de PERMANOVA sugere uma diferença estatisticamente significante entre os grupos (R^2 = 0,10, p = 9,99x10-5), o que também foi confirmado pela análise de similaridade - ANOSIM (R = 0,27, p < 0,001).

As amostras de rios temperados se agruparam com amostras de solo da floresta amazônica formando um grupo fechado. Quatro estações do rio Amazonas, representando 13,2% dos metagenomas do rio Amazonas, foram agrupadas no NMDS com amostras de bacias hidrográficas do Canadá (7% dos metagenomas dessa região) (Figura 3.6). Essas amostras do rio Amazonas foram representadas, respectivamente por amostras provenientes de 2 estações na seção à jusante e 3 estações da seção estuário. As amostras do rio Amazonas agruparam-se com 4 amostras de bacias protegidas (14% do total) e 5 amostras sob efeito de áreas agrícolas (12,8% do total) do Canadá. Assim, fica evidente que mesmo havendo uma certa sobreposição entre esses habitats, ela ocorre em uma pequena extensão em relação ao total de amostras observadas para ambos os ambientes.

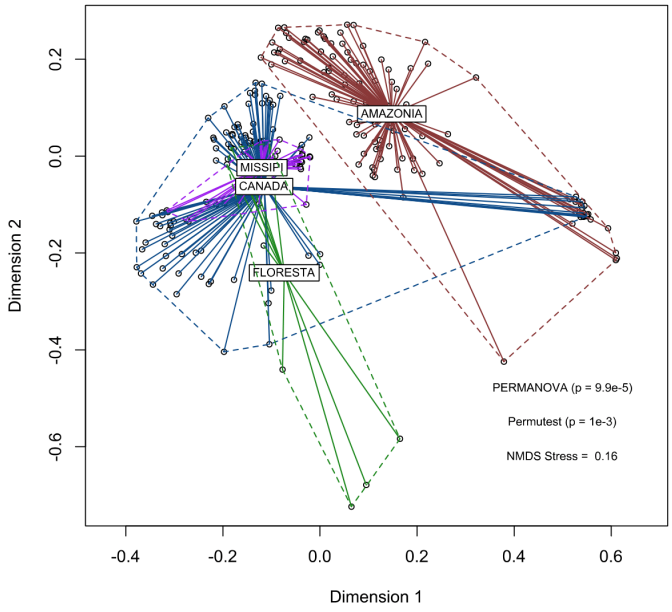

Figura 3.6. A diversidade de k-mers inferida utilizando o programa SIMKA, testando-se o efeito de filtro ambiental. O NMDS com delimitação de polígonos por amostras revela as amostras do rio Amazonas (AMAZONIA), solo florestal amazônico (FLORESTA), bacias hidrográficas do Canadá (CANADA) e rio Mississipi (MISSISSIPI). (Fonte: Próprio autor)

3.3.3 Composição taxonômica do AMnrGC

Quase 47,8% dos genes presentes no AMnrGC não foram atribuídos a uma taxonomia ou função. Além disso, ~1,6% dos genes do AMnrGC foram previamente encontrados em estudos metagenômicos, mas mal caracterizados por não serem atribuídos a um determinado táxon, e foram nomeados aqui como genes "Metagenômicos" (Figura 3.4). Os genes anotados exclusivamente através de perfis de cadeias ocultas de Markov, baseados em alinhamentos estatísticos sem ortologia direta

(também chamados de “Sem assinatura”), representaram 13,3% do AMnrGC. Isso mostra a pouca compreensão do ecossistema amazônico, onde a maioria das proteínas (61,11%) não possui ortólogos nas principais bases de dados de referência. Genes procarióticos (35,7% genes bacterianos e 0,6% genes de arquéias) formam a maior parte do AMnrGC, com apenas 0,3% dos genes provenientes de origem eucariótica e 0,6% de origem viral.

3.3.4 Análise do perfil metabólico da microbiota aquática da bacia amazônica

A superclasse COG de “processos metabólicos” é definida como aquela com funções pertencentes à produção e conversão de energia; metabolismo e transporte de aminoácidos, nucleotídeos, carboidratos, coenzimas, lipídios, transporte de íons inorgânicos, e por fim, a biossíntese, transporte e catabolismo de metabólitos secundários. Esta superclasse foi a mais importante obtida dentre as presentes no AMnrGC (35,8% dos genes anotados com bases de dados COG e eggNOG) como mostrado na Figura 3.7. Como limitação experimental não obtivemos amostras de metagenomas da montante do rio Amazonas na fração de microrganismos associados à partículas, dessa forma, esta seção foi eliminada das comparações realizadas nos microrganismos associados à partículas.

A classe COG mais representada no AMnrGC foi relativa a genes com função desconhecida (também dita classe “S”), compondo 21,43% de proteínas anotadas com classes COG. Isto revela que mesmo entre os genes anotados nem sempre as suas

funções estão disponíveis e que estes foram abundantes no microcosmo da água doce do rio Amazonas, mantendo um potencial biotecnológico inexplorado.

As funções "*core"* bacterianas foram aqui definidas como aquelas codificadas por muitas espécies e que estão envolvidas na homeostase de todo o ecossistema. Este *core* bacteriano representou cerca de 8% das funções identificadas com os bancos de dados KEGG e PFAM. As 100 funções mais abundantes observadas no *core* de funções procarióticas foram funções de "manutenção" envolvidas nas principais vias metabólicas (por exemplo, o metabolismo de carboidratos, o *quorum sensing* e o transporte e metabolismo de aminoácidos) e complexos protéicos importantes (por exemplo, RNA e DNA polimerases e ATP sintase). Interessantemente, as vias relacionadas às funções do tipo "*core"* foram relacionadas ao metabolismo acetogênico e metilotrofia, enquanto as funções gênicas do tipo "não *core"* foram associadas ao metabolismo de compostos xenobióticos e na produção, consumo e transporte de produtos do metabolismo secundário.

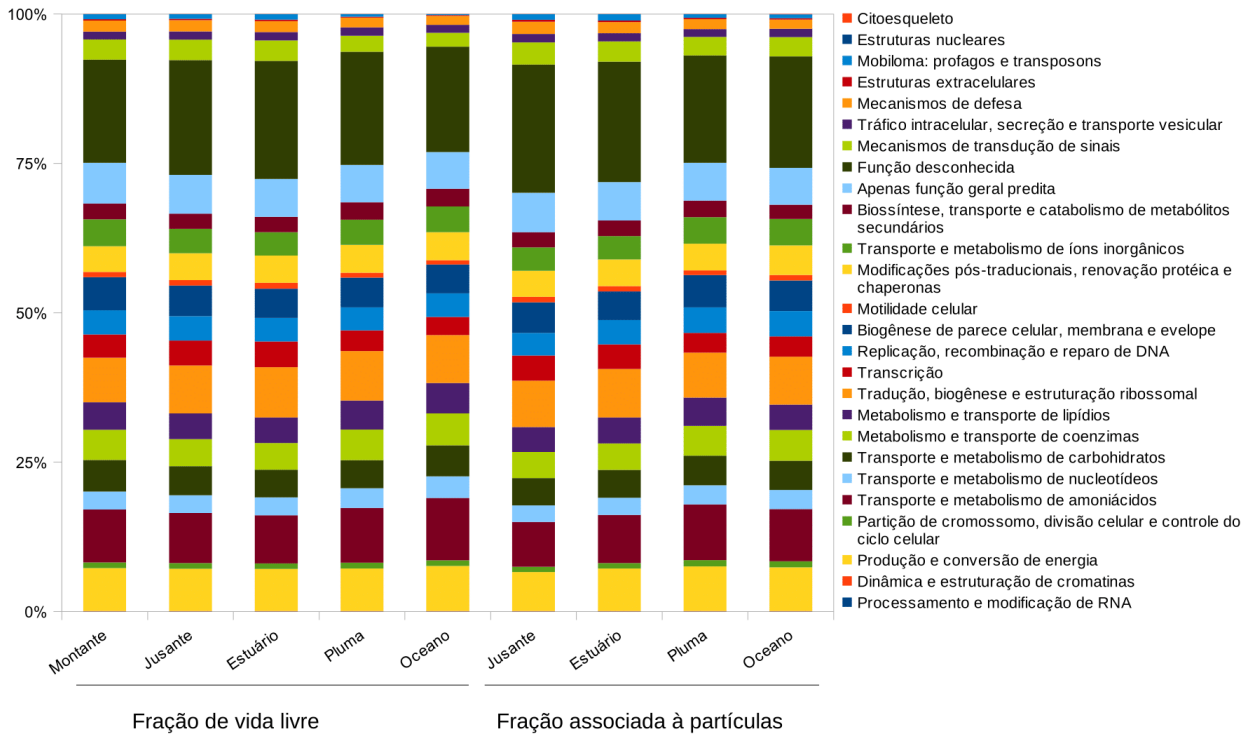

Figura 3.7. Perfil funcional de seções e frações da bacia amazônica. Foram utilizadas as funções eggNOG e COG, codificadas como classes COG por seção. É possível observar a classe correspondente ao metabolismo e transporte de aminoácidos respondendo diferencialmente ao efeito das frações e seções do rio. A classe de função desconhecida ocupa uma quantidade maior de genes na microbiota associada a partículas. (Fonte: Próprio autor)

3.3.5 Maquinaria bioquímica de degradação de OM terrestre

Um total de 6.516 genes do AMnrGC (representando 0,17% do total) foram identificados como a maquinaria de degradação da OM terrestre do microbioma do rio Amazonas, sendo divididos em: degradação da celulose (143 genes), degradação da hemicelulose (92 genes), oxidação da lignina (73 genes), transporte e metabolismo de compostos aromáticos derivados da lignina (2.324 genes) e transporte de tricarboxilatos (3.884 genes). As principais funções estão representadas na Figura 3.8.

3.3.6 Estilo de vida microbiano influencia a maquinaria de degradação da OM terrestre

A partir dos resultados de uma análise de variância de 1 fator (ANOVA), que foi realizada comparando-se o número de genes relacionados à maquinaria de degradação de OM terrestre obtidos por metagenoma, verificou-se não haver diferença significativa no número de funções associadas à degradação da celulose e hemicelulose, bem como à oxidação da lignina nas frações testadas (Apêndice 5). No entanto, o teste de variância dentro e entre os grupos mostrou micróbios de vida livre e microbiota associada a partículas como diferentes grupos fechados (Apêndice 6).

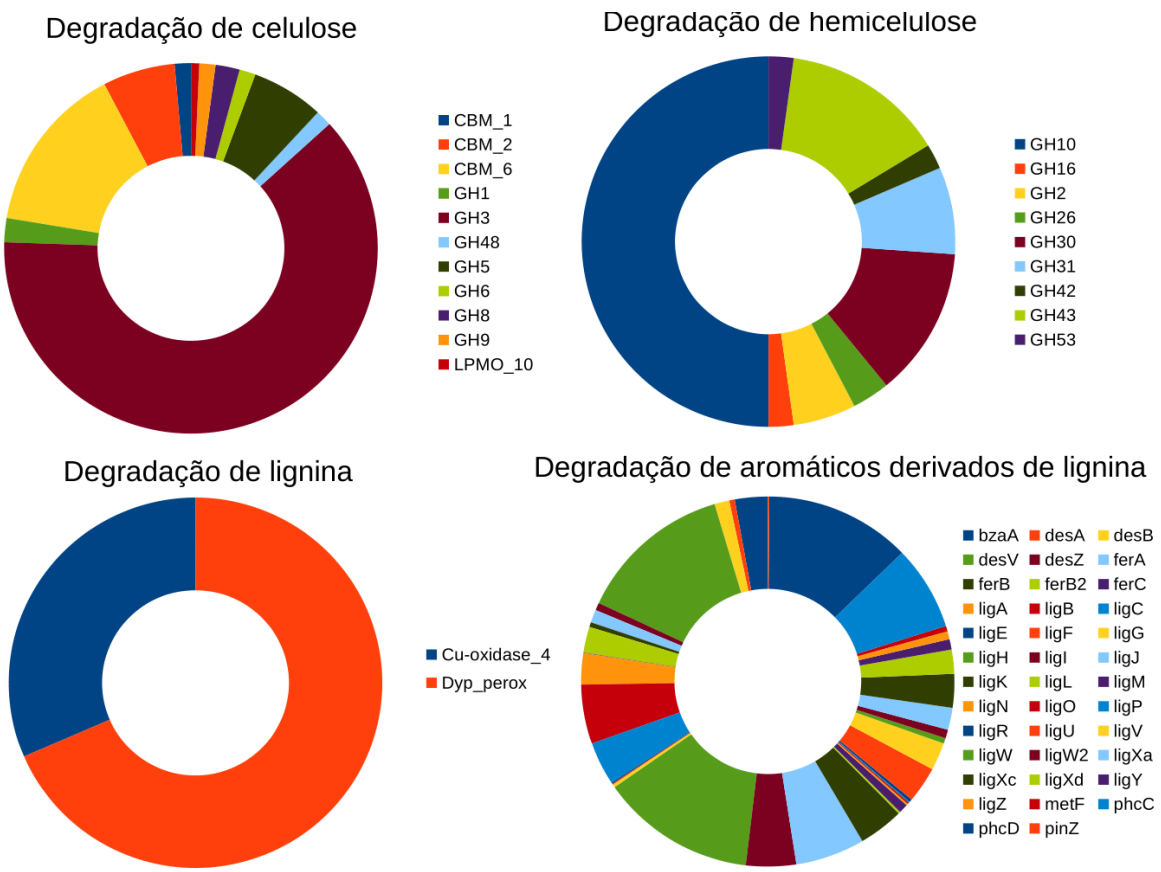

Figura 3.8. Perfis de degradação de diferentes etapas da degradação de OM terrestre. O número de genes atribuídos como diferentes famílias de enzimas foi usado para produzir os gráficos acima. (Fonte: Próprio autor)

O número de genes associados à localização periplasmática diminui substancialmente na transição para águas salobras, sendo substituído por proteínas de superfície celular e enzimas citoplasmáticas nos oceanos (Figura 3.9). Outra característica é a predominância da degradação da celulose sobre a oxidação da lignina e a degradação da hemicelulose nos oceanos (Figura 3.9). Nas águas doces, as funções parecem funcionar em equilíbrio, onde a oxidação da lignina e a degradação da hemicelulose são dominantes. Além disso, as seções do rio também mostraram uma β-dispersão não significativa ($p_{PERMUTEST} > 0.05$), com uma PERMANOVA significativa ($p < 0,05$), reforçando mais uma vez a divisão aqui estabelecida, também em termos do aparato de degradação da OM terrestre (Apêndice 6), da influência das seções da bacia amazônica e do tipo de funções bioquímicas observadas em relação a degradação de OM terrestre.

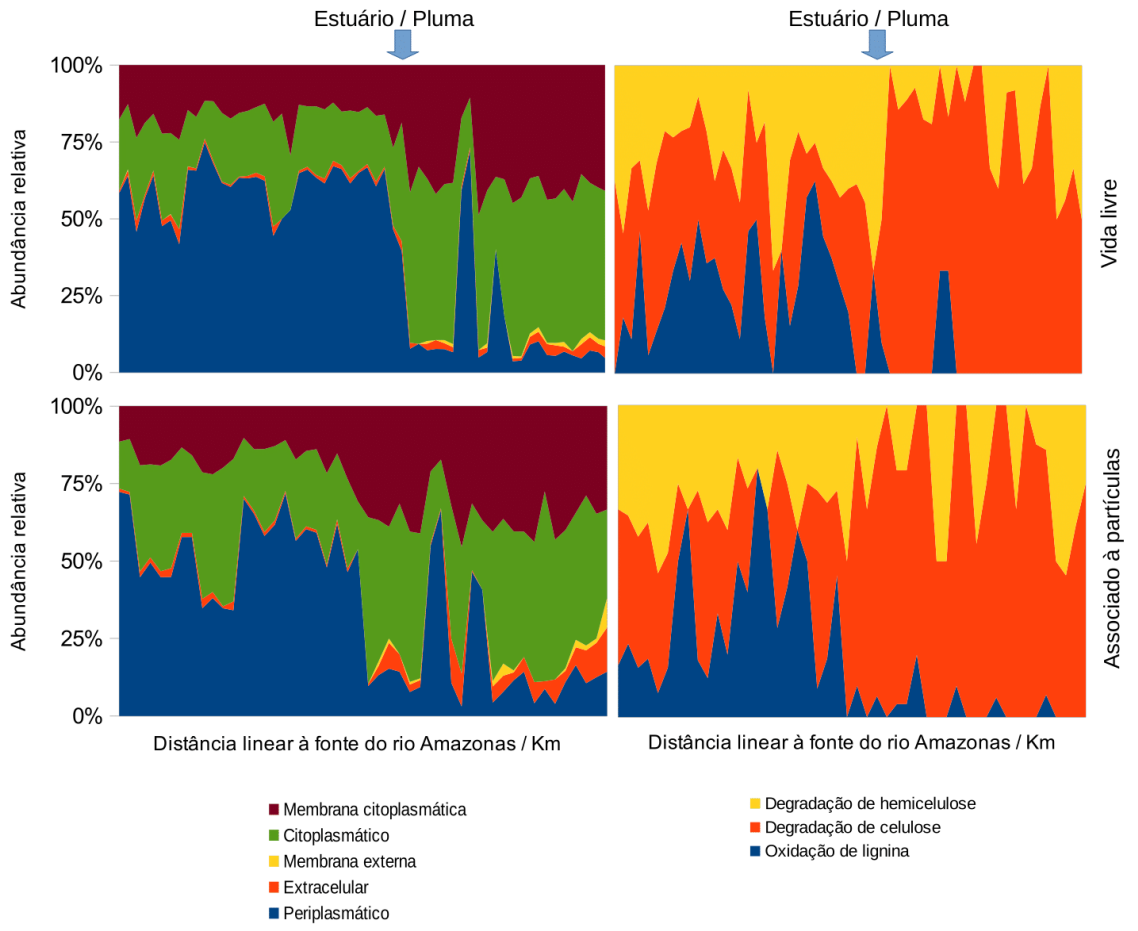

Figura 3.9. Perfil funcional do rio Amazonas por estilo de vida e localização subcelular. O número de genes associados a uma função específica por metagenoma foi usado para construir os gráficos. A localização subcelular (a) e as principais funções de desconstrução dos biopolímeros (b) são mostradas pelo estilo de vida e distância linear crescente da fonte do rio Amazonas (da esquerda para a direita nos gráficos). (Fonte: Próprio autor)

3.3.7 A oxidação de lignina e a degradação de celulose e hemicelulose

A oxidação da lignina é uma função mediada principalmente por peroxidases degradadoras de corantes (DyPs) e predominantemente associadas a águas doces, independentemente do estilo de vida microbiano. Apenas lacases e peroxidases foram encontradas no ecossistema do rio Amazonas, não tendo sido encontradas descarboxilases de ácido fenólico ou glioxal oxidases.

Como a degradação da celulose e da hemicelulose parece substituir quase que totalmente a oxidação da lignina em águas salobras (Figura 3.9), decidimos investigar a composição dessas funções sobre o estilo de vida microbiano (Figura 3.10). A degradação da hemicelulose parece ser homogênea em ambas as frações, embora a família GH53 seja encontrada apenas em micróbios de vida livre.

Enquanto isso, a degradação da celulose tende a formar mais padrões. Um dos mais interessantes foi o estabelecimento de uma maior abundância do *motif* de ligação à celulose 2 (CBM2) nas proteínas encontradas nos micróbios de vida livre, assim como, por enzimas das famílias GH1 e GH6, que também estão associadas à água doce. O *motif* CBM1 foi associado as proteínas de micróbios associados a partículas em águas salobras, assim como a família GH9.

De um modo geral, há uma família dominante em ambas as vias de desconstrução de biopolímeros. A família GH10 parece ser a principal enzima na degradação da hemicelulose, em ambas as frações e em todas as seções. Foi observada dominância ubíqua semelhante por parte da família GH3 na degradação da celulose. As monooxigenases polissacarídicas líticas funcionam tanto na desconstrução da celulose

como na hemicelulose, interessantemente presentes apenas na seção do rio à montante na fração de vida livre.

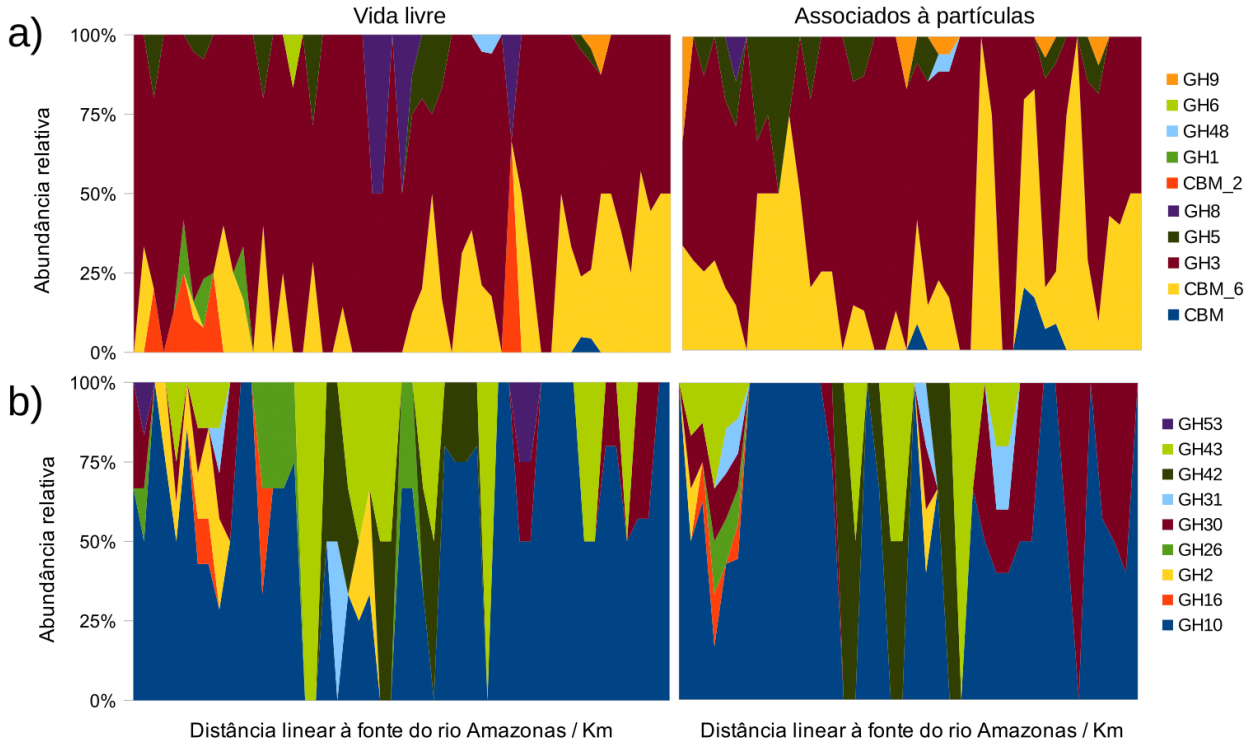

Figura 3.10. Perfil funcional da maquinaria de degradação de celulose e hemicelulose no rio Amazonas por estilo de vida microbiano. Principais abundâncias das famílias de proteínas na degradação da celulose (a) e hemicelulose (b) no rio Amazonas pela distância linear crescente da fonte do rio Amazonas (da esquerda para a direita nos gráficos). (Fonte: Próprio autor)

3.3.8 Degradação dos compostos aromáticos derivados de lignina

Todas as funções relacionadas ao metabolismo de compostos aromáticos derivados de lignina foram encontradas, exceto o gene ligD, uma Cα-desidrogenase para os isômeros αR dos β-aril éteres. A via completa de degradação de compostos derivados de lignina (Figura 3.8) foi verificada com 772 e 449 genes pertencentes, respectivamente, às vias de funilamento de diaris e monoaris, seguido por 346 genes que compõem as etapas de O-desmetilação e metabolismo tipo C1 e, finalmente, a via de clivagem do anel de protocatecuato com 713 genes. Quatro famílias de genes (ligH, desV, phcD e phcC) representam em conjunto 46,8% de todos os genes relacionados com o metabolismo de compostos derivados de lignina. Esses 4 genes são representantes dos três principais passos do metabolismo da lignina intracelularmente.

Para entender melhor o uso dos substratos gerados a partir da oxidação de lignina no rio Amazonas, os genes por seção do rio foram divididos em substrato preferencial (Figura 3.11a) e o número de genes foi normalizado pela média de genes encontrados em todos os trechos do rio Amazonas. É possível observar uma degradação preferencial do pinoresinol na pluma e estações à jusante de Manaus, e dos β-aril éteres (β-O-4) no estuário, enquanto que as estações no oceano não apresentaram qualquer preferência. A seção à montante, apesar do número geralmente baixo de genes, mostrou uma clara preferência por dois substratos: fenilcoumaran (β-5) e ferulato. De um modo geral, uma preferência por comunidades degradando diaris pôde ser observada no estuário e pluma, enquanto a seção à montante tem uma microbiota mais adaptada a ambos, diaris e monoaris.

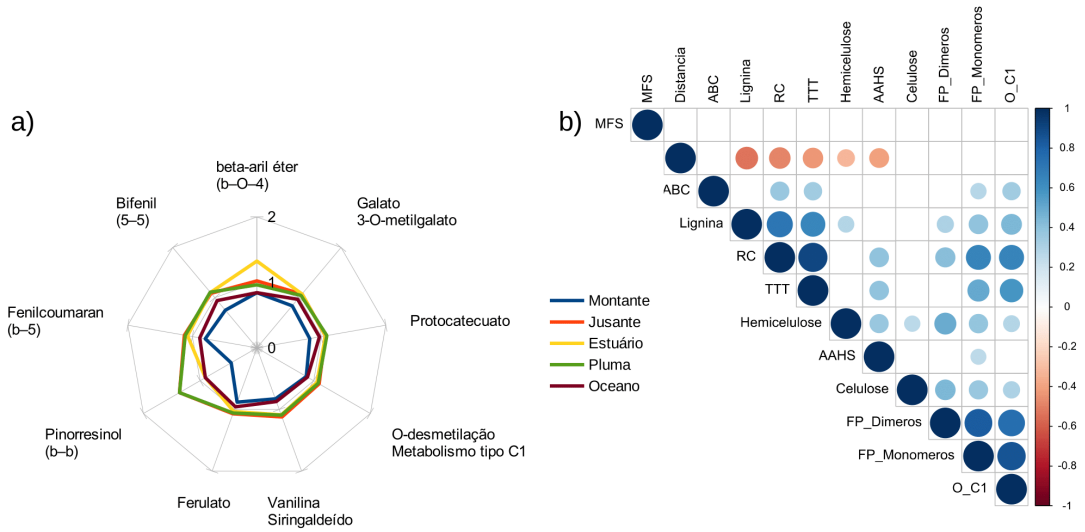

Figura 3.11. Processamento de compostos aromáticos derivados de lignina no rio Amazonas. O número de genes associados a uma via específica de processamento de compostos aromáticos derivados de lignina foi normalizado pela média dos genes encontrados em todos os trechos do rio Amazonas e foi plotada em relação aos substratos usados nessas vias (a). A correlação do número de genes associados à oxidação da lignina (lignina), desconstrução da celulose e hemicelulose (celulose e hemicelulose, respectivamente), sistemas de transporte (AAHS, MFS, ABC e TTT), vias de processamento de compostos aromáticos derivados da lignina (RC: clivagem do anel de protocatecuato; O_C1: vias de desmetilação/metabolismo C1; vias de funilização de dímeros - FP_Dimeros e Monômeros - FP_Monomeros), e a distância linear das amostras da fonte do rio Amazonas (Distancia) são mostrados (b). Apenas correlações significativas ($p < 0,05$) foram desenhadas. (Fonte: Próprio autor)

A existência de uma potencial degradação diferencial de determinados substratos gerados a partir da oxidação de lignina em alguns sítios sugere uma possível correlação entre as múltiplas variáveis aqui tomadas para entender a maquinaria bioquímica da degradação da OM terrestre. As correlações significativas foram plotadas em um correlograma (Figura 3.11b). A distância linear das amostras da nascente do rio Amazonas apresentou coeficientes de correlação negativos com a oxidação da lignina, degradação da hemicelulose, via de clivagem do anel de protocatecuato, transporte tripartido de tricarboxilatos e transportadores do tipo AAHS.

A diversidade da maquinaria bioquímica relacionada à degradação de compostos aromáticos derivados da lignina apresenta uma correlação positiva com a diversidade do aparato de oxidação da lignina. Embora a oxidação da lignina e as vias de degradação da hemicelulose tenham apresentado uma pequena correlação positiva, os genes de degradação da celulose não mostraram qualquer correlação com a oxidação da lignina, apenas uma correlação positiva menor (mas significativa) com a degradação da hemicelulose. Portanto, a substituição de funções da lignina pela degradação da celulose não é proporcional ao longo do curso do rio. Os números de genes relacionados com a degradação da hemicelulose foram positivamente correlacionados com as vias de degradação dos compostos aromáticos derivados da lignina, exceto aquelas relacionadas à clivagem do anel de protocatecuato, em vez da oxidação da lignina. Isto poderia indicar um potencial efeito de *priming* do metabolismo de hemicelulose sobre a degradação de derivados de lignina, principalmente nas vias de afunilamento para mono- e diaris.

Embora haja uma aparente preferência pela degradação de certos substratos por seção de rio, o correlograma (Figura 3.11b) não apresentou nenhum tipo de

correlação entre os diferentes tipos de vias de funilamento (FP_Dimeros e FP_Monomeros) e a distância linear da amostra da fonte do rio Amazonas.

3.3.9 Sistemas transportadores usuais e novas descobertas

Compostos aromáticos derivados de lignina precisam ser transportados do ambiente extracelular para o citoplasma antes de sua degradação. Comumente, os transportadores associados à degradação de lignina (transportador MFS, família AAHS e transportadores ABC) foram encontrados no AMnrGC, enquanto outros (família MHS, superfamília ITS e transportador TRAP) não foram encontrados. O sistema MFS não mostrou correlação com nenhuma das variáveis testadas. Apenas o sistema de transporte AAHS foi negativamente correlacionado à distância linear das amostras da nascente do rio Amazonas, as demais famílias não apresentaram nenhum tipo de correlação com a distância (Figura 3.11b). Enquanto isso, os transportadores AAHS e ABC também mostraram correlação positiva com a via de afunilamento dos monoaris e da via de clivagem do anel de protocatecuato, em que o sistema ABC também se correlaciona positivamente com a O-desmetilação e o metabolismo do tipo C1.

O sistema tripartido de transporte de tricarboxilatos (TTT) correlacionou-se positivamente com os transportadores AAHS e ABC (Figura 3.11b), significando uma possível complementaridade de funções. Além disso, o sistema TTT mostrou uma correlação positiva com a oxidação da lignina e a degradação da hemicelulose. Outra característica importante do sistema TTT no rio Amazonas é a sua aparente redução à medida que o rio ruma para o oceano.

Plotando o número normalizado de genes agrupados pelos substratos transportados por seus ortólogos (Figura 3.12), foi possível observar uma preferência dos micróbios da seção à montante de Manaus por tricarboxilatos. Assim como, o transporte de galato reduzido nas seções de água doce. Por sua vez, o transporte de p-cumarato e ferulato/Di-hidroferulato também aparece reduzido nas seções de água doce e oceano, exceto no estuário e pluma. Os outros substratos aparecem similarmente transportados em todas as seções, excetuando-se à montante de Manaus, que aparece, de um modo geral, subrepresentada na Figura 3.12.

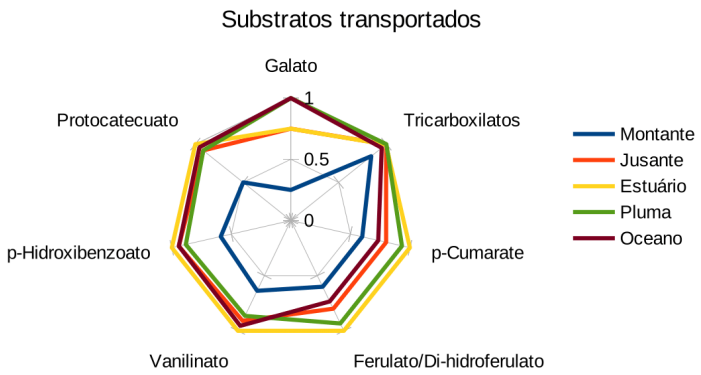

Figura 3.12. Perfil do transporte de compostos aromáticos derivados de lignina nas seções do rio Amazonas. O número de genes relacionados a diferentes famílias transportadoras normalizadas pelo número máximo de genes encontrados em todos os locais está em (a). Os substratos transportados pelos diferentes genes foram observados em (b) também normalizados pelo valor máximo dos genes obtidos em todos os locais. (Fonte: Próprio autor)

3.4 Discussão

Neste trabalho, nós relatamos o primeiro catálogo de genes não-redundantes de microrganismos aquáticos da bacia do rio Amazonas, com um elevado número de genes provenientes de fontes bacterianas. Uma análise da diversidade genética sugeriu que a mesma pode ser derivada de processos evolutivos locais. Os genes relacionados à degradação da OM terrestre apresentaram padrões investigados enquanto sua presença em diferentes frações e seções de rios. Observou-se que o estilo de vida e a distribuição espacial dos genes influenciam a maquinaria bioquímica disponível para a degradação da OM terrestre.

