UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Biologia

MELLINE FONTES NORONHA

METAGENÔMICA COMPARATIVA DE COMUNIDADES MICROBIANAS DE SOLOS DE BIOMAS GLOBAIS

CAMPINAS
2016

MELLINE FONTES NORONHA

METAGENÔMICA COMPARATIVA DE COMUNIDADES MICROBIANAS EM SOLOS DE BIOMAS GLOBAIS

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Genética e Biologia Molecular, na Àrea de Bioinformática

Supervisor/Orientador: VALÉRIA MAIA MERZEL

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL TESE DEFENDIDA PELA ALUNA MELLINE FONTES NORONHA, E ORIENTADA PELA DRA. VALÉRIA MAIA MERZEL.

CAMPINAS
2016

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

N789m

Noronha, Melline Fontes, 1985-
Metagenômica comparativa de comunidades microbianas de solos de biomas globais / Melline Fontes Noronha. – Campinas, SP : [s.n.], 2016.

Orientador: Valéria Maia Merzel.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Metagenômica. 2. Solos. 3. Ecossistema. 4. Carboidratos. 5. Drogas - Resistência em microrganismos. I. Oliveira, Valéria Maia de,1966-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Comparative metagenomics of microbial communities from global biome soils
Palavras-chave em inglês:
Metagenomics
Soils
Ecosystem
Carbohydrates
Drug resistance in microorganisms
Área de concentração: Bioinformática
Titulação: Doutora em Genética e Biologia Molecular
Banca examinadora:
Valéria Maia Merzel [Orientador]
Rodrigo Mendes
Rodrigo Gouvêa Taketani
Lucélia Cabral
Marcelo Falsarella Carazzolle
Data de defesa: 15-12-2016
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, 15/12/2016.

COMISSÃO EXAMINADORA

Profa. Dra. Valéria Maia Merzel

Dr. Rodrigo Mendes

Dr. Rodrigo Gouvêa Taketani

Dra. Lucélia Cabral

Dr. Marcelo Falsarella Carazzolle

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

AGRADECIMENTOS

À minha orientadora Dra. Valéria Maia Merzel, primeiramente por me aceitar como sua aluna faltando menos de 2 anos para o fim do meu doutorado, foi mesmo um desafio! Obrigada por toda a sua contribuição, dedicação, compreensão e colaboração para o desenvolvimento deste trabalho.

Ao meu namorado, colega e co-autor deste trabalho, Gileno, pela paciência, dedicação e contribuição ao longo desse trabalho. Obrigada por me apoiar sempre nesse trabalho tão intenso.

Ao meu supervisor Dr. Jack Gilbert, por me aceitar em seu laboratório e pela sua contribuição nesse trabalho. Agradeço também pelos conselhos que me guiam até hoje. Obrigada a todos os colegas do Argonne National Laboratory, em especial a minha colega de sala e amiga Iratxe Zarranaoindia, por todo ensinamento, carinho e companherismo.

Aos meus colegas da DRM pela troca de conhecimentos e pelos momentos de descontração. E ao CPQBA, por toda a infraestrutura. Um obrigada especial ao Gustavo por me ajudar a manter o servidor longe das *Leis de Murphy*.

Aos meus colegas da bioinformática do LGE por todos os ensinamentos e pela amizade. Em especial a minha amiga Luciana Mofatto, por todos esses anos de amizade, me dando conselhos e ouvindo meus desabafos, além da troca de conhecimentos e da parceria. Agradeço também ao Marcelo Carazzolle por todos os ensinamentos dados todos esses anos e a Juliana pela amizade e apoio artístico em nossos trabalhos.

À minha família, especialmente ao amor e apoio incondicional da minha mãe e da minha querida avó Lais (*in memoriam*). Sei que se estivesse aqui, ficaria muito orgulhosa de ver a conclusão do meu trabalho. As minhas pequeninas irmãs, Bia e Manu, por tentarem compreender que a vida da sua irmã mais velha não é só brincar, e que ficar longe delas também é dificil para mim. Ao meu pai e a Kátia, por todo apoio. Aos meus outros avós Antônio (*in memoriam*), Selma, Walter e Ângela.

À minha família americana, em especial a *host mom* Maryann Calabrese, por me adotar como sua filha. Sei que esse amor é sincero e recíproco.

Aos meus amigos e amigas que direta ou indiretamente contribuiram para a conclusão deste trabalho, em especial a Daniele, Amanda, Rafaela, Claryanne, Dayana, Geórgia e Luana.

À Petrobrás, CAPES e CNPQ pelo apoio financeiro.

“O fracasso é apenas uma oportunidade para recomeçar com mais inteligência”

Henry Ford

RESUMO

Um bioma é uma unidade geográfica caracterizada de acordo com seu tipo de vegetação, macro-clima, solo e altitude específica. Em contraste, o microbioma é a totalidade de micro-organismos e seus genomas coletivos presentes em um dado ambiente. Embora as comunidades microbianas de solo demonstrem grande variação quando comparadas em diferentes escalas espaciais em um mesmo espaço amostral, o microbioma desses solos parecem ser direcionados de acordo com algumas características biogeográficas. A fim de investigar a convergência taxonômica e funcional em solos pertencentes ao mesmo tipo de bioma, trinta metagenomas disponíveis publicamente foram selecionados a partir de 11 biomas globais e agrupados com base nas características da vegetação (ou seja, floresta, pastagem, tundra, semi-árido e deserto). As análises funcionais revelaram um padrão entre os grupos de biomas, nos quais biosíntese de proteínas, metabolismo central dos carboidratos, e resistência a antibióticos foram os metabolismos com diferenças estatisticamente mais significativas baseado na anotação dos subsistemas do SEED. A fim de proporcionar uma melhor resolução analítica desses metabolismos, as sequências metagenômicas foram anotadas utilizando os bancos de dados de enzimas ativas em carboidratos (Cazy), de genes de resistência a antibióticos (ARDB) e, adicionalmente, de proteínas de choque térmico (HSPIR). A análise de enzimas ativas em carboidratos mostrou que a degradação da biomassa, metabolismo da sacarose e do amido, síntese da parede celular e degradação de alginato foram mais abundantes em solos florestais e em pastagens. Como esperado, genes de resistência à dessecação e a outros estresses foram mais abundantes em solos de desertos e semi-áridos. Genes de resistência a antibióticos (ARGs) foram predominantes em solos florestais e de pastagem, onde a resistência a multidrogas a partir de bombas de efluxo foram as classes mais abundantes, com a maior parte das sequências anotadas taxonomicamente como afiliadas a Proteobacteria. Proteínas de choque térmico (HSPs) foram mais abundantes em solos de tundra, semi-áridos e desertos. Embora HSP70 e Hsp100 tenham se mostrado uniformemente distribuídas entre os biomas, HSP90 foi estatisticamente mais abundante em solos de florestas e pastagens em comparação aos outros grupos de biomas. Ainda, HSP60 e HSP20, que foram predominantemente anotadas para o domínio de arquéias, foram mais abundantes nos solos de deserto salino. Nossos resultados sugerem que as condições ambientais locais regem o enriquecimento de funções específicas importantes para a sobrevivência dos micro-organismos nesses ecossistema

ABSTRACT

Biome is a geographical unit characterized according to its vegetation type, macro-climate, soil, and specific altitude. In contrast, a microbiome is the totality of microorganisms and their collective genetic material present in a given environment. Although soil microbial communities have shown to vary across many spatial scales, soils between ecosystems showed to be driven by some biogeographical trends. In order to investigate taxonomic and functional convergence within soils from the same biome type, thirty publically-available metagenomes from 11 globally distributed biomes were selected and clustered by biome groups (i.e. forest, grassland, tundra, semiarid and desert) based on vegetation features. Functional analyses revealed a close pattern among biome groups, in which protein biosynthesis, central carbohydrate metabolism, and antibiotic resistance were the most statistically different metabolisms annotated by SEED subsystems. In order to provide a better analytical resolution of those metabolisms, metagenomic reads were annotated using the Carbohydrate-Active enZYmes database (Cazy), Antibiotic Resistance gene DataBase (ARDB) and, additionally, the Heat Shock Protein Information Resource (HSPIR). Carbohydrate-active enzyme analyses showed that biomass degradation, sucrose and starch metabolism, cell wall biosynthesis and alginate degradation were overrepresented in forest and grasslands soils. As expected, desiccation and other stress resistance genes were more abundant in deserts and semiarid soils. Antibiotic Resistance Genes (ARGs) were prevalent in forest and grassland soils, where multidrug efflux pumps were the most abundant ARG class, with the majority of the reads assigned to Proteobacteria. Heat Shock Proteins (HSPs) were more abundant in tundra, semiarid and desert soils. Although HSP70 and HSP100 were uniformly distributed across biomes, HSP90 were overrepresented in forest and grassland soils when compared to others biomes groups. HSP60 and HSP20, which are predominantly from Archaea, were more abundant in the saline desert soils. Our results suggest that local environmental conditions select for the enrichment of specific functions important for microbial survival in those ecosystems.

SUMÁRIO

1. INTRODUÇÃO

Bioma é uma região geográfica caracterizada de acordo com sua vegetação, macroclima, solo e altitude específicos. Por outro lado, o microbioma é o conjunto de microrganismos que coexistem em um determinado ambiente ou nicho, em uma escala menor. Até recentemente, as comunidades microbianas não eram incluídas como um fator que contribuísse para a classificação dos biomas. Entretanto, apesar das comunidades microbianas de solos mostrarem-se bastantes heterogêneas quando comparadas em diferentes escalas espaciais, a composição do microbioma de ecossistemas específicos tem se mostrado relacionada a fatores biogeográficos tais como o pH, vegetação, disponibilidade de água e nutrientes (O'Brien *et al.*, 2015).

Nos últimos anos, alguns trabalhos vêm abordando uma possível correlação entre o microbioma e as características do ambiente no qual essa comunidade está inserida. Em um trabalho pioneiro, Lauber e colaboradores sequenciaram uma região do gene RNA ribossomal 16S de microrganismos de solo em uma escala continental. Nesse trabalho, eles concluiram que o pH era o fator ambiental que mais influenciava na diferenciação entre as comunidades bacterianas e que, ambientes que possuiam pH neutro apresentavam maior diversidade de microrganismos quando comparadas a condições mais alcalinas ou ácidas (Lauber *et al.*, 2009). Já em um outro trabalho realizado em escala continental na China, baseado também no sequenciamento de amplicons de RNAr 16S, a disponibilidade de água foi o fator que mais influenciou a estrutura e estabilidade da comunidade microbiana dos solos estudados (B. Ma *et al.*, 2016). Apesar de apresentarem fatores fisico-quimicos distintos, esses trabalhos evidenciam que fatores abióticos podem influenciar a estrutura microbiana de um ecossistema.

Em uma abordagem taxonômica e funcional, Fierer e colaboradores realizaram o sequenciamento *shotgun* do metagenoma de diferentes solos, os quais foram classificados em amostras de solos desérticos (quente e polar) e não desérticos (floresta, campo e tundra). Nesse trabalho, diferenças significativas foram encontradas na comparação entre esses ambientes, como a maior abundância de genes relacionados à osmorregulação em desertos, e de genes relacionados à resistência a antibióticos e lise celular em regiões não desérticas (Noah Fierer *et al.*, 2012). Esse trabalho foi de grande importância para melhor compreender a relação entre o microbioma e os biomas em que estão inseridos. Entretanto, classificar os solos apenas em desertos e não desertos

não permite explorar as variações menores encontradas, por exemplo, entre a vegetação de florestas e de campos, como as árvores de grande porte encontradas em florestas que podem influenciar o microbioma ao seu redor. Além disso, a análise funcional das amostras foi pouco explorada entre as amostras.

A fim de ampliar e detalhar as análises realizadas por Fierer e seu colegas, selecionamos 30 amostras de metagenoma de solos de 13 biomas terrestres diferentes a partir de bancos de dados públicos. Além disso, para realizar uma análise mais detalhada acerca das características encontradas nesses biomas, classificamos essas amostras em 5 grupos de biomas distintos (floresta, pastagem, semi-árido, deserto e tundra). Para uma maior resolução funcional dos mecanismos que moldam essas comunidades microbianas, bancos de dados específicos foram utilizados no estudo de genes relacionados à resistência a antibióticos, à degradação de carboidratos e às proteínas de choque térmico (*Heat Shock proteins*).

2. OBJETIVOS

Este trabalho teve como objetivo geral a comparação taxonômica e funcional do metagenoma microbiano de solos representativos de diferentes biomas do mundo.