3.4.1 Catálogo de genes microbianos não redundantes da bacia amazônica (AMnrGC)

O AMnrGC nos permite expandir consideravelmente nossa compreensão dos microbiomas de água doce tropical. Os genes desconhecidos estão concentrados principalmente em bactérias de água doce e, quando em águas salobras, principalmente naquelas associadas a partículas. Isto se deve, provavelmente, ao melhor entendimento dos sistemas procarióticos oceânicos, que possuem mais informação disponível. Pouco menos da metade dos mais de 3,7 milhões de genes observados no AMnrGC ainda são desconhecidos, indicando que este ambiente tem um grande potencial de diversidade microbiana. Os genes não anotados (48% do AMnrGC) apontam para uma

biodiversidade ainda não descoberta, por conta da falta de genes ortólogos nos bancos de dados de referência.

As distâncias ecológicas, substituindo as contagens de espécies pela contagem de k-mers utilizando-se o método SIMKA, captam quantitativamente e qualitativamente a estrutura biológica essencial das comunidades microbianas, sendo altamente correlacionadas com perfis taxonômicos baseados na estratégia de alinhamento de todas as sequências (BENOIT et al., 2016). A distância de Jaccard calculada a partir da matriz de ausência e presença de k-mer mostrou que a diversidade biológica observada nos metagenomas da bacia do rio Amazonas está muito distante de outros rios de água doce utilizados nesta comparação, assim como daquela observada em solo da floresta amazônica. No entanto, o pequeno grupo de amostras co-agrupadas revela uma pequena porção das amostras do rio Amazonas semelhantes às das bacias hidrográficas do Canadá. A grande maioria das amostras do rio Amazonas, por sua vez, mostra o grupo, de um modo geral, diferente de outros ambientes. Estes resultados corroboram com a ideia de diversidade funcional e genética observada no AMnrGC, que pode refletir processos locais de evolução selecionados pelo filtro ambiental. Assim, também reforça as descobertas como novas e importantes no contexto da estrutura microbiana e da dinâmica bioquímica da microbiota dos corpos hídricos de água doce.

A atribuição taxonômica é uma tarefa difícil e depende de ortólogos disponíveis em bancos de dados de referência. Mais da metade dos nossos genes não possuem uma correspondência nos bancos de dados mais completos até o momento, permanecendo desconhecidos. O esquema de filtragem evitou uma contribuição excessiva de eucariotos, representando uma fração menor que ~ 0.3% do AMnrGC, provavelmente proveniente de restos celulares ou DNA livre suspenso. Entretanto, a

extensão real da contribuição eucariótica é desconhecida, uma vez que há uma enorme quantidade de genes não anotados. Genes provenientes de amostras da pluma e oceano em ambas as frações (vida livre e associada a partículas) revelaram um maior conteúdo de genes de eucariotos. Um trabalho recente (CARRADEC et al., 2018) envolvendo os dados da expedição *TARA-Oceans*, revelou uma diversidade gênica associada aos potenciais bioquímico e ecológico de eucariotos de água salobra muito importante. Trabalhos futuros envolvendo o catálogo AMnrGC poderão desvendar este conteúdo genético eucariótico.

3.4.2 Estrutura espacial dos genes e seu potencial metabólico

Observamos uma fração relativamente grande de genes compartilhados por todos os tipos de água, o que sugere um conjunto comum de funções presentes em todo o microbioma da bacia do rio Amazonas. O AMnrGC é composto predominantemente por genes microbianos de água doce, com contribuições exclusivas da pluma e águas salobras, significando uma maior diversidade genética advinda de águas doces. A zona de pluma, onde as águas oceânicas se misturam com a água doce do rio, revelou-se uma zona de maior diversidade genética do que o oceano, reforçando a ideia de uma clara separação evolutiva entre linhagens de água doce e marinha (LOGARES et al., 2009). Curiosamente, as maiores quantidades de genes pertencentes à fração associada a partículas poderiam estar associadas às células maiores e às maiores quantidades de DNA, que podem sobrerepresentar esses micróbios (RIECK et al., 2015).

O padrão de agrupamento da diversidade de genes sugere a salinidade como um fator chave na determinação da colonização espacial dos micróbios. Além disso, a pluma representa uma zona de transição, que em termos de diversidade genética é similar às amostras oceânicas (Figura 3.5). Autores anteriores (LOGARES et al., 2009) propõem que em ambientes como esses haja uma população microbiana complexa distribuída de forma heterogênea, onde o rápido crescimento populacional promove o surgimento de populações localmente adaptadas e evita o estabelecimento de populações imigrantes. Ainda é importante mencionar o ambiente físico único da pluma do rio Amazonas, que consiste em um corredor de conectividade para várias espécies que crescem sob baixa luminosidade e altos níveis de partículas estabelecidos por um sistema de recifes (MOURA et al., 2016).

As funções COG dentro da superclasse “Metabolismo” foram as mais abundantes no AMnrGC, bem como no rio Mississippi em sua porção superior (STALEY et al., 2014b). A distribuição similar das funções COG foi observada entre diferentes seções do rio Amazonas (Figura 3.7), sugerindo que um conjunto de "características funcionais *core*" é conservado em todo o rio, como previamente inferido para o rio Mississippi (STALEY et al., 2014b). As funções “*core*” bacterianas representaram menos de 8% das funções anotadas com KEGG e PFAM e nos ajudaram a entender o metabolismo procariótico conservado. Este, por sua vez, composto principalmente por funções do metabolismo de carboidratos e vários sistemas de transportadores, principalmente do tipo ABC. Isso sugere que uma maquinaria sofisticada para processar carbono, produzir e economizar energia poderia ser uma estratégia geral dos micróbios de água doce, não somente daqueles da Amazônia.

As vias não essenciais (*non-core*) revelam traços de adaptação específicos para ambientes quimicamente complexos, com biodegradação de compostos xenobióticos e metabolismo secundário. Microrganismos associados a partículas e de vida livre parecem possuir o mesmo metabolismo central em potencial, com poucas diferenças, significando uma maneira similar de lidar com os compostos provenientes do meio ambiente.

A classe COG de genes de função desconhecida (S) foi sobrerepresentada e pode guardar um potencial biotecnológico substancial e inexplorado. Proteínas previamente caracterizadas a partir do ambiente microbiano da Amazônia também puderam ser identificadas no AMnrGC, como por exemplo, β-glucosidases (BERGMANN et al., 2014; TOYAMA et al., 2018) e α-manosidases (MATSUDA et al., 2011), o que reforça as potenciais aplicações do AMnrGC na descoberta e exploração de novas enzimas.

3.4.3 *Destino da matéria orgânica (OM) terrestre e o efeito* priming *no rio Amazonas*

O destino do carbono no ecossistema aquático da Amazônia está mais relacionado às vias acetogênica e metanogênica, corroborando com achados anteriores (SATINSKY; SMITH; SHARMA; WARD; et al., 2017), em que o metabolismo de C1 (metano monooxigenase - mmoB e enzima de ativação de formaldeído - fae) foi visto como altamente expresso.

Existem muitas vias microbianas possíveis para degradar as principais fontes de carbono (lignina, celulose e hemicelulose) na bacia do rio Amazonas, como

observado anteriormente (WARD et al., 2013). No rio Amazonas, outros autores (SEIDEL et al., 2016) também sugerem vários processos químicos, como remineralização microbiana, sorção, fotodegradação, mesmo em pequena extensão (REMINGTON; KRUSCHE; RICHEY, 2011), e troca lateral com várzeas ou solos, além da degradação microbiana atuando também na OM terrestre. A lignina é altamente resistente à degradação (KÖGEL-KNABNER, 2002) e seu papel é impedir que as enzimas microbianas degradem os polissacarídeos lábeis da parede celular (PAULY; KEEGSTRA, 2008). Sua via de degradação envolve um passo de oxidação determinante, mediado no rio Amazonas por lacases e DYPs. Desta forma, o efeito priming parece determinante para a degradação eficiente da lignina (WARD et al., 2016), assim como, parece aumentar a degradação da celulose/hemicelulose previamente protegida por ela.

Por meio do NMDS calculado com as abundâncias de famílias gênicas (Apêndice 6) foi possível observar os micróbios associados a partículas e de vida livre formando agrupamentos separados. Isso significa que o conteúdo gênico por metagenoma varia entre frações de tamanho, mas não em termos de funções, e a distribuição dessas funções ao longo do curso do rio tem um padrão similar em ambos os estilos de vida. Na água doce, a oxidação da lignina e a degradação da hemicelulose são dominantes, o que estabeleceria um *continuum* de degradação da OM terrestre, mostrando que a distância é uma característica importante a ser considerada neste tipo de estudo.

A seção oceânica não mostra uma clara preferência por quaisquer substratos, sendo provavelmente menos propensa a degradar os compostos aromáticos derivados de lignina. A oxidação da lignina, a degradação da hemicelulose, a via de clivagem do anel

de protocatecuato, o sistema TTT e os transportadores do tipo AAHS tendem a reduzir seus níveis à medida que as estações se aproximam do oceano, reforçando a ideia de uma substituição das funções de degradação de lignina pelas de degradação de celulose nos oceanos (Figura 3.9). A abundância de genes relacionados à degradação dos compostos aromáticos derivados de lignina também apresentam uma correlação positiva com as funções de oxidação da lignina, como esperado.

A degradação de celulose foi principalmente mediada pela família de GH3, uma família de enzimas com 2 domínios globulares e que são exo-atuantes, mostrando que a degração é favorecida nas extremidades do polímero. Por sua vez, a degradação de hemicelulose foi mediada pelas GH10, que possuem uma relação com as GH11 que não foram encontradas. Geralmente, as famílias GH10 e GH11 são encontradas juntas e tem uma maior relação com fungos. No entanto, as GH10 possuem uma melhor performance em relação às GH11, mesmo estas sendo mais ativas cineticamente. Uma explicação seria que as GH10 são mais termoestáveis, além de ter maior acessibilidade ao xilano, preferindo-o quando acetilado (HU; SADDLER, 2018).

As vias de degradação da hemicelulose/celulose foram positivamente correlacionadas com as vias de degradação dos compostos aromáticos derivados da lignina, em vez da oxidação da lignina. Isto sugere um potencial efeito de iniciação impulsionado por comunidades microbianas degradadoras de celulose/hemicelulose. Apesar da aparente preferência pela degradação de certos substratos por seção de rio, não foi encontrada correlação entre as vias de afunilamento e a distância linear da amostra à nascente do rio Amazonas, o que pode representar uma variação desproporcional do número de genes ao longo do rio. Em outras palavras, essas populações deveriam formar zonas de ação ao longo do curso do rio e suas sobras

representariam o principal elo de funcionamento entre elas, o que provavelmente, afeta e auxilia as comunidades à jusante a se expandirem.

Os compostos aromáticos derivados de lignina precisam ser transportados do ambiente extracelular para o citoplasma antes de sua degradação. Sistemas diferentes podem funcionar nesses compostos, os quais foram resumidos por outros autores (KAMIMURA et al., 2017; PORETSKY et al., 2010). A degradação de compostos aromáticos derivados da lignina acaba em produtos que podem ser convertidos em piruvato ou oxaloacetato, substratos para o TCA (KAMIMURA et al., 2017), assim como os produtos da degradação de celulose e hemicelulose. Descobertas recentes sugerem que o sistema TTT pode estar relacionado ao transporte de compostos potencialmente produzidos como subprodutos da degradação de OM terrestre (HOSAKA et al., 2013; ROSA et al., 2017). Pela primeira vez, o sistema TTT pode ser conectado com um sistema funcional, neste caso com outros transportadores (dos tipos AAHS e ABC) e a funções relacionadas à oxidação de lignina e degradação de hemiceluloses, reforçando a ideia de seu papel na degradação da OM terrestre, e sugerindo um papel de coadjuvante no efeito *priming*.

O sistema TTT no rio Amazonas, e a sua aparente redução frente ao oceano, provavelmente está relacionado com a natureza das proteínas de ligação ao substrato, necessárias a este sistema, que mudam a partir dos principais materiais degradados no ambiente, neste caso, no oceano. A degradação dos compostos aromáticos derivados da lignina provavelmente estaria atuando nos estágios finais de degradação da via de afunilamento dos monoáris, passando pela O-desmetilação e metabolismo C1, e pela via de clivagem do anel de protocatecuato.

Com base em todas as descobertas acima mencionadas, propomos um modelo de efeito *priming* atuando em complexos lignocelulósicos do sistema fluvial amazônico (Figura 3.13), onde comunidades degradadoras de hemicelulose geram compostos facilmente assimiláveis por comunidades degradadoras de lignina, proporcionando carbono e energia para si e para as comunidades relacionadas. Por sua vez, comunidades degradadoras de lignina liberam ectozimas que geram compostos aromáticos derivados de lignina. Estes compostos são em menor extensão fotodegradados para gerar moléculas alicíclicas ricas em carboxila (CRAMs), que são novamente assimilados pelas comunidades degradadoras de celulose como nutrientes limitantes. Esses compostos derivados de lignina podem, ainda, retornar às comunidades degradadoras de lignina, onde são consumidos principalmente por uma comunidade especializada em uso de diaris, e outra especializada no consumo de monoáris. Isso ainda leva a um submodelo de *priming* dentro dessas comunidades, onde há uma retroalimentação positiva com as sobras de uma comunidade para outra, fornecendo não apenas carbono e energia estruturais, mas também nutrientes limitantes.

A degradação dos compostos de lignina e celulose sofre forte regulação pelo consumo de fontes de carbono facilmente decomponíveis (KLOTZBÜCHER et al., 2011). As proteínas de ligação ao substrato (TctC) do sistema TTT foram super-representadas no AMnrGC, algo também observado anteriormente no rio Amazonas (GHAI; RODRIGUEZ-VALERA; et al., 2011), o que poderia refletir os tricarboxilatos como uma fonte comum de carbono nesse rio. Sua presença, nestes termos, é um indicativo de que eles poderiam estar atuando no uso de fontes alternativas de carbono geradas pela degradação de polímeros biológicos complexos, regulando as vias de degradação por meio de *feedback*. Além disso, também podem estar realizando um

papel na repressão ao catabolismo de ligninas, celuloses e hemiceluloses; uma vez que representam uma fonte de carbono mais facilmente assimilável que as demais, reforçando a hipótese de seu papel na regulação da degradação da celulose e lignina.

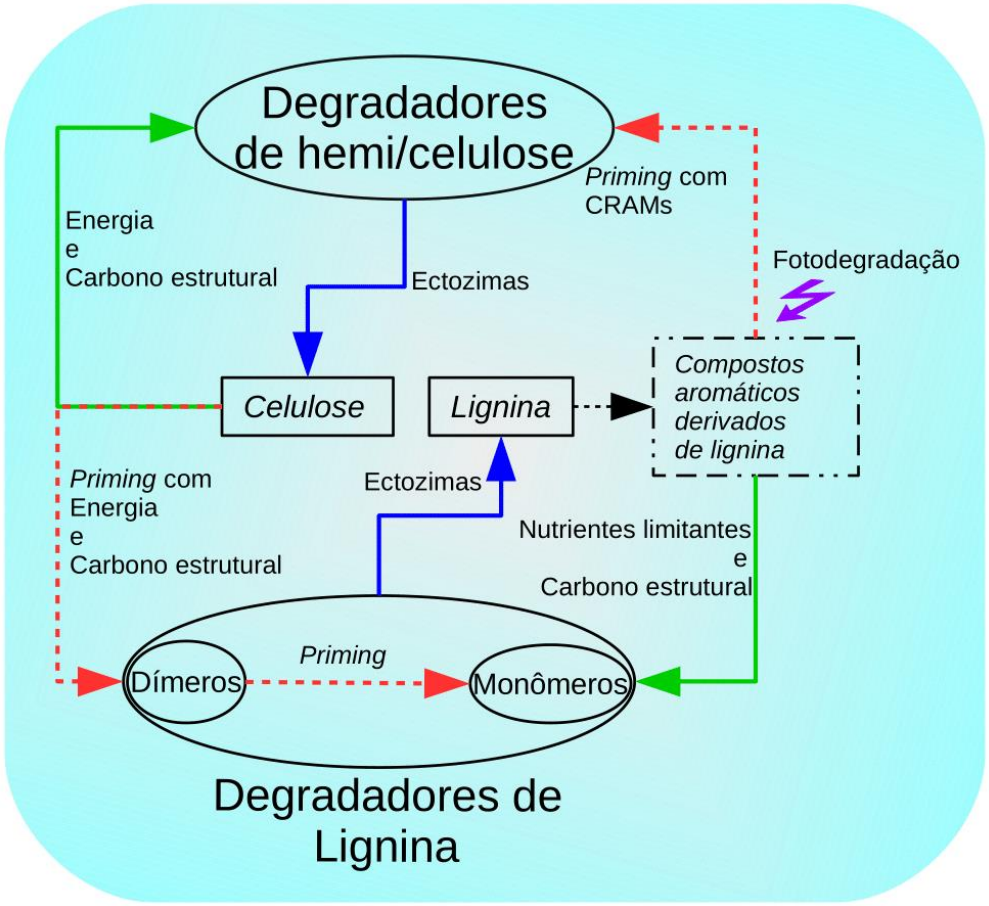

Figura 3.13. Efeito *priming* sobre as comunidades do rio Amazonas e sua interação. Modelo esquemático do efeito de priming do complexo ligno-celulose no rio Amazonas baseado nos genes presentes no AMnrGC. Neste modelo há duas comunidades diferentes degradando lignina, uma que degrada monômeros e outra que utiliza mais dímeros. Essas comunidades se apoiam mutuamente por suas sobras e são sustentadas pelos degradadores da celulose e hemicelulose, que fornecem carbono estrutural e

energia. Os produtos de fotodegradação de compostos aromáticos derivados de lignina representam uma pequena contribuição para a produção de CO_2 (menos de 1%), pois as moléculas alicíclicas ricas em carboxila (CRAM) retroalimentam os degradadores de celulose através do transporte mediado pelo sistema tripartido de transporte de tricarboxilatos (TTT). Setas vermelhas tracejadas significam efeito de *priming*, setas verdes significam *feedback* natural, setas azuis indicam degradação de um composto, seta preta significa a transformação de substratos em produtos. (Fonte: Próprio autor)

Cabe aqui ressaltar ainda que o potencial demonstrado por essa microbiota em degradar material lignocelulósico pode ser altamente aproveitado na indústria têxtil, sucroalcooleira e de papel, que se baseiam na premissa de modificação dos materiais lignocelulósicos por meio de coquetéis enzimáticos ou cepas especiais. Dessa forma, o AMnrGC além de revelar traços importantes da ecologia destes microrganismos, também revela um potencial biotecnológico inexplorado para a indústria que também pode utilizar-se destes mecanismos.

3.5 Conclusão

Usamos 106 metagenomas disponíveis publicamente, de 30 estações da bacia do rio Amazonas, para produzir um catálogo de genes microbianos não redundantes desta bacia (AMnrGC). O presente catálogo contém mais de 3,7 milhões de genes microbianos, com metade deles desconhecidos. O teste de diversidade de k-mers mostrou os metagenomas do rio Amazonas formando um grupo fechado e separado em

comparação aos metagenomas do solo da floresta amazônica, bacias hidrográficas do Canadá e rio Mississippi, reforçando a ideia de que diversidade genética local seja decorrente de processos evolutivos ocorridos nesta região peculiar. Uma estruturação espacial complexa de genes foi observada com clara predominância de genes bacterianos exclusivamente de água doce. O AMnrGC contém a mais extensa biodiversidade genética já compilada para um ecossistema de água doce. Ele revelou importantes *insights* sobre os fluxos de carbono da terra para o oceano, como um modelo de efeito *priming* baseado na interação de comunidades degradadoras de celulose/hemicelulose e comunidades degradadoras de lignina, dividindo os degradadores de lignina em degradadores de monoaris e diaris. Outra contribuição importante é a descoberta do sistema TTT participando na degradação da OM terrestre. O AMnrGC também forneceu uma visão global da diversidade funcional e genética da bacia do rio Amazonas, permitindo futuros estudos envolvendo genes microbianos e seu uso em aplicações biotecnológicas.

Capítulo 4 - Evidências genômicas e o modelo de um potencial efeito *priming* amazônico

4.1 Introdução e objetivos

Assinaturas heterotróficas foram coletadas de bactérias do rio Amazonas (DOHERTY et al., 2017; GHAI; RODRIGUEZ-VALERA; et al., 2011; SATINSKY; CRUMP; et al., 2014; SATINSKY; SMITH; SHARMA; WARD; et al., 2017), um sistema túrbido e rico em material particulado orgânico. Estudos anteriores (GHAI; RODRIGUEZ-VALERA; et al., 2011; SANTOS-JÚNIOR et al., 2017; SATINSKY et al., 2015; TOYAMA et al., 2016, 2017) indicaram que a microbiota do rio Amazonas é amplamente composta por Actinobacteria e Proteobacteria, principalmente dos grupos Polinucleobacter e Methylophilaceae. Esses táxons estão marcadamente relacionados, em lagos, ao estabelecimento de estilos de vida mixotróficos, onde há geração de energia por fototrofia, ao mesmo tempo que há heterotrofia, além de serem capazes de degradar compostos orgânicos complexos (NEWTON et al., 2011). Este ambiente é ocupado principalmente por microrganismos especialistas na degradação de matéria orgânica complexa (SATINSKY et al., 2015).

No capítulo anterior, o catálogo de genes microbianos não redundantes da bacia Amazônica (AMnrGC) revelou uma estratificação no processamento da matéria orgânica pelo tipo de água (doce ou salgada), assim como uma maquinaria bioquímica específica, principalmente formada por glicosil-hidrolases e lacases, para degradação de OM terrestre. Observou-se também uma correlação entre os sistemas tripartido de transporte de tricarboxilatos (TTT) e a degradação de lignina e hemicelulose, o que poderia representar o processamento de fontes alternativas de carbono oriundas da degradação de OM complexa na forma de tricarboxilatos. No modelo de efeito de *priming* proposto no capítulo 3, duas populações diferentes, uma especializada em

degradação de lignina e outra como degradadora de celulose, interagem para acelerar a degradação de carbono orgânico terrestre. Embora as principais funções envolvidas nos processos acima mencionados já tenham sido estabelecidas, a informação do AMnrGC não possui o contexto genômico dessas descobertas. Assim, o conjunto específico de genes que cada tipo de micróbio carrega, a regulação desses genes e a possibilidade de interação entre conjuntos de genes pertencentes a diferentes organismos, inibindo ou promovendo seu crescimento ainda não foi esclarecida. Neste capítulo, abordamos as seguintes questões-chave:

- Os genomas populacionais (PGs) recuperados do rio Amazonas têm adaptações à degradação da OM terrestre?

- O sistema tripartido de transporte de tricarboxilatos (TTT) está acoplado à maquinaria de degradação de lignina/celulose?

- Esses organismos economizam carbono intracelular nos períodos de restrição?

Nossa hipótese é que, como proposto no capítulo anterior, haja duas populações de PGs, uma degradando a lignina e outra, principalmente a celulose. Além disso, como essas características estão correlacionadas ao sistema TTT, devem ser encontradas nos genomas das espécies degradadoras de celulose, a fim de serem iniciadas pelas CRAMs geradas pela fotodegradação de subprodutos da degradação da lignina. A estocagem de carbono é um mecanismo difundido em diversos táxons (BRIGHAM et al., 2011; LENZ; MARCHESSAULT, 2005; QUELAS et al., 2016), e deveria ocorrer mesmo em sistemas com abundantes fontes de carbono, como a Amazônia. Sabe-se que os corpos d'água amazônicos podem diluir as suas águas

durante a estação chuvosa, num pulso de inundação (AFFONSO; BARBOSA; NOVO, 2011; MONTEIRO; PEREIRA; JIMÉNEZ, 2016), que dilue também os nutrientes e, provavelmente, promove a competição microbiana. Nesse sentido, levantamos a hipótese de que organismos com taxas de crescimento mais altas, com mais necessidades de carbono, provavelmente deverão armazená-lo.

4.2 Material e métodos

Neste capítulo utilizou-se os metagenomas presentes no Apêndice 1. Estes metagenomas foram produzidos por outros estudos, para mais detalhes vide seção 3.2 da presente tese. As *reads* foram previamente tratadas e montadas como explicado nos itens 3.2 e 3.2.1 da presente tese. Utilizou-se o mesmo sistema de divisão da bacia amazônica em 5 seções.

4.2.1 Binning *de sequências e construção de PGs*

As *reads* já filtradas com relação à qualidade e comprimento foram mapeadas contra os *contigs* obtidos na montagem por meio do uso dos programas BWA versão 0.7.12-r1039 (LI, H.; DURBIN, 2009) e SamTools versão 1.3.1 (LI, H. et al., 2009). Os *contigs* de cada grupo foram então agrupados com auxílio do programa Metabat versão 2.12.1 (KANG et al., 2015) com as configurações fixadas no modo "*superspecific*". Os *contigs* com valores extremos (*outliers*) em termos de composição de k-mers e conteúdo GC foram eliminados usando o programa RefineM versão 0.0.23 (PARKS et al., 2017) com configuração padrão.

As *bins* refinadas foram avaliadas quanto à sua completude, contaminação e heterogeneidade da cepa utilizando-se o *software* CheckM versão 1.0.11 (PARKS et al., 2015) em dois módulos distintos: "*lineage_wf*" e "*ssu_finder*". Os 16S rRNA genes extraídos de cada *bin* foram mapeados contra o banco de dados SILVA SSU Ref NR99 versão 123 (QUAST et al., 2013; YILMAZ et al., 2014) com o programa Usearch versão 9.2 (EDGAR, 2010). Utilizou-se como parâmetro de corte uma identidade mínima de 97% e cobertura maior que 70%. Os *contigs* foram removidos de uma *bin* se eles possuíssem mais de 98% de identidade com uma correspondência no banco de dados SILVA proveniente de classificação taxonômica incongruente com a assinatura taxonômica obtida com o módulo "*tree*" do programa CheckM (PARKS et al., 2015).

Como sugerido por outros autores (BOWERS et al., 2017; PARKS et al., 2017) somente *bins* com uma completude superior a 50%, contaminação inferior a 10%, qualidade estimada geral superior ou igual a 50% (definida por Parks et al. (PARKS et al., 2017) como a completude – 5 x contaminação) e sendo composto por menos de 500 *contigs* cujo N50 seja superior ou igual a 10 Kpb seguiu nos processos posteriores.

A identidade média de aminoácidos (AAI) foi calculada dentre os PGs por meio do *software* CompareM versão 0.0.13 (PARKS, 2016) com a função "*aai_wf*". PGs com AAI ≥ 99,5% foram considerados redundantes. O par de PGs nessa condição foi avaliado separadamente e o mais completo com menor contaminação e heterogeneidade de cepa foi mantido nos processos seguintes. Os PGs foram depositados no *European Nucleotide Archive* (ENA) no projeto PRJEB25176.

4.2.2 Análise de similaridade de PGs

A identidade média de nucleotídeos (ANI) foi calculada usando a ferramenta FastANI (JAIN et al., 2017) com o presente conjunto de dados de PGs contra:

- 957 PGs da expedição *TARA-Oceans* (EREN, 2017);
- 18 genomas do grupo Verrucomicrobia de diversos reservatórios de água doce (CABELLO-YEVES; GHAI; et al., 2017);
- 35 PGs do lago Baikal (CABELLO-YEVES et al., 2018);
- 2 PGs de *Synechococcus* de água doce (CABELLO-YEVES; HARO-MORENO; et al., 2017);
- 3.087 genomas UBA do banco de dados de taxonomia genômica – GTDB (PARKS et al., 2018); e
- 7.520 genomas referência de alta-qualidade do NCBI.

Valores de ANI maiores que 96,5% e uma fração alinhada (AF) maior que 60% foram mantidos e a probabilidade do par de medidas AF e ANI significar cepas coespecíficas foi calculada seguindo as recomendações de Varghese et al. (VARGHESE et al., 2015). Apenas genomas com um $p \geq 0,9$ foram reportados.

4.2.3 Estimativa da abundância dos PGs

As *reads* previamente filtradas por qualidade dos 106 metagenomas foram mapeadas contra os PGs refinados utilizando-se o programa BWA versão 0.7.12-r1039 (LI, H.; DURBIN, 2009) e o programa sambamba versão 0.6.6 (TARASOV et al.,

2015). Após o mapeamento, os arquivos BAM foram filtrados com um *script* caseiro em linguagem PERL (uma cortesia de Amin Madoui – Genescope, França) para filtrar as correspondências que tivessem uma identidade mínima de 97% e cobertura mínima da *read* de 80%. A abundância de cada PG por metagenoma foi dada como o número de *reads* recrutadas por ele, enquanto a abundância por seção do rio Amazonas e das frações (vida livre e associado a partículas) foi dado como a média arredondada das abundâncias por metagenoma. As abundâncias dos PGs foram então normalizadas para *Z-scores* com a seguinte fórmula:

$$z = \frac{(x - \mu)}{\sigma}$$

Onde z representa o *Z-score*, μ representa a média das abundâncias por coluna da matriz (isto é, por seção do rio) e σ representa o desvio padrão obtido no cálculo da média.

Com os valores normalizados construiu-se um *heatmap* com o auxílio dos seguintes pacotes implementados em linguagem R: ggplot2 (WICKHAM, 2009), gplots (WARNES et al., 2016) e ColorBrewer (NEUWIRTH, 2014).

4.2.4 Predição de genes e anotação

Os genes de cada *bin* foram preditos utilizando o programa Prodigal versão 2.6.3 (HYATT et al., 2010) no modo “normal”. A anotação de genes foi feita por meio do programa Prokka versão 1.11 (SEEMANN, 2014).

Os genes previamente preditos e anotados foram submetidos à anotação com os bancos de dados referência: KEGG versão 2015-10-12 (KANEHISA et al., 2012, 2017), COG versão 2014 (TATUSOV et al., 2003) e UniProtKB versão 2016-08 (UNIPROT CONSORTIUM, 2015) utilizando-se o algoritmo Blastp implementado no programa Diamond versão 0.9.17 (BUCHFINK; XIE; HUSON, 2014) com uma cobertura de busca maior que 50%, identidade mínima de proteínas de 45% e um e-*value* máximo de $1e^{-5}$ com *Score* maior que 50.

Também se anotou os genes encontrados com perfis estatísticos do tipo HMM por meio do programa HMMSearch versão 3.1b1 (EDDY, 2011), com os seguintes bancos de dados: dbCAN versão 5 (YIN et al., 2012), PFAM versão 30 (FINN et al., 2016) e eggNOG versão 4.5 (HUERTA-CEPAS et al., 2016). Os resultados desta busca foram filtrados para um e-*value* < $1e^{-5}$ e probabilidade posterior ao alinhamento de resíduos > 0.75, desconsiderando-se os domínios sobrepostos. As anotações dos PGs estão disponíveis no link: 10.5281/zenodo.1484510.

4.2.5 Inferência filogenética

A inferência filogenética foi realizada similarmente ao descrito anteriormente por outros autores (PARKS et al., 2017). Resumidamente, a árvore filogenética construída com os PGs foi inferida a partir da concatenação de 43 famílias de marcadores proteicos (Apêndice 7). As proteínas foram identificadas e alinhadas usando HMMER versão 3.1b1 (EDDY, 2011), onde as colunas representadas por < 50% dos

táxons ou sem um aminoácido comum a ≥ 25% dos táxons foram removidas. Os marcadores foram presentes em cópia única em ≥ 76,78% dos PGs.

O multi-alinhamento foi feito concatenando-se os marcadores de nossos PGs e o banco de dados dos marcadores de-replicados dos genomas do banco GTDB (PARKS et al., 2018) e dos genomas de referência RefSeq/GenBank versão 76 (PARKS et al., 2017). Ele foi recuperado com o auxílio da opção "*tree_qa*" do programa CheckM (PARKS et al., 2015). As árvores foram inferidas com o programa FastTree versão 2.1.7 (PRICE; DEHAL; ARKIN, 2009) sob os modelos JTT + CAT e valores de suporte determinados usando 100 pseudo réplicas de *bootstrap* não paramétricas. A renderização de árvores em formato *Newick* foi feita por meio do programa Dendroscope versão 3 (HUSON et al., 2007).

4.2.6 Análise funcional

As vias metabólicas foram previstas para cada PG usando um mapeamento de KOs, obtido na anotação dos genes com o banco de dados KEGG, através do programa MinPath versão 1.4 (YE; DOAK, 2009). As rotas mantidas pelo MinPath foram avaliadas enquanto sua completude, calculada pela divisão do número total de famílias de proteínas envolvidas na via correspondente naquele PG pelo número total de famílias envolvidas no registro KEGG. Esses valores foram usados para gerar um gráfico de bolhas com o pacote ggplot2(WICKHAM, 2009), implementado em linguagem R.

4.2.7 Metabolismo do carbono

Para investigar melhor o metabolismo do carbono, procuramos os PGs contendo as vias para:

(i) degradação da matéria orgânica terrestre: lacases (PF02578) e anotações coerentes de glicosil-hidrolases (GH) obtidas com bancos de dados dbCAN e PFAM;

(ii) genes das vias de degradação de compostos aromáticos derivados de lignina (Apêndice 3);

(iii) transporte tripartido de tricarboxilatos (TTT): TctA (PF01970), TctB (PF07331) e TctC (PF03401);

(iv) via de síntese poli-hidroxi-butirato e poli-hidroxi-alcanoato (PHB/A): phaA (anotado simultaneamente como PF00108 e COG0183), phaB (anotado simultaneamente como PF00106 e COG1028), phaC (anotado simultaneamente como PF07167 e COG3243), phaE (PF09712) e phaR (anotados simultaneamente como PF05233 e COG5394).

4.2.8 Expressão gênica

Trabalhos anteriores (SATINSKY; CRUMP; et al., 2014; SATINSKY et al., 2015) disponibilizaram 108 metatranscriptomas das comunidades procarióticas das águas do rio Amazonas e de sua pluma na plataforma iMicrobe (https://www.imicrobe.us/) sob o projeto CAM_P_0001194. No entanto, estes

metatranscriptomas foram feitos apenas para a região compreendendo desde a seção à jusante até o oceano, não incluindo seções à montante de Manaus. Também é interessante mencionar que eles cobriram ambas as frações testadas por metagenômica (microrganismos de vida livre e associados a partículas). Uma vez que metatranscriptomas de seções à montante não estão disponíveis, foram avaliados apenas os PGs, contendo vias avaliadas completas, advindos das seções a partir da jusante da cidade de Manaus.

As *reads* dos metatranscriptomas passaram por um pré-tratamento de qualidade similar ao exposto no item 3.2.1. O mapeamento dos transcritos contra os genes anotados de cada PG foi feito utilizando-se o programa BWA versão 0.7.12-r1039 (LI, H.; DURBIN, 2009) e o programa sambamba versão 0.6.6 (TARASOV et al., 2015). A expressão gênica foi estimada como o número de transcritos por milhão (TPM) calculado com o *software* eXpress versão 1.5.137 (ROBERTS; PACHTER, 2012), desconsiderando-se a correção de *bias*. Os gráficos foram feitos utilizando a média dos TPM calculados por seção e fração.

4.3 Resultados

4.3.1 Binning *de genomas populacionais (PGs) a partir das co-montagens do rio Amazonas*

Nosso conjunto de dados original continha 106 metagenomas de 30 estações distintas (Figura 3.1). Foram obtidas 30 co-montagens com comprimento mínimo de *contigs* de 1 Kpb, totalizando um comprimento de 5,54 Gbp (Apêndice 4). As co-

montagens foram agrupadas em 1443 PGs putativos usando o programa MetaBAT. Esses PGs foram avaliados pelo CheckM e apenas 54 PGs (3,74%) foram selecionados após filtragem por N50 mínimo de 10 Kpb, máximo de 500 *contigs* e qualidade geral mínima de 50%, com completude mínima e contaminação máxima de 50% e 10%, respectivamente (Apêndice 8).

A AAI revelou três PGs redundantes (Tabela 4.1), que foram eliminados da análise e representados, em nosso conjunto de dados, pelo PG mais completo e menos contaminado dessa linhagem.

Foi possível recuperar 25 genomas completos da seção à montante de Manaus, 9 da seção à jusante de Manaus, 6 do estuário, 9 da pluma e 2 do oceano costeiro (Figura 4.1). A composição taxonômica dos mesmos revelou predomínio de Proteobacteria (39% dos genomas recuperados), Bacteroidetes (15,7%) e bactérias não classificadas (15,7%). As cianobactérias foram representadas por 4% dos nossos genomas recuperados. Apenas dois genomas de arquéias foram recuperados: um pertencente a Thaumarchaeota (da seção à montante) e outro de Euryarchaeota (da pluma).

Tabela 4.1. Identidade média de aminoácidos (AAI) dos PGs recuperados do rio Amazonas. Para remover os PGs redundantes do nosso conjunto de dados, foi utilizado o *software* CompareM. Inferiu-se a média dos valores de AAI (μ_{AAI}) por par de genomas (PG A e B), assim como os desvios-padrão (σ_{AAI}) e suas frações ortólogas (OF). (Fonte: Próprio autor)

PG A	# Genes A	PG B	# Genes B	# ortólogos	μ_{AAI}	σ_{AAI}	OF
AM_1413	2083	AM_1606	2088	1961	99,98	0,32	94,14
AM_1804	2154	AM_2804	2189	2092	99,93	0,48	97,12
AM_0750	5030	AM_0616	5208	4600	99,84	1,19	91,45

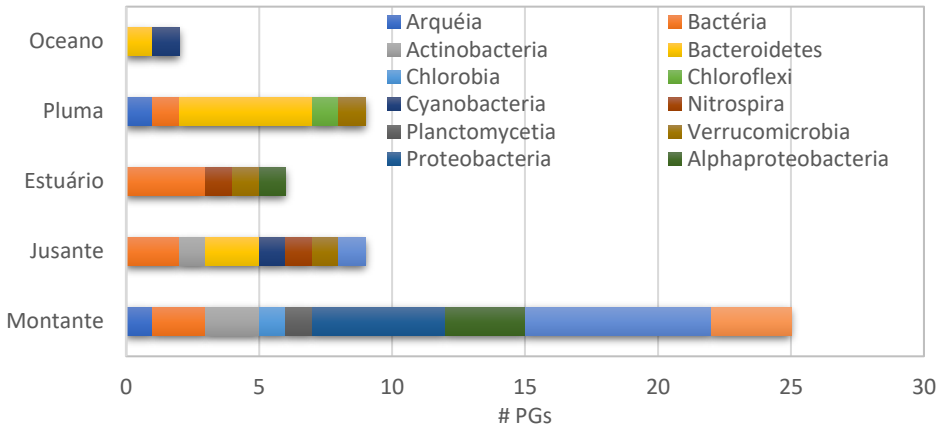

Figura 4.1. Genomas populacionais recuperados por seção do rio Amazonas. (Fonte: Próprio autor)

Nosso conjunto de dados foi composto por 49% de genomas de alta qualidade e 51% de genomas de qualidade média, determinados conforme recomendado por outros autores (BOWERS et al., 2017). O teor médio de GC foi de 50,25 ± 9,62%, com tamanho variando de 0,53 a 7,87 Mbp.

Os genomas populacionais tem sua completude inferida a partir do número de genes que ocorrem em cópia única na maioria dos genomas bacterianos até hoje isolados. No caso de se encontrar várias cópias desses genes em um genoma populacional, sugere-se que haja uma "contaminação". A natureza da contaminação, pode ser coespecífica (identificada por um alto grau de heterogeneidade da cepa), ou pode ser devido a uma divergência na origem dos *contigs*, o que é mais preocupante. Em todos os casos, o máximo de contaminação neste conjunto de dados foi determinado em 6,78%, com 52,9% de PGs apresentando uma contaminação abaixo de 1% ou igual a zero.

Um resumo das principais características dos PGs apresentados neste estudo é mostrado na Tabela 2 e descrito com mais detalhes no Apêndice 9. Embora não seja muito preciso, o número de tRNAs e aminoácidos codificados por eles pode ser usado para predizer a integridade do genoma. Nesse sentido, os PGs aqui apresentados apresentaram 24 a 55 genes codificando tRNAs para 15 a 20 aminoácidos diferentes, revelando seu bom nível de completude.

Tabela 4.2. Principais informações dos genomas produzidos neste estudo. Os genomas montados a partir dos metagenomas (PGs) foram descritos em termos de seção de onde vieram, composição (teor GC%), tamanho em Mpb, completude (C), contaminação (Cx), qualidade e taxonomia dos marcadores presentes de acordo com a avaliação do programa CheckM. (Fonte: Próprio autor)

PG	Seção	GC	Tam. (Mpb)	C	Cx	Qualidade	Taxonomia
AM_0118	Montante	37.77	1.35	81.87	0	Média	Streptomycetaceae
AM_0219	Montante	36.37	1.05	76.13	0.71	Média	*Methylopumilus* sp1
AM_0226	Montante	63.28	2.99	98.79	0.34	Alta	Rhizobiales
AM_0228	Montante	56.89	2.92	97.56	2.42	Alta	Burkholderiales
AM_0233	Montante	53	2.98	96.26	0	Alta	Proteobacteria
AM_0240	Montante	60.07	3.02	91.32	0.89	Alta	Bactéria
AM_0244	Montante	41.41	2.10	93.23	0.76	Alta	Gamma-proteobacteria
AM_0256	Montante	53.45	2.66	88.24	2.15	Média	Bactéria
AM_0268	Montante	52.62	2.91	86.69	0.42	Média	Bactéria
AM_0275	Montante	52.7	2.41	72.7	1.41	Média	Rhizobiales
AM_0466	Montante	41.13	1.03	62.5	0	Média	Bactéria
AM_0507	Montante	36.3	1.00	79.25	0.64	Média	Beta-proteobacteria
AM_0510	Montante	62.61	2.00	98.08	1.1	Alta	Bactéria
AM_0519	Montante	65.04	4.82	96.74	0.07	Alta	Gamma-proteobactéria
AM_0528	Montante	48.07	2.08	93.43	0.34	Alta	Rhizobiales
AM_0546	Montante	58.47	1.47	70.69	0.07	Média	Actinobacteria
AM_0608	Montante	38.52	3.23	99.28	1.07	Alta	Moraxellaceae
AM_0615	Montante	36.39	1.56	99.51	0.97	Alta	Thaumarchaeota
AM_0616	Montante	59.76	5.51	98.66	0.59	Alta	Beta-proteobactéria
AM_0619	Montante	42.57	2.83	91.18	2.52	Alta	Bactéria
AM_0621	Montante	56.53	7.87	90.33	0	Alta	Bactéria
AM_0630	Montante	52.99	2.89	95.43	2.17	Alta	Proteobacteria
AM_0643	Montante	56.96	2.89	91.49	2.2	Alta	Burkholderiales

AM_0729	Montante	53.85	2.45	84.4	1.26	Média	Bactéria
AM_0764	Montante	50.2	3.89	86.77	1.1	Média	Bactéria
AM_0832	Jusante	44.61	3.47	96.39	1.64	Alta	Bacteroidetes
AM_0849	Jusante	65.31	2.11	86.21	6.78	Média	Actinobactéria
AM_0854	Jusante	52.59	0.53	73.22	0	Média	Bactéria
AM_0876	Jusante	52.57	4.47	90.28	2.88	Alta	Bactéria
AM_0902	Jusante	37.07	4.06	96.49	0	Alta	Cyanobacteria
AM_0936	Jusante	62.13	2.69	86.63	4.45	Média	Bactéria
AM_1003	Jusante	59.74	6.26	95.93	0.91	Alta	Beta-Proteobacteria
AM_1104	Jusante	59.11	3.00	78.96	3.08	Média	Bactéria
AM_1111	Jusante	38.28	2.82	90.95	6.68	Média	Bacteroidetes
AM_1205	Estuário	47.76	0.71	65.39	0	Média	Bactéria
AM_1312	Estuário	61.53	2.17	85.71	1.82	Média	Bactéria
AM_1409	Estuário	39.04	0.61	70.98	1.72	Média	Bactéria
AM_1503	Estuário	67.21	2.09	63.57	0.73	Média	Bactéria
AM_1603	Estuário	63.07	4.54	98.42	1.19	Alta	Sphingomonadales
AM_1606	Estuário	55.47	2.24	96.4	0.61	Alta	Bactéria
AM_1801	Pluma	45.14	1.20	67.56	0	Média	Euryarchaeota
AM_1811	Pluma	43.37	2.27	85.47	3.72	Média	Bactéria
AM_2104	Pluma	60.62	2.11	84.85	4.55	Média	Bactéria
AM_2116	Pluma	46.6	0.69	52.31	0	Média	Bactéria
AM_2124	Pluma	40.54	1.53	87.19	1.49	Média	Flavobacteriaceae
AM_2202	Pluma	47.45	1.80	97.31	0	Alta	Bactéria
AM_2207	Pluma	43.09	1.79	85.25	1.19	Média	Flavobacteriaceae
AM_2208	Pluma	53.01	1.62	92.99	0.36	Alta	Bactéria
AM_2324	Pluma	37.43	3.41	90.78	0.99	Alta	Bacteroidetes
AM_2502	Oceano	39.43	1.43	89.99	1.18	Média	Flavobacteriaceae
AM_2804	Oceano	33.5	3.29	93.56	0.27	Alta	Cyanobacteria

4.3.2 Similaridade de genomas

Através da taxonomia completa inferida pelo programa CheckM no modo "*tree"* aliada à árvore filogenética calculada pelo programa FastTree e a ANI inferida (Apêndices 8d e 10) foi possível identificar pelo menos 10 PGs provenientes de linhagens coespecíficas: *Richelia intracellularis* A (AM_2804), *Trueperella pyogenes* (AM_0546), *Acinetobacter junii* (AM_0608), *Methylopumilus sp1* (AM_0507 e AM_0219), *Coccinistipes* sp. (AM_2208), *Sphingobium* sp2 (AM_1603), o gênero desconhecido UBA11236 (AM_1606), *Xanthomonas fuscans* subsp. *fuscans* (AM_0519) e uma espécie de *Rokubacteria* GWA2-73-35 sp1 (AM_2207).

Nenhum PG da expedição *TARA-Oceans* (EREN, 2017) ou qualquer outro ambiente de água doce (CABELLO-YEVES et al., 2018; CABELLO-YEVES; GHAI; et al., 2017; CABELLO-YEVES; HARO-MORENO; et al., 2017) pode ser agrupado em cepas coespecíficas com os PGs isolados neste estudo. Por volta de 80% dos genomas aqui reportados foram classificados como microrganismos taxonomicamente desconhecidos ou novos, isto pois, não possuem uma classificação forte o suficiente para serem tidos como tendo sido descritos ou verificados anteriormente. A resolução mais acurada o possível para os outros PGs foi apenas em nível de reino em 14% dos casos, em nível de filo em 14% dos casos, em nível de classe em 8% dos casos, de ordem em 16% dos casos, de família em 26% deles e de gênero em apenas 4%. Isto mostra que há um alto grau de novidade e diversidade até então desconhecida associado ao presente conjunto de dados.

Os PGs pertencentes a cepas coespecíficas foram comparados (Apêndice 10) mostrando maior completude, na maioria dos casos, excluindo-se AM_0519 e AM_0546, que apresentaram menor número de proteínas. A diferença mínima em GC% e densidade de codificação de proteína reforçou os resultados apresentados pela ANI. Grupos como Cyanobacteria, que apresentam um decaimento do genoma ativo, também apresentaram menores densidades de codificação. A maior variação ocorreu entre as cepas coespecíficas de *Sphingobium* sp2, onde AM_1603 apresentou 875 proteínas a mais que a referência (UBA11613), e quando o verificado com a cepa representativa (UBA5915) ainda apresentou melhor estatística "p". Enquanto o genoma UBA5915 é 96,48% completo com 4570 proteínas, o PG AM_1603 apresenta 98,42% de integridade contendo 38 proteínas a mais.

4.3.3 Abundância dos PGs

A abundância de genomas por seção do rio Amazonas (Figura 4.2) revelou uma clara preferência dos PGs por seções e frações específicas, reforçando o esquema de isolamento de genomas por co-montagem.

Analisando-se os padrões dos dendrogramas laterais, a segregação de seções, com amostras de água doce segregadas de amostras de água salobra, valida a divisão do rio adotada aqui. Apesar disso, amostras da pluma de microrganismos associados a partículas foram segregadas conjuntamente com amostras oceânicas. As seções à jusante e estuário formam uma ramificação única, que deriva a partir de um nó em que se insere a seção à montante.

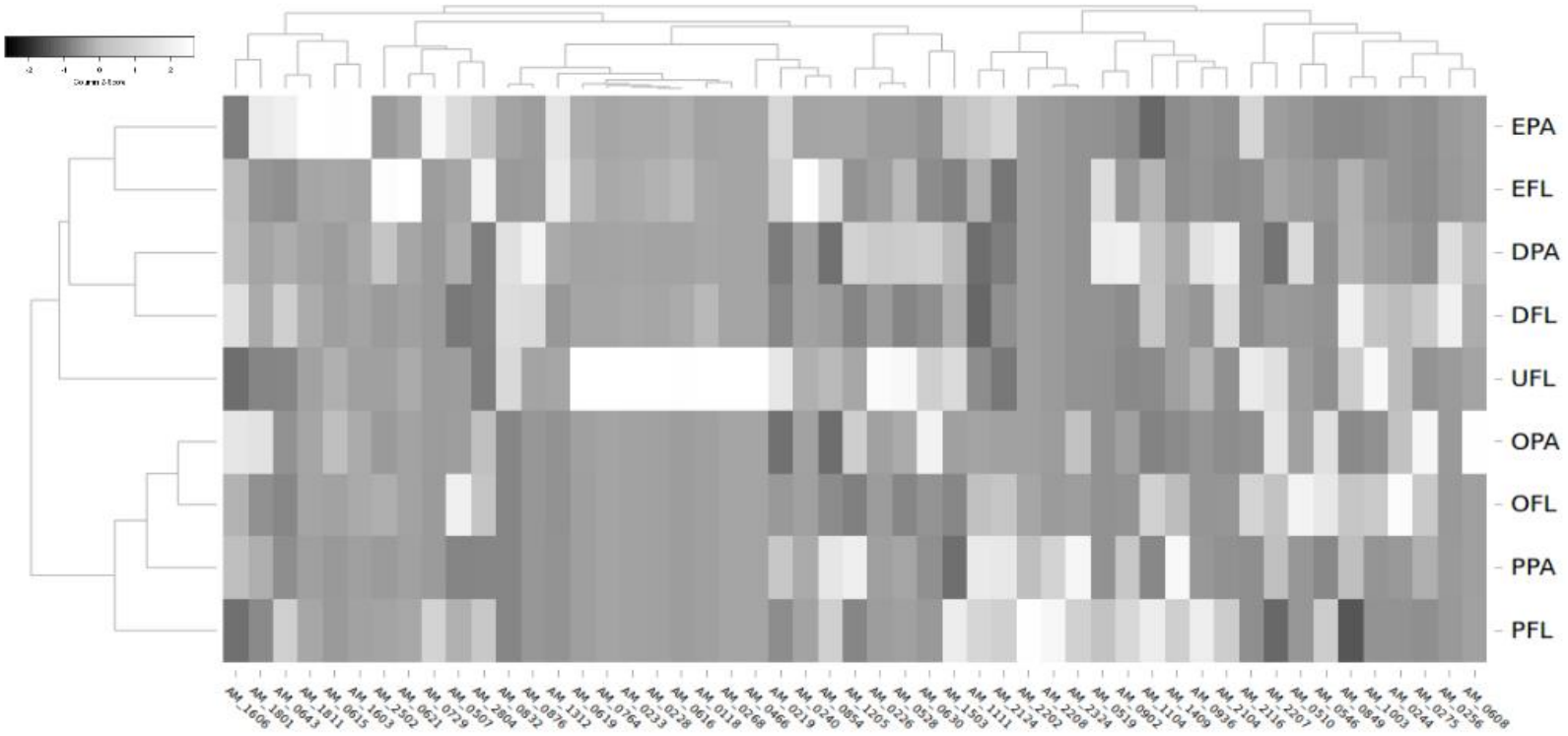

Figura 4.2. Abundância dos diferentes PGs por seção do rio Amazonas. A abundância média de cada um dos PGs foi calculada para os metagenomas provenientes de diferentes seções (U - Montante, D - Jusante, E - Estuário, P - Pluma, O - Oceano) e frações (vida livre - FL e associada a partículas - PA). A escala de cores é baseada no *Z-score* calculado por colunas e o agrupamento foi feito usando-se a distância euclidiana com ligação completa. (Fonte: Próprio autor)

4.3.4 Potencial metabólico

O potencial metabólico (Figura 4.3) foi verificado utilizando-se a completude da via bioquímica avaliada. As vias de fixação do carbono e da fosforilação oxidativa foram utilizadas como controles de integridade do genoma, uma vez que devem ser mantidas na maioria dos organismos. Outro marcador foi a via de fotossíntese mediada por proteínas de antena para avaliar o potencial de metabolismo fotossintético dos genomas, validando aqueles pertencentes à Cyanobacteria (AM_2804 e AM_0902).

Microrganismos tropicais, geralmente, possuem um maquinário complexo associado à degradação de compostos aromáticos (DAS; CHANDRAN, 2011; SEO; KEUM; LI, 2009). A degradação de xenobióticos mediada pelo citocromo P450 não foi o principal mecanismo associado à degradação dos mesmos (Figura 4.3). De fato, o isolamento de xenobióticos por grupos de substratos revelou importantes padrões de degradação. Foi observado que a dioxina pode ser degradada apenas por *Sphingobium* sp2 (AM_1603) e Rhizobiales (AM_0275), bem como, naftaleno e aminobenzeno podem ser degradados principalmente por Rhizobiales (AM_0275). Cloroalcano/alcenos são substratos de difícil degradação, Cyanobacteria (AM_0902 e AM_2804) e Chlorobiaceae (AM_0510) estão mais propensos a degrada-los.

Figura 4.3. Perfil funcional dos PGs da bacia do rio Amazonas. As vias bioquímicas de fixação de carbono e fosforilação oxidativa foram utilizadas como controles de integridade do genoma para superar as diferenças de completude. Vias de degradação de diversos compostos aromáticos, considerados xenobióticos, foram avaliadas por seus substratos. A completude da via foi mostrada proporcionalmente ao tamanho do círculo, bem como o táxon a que pertence cada PG pela cor de seu círculo. (Fonte: Próprio autor)

Metabolismos alternativos são muito importantes para ligar a microbiota que vive em diferentes habitats, por exemplo, microrganismos de ambientes aeróbicos e anaeróbicos. Neste caso, alguns subprodutos metabólicos podem desempenhar papéis importantes no fluxo de nutrientes, como o metano no fluxo de carbono. O metano é uma fonte alternativa de carbono gerada por diversos gêneros metanogênicos que vivem em condições anaeróbicas, como *Methanobacterium, Methanoculleus, Methanocorpusculum, Methanobacterium, Methanosaeta* (ZIGANSHIN et al., 2016).

Não foram encontrados organismos metanogênicos, seguindo a principal sonda para sua detecção, o gene da subunidade α da coenzima-metil-redutase α (mcrA). Foi possível observar que a metilotrofia está relacionada a um conjunto reduzido de genes do metabolismo C1 na arqueia metanotrófica AM_0615 da seção à montante de Manaus, que contém o conjunto completo de genes da metano monooxigenase (pmoA-C), uma sonda para esse tipo de organismos (MCDONALD, I R; MURRELL, 1997). Outros organismos metanotróficos de origem bacteriana também puderam ser identificados por meio do uso desta sonda, mostrando que AM_0621 (Planctomycetes), AM_0876 (bactéria desconhecida), AM_1503 (Opitutaceae) e AM_1811 (Opitutaceae) também podem estar envolvidos nesta via. Os organismos metanotróficos atuam no *turnover* de compostos de carbono único que afetam o fluxo de carbono global.

O fluxo de nitrogênio sobre o ecossistema da Amazônia foi verificado por meio da completude do metabolismo de nitrogênio. Genes envolvidos com a oxidação de amônio anaeróbico não foram observados em nosso conjunto de genomas. O potencial para nitrificação foi observado apenas no genoma AM_0615, uma Thaumarchaeota capaz de oxidar o metano/amônio por meio da metano monooxigenase particulada.

Genes envolvidos com a fixação de nitrogênio foram observados em PGs isolados de todas as seções, exceto no estuário. Os táxons envolvidos principalmente no metabolismo de nitrogênio foram Cyanobacteria, Chlorobia e Thaumarchaeota.

O metabolismo do enxofre foi identificado em 18 PGs, capazes de captar o sulfato extracelular e reduzi-lo pela via assimilatória. Esta via é predominante nos PGs provenientes das seções à montante (19,6%) e jusante de Manaus (7,8%). Os principais táxons envolvidos no uso e reciclagem de enxofre no ecossistema amazônico são bactérias desconhecidas (22,2%), cianobactérias (16,7%), beta-proteobactérias (22,2%) e outros táxons (38,9%).

O metabolismo do fosfonato e fosfinato mostra o potencial de produtos naturais contendo ligações carbono-fósforo, produzidas principalmente por Actinobacteria, que possuem um alto nível de estabilidade e são capazes de substituir estericamente os ésteres e carboxilatos de fosfato (PECK; GAO; VAN DER DONK, 2012). Apesar disso, poucas enzimas foram encontradas e diferentemente de Actinobacteria, o principal táxon responsável por essa função no rio Amazonas, parece ser o grupo das Nitrospiraceae. Apenas os genomas AM_0226, AM_1312 e AM_1104 possuem enzimas pertencentes a essa via. A enzima fosfonoacetaldeído hidrolase está presente em todos esses genomas. O gene da fosfoenolpiruvato mutase está presente apenas em AM_1312. Estes organismos possuem no máximo duas enzimas desta via, sugerindo um papel potencial para seus produtos intermediários. PGs pertencentes às seções à montante de Manaus e estuário foram os principais responsáveis por esta via.

4.3.5 A degradação de celulose e lignina está acoplada nos PGs

A degradação da matéria orgânica terrestre é uma função importante que ocorre em duas etapas: degradação da lignina mediada por lacases e degradação da celulose mediada por famílias específicas de glicosil-hidrolases, nomeadamente GH1, GH3, GH5, GH6, GH8, GH9, GH12, GH45, GH48, GH51 e GH74 (SUKHARNIKOV et al., 2011). Mais da metade dos PGs (52,9%) não possuem a capacidade de degradar a matéria orgânica terrestre, devido à ausência de genes celulolíticos e lignolíticos (Figura 4.4). Lacases estão presentes em todos os táxons, exceto Bacteroidetes. Pela análise desta diversidade gênica (mínimo de duas famílias de proteínas apresentando duas variantes proteicas), identificamos AM_0519 (*Xanthomonas fuscans* subsp. *aurantifolii*), AM_0876 e AM_0936 (ambas bactérias desconhecidas), e AM_1603 (*Sphingobium* sp2) como os PGs mais propensos à degradação da matéria orgânica terrestre. Verificou-se um acoplamento entre a presença de lacases e GHs nos genomas. Todos os genomas apresentando potencial de degradação de lignina, também apresentaram alguma família de glicosil-hidrolase relacionada à degradação da celulose. Além disso, há poucos genomas de degradadores exclusivos de celulose (~ 20%).

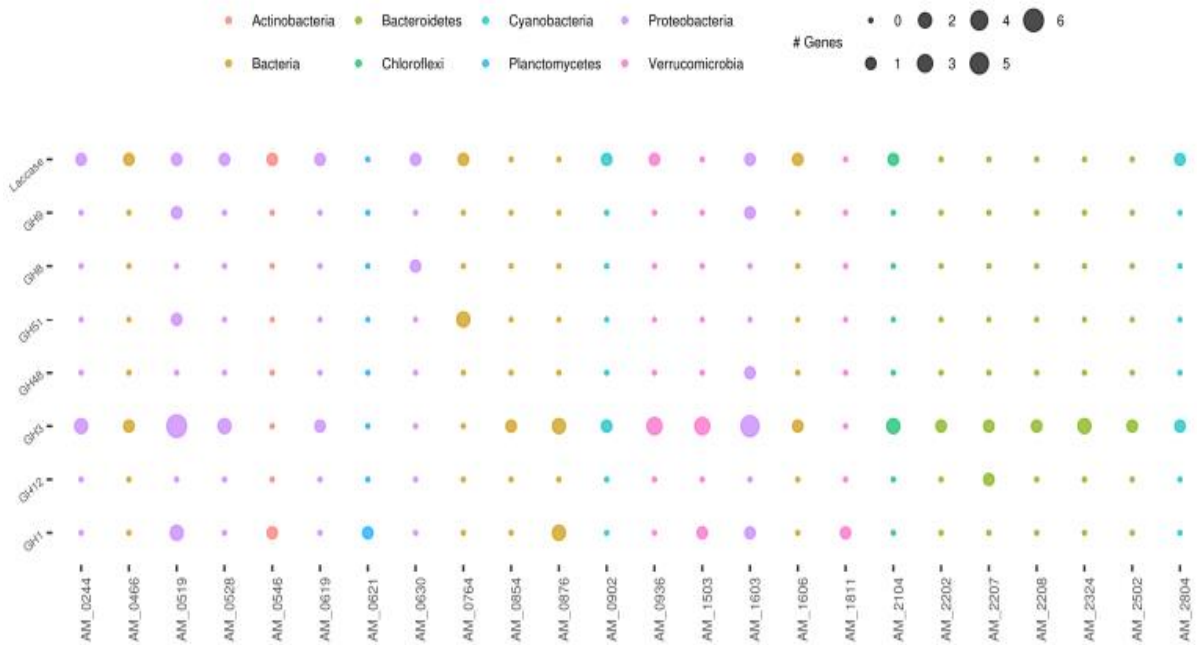

Figura 4.4. Perfil dos genomas populacionais contendo genes para oxidação de lignina e degradação de celulose. As principais famílias de glucosil-hidrolases (GHs) que se sabe terem lugar nesta importante via foram pesquisadas em cada um dos PGs e o número de genes por família de proteínas é mostrado. A completude da via foi mostrada proporcionalmente ao tamanho do círculo, bem como o táxon a que pertence cada PG pela cor de seu círculo. (Fonte: Próprio autor)

Após a oxidação da lignina formam-se compostos menores de natureza aromática que tem que ser internalizados via sistemas de transporte transmembranar. Uma avaliação quanto aos mecanismos deste transporte foi realizado e é mostrado na Tabela 4.3.

Tabela 4.3. Transporte de derivados aromáticos de lignina. As células mostrando um número de genes maior que zero foram sombreadas. (Fonte: Próprio autor)

Gene	Ortólogo	AM_0226	AM_0228	AM_0233	AM_0275
PcaX_NBD	CAC49876.1	0	1	0	0
CouS_NBD	CAE27232.1	1	3	4	1
CouT_TMD_NBD	CAE27233.1	1	2	3	0
Gene	**Ortólogo**	**AM_0616**	**AM_0630**	**AM_0643**	**AM_0750**
PcaX_NBD	CAC49876.1	0	0	1	0
CouS_NBD	CAE27232.1	1	5	3	1
CouT_TMD_NBD	CAE27233.1	0	2	2	0
Gene	**Ortólogo**	**AM_0849**	**AM_0902**	**AM_1003**	
PcaX_NBD	CAC49876.1	0	0	1	
CouS_NBD	CAE27232.1	1	1	0	
CouT_TMD_NBD	CAE27233.1	0	0	1	

Dos PGs que apresentaram mecanismos de internalização, apenas dois deles (AM_0630 e AM_0902) poderiam oxidar a lignina. Dessa forma, verifica-se que talvez o mecanismo de oxidação de lignina seja acoplado ao consumo de celulose, entretanto, desacoplado ao consumo da própria lignina. Os genes relativos ao processamento dos compostos aromáticos produzidos a partir da degradação de lignina também foram avaliados. Na Tabela 4.4 é possível observar que apenas outros dois PGs (AM_0519 e AM_1603), que possuíam a capacidade de oxidar lignina, mostraram-se capazes de metabolizar compostos aromáticos derivados dela. Os outros PGs que se mostraram

capazes de degradar estes compostos não possuíam em seus genomas genes de degradação de celulose ou oxidação de lignina. Isto sugere que a oxidação de lignina está, provavelmente, desacoplada de seu consumo.

Tabela 4.4. Degradação de compostos aromáticos derivados de lignina. O número de genes por PG verificado é mostrado de acordo com os genes utilizados em diferentes vias de degradação desses compostos. As células contendo um número maior que zero foram sombreadas. (Fonte: Próprio autor)

Gene	Ortólogo	AM_0228	AM_0275	AM_0519	AM_0616
desB	BAK65008.1	0	0	1	0
ferB2	BAK65462.1	0	1	0	0
ligB	BAK65925.1	0	1	0	0
ligU	BAK65931.1	0	0	0	1
ligXc	BAK65525.1	1	0	0	0
ligZ	BAK65447.1	0	1	0	0
phcD	BAK65625.1	0	0	0	1
Gene	**Ortólogo**	**AM_0643**	**AM_0750**	**AM_1003**	**AM_1603**
ligA	BAK65926.1	0	0	0	1
ligB	BAK65925.1	0	0	0	1
ligC	BAK65924.1	0	0	0	1
ligI	BAK65932.1	0	0	0	1
ligJ	BAK65927.1	0	0	1	1
ligK	BAK65930.1	0	0	0	1
ligR	BAK65929.1	0	0	0	1
ligU	BAK65931.1	0	1	1	1
ligXc	BAK65525.1	1	0	0	2
ligXd	BAK66795.1	0	0	0	1
phcC	BAK65623.1	0	0	1	0
phcD	BAK65625.1	0	1	1	0

4.3.6 A expressão gênica revela um perfil de degradação de celulose mediado por GH3

Selecionou-se dois PGs degradadores de lignocelulose, provenientes das seções à jusante de Manaus (AM_0876) e estuário (AM_1603), para realizar-se uma análise de expressão gênica (Figura 4.5). Os perfis de expressão gênica não revelaram expressão associada a esses genes em ambientes de água salobra, reforçando a especificidade desses organismos a ambientes de água doce, ou uma alteração devido a mudanças de seus metabolismos em decorrência de efeitos ambientais.