Objetivos Específicos:

2.1.1. Avaliar se o perfil de enzimas envolvidas na degradação de carboidratos apresenta diferenciação específica com base na vegetação encontrada em cada bioma;

2.1.2. Correlacionar o perfil do resistoma (conjunto de genes de resistência a antibióticos) e os respectivos *taxa* que possivelmente conferem essas resistências em cada um dos biomas

2.1.3. Identificar os perfis de proteína de choque térmico em cada um dos biomas.

3. REVISÃO BIBLIOGRÁFICA

3.1. Solos

Solos são conhecidos por serem os ambientes que abrigam a maior diversidade de microrganismos entre todos os ambientes/ecossistemas da Terra. Em 1990, uma análise independente de cultivo utilizando hibridização de DNA estimou aproximadamente 10.000 espécies de bactérias em solos (Torsvik, Goksyr e Daae, 1990; Roesch *et al.*, 2007). Entretanto, com o advento das técnicas moleculares de sequenciamento de alto desempenho (*high throughput sequencing*) nas últimas décadas, essa estimativa mostrou-se bastante subestimada em relação à diversidade bacteriana fenomenal encontrada em solos. Segundo o trabalho de Raynaud & Nunan, um grama de solo pode abrigar até 10^{10} células bacterianas, contendo uma riqueza estimada entre $4\text{-}10^3$ a $5\text{-}10^4$ espécies (Raynaud e Nunan, 2014).

A diversidade bacteriana encontrada nos diferentes solos parece variar em função de alguns fatores ambientais como pH, nitrogênio, comunidades de plantas ou uso do solo, e a biomassa bacteriana do solo (carbono orgânico do solo). Com isso, há uma grande variação do microbioma entre diferentes regiões geográficas (Raynaud e Nunan, 2014). No trabalho apresentado por Lauber e colaboradores, regiões 16S ribossomais de amostras de solos foram extraídas e correlacionadas a fatores físico-quimicos analisados dessas mesmas amostras. Esse trabalho concluiu que o pH foi o fator que melhor explicou a variabilidade associada as alterações observadas na estrutura filogenética (Lauber *et al.*, 2009). Já em uma análise de comunidades microbianas em solos submetidos a diferentes gradientes de nitrogênio, a abundância dos membros dos filos Proteobacteria e Bacteroidetes aumentou enquanto que a dos membros de Acidobacteria diminuiu com o aumento de nitrogênio (N. Fierer *et al.*, 2012). Além disso, a vegetação também é uma fator ambiental que pode influenciar a diversidade microbiana, como mostrado no trabalho comparando solos de tundra em diferentes sazonalidades (congelado e descongelado – com vegetação rasteira) (Shi, Xiang, *et al.*, 2015). Ainda, dados da literatura demonstraram que a agricultura pode exercer grande impacto sobre a diversidade taxonômica e funcional em solos de cerrado (Souza *et al.*, 2016) (Souza *et al.*, 2016) e da floresta Amazônica (Navarrete *et al.*, 2015).

3.2. Biomas

Bioma é uma região geográfica caracterizada de acordo com sua vegetação, macroclima, solo e altitude específicos. Já o ecossistema é definido como o relacionamento entre os seres vivos e o ambiente físico em que se encontram. Portanto, um bioma seria formado por um conjunto de ecossistemas. Dentre os biomas terrestres, estes podem ser classificados entre regiões árticas, sub-árticas, temperadas e tropicais (Figura 1), sendo diferenciados principalmente pela temperatura e umidade. As regiões árticas são aquelas localizadas próximas aos pólos portanto com menor radiação solar, as regiões temperadas seriam intermediárias (entre árticas e tropicais) e as tropicais, mais próximas à linha do equador, com maior incidência solar.

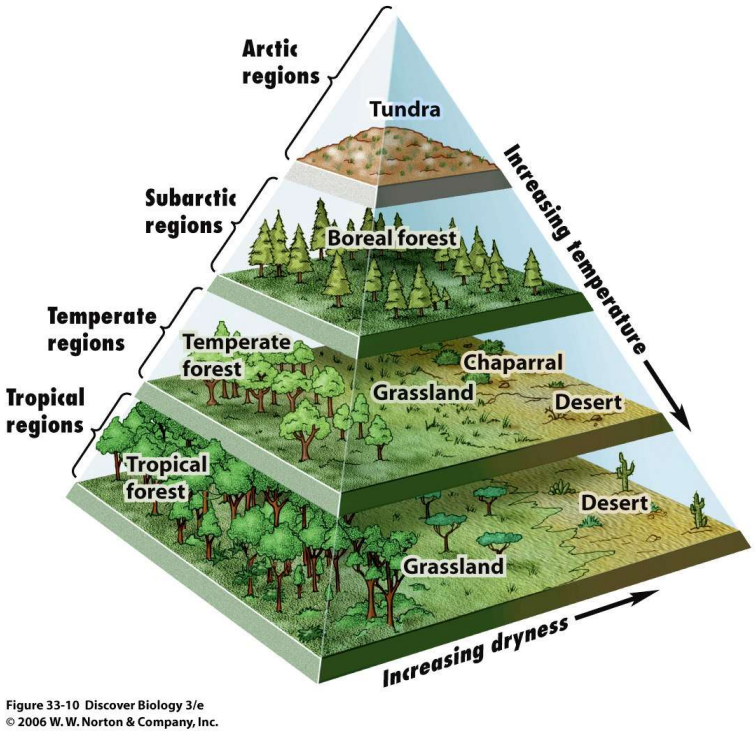


Figura 1. Classificação dos biomas em regiões árticas, sub-árticas, temperadas e tropicais, de acordo com humidade e temperatura média (Loo, 2010).

As vegetações nessas regiões também são bastante diferenciadas entre si. A tundra é formada por uma camada de gelo durante a maior parte do ano, e no verão, com o derretimento do gelo (parcial ou total), é predominantemente composta por arbustos, gramíneas, musgos e líquens. Já na floresta boreal, o gelo é completamente derretido no verão, permitindo assim uma vegetação de grande porte formada por larícios, abetos, pinheiros e espruces. As florestas temperadas podem ser classificadas, baseada no aspecto da vegetação, em coníferas (árvores em

forma de cone) ou decíduas (árvores perdem folhas no outono), e possuem as quatro estações do ano bem definidas, permitindo a existência de árvores como o carvalho, os bordos, as faias e as nogueiras. A vegetação pastagem/estepe é predominantemente formada por planícies com poucas árvores, compostas por herbáceas e pequenos bosques. Chaparral ou mediterrâneo, similar à Caatinga, é uma região sazonal com árvores espassadas relativamente densas, entretanto com baixo desenvolvimento devido aos longos períodos de seca. As florestas tropicais são bastante densas, com clima constantemente quente e úmido. E por fim, os desertos apresentam pouca precipitação pluviométrica, podendo ser formados por dunas ou vegetação espassada com gramíneas e pequenos arbustos (Figura 2).

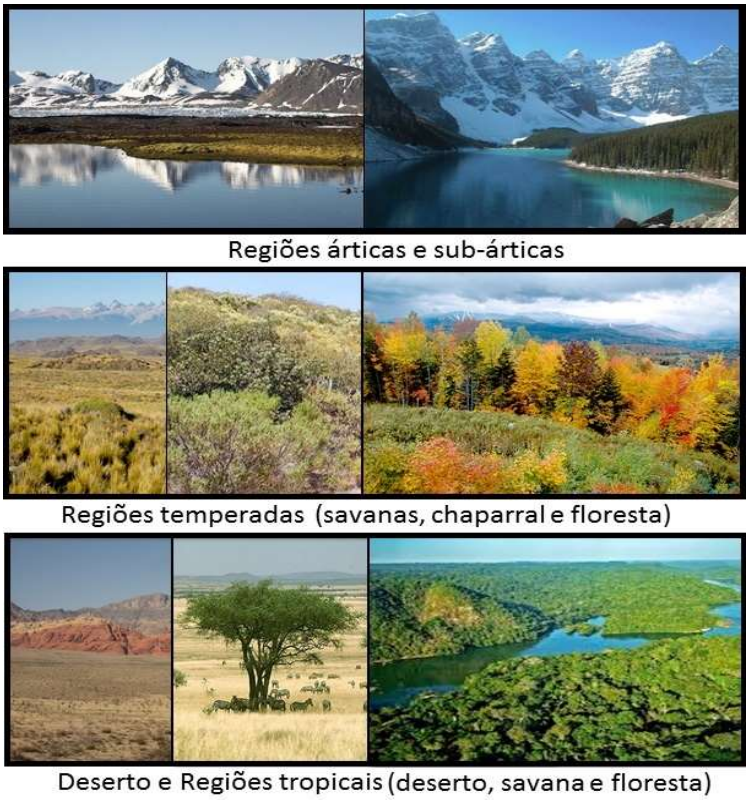


Figura 2. Representação visual dos biomas divididos por regimes (sub-)árticas, temperadas e tropicais (George, Kathryn e Sarah, sem data; John Muir School, sem data; Louro, sem data; The nature conservancy, sem data; The Osprey Brand Team, 2009; Aborn e Berbine, 2010; Chinwuba, 2014; Belezas Naturais, 2015).

A importância dos diferentes biomas terrestres para o planeta vem sendo alvo de muitos estudos. Em 2004, através de uma simulação realizada excluindo os diferentes tipos de biomas da Terra, baseado no modelo CCM3-IBIS, previu-se que a destruição de florestas tropicais, temperadas e/ou boreais poderia acarretar um esfriamento durante o inverno e a primaveira, e um aquecimento global durante o verão. Além disso, a remoção de pastagens e estepes poderia causar

aquecimento e baixa humidade na atmosfera durante o verão, principalmente nos Estados Unidos. Já os arbustos e a vegetação da tundra poderiam influenciar principalmente a Austrália, devido à redução de esfriamento latente, e a Sibéria central, onde o transporte de ar frio seria reforçado (Snyder, Delire e Foley, 2004). Estudos nesse âmbito vêm sendo realizados principalmente com questões relativas ao aquecimeno global. Recentemente, um estudou baseado em projeção ressaltou a importância da manutenção de ambientes naturais (limitando o uso da terra em plantios) para evitar os impactos das mudanças climáticas na América Latina (Boit *et al.*, 2016).

3.3. Microbiomas

Da mesma forma que os fatores abióticos influenciam e são influenciados pela vegetação, fatores bióticos, como a interação entre as plantas e os microrganismos, também podem influenciar um dado ecossistema. São alguns exemplos: a maior germinação da cana-de açúcar identificada após a inoculação de uma bactéria isolada a partir da própria rizosfera de cana (Beneduzi *et al.*, 2013), a fixação de nitrogênio identificada em 7 bactérias endofíticas no arroz (Hongrittipun, Youpensuk e Rerkasem, 2014), e um nível menos severo de patogenicidade observado na contaminação do fungo *Fusarium oxysporum* f. sp. em tomates, quando este estava associado ao aparecimento de *Pseudomonas fluorescens* WCS365 (Kamilova *et al.*, 2006). Essa associação do microbioma com o seu ecossistema não necessariamente está relacionada a um ou poucos microrganismos, como no trabalho onde a degradação das fibras das plantas e a fermentação dos produtos em acetato foi identificada como sendo realizada por um conjunto de simbiontes do cupim, sendo eles bactérias, protozoários e arquéias. Além disso, esse trabalho resaltou a contribuição desse conjunto de simbiontes para o ciclo do nitrogênio em solos através da mineralização de componentes do húmus (Brune, 2014). Esses trabalhos apontam uma forte relação do ecossistema com os microrganismos que o habitam, portanto, identificar esses microrganismos e a influência que eles exercem no seu habitat são fatores importantes no estudo de microbiomas.

3.4. Metagenômica

A metagenômica é o estudo do conjunto de genomas recuperado diretamente a partir de amostrais ambientais. Estima-se que apenas 1-10% dos microrganismos sejam cultiváveis utilizando as técnicas padrões de cultivo laboratorial, portanto extrair o DNA diretamente do ambiente permite o acesso a dados genéticos de organismos ainda não cultivados. Em 1987, uma

metodologia de filotipagem foi demonstrada em um trabalho de Carl Woese. Essa metodologia utiliza regiões dos genes RNA ribossomais 16S e 5S a partir do DNA extraido diretamente da amostra ambiental para diferenciar grupos filogenéticos (Woese, 1987). A partir dessa metodologia e do surgimento da tecnologia da reação em cadeia da polimerase (PCR), outras técnicas moleculares foram desenvolvidas e adaptadas para o estudo da ecologia microbiana, como DGGE (*Denaturing Gradient Gel Electrophoresis*), TGGE (*Temperature Gradient Gel Electrophoresis*), RFLP (*Restriction Fragment Length Polymorphism*), PCR-ELISA, entre outros. Entretanto, essas técnicas são limitadas a identificar os grupos filogenéticos com base em genes marcadores especificos, não conseguindo acessar o material genético total dos microrganismos. O primeiro trabalho baseado na extração e clonagem do DNA total de uma amostra ambiental foi relatada em comunidades microbianas marinhas pelo grupo do Schmidt e colaboradores (Schmidt, DeLong e Pace, 1991). Já em 1998, surgiu o termo metagenômica definido como o estudo dos genomas coletivos de uma determinada amostra ambiental (Handelsman *et al.*, 1998).