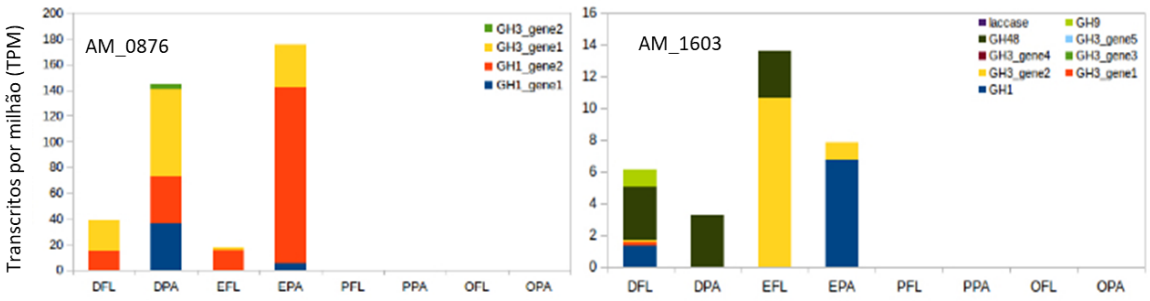

Figura 4.5. Perfil de expressão gênica de 2 PGs representativos da degradação de celulose e oxidação de lignina. Dois PGs provenientes da seção à jusante e estuário foram selecionados para experimentos de expressão gênica. AM_0876 é mostrado como exemplo de um genoma de um organismo com maquinaria menos diversificada para degradação, enquanto AM_1603 mostra um exemplo de um micro-organismo altamente adaptado à degradação da celulose. As diferentes variantes gênicas por família de proteínas foram numeradas de acordo com sua posição nos *contigs* após a predição de genes. As expressões gênicas foram dadas como TPM médio por seção e fração.

Legenda: FL - fração de vida livre; PA - fração associada à partícula; D – Jusante; E - Estuário; P - Pluma; O - Oceano costeiro. (Fonte: Próprio autor)

O PG AM_0876, uma bactéria desconhecida, apresentou maiores expressões associadas a frações associadas a partículas, principalmente no estuário (Figura 4.5). É possível observar a maior diversidade de genes na seção à jusante, enquanto os genes mais expressos sempre foram GH1 e variantes de GH3 2. O maquinário de degradação da matéria orgânica terrestre de *Sphingobium* sp2 (AM_1603) é menos expresso de maneira geral quando comparado a AM_0876, apresentando uma diferença nos níveis de TPM de duas ordens de magnitude (Figura 4.5). Assim, *Sphingobium* sp2 também não demonstrou expressão de nenhum dos genes avaliados em ambientes com água salobra. No entanto, foi expresso principalmente na seção de estuário, na fração de vida livre. Os genes GH1 e GH3 variante 2 foram encontrados principalmente em estuário, enquanto o gene GH48 não foi encontrado na fração associada à partícula no estuário.

4.3.7 Desacoplamento do sistema TTT da degradação de OM terrestre

O fluxo de carbono também é afetado pelos sistemas de transporte, uma vez que eles permitem o uso de fontes alternativas de carbono. O sistema do rio amazonas possui duas principais formas de carbono, uma complexa composta por material de origem alóctone, e materiais menos complexos, como ácidos húmicos e tricarboxilatos. Os transportadores tripartidos de tricarboxilatos (Figura 4.6), também conhecidos como TTT, utilizam proteínas de ligação ao substrato para sequestrar os seus ligandos do

ambiente circundante da célula, sendo considerados transportadores secundários ativos de procariotos (WINNEN; HVORUP; SAIER, 2003).

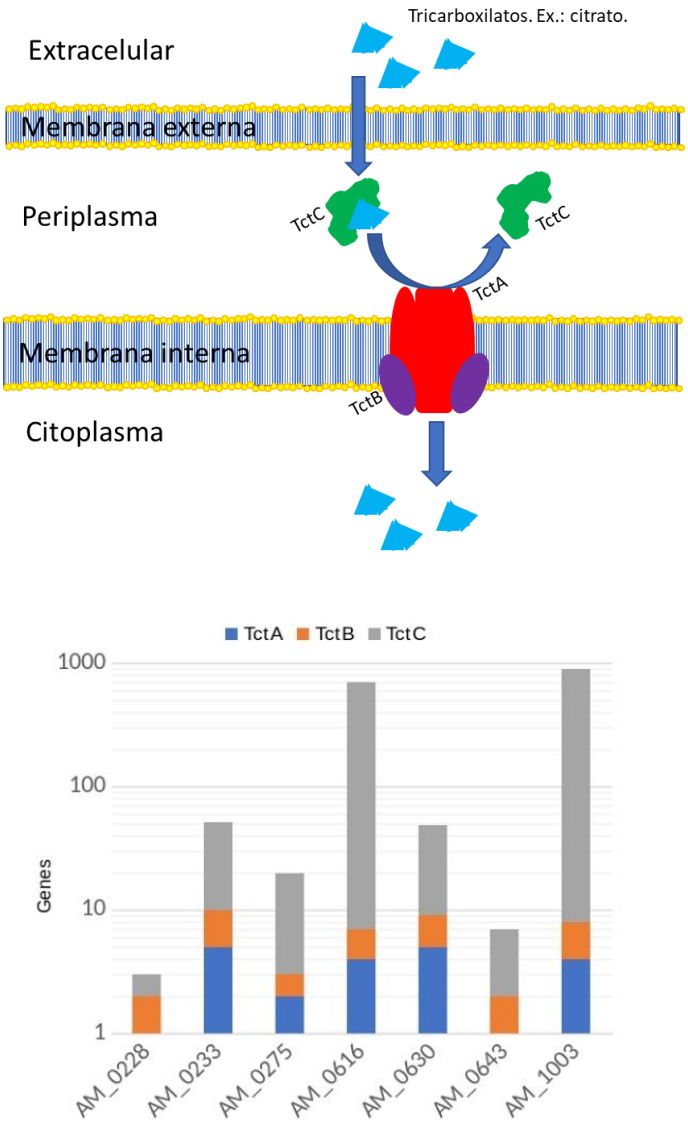

Figura 4.6. Via envolvida no uso de tricarboxilatos e os PGs relacionados a ela. Os PGs da bacia hidrográfica do rio Amazonas foram pesquisados quanto ao potencial para consumir tricarboxilatos via sistema tripartido de transporte de tricarboxilados. (Fonte: Próprio autor)

Neste trabalho, os organismos que possuem o sistema TTT completo, sendo capazes de usar uma ampla gama de fontes de carbono, foram chamados de usuários de fontes alternativas de carbono (UFACs). Apenas sete PGs mostraram-se potencialmente envolvidos nesta via (Figura 4.6), com uma pequena diversidade gênica associada aos genes *tctA* e *tctB,* elementos formadores do canal membranar desse sistema (Figura 4.6). A taxonomia dos PGs contendo o sistema TTT completo foi verificada e apenas uma alfa-Proteobacteria (AM_0275) foi relatada, enquanto as demais foram identificadas como Beta-proteobactérias, principalmente da família Burkholderiales.

Uma característica importante deste sistema é a existência de muitas variantes de proteínas de ligação ao substrato (*tctC*), enquanto apenas alguns genes da porção ligada à membrana (*tctA* e *B*) são necessários. Isto é especialmente mostrado na Figura 4.6, onde o *tctC* foi encontrado desde dezenas a centenas de variantes gênicas ao longo de um único genoma. Micróbios contendo sistemas TTT completos não foram encontrados como sendo agentes degradadores de celulose ou oxidadores de lignina, exceto o PG AM_0630, um membro de Burkholderiales contendo lacase e genes de GH8. Interessantemente, todos os organismos que contiveram o sistema TTT completo (exceto AM_0630 e AM_0233) também possuem maquinaria para degradação de compostos aromáticos derivados da oxidação de lignina. Isso sugere uma conexão potencial desses organismos na estrutura microbiana que domina esse ambiente, de forma que haja uma especificação de funções ecológicas.

O PG AM_1003 foi recuperado da seção à jusante e possui o maior sistema TTT representado em nosso conjunto de dados (4 *tctA*, 4 *tctB* e 894 *tctC*) e foi utilizado para uma análise de expressão gênica. Os perfis de expressão gênica não revelaram a presença do sistema TTT desse genoma em águas salobras, sugerindo sua preferência

por água doce (Figura 4.7). A expressão gênica também não foi observada na fração associada a partículas da seção estuarina. Foi observada uma expressão preferencial de algumas variantes dos genes *tctA* e *tctB*. Evidências da preferência por frações de vida livre em água doce foram observadas para os genes *tctC*, bem como, menores taxas de expressão (a maioria das variantes gênicas tem < 10 TPM). Poucas variantes foram expressas em níveis superiores a 10 TPM, com apenas duas delas sendo expressas em níveis mais altos (> 100 TPM), especificamente na seção de vida livre do estuário, sendo candidatas interessantes para uma análise mais aprofundada no futuro.

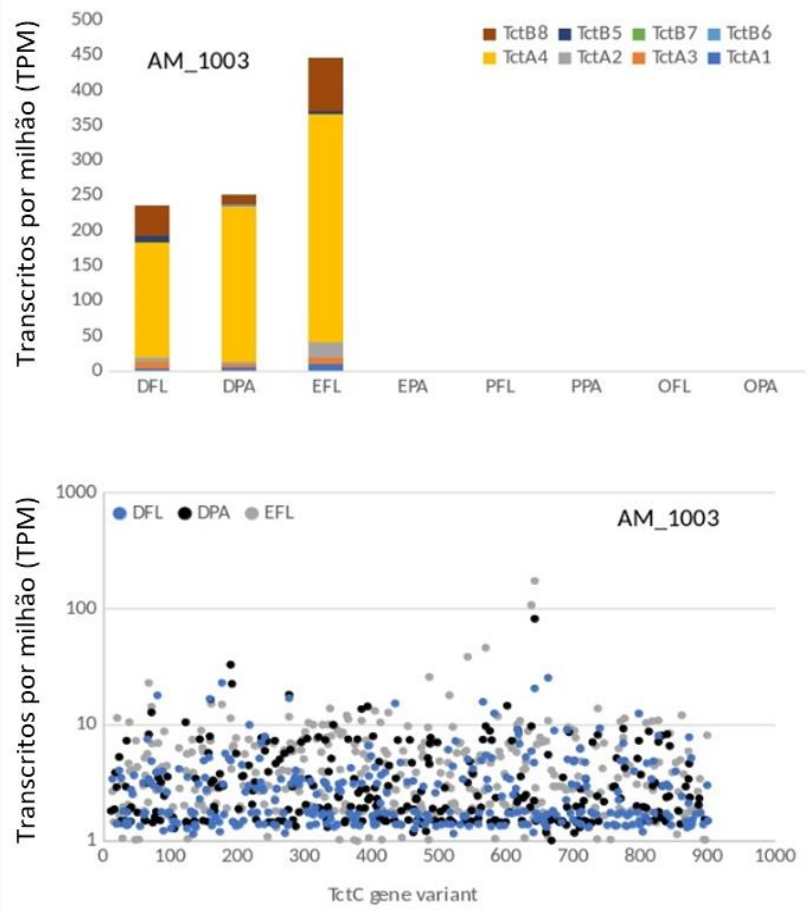

Figura 4.7. Expressão gênica do PG com maior sistema TTT encontrado neste estudo. O genoma do organismo AM_1003, que se acredita ser um membro da classe Beta-Proteobacteria, teve sua expressão gênica analisada quanto ao sistema TTT. Os

genes que compõem o sistema tripartido de transporte de tricarboxilato (genes *tctA*, *B* e *C*) são mostrados separadamente. Legenda: FL – Fração de vida livre; PA – Fração associada a partículas; D – Jusante; E – Estuário; P – Pluma; O – Oceano. (Fonte: Próprio autor)

4.3.8 Armazenamento de carbono: o sistema TTT está acoplado à via de síntese do PHB

Um dos sistemas mais importantes de armazenamento de carbono é o metabolismo de poli-hidroxi-butiratos (PHB) (Figura 4.8), uma característica interessante no fluxo de carbono, mas pouco explorada. Aqui, as enzimas de biossíntese de PHB foram pesquisadas nos PGs para avaliar o seu potencial para armazenar carbono neste biopolímero (Figura 4.8). Quase todos os PG que apresentam a via completa (*phaA-C*) também possuíram o sistema TTT, com exceção de AM_0528 e AM_1603, que em vez disso, não têm o sistema TTT e degradam a OM terrestre. Observou-se que a principal diversidade de genes está relacionada aos passos iniciais da biossíntese de PHB, enquanto um número limitado de genes codifica os últimos passos realizados pelos genes *phaC* e *phaE*. O gene *phaR* está presente apenas em 7 dos 13 PGs que se acredita serem capazes de produzir PHB, enquanto apenas o PG AM_1111 pode ser capaz de produzir PHAB, uma vez que também contém o gene *phaE*. Não foram encontrados PGs contendo o gene *cat2*, também conhecido como *orfZ*. O maior número de genes relacionados à essa via foi encontrado no degradador de OM terrestre AM_1603, representante da espécie *Sphingobium* sp2.

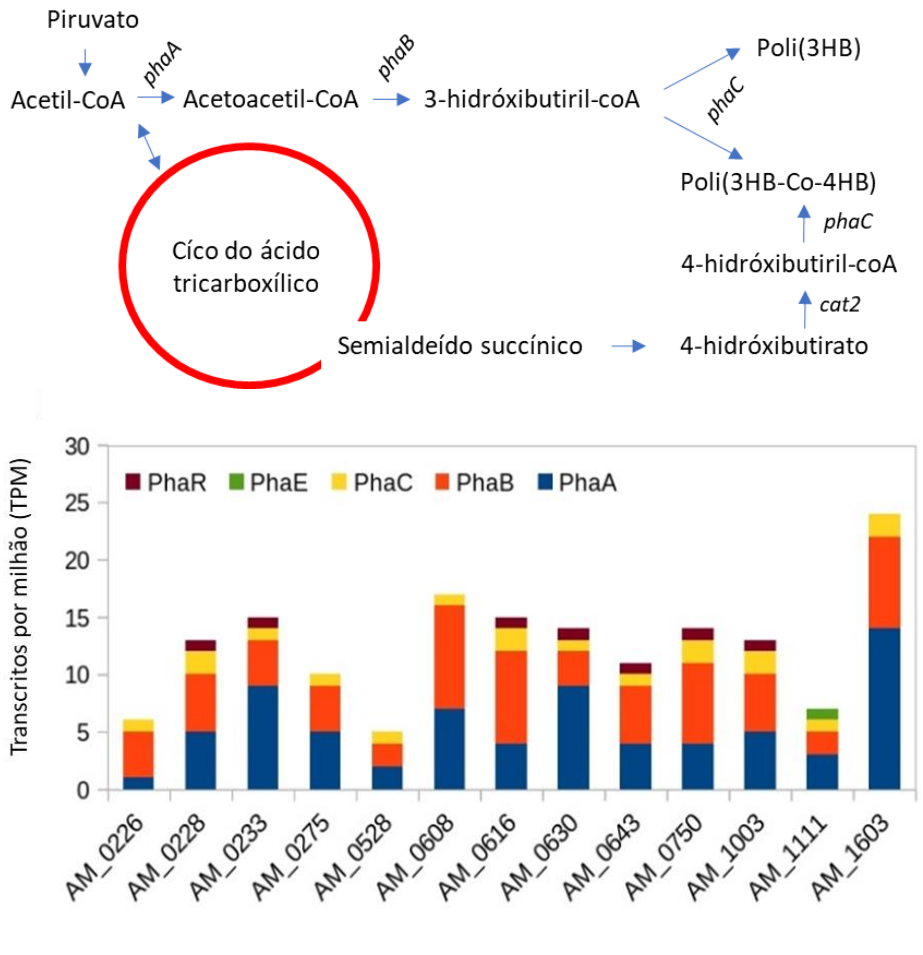

Figura 4.8. Via de armazenamento de carbono por biossíntese de PHB. A produção de polihidroxibutirato/alcanoato foi investigada para avaliar o potencial de armazenamento de fontes de carbono nesses micróbios. Os PGs com o potencial para produzir esses compostos são mostrados. A enzima *phaE* permite, quando em presença da enzima *phaC*, a biossíntese de Poly3HB-co-4HB, além da proteína *phaR* que regula a acumulação de PHB em grânulos maiores no interior da célula. (Fonte: Próprio autor)

Para entender melhor como a biossíntese de PHB pode ocorrer ao longo do rio Amazonas e quais são as diferenças entre UFACs e organismos degradadores de OM

terrestre, avaliou-se a expressão gênica dos genes da via do PHB nos dois PGs que melhor representaram essa via, os genomas AM_1003 e AM_1603 (Figura 4.9).

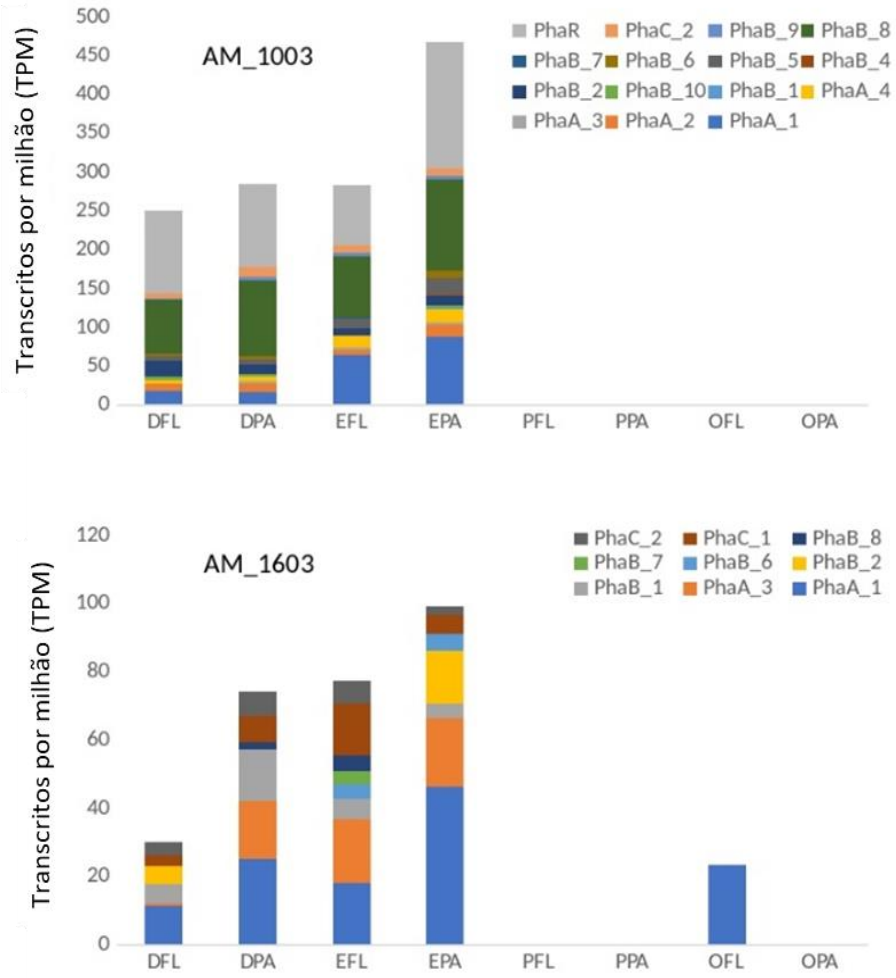

Figura 4.9. Perfis de expressão gênica da via do PHB em dois PGs representantes dos grupos: UFAC (AM_1003) e degradadores de OM terrestre (AM_1603). De maneira geral, as diferentes variantes gênicas por família de proteínas foram numeradas de acordo com sua posição nos *contigs* durante a predição gênica. Legenda: FL – Fração de vida livre; PA – Fração associada a partículas; D – Jusante; E – Estuário; P – Pluma; O – Oceano. (Fonte: Próprio autor)

O genoma AM_1003 foi caracterizado como UFAC e apresentou maior expressão de genes relacionados à via do PHB quando comparado ao AM_1603, um degradador de OM terrestre. Apesar do PG AM_1603 apresentar mais genes relacionados a essa via, a maioria deles não foi expressa. Ambos os PGs não tiveram expressão de genes dessa via em ambientes com água salobra, mostrando sua preferência por água doce. A via do PHB é administrada por variantes específicas dos genes *phaA-C* em ambos os organismos, enquanto a expressão gênica de *phaA* e *phaB* é muito maior que a do gene *phaC*. Apenas o PG AM_1003 possui o gene *phaR* considerado essencial para o acúmulo de PHB em altos níveis a partir de substratos de 1 a 5 carbonos. O gene *phaR* foi transcrito em altos níveis, quando comparado às outras enzimas dessa via metabólica (Figura 4.9).

4.4 Discussão

4.4.1. Os genomas populacionais da bacia do rio Amazonas

Cinquenta e dois PGs (3,6% do total de PGs recuperados dos *contigs* metagenômicos da bacia do rio Amazonas) foram selecionados após as etapas de filtragem, refinamento e eliminação de PGs redundantes (Apêndice 8). Essa taxa de recuperação de PGs, após a filtragem por qualidade, é semelhante à obtida por outros autores (HUGERTH et al., 2015; PARKS et al., 2017). Isso mostra que os algoritmos de categorização precisam de melhorias, de modo que mais PGs de qualidade possam ser recuperados, reduzindo-se a categorização não específica.

De um modo geral, os PGs aqui reportados estavam de acordo com padrões previamente estabelecidos (BOWERS et al., 2017), sendo considerados genomas-esboço de alta qualidade. A maior taxa de recuperação de PG da seção à montante de Manaus (Figura 4.1) reflete a cobertura mais profunda dos metagenomas dessa seção, sugerindo que metagenomas com cobertura mais profunda podem ser melhores para gerar genomas populacionais, quando comparados a metodologias que adotam números maiores de réplicas de metagenomas com baixa cobertura. Além disso, os metagenomas que compõe essa seção são aqueles produzidos pelo nosso laboratório.

Interessantemente, o maior número de PGs recuperado na montante poderia ser associado a uma maior quantidade de material particulado e nutrientes das águas brancas do rio Solimões, que após a mistura das águas do rio Negro, um rio de águas pretas, dilui a carga em suspensção e os nutrientes, reduzindo assim o número de microrganismos nas estações seguintes (seções da jusante à estuário). Já observa-se que nas estações na pluma e oceano, uma barreira como a *salinidade* compõe um dos principais fatores estruturadores das comunidades microbianas nestas regiões, onde há uma sazonalidade devido à produção primária e sedimentária.

A composição taxonômica dos genomas recuperados (Figuras 4.1 e 4.2) foi semelhante àquela observada por outros autores utilizando a análise de *reads* (DOHERTY et al., 2017; SATINSKY; SMITH; SHARMA; WARD; et al., 2017). A predominância de Proteobacteria (39% dos genomas recuperados) reforça os resultados de outros estudos (GHAI; RODRIGUEZ-VALERA; et al., 2011; TOYAMA et al., 2017) que indicam a microbiota do rio Amazonas amplamente ocupada pelos grupos Actinobacteria e Proteobacteria. A seção a montante possuía o conjunto taxonomicamente mais diverso de PGs, quando comparado com as outras, o que pode

estar relacionado a uma cobertura mais profunda de seus metagenomas aliada a uma maior diversidade microbiana que decresce após a junção dos rios Solimões e Negro (seção à jusante de Manaus), um provável efeito da interface de mistura estratificada.

O *continuum* salino também mostrou ser um fator principal para a composição de PGs, onde um menor número de táxons pode ser recuperado e diferentes composições podem ser observadas entre pluma e oceano, como previamente verificado por Satinsky et al. (SATINSKY; CRUMP; et al., 2014; SATINSKY; ZIELINSKI; et al., 2014). Embora *Synechococcus* tenha sido relatado como um gênero fototrófico dominante nas águas do rio Amazonas (SATINSKY; CRUMP; et al., 2014; SATINSKY et al., 2015), outras cianobactérias (nomeadamente *Anabaena* e *Richelia*) foram mais abundantes neste ambiente, uma vez que pudemos recuperar seus PGs.

Comunidades de arquéias mostram-se claramente segregadas, com dominância das Thaumarchaeota em ambientes de água doce, enquanto Euryarchaeota está presente em água salobra. Isto reforça o que já foi observado anteriormente (SANTOS-JÚNIOR et al., 2017).

4.4.2 Cepas coespecíficas e sua importância ambiental

Cepas coespecíficas (Apêndice 10) mostram alta completude e similaridade com seus genomas referência. Nossas comparações foram realizadas usando milhares de genomas de alta qualidade depositados em bancos de dados de referência. No entanto, apenas 19,2% dos genomas aqui relatados foram agrupados em cepas coespecíficas, revelando uma grande novidade associada a eles. PGs deste estudo também mostraram

ser melhores referências para algumas espécies do que aqueles depositados nas bases de dados de referência, por exemplo, a cepa coespecífica de *Sphingobium* sp2.

Por meio de abordagens de similaridade genômica, foi possível identificar os dois principais gêneros tidos como responsáveis pelas assinaturas fotossintéticas no ambiente da bacia amazônica: *Richelia* e *Anabaena*. *Richelia intracellularis* foi previamente identificado por análise de *reads* (HILTON et al., 2015), e embora tenha sido relatado anteriormente ocupando preferencialmente a seção da pluma (DEL VECCHIO; SUBRAMANIAM, 2004), verificou-se que era abundante nas seções do estuário até oceano. Enquanto isso, o genoma populacional relacionado à *Anabaena*, AM_0902, foi mais abundante nas seções à jusante e na pluma. Estas descobertas reforçam a ideia de uma crescente assinatura autotrófica em direção ao oceano, principalmente devido à menor presença de materiais particulados (HILTON et al., 2015; SATINSKY et al., 2015; SATINSKY; SMITH; SHARMA; LANDA; et al., 2017; SATINSKY; SMITH; SHARMA; WARD; et al., 2017). Outro papel desempenhado por eles é o ciclo do nitrogênio, uma vez que esses gêneros são considerados diazotróficos importantes, como observado anteriormente (HILTON et al., 2015).

Grupos desconhecidos foram recuperados como cepas coespecíficas, como no caso do gênero desconhecido UBA11236 (AM_1606), no qual não há informações associadas. Alguns deles foram parcialmente estudados, como a cepa coespecífica AM_2207 da GWA2-73-35 sp1, algumas espécies caracterizadas do filo candidato a Rokubacteria. Embora poucas informações disponíveis, Rokubateria é reconhecido como um filo frequente em estudos metagenômicos com crescimento realizado a partir da degradação de acetato ou de ácidos graxos. Sua característica única é codificar múltiplas oxirredutases nitríticas, com essas bactérias provavelmente formando

comunidades metabólicas interconectadas (HUG et al., 2016). Isso sugere o PG AM_2207 como um dos principais participantes do ciclo de nitrogênio nas comunidades microbianas da bacia do rio Amazonas. Outro exemplo de cepas específicas com táxons não-descritos é o PG AM_2208, uma cepa coespecífica de *Coccinistipes* sp. Este gênero é representado por um isolado do ambiente marinho em 2006 ainda não publicado. Há poucas informações disponíveis sobre este gênero do filo Bacteroidetes. Mas aqui, podemos inferir mais informações sobre sua abundância e relatar um genoma desse organismo que pode ser usado para prever características fenotípicas e ajudar a entender melhor sua importância no ecossistema.

Embora dois PGs (AM_0507 e AM_0219) tenham sido cepas coespecíficas de *Candidatus* Methylopumilus sp1, observou-se que ambos diferem em um grau relativamente alto para serem considerados iguais. Desta forma, sugerimos que ambos sejam considerados cepas distintas. *Candidatus* Methylopumilus é um gênero composto por espécies metilotróficas originalmente isoladas de plâncton de lagos de água doce (SALCHER et al., 2015). Estes organismos geralmente possuem bacteriorrodopsinas, permitindo que haja produção de energia com a presença de luz (SALCHER et al., 2015), mas isso não foi verificado para AM_0507 e AM_0219, talvez devido à sua integridade de apenas ~80%. Esses dois genomas apresentam uma distribuição de abundância complementar, sendo AM_0507 mais abundante em águas salobras, principalmente em frações livres, enquanto AM_0219 está mais presente nos trechos à montante e estuário (Figura 4.2).

Dois genomas (AM_0608 e AM_1603) foram relacionados à espécies altamente ativas na mineralização de compostos orgânicos que impactam de diferentes formas as fontes de carbono no ambiente. A cepa específica *Acinetobacter junii*

AM_0608, apesar de estar relacionada a vários casos de infecções nosocomiais (GERISCHER, 2008), é uma espécie imóvel, Gram-negativa, oxidase negativa do filo Gama-Proteobacteria. O gênero *Acinetobacter* foi encontrado anteriormente na bacia do rio Amazonas (GHAI; RODRIGUEZ-VALERA; et al., 2011; TOYAMA et al., 2017) e revelou-se mais abundante nas frações associadas às partículas oceânicas (Figura 4.2). Este gênero é um importante grupo de organismos envolvidos na mineralização de compostos aromáticos (JUNG; BAEK; PARK, 2010). Outro gênero relatado foi *Sphingobium* sp2 como uma cepa específica de AM_1603, conhecida por degradar vários compostos aromáticos e halogenados (por exemplo, herbicidas como o ácido (RS)-2-(4-cloro-2-metilfenoxi) propiônico (GAI et al., 2011; ZIPPER et al., 1996). Este micróbio foi mais abundante na fração associada a partículas do estuário, com uma distribuição regular sobre outras seções (Figura 4.2). Isto mostra uma preferência desta espécie de *Sphingobium* por água doce em um estilo de vida associado a partículas.

Houve cepas coespecíficas de patógenos, como *Trueperella pyogenes* e *Xanthomonas fuscans* subsp. *aurantifolii*, respectivamente AM_0546 e AM_0519. O PG AM_0546, apesar de estar relacionado a um patógeno animal oportunista, pode ser considerado um micróbio potencialmente degradador da OM terrestre (Figuras 4.3 e 4.4). *T. pyogenes* é geralmente hospedado em vias respiratórias superiores, urogenitais e gastrointestinais (MACHADO; BICALHO, 2014), onde a OM terrestre trafega durante a digestão, explicando em parte este achado. No entanto, o PG AM_0546 também possui o gene da listeriolisina O, codificando um dos mais importantes fatores de virulência envolvidos em suas infecções. Isso mostra que este organismo ao mesmo tempo pode funcionar como um degradador da OM terrestre no meio ambiente, e no caso de encontrar um hospedeiro, pode se transformar em patógeno. Evidências de sua

preferência por águas salobras (Figura 4.2) também sugerem essa cepa como uma espécie diferente, mais adaptada ao estilo de vida oceânico.

A linhagem patogênica AM_0519 também se mostrou degradadora da OM terrestre (Figuras 4.3 e 4.4), principalmente nas seções à jusante de Manaus, estuário e pluma, com migração da fração associada à partícula para vida livre entre as seções à jusante até a pluma. *Xhantomonas fuscans* subsp. *aurantifolii* é um tipo de patógeno de *Citrus* sp., causando o cancro cítrico tipo C (MOREIRA et al., 2010). Seu estilo de vida aliado à sua principal maquinaria genômica revela uma potencial introdução desse organismo por transferência mássica de material terrestre contaminado. Este organismo pode ser adaptado a esta mudança de habitat, transformando-se em um decompositor, uma vez que é difundido nas águas à jusante de Manaus até a pluma, abandonando as partículas e tornando-se um organismo de vida livre. Entretanto, mais evidências são necessárias para suportar tal teoria.