Nos últimos anos, com o advento das novas tecnologias de sequenciamento (*high throughput sequencing*), tem sido possível extrair todo o DNA presente em uma dada amostra e realizar o sequenciamento diretamente, sem a utilização de vetores e hospedeiros. Essas novas tecnologias são conhecidas como tecnologias de sequenciamento de segunda geração, sendo a pioneira a plataforma 454 (Roche - descontinuada) em 2005, seguida pela Illumina (Solexa) em 2006, a SoliD em 2007 e a Ion Torrent em 2010. Apesar dessas tecnologias serem distintas entre si, todas são baseadas em amplificação das moléculas por clonagem, limitando assim o tamanho dos *reads* nesse tipo de sequenciamento (Fatih Ozsolak, 2013). Mais recentemente, surgiram as tecnologias de terceira geração, sem necessidade de clonagem, como a Helicos BioSciences, Nanopore e Pacific Biosciences. Essas tecnologias permitem obter uma maior precisão na leitura dos fragmentos e também um maior comprimento dos *reads* (Fatih Ozsolak, 2013). Portanto, o advento dessas novas tecnologias de sequenciamento trouxe, e vem trazendo, grandes avanços principalmente para análises de amostras de sequenciamento *shotgun*, permitindo que em poucos dias, ou até mesmo em horas, seja possível acessar o material genético total de um ambiente.

O sequenciamento do metagenoma de amostras ambientais vem sendo bastante utilizado para entender a dinâmica das comunidades microbianas em diferentes ambientes. Uma abordagem é o sequenciamento de marcados filogenéticos, como regiões do gene RNAr 16S para a identificação de bactérias e de regiões ITS para fungos. Esse método chamado de *amplicon*

sequencing ou sequenciamento de amplicons vem sendo amplamente aplicado no estudo da composição e diversidade de taxa microbianos em habitats diversos, como no estudo do microbioma do intestino humano (Qin *et al.*, 2010), no acompanhamento da variação dos grupos taxonômicos em solos contaminados por óleo (Peng, Zi e Wang, 2015) (Peng, Zi e Wang, 2015), na comparação do microbioma antes e após renovação da água em um aquário (LaPointe *et al.*, 2015), dentre inúmeros outros exemplos.

O sequenciamento de *amplicons* permite a identificação das taxonomias associadas àquele ambiente com um custo relativamente baixo. Entretanto, esse tipo de análise limita o conhecimento sobre o potencial metabólico daquela comunidade, sendo necessário portanto o sequenciamento de todo o material genético proveniente da amostra ambiental. Em um trabalho sobre ambientes euxínicos modernos foi possível correlacionar a biodiversidade com as características funcionais do ambiente através da metagenômica *shotgun* (Llorens-Marès *et al.*, 2015). Ainda, uma análise comparativa do resistoma encontrado em amostras de fezes de porco, galinha e de humanos (L. Ma *et al.*, 2016) e respostas da estrutura da comunidade microbiana (taxonômica e funcional) em áreas de desmatamento na Amazônia (Navarrete *et al.*, 2015) são outros exemplos de trabalhos que com sucesso associaram parâmetros ambientais com dados funcionais e taxonômicos a partir de dados metagenômicos. Além disso, o sequenciamento *shotgun* de amostras ambientais permitiu a montagem de genomas *draft* de microrganismos não cultivados, como da SAR324 Bacterium lautmerah10, através da metagenômica de amostras do Mar Vermelho (Haroon, Thompson e Stingl, 2016) e de diversos genomas *draft*, incluindo o de *Butyrivibrio fibrisolvens*, a partir de dados de metagenômica de rumen bovino (Hess *et al.*, 2011).

3.5. Metagenômica de solos

Os solos são ambientes complexos e de alta diversidade microbiana. Entretanto, a contribuição desses microrganismos para a manutenção da estabilidade dos ecossistemas ainda não foi totalmente compreendida. Estudos biogeográficos de ambientes de solo tem ajudado a determinar os fatores ecológicos chave (seleção, dispersão, especiação) que definem a diversidade e a composição microbiana dentro de cada ecossistema (Pasternak *et al.*, 2013; Ranjard *et al.*, 2013; Shi, Grogan, *et al.*, 2015). Uma nova abordagem promissora baseada em re-montagem de genomas em gradientes biogeográficos sugere que em breve será possível deduzir as pressões seletivas locais que moldam a especiação de algumas espécies bacterianas em particular em ecossistemas complexos (Sangwan *et al.*, 2016). Tal pressão seletiva pode ocorrer em diferentes

escalas espaciais, incluindo uma escala de sub-centímetros, levando a mudanças na abundância relativa de taxa de bactérias específicas em amostras altamente próximas (O'Brien *et al.*, 2015). No entanto, apesar desta aparente alta diversidade microbiana em solos próximos, foi possível encontrar padrões da composição da comunidade microbiana entre solos de ecossistemas similares em amostras distribuídas globalmente. Isto sugere que a menor resolução espacial, a seleção e a especiação (e possivelmente a limitação da dispersão) podem agir para formar comunidades microbianas discretas dentro de cada habitat (O'Brien *et al.*, 2015).

Além de identificar padrões taxonômicos, alguns trabalhos vem buscando encontrar também possíveis padrões funcionais em amostras de ecossistemas específicos. Com o objetivo de identificar os principais metabolismos associados aos diferentes biomas, terrestres ou aquáticos, 87 conjuntos de dados de sequenciamento *shotgun* (microbioma e viroma) agrupados em 9 biomas distintos foram investigados: subterrâneo, associados a animais terrestres, marinho, água doce, lagoas hipersalinas, associados a corais, microbialitos, associados a aquacultura/peixes e associados a mosquitos. Baseado nessas análises, os autores encontraram evidências de quais metabolismos são importantes para cada comunidade microbiana de acordo com dados ambientais em que estão inseridos. Por exemplo, eles observaram que os subsistemas envolvidos na respiração e do metabolismo protéico permitiram diferenciar o microbioma associado aos corais do microbioma de animais terrestres. Por outro lado, eles observaram que os genes de virulência foram proporcionalmente mais abundantes na microbiota associada a organismos hospedeiros (simbiontes) do que nos micróbios de vida livre. Em particular, dois subsistemas, metabolismo alcanossulfonato e purina, foram sobre-representados em metagenomas associados a peixes. Portanto, eles concluiram que a maior parte da diversidade funcional foi mantida em todas as comunidades, mas a ocorrência relativa de metabolismos e as características que diferenciam os metagenomas permitiram inferir as condições bioquímicas encontradas em cada ambiente (Dinsdale *et al.*, 2008).

Em 2009, Lauber e colaboradores obtiveram amostras de 88 solos de diferentes biomas terrestres distribuídos ao longo da América do Norte e do Sul. Nesse trabalho, foi realizada a análise de diversidade com base no sequenciamento parcial do gene RNAr 16S do metagenoma dessas amostras e, posteriormente, a estrutura dessas comunidades microbionas foi correlacionada com fatores físico-quimicos analisados nestas mesmas amostras. Eles observaram que as amostras de solos dos biomas eram todas dominadas por cinco grandes grupos taxonômicos: Acidobacteria,

Actinobacteria, Proteobacteria, Bacteroidetes e Firmicutes. Além disso, concluiram que o pH foi a característica físico-quimica que melhor explicou a variabilidade dos dados, corroborando com evidências encontradas em trabalhos prévios, inclusive utilizando metodologias diferentes. Os autores concluiram que o pH do solo muitas vezes está direta ou indiretamente relacionado com uma série de características do solo, como por exemplo disponibilidade de nutrientes, a solubilidade do metal catiônico, características de C orgânico, regime de umidade do solo e salinidade. Isto poderia explicar porque o pH era identificado como influenciador do microbioma em muitos conjuntos de dados (Lauber *et al.*, 2009).

Em um trabalho pioneiro realizado por Fierer e colaboradores, amostras de solos de diferentes biomas terrestres foram analisadas a partir do sequenciamento de *amplicons* (RNAr 16S), como também *shotgun,* com o objetivo de avaliar a estrutura taxonômica e funcional dessas comunidades. Neste contexto, amostras de 16 solos distintos foram coletadas, sendo três amostras de solos de desertos quentes, seis de desertos polares, quatro de florestas temperadas e tropicais, uma de campo (pastagem), uma de tundra e uma de floresta boreal. Essas amostras foram posteriormente agrupadas em amostras desérticas (desertos quentes e polares) e não desérticas (florestas, campos e tundras). Uma maior abundância relativa de genes associados aos metabolismos de nitrogênio, potássio e de enxofre foi encontrada nos solos não desérticos. Por outro lado, a exposição frequente a estresses ambientais pode explicar a maior abundância relativa de genes associados com dormência/esporulação, proteínas de estresse e metabolismo de aminoácidos (relacionados à osmorregulação em bactérias) em solos desérticos. Além disso, genes de resistência a antibióticos e outros genes possivelmente associados com a competição demonstraram uma maior abundância em solos desérticos em relação a solos não desérticos. Ainda, proteínas envolvidas na clivagem de peptidoglicanos bacterianos (associados à lise celular) foram mais abundantes em solos não desérticos. Este trabalho demonstrou o uso bem sucedido de abordagens metagenômicas para compreender como a diversidade filogenética e funcional microbiana varia entre os diferentes biomas terrestres (Noah Fierer *et al.*, 2012).

Em um trabalho mais recente, 33 metagenomas (sequenciamento *shotgun*) disponíveis publicamente e agrupados em florestas, pastagens, desertos, solos árticos e sedimentos de mangue foram analisados. Do total de amostras, catorze são de pastagem, sete de solo de floresta, nove de deserto, dois de solo do Ártico, e uma de sedimento de mangue. A anotação funcional revelou um enriquecimento significativo nos módulos (KEGG) identificados no microbioma de desertos em

comparação com a microbiota de pastagens e de solos florestais. Dentre estes, alguns módulos envolvidos no ciclo redutivo pentose fosfato, que é a principal via para a conversão de CO_2 atmosférico em compostos orgânicos. Por outro lado, alguns módulos metabólicos associados com metabolitos de origem vegetal (via do chiquimato e do sistema de transporte de aminoácidos ramificados) foram significativamente sobre-representados em pastagens e solos florestais quando comparados com amostras de desertos, o que já era esperado uma vez que solos de desertos apresentam pouca ou nenhuma vegetação (Xu *et al.*, 2014).

Neste trabalho de tese, realizamos uma análise comparativa de conjuntos de dados derivados do sequenciamento *shotgun* de solos de diversos biomas (agrupados em florestas, pastagens, semi-áridos, desertos e tundra). Amostras de biomas que não haviam sido anteriormente utilizadas, como Cerrado (savana tropical) e solos de regiões semi-áridas, foram incluídas em nossas análises, abrangendo e aprofundando ainda mais a análise dos diferentes biomas globais. Ainda, além de utilizar a anotação taxonômica e funcional dada a partir de bancos de dados gerais, como SEED (Overbeek *et al.*, 2005), o KEGG (Kanehisa e Goto, 2000; Kanehisa *et al.*, 2016) e COG (Tatusov *et al.*, 2000), esses dados foram anotados funcionalmente utilizando também bancos de dados específicos. Portanto, segue nos próximos itens uma breve revisão sobre enzimas de degradação de carboidratos pelo banco de dados do CAZy (Cazymes), genes de resistência a antibióticos pelo banco de dados do ARDB e de proteínas de choque térmico utilizando o HSPIR.

3.6. Carbohydrate-Active enZYmes - *Cazymes*

Carboidratos são moléculas orgânicas que formam a principal fonte de energia dos seres vivos. Entretanto, para utilizar essas moléculas como fonte energética, é necessário um conjunto de enzimas responsáveis pela quebra da ligação entre carboidratos complexos e glicoconjugados. Portanto, buscar quais enzimas relacionadas à degradação de carboidratos estão presentes nos genomas dos microrganismos de uma comunidade microbiana, pode ajudar a entender quais as possíveis fontes energéticas disponíveis em cada microbioma.

O banco de dados de enzimas de degradação de carboidratos, ou enzimas ativas em carboidratos, mais utilizado atualmente é o *Carbohydrate-Active enZYmes database* (CAZy). O CAZy é um banco de dados público online (http://www.cazy.org) criado em 1998 com o objetivo de armazenar dados genômicos, estruturais e bioquímicos de enzimas ativas em carboidratos, denominadas *CAZymes*. O banco foi dividido em dois módulos. O primeiro é formado por enzimas que catalizam a quebra, a biossíntese ou a modificação de carboidratos ou glicoconjugados, enquanto que o segundo módulo é formados por enzimas que aderem a carboidratos. As *CAZymes* do primeiro módulo são divididas em 5 classes: glicosídeo hidrolases ou GHs (hidrólise e/ou rearranjo das ligações glicosídicas), glicosiltransferases ou GTs (formação de ligações glicosídicas), polissacarídeo liases ou PLs (clivagem não-hidrolítica de ligações glicosídicas), esterases ou CEs (hidrólise de ligação éster) e enzimas de atividades auxiliares ou AAs (enzimas redox que funcionam em conjunto com as CAZymes). O segundo módulo é formado apenas por uma classe denominada CBM ou módulo de ligação a carboidratos. Cada uma dessas classes ainda é subdividida em famílias, totalizando mais de 350 famílias entre todas as classes (Levasseur *et al.*, 2013; Lombard *et al.*, 2014).