4.4.3 Potencial metabólico

Rhizobiales (AM_0275) foram encontrados com potencial para degradar dioxina, naftaleno e aminobenzeno, enquanto *Sphingobium* sp2 (AM_1603) foi principalmente associado à degradação de dioxina e cloroalcanos/alcenos, mais provavelmente degradados por cianobactérias (AM_0902 e AM_2804) e Chlorobiaceae (AM_0510). Esses padrões de degradação mostram organismos à jusante de Manaus com um maquinário mais sofisticado para lidar com xenobióticos do que aqueles presentes nas seções após a jusante de Manaus até o oceano. De fato, parece que o membro Rhizobiales AM_0275, é um dos organismos mais interessantes para ser usado

em processos de biorremediação ou capaz de povoar essas águas em caso de contaminação na seção à jusante, enquanto Cianobacteria e Clorobiacea parecem estar mais focados em compostos halogenados. Apesar do gênero *Sphingobium* ser considerado um grupo de degradadores altamente potentes, o PG AM_1603 não apresentou potencial de degradação de uma grande variedade de xenobióticos, sugerindo que esta cepa possa ter adaptações que a diferenciam das cepas desse gênero. Deve-se notar que essas discussões foram feitas sobre comparações de dados genômicos com as vias e organismos presentes no banco de dados referência KEGG e não atribuem completamente as vias putativas de um organismo particular, apenas revelam um potencial levando em conta o conhecimento atual.

A ausência de organismos metanogênicos pode ser explicada pelo metabolismo anaeróbico desses organismos e pela amostragem feita em zonas consideradas oxigenadas. Esperava-se que o Thaumarchaeota AM_0615 fosse metanotrófico, uma vez que esse grupo é marcado por adaptações à quimioautotrofia, via oxidação anaeróbica de amônio (SWAN et al., 2014). Este organismo também pode ser importante para os ciclos de nitrogênio e carbono, desempenhando um papel fundamental neste ambiente (PESTER; SCHLEPER; WAGNER, 2011). Outro possível metanótrofo foi um membro Planctomycetes (AM_0621), um filo bacteriano menos conhecido, que é sabidamente importante para a ciclagem de nitrogênio e carbono (TADONLÉKÉ, 2007). O PG AM_0621 foi encontrado principalmente em amostras de vida livre no estuário, o que pode significar uma intersecção com o excesso de nitrogênio ciclado pelas cianobactérias, previamente descritas por outros autores. (HILTON et al., 2015; SATINSKY et al., 2015; SATINSKY; SMITH; SHARMA; WARD; et al., 2017). Duas Verrucomicrobia (AM_1503 e AM_1811) também tinham

uma assinatura metanotrófica em seus genomas. Verrucomicrobia metanotróficas foram previamente encontradas em fontes ácidas e geotermais (SHARP et al., 2014) sugerindo agora uma nova distribuição delas também em ambientes não extremos.

O ciclo de nitrogênio no ecossistema da Amazônia não foi amplamente verificado em nosso conjunto de dados. No entanto, os genomas de Cyanobacteria, Chlorobia e Thaumarchaeota mostraram um potencial para realizar esta via. Isso foi sugerido anteriormente (HILTON et al., 2015) e pode estar relacionado a uma baixa abundância de organismos diazotróficos, devido ao alto nível de materiais particulados e à difícil penetração da luz. Assim, as presentes descobertas reforçam a função do grupo das Cyanobacterias no ciclo do nitrogênio, sugerindo também uma parceria com os organismos do grupo dos Thaumarchaeota.

O ciclo de enxofre revelou-se mais intenso na seção à montante, principalmente devido à táxons desconhecidos, Beta-proteobacteria e Cyanobacteria, algo já observado por outros autores (SATINSKY; SMITH; SHARMA; LANDA; et al., 2017; SATINSKY; SMITH; SHARMA; WARD; et al., 2017). As componentes chaves parecem ser a absorção extracelular de sulfato e a redução de sulfato assimilatório. Esta aparente falta de mais componentes é devido às características anaeróbicas deste ciclo (PFENNIG; WIDDEL, 1982) e nossa amostragem ter ocorrido em zonas oxigenadas. Os principais táxons que realizam o ciclo do enxofre crescem em condições anóxicas. Alguns táxons, por exemplo *Thiobacillus*, podem mediar o ciclo do enxofre sobre a interface óxico-anóxica (JØRGENSEN, 1982). No entanto, esses organismos ou organismos relacionados a eles não foram encontrados em nosso estudo. Assim, sugerimos uma investigação mais aprofundada sobre a microbiota de sedimentos para

uma melhor compreensão do ciclo biogeoquímico deste elemento no ecossistema amazônico.

4.4.4 Organismos degradadores de OM terrestre

A degradação da OM terrestre revelou-se uma função-chave para essa microbiota, uma vez que quase metade das PGs a apresentou (Figura 4.4). Isto é concordante com o fato de que a OM terrestre (celulose, hemicelulose e lignina) pode representar 80% de uma biomassa florestal que termina no rio (BOERJAN; RALPH; BAUCHER, 2003; BOSE et al., 2009; MARTENS; REEDY; LEWIS, 2004). Este fato sugere uma população de especialistas ajudando uma população ligeiramente maior de espécies generalistas que não podem usar o carbono orgânico desses materiais em estado bruto. Entre as espécies que possuem algumas famílias de proteínas envolvidas na oxidação de lignina, parecem estar presentes em uma cópia única em quase todos os táxons, exceto Bacteroidetes. Essa diversidade gênica é muito menor do que a verificada para organismos do solo (LÓPEZ-MONDÉJAR et al., 2016), mostrando que, apesar de terem capacidade de degradar lignina e celulose, essa capacidade deve ser relativamente mais limitada.

Os PGs de degradadores de OM terrestre parecem possuir a habilidade e potencial dependência da OM terrestre para sobreviver, uma vez que foram obtidas de amostras de água doce e não foram abundantes em ambientes com água salobra. Outra conclusão possível é de que há uma redução da atividade de degradação da OM terrestre à medida que o curso do rio atinge o oceano, uma vez que o número de PG recuperados

com potencial de degradação é reduzido nas seções à jusante e estuário. Entretanto, é importante ressaltar a possibilidade de a OM terrestre não ser degradada e tornar-se um material recalcitrante em ambientes de água salobra, uma vez que as PG recuperadas dessas regiões apresentavam pequeno potencial de degradação da OM terrestre aliada a uma baixa diversidade genética.

Os perfis de expressão gênica mostram os organismos avaliados (AM_0876 e AM_1603) como organismos de água doce, fato que sustenta a ideia de serem utilizados como modelos de organismos degradadores de OM terrestre na bacia do rio Amazonas. O PG AM_0876, uma bactéria desconhecida, apresentou maiores expressões associadas a frações associadas a partículas, principalmente no estuário (Figura 4.5). Isso é consonante com observações de outros autores em que a maior degradação da OM terrestre está associada a uma tendência de agregação celular (CORNO et al., 2015). A diversidade gênica e os padrões de expressão ainda sugerem que o repertório de enzimas usado por AM_0876 na degradação da OM terrestre pode estar funcionando de uma maneira indutível ou que outras enzimas que não as variantes 2 dos genes de GH1 e GH3 poderiam ser reguladas por *feedback*.

A maquinaria de degradação da OM terrestre do *Sphingobium* sp2 (AM_1603) é uma ordem de grandeza menos expressa de maneira geral quando comparada a do PG AM_0876 (Figura 4.5). Isso poderia sugerir que, apesar de ter uma maquinaria geneticamente diversa, o *Sphingobium* sp2 seja fastidioso ou menos responsivo a OM terrestre, uma vez que pode usar outras fontes de carbono. A expressão gênica concentrada na fração de vida livre da seção estuarina traz a ideia de que os organismos degradadores da OM terrestre formam consórcios quando associados às partículas. Seu perfil de expressão gênica sugere uma organização complexa dessas variantes gênicas

sobre o genoma, seguida por uma rede de regulação de expressão bastante truncada. Assim, as mudanças no perfil compostas principalmente pelas enzimas GH1 e GH3, e a ausência da GH48 na fração associada às partículas do estuário poderiam refletir a perda de carbono terrestre dentro do rio. Estas são evidências de um material orgânico mais complexo na seção à jusante, enquanto a OM terrestre no estuário deve ser mais facilmente degradada. De fato, isso reforça a ideia da mudança da natureza da OM ao longo do curso do rio, em que porções à montante e jusante possuem origem alóctone que é substituída por fontes de produção primária na foz do rio (SATINSKY; SMITH; SHARMA; WARD; et al., 2017; SEIDEL et al., 2016).

4.4.5 O papel do sistema transportador tripartido de tricarboxilatos no uso de fontes alternativas de carbono

Embora outros estudos relatem uma enorme quantidade de novos elementos TTT no ecossistema microbiano do rio Amazonas (GHAI; RODRIGUEZ-VALERA; et al., 2011), uma pequena parte dos nossos PGs tinha o sistema TTT (13,5%). Seis dos sete PGs descritos como contendo o sistema TTT pertencem a Beta-proteobactérias, principalmente a família Burkholderiales. Isso era esperado, uma vez que este sistema de transporte foi descrito pela primeira vez e provou ser mais presente no grupo das Beta-proteobactérias (ANTOINE et al., 2003), com algumas exceções.

Encontrou-se um arsenal completo de diferentes genes de proteínas de ligação ao substrato (*tctC*), de dezenas a centenas de variantes genéticas ao longo de um único genoma. Esta evidência concorda com outros autores (ANTOINE et al., 2003) que descobriram esta família de proteínas em *Bordetella pertussis*, que possuía 90 cópias do

gene, com 79 delas intactas e funcionais. Outras bactérias foram encontradas contendo entre 1 e 11 genes *tctC*. Como existe uma alta afinidade de cada gene variante ao seu substrato, há evidências de um uso de múltiplos substratos através deste sistema (ANTOINE et al., 2003, 2005; WINNEN; HVORUP; SAIER, 2003). Pode ser que em trabalhos futuros, haja a identificação de *tctC* responsáveis pelo transporte de outros substratos, além daqueles atualmente conhecidos, reconhecendo genes que desempenhem um papel importante nas comunidades microbianas da Amazônia. Os genes dos elementos de membrana (*tctA* e *tctB*) mostraram-se menos diversificados, com poucas variantes por genoma. Isto explica a eficiência aparente do sistema, onde são necessários poucos elementos de membrana para receber vários substratos diferentes.

A especialização nesta função de transporte parece ser um traço de organismos não-degradadores de OM terrestre, o que poderia sugerir uma potencial compartimentalização do uso de carbono na estrutura microbiana que domina este ambiente. Esta é uma característica comum em micróbios heterotróficos do oceano profundo e a especialização em sistemas de transporte parece ser uma das características mais promissoras do uso de OM dissolvida na coluna de água oceânica (BERGAUER et al., 2018).

A análise da expressão gênica do PG AM_1003 mostrou a expressão preferencial de algumas variantes de *tctA* e *tctB*, o que poderia implicar em um controle gênico mediado por condições específicas, onde a resposta basal corre sobre variantes gênicas específicas. A sobre-representação de variantes do gene *tctC* sugere uma maquinaria bioquímica rica para lidar com este painel de substratos putativos. Outro fato que reforça essa capacidade ampliada do uso de fontes alternativas de carbono é o

fato de variantes do gene *tctC* serem pouco expressos, apresentando expressão preferencial sob condições ambientais específicas, que devem ser investigadas. Em resumo, o TTT revelou-se um sistema importante no processamento de OM e representa uma alternativa para a degradação direta da OM terrestre.

4.4.6 Armazenamento de carbono via PHB

O metabolismo de polihidróxi-butiratos (PHB) (Figura 4.8) é uma característica interessante para os genomas microbianos em contextos ambientais e biotecnológicos, uma vez que este polímero apresenta diversas aplicações industriais (LENZ; MARCHESSAULT, 2005; POLI et al., 2011). Organismos especializados em extrair carbono do meio ambiente podem apresentar dois comportamentos principais, utilizá-lo ou simultaneamente armazená-lo. Um polímero alternativo pode ser produzido na via do PHB com a presença da enzima PhaE e da enzima PhaC, gerando um polímero híbrido contendo monômeros de 4-hidroxibutiril e 3-hidroxibutiril (PHAB), com diferentes propriedades. Para isso, o semi-aldeído succínico proveniente do ciclo de TCA precisa ser convertido pela enzima Cat2 (OrfZ) em moléculas de 4-butiril. Este metabolismo é uma rede genética complexa, onde o sistema de regulação gênica é mediado pelo gene *phaR*, um repressor transcricional. O *phaR* controla a expressão de enzimas biossintéticas, *phasins* e PHAB, podendo atuar como regulador global da alocação e simbiose de carbono em excesso (QUELAS et al., 2016).

A via completa (*phaA-C*) foi relacionada aos organismos com sistema TTT, sugerindo primeiramente um acoplamento desse sistema, e em segundo lugar um papel

importante realizado pelas Betaproteobactérias no armazenamento do carbono. A redundância do sistema de biossíntese do PHB é muito inferior à apresentada pelo sistema TTT. As reações limitantes da via são realizadas por enzimas codificadas por genes únicos ou poucas variantes genéticas, sugerindo a ruptura da via se elas forem perdidas. O gene *phaR* revelou estar em mais da metade dos PGs apresentando a via, o que mostra que esses organismos são potenciais acumuladores de PHB. No entanto, acredita-se que apenas um deles (AM_1111) produz PHAB, apesar de não possuir o gene *cat2* (Figura 4.8).

Análises de expressão gênica (Figura 4.9) revelam que UFACs apresentam maior expressão de genes relacionados à via do PHB quando comparados a organismos degradadores de OM terrestre, embora possuam menor redundância dos genes dessa via. Outro fato importante é que a via de biossíntese de PHB é administrada por variantes específicas dos genes *phaA-C* em ambos os organismos, enquanto a expressão gênica de *phaA* e *phaB* é muito maior que do que a do gene *phaC*. O gene *phaR* do PG AM_1003, um UFAC, foi altamente expresso, algo incomum para os reguladores transcricionais. Entretanto, outros autores (KOROTKOVA; CHISTOSERDOVA; LIDSTROM, 2002) também sugerem que o *phaR* é um agente condutor do fluxo de acetil-coA para a síntese de PHB, permitindo o acúmulo de PHB em altos níveis a partir de substratos contendo de 1 a 5 carbonos. Nesse sentido, organismos que degradam a OM terrestre e também realizam armazenamento de carbono devem armazenar quantidades menores de PHB em um mesmo lapso de tempo, quando comparados a organismos que utilizam o sistema TTT para recuperar suas fontes de carbono do meio ambiente.

4.4.7 Releitura do efeito priming *no rio Amazonas*

Com base no exposto neste capítulo revisou-se o modelo do efeito *priming* proposto no capítulo 3 do presente trabalho. Neste modelo (Figura 4.10), levou-se em conta o contexto genômico dos genes, em que há um acoplamento da oxidação de lignina e degradação de celulose e hemicelulose. Diferentemente do modelo previamente proposto, aqui sabemos que há um consórcio que degrada diretamente a matéria orgânica terrestre, composto por duas subcomunidades, uma exclusivamente degradadora de celulose e outra que além disso também degrada a lignina.

Essas comunidades devem trabalhar juntas com o intuito de degradar celulose, para que recebam carbono estrutural e energia. Entretanto, uma vez que não verificamos os aparatos de internalização e metabolização dos compostos aromáticos derivados de lignina, é esperado que esses microrganismos não consumam esses compostos e, portanto, degradem a lignina, pois esta impede a ação das celulases estericamente.

Uma vez que a lignina é degradada, gerando compostos fenólicos, o consórcio microbiano degradador de OM terrestre sofre uma inibição. Outros autores reportam um efeito inibitório significante desses compostos em relação a celulases (QIN, Lei et al., 2016) e à dinâmica de fermentação anaeróbica (MONLAU et al., 2014; XUE et al., 2018) e culturas puras aeróbicas (ASTON et al., 2016).

As comunidades UFAC consomem esses compostos fenólicos derivados de lignina, eliminando a inibição por *feedback* sofrido pelas comunidades degradadoras de OM terrestre. Essas comunidades também recebem carbono estrutural e energia a partir da degradação de celulose, promovida pelo consórcio degradador. Por sua vez, as

UFAC se caracterizam pelo acoplamento do sistema TTT e os sistemas de degradação dos compostos fenólicos derivados da lignina, assim como, a presença do sistema de armazenamento de carbono intracelular mediado pela biossíntese de PHB. Essas comunidades ainda realizam de modo intracelular a produção desse polímero e seu consumo de acordo com o ambiente extracelular, regulando assim seu próprio crescimento e expansão clonal. Dessa forma, o efeito *priming* no rio Amazonas parece ser devido à eliminação da inibição decorrente da produção de compostos fenólicos derivados da lignina. A degradação de celulose é mediada principalmente por enzimas GH3 e GH1, enquanto a oxidação da lignina ocorre via lacases e DYPs.

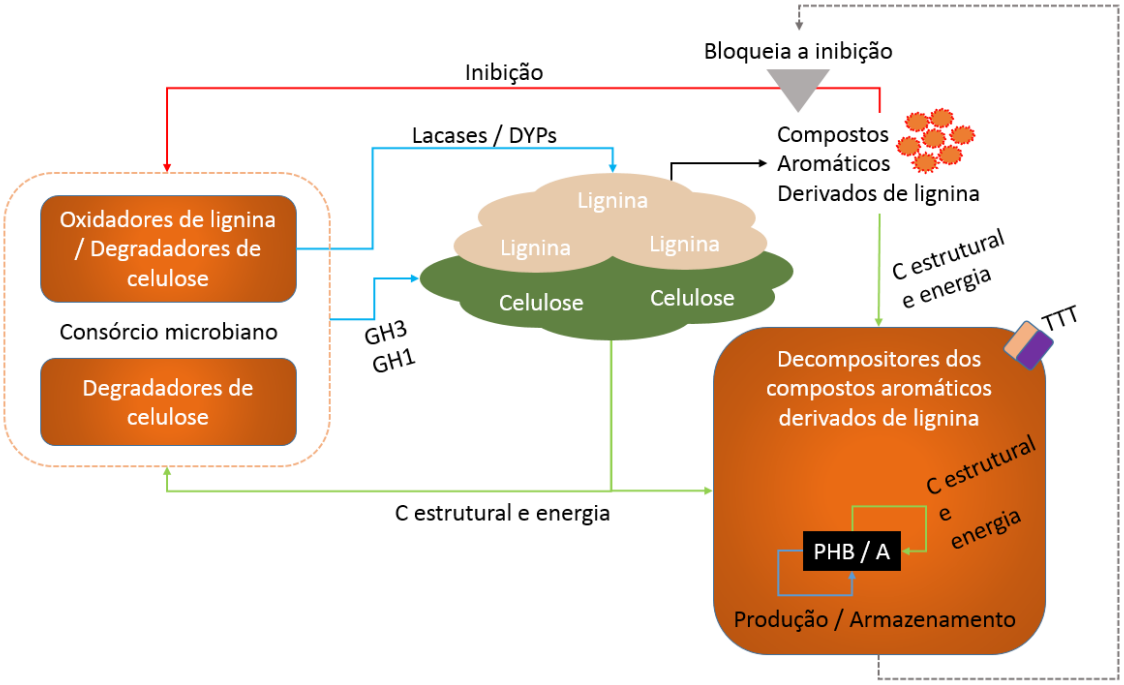

Figura 4.10. Esquema de efeito *priming* a partir das evidências genômicas. As setas verdes mostram relações de efeito benéfico para as comunidades, as setas azuis representam a secreção de ectozimas ou degradação de compostos, enquanto as setas

negras mostram a conversão do substrato em um produto. A seta vermelha mostra uma relação de inibição de crescimento e a seta cinza seu bloqueio. (Fonte: Próprio autor)

4.5 Conclusão

Foi possível recuperar 52 PGs não redundantes de 106 metagenomas previamente depositados, obtidos pelo sequenciamento Illumina do DNA ambiental total de 30 estações na bacia do rio Amazonas. Este é o primeiro relato do estabelecimento de PGs procarióticos na bacia do rio Amazonas. PGs aqui relatados mostraram maquinarias completas para processar OM terrestre, além de um conjunto distinto de espécies UFAC, via sistema TTT. Esses genomas ainda revelaram uma tendência do sistema TTT estar acoplado ao armazenamento de carbono via síntese de PHB e ao uso de compostos aromáticos derivados da lignina, embora haja exceções.

As beta-proteobactérias mostraram-se o grupo envolvido principalmente nos processos de fluxo de carbono na bacia do rio Amazonas, apesar de estabelecerem uma rede estruturada espacialmente com outros organismos ao longo do curso do rio. Foi possível observar a tendência de organismos degradadores da OM terrestre, estabelecendo consórcios de degradação.

Sob a nova ótica do contexto genômico, o modelo de efeito *priming* previamente estabelecido no capítulo 3 pôde ser revisto e teve melhorias adicionadas ao seu escopo. Deste modo, determinamos os principais mecanismos de degradação e qual o potencial efeito *priming* que regula a degradação de OM terrestre no rio.

Em resumo, esses PGs derivados da bacia do rio Amazonas são alvos interessantes para estudos visando uma maior compreensão da bacia do rio Amazonas,

uma vez que representam os organismos mais abundantes nesse sistema e podem ajudar a construir um painel mais completo de ecossistemas tropicais aquáticos.

Capítulo 5 – Desenvolvimento do dispositivo automático de *binning* por referência (BEAF)

5.1 Introdução

As novas tecnologias de sequenciamento em larga escala reduziram rapidamente os custos de sequenciamento (MARDIS, 2011; MUIR et al., 2016) permitindo com que diversas áreas dependentes de sequenciamento massivo fossem impulsionadas. O aumento do volume de dados gerado criou demandas por ferramentas mais rápidas e por análises bioinformáticas mais profundas (MUIR et al., 2016). Diversas plataformas têm sido desenvolvidas para este fim, como o UPARSE (EDGAR, 2013) e IMP (NARAYANASAMY et al., 2016).

Há diversos problemas com as aplicações atualmente disponíveis, como: longo tempo de processamento e a requisição de altos níveis de processamento (NARAYANASAMY et al., 2016). Além desses problemas, tecnicamente há alguns pontos que devem ser levados em conta. Embora permitam uma melhora na velocidade e diminuição do custo de processamento, algumas estratégias adotadas por essas aplicações, como por exemplo, aquelas heurísticas (perfis HMM) e as por frequência de k-mers, podem inserir *bias* nos resultados. O grande problema de aplicações deste tipo é que a cada execução, elas tendem a apresentar resultados diferentes (EDDY, 2011; FINN; CLEMENTS; EDDY, 2011; LEE; PARK, 2014; YOON, 2009).

Por fim, o grande número de aplicações necessárias para o tratamento completo dos dados, aliado à necessidade de especialização para lidar com eles, exclui a maioria dos usuários que não podem lidar com *pipelines* que não sejam totalmente automatizados.

Considerando esses desafios, desenvolvemos o BEAF, uma *pipeline* automatizada capaz de trabalhar em diferentes conjuntos de dados, executando todas as etapas, desde a edição de *reads* até a montagem e a anotação de ORFs. O BEAF é mais simples e fácil de usar, não exigindo conhecimentos específicos de bioinformática, além de funcionar adequadamente em computadores portáteis de uso pessoal.

5.2 Material e métodos

O programa BEAF foi construído com o auxílio do biotecnologista Guilherme Bonotti Coppini, meu orientado direto durante sua iniciação científica, que inclusive apresentou seu trabalho de conclusão de curso de graduação em biotecnologia, intitulado "Desenvolvimento de um módulo de análises taxonômicas do gene 16S rRNA de metagenomas para o programa BEAF (*Referenced **B**inning **E**ngine for **A**utonomous **F**inding*): BEAF16S" e recebeu menção honrosa por parte do CoPICT (Coordenadoria dos Programas de Iniciação Científica e Tecnológica) da UFSCar por este trabalho.

O programa BEAF foi projetado como um *workflow* para a análise em todos os estágios de dados de sequenciamento em larga escala, especialmente metagenômicos, realizando todas as etapas do processo, desde a limpeza de *reads* até a busca por ORFs nos *contigs* finais, de forma automatizada e mostrando baixo custo de processamento. Ele consegue categorizar as *reads* chegando até mesmo ao nível de sequências de genomas, proteínas preditas ou genes, utilizando-se para tal de buscas referenciadas (Figura 1). O programa consiste basicamente de um *script* modularizado escrito para ambiente *bash*/*shell*, com etapas em *Python* e *Perl* e se apresenta disponível na plataforma GitHub: https://github.com/celiosantosjr/BEAF. Seus códigos foram

testados e executados em um laptop com sistema operacional Ubuntu 18.1, contendo quatro núcleos *hyper-threaded*, 16 GB de RAM. Para otimizações e comparação, o código também foi rodado em um servidor do Laboratory of Microbial Processes & Biodiversity, coordenado pelo Prof. Dr. Hugo Sarmento associado ao Departamento de Hidrobiologia – DHB/UFSCar, com 24 processadores, 64 GB de RAM e sistema operacional Linux Ubuntu 14.04 LTS

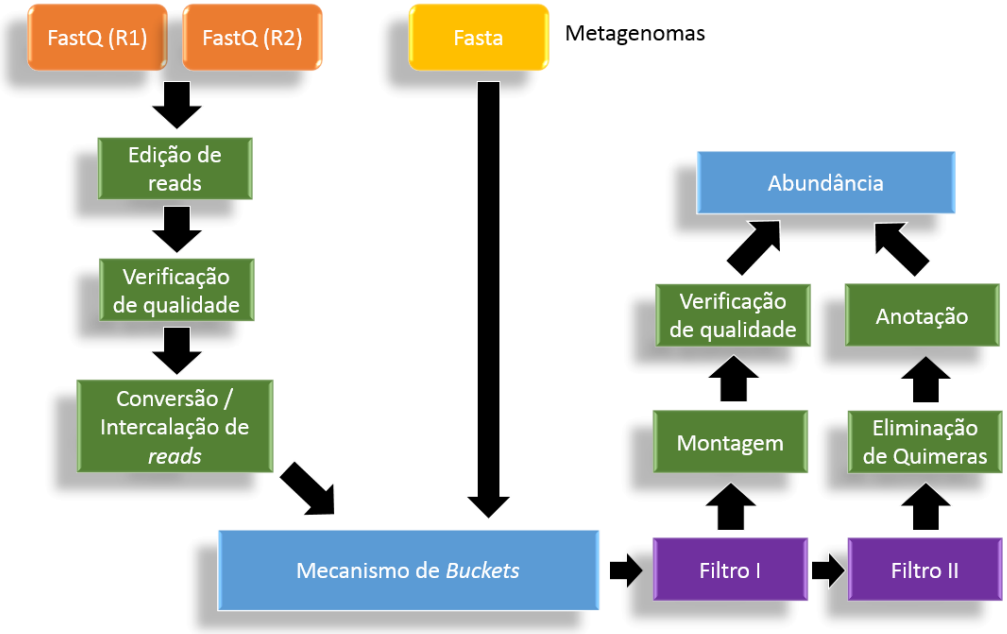

Figura 5.1. Funcionamento do programa BEAF. Utilizando um banco de dados referência (contendo genomas, genes ou proteínas), o programa encontra as *reads* ortólogas e as monta em *contigs*. Esses *contigs* então são varridos para que se encontrem as ORFs e por fim há uma anotação das mesmas. (Fonte: Próprio autor)

O *input* do programa permite a configuração de trabalhos em batelada, sendo uma tabela organizada em 7 colunas, respectivamente:

1. T1: discrimina-se 'G' para analisar genomas, ‘T’para a análise de taxonomia, 'P' para analisar proteínas e 'N' com sequências nucleotídicas de proteínas.
2. T2: discrimina-se 'R' para arquivos FASTQ de bibliotecas do tipo *paired-end*; 'I' para um arquivo FASTQ intercalado ou de bibliotecas do tipo *single-end*; ou 'F' para um arquivo FASTA.
3. R1: direciona o BEAF para o caminho completo do arquivo de sequenciamento, no caso o R1 (sequências *forward*).
4. R2: direciona o BEAF para o caminho completo do arquivo de sequenciamento, no caso o R2, no caso de arquivos de amostra intercalados, o usuário deve inserir 'NA' neste campo, caso esteja no modo taxonômico.
5. Ref: o endereço completo do arquivo de referência para o uso do BEAF, ou “NA” no caso do uso de apenas sub-referências.
6. SubRef: indica uma pasta, contendo arquivos de sub-referência, se a análise é uma categorização de genomas, o usuário deve digitar 'NA' ou inserir cepas da espécie desejada.
7. Out: designa o nome desejado para a pasta de saída gerada.

Esta configuração facilita o processo para o usuário e permite uma análise funcional dos dados de forma ampla, uma vez que múltiplos dados em diferentes níveis (genomas, proteínas e genes) podem ser coletados numa única execução do programa, sem a necessidade de reinicialização ou configuração constante do programa. O BEAF realiza todos os processos sem necessidade de qualquer interação do usuário, permitindo

o acompanhamento do progresso por meio de mensagens na tela do *prompt* de comando.

Inicialmente, o programa foi dividido em 3 modos:

- Genomas: busca *reads* homólogas ao (s) genoma (s) de referência, depois realizando a montagem das mesmas e comparação dos *contigs* obtidos com o genoma referência;
- Genes e proteínas: busca *reads* homólogas às sequências referência das famílias de genes ou proteínas, montando-as e predizendo possíveis ORFs;
- Taxonomia: também chamado 16S, identifica *reads* com identidade com bancos de dados de genes âncoras filogenéticas (por exemplo 16S, 18S e ITS), realizando assim uma análise do perfil taxonômico da amostra, com dados da diversidade e abundância de espécies.

Desta forma há uma ordem de trabalho no BEAF, em que primeiro se executa o modo 16S, com o fim de identificar-se as espécies mais abundantes na amostra. Posteriormente, obtém-se os genomas referência dessas espécies e, então, executa-se o modo genômico, obtendo-se assim maior eficiência do processo. O modo de genes/proteínas é inerente à identificação de espécies e genomas.

Diversos mecanismos foram implementados com o fim de reduzir o tempo de processamento e aumentar a eficiência. Um deles foi a divisão do arquivo de *reads* após sua edição e conversão para o formato FASTA em arquivos menores, também

conhecidos como *buckets* de no máximo 256 MiB. A pesquisa de homólogos é realizada em 2 níveis, visando-se a redução do tempo de busca:

1) 1° filtro: Nesta etapa os *buckets* são buscados com o algoritmo do programa USEARCH versão 10 (EDGAR, 2010), no caso de genes vs. genes e no caso de genes vs. proteínas. Este filtro consiste nos seguintes parâmetros de corte: identidade (95% para genomas e 25% para famílias de proteínas/genes) e e-*value* (1e-20 para genomas e 1e-5 para famílias de proteínas/genes) em ambas as fitas. Esta etapa visa a eliminação do maior número de *reads* não homólogas possível, reduzindo assim o *dataset* para a próxima etapa, que tem uma maior acurácia;
2) 2° filtro: Nesta etapa a busca por *reads* homólogas é realizada com níveis de corte mais altos o possível (identidade de nucleotídeos mínima de 90-97%; identidade de aminoácidos mínima de 45%, comprimento mínimo de alinhamento de 25 aminoácidos e cobertura de no mínimo 50% do comprimento da sequência), eliminando-se assim qualquer ruído e aumentando a acurácia dos resultados. Neste filtro é possível utilizar sub-referências de genomas, para identificação de cepas; ou mesmo de subfamílias de proteínas/genes, além do modo taxonômico convencional que o utiliza para eliminar resultados falso-positivos.

Para a recuperação de genomas, as *reads* são montadas com o programa SPADES (BANKEVICH et al., 2012) e a qualidade dos *contigs* é posteriormente testada contra o genoma de referência por meio do programa QUAST (GUREVICH et al., 2013).

Para genes, proteínas e no modo taxonômico, o próximo passo envolve a montagem de *reads* usando o programa SPADES (BANKEVICH et al., 2012) no modo "*only assembler*". Abundância de acertos é calculada como correspondências por milhão de *reads* (ppm). Então, para ambas as análises, a previsão de ORFs ocorre usando um *script perl* descrito por outro autor (SENALIK, 2013).

Existe um sistema recursivo que testa a eficácia da montagem e reanalisa a lista de k-mers quando esta falha. No entanto, pode-se perder novos arranjos do genoma ou variantes de proteínas/genes, uma vez que a análise é baseada em uma pesquisa de similaridade. As ORFs encontradas dessa forma são reanalisadas por meio de nova busca de homólogos contra o banco de dados, de modo que as correspondências sejam duplamente confirmadas.

Se os arquivos de referência ou sub-referência forem grandes bancos de dados é uma opção viável reduzir o tempo de cálculo, reutilizando bancos de dados mantendo-os já formatados no estilo BLAST com o parâmetro "--KeepBlastsDBs". O BEAF testará o arquivo de entrada antes de iniciar os procedimentos e indicará se ele falha no teste, indicando suas possíveis correções para o usuário.

Desta forma o BEAF gera um *output* completo para o usuário contendo:

1) Todos os modos: Arquivo contendo o resumo de todas as execuções: Log.tsv;
2) Modo genoma: Arquivo contendo o resumo da análise: Log.txt; Arquivo contendo as leituras identificadas como ocorrências: hits.fa.gz; Arquivo contendo os *contigs* ou *scaffolds* obtidos na montagem dos hits:

scaffolds.fa.gz; Arquivo contendo a avaliação da montagem: assessment.tar.gz; Pasta contendo os resultados da qualidade das *reads* obtido com o programa FASTQC (ANDREWS, 2017) inicial e final: FASTQCresults; Arquivo contendo as ORFs identificadas: ORFs.fa.gz.

3) Modo de genes/proteínas: Pasta contendo os resultados da análise de qualidade de *reads* pelo programa FASTQC (ANDREWS, 2017) antes e depois da edição de *reads*: FASTQCresults; Arquivo contendo o resumo da análise: Log.txt; Arquivo contendo apenas a tabela *SubReference*: SubRefs.tsv; Arquivo contendo as leituras identificadas como ocorrências: hits.fasta.gz; Pasta contendo os resultados do Blastx/n: blast_hits; Pasta contendo sequências após a filtragem com o Blast: read_hits; Pasta contendo *contigs* após a montagem: contigs; Pasta contendo arquivos identificados das ORFs por *contigs*: ORFs.

5.3 Resultados e Discussão

Testou-se exaustivamente o programa BEAF, por meio da análise de duas comunidades simuladas, a primeira contendo 73 genomas (NARAYANASAMY; JAROSZ; WILMES, 2016) e outra obtida por sequenciamento Illumina contendo o material genômico de 26 organismos (SINGER et al., 2016). Para os testes com a pipeline BEAF utilizou-se por convenção no modo de taxonomia o arquivo R1 (fita *forward*) gerado pelo sequenciamento Illumina™. Os resultados apresentados a seguir são um exemplo da aplicação do programa BEAF e dos resultados possíveis de se obter a partir dessas análises.

5.3.1 Taxonomia

Os testes com diferentes níveis de clusterização do banco de dados referência *Greengenes* (DESANTIS et al., 2006) para as buscas por homologia com a comunidade simulada IMP-Narayanasamy (NARAYANASAMY; JAROSZ; WILMES, 2016) mostram que o banco referência utilizado causa pouca influência no resultado final, assim há uma variação muito baixa do número de sequências e unidades taxonômicas operacionais (OTUs) encontradas em diferentes condições.

Diferentes níveis de clusterização afetam consideravelmente o tempo de execução do programa, assim avaliou-se a porcentagem de cada tipo de acerto e erro (Figura 5.2). As condições que garantem um baixo tempo de processamento e resultados satisfatórios foram encontradas quando utilizou-se o banco de dados referência do *Greengenes* (DESANTIS et al., 2006) clusterizado a 97% de identidade para a primeira busca por homologia, e a 100% de identidade para a segunda busca, apresentando um número maior de ortólogos classificados corretamente selecionando-se, assim, a condição padrão de análise do programa.

É importante mencionar-se que o programa BEAF também pode utilizar outras bases de dados referência, uma vez que é uma *pipeline* flexível e facilmente ajustável às necessidades do usuário. A alteração da base de dados é feita por meio da indicação do banco de dados a ser utilizado no arquivo de configuração inicial, extraindo-se os *index* para identificação de OTUs e táxons no novo banco (isso pode ser feito através do instalador disponível *online* juntamente com o programa BEAF).

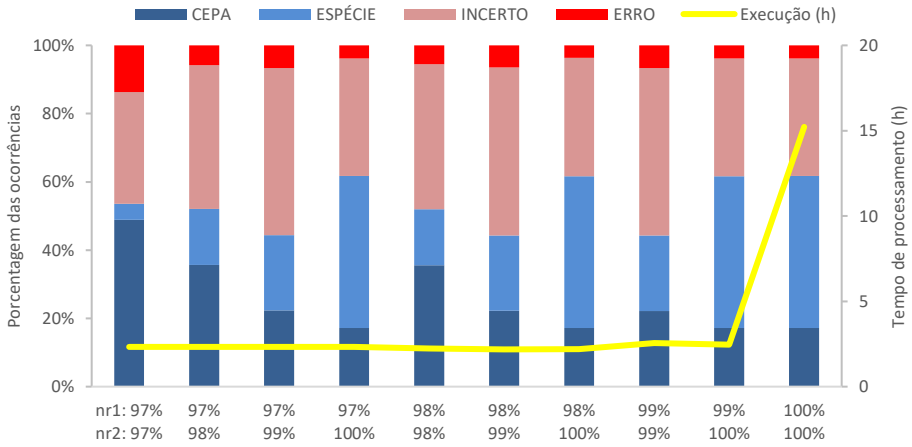

Figura 5.2. Diferentes níveis de clusterização da base de dados referência e seu efeito na busca por ortólogos com o programa BEAF. O banco de dados Greengenes foi utilizado com duas clusterizações ao nível de identidade indicado no gráfico para as buscas com os filtros 1 e 2. O grau de erros, acertos (em nível de espécie e cepa) e de resultados incertos foi mostrado. Fonte: Adaptado de Coppini (COPPINI, 2018).

Testou-se combinações pareadas de *cut-off* por identidade para o primeiro filtro (90% e 95% de identidade) e para o segundo filtro (95%, 99% e 97% de identidade), a fim de se realizar uma avaliação dos melhores parâmetros de filtragem por homologia. Analisando-se a porcentagem de erros e acertos nas condições testadas observou-se uma pequena influência de parâmetros como *maxaccepts* e *maxrejects*, havendo uma diferença inferior a 0,07% nas proporções de acertos e erros finais, dessa forma estes parâmetros foram estabelecidos como 5000, a fim de se acelerar as buscas. As melhores identidades para os filtros de busca foram, respectivamente, 90% e 97% (Figura 5.3), devido a maior sensibilidade de seus resultados a nível da identificação de cepas, espécies e gênero. Observou-se uma variação muito pequena na sensibilidade dos resultados do programa BEAF quando utilizadas diferentes condições de teste,

mostrando que há um alto nível de reprodutibilidade e uniformidade dos resultados. Esta característica também foi inferida quando o programa foi testado em máquinas diferentes, ainda obtendo-se resultados consistentes.

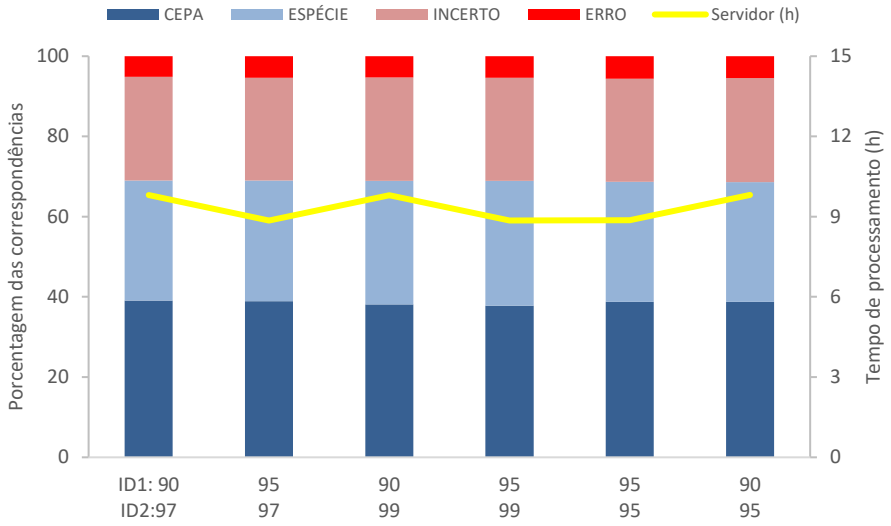

Figura 5.3. Performance do programa BEAF em modo taxonômico alterando-se o a identidade de busca de ortólogas na base de dados referência nos filtros de homologia 1 e 2. Os erros são tidos como resultados não confiáveis, enquanto os resultados incertos são resultados que podem ter sido indevidamente classificados, por conta de inconsistências durante a anotação do banco de referência. Os resultados corretos foram dados separadamente em nível de espécie e cepa. Fonte: Adaptado de Coppini (COPPINI, 2018).

Utilizando-se as configurações ótimas determinadas pelos pré-testes, o programa BEAF foi utilizado para avaliar os microrganismos componentes da comunidade simulada MBARC-26, composta por 26 espécies. Apenas 65.4% das

espécies foram encontradas pelo programa em nível de cepa, no entanto, 92.3% dos microrganismos componentes da comunidade simulada foram identificados, se também considerarmos espécies sem a informação de suas cepas. Há duas espécies que não puderam ser identificadas na amostra, *Natronobacterium gregoryi* e *Nocardiopsis dassonvillei*, sendo que *N. dassonvillei* também não pôde ser identificada pelo grupo de pesquisadores que caracterizou a comunidade simulada MBARC-26 (SINGER et al., 2016), indicando que esta espécie estava ausente na comunidade.

Os erros mais frequentes observados ao longo da execução do programa BEAF no modo taxonômico foram associados a anotação ineficiente dos bancos de dados referência, um problema recorrente na literatura (MCDONALD, Daniel et al., 2012; NELSON; MORRISON; YU, 2011; PETTENGILL; RAND, 2017; WERNER et al., 2012; ZHANG; SHAO; YE, 2012). A anotação dos bancos de dados referência não tem um padrão fixo, podendo apresentar divergências de classificação taxonômica entre diferentes bases de dados (MCILROY et al., 2015). Por consequência, há uma anotação com um grau de incerteza da taxonomia das OTUs em nível de espécie. Um exemplo de má anotação foi o microrganismo *Thermobacillus xylanilyticus*, componente da MBARC-26, cujas sequências representativas foram mal anotadas no banco de dados referência *Greengenes*, em que consta uma taxonomia apenas até o nível de gênero *Thermobacillus* "espécie desconhecida". Uma vez que o sistema implementado no programa BEAF modo taxonômico mostrou-se funcional, com erros reportados devido a inconsistências do banco de dados referência, há uma necessidade da geração de uma base de dados com anotação consistente, padronizada e sequências pouco redundantes.

A análise no modo taxonômico do programa BEAF, dada pelos *contigs* complementa a análise por contagem de *reads*, e permite a confirmação de parte dos

resultados, entretanto, apresenta alta taxa de falsos negativos, não devendo ser utilizada de forma independente (Figura 5.4).

Uma comparação dos resultados obtidos com o programa BEAF no modo taxonômico e dos resultados obtidos pela *pipeline* miTAGs (LOGARES et al., 2014) para a comunidade simulada MBARC-26 (SINGER et al., 2016) com a base de dados referência *Greengenes* (DESANTIS et al., 2006), foi realizada com fins de validação técnica (Figura 5.4). Apesar do aparente ganho de tempo de execução com a *pipeline* miTAGs, o programa BEAF teve uma melhor performance à nível de acertos (Figura 5.4). Outro ponto, é que o ganho por processador com o programa BEAF, sugere uma maior capacidade de paralelização, ou seja, aumentando-se o número de processadores utilizados, pode-se reduzir ou até mesmo eliminar-se a diferença de tempos de execução entre as 2 *pipelines*. Os resultados das duas *pipelines* apresentam alta correlação (R^2 = 0.9875), e o teste de similaridade de Kolmogorov-Smirnov entre essas distribuições, sugere que não há diferença estatística (p-*value* = 0,950), significando uma mesma proporção dentro das espécies identificadas entre as duas metodologias e, portanto, uma conformidade dos seus resultados.

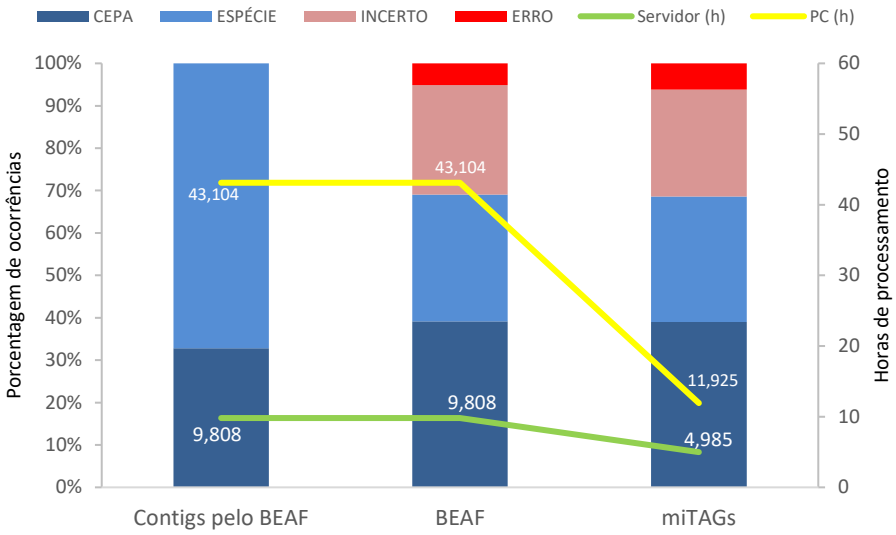

Figura 5.4. Comparação entre as *pipelines* BEAF-16S e miTAGs. A primeira coluna indica os resultados encontrados em termos de *contigs* obtidos no módulo taxonômico do BEAF, enquanto a segunda mostra os dados referentes à contagem das *reads*. O tempo de execução das *pipelines* foi contabilizado como os valores nas linhas amarela, para o tempo de execução em um servidor utilizando 21 processadores, e marrom, para o tempo de execução em um laptop pessoal utilizando 7 processadores. Fonte: Adaptado de Coppini (COPPINI, 2018).

5.3.2 Recuperação de genomas

Tentou-se a recuperação dos 26 genomas constituintes da comunidade simulada MBARC-26 a partir dos genomas-referências disponíveis no NCBI. A recuperação dos genomas foi verificada como um módulo bem-sucedido do programa BEAF, em que se recuperou mais de 90% dos genomas de 25 das 26 espécies que compuseram a comunidade MBARC-26 (Figura 5.5).

De modo geral, o programa BEAF recuperou frações genômicas menores de todos os outros organismos (Figura 5.5) quando comparado aos dados de Singer et al. (SINGER et al., 2016). No entanto, há uma diferença entre as duas abordagens, em que o programa BEAF obtém os *contigs* do genoma, pela montagem das *reads* com o programa SPADES (BANKEVICH et al., 2012), e avalia a fração genômica recuperada com o programa QUAST (GUREVICH et al., 2013). Singer et al. (SINGER et al., 2016), por sua vez, utilizaram-se de um mapeamento de *reads* para obter essa estatística. Uma vez que a porcentagem de *reads* montadas ao nível de *contigs* tende a ser muito inferior ao total de *reads* que inicia o processo, esta informação não seria contabilizada nos cálculos de fração genômica do programa BEAF, subestimando esses valores.

Apenas 0,004% do genoma de *Nocardiopsis dassonvillei* pôde ser recuperado, remetendo provavelmente ao efeito de artefatos de técnica, algo observado anteriormente por outros autores (SINGER et al., 2016). Dentre os outros microrganismos componentes da comunidade simulada MBARC-26, a menor fração genômica recuperada pertenceu ao *Terriglobus roseus*, com 91,04% recuperado. Algo similar foi anteriormente observado por Singer et. al (SINGER et al., 2016), que obteve níveis de recuperação desse genoma de 83,95%.

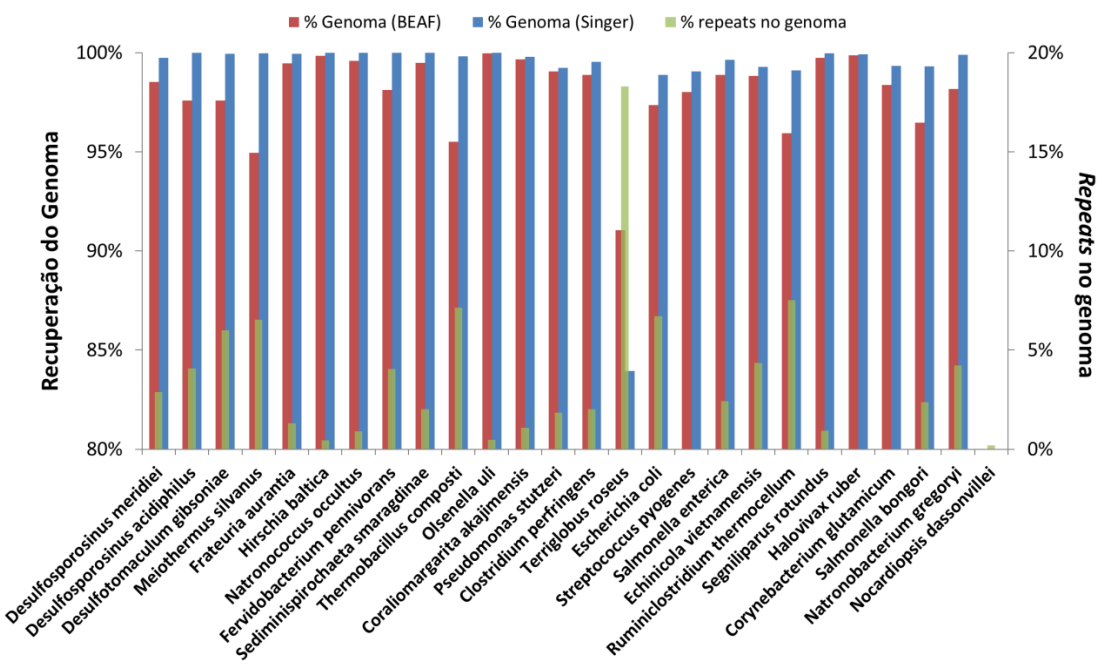

Figura 5.5. Comparação da fração genômica recuperada dos organismos que compõe a comunidade simulada MBARC-26 pelo programa BEAF e Singer et al. (SINGER et al., 2016). A porcentagem dos genomas recuperados por Singer et. al (2016) é maior por se tratar de um cálculo com base no mapeamento direto de *reads*, alinhando-as diretamente ao genoma referência, diferentemente do programa BEAF, onde a fração genômica foi calculada com base em *contigs* montados e alinhados contra a referência. Fonte: Adaptado de Coppini (COPPINI, 2018).

Uma explicação possível para uma melhor recuperação do genoma de *Terriglobus roseus*, uma vez que seu tamanho e conteúdo GC não são discrepantes dos de outras espécies da MBARC-26, é que há uma maior porcentagem de regiões de repetição neste genoma, chegando a ser o dobro do de outras espécies. Apesar de não haver uma repetição deste padrão de recuperação de acordo com as taxas de repetição genômica, observou-se que os organismos com genomas repetitivos foram aquelas cujas

variações entre a porcentagem de recuperação obtida por Singer et al. (SINGER et al., 2016) e através da *pipeline* BEAF foram as mais pronunciadas (Figura 5.5).

5.3.3 Binning *de famílias de proteínas*

Testou-se o mecanismo de *binning* para 3 famílias de proteínas do banco de dados referência: buk (quinase de butiratos), but (4-hidróxibutirato co-A transferase) e hydA (hidrogenase de ferrodoxina) advindas do banco de dados Fungene (http://fungene.cme.msu.edu/). Esses bancos foram usados para uma análise da comunidade simulada de Narayanasami et al. (NARAYANASAMY; JAROSZ; WILMES, 2016), que contém genomas de 73 espécies de bactérias codificado em 9.935.135 *paired reads* após o processo de edição por qualidade.

A análise feita com o programa BEAF utilizou sete núcleos de processamento e demorou cerca de 5,23 h para ser completada em um laptop de uso pessoal com 16 GiB de memória RAM e sistema operacional Ubuntu 18. No total, 42.371 ocorrências (2.134 ppm) foram selecionadas para o segundo filtro de homologia e análise específica por família de proteína. Nossos resultados mostram que o primeiro filtro de homologia é capaz de acelerar o processo de busca de ortólogos em 2,25 vezes na segunda filtragem dos dados.

Após a segunda filtragem, as famílias de proteínas testadas tiveram diferentes números de ocorrências associadas: buk (3.648), but (2.774) e hydA (17.048). Isso mostra que de acordo com sua abundância nos genomas de origem os genes e proteínas são categorizados de modo diferente. A partir da análise de montagem que o programa BEAF performa através do *software* SPADES (BANKEVICH et al., 2012) obtivemos o

seguinte número de *contigs* distribuídos de acordo com a família de proteínas referência: buk (47), but (22) e hydA (220). Esse resultado sugere que um maior número de *reads* obtido no segundo filtro pode se relacionar a um maior número de *contigs* final. As estatísticas das montagens mostram que os *contigs* obtidos foram relativamente longos (comprimento médio de 904 pb com os maiores *contigs* variando entre 3.617 pb e 1.594 pb. Essas estatísticas de montagem refletem um padrão de qualidade elevado para o conjunto de dados final, que indica uma acurácia associada aos resultados obtidos.

Esses *contigs* foram avaliados quanto a presença de ORFs completas maiores que 150 pb, cada família de proteínas conseguiu recrutar um número diferente de ORFs: buk (108), but (56) e hydA (490). Essas ORFs tiveram um comprimento médio de 459 pb, com seu comprimento máximo variando de 1.125 pb a 1.953 pb. Essas estatísticas revelam que as proteínas preditas encontradas tinham uma estrutura grande o suficiente para possuir atividade enzimática.

Um alinhamento com as proteínas preditas de tamanho superior a 500 resíduos encontradas, utilizando-se a família de proteínas buk como referência revelou que diversos *motifs* estavam presentes em *tandem*, sendo conservados nas diferentes proteínas. Nenhuma delas apresentou tendências quiméricas, de acordo com as análises e anotações realizadas via Blastp que o programa BEAF realiza. Em contrapartida, pudemos encontrar durante a anotação apenas 44% das proteínas preditas como ortólogas das proteínas referências do banco de dados buk, representando um total de 39 sequências diferentes daquele banco. Em uma análise posterior com o *software* Interproscan (https://www.ebi.ac.uk/interpro/sequence-search), foi possível detectar que as enzimas com mais de 900 resíduos apresentaram, geralmente, os dois sítios típicos da

butirato quinase (186-203) e acetato quinase (18-24). Esses resultados demonstram que o programa BEAF foi eficiente para a categorização de *reads* ao nível de genes e proteínas preditas com alta confiabilidade e resulta num esforço importante para a bioprospecção de projetos metagenoma.

5.3.4 Prospecção de famílias de genes

Realizou-se dois testes neste módulo: um deles com um banco de dados contendo sequências referência para elementos de transposição de bactérias (3.454 sequências), protistas (2.497 sequências) e fungos (373 sequências); e outro contendo apenas o gene da DNA polimerase subunidade beta (4.350 sequências). Ambos os bancos de dados referência utilizados neste estudo foram obtidos por meio do *download* das sequências a partir do NCBI.

No teste envolvendo o banco de dados do gene da DNA polimerase subunidade beta (DNApolB), o programa BEAF levou cerca de 2,1 h de processamento com 7 núcleos de processamento em um laptop de uso pessoal com 16 GiB de memória RAM e sistema operacional Ubuntu 18. Um total de 308.471 ocorrências foram selecionadas a partir da primeira filtragem dos dados, representando uma abundância de 15.539 ppm. Essa alta abundância inicial era esperada já que o gene da DNApolB é considerado um gene do tipo *housekeeping* e, portanto, presente em todos os genomas avaliados.

A análise com o banco de dados da DNApolB levou a um número final de ocorrências de 703 *reads* depois da segunda filtragem, reduzindo a abundância para 35 ppm. Obteve-se 8 *contigs* após o processo de montagem dessas *reads*, cujo comprimento médio foi de 1.140 pb e o comprimento máximo de 1.375 pb. A predição

de ORFs gerou 36 ORFs, cujo comprimento médio foi de 284 pb e máximo de 861 pb. Essas ORFs foram filtradas para um comprimento mínimo de 800 pb, encontrando-se apenas 4 finais. As ORFs selecionadas foram traduzidas e alinhadas mostrando uma identidade de 100% com conservação de comprimento (861 pb) e completa de motivos protéicos. Obteve-se basicamente 4 cópias do gene em diferentes contextos genômicos, uma vez que os *contigs* continham exatamente a mesma proteína, mas eram distintos entre si.

No teste envolvendo elementos de transposição, os dados relativos a ORFs e proteínas preditas não podem ser aplicados, pois não há uma identificação direta delas nesse contexto. Portanto, os elementos transponíveis identificados foram analisados quanto à sua identificação em termos de anotação e alinhamentos. O programa BEAF trabalhando nas mesmas condições já citadas levou cerca de 3,99 h para terminar a análise. Ao todo, 308.471 ocorrências foram selecionadas a partir do primeiro filtro e apenas 135.675 foram selecionadas no segundo filtro, contabilizando uma abundância de 6834 ppm. Elementos genômicos móveis tendem a se apresentar em *tandem* de modo repetitivo, o que pode aumentar sua representação durante o sequenciamento.

Na análise por subreferência, não foi encontrada nenhuma ocorrência para protistas e apenas 36 para fungos, o que reflete a composição da comunidade simulada, que não possui organismos eucarióticos. Obteve-se 687 *contigs* para a montagem das *reads* filtradas com a subreferência de elementos de transposição bacterianos, com comprimento médio de 827 pb e máximo de 9.498 pb. Esses *contigs* renderam cerca de 1.339 ORFs, com comprimento médio de 367 pb e máximo de 2.451 pb. Na análise da anotação destes *contigs*, foram identificados 634 sequências ortólogas com 260 sequências de elementos de transposição do nosso banco de dados. Diferentemente das

análises anteriores, observou-se sinais de ortologia conflitantes, em que diferentes referências de elementos de transposição de espécies distintas possuíram identidade com o mesmo *contig* (~5%), sugerindo sinais de quimerismo nessas sequências, o que cria a necessidade de sua filtragem *a posteriori*.

5.4 Conclusão

O BEAF revelou-se um fluxo de trabalho integrado e eficiente na análise de dados do tipo WGS, tanto com arquivos no formato FASTA, quanto FASTQ de bibliotecas *singled-end* e *paired-end*. Ao contrário das *pipelines* anteriores, o BEAF é uma aplicação que utiliza pouca memória e processamento, podendo funcionar rapidamente na categorização de genomas e prospecção referenciada de genes e proteínas. Além disso, o BEAF é uma aplicação flexível devido à sua configuração automatizada. Os arquivos de *log* fornecem relatórios em cada estágio analítico. O construtor de referências integrado com um montador eficiente e uma avaliação de qualidade dos *contigs*, torna o BEAF uma ferramenta valiosa. A implementação dessa pipeline permite que os usuários concluam, de maneira rápida e eficiente, as prospecções primárias em metagenomas e sequenciamento de genomas obtidos com tecnologias de alto rendimento. A bioprospecção de sequências pelo sistema de subreferências de genes e proteínas foi eficiente reiterando a utilização deste programa neste tipo de abordagem em dados metagenômicos economizando substancialmente o tempo de análise de grandes conjuntos de dados.

Capítulo 6 – Sistema de gerenciamento de genes da bacia do rio Amazonas: AGSSY

6.1 Introdução

No presente trabalho, em seu capítulo 3, foi apresentado um catálogo de genes microbianos não redundantes da bacia do rio Amazonas. Este catálogo de genes, com seus mais de 3,7 milhões de genes, possui preciosas informações tanto de cunho biotecnológico quanto ecológico. Utilizando essas informações foi possível analisar as comunidades do rio Amazonas e tirar conclusões importantes sobre seus nichos, entretanto, ainda não se abordou as questões pertinentes à bioprospecção.

A bioprospecção pode ser em termos funcionais ou de ortologia, a primeira utiliza-se da busca direta de termos nos bancos de dados anotados, como: classes funcionais, cofatores ou nome de enzimas diretamente; já a segunda utiliza-se da busca de sequências diretamente por meio de algoritmos de alinhamento e com base nos scores, atribuindo-se funções de interesse. Nesses dois casos a melhor metodologia é uma abordagem polifásica onde pode-se por meio de etapas envolvendo os dois princípios confirmar os dados obtidos. O sistema do catálogo de genes AMnrGC é interessante, pois apresenta uma anotação completa com sete bancos de dados referência diferentes, além de ter a possibilidade da busca direta de sequências, assim, ele é ideal para bioprospecções típicas.

Pensando nisso, o presente capítulo vai abordar um sistema de busca e gerenciamento de genes e proteínas preditas do AMnrGC que torna a bioprospecção acessível e rápida, também facilitando quaisquer estudos que envolvam a busca por classes de proteínas específicas e suas características.

6.2 Implementação

O sistema de gerenciamento de genes do rio Amazonas, ou ***A****mazon River* ***Ge****nes* ***S****earching* ***Sys****tem* – AGSSY, foi produzido para lidar com essas questões. Ele consiste num esquema típico de *script* feito em linguagem híbrida *bash*/*perl* e *bash*/*python* que atua de modo interativo com o usuário. O código do programa apresenta-se disponível no Apêndice 11.

O AGSSY funciona com uma plataforma gráfica (Figura 6.1) que permite inicialmente a busca de: (1) códigos de *clusters* ortólogos, famílias de proteínas, função celular, ou termos e palavras-chave; (2) sequências via alinhamento estatístico; (3) perfis de abundância gênica por metagenoma analisado e (4) sequências de proteínas e nucleotídeos com base nos seus códigos de identificação.

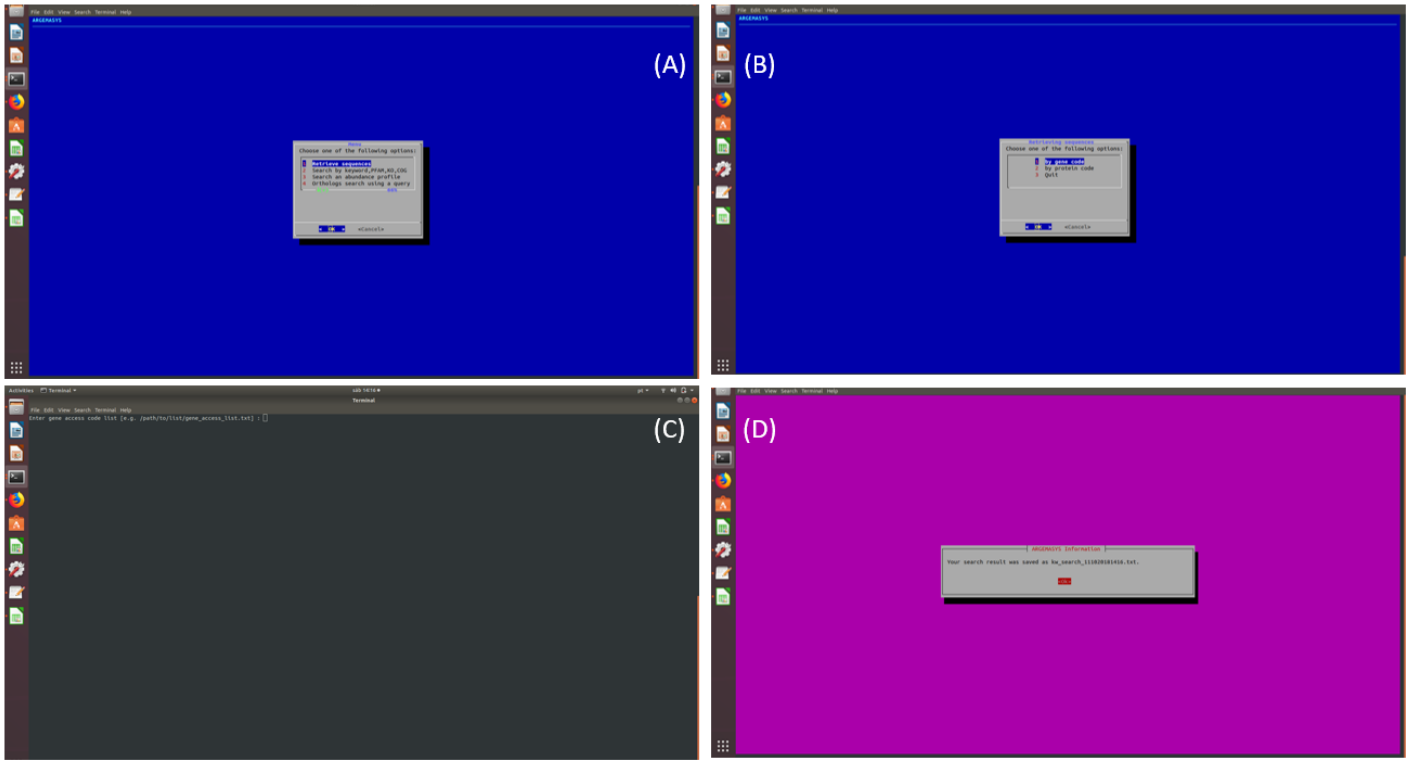

Figura 6.1. Telas do programa AGSSY. Em (A) temos o menu principal, o qual selecionando-se uma opção, redireciona o usuário até um sub-menu (B), e depois a uma janela de *input*, em que o programa pede seu arquivo de interesse de acordo com a opção. Por fim, após o processamento, o programa informa o *output* gerado para o usuário e pede a confirmação da mensagem (D). (Fonte: Próprio autor)

O AGSSY foi desenhado para trabalhar numa ordem específica, em um esquema polifásico. Primeiramente o usuário seleciona a busca por termos, onde há um sub-menu com 2 opções: (1) busca em um banco de dados específico ou (2) busca geral usando uma lista de palavras-chave. Caso o usuário escolha a primeira opção, então o programa o redireciona a um outro sub-menu, contendo desta vez mais 7 opções: (1) PFAM *numbers or terms*; (2) KO *numbers*; (3) GO *numbers*; (4) COG *numbers*; (5) eggNOG *clusters or terms*; (6) dbCAN *families or terms*, e (7) *Quit*. Todas estas opções de 1 a 6, quando selecionadas, redirecionam o usuário para uma tela do *bash* em que o AGSSY pede uma palavra-chave ou código a ser digitado diretamente. A partir desse termo, há uma busca mediada pelo dispositivo *zgrep* do Linux e o retorno de uma tela contendo o nome do arquivo gerado com as informações obtidas, o usuário então pressiona a tecla *Enter* e encerra o programa. Caso o usuário opte pela opção de busca geral usando uma lista de palavras-chave, o programa o direciona a uma janela negra do *bash* em que pede o endereço completo do arquivo contendo as palavras-chave de interesse. Uma vez preenchido o requerimento é analisado da mesma forma que o anterior, por meio do dispositivo *zgrep* e o programa edita e organiza os resultados para que tudo seja entregue ao usuário em um único arquivo com os resultados separados por banco de dados utilizado na anotação. Novamente, é gerado um arquivo de saída que é indicado numa nova janela do programa em que é requerido ao usuário pressionar a tecla *Enter* para que o programa seja encerrado.