A utilização do banco de dados do CAZy para buscar possíveis enzimas ativas de carboidratos em dados metagenômicos é bastante difundida em diversas áreas. Com o objetivo de compreender o papel das enzimas ativas de carboidratos na alimentação humana, um trabalho comparou o perfil de Cazymes encontradas no microbioma do intestino de 448 indivíduos. Foram encontradas diferenças nos perfis de *CAZymes* em crianças e recém-nascidos quando comparados ao de adultos, principalmente na abundância de enzimas de degradação de carboidratos simples (como a lactose e sacarose). Ainda, foi observada diferenças de perfis dessas enzimas baseadas na localização geográfica dos indivíduos. Além disso, encontraram um *core* de 89 famílias de *CAZymes* presente em 85% dos indivíduos (Bhattacharya, Ghosh e Mande, 2015). Em um trabalho

mais biotecnológico, a análise do perfil de *CAZymes* revelou um *core* de enzimas altamente conservadas dentro do grupo de amostras lignocelulolíticas, independentemente das suas composições taxonômicas (Mhuantong *et al.*, 2015). Outro trabalho mostrou as alterações do perfil de *CAZymes* ocorridas em uma região de floresta desmatada em relação à região de floresta natural. Neste trabalho foram identificadas 41 famílias de genes consistentemente afetadas pelo desmatamento, incluindo famílias envolvidas na degradação de lignina, celulose, hemicelulose e pectina, alterando o ciclo de carbono e de outros nutrientes no subsolo (Cardenas *et al.*, 2015). Além desses, há diversos outros trabalhos utilizando o banco de dados do CAZy para análise *in silico* de dados metagenômicos em diversos ambientes, como em comunidades de biofilme marinho (Sanli *et al.*, 2015), em sedimentos de mangues (Thompson *et al.*, 2013), entre outros.

3.7. Genes de resistência a antibióticos

Antibióticos são compostos antimicrobianos ou drogas capazes de combater uma infecção causada por microrganismos. Desde a descoberta acidental da penicilina em 1928 por Alexander Fleming, vários antibióticos vêm sendo utilizados para tratamento de infeções humanas e em animais, e também na agricultura (Pareek *et al.*, 2015). Por outro lado, sua importância no ambiente natural ainda não foi totalmente elucidada, sendo por muitos trabalhos sugerido que a sua produção pode estar relacionada com processos de competição, sinalização, e, em menor grau, na predação (Leisner, Jørgensen e Middelboe, 2016) .

O uso e o descarte irracional dessas substâncias, contaminando o meio ambiente, vem causando uma seleção e propiciando o desenvolvimento de cepas multiresistentes. A resistência microbiana a certos tipos de antibióticos é algo que existe naturalmente, entretanto o contato com antimicrobianos pode desconvolver um novo mecanismo de resistência em bactérias (antes não resistentes) a partir da transferência horizontal de genes. O surgimento de resistência a novos antibióticos por bactérias é preocupante, pois esse processo vem acumulando e acelerando ao longo do tempo (Pareek *et al.*, 2015). Embora as bactérias vêm se tornando cada vez mais resistentes, a descoberta de novas classes de antibióticos ocorre de forma incipiente (Ling *et al.*, 2015).

Com a crescente atenção das pesquisas para as questões envolvendo a resistência a antibióticos pelos microrganismos, foram desenvolvidas recentemente bases de dados de genes de resistência a antibióticos, como o ARDB (Liu e Pop, 2009). Essas bases vêm sendo utilizadas para

buscar genes de resistência a partir de sequências provenientes de sequenciamento de dados amostrais, como dados metagenômicos. Em um desses trabalhos, duas dessas bases de dados foram utilizadas para identificar possíveis genes de resistência a antibióticos (ARGs) em amostras de fezes humanas, de frangos e de porcos. Um grande número de sequências relacionadas à resistência a tetraciclina, multidrogas, eritromicina e aminoglicosídeo foram encontradas, sendo as amostras de fezes de frangos adultos as que exibiram maior número de ARGs detectadas (L. Ma *et al.*, 2016). Em outro trabalho, ARGs encontrados no intestino humano de indivíduos de diferentes países foram relacionados com o uso de antibióticos em humanos e animais em cada região. Eles encontraram que as amostras de indivíduos de países como Espanha, Itália e França apresentaram uma maior abundância de ARGs quando comparado a amostras de países como Dinamarca, Estados Unidos e Japão (Forslund *et al.*, 2013). Já em um trabalho de Li e colaboradores, foi possível observar um maior número de ARGs identificados em ambientes com ação antropogênica, baseado na análise de 50 amostras ambientais incluindo água doce, fezes humanas, solos, lodo, etc (Li *et al.*, 2015). Diversos outros trabalhos vêm utilizando a identificação de ARGs, principalmente com o intuito de entender a relação do resistoma ao uso de antibióticos nas diversas áreas de sua aplicação.

3.8. Proteínas de choque térmico (*Heat Shock Proteins*)

As proteínas de choque térmico (HSPs) são uma família altamente conservadas de chaperonas, que , apesar do nome remeter a proteínas envolvidas com variação térmica, também estão envolvidas em uma grande variedade de estresses bióticos e exposição à luz UV (Park e Seo, 2015). Elas possuem um papel importante em funções como dobramento/desdobramento de proteínas, montagem de complexos multiproteícos, transporte/seleção correta de proteínas em compartimentos subcelulares, controle do ciclo celular e de sinalização, e proteção das células contra estresse/apoptose (Li e Srivastava, 2004).

Apesar das HSPs serem conhecidas há bastante tempo, apenas recentemente um banco de dados reuniu informações sobre essas proteínas, o HSPIR – *Heat Shock Protein Information Resource* (Ratheesh Kumar *et al.*, 2012). Neste banco, essas proteínas foram divididas em famílias nomeadas por seu peso molecular expresso em kilodaltons, como HSP20 (small Hsp), Hsp40 (J-Proteins), HSP60 (Chaperoninas), HSP70, HSP90 e Hsp100. Recentemente, alguns trabalhos vêm estudando essas proteínas *in silico* em organismos ou microbiomas associados a ambientes de estresse, como no trabalho de transcriptômica da folha de árvore peônia exposta a temperaturas

elevadas, onde foram identificadas 24 potenciais *HSPs* (Zhang *et al.*, 2015), em uma análise de *small HSP* (sHsp) em vírus marinhos e em cianobactérias hospedeiras, e em um estudo das comunidades microbianas na Antártica e no Ártico envolvendo diversos genes relacionados a estresse, incluindo as *HSPs*, que foram encontradas em grande abundância principalmente as *cold shock proteins* Csps (Hsp100), DnaK (Hsp70) e DnaJ (Hsp70) (Varin *et al.*, 2012) .

4. MATERIAL E MÉTODOS

4.1. As amostras

Em 2001, Olson e seus colaboradores publicaram um trabalho onde dividiram a Terra em ecorregiões denominadas biomas. Essas ecorregiões foram delimitadas com base em mapas prévios de vegetação, plantas e animais, fatores bióticos, entre outros. Essa classificação foi definida como clássica e contempla 14 biomas terrestres distintos (Figura 3).

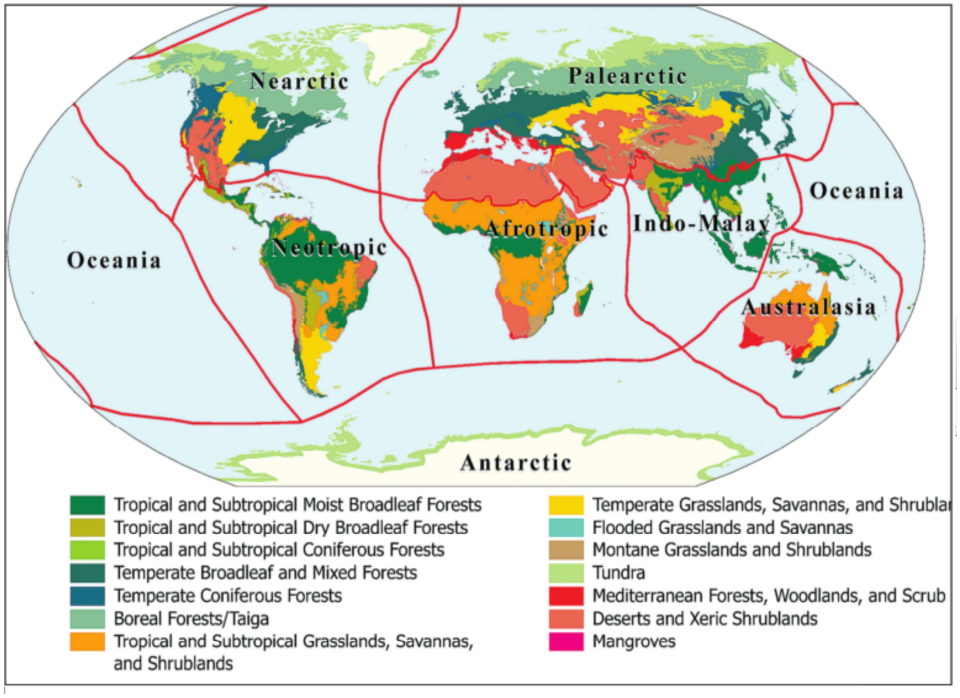


Figura 3. Biomas terrestres baseados na classificação de Olson (Olson et al., 2001).

Baseado nesta classificação, amostras provenientes de 11 dos 14 biomas de Olson foram utilizadas nesse trabalho. Trinta conjuntos de dados derivados do sequenciamento *shotgun* de metagenoma de solos foram selecionados representando esses 11 biomas: (1) floresta tropical (3 réplicas), (2) savana tropical (3 réplicas), (3) floresta temperada conífera (1 réplica), (4) floresta temperada decídua (1 réplica), (5) savana temperada (3 réplicas), (6) floresta mediterrânea (2 réplicas), (7) desertos (quentes e frios; 3 réplicas cada), (8) floresta tropical seca/semi-árido (3 réplicas), (9) tundra (3 réplicas), (10) floresta boreal (1 réplica) e (11) savana alagada/deserto salino (3 réplicas). Exceto as amostras de floresta tropical seca/semi-árido (8), todas as amostras são provenientes de banco de dados públicos MG-RAST (Meyer *et al.*, 2008) e do EBI (Mitchell *et al.*, 2016).

Amostras de floresta boreal, estepe temperado (prairie), floresta temperada conífera, floresta temperada decídua, e desertos quentes e frios foram derivadas do trabalho de Fierer e colaboradores (Noah Fierer *et al.*, 2012). Amostras de solo da Amazônia foram utilizadas para representar o bioma de floresta tropical (Mendes *et al.*, 2015). Amostras de solo do bioma da tundra foram selecionadas do trabalho de Hultman e colegas (Hultman *et al.*, 2015). Já as amostras de estepe temperado (EUA) utilizadas foram derivadas do trabalho de Cline e Zak (Cline e Zak, 2015). Amostra do solo de Cerrado brasileiro foram utilizadas para representar o bioma da savana temperada (Souza *et al.*, 2016). O bioma Mediterrâneo foi representado por amostras de solo da Austrália do projeto "Biomes of Australian Soil Environments database" (Bissett *et al.*, 2016). Amostras do deserto salino sazonal (savana alagada) localizado em Kutch Rann na Índia foram também selecionadas (Pandit *et al.*, 2014). E por último, amostras de solo da Caatinga, que fazem parte de um trabalho de nosso laboratório, foram utilizadas nesse trabalho. Essas amostras foram coletadas em 2014, em Petrolina, em uma reserva deste bioma na Estação experimental da Embrapa semi-árido. O mapa mostrando o local onde cada amostra foi coletada, assim como número de réplicas e o bioma em que se encontra pode ser observado na Figura 4. Maiores informações sobre as amostras como ID nos bancos de dados, local exato da coleta, ano da coleta e número de sequências estão disponíveis na Tabela 1.

Figura 4. Mapa visual do local onde as amostras foram coletadas, qual bioma essa representa (em parênteses) e o número de réplicas representado pelo número de marcadores na figura.

Tabela 1. Informações sobre as amostras utilizadas nesse trabalho.

Bioma	Banco de dados	ID	Plataforma	Latitude	Longitude	Ano da coleta	Localização
Floresta tropical	MG-RAST	4497389.3	Illumina	-11.68	-55.83	2009	Floresta Amazônica, Brazil
Floresta tropical	MG-RAST	4497390.3	Illumina	-11.68	-55.83	2009	Floresta Amazônica, Brazil
Floresta tropical	MG-RAST	4497397.3	Illumina	-11.68	-55.83	2009	Floresta Amazônica, Brazil
Deserto quente	MG-RAST	4477805.3	Illumina	34.9	-115.65	2010	Deserto Mojave, EUA
Deserto quente	MG-RAST	4477872.3	Illumina	35.38	-105.93	2010	Deserto Chihuahuan, EUA
Deserto quente	MG-RAST	4477873.3	Illumina	34.33	-106.73	2010	Deserto Chihuahuan, EUA
Boreal	MG-RAST	4477876.3	Illumina	64.8	-148.25	2010	Bonanza Creek LTER, Alasca
Floresta temperada decídua	MG-RAST	4477877.3	Illumina	34.61	-81.66	2010	Floresta experimental Calhoun, EUA
Floresta temperada conífera	MG-RAST	4477899.3	Illumina	35.96	-79.08	2010	Floresta Duke, EUA
Prairie	MG-RAST	4477804.3	Illumina	39.1	-96.6	2010	Konza prairie LTER, EUA
Tundra	MG-RAST	4470253.3	Illumina	64.67	-157.83	2009	Estação experimental de turfeiras do Alasca, Alasca
Tundra	MG-RAST	4470255.3	Illumina	64.67	-157.83	2009	Estação experimental de turfeiras do Alasca, Alasca
Tundra	MG-RAST	4470378.3	Illumina	64.67	-157.83	2009	Estação experimental de turfeiras do Alasca, Alasca
Deserto polar	MG-RAST	4477900.3	Illumina	-78.03	163.87	2010	Vale de Garwood, Antártica
Deserto polar	MG-RAST	4477901.3	Illumina	-77.73	162.31	2010	Vale do lago Bonney, Antártica
Deserto polar	MG-RAST	4477902.3	Illumina	-77.61	163.25	2010	Vale do lago Fryxell, Antártica
Cerrado	MG-RAST	4578924.3	Ion torrent	-15.60	-47.74	2014	Estação experimental Embrapa cerrados, Brasil
Cerrado	MG-RAST	4578925.3	Ion torrent	-15.60	-47.74	2014	Estação experimental Embrapa cerrados, Brasil
Cerrado	MG-RAST	4577669.3	Ion torrent	-15.60	-47.74	2014	Estação experimental Embrapa cerrados, Brasil
Caatinga	MG-RAST	4661239.3	Illumina	-09.03	-40.18	2014	Estação experimental Embrapa semi-áridos, Brasil
Caatinga	MG-RAST	4661237.3	Illumina	-09.03	-40.18	2014	Estação experimental Embrapa semi-áridos, Brasil
Caatinga	MG-RAST	4661235.3	Illumina	-09.03	-40.18	2014	Estação experimental Embrapa semi-áridos, Brasil
Deserto salino	MG-RAST	4543020.3	Ion torrent	23.94	70.18	2013	Kutch Rann, Índia
Deserto salino	MG-RAST	4543022.3	Ion torrent	23.94	70.18	2013	Kutch Rann, Índia
Deserto salino	MG-RAST	4543022.3	Ion torrent	23.94	70.18	2013	Kutch Rann, Índia
Estepe temperado	MG-RAST	4541649.3	Illumina	45.40	-93.20	2013	Cedar Creek Ecosystem Science Reserve, EUA
Estepe temperado	MG-RAST	4541650.3	Illumina	45.40	-93.20	2013	Cedar Creek Ecosystem Science Reserve, EUA
Estepe temperado	MG-RAST	4541651.3	Illumina	45.40	-93.20	2013	Cedar Creek Ecosystem Science Reserve, EUA
Mediterrâneo	EBI	ERR671917	Illumina	-35.12	141.99	2012	Walpeup, Austrália
Mediterrâneo	EBI	ERR671916	Illumina	-35.12	141.99	2012	Walpeup, Austrália

4.2.Tratamento de qualidade

Primeiramente, todos os dados brutos do sequenciamento foram extraídos dos seus respectivos bancos de dados. Mesmo para as amostras provenientes do banco de dados do MG-RAST, os dados brutos foram submetidos ou resubmetidos na plataforma online do MG-RAST para o tratamento de qualidade. Essa resubmissão no banco de dados foi feita para uma melhor padronização das análises, dado que cada amostra foi anotada pelo MG-RAST em anos diferentes e sabemos que essa plataforma vem sofrendo constantes atualizações. O pipeline de tratamento de qualidade dessa plataforma inicia por um pré-processamento dos dados utilizando o software SolexaQ (Cox, Peterson e Biggs, 2010). Posteriormente, duas etapas de remoção de *reads* artificiais são utilizadas: 1) *dereplication*, onde k-mers dos primeiros 20 pb são processados e utilizados para cálculo de erro de inferência dessas *reads* artificiais através do sofwate DRISEE (Keegan *et al.*, 2012); 2) os *reads* são alinhados contra o genoma humano para remover possíveis contaminações com DNA humano, utilizando o software bowtie (Langmead *et al.*, 2009). Essas informações foram extraídas do manual do MG-RAST (Wilke *et al.*, 2015). Após essas etapas de qualidade descritas, 135 Gb de dados processados foram utilizados para posteriores análises.

4.3. Normalização das amostras

Pelo fato de ser um trabalho que depende de muitas amostras de solos naturais, coletados em diversos biomas e utilizando dados de sequenciamento em larga escala, seria necessário bastante tempo e grandes volumes de recursos financeiros para coletar e sequenciar todas as amostras. Portanto, utilizamos dados já disponíveis provenientes de bancos de dados públicos. Entretanto, apesar das buscas em diversos bancos de dados por amostras que seguissem a mesma metodologia, alguns biomas são muito pouco estudados e poucos dados de sequenciamento *shotgun* estão disponíveis.

Das 13 amostras (30 contando com as réplicas), a maior parte foi sequenciada utilizando a plataforma Illumina, sendo que 2 amostras foram sequenciadas utilizando a plataforma Ion torrent. Em um trabalho comparando sequenciamento das mesmas amostras com tecnologias de sequenciamento distintas (Ion torrent, Illumina e Pacific bioscience), os autores concluíram que, para os organismos ricos em regiões GC, uma cobertura similar das sequências foi encontrada com as três tecnologias de sequenciamento analisadas. Entretanto, para genomas ricos em regiões AT, houve

uma cobertura mais baixa de sequenciamento utilizando a plataforma Ion Torrent (Quail *et al.*, 2012). Apesar desse pequeno viés apresentado pela plafatorma Ion Torrent, a maior parte das análises nesse trabalho foi realizada utilizando as amostras agrupadas de acordo com a vegetação (floresta, pastagens, desertos, etc), o que minimizaria os possíveis viéses em algumas amostras.

Além disso, outras medidas foram utilizadas para minimizar as diferenças entre as amostras. O tamanho médio de pares de base gerados entre as amostras variou de 100 a 250 pb, como também variou o número de réplicas e o número de sequências geradas pelo sequenciamento entre as amostras. Portanto, desenvolmemos algoritmos para a normalização desses dados utilizando a linguagem *perl*. O primeiro algoritmo normalizou as amostras (individuais – 30 amostras) em aproximadamente 1.600.000 sequências, as quais foram utilizadas para comparação individual de cada réplica. Ainda, as réplicas de cada bioma foram agrupadas em um único arquivo (13 amostras) e normalizadas em 3.750.000 sequências. Essas duas normalizações foram necessárias pois as amostras não possuem o mesmo número de réplicas. Para as análises entre biomas, utilizamos as réplicas agrupadas. Por outro lado, quando especificamos cada réplica, utilizamos a normalização em 1.600.000 (Tabela 2). Essa normalização foi realizada utilizando seleção aleatória de sequências, garantindo a não duplicação de *reads*. Um outro algoritmo foi desenvolvido para normalização dos *reads* com tamanho médio de 100 pb. Nesse *script*, foi realizado o cálculo do tamanho médio dos *reads* e posteriormente esses reads eram trimados até atingirem a média de 100 pb. Essa trimagem era feita através da extração de uma subsequência de 100 pb a partir de uma posição inicial randômica dentro de cada *read*. Após a normalização, os resultados foram avaliados e concluimos que as normalizações foram eficazes para diminuir os vieses entre as amostras, o que pode ser observado na (Tabela 3). Os valores apresentados por amostra normalizada, independentemente da plataforma ou tamanho dos *reads*, seguiram um mesmo padrão, o que não acontecia antes da normalização dos mesmos. Apesar da perda de informação biológica que geralmente ocorre com a normalização, foi necessária a utilização desta estratégia para viabilizar a comparação entre dados gerados a partir de diferentes tecnologias de sequenciamento.

Tabela 2. Número de *reads* antes e após a normalização em cada amostra.

Amostra	Número de *reads* antes da normalização	Número de *reads* após a normalização
Floresta_tropical_1	15.155.716	1.625.000
Floresta_tropical_2	18.983.799	1.625.000
Floresta_tropical_3	16.852.525	1.625.000
Boreal	6.271.219	1.625.000
F_temp_conífera	3.769.328	1.625.000
F_temp_decídua	6.061.283	1.625.000
Prairie	5.118.954	1.625.000
Estepe_temperado_1	58.786.973	1.625.000
Estepe_temperado_2	49.341.941	1.625.000
Estepe_temperado_3	57.821.781	1.625.000
Cerrado1	4.467.532	1.625.000
Cerrado2	4.763.523	1.625.000
Cerrado3	4.938.000	1.625.000
Caatinga1	36.023.156	1.625.000
Caatinga2	36.263.486	1.625.000
Caatinga3	35.028.639	1.625.000
Mediterrâneo_1	88.566.066	1.625.000
Mediterrâneo_2	88.660.538	1.625.000
Deserto_quente_1	5.635.708	1.625.000
Deserto_quente_2	6.540.487	1.625.000
Deserto_quente_3	10.657.707	1.625.000
Deserto_polar_1	7.472.298	1.625.000
Deserto_polar_2	4.590.211	1.625.000
Deserto_polar_3	9.078.616	1.625.000
Deserto_salino_1	1.952.467	1.625.000
Deserto_salino_2	2.038.597	1.625.000
Deserto_salino_3	1.625.785	1.625.000
Tundra_1	15.022.182	1.625.000
Tundra_2	18.806.543	1.625.000
Tundra_3	36.348.248	1.625.000

Tabela 3. Informações sobre a normalização nos grupos de biomas.

Bioma	Número de *reads* antes da normalização	Número de *reads* após a normalização	Número de *orfs* preditas	Número de CAZymes identificadas	Número de ARGs identificadas	Número de HSPs identificadas
Floresta tropical	50.992.040	3.750.000	3.207.346	32.893	240	2.096
Boreal	6.271.219	3.750.000	3.345.395	35.678	332	2.633
Floresta temperada conífera	3.769.328	3.750.000	3.339.110	37.254	366	2.730
Floresta temperada decídua	6.061.283	3.750.000	3.342.024	37.036	446	2.586
Prairie	5.118.954	3.750.000	3.319.927	40.430	268	2.963
Estepe temperado (100 pb)	165.950.695	3.750.000	3.535.557	44.352	326	3.045
Estepe temperado (150 pb)*	165.950.695	3.750.000	3.615.700	80.690	-	-
Cerrado (100 pb)	14.169.055	3.750.000	3.320.697	37.871	255	2.303
Cerrado (250 pb)*	14.169.055	3.750.000	3.534.283	103.315	-	-
Caatinga	107.315.281	3.750.000	3.337.131	43.774	128	3.150
Mediterrâneo (100 pb)	177.226.604	3.750.000	3.461.398	40.331	134	3.530
Mediterrâneo (150 pb)*	177.226.604	3.750.000	3.542.721	72.569	-	-
Deserto quente	22.833.902	3.750.000	3.193.607	35.697	179	3.236
Deserto polar	21.141.125	3.750.000	3.236.202	33.225	96	3.497
Deserto salino (100 pb)	5.616.849	3.750.000	2.935.002	32.295	30	3.981
Deserto salino (150 pb)*	5.616.849	3.750.000	2.979.357	43.679	-	-
Tundra	70.176.973	3.750.000	3.544.628	50.231	167	3.262

* Dados normalizados em 3.750.000 *reads*

4.4.Anotação taxonômica e funcional (SEED)

Os dados normalizados foram resubmetidos à plataforma do MG-RAST (http://metagenomics.anl.gov) para nova anotação. O banco de dados do SEED foi utilizado para anotação taxonômica e, juntamente com os bancos de dados do COG (Clusters of Orthologous Groups) e KEGG (Kyoto Encyclopedia of Genes and Genomes), para a anotação funcional das sequências. Escolhemos o SEED como principal banco de dados pelo fato de ter sido desenvolvido principalmente para anotação de dados metagenômicos e estar subdividido em subsistemas (3 níveis), de acordo com o papel funcional e via metabólica na qual os genes anotados estão envolvidos. As sequências da anotação taxonômica e funcional foram anotadas utilizando o BlastX com parâmetro de *best hit* com 50 pb como menor tamanho do alinhamento e $E<1x10^{-5}$ como valor de corte para o E-value. O gráfico de dispersão (Scatter plot), comparando par a par os grupos de biomas, foi gerado utilizando a estatística de *teste t* de *Welch*. As análises estatísticas foram realizadas utilizando o software STAMP - *Statistical Analysis of Metagenomic Profiles* (Parks *et al.*, 2014).

Para a análise de agrupamento hierárquico (UPGMA), a matriz de similaridade de *Bray-Curtis* foi calculada utilizando a anotação dos subsistemas do banco de dados SEED no software PAST - *Paleontological statistics software package* (Hammer, Harper e Ryan, 2001). Já a análise de correlação de *Spearman* foi realizada por comparação par a par entre as amostras dentro de cada grupo de bioma. O cálculo de correlação par a par por *Spearman* e a média e o desvio padrão foram calculados usando o ambiente R.