Depois da busca por termos, deve-se fazer a busca por ortólogos, em que se seleciona a opção 4 do menu inicial (*Orthologs search using a query*). O usuário é direcionado a um sub-menu no qual é questionado sobre qual tipo de *input* ou *output* ele deseja: (1) Nucleotídeos e (2) Proteínas. Independentemente da opção selecionada, o

usuário será direcionado a uma tela do *bash* em que lhe é requerido o endereço completo do arquivo do tipo FASTA em que estão contidas as sequências de interesse. Caso a opção selecionada seja a (1), o AGSSY faz uma busca por meio do algoritmo *nhmmer* implementado no programa HMMER versão 3.1b1 (EDDY, 2011), com e-*value* máximo de 1e-5, retornando uma tabela de saída no modelo do DFAM (HUBLEY et al., 2016). Caso a busca seja por proteínas, o algoritmo *phmmer* também implementado no programa HMMER versão 3.1b1 (EDDY, 2011) será utilizado, com e-*value* máximo de 1e-5, retornando uma tabela de saída no modelo de tabela de domínios (*DOMTBLOUT*).

A busca por ortólogos é feita por meio de alinhamentos estatísticos utilizando-se perfis HMM-like, por duas principais razões: primeiro, porque eles levam em conta desvios relativos à mutação e são melhores em alinhamentos de ortólogos divergentes; e segundo, pois o tamanho dos bancos de dados do AMnrGC são muito grandes para uma busca formal por alinhamento, que demoraria de horas a dias, com um número relativamente pequeno de sequências referência. Assim, reunindo-se as duas razões adotamos os algoritmos implementados no programa HMMER que se tem mostrado precisos e rápidos o suficiente para buscas pequenas e relativamente longas.

Uma vez gerados os arquivos de busca por ortólogos, o usuário com comandos simples pode cruzar os dados de ambos os resultados e determinar quais genes e proteínas foram identificados por ambas as metodologias, abaixo um exemplo de código:

(1) Com o arquivo obtido da busca de termos no PFAM por exemplo:

```
awk '{print $1}' file | sort | uniq > list1
```

(2) Com o arquivo obtido na busca de ortólogos - proteínas:

awk '{print $1}' file | sort | uniq > list2

(3) Uma vez que o AMnrGC tem um padrão de diferenciação de genes e proteínas correspondentes pelo mesmo código, com a diferença de que as últimas possuem o sufixo "_1", pode-se também obter a lista de genes observados em comum entre 1 e 2:

sed 's/..$//' list2 | sort | uniq > list2.1

comm -1 -2 list1 list2.1 > common.genes.list

Uma vez com essas listas, basta selecionar-se a opção (1) "*Retrieve sequences*" do menu principal e depois selecionar-se respectivamente (1) "*by gene code*" ou (2) "*by protein code"*. Qualquer das opções que for selecionada no sub-menu, redireciona o usuário para uma tela do *bash* em que o programa pede pelo endereço completo do arquivo do tipo lista de códigos de acesso. Uma vez fornecido o arquivo, o programa busca e extrai do AMnrGC as sequências desejadas por meio de um código de *Perl* e são armazenadas num arquivo FASTA indicado na tela final do programa, em que o usuário é compelido a pressionar a tecla *Enter* para finalizá-lo.

Por fim, com a lista de códigos de acesso aos genes, pode-se ainda extrair os perfis de abundância deles para cada um dos 106 metagenomas utilizados na produção do AMnrGC. Ao selecionar-se a opção (3) *"Search an abundance profile"* pode-se selecionar o perfil de abundância para um gene (1) ou a opção de uma lista de genes (2) no sub-menu. Ao selecionar-se a opção 3-1, digita-se diretamente na janela do programa o código do gene e o AGSSY gera um arquivo do tipo tabela com 107 colunas e 2 linhas, a primeira um cabeçalho e a segunda contendo na primeira coluna o nome do

gene e nas outras colunas o valor de TPM (Transcriptos por milhão de *reads*) contido em cada metagenoma. Caso a opção selecionada seja 3-2, digita-se o nome da lista de genes e a busca gera um arquivo similar ao anterior. Em ambos os casos é gerado um arquivo final que é indicado pelo programa, de modo similar aos outros módulos.

Para um exemplo do funcionamento do AGSSY utilizamos a busca do gene LacZ de *Escherichia coli,* um bom exemplo de uma enzima com atividade interessante e potencial biotecnológico. As beta-galactosidades foram então buscadas do seguinte modo:

(A) Selecionou-se a opção de busca por termos e palavras-chave, depois a busca por uma lista de termos em todos os bancos de dados ao mesmo tempo, buscando-se pelo termo "*beta-galactosidase*".

(B) Gerada a lista A, selecionamos os resultados do banco de dados KEGG para gerar-se a lista 1, como o procedimento anteriormente mostrado sugere.

(C) A busca por ortólogos na base de nucleotídeos foi feita com o gene "NC_000913.3:c366305-363231" e a de proteínas foi feita com a proteína "NP_414878.1" ambos de *Escherichia coli* str. *K-12* substr. MG1655.

(D) As tabelas geradas foram usadas numa comparação entre elas.

(E) Os resultados finais foram buscados por meio do dispositivo de busca direta do gene e proteínas, e por fim, alinhados com as proteínas e genes de interesse utilizados na busca por meio da ferramenta Multalin (http://multalin.toulouse.inra.fr/multalin/).

(F) Os perfis de abundância foram usados para geração de um *HeatMap* que mostra a abundância de cada um desses genes por metagenoma.

6.3 Resultados e discussão

A busca por palavras-chave rendeu diferentes números de genes/proteínas por banco de dados avaliado (Tabela 6.1). Observa-se também que com uma busca rápida do termo *beta-galactosidase* na internet também pode-se inferir diferentes termos que remetam à mesma enzima, como por exemplo seus códigos de KO (KO1190), COG (COG3250), GO – *gene onthology* (GO:0004565, GO:0009341), eggNOG (ENOG4105CNT) e PFAM (PF02929, PF00703, PF02836, PF02837). Neste exemplo utilizamos apenas o nome da enzima, e, por conseguinte algumas análises que não contém o termo ou expressão com o nome da enzima em seu output, como no caso da anotação pelo banco de dados PFAM, os resultados são muito reduzidos. Alterando-se a metodologia de busca para códigos específicos por banco de dados, espera-se haver um ganho substancial nas descobertas.

Na busca direta utilizando-se a sequência referência do gene *lacZ* obtivemos 96 ortólogos e com a sequência da proteína, obtivemos 163 ortólogos. Esses ortólogos foram cruzados e observou-se que 47 ortólogos eram comuns a ambas as listas. Esses 47 ortólogos foram por sua vez verificados nas anotações obtidas com a busca por termos (Tabela 6.1), e observou-se que o único banco de dados referência que gerou anotações com sobreposição aos resultados de busca direta de ortólogos foi o KEGG, com 25 ocorrências.

Tabela 6.1. Diferente número de proteínas encontrado na busca unificada em diversos bancos de dados com o termo "*beta-galactosidase*". Observa-se uma variação decorrente da diferença no modelo de estrutura da anotação. Os bancos de dados Uniprot, PFAM e dbCAN foram eliminados da tabela, pois não retornaram nenhum resultado durante a busca por palavras-chave. (Fonte: Próprio autor)

Bancos de dados	Ocorrências na busca por termos	Comum com a busca de ortólogos
KEGG	138	25
COG	241	0
eggNOG	46	0

Optou-se por captar os 25 ortólogos confirmados e recuperar suas sequências. O AGSSY recuperou as 25 sequências de proteínas e genes num período de 60 s. Com as sequências de aminoácidos realizamos um alinhamento com sua referência (Apêndice 12). Observa-se que as proteínas obtidas têm comprimento variável, o que faz sentido devido a forma como foram obtidas no catálogo. Entretanto, apesar desse comprimento relativamente irregular, as regiões alinhadas mostram uma conservação dos sítios de ligação ao substrato e dos sítios ativos, com a conservação global do resíduo E569 e P564. Isso sugere, potencial de que estejam funcionais, mesmo com o gene relativamente alterado em termos de tamanho.

Com os códigos dos ortólogos obtidos nessa abordagem, extraiu-se o perfil de abundância de acordo com os 106 metagenomas. O perfil de abundância serviu como *input* do programa HeatMapper (http://www1.heatmapper.ca/expression/) – Figura 6.2.

No *heatmap* é possível observar os genes mais abundantes de modo geral, aqueles contidos em um pequeno grupo de metagenomas e outros que estão em pequena abundância em todos eles. Assim, além de possibilitar conhecer quais os genes com potencial biotecnológico, o *software* AGSSY também auxilia na geração dos dados de distribuição geográfica dos mesmos.

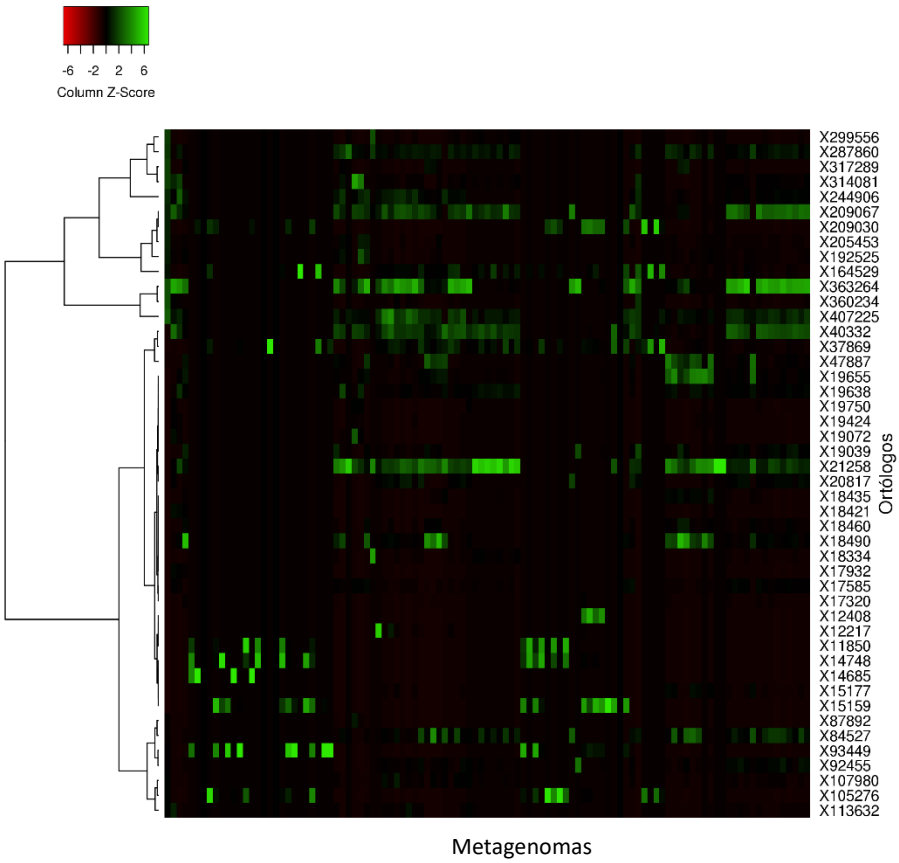

Figura 6.2. *Heatmap* da abundância dos ortólogos encontrados pelo AGSSY. A abundância em TPM foi obtida com o AGSSY e usada para calcular o Z-*score* por metagenoma utilizado no desenho deste mapa. No eixo "x" foi representado os metagenomas, sem identificá-los devido a noções de escala. Em y estão representados

os ortólogos com o prefixo típico “AM_AGSSY_” substituído por X. (Fonte: Próprio autor)

6.4 Conclusão

Num modelo de busca, o gene *lacZ* foi utilizado como referência e encontramos cerca de 25 ortólogos com *motifs* que garantem sua completude funcional e conservação estrutural. Os genes encontrados puderam ser analisados quanto a sua abundância e distribuição geográfica dentro dos 106 metagenomas. As anotações mostraram-se consistentes com a busca de ortólogos via alinhamento estatístico, entretanto, observou-se que a natureza da sequência referência influencia na consistência do conteúdo de ortólogos também encontrados nas anotações com os diversos bancos de dados. Assim, o AGSSY mostrou-se um sistema eficiente e rápido para a bioprospecção de genes e proteínas. Ele possibilitou um conjunto de operações importantes para a busca eficiente de funções e enzimas dentro do AMnrGC. Essas buscas quando orientadas por um objetivo consolidado podem levar a descoberta de enzimas e proteínas com novas funções e propriedades que tem elevado valor agregado na indústria de biotecnologia. Outro ponto forte mostrado pelo AGSSY foi sua portabilidade e facilidade de uso, em que o usuário tem uma interação limitada por uma interface gráfica que permite decisões simples e direcionadas que tornam o uso mais democrático. Dessa forma, recomendamos o uso do AGSSY para trabalhos futuros utilizando os genes presentes no AMnrGC.

CONCLUSÕES E PERSPECTIVAS FUTURAS

Ao início do presente trabalho nos propusemos a responder a seis questões de pesquisa importantes para o tema abordado, agora vamos resumir os achados dispostos para cada uma delas:

Q1 A diversidade gênica presente nos microrganismos do rio Amazonas reflete processos evolutivos locais ou é similar a outros sistemas fluviais?

A partir de uma análise de diversidade de k-mers, fizemos a comparação do rio Amazonas, com os rios Mississipi e sistemas da bacia hidrográfica do Canadá (ambos de clima temperado), e também com o solo da floresta amazônica. O que verificamos é que os microrganismos presentes no rio Amazonas se diferenciam em termos de diversidade de k-mers dos demais ambientes testados, além de sua aparente similaridade com alguns pontos dos sistemas temperados do Canadá, que ao fim parecem se dever a um efeito antrópico similarmente distribuído dentro de ambos os sistemas. Outro ponto relevante a se observar é que também foi verificada uma baixa influência de microrganismos de solo nos dados coletados para o rio Amazonas, reforçando que os achados apresentados nesta tese são basicamente derivados de microrganismos aquáticos do sistema fluvial da bacia do Amazonas.

Assim, concluímos que a microbiota do rio Amazonas, altamente heterotrófica, possui sim diversidade genética derivada de processos locais de evolução que se diferenciam de outros sistemas de água doce e solo, e que

talvez possam ser representativos de outros sistemas tropicais até então não estudados.

Q2. Há uma estrutura espacial na ocorrência de genes ao longo do curso do rio Amazonas que poderia indicar especialização metabólica?

No capítulo 3 podemos verificar facilmente que há uma estrutura espacial, que inclusive identifica um zoneamento de metabolismos, em que nas águas doces há uma preponderância da degradação de lignina e hemicelulose, enquanto que nas águas salobras e pluma, essa dominância se inverte e a degradação de celulose parece ser responsável pela maior parte do metabolismo microbiano. Também foi observado que dentre os produtos gerados pela oxidação da lignina, há uma preferência de transporte e consumo de monoaris pelas comunidades à jusante de Manaus, enquanto microrganismos à montante tendem a consumir diaris.

A salinidade se mostrou uma barreira importante para esses micróbios, fato que já havia sido previamente observado por outros autores (SATINSKY et al., 2015). A região à montante também parece ser mais diversa que as porções do rio Amazonas à jusante da cidade de Manaus. Essa diversificação à montante provavelmente é devido a alterações da água após a mistura das águas brancas e pretas que dilui os nutrientes e altera o pH, fazendo com que a população microbiana seja selecionada ao longo do curso do rio.

Assim, concluímos que apesar dos perfis metabólicos serem mantidos, as funções não-centrais desse metabolismo se alteram para o metabolismo secundário e degradação de xenobióticos, que podem inclusive sofrer um zoneamento de funções. Com relação à degradação de OM terrestre, parece que os microrganismos de água doce e salgada alternam os seus substratos preferenciais, sendo a degradação de celulose associada a ambientes marinhos e a degradação de lignina e hemicelulose a ambientes de água doce.

Q3. Quais as principais famílias de proteínas associadas à degradação de OM terrestre?

Nos capítulos 3 e 4 fazemos uma revisão dessa questão. No capítulo 3, verificamos um perfil geral das enzimas utilizadas no processamento de celulose, hemicelulose e lignina. Essas enzimas foram classificadas em termos de abundância e localização geográfica. Verificamos que há uma preferência para que a degradação de lignina seja mediada principalmente por DYPs e lacases, a degradação de celulose via GH3 e GH1, e a de hemicelulose via GH10.

No capítulo 4 vimos que esses sistemas de degradação estavam acoplados, de modo que microrganismos que oxidavam lignina tinham a capacidade de degradar celulose, através das mesmas enzimas já citadas.

Interessantemente, a degradação dos compostos derivados de lignina se mostrou desacoplada do sistema de oxidação extracelular da mesma. Ele foi reconhecido em uma comunidade secundária que aparentemente utiliza o sistema TTT de modo integrado e tem a capacidade de armazenar carbono via produção e degradação de PHB/A.

Essas populações se mostraram diversas em termos taxonômicos, sendo que a população de degradadores parece funcionar na forma de um consórcio de espécies de diversos grupos, inclusive contendo diversos genomas não identificados de bactérias. Enquanto isso, a comunidade de degradadores secundários é principalmente composta por beta-proteobactérias, em especial do grupo Burkholderiales.

Q4. Há mecanismos alternativos promovendo o uso de fontes de carbono secundárias ou o armazenamento de carbono na microbiota da bacia Amazônica?

Sim. No capítulo 4 está relatada a descoberta, em primeira mão, do sistema TTT sendo associado a sistemas de degradação de compostos aromáticos derivados de lignina e também à via do PHB, mostrando que sistemas utilizados pelos UFAC são importantes para o funcionamento do ecossistema no micro-habitat aquático do rio Amazonas. Esses sistemas mostraram-se completos e ainda se pode inclusive verificar nos dados expostos

naquele capítulo, os perfis de expressão gênica, mostrando-se que há uma clara preferência destes pela água doce.

Q5. Há evidências de um *priming effect* e quais os mecanismos para tal?

No presente trabalho propusemos dois modelos de efeito *priming* para explicar a degradação de OM terrestre no rio Amazonas (Capítulos 3 e 4). Esses modelos foram tomados utilizando-se dados distintos, uma vez que o primeiro deles foi feito com base em dados gênicos e o segundo com dados genômicos. O contexto genômico auxiliou numa evidente melhoria do modelo de modo que propusemos que o segundo modelo seja o mais adequado para futuras adequações da teoria ao ambiente amazônico. De acordo com as nossas observações, provavelmente o efeito *priming* no rio Amazonas se deve à eliminação da inibição que a comunidade degradadora de OM terrestre sofre pelos compostos fenólicos derivados da lignina.

Brevemente, há um consórcio que degrada diretamente a OM terrestre. Essas comunidades recebem carbono estrutural e energia da degradação da celulose, que só pode ocorrer uma vez que a lignina que a envolve for oxidada. A lignina degradada gera compostos fenólicos que inibem o consórcio microbiano degradador de OM terrestre. As comunidades secundárias (UFAC) consomem esses compostos fenólicos derivados de lignina, eliminando a inibição. Essas comunidades também recebem carbono

estrutural e energia a partir da degradação de celulose, sendo promovidas pelo consórcio degradador.

Q6. As inferências obtidas por meio do uso do catálogo de genes são mantidas quando avaliado o contexto genômico?

Não. Elas tendem a se alterar um pouco, devido a alguns ajustes finos. Esta questão conecta-se com as questões 4 e 5. A sua resposta é que os dados do contexto genômico somam-se às observações gênicas e reforçam os achados anteriores, adicionando-se certa complexidade necessária ao modelo que auxilia numa maior representação da realidade. O contexto da genômica auxiliou na reavaliação dos dados e reestruturação de algumas respostas, como por exemplo, que os sistemas de oxidação de lignina e degradação de celulose funcionariam acoplados num mesmo organismo; e que, além disso, os mecanismos do consumo dos compostos aromáticos gerados na oxidação da lignina são desacoplados do seu processo de produção.

Q7. Pode-se produzir um software que realize as análises básicas de metagenomas sem a necessidade de conhecimentos amplos na área de bioinformática?

Sim. De fato para atender às necessidades dos usuários do AMnrGC foi desenvolvido no capítulo 6, o AGSSY um sistema intuitivo para bioprospecção que mostrou-se eficiente e rápido na gestão dos genes do catálogo. Enquanto isso, nas análises de metagenomas ainda no nível de *reads* foi desenvolvido no capítulo 5, o programa BEAF que realiza análises desde taxonomia com um banco de dados referência até mesmo a categorização de genomas e busca de genes e famílias de proteínas neste tipo de arquivo. Ele funciona sem a necessidade de interação com o usuário de forma automática e supre a demanda de *softwares* de simples execução à baixos custos de processamento, uma demanda de pesquisados de países subdesenvolvidos. O BEAF mostrou-se competitivo com outros programas similares, apresentando um rendimento e resultados bastante satisfatórios.

Uma vez que as questões aqui propostas foram respondidas, ainda gostaríamos de propor novos rumos para este trabalho no futuro. Um dos mais proeminentes seria o sequenciamento dos genomas aqui obtidos como PGs, mas desta vez por meio de novas coletas utilizando-se a tecnologia de SAGs (*Single Amplified Genome*), que permitirá uma maior resolução e estudos de mutações pontuais dentro da população microbiana do rio Amazonas.

Outra questão interessante é a verificação do efeito *priming*, proposto nos capítulos 3 e 4, por meio de outros métodos, como também associar os dados de metatranscriptômica aos achados aqui apresentados. Além disso, também realizar-se experimentos ambientais, com a finalidade de se verificar a veracidade desta teoria.

Um ponto que permanece em aberto é o potencial biotecnológico do catálogo de genes proposto no presente trabalho, que por meio do programa AGSSY, pode ser mais facilmente explorado, chegando-se a novos genes e enzimas cujo potencial biotecnológico só tende a agregar à produção industrial de bens e serviços, além de representar potenciais novas terapias e tratamentos.

REFERÊNCIAS

AFFONSO, AG; BARBOSA, C.; NOVO, EMLM. Water quality changes in floodplain lakes due to the Amazon River flood pulse: Lago Grande de Curuaí (Pará). **Brazilian Journal of Biology**, v. 71, n. 3, p. 601–610, 2011.

AHMAD, Mark et al. Identification of DypB from Rhodococcus jostii RHA1 as a Lignin Peroxidase. **Biochemistry**, v. 50, n. 23, p. 5096–5107, 2011.

ANDREWS, Simon. **Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data**. Disponível em: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Acesso em: 8 nov. 2017.

ANTOINE, Rudy et al. Overrepresentation of a gene family encoding extracytoplasmic solute receptors in Bordetella. **Journal of bacteriology**, v. 185, n. 4, p. 1470–4, 2003.

ANTOINE, Rudy et al. The Periplasmic Binding Protein of a Tripartite Tricarboxylate Transporter is Involved in Signal Transduction. **Journal of Molecular Biology**, v. 351, n. 4, p. 799–809, 2005.

ARNOSTI, Carol. Microbial Extracellular Enzymes and the Marine Carbon Cycle. **Annual Review of Marine Science**, v. 3, n. 1, p. 401–425, 2011.

ASTON, John E. et al. Degradation of phenolic compounds by the lignocellulose deconstructing thermoacidophilic bacterium Alicyclobacillus Acidocaldarius. **Journal of Industrial Microbiology & Biotechnology**, v. 43, n. 1, p. 13–23, 2016.

AYRES, J.M. **As matas de várzea do Mamirauá: Médio Rio Solimões**. 2. ed. Brasília, DF: CNPq; Tefé, AM: Sociedade Civil Mamirauá, 1995.

BAHRAM, Mohammad et al. Structure and function of the global topsoil microbiome. **Nature**, v. 560, n. 7717, p. 233–237, 2018.

BAKER, Brett J et al. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. **Microbiome**, v. 3, n. 1, p. 14, 2015.

BANKEVICH, Anton et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of computational biology : a journal of computational molecular cell biology**, v. 19, n. 5, p. 455–477, 2012.

BATTIN, Tom J. et al. Biophysical controls on organic carbon fluxes in fluvial networks. **Nature Geoscience**, v. 1, n. 2, p. 95–100, 2008.

BENNER, Ronald et al. Bacterial carbon metabolism in the Amazon River system. **Limnology and Oceanography**, v. 40, n. 7, p. 1262–1270, 1995.

BENNER, Ronald; MORAN, Mary Ann; HODSON, Robert E. Biogeochemical cycling of lignocellulosic carbon in marine and freshwater ecosystems: Relative contributions of procaryotes and eucaryotes1. **Limnology and Oceanography**, v. 31, n. 1, p. 89–100, 1986.

BENOIT, Gaëtan et al. Multiple comparative metagenomics using multiset k-mer counting. **PeerJ Computer Science**, v. 2, p. e94, 2016.

BERGAUER, Kristin et al. Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 3, p. E400–E408, 2018.

BERGMANN, Jessica C. et al. Discovery of two novel β-glucosidases from an Amazon soil metagenomic library. **FEMS Microbiology Letters**, v. 351, n. 2, p. 147–155, 2014.

BIANCHI, Thomas S. The Role of Terrestrially Derived Organic Carbon in the Coastal Ocean: A Changing Paradigm and the Priming Effect. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 49, p. 19473–19481, 2011.

BINGEMANN, C. W.; VARNER, J. E.; MARTIN, W. P. The effect of the addition of organic materials on the decomposition of an organic soil. **Soil Sciences Society American Proceedings**, v. 17, p. 34–38, 1953.

BOERJAN, Wout; RALPH, John; BAUCHER, Marie. Lignin Biosynthesis. **Annual Review of Plant Biology**, v. 54, n. 1, p. 519–546, 2003.

BOSE, Samar K. et al. Lignin content versus syringyl to guaiacyl ratio amongst poplars. **Bioresource Technology**, v. 100, n. 4, p. 1628–1633, 2009.

BOWERS, Robert M et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. **Nature Biotechnology**, v. 35, n. 8, p. 725–731, 2017.

BRIGHAM, Christopher J. et al. Bacterial Carbon Storage to Value Added Products. **Journal of Microbial & Biochemical Technology**, v. s3, 2011.

BROWN, Margaret E.; BARROS, Tiago; CHANG, Michelle C. Y. Identification and Characterization of a Multifunctional Dye Peroxidase from a Lignin-Reactive Bacterium. **ACS Chemical Biology**, v. 7, n. 12, p. 2074–2081, 2012.

BRUMM, Phillip J. Bacterial genomes: what they teach us about cellulose degradation. **Biofuels**, v. 4, n. 6, p. 669–681, 2013.

BUCHFINK, Benjamin; XIE, Chao; HUSON, Daniel H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, v. 12, n. 1, p. 59–60, 2014.

BURDIGE, D.J. The burial of terrestrial organi carbon in marine sediments: A reassessment. **Global Biogeochemical Cycles**, 2005.

CABELLO-YEVES, Pedro J. et al. Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. **Applied and Environmental Microbiology**, v. 84, n. 1, p. e02132-17, 2018.

CABELLO-YEVES, Pedro J.; HARO-MORENO, Jose M.; et al. Novel Synechococcus Genomes Reconstructed from Freshwater Reservoirs. **Frontiers in Microbiology**, v. 8, p. 1151, 2017.

CABELLO-YEVES, Pedro J.; GHAI, Rohit; et al. Reconstruction of Diverse Verrucomicrobial Genomes from Metagenome Datasets of Freshwater Reservoirs. **Frontiers in Microbiology**, v. 8, p. 2131, 2017.

CARRADEC, Quentin et al. A global ocean atlas of eukaryotic genes. **Nature Communications**, v. 9, n. 1, p. 373, 2018.

CASTELLO, Leandro et al. The vulnerability of Amazon freshwater ecosystems. **Conservation Letters**, v. 6, n. 4, p. 217–229, 2012.

CHERRIER, J. et al. Radiocarbon in marine bacteria: evidence for the ages of assimilated carbon. **Limnology and Oceanography**, v. 44, p. 730–736, 1999.

COLE, J. J. et al. Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget. **Ecosystems**, v. 10, n. 1, p. 172–185, 2007.

COPPINI, Guilherme Bonotti. **Desenvolvimento de um módulo de análises taxonômicas do gene 16S rRNA de metagenomas para o programa BEAF (Referenced Binning Engine for Autonomous Finding): BEAF16S**. 2018. 60 f. Trabalho de Conclusão de Curso – Universidade Federal de São Carlos, São Carlos, SP, 2018.

CORNO, G et al. Interspecific interactions drive chitin and cellulose degradation by aquatic microorganisms. **Aquatic Microbial Ecology**, v. 76, n. 1, p. 27–37, 2015.

CRAGG, Simon M et al. Lignocellulose degradation mechanisms across the Tree of Life. **Current Opinion in Chemical Biology**, v. 29, p. 108–119, 2015.

DAS, Nilanjana; CHANDRAN, Preethy. Microbial Degradation of Petroleum Hydrocarbon Contaminants: An Overview. **Biotechnology Research International**, v. 2011, p. 1–13, 2011.

DE HAAN, H. Effect of benzoate on microbial decomposition of fulvic acids in Tjeukemeer (Netherlands). **Limnology and Oceanography**, v. 22, p. 38–44, 1977.

DEL VECCHIO, Rossana; SUBRAMANIAM, Ajit. Influence of the Amazon River on the surface optical properties of the western tropical North Atlantic Ocean. **Journal of Geophysical Research**, v. 109, n. C11, p. C11001, 2004.

DELWICHE, C.F.; GRAHAM, L. E.; THOMSON, N. Lignin-like compounds and sporopollenin coleochaete, an algal model for lands plant ancestry. **Science**, v. 245, p. 399–401, 1989.

DESANTIS, T Z et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Applied and environmental microbiology**, v. 72, n. 7, p. 5069–5072, 2006.

DICK, Gregory J et al. Community-wide analysis of microbial genome sequence signatures. **Genome Biology**, v. 10, n. 8, p. R85, 2009.

DIXON, Philip. VEGAN, a package of R functions for community ecology. **Journal of Vegetation Science**, v. 14, n. 6, p. 927–930, 2003.

DOHERTY, Mary et al. Bacterial Biogeography across the Amazon River-Ocean Continuum. **Frontiers in Microbiology**, v. 8, p. 882, 2017.

EDDY, Sean R. Accelerated Profile HMM Searches. **PLoS Computational Biology**, v. 7, n. 10, p. e1002195, 2011.

EDGAR, Robert C. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, v. 26, n. 19, p. 2460–2461, 2010.

EDGAR, Robert C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. **Nature Methods**, v. 10, p. 996, Agosto 2013.

EREN, A. Murat. TARA-NON-REDUNDANT-MAGs. **figshare**, 2017. Disponível em: <https://figshare.com/articles/TARA-NON-REDUNDANT-MAGs/4902923>.

ERTEL, John R. et al. Dissolved humic substances of the Amazon River system1. **Limnology and Oceanography**, v. 31, n. 4, p. 739–754, 1986.

FARJALLA, Vinicius F. Are the mixing zones between aquatic ecosystems hot spots of bacterial production in the Amazon River system? **Hydrobiologia**, v. 728, n. 1, p. 153–165, 2014.

FIELD et al. Primary production of the biosphere: integrating terrestrial and oceanic components. **Science (New York, N.Y.)**, v. 281, n. 5374, p. 237–40, 1998.

FINN, Robert D et al. The Pfam protein families database: towards a more sustainable future. **Nucleic acids research**, v. 44, n. D1, p. D279–85, 2016.

FINN, Robert D; CLEMENTS, Jody; EDDY, Sean R. HMMER web server: interactive sequence similarity searching. **Nucleic acids research**, v. 39, n. Web Server issue, p. W29–W37, 2011.