A rede de co-ocorrência (bipartida) foi gerada pelo software CoNET (versão 1.1.0 beta) (Faust *et al.*, 2012) utilizando a correlação de Spearman $\rho>0.8$, Pearson $\rho>0.8$ e Fisher's Z com *p*-value de corte < 0.05. Ainda, para reduzir possíveis falso-positivos, um teste múltiplo de ajustamento utilizando o método de Bonferroni foi aplicado.

4.5. Anotação funcional de enzimas ativas de carboidratos

Baseado nas análises estatísticas da anotação funcional a partir do banco de dados do SEED, a via de metabolismo central de carboidratos mostrou se diferenciar substancialmente entre os grupos de biomas estudados. Entender o perfil de enzimas de degradação de carboidratos pode fornecer evidências sobre quais carboidratos podem estar disponíveis em cada grupo de biomas.

Portanto, neste trabalho, utilizamos o perfil de HMM (*Hidden Markov models*) das famílias do CAZy para anotação das sequências. Esse perfil está disponível na base de dados do dbCAN (Yin *et al.*, 2012), que é um banco de dados para anotação automática das CAZymes e de mais três módulos de celulossomas: doquerina, coesina e SLH (homologia na camada S).

Primeiramente, o quadro aberto de leitura (*Open Reading Frame - ORF*) das sequências provenientes dos dados normalizados de cada bioma foi predito e convertido de sequências de nucleotídeos em aminoácidos utilizando o programa FragGeneScan 1.30 (Rho, Tang e Ye, 2010). Posteriormente, utilizamos o software HMMER3.1b2 (Finn, Clements e Eddy, 2011) para comparar os perfis *HMM* provenientes do banco de dados do dbCAN com as sequências de aminoácidos dos conjuntos de dados deste estudo, utilizando o valor de corte de 1e-5 (sequência completa). Esses perfis foram baixados e anotados localmente para os conjuntos de dados de cada bioma em separado.

A análise estatística foi realizada através do pacote do STAMP - Statistical Analysis of Metagenomic Profiles (Parks *et al.*, 2014), utilizando as anotações dos conjuntos de dados separados por biomas, sendo o metadados utilizados para agrupar os dados em grupos de biomas. A estatística ANOVA e um teste *post hoc* (Tukey-Kramer) foram aplicados para comparações múltiplas de grupos de biomas (p <0.05). A partir dos resultados encontrados na comparação múltipla dos conjuntos de dados, foram selecionadas todas as CAZymes estatisticamente significantes em cada grupo de biomas e estas foram analisadas manualmente a partir de referências bibliográficas disponíveis na literatura. A partir dessa triagem manual, as principais vias e/ou metabolismos foram selecionados e detalhados nesse trabalho.

Para a anotação taxonômica de algumas das sequências identificadas como possíveis CAZymes, foi realizado um Blastp (Altschul *et al.*, 1997) contra o banco de dados NR do NCBI (e-value 1e-5) e posteriormente o arquivo de saída do programa foi utilizado como entrada para anotação taxonômica utilizando o algoritmo LCA (*Lower Common Ancestor*) pelo programa MEGAN5 (Huson, Mitra e Ruscheweyh, 2011).

A rede de co-ocorrência (bipartida) foi gerada pelo software CoNET (versão 1.1.0 beta) (Faust *et al.*, 2012) utilizando a correlação de Spearman ρ>0.9, Pearson ρ>0.9 e Fisher's Z com p-value de corte < 0.05. Ainda para reduzir possíveis falso-positivos, um teste múltiplo de

ajustamento utilizando o método de Bonferroni foi aplicado.

4.6. Anotação funcional de genes de resistência a antibióticos

Assim como no caso do metabolismo central de carboidratos, a análise estatística da anotação funcional a partir do banco de dados do SEED revelou uma diferença estatisticamente significativa na abundância dos genes relacionados à resistência a antibióticos entre os grupos de biomas estudados. Apesar da anotação funcional do MG-RAST pelo banco de dados do SEED já fornecer informações sobre os genes relacionados à resistência a antibióticos (ARGs) pelo subsistema "Resistance to antibiotics and toxic compounds", utilizamos um banco de dados específico para identificar e obter mais informações sobre os genes de resistência a antibióticos.

As sequências de aminoácidos, com *ORFs* preditas e traduzidas no passo anterior via FragGeneScan, dos conjuntos de dados de biomas normalizados foram utilizadas nesse análise. Essas sequências de aminoácidos foram preditas como possíveis genes de resistência a antibióticos utilizando o programa BlastP (Altschul *et al.*, 1997; Edgar, 2010). Essa predição foi feita alinhando as nossas sequências contra sequências disponíveis no banco de dados do ARDB - *Antibiotic Resistance Genes Database* (Liu e Pop, 2009), utilizando um valor de corte (e-value) de 1e-5. Esse banco de dados é um banco específico para anotação de ARGs, sendo ele subdividido em 380 tipos de resistências agrupados em 95 diferentes classes.

Após as sequências serem preditas via BlastP, um outro parâmetro foi utilizado para filtar os ARGs identificados. Selecionamos apenas sequências que exibiram identidade maior ou igual a 80% no alinhamento contra as sequências do banco de dados do ARDB e um tamanho mínimo de 25 aminoácidos, resultando em 2.775 possíveis ARGs de um total de 74.986 anteriormente identificados. Essa filtragem foi realizada utilizando comandos *awk* do linux. Além disso, para anotação das classes relativas a cada ARG e formatação do arquivo para um formato de entrada para a criação da rede, foram desenvolvidos scripts na linguagem *perl*.

Na primeira análise dos dados, foi gerada uma rede bipartida de correlação entre os biomas e os tipos de genes de resistência a antibióticos encontrados em cada um deles, mostrando a abundância (através da espessura da aresta) e a qual grupo de bioma aquele bioma pertente (através das cores das arestas). Essa rede foi gerada utilizando o software Cytoscape 3.3.0 (Shannon, 2003). As sequências identificadas como possíveis ARGs (com filtros) foram anotadas taxonomicamente utilizando o banco de dados NR do NCBI e o programa BlastX (a partir das

sequências de nucleotídeos). A anotação final foi obtida com o software MEGAN5 (Huson, Mitra e Ruscheweyh, 2011) através do algoritmo LCA (*Lower Common Ancestor*) utilizando os parâmetros *default* (*maximum number of matches per read: 5, min support: 5, min score: 35*). Posteriormente, a anotação taxonômica de cada sequência foi correlacionada ao bioma e tipo, classe e resistência a antibióticos correspondente. Essa tabela contendo todas as informações cruzadas foi construída através de um *script* desenvolvido *in-house* na linguagem *perl*. Essa informações são bastante interessantes, pois permitiram relacionar quais grupos taxonômicos carreiam os genes de resistência e em quais biomas, além da informação sobre quais tipos de resistência são mais encontrados em determinado bioma ou grupo de bioma

Para avaliar a co-ocorrência de classes de ARGs e as taxonomias às quais estão relacionados, uma rede de co-ocorrência foi gerada a partir de uma análise realizada pelo software CoNET (versão 1.1.0 beta) (Faust *et al.*, 2012) utilizando a correlação de Spearman ρ>0.8, Pearson ρ>0.9 e Fisher's Z com *p*-value de corte < 0.05. Ainda, para reduzir possíveis falso-positivos, um teste múltiplo de ajustamento utilizando o método de Bonferroni foi aplicado.

O agrupamento hierárquico (UPGMA) derivado de uma matriz de similaridade de Bray-Curtis foi calculado a partir da abundância de ARGs em cada conjunto de dados de biomas (coeficiente de correlação 0.92), usando o software PAST. Para gerar o gráfico ternário, os grupos de biomas de semi-árido e de deserto foram agrupados em um único grupo, chamado aqui de semi-árido/deserto. Nesta análise não foram mostrados os resultados do bioma da tundra e do deserto salino, que também foi excluído pois, baseado no resultado mostrado no UPGMA, apresenta resultados bem distintos dos biomas semi-árido e deserto para a anotação de ARGs, sendo um *outgroup* nesta análise. O gráfico ternário foi realizado com base em valores médios de anotação taxonômica das sequências de ARGs e da abundância das ARGs em cada grupo de bioma (floresta, pastagem e semiárido/deserto), gerados através do software JMP® Statistical Discovery Software (SAS Institute Inc, sem data).

4.7.Anotação funcional de proteínas de choque térmico

Para a anotação dos possíveis genes que codificam as proteínas de choque térmico, foi utilizado o banco de dados HSPIR - *Heat Shock Protein Information Resource* (Ratheesh Kumar *et al.*, 2012). Esse banco de dados manualmente acurado é exclusivo para identificação de proteínas de choque térmico das 6 maiores famílias de HSP: Hsp100, 90, 70, 60, 40 e 20 (shsp) e, adicionalmente, uma família PAM16 (não utilizada neste trabalho). Os perfis HMM de todas as famílias de HSP disponíveis foram baixados e analizados em um servidor local. Os perfis foram comparados às sequências dos conjuntos de dados normalizados para cada bioma utilizando o software HMMER3.1b2 (Finn, Clements e Eddy, 2011) com valor de corte de 1e-5.

Para anotar taxonomicamente algumas sequências de HSP20 e HSP40 do conjunto de dados do deserto salino, foi realizado um BlastX (Edgar, 2010) das sequências de nucleotídeo anotadas para essas HSPs separadamente contra o banco de dados NR do NCBI. Posteriormente, esse resultado foi utilizado como arquivo de entrada para o software MEGAN5 (Huson, Mitra e Ruscheweyh, 2011) para uma anotação mais acurada utilizando o algoritmo LCA (*Lower Common Ancestor*) e os parâmetros *default* (*maximum number of matches per read: 5, min support: 5, min score: 35*).

5. RESULTADOS E DISCUSSÃO

Os resultados nesse trabalho foram apresentados e discutidos de duas formas: em algumas análises a comparação foi feita entre os biomas já descritos, ao passo que em outras análises a comparação foi feita entre os chamados grupos de biomas. Esses grupos são formados por biomas que possuem uma mesma característica de vegetação: (a) Florestas, formadas pelos biomas floresta tropical, floresta boreal, floresta temperada conífera e floresta temperada decídua (representados pela cor verde); (b) Pastagens, formada pelos biomas *prairie*, estepes temperados e Cerrado (representados na cor laranja); (c) Semi-áridos, formado pela Caatinga e pelo Mediterrâneo (representados pela cor azul); (d) Desertos, quente, frio e salino (representados pela cor vermelha); e (e) Tundra, na cor azul ciano (Figura 5).

Figura 5. Clusterização dos grupos de biomas a partir das características da vegetação apresentada (*Prairie*, sem data; The nature conservancy, sem data; GoKutch, sem data; Louro, sem data; Silvey, 2009; The Osprey Brand Team, 2009; Eisner, 2011; Plantier, 2013; Melo, 2014; Resolute forest products, 2014; Belezas Naturais, 2015; Cruz, 2015; Ferrari, 2015).

5.1. Diversidade taxonômica das comunidades microbianas dos solos

Primeiramente, as abundâncias relativas de fungos, bactérias e arquéias foram caracterizadas para todos os biomas. As amostras de deserto salino exibiram a maior abundância de arquéias. Com relação aos fungos, resultados estatisticamente significantes mostraram que solos de florestas e de pastagens têm uma maior proporção do filo Ascomycota que solos de semi-áridos e desertos. Por outro lado, solos de semi-áridos possuem uma maior abundância de arquéias que solos de florestas e pastagens (Figura 6).

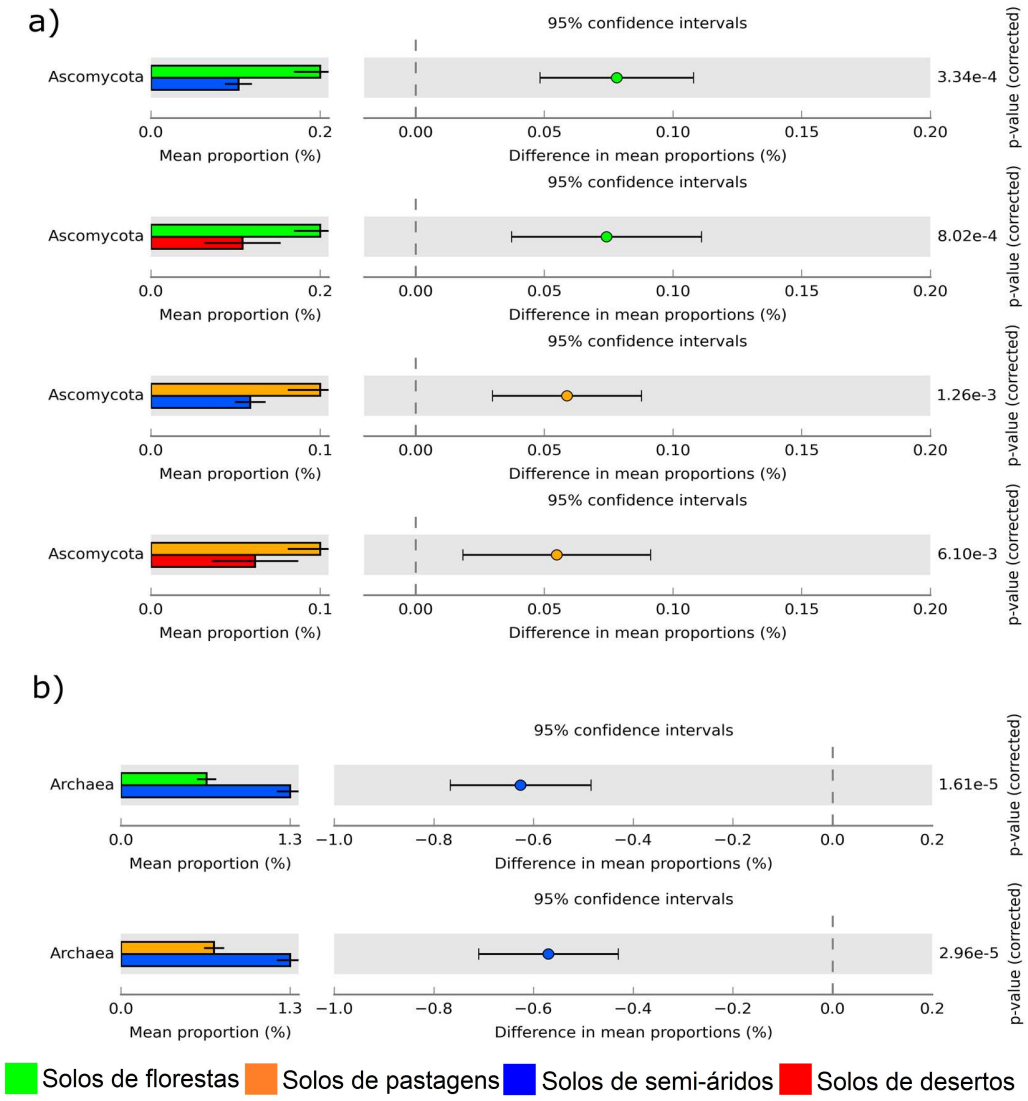


Figura 6. Resultados estatisticamente significantes da comparação da abundância relativa de fungos e arquéias entre grupos de biomas (Welch´s t-test; p-value<0.05).

A análise taxonômica das bactérias revelou uma maior proporção dos filos Proteobacteria, Verrucomicrobia e Acidobacteria em florestas, pastagens e tundra quando comparadas aos outros grupos de biomas. O filo Chlamydiae foi encontrado como mais abundante em florestas e pastagens quando comparado a solos de desertos e de semi-áridos. Por outro lado, Chloroflexi, Firmicutes e Deinococcus-Thermus estão significantemente mais abundantes em solos semi-áridos quando comparados a solos de florestas, pastagens e desertos. Bacteroidetes, Fusobacteria, Dictyoglomi, Chlorobi, Spirochaetes, Deferribacteres, Synergistetes, Aquificae e Thermotogae foram significantemente mais abundantes em solos de desertos quando comparado aos outros grupos de biomas. Além disso, uma maior proporção de Tenericutes foi encontrado em solos de desertos e tundra em relação aos outros grupos de biomas. Cianobactéria apresentou abundância três vezes maior em solos de desertos quentes e frios quando comparado ao grupo de biomas florestais. Esse foi um resultado esperado, já que as cianobactérias desempenham um papel importante em ambientes oligotróficos áridos, fornecendo estabilidade ao solo, retenção da humidade e fertilidade (Makhalanyane *et al.*, 2015) (Figura 7 e Figura 8).

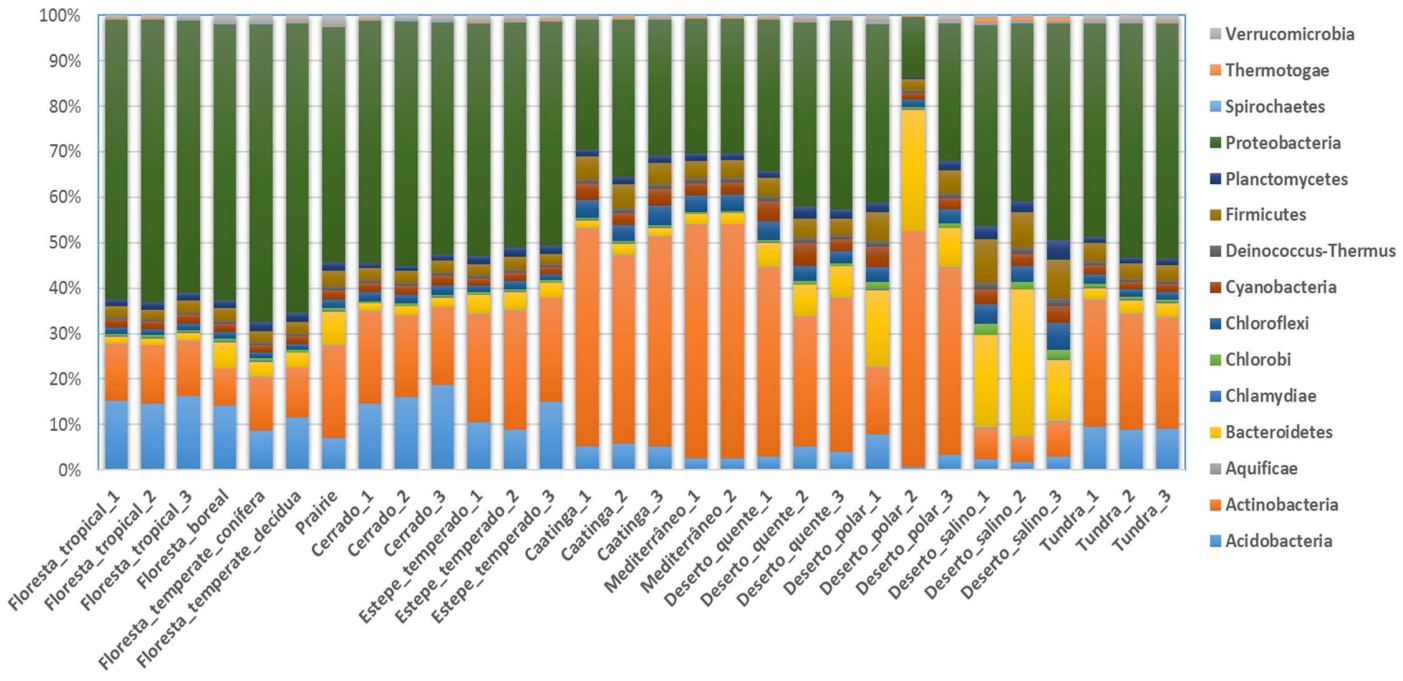


Figura 7. Proporção de cada taxonomia encontrada nas amostras para cada amostra baseada na anotação taxonômica utilizando o banco de dados do SEED (em nível de filo).

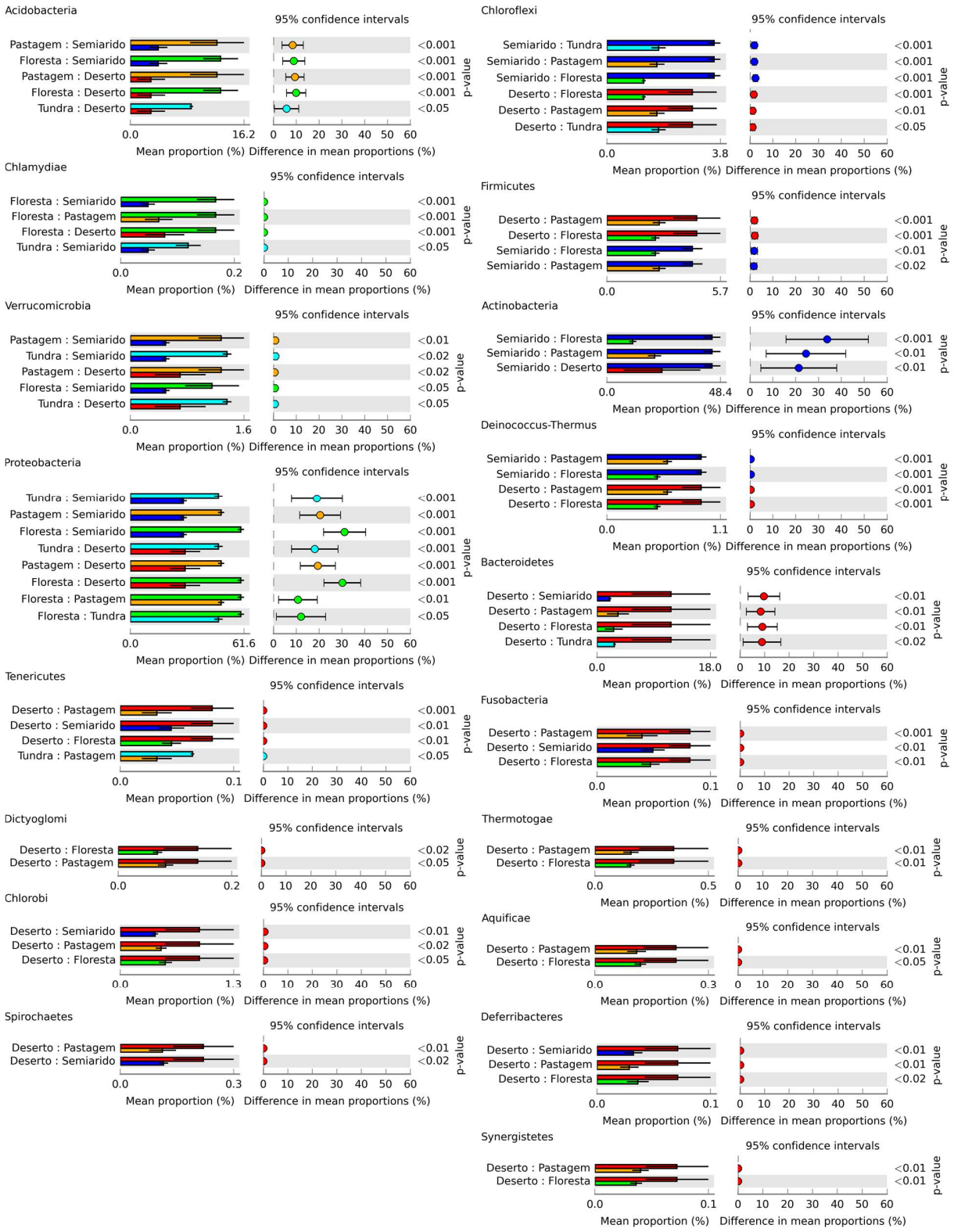


Figura 8. Comparação da abundância relativa dos filos de bactérias entre os grupos de biomas (ANOVA; Tukey-Kramer; p-value <0.05).

Na análise de componentes principais (PCA) da composição taxonômica das comunidades microbianas dos diferentes biomas podemos observar a correlação das amostras dentro dos grupos de biomas, principalmente no grupo de florestas, pastagens e semi-áridos (Figura 9). Amostras de solos de grupos de florestas, pastagens e tundra (regiões não desérticas) possuem uma maior abundância do filo Proteobacteria, seguido de Acidobacteria e Actinobacteria e, portanto, foram agrupadas mais próximas no PCA. As amostras de regiões semi-áridas e de deserto quente possuem uma maior abundância de sequências anotadas como Actinobacteria e desertos frios e salinos de sequencias de Bacteroidetes (Figura 7 e Figura 9). Essas análises corroboram com os agrupamentos encontrados no PCA. Inesperadamente, as amostras do bioma tundra agruparam junto as amostras de biomas de pastagens. As amostras de tundra foram coletadas durante o verão, quando gelo derrete e uma vegetação monocotiledônea surge (Virtanen *et al.*, 2015), portanto isto poderia explicar o fato de amostras da tundra agruparem com as amostras de pastagens. Os biomas da Caatinga e Mediterrâneo foram agrupadas bem próximas, o que é coerente com o fato de que ambas as regiões são semi-áridas com clima sazonal (verão seco e inverno chuvoso) e vegetação xerófila. As amostras de desertos não apresentaram uma clusterização tão próxima, uma vez que apresentam características extremófilas bem distintas. Entretanto, como o objetivo desse trabalho é encontrar padrões entre estes desertos, independentemente do tipo de estresse a que esses microrganismos estão submetidos, estas amostras foram agrupadas em grupos de desertos.

Avaliando apenas a composição bacteriana em nível de filos entre os diferentes biomas, também foi observado um forte padrão de clusterização baseado no agrupamento hierárquico (UPGMA – Bray-Curtis) de amostras de acordo com as características da vegetação apresentada nesses biomas (Figura 10).