FU, Limin et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics**, v. 28, n. 23, p. 3150–3152, 2012.

GAGNE-MAYNARD, William C. et al. Evaluation of Primary Production in the Lower Amazon River Based on a Dissolved Oxygen Stable Isotopic Mass Balance. **Frontiers in Marine Science**, v. 4, p. 26, 2017.

GAI, Zhonghui et al. Genome sequence of *Sphingobium* yanoikuyae XLDN2-5, an efficient carbazole-degrading strain. **Journal of bacteriology**, v. 193, n. 22, p. 6404–5, 2011.

GALL, Daniel L. et al. Stereochemical Features of Glutathione-dependent Enzymes in the *Sphingobium* sp. Strain SYK-6 β-Aryl Etherase Pathway. **Journal of Biological Chemistry**, v. 289, n. 12, p. 8656–8667, 2014.

GARZA, D. R.; DUTILH, B. E. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. **Cellular and Molecular Life Sciences**, v. 72, p. 4287–4308, 2015.

GEORGES, Anna A et al. Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. **The ISME Journal**, v. 8, n. 6, p. 1301–1313, 2014.

GERISCHER, Ulrike. **Acinetobacter molecular microbiology**. [S.l.]: Caister Academic Press, 2008.

GHAI, Rohit; RODRIGUEZ-VALERA, Francisco; et al. Metagenomics of the water column in the pristine upper course of the Amazon river. **PLoS ONE**, v. 6, n. 8, p. e23785, 2011.

GHAI, Rohit; PAŠIĆ, Lejla; et al. New Abundant Microbial Groups in Aquatic Hypersaline Environments. **Scientific Reports**, v. 1, n. 1, p. 135, 2011.

GONZALO, Gonzalo de et al. Bacterial enzymes involved in lignin degradation. **Journal of Biotechnology**, v. 236, p. 110–119, 2016.

GROSSART, Hans-Peter et al. Comparison of cell-specific activity between free-living and attached bacteria using isolates and natural assemblages. **FEMS Microbiology Letters**, v. 266, n. 2, p. 194–200, 2007.

GUENET, B. et al. Primming Effect: Briging the Gap between Terrestrial and Aquatic Ecology. **Ecology**, v. 91, n. 10, p. 2850–2861, 2010.

GUREVICH, Alexey et al. QUAST: Quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 2013.

HERNES, P.J.; BENNER, R. Photochemical and microbial degradation of dissolved lignin phenols: implications for the fate of terrigenous dissolved organic matter in marine environments. **Journal of Geophysical Research: Oceans**, v. 108, p. 3291, 2003.

HILTON, Jason A et al. Metatranscriptomics of N2-fixing cyanobacteria in the Amazon River plume. **The ISME journal**, v. 9, n. 7, p. 1557–69, 2015.

HORVATH, RS. Microbial co-metabolism and the degradation of organic compounds in nature. **Bacteriological reviews**, v. 36, p. 146–155, 1972.

HOSAKA, Masaru et al. Novel tripartite aromatic acid transporter essential for terephthalate uptake in Comamonas sp. strain E6. **Applied and environmental microbiology**, v. 79, n. 19, p. 6148–55, 2013.

HU, Jinguang; SADDLER, Jack N. Why does GH10 xylanase have better performance than GH11 xylanase for the deconstruction of pretreated biomass? **Biomass and Bionenergy**, v. 110, p. 13-16, 2018.

HUBLEY, Robert et al. The Dfam database of repetitive DNA families. **Nucleic Acids Research**, v. 44, n. D1, p. D81–D89, 2016.

HUERTA-CEPAS, Jaime et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. **Nucleic Acids Research**, v. 44, n. D1, p. D286–D293, 2016.

HUG, Laura A. et al. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. **Environmental Microbiology**, v. 18, n. 1, p. 159–173, 2016.

HUGERTH, Luisa W. et al. Metagenome-assembled genomes uncover a global brackish microbiome. **Genome Biology**, v. 16, n. 1, p. 279, 2015.

HUTALLE-SCHMELZER, Kristine Michelle L. et al. Enrichment and cultivation of pelagic bacteria from a humic lake using phenol and humic matter additions. **FEMS Microbiology Ecology**, v. 72, n. 1, p. 58–73, 2010.

HYATT, Doug et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics**, v. 11, n. 1, p. 119, 2010.

JAIN, Chirag et al. High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. **bioRxiv**, p. 225342, 2017.

JØRGENSEN, B B. Ecology of the bacteria of the sulphur cycle with special reference to anoxic-oxic interface environments. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 298, n. 1093, p. 543–61, 1982.

JUNG, Jaejoon; BAEK, Jeong-Hun; PARK, Woojun. Complete genome sequence of the diesel-degrading Acinetobacter sp. strain DR1. **Journal of bacteriology**, v. 192, n. 18, p. 4794–5, 2010.

JUNK, W. J. Wetlands of tropical South America. **Wetlands of the World**. Kluwer Publishers, The Netherlands: Whigham, D.F.; Hejnym, S.; Dykyjova, D., 1993. p. 679–739.

JUNK, W. J.; BAYLEY, P.B.; SPARKS, R.E. The flood pulse concept in river-floodplain systems. **Canadian Journal of Fishers and Aquatic**, v. 106, p. 110–127, 1989.

KAMIMURA, Naofumi et al. Bacterial catabolism of lignin-derived aromatics: New findings in a recent decade: Update on bacterial lignin catabolism. **Environmental Microbiology Reports**, v. 9, n. 6, p. 679–705, 2017.

KANEHISA, Minoru et al. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic acids research**, v. 40, n. Database issue, p. D109–14, 2012.

KANEHISA, Minoru et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. **Nucleic Acids Research**, v. 45, n. D1, p. D353–D361, 2017.

KANG, Dongwan D. et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. **PeerJ**, v. 3, 2015.

KARLEN, Steven D. et al. Monolignol ferulate conjugates are naturally incorporated into plant lignins. **Science Advances**, v. 2, n. 10, 2016. Disponível em: <http://advances.sciencemag.org/content/2/10/e1600393>.

KENDALL, C.; STEVEN, R. S.; KELLY, V. J. Carbon and nitrogen isotopic compositions of particulate organic matter in four large river systems across the Unitated States. **Hydrological Processes**, v. 15, p. 1301–1346, 2001.

KLOTZBÜCHER, Thimo et al. A new conceptual model for the fate of lignin in decomposing plant litter. **Ecology**, v. 92, n. 5, p. 1052–1062, 2011.

KÖGEL-KNABNER, Ingrid. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. **Soil Biology and Biochemistry**, v. 34, n. 2, p. 139–162, 2002.

KONG, Hyesuk et al. Species-specific distribution of a modular family 19 chitinase gene in Burkholderia gladioli. **FEMS Microbiology Ecology**, v. 37, n. 2, p. 135–141, 2001.

KOROTKOVA, Natalia; CHISTOSERDOVA, Ludmila; LIDSTROM, Mary E. Poly-beta-hydroxybutyrate biosynthesis in the facultative methylotroph methylobacterium extorquens AM1: identification and mutation of gap11, gap20, and phaR. **Journal of bacteriology**, v. 184, n. 22, p. 6174–81, 2002.

KUZYAKOV, Y.; FRIEDEL, J. K.; STAHR, K. Review of mechanisms and quantification of priming effects. **Soil Biology and Biochemistry**, v. 32, p. 1485–1498, 2000.

LARAQUE, Alain; GUYOT, Jean Loup; FILIZOLA, Naziano. Mixing processes in the Amazon River at the confluences of the Negro and Solimões Rivers, Encontro das Águas, Manaus, Brazil. **Hydrological Processes**, v. 23, n. 22, p. 3131–3140, 2009.

LEE, Kyung-Eun; PARK, Hyun-Seok. A review of three different studies on hidden markov models for epigenetic problems: a computational perspective. **Genomics & informatics**, v. 12, n. 4, p. 145–150, 2014.

LENZ, Robert W.; MARCHESSAULT, Robert H. Bacterial Polyesters: Biosynthesis, Biodegradable Plastics and Biotechnology. **Biomacromolecules**, v. 6, n. 1, p. 1–8, 2005.

LI, Dinghua et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. v. 102, p. 3–11, 2016.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 2009.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, n. 13, p. 1658–1659, 2006.

LIANG, Yan-Ling et al. Isolation, Screening, and Identification of Cellulolytic Bacteria from Natural Reserves in the Subtropical Region of China and Optimization of Cellulase Production by Paenibacillus terrae ME27-1. **BioMed Research International**, v. 2014, p. 1–13, 2014.

LLADÓ, Salvador; LÓPEZ-MONDÉJAR, Rubén; BALDRIAN, Petr. Forest Soil Bacteria: Diversity, Involvement in Ecosystem Processes, and Response to Global Change. **Microbiology and molecular biology reviews : MMBR**, v. 81, n. 2, p. e00063-16, Abril 2017.

LOGARES, Ramiro et al. Infrequent marine–freshwater transitions in the microbial world. **Trends in Microbiology**, v. 17, n. 9, p. 414–422, 2009.

LOGARES, Ramiro et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. **Environmental Microbiology**, v. 16, n. 9, p. 2659–2671, 2014.

LÓPEZ-MONDÉJAR, Rubén et al. Cellulose and Hemicellulose Decomposition by Forest Soil Bacteria Proceeds by the Action of Structurally Variable Enzymatic Systems. **Scientific Reports**, v. 6, 2016.

MACHADO, Vinicius S.; BICALHO, Rodrigo C. Complete Genome Sequence of Trueperella pyogenes, an Important Opportunistic Pathogen of Livestock. **Microbiology Resource Announcements**, v. 2, n. 2, 2014. Disponível em: <https://mra.asm.org/content/2/2/e00400-14>.

MAJUMDAR, Sudipta et al. Roles of Small Laccases from Streptomyces in Lignin Degradation. **Biochemistry**, v. 53, n. 24, p. 4047–4058, 2014.

MAKARIEVA, A.M. et al. Where do winds come from? A new theory on how water vapor condensation influences atmospheric pressure and dynamics. **Atmospheric Chemistry and Physics**, v. 13, p. 1039–1056, 2013.

MAKARIEVA, A.M.; GORSHKOV, V.G. Why Does Air Passage over Forest Yield More Rain? Examining the Coupling between Rainfall, Pressure, and Atmospheric Moisture Content. **Journal of Hydrometeorology**, v. 15, n. 1, p. 411–426, 2014.

MALHI, Yadvinder et al. Climate change, deforestation, and the fate of the Amazon. **Science**, v. 319, n. 5860, p. 169–72, 2008.

MARDIS, Elaine R. A decade's perspective on DNA sequencing technology. **Nature**, v. 470, p. 198, 2011.

MARTENS, Dean A.; REEDY, Thomas E.; LEWIS, David T. Soil organic carbon content and composition of 130-year crop, pasture and forest land-use managements. **Global Change Biology**, v. 10, n. 1, p. 65–78, 2004.

MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, v. 17, n. 1, p. 10, 2011.

MARTONE, Patrick T. et al. Discovery of Lignin in Seaweed Reveals Convergent Evolution of Cell-Wall Architecture. **Current Biology**, v. 19, n. 2, p. 169–175, 2009.

MASAI, Eiji; KATAYAMA, Yoshihiro; FUKUDA, Masao. Genetic and Biochemical Investigations on Bacterial Catabolic Pathways for Lignin-Derived Aromatic Compounds. **Bioscience, Biotechnology, and Biochemistry**, v. 71, n. 1, p. 1–15, 2007.

MATSUDA, Kana et al. Heterologous Expression, Purification, and Characterization of an α-Mannosidase Belonging to Glycoside Hydrolase Family 99 of Shewanella amazonensis. **Bioscience, Biotechnology, and Biochemistry**, v. 75, n. 4, p. 797–799, 2011.

MAYER, Lawrence M. et al. Organic matter in small mesopores in sediments and soils. v. 68, n. 19, p. 3863–3872, 2004.

MAYORGA, Emilio et al. Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. **Nature**, v. 436, p. 538, 2005.

MCDONALD, Daniel et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. **The ISME journal**, v. 6, n. 3, p. 610–618, 2012.

MCDONALD, I R; MURRELL, J C. The particulate methane monooxygenase gene pmoA and its use as a functional gene probe for methanotrophs. **FEMS microbiology letters**, v. 156, n. 2, p. 205–10, 1997.

MCILROY, Simon Jon et al. MiDAS: the field guide to the microbes of activated sludge. **Database : the journal of biological databases and curation**, v. 2015, p. bav062–bav062, 2015.

MELO, Michaela L. et al. Flood pulse regulation of bacterioplankton community composition in an Amazonian floodplain lake. **Freshwater Biology**, v. 00, p. 1–13, 2018.

MENDE, Daniel R. et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. **Nature Microbiology**, v. 2, n. 10, p. 1367–1373, 2017.

MEYER, Kyle M. et al. Conversion of Amazon Rainforest to Agriculture Alters Community Traits of Methane-cycling Organisms. **Molecular Ecology**, v. 26, n. 6, p. 1547–1556, 2017.

MIEDES, Eva et al. The role of the secondary cell wall in plant resistance to pathogens. **Frontiers in plant science**, v. 5, p. 358–358, 2014.

MIKHAILOV, V. N. Water and sediment runoff at the Amazon River mouth. **Water Resources**, v. 37, n. 2, p. 145–159, 2010.

MIKHEENKO, Alla; SAVELIEV, Vladislav; GUREVICH, Alexey. MetaQUAST: Evaluation of metagenome assemblies. **Bioinformatics**, v. 32, n. 7, p. 1088–1090, 2016.

MOLDREUP, P. et al. TORTUOSITY, DIFFUSIVITY, AND PERMEABILITY IN THE SOIL LIQUID AND GASEOUS PHASES. **Soil Sciences Society of America Journal**, v. 65, n. 3, p. 613–623, 2001.

MONLAU, F. et al. Do furanic and phenolic compounds of lignocellulosic and algae biomass hydrolyzate inhibit anaerobic mixed cultures ? A comprehensive review. **Biotechnology Advances**, v. 32, n. 5, p. 934–951, 2014.

MONTEIRO, Marcela C.; PEREIRA, Luci C. C.; JIMÉNEZ, José A. The Trophic Status of an Amazonian Estuary Under Anthropogenic Pressure (Brazil). **Journal of Coastal Research**, v. 75, n. sp1, p. 98–102, mar. 2016.

MOREIRA, Leandro M et al. Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of *Xanthomonas fuscans* subsp. *aurantifolii*. **BMC Genomics**, v. 11, n. 1, p. 238, 2010.

MOURA, R. L. et al. An extensive reef system at the Amazon River mouth. **Science Advances**, v. 2, n. 4, p. e1501252–e1501252, 2016.

MUIR, Paul et al. The real cost of sequencing: scaling computation to keep pace with data generation. **Genome Biology**, v. 17, n. 1, p. 53, 2016.

MULLER, Emilie E. L. et al. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. **Nature Communications**, v. 5, p. 5603, 2014.

NARAYANASAMY, Shaman et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. **Genome Biology**, v. 17, n. 1, p. 260, 2016.

NARAYANASAMY, Shaman; JAROSZ, Yohan; WILMES, Paul. **IMP simulated mock community data set**. Disponível em: <https://doi.org/10.5281/zenodo.160261>.

NELSON, Michael C.; MORRISON, Mark; YU, Zhongtang. A Meta-Analysis of the Microbial Diversity Observed in Anaerobic Digesters. **Bioresource Technology**, v. 102, n. 4, p. 3730–3739, 2011.

NEUWIRTH, Erich. **CRAN - Package ColorBrewer Palettes**. [S.l.]: Comprehensive R Archive Network (CRAN), 2014. Disponível em: <https://cran.r-project.org/web/packages/RColorBrewer/index.html>. Acesso em: 7 nov. 2017.

NEWTON, Ryan J et al. A guide to the natural history of freshwater lake bacteria. **Microbiology and molecular biology reviews : MMBR**, v. 75, n. 1, p. 14–49, 2011.

NOBRE, A.D. **O Futuro Climático da Amazônia, Relatório de Avaliação Científica**. São José dos Campos - SP, Brasil: Patrocinado por ARA, CCST-INPE, e INPA, 2014. Disponível em: <https://www.socioambiental.org/sites/blog.socioambiental.org/files/futuro-climatico-da-amazonia.pdf>.

NYCHKA, Douglas et al. **fields: Tools for spatial data**. Boulder, CO, USA: University Corporation for Atmospheric Research, 2017. Disponível em: <www.image.ucar.edu/nychka/Fields>.

PAN, Hudan et al. A Gene Catalogue of the Sprague-Dawley Rat Gut Metagenome. **GigaScience**, v. 7, n. 5, 2018.

PARKS, Donovan H et al. A proposal for a standardized bacterial taxonomy based on genome phylogeny. **bioRxiv**, p. 256800, 2018.

PARKS, Donovan H et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. **Genome research**, v. 25, n. 7, p. 1043–55, 2015.

PARKS, Donovan H. **CompareM**. Disponível em: <https://github.com/dparks1134/CompareM/blob/master/users_guide.pdf>. Acesso em: 24 out. 2017.

PARKS, Donovan H et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. **Nature Microbiology**, v. 2, n. 11, p. 1533–1542, 2017.

PAULY, Markus; KEEGSTRA, Kenneth. Cell-wall carbohydrates and their modification as a resource for biofuels. **The Plant Journal**, v. 54, n. 4, p. 559–568, 2008.

PAYNE, Christina M. et al. Fungal Cellulases. **Chemical Reviews**, v. 115, n. 3, p. 1308–1448, 2015.

PECK, Spencer C.; GAO, Jiangtao; VAN DER DONK, Wilfred A. Discovery and Biosynthesis of Phosphonate and Phosphinate Natural Products. **Methods in Enzymology**, v. 516, p. 101–123, 2012.

PESTER, Michael; SCHLEPER, Christa; WAGNER, Michael. The Thaumarchaeota: An emerging view of their phylogeny and ecophysiology. **Current Opinion in Microbiology**, v. 14, n. 3, p. 300–306, 2011.

PETTENGILL, James B.; RAND, Hugh. Segal's Law, 16S RRNA Gene Sequencing, and the Perils of Foodborne Pathogen Detection within the American Gut Project. **PeerJ**, v. 5, p. e3480, 2017.

PFENNIG, N; WIDDEL, F. The bacteria of the sulphur cycle. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 298, n. 1093, p. 433–41, 1982.

PIEDADE, M.T.F.; JUNK, W. J.; PAROLIN, P. The flood pulse and photosynthetic response of trees in a white water floodplain (várzea) of the Central Amazon, Brazil. **Verh. Internat. Verein. Limnol.**, v. 27, p. 1–6, 2000.

POLI, Annarita et al. Synthesis, production, and biotechnological applications of exopolysaccharides and polyhydroxyalkanoates by archaea. **Archaea (Vancouver, B.C.)**, v. 2011, p. 693253, 2011.

PORETSKY, Rachel S. et al. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. **Environmental Microbiology**, v. 12, n. 3, p. 616–627, 2010.

QIN, Junjie et al. A human gut microbial gene catalogue established by metagenomic sequencing. **Nature**, v. 464, n. 7285, p. 59–65, 2010.

QIN, Lei et al. Inhibition of lignin-derived phenolic compounds to cellulase. **Biotechnology for Biofuels**, v. 9, n. 1, p. 70, 2016.

QUAST, Christian et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic acids research**, v. 41, n. Database issue, p. D590-6, 2013.

QUELAS, J I et al. Regulation of Polyhydroxybutyrate Synthesis in the Soil Bacterium Bradyrhizobium diazoefficiens. **Applied and environmental microbiology**, v. 82, n. 14, p. 4299–308, 2016.

QUINLAN, Aaron R; HALL, Ira M. BEDTools: A flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, n. 6, p. 841–842, 2010.

RAHMANPOUR, Rahman; BUGG, Timothy D.H. Characterisation of Dyp-type peroxidases from Pseudomonas fluorescens Pf-5: Oxidation of Mn(II) and polymeric lignin by Dyp1B. **Archives of Biochemistry and Biophysics**, v. 574, p. 93–98, 2015.

RAYMOND, Petter A.; BAUER, J. E. Riverine export of aged terrestrial organic matter to the North Atlantic Ocean. **Nature**, v. 409, p. 497–500, 2001.

REMINGTON, Sonya; KRUSCHE, Alex; RICHEY, Jeff. Effects of DOM Photochemistry on Bacterial Metabolism and CO2 Evasion during Falling Water in a Humic and a Whitewater River in the Brazilian Amazon. **Biogeochemistry**, v. 105, n. 1, p. 185–200, 2011.

RICE, P; LONGDEN, I; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends in genetics : TIG**, v. 16, n. 6, p. 276–7, 2000.

RIECK, Angelika et al. Particle-Associated Differ from Free-Living Bacteria in Surface Waters of the Baltic Sea. **Frontiers in microbiology**, v. 6, p. 1297, 2015.

ROBERTS, Adam; PACHTER, Lior. Streaming fragment assignment for real-time analysis of sequencing experiments. **Nature Methods**, v. 10, n. 1, p. 71–73, 2012.

ROSA, Leonardo T. et al. Structural basis for high-affinity adipate binding to AdpC (RPA4515), an orphan periplasmic-binding protein from the tripartite tricarboxylate transporter (TTT) family in Rhodopseudomonas palustris. **The FEBS Journal**, v. 284, n. 24, p. 4262–4277, 2017.

SALCHER, Michaela M et al. The ecology of pelagic freshwater methylotrophs assessed by a high-resolution monitoring and isolation campaign. **The ISME journal**, v. 9, n. 11, p. 2442–53, 2015.

SANCHEZ, C. Lignocellulosic residues: Biodegradation and bioconversion by fungi. **Biotechnology Advances**, v. 27, p. 185–194, 2009.

SANTOS-JÚNIOR, Célio Dias et al. Metagenome Sequencing of Prokaryotic Microbiota Collected from Rivers in the Upper Amazon Basin. **Genome Announcements**, v. 5, n. 2, p. e01450–16, 2017.

SATINSKY, Brandon M; SMITH, Christa B; SHARMA, Shalabh; LANDA, Marine; et al. Expression patterns of elemental cycling genes in the Amazon River Plume. **The ISME Journal**, v. 11, n. 8, p. 1852–1864, 2017.

SATINSKY, Brandon M. et al. Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. **Microbiome**, v. 3, n. 1, p. 39, 2015.

SATINSKY, Brandon M; CRUMP, Byron C; et al. Microspatial gene expression patterns in the Amazon River Plume. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 30, p. 11085–90, 2014.

SATINSKY, Brandon M.; SMITH, Christa B.; SHARMA, Shalabh; WARD, Nicholas D.; et al. Patterns of Bacterial and Archaeal Gene Expression through the Lower Amazon River. **Frontiers in Marine Science**, v. 4, p. 253, 2017.

SATINSKY, Brandon M; ZIELINSKI, Brian L; et al. The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. **Microbiome**, v. 2, p. 17, 2014.

SAWAKUCHI, Henrique O. et al. Carbon Dioxide Emissions along the Lower Amazon River. **Frontiers in Marine Science**, v. 4, p. 76, 2017.

SCHLESINGER, W. H. Carbon balance in terrestrial detritus. **Anuual Reviews in Ecology of Systems**, v. 8, p. 51–81, 1977.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, v. 30, n. 14, p. 2068–2069, 2014.

SEIDEL, Michael et al. Seasonal and spatial variability of dissolved organic matter composition in the lower Amazon River. **Biogeochemistry**, v. 131, n. 3, p. 281–302, 2016.

SENALIK, Douglas. **bb.orffinder - current version 1.3.0**. Disponível em: <https://github.com/dsenalik/bb/blob/master/bb.orffinder>. Acesso em: 7 nov. 2018.

SEO, Jong-Su; KEUM, Young-Soo; LI, Qing. Bacterial Degradation of Aromatic Compounds. **International Journal of Environmental Research and Public Health**, v. 6, n. 12, p. 278–309, 2009.

SHARP, Christine E. et al. Distribution and diversity of Verrucomicrobia methanotrophs in geothermal and acidic environments. **Environmental Microbiology**, v. 16, n. 6, p. 1867–1878, 2014.

SHARPTON, Thomas J. An introduction to the analysis of shotgun metagenomic data. **Frontiers in Plant Science**, v. 5, p. 209, 2014.

SILVA, Bruno S de O et al. Virioplankton Assemblage Structure in the Lower River and Ocean Continuum of the Amazon. **mSphere**, v. 2, n. 5, 2017.

SINGER, Esther et al. Next generation sequencing data of a defined microbial mock community. **Scientific Data**, v. 3, p. 160081, 2016.

SIOLI, H. The Amazon and its main affluents: Hydrography, morphology of the river courses, and river types. **The Amazon: Limnology and landscape ecology of a mighty tropical river and its basin**. Dordrecht: Springer Netherlands: Sioli, H., 1984. p. 127–165.

STALEY, Christopher et al. Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. **Frontiers in Microbiology**, v. 5, p. 524, 2014a.

STALEY, Christopher et al. Core Functional Traits of Bacterial Communities in the Upper Mississippi River Show Limited Variation in Response to Land Cover. **Frontiers in Microbiology**, v. 5, 2014b. Disponível em: <https://www.frontiersin.org/articles/10.3389/fmicb.2014.00414/full>. Acesso em: 16 ago. 2018.

STREIT, W. R.; SCHMIDTZ, R. A. Metagenomics – the key to the uncultured microbes. **Current Opinion in Microbiology**, v. 7, n. 5, p. 492–498, 2004.

SUBRAMANIAM, A et al. Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. **Proceedings of the National Academy of Sciences of the United States of America**, v. 105, n. 30, p. 10460–5, 2008.

SUN, Shulei et al. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. **Nucleic acids research**, v. 39, n. Database issue, p. D546–51, 2011.

SUNAGAWA, S. et al. Structure and function of the global ocean microbiome. **Science**, v. 348, n. 6237, p. 1261359–1261359, 2015.

SWAN, Brandon K et al. Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. **PloS one**, v. 9, n. 4, p. e95380, 2014.

TADONLÉKÉ, Rémy D. Strong coupling between natural Planctomycetes and changes in the quality of dissolved organic matter in freshwater samples. **FEMS microbiology ecology**, v. 59, n. 3, p. 543–55, 2007.

TARASOV, Artem et al. Sambamba: fast processing of NGS alignment formats. **Bioinformatics**, v. 31, n. 12, p. 2032–2034, 2015.

TATUSOV, Roman L et al. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, v. 4, n. 1, p. 41, 2003.

TESSLER, Michael et al. A Global eDNA Comparison of Freshwater Bacterioplankton Assemblages Focusing on Large-River Floodplain Lakes of Brazil. **Microbial Ecology**, v. 73, n. 1, p. 61–74, 2017.

TOYAMA, Danyelle et al. A snapshot on prokaryotic diversity of the Solimões River basin (Amazon, Brazil). **Genetics and Molecular Research**, v. 16, n. 2, p. gmr16029567, 2017.

TOYAMA, Danyelle et al. A novel β-glucosidase isolated from the microbial metagenome of Lake Poraquê (Amazon, Brazil). **Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics**, v. 1866, n. 4, p. 569–579, 2018.

TOYAMA, Danyelle. **Metagenoma da Amazônia: Busca por genes de interesse biotecnológico**. 2016. 125 f. Doctoral thesis – Federal University of Sao Carlos, 2016.

TOYAMA, Danyelle et al. Metagenomics Analysis of Microorganisms in Freshwater Lakes of the Amazon Basin. **Genome Announc**, v. 4, n. 6, p. 1440–16, 2016.

TULLY, Benjamin J. et al. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. **PeerJ**, v. 5, p. e3558, 2017.

UNIPROT CONSORTIUM. UniProt: a hub for protein information. **Nucleic Acids Research**, v. 43, n. D1, p. D204–D212, 2015.

VAN DEN BRINK, Joost; DE VRIES, Ronald P. Fungal enzyme sets for plant polysaccharide degradation. **Applied Microbiology and Biotechnology**, v. 91, n. 6, p. 1477–1492, 2011.

VAN ROSSUM, Thea et al. Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. **Frontiers in Microbiology**, v. 6, p. 1405, 2015.

VARGHESE, Neha J. et al. Microbial species delineation using whole genome sequences. **Nucleic Acids Research**, v. 43, n. 14, p. 6761–6771, 2015.

VICUÑA, Rafael. Bacterial degradation of lignin. **Enzyme and Microbial Technology**, v. 10, n. 11, p. 646–655, 1988.

VIDAL, Luciana O. et al. Hydrological pulse regulating the bacterial heterotrophic metabolism between Amazonian mainstems and floodplain lakes. **Frontiers in Microbiology**, v. 6, p. 1054, 2015.

VOELKER, Steven L. et al. Reduced wood stiffness and strength, and altered stem form, in young antisense 4CL transgenic poplars with reduced lignin contents. **New Phytologist**, v. 189, n. 4, p. 1096–1109, 2011.

VOLLMERS, John; WIEGAND, Sandra; KASTER, Anne-Kristin. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! **PLOS ONE**, v. 12, n. 1, p. e0169662, 2017.

WARD, Nicholas D. et al. Degradation of terrestrially derived macromolecules in the Amazon River. **Nature Geoscience**, v. 6, n. 7, p. 530–533, 2013.

WARD, Nicholas D. et al. The reactivity of plant-derived organic matter and the potential importance of priming effects along the lower Amazon River. **Journal of Geophysical Research: Biogeosciences**, v. 121, n. 6, p. 1522–1539, 2016.

WARNES, Gregory R. et al. **gplots: Various R Programming Tools for Plotting Data**. Disponível em: <https://cran.r-project.org/web/packages/gplots/index.html>. Acesso em: 6 nov. 2017.

WEI, Taiyun; SIMKO, Viliam. **R package "corrplot": Visualization of a Correlation Matrix**. Disponível em: <https://github.com/taiyun/corrplot>. Acesso em: 24 out. 2017.

WERNER, Jeffrey J. et al. Impact of Training Sets on Classification of High-Throughput Bacterial 16s RRNA Gene Surveys. **The ISME Journal**, v. 6, n. 1, p. 94–103, 2012.

WICKHAM, Hadley. **Ggplot2 : elegant graphics for data analysis**. [S.l.]: Springer, 2009.

WINNEN, Brit; HVORUP, Rikki N.; SAIER, Milton H. The tripartite tricarboxylate transporter (TTT) family. **Research in Microbiology**, v. 154, n. 7, p. 457–465, 2003.

WISSMAR, R. C. et al. Plankton Metabolism and Carbon Processes in the Amazon River, Its Tributaries, and Floodplain Waters, Peru-Brazil, May-June 1977. **Ecology**, v. 62, n. 6, p. 1622–1633, 1981.

XUE, Saisi et al. Water-Soluble Phenolic Compounds Produced from Extractive Ammonia Pretreatment Exerted Binary Inhibitory Effects on Yeast Fermentation Using Synthetic Hydrolysate. **PLOS ONE**, v. 13, n. 3, p. e0194012, 2018.

YE, Yuzhen; DOAK, Thomas G. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. **PLoS Computational Biology**, v. 5, n. 8, p. e1000465, 2009.

YILMAZ, Pelin et al. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. **Nucleic Acids Research**, v. 42, n. D1, p. D643-8, 2014.

YIN, Yanbin et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. **Nucleic acids research**, v. 40, n. Web Server issue, p. W445–51, 2012.

YOON, Byung-Jun. Hidden Markov Models and their Applications in Biological Sequence Analysis. **Current genomics**, v. 10, n. 6, p. 402–415, 2009.

YU, Nancy Y. et al. PSORTb 3.0: Improved Protein Subcellular Localization Prediction with Refined Localization Subcategories and Predictive Capabilities for All Prokaryotes. **Bioinformatics**, v. 26, n. 13, p. 1608–1615, 2010.

ZHANG, Tong; SHAO, Ming-Fei; YE, Lin. 454 Pyrosequencing Reveals Bacterial Diversity of Activated Sludge from 14 Sewage Treatment Plants. **The ISME Journal**, v. 6, n. 6, p. 1137–1147, 2012.

ZIGANSHIN, Ayrat M. et al. Comparative Analysis of Methanogenic Communities in Different Laboratory-Scale Anaerobic Digesters. **Archaea**, v. 2016, p. 1–12, 2016.

ZIPPER, C et al. Complete microbial degradation of both enantiomers of the chiral herbicide mecoprop [(RS)-2-(4-chloro-2-methylphenoxy)propionic acid] in an enantioselective manner by Sphingomonas herbicidovorans sp. nov. **Applied and environmental microbiology**, v. 62, n. 12, p. 4318–22, 1996.

APÊNDICES

Devido ao grande volume ocupado pelas tabelas e dados apresentados neste tópico, os 12 apêndices referidos no texto estão disponíveis *online* permanentemente no link. E podem ser citados conforme segue:

Célio Dias Santos Júnior. (2019). Material suplementar da tese intitulada "Degradação de matéria orgânica terrestre por microrganismos do rio Amazonas - Metagenômica e Genômica Populacional" [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2530038.