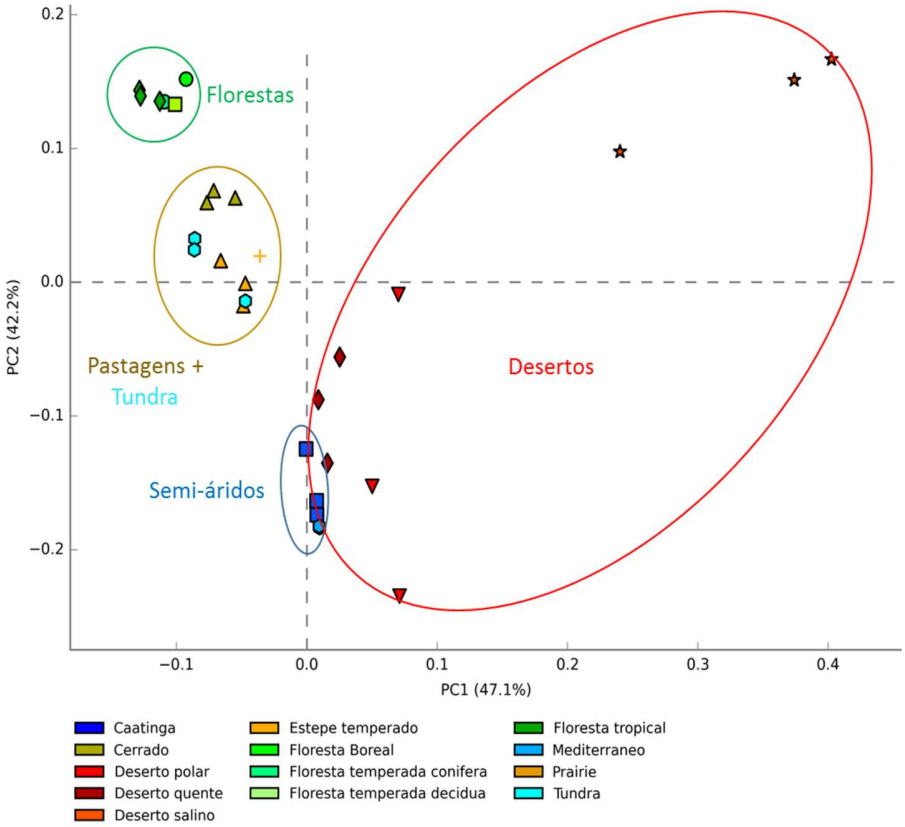


Figura 9. Análise de componentes principais da estrutura taxonômica das comunidades microbianas de todas as amostras.

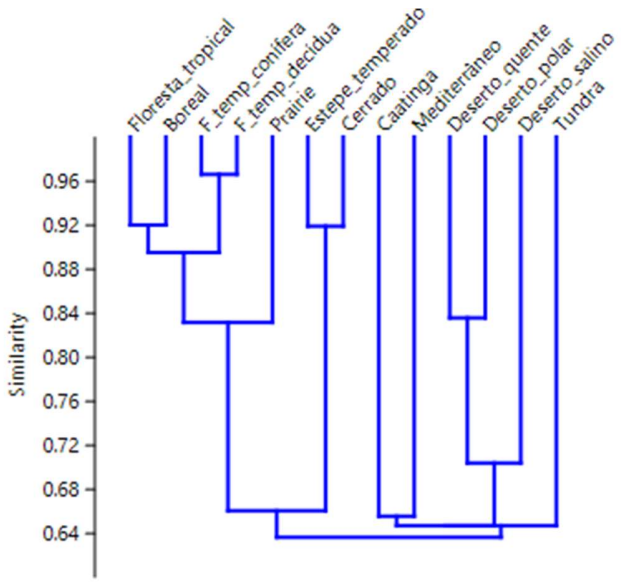


Figura 10. Agrupamento hierárquico (UPGMA) a partir das abundância dos filos (bactérias) nos diferentes biomas.

5.2. Diversidade funcional das comunidades microbianas de solos

Duas abordagens foram utilizadas na comparação do potencial funcional encontrado nas comunidades microbianas dos solos dos biomas analisados. Primeiramente, analisamos funcionalmente a correlação entre as amostras dentro de cada grupo de bioma, utilizando a anotação dada por diferentes bancos de dados. Posteriormente, assim como na análise taxonômica, uma análide de componentes principais (PCA) foi gerada a partir da anotação funcional através do banco de dados do SEED. Entretanto, ao invés de classificar as amostras por cada bioma, elas foram classificadas e coloridas por grupos de biomas.

Uma comparação par a par utilizando a correlação de Spearman foi calculada para quantificar a semelhança das amostras dentro de cada grupo de biomas definido, com base nos dados de anotação funcional dos bancos de dados SEED, COG e KEGG em nível de função. Podemos observar na Figura 11 que há uma maior correlação entre as amostras de florestas, pastagens e semiáridos, entretanto uma menor correlação é observada em solos desérticos. Esse resultado já era esperado, em função da variabilidade de estresses encontrada entre os solos desérticos amostrados, tais como temperaturas quentes e frias e hipersalinidade. As amostras de tundra não fizeram parte dessas análises por serem 3 réplicas de uma mesma amostra. Ainda, o número e a função das categorias diferem entre os bancos de dados, sendo o banco de dados SEED o que possui o maior número de categorias funcionais, seguido do banco de dados do KEGG e do COG. Essas diferentes categorizações das enzimas podem explicar os diferentes resultados obtidos entre os bancos de dados.

A análise de componentes principais com base em estatísticas de análise de variância (ANOVA) e no teste de Tukey Kramer da composição funcional da comunidade mostrou clara separação de grupos de biomas distintos (Figura 12). Assim como os resultados encontrados no PCA baseada na anotação taxonômica, é possível observar o agrupamento de amostras de acordo com as características da vegetação. Há também uma maior proximidade entre as amostras não desérticas (florestas, pastagens e tundra), quando comparadas às amostras desérticas (desertos e semi-áridos).

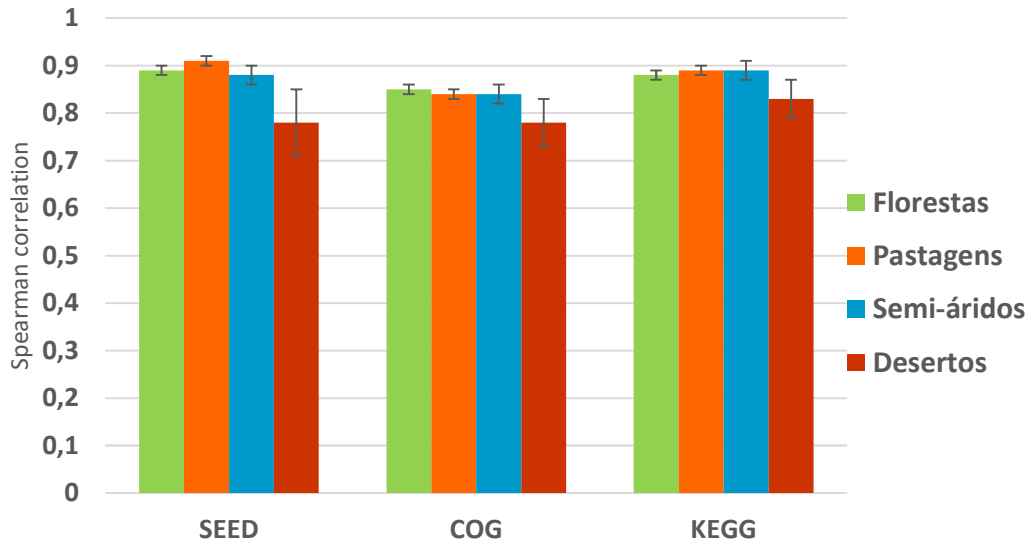


Figura 11. Correlação de Spearman entre as amostras de solos dentro de cada grupo de bioma baseada na anotação dos bancos de dados do SEED, COG E KEGG.

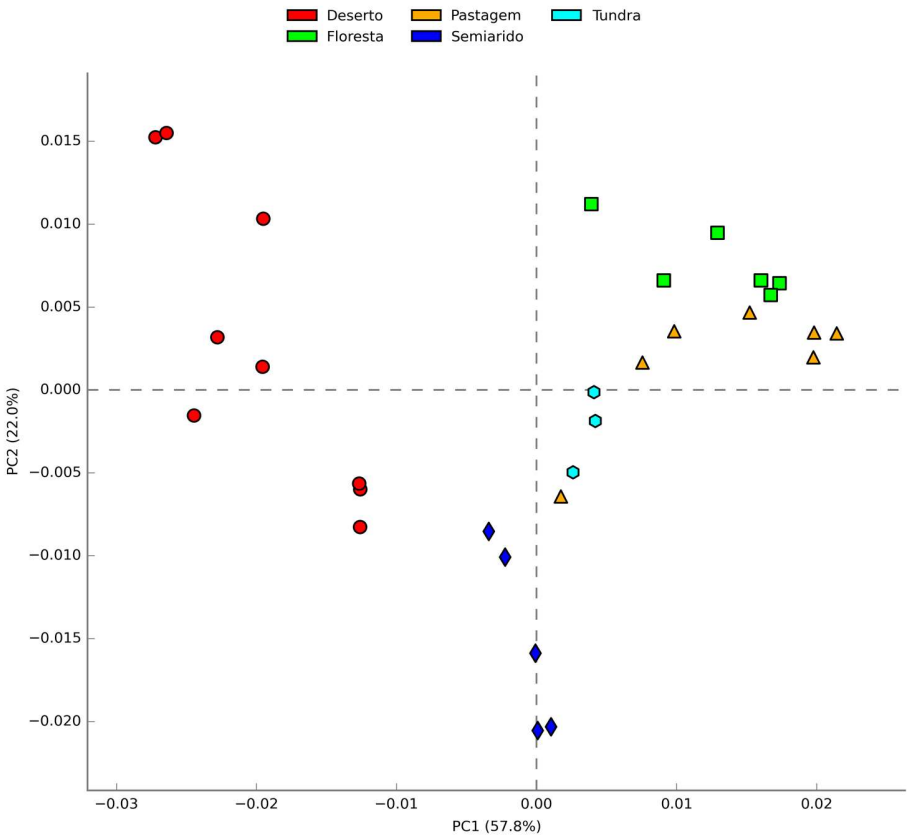


Figura 12. Análise de componentes principais baseada na anotação funcional dos dados metagenômicos dos grupos de biomas.

Com base nesses resultados, comparamos funcionalmente amostras de ambientes desérticos (neste caso, desertos e regiões semi-áridas) e não desérticos (florestas, pastagens e tundra) utilizando o banco de dados do SEED (subsistemas do nível 1). Podemos observar na Figura 13 que os subsistemas mais abundantes em solos não desérticos foram os envolvidos com metabolismo de compostos aromáticos, de potássio e enxofre (assimilação de enxofre orgânico), sinalização celular, transporte de membrana, motilidade, regulação e sinalização celular, e metabolismo secundário (como metabolismo de aminoácidos aromáticos), além de resistência de antibióticos (subsistema *Virulence, disease and defense*) e estresse oxidativo (*Stress response*). As amostras de solos desérticos apresentaram uma maior abundância de sequências anotadas para o metabolismo de proteínas, RNA e DNA (incluindo metabolismo de reparação de DNA), divisão e ciclo celular, etc. A abundância desses subsistemas demonstra claramente o quão adaptados os metabolismos dos microrganismos estão para sobreviver em cada bioma específico.

Ainda, um gráfico de dispersão foi utilizado para compararmos estatisticamente os metabolismos diferentes entre os grupos de biomas (par a par). Três grupos funcionais foram responsáveis pela diferenciação entre os grupos de biomas baseado na comparação dos subsistemas nível 2 do SEED: (1) a resistência aos antibióticos e compostos tóxicos, (2) a síntese proteica, e (3) o metabolismo central de carboidratos (Figura 14). Sequências relacionadas com a resistência aos antibióticos e compostos tóxicos apresentaram maior proporção em solos de florestas e de pastagens, corroborando os resultados observados em trabalho anterior em que amostras de solos não-desérticos foram comparados a solos desérticos (Noah Fierer *et al.*, 2012). Esta proporção mais alta ocorreu principalmente devido aos metabolismos de resistência ao cobalto, zinco e cádmio e módulos de bomba de efluxo de resistência a multidrogas (nível 3). Por outro lado, o subsistema de biossíntese de proteínas mostrou-se mais abundante em amostras de deserto e de regiões semi-áridas quando comparadas a amostras de solos não desérticas. Esse subsistema foi bastante representado pelo módulo de GTPase universal, nos quais *GTP binding protein Hflx*, *GTP binding and nucleic acid-binding protein YchF* e *Methionyl tRNA synthetase* foram os mais abundantes. A proteína YchF foi recentemente relacionada à regulação da resposta a estresses oxidativos em *E. coli* (Wenk *et al.*, 2012). A proteína Hflx foi relacionada ao processo da biogênese do ribossomo durante condições de estresses (Shields, Fischer e Wieden, 2009), tendo sido observada maior abundância dessa proteína durante estresses induzidos por radiação gama (Basu e Apte, 2012). Portanto, a maior a abundância dessas enzimas envolvidas na resistência a diferentes condições de

estresses podem representar fatores chaves na fisiologia e adaptação da microbiota em solos de biomas com condicões climáticas extremas.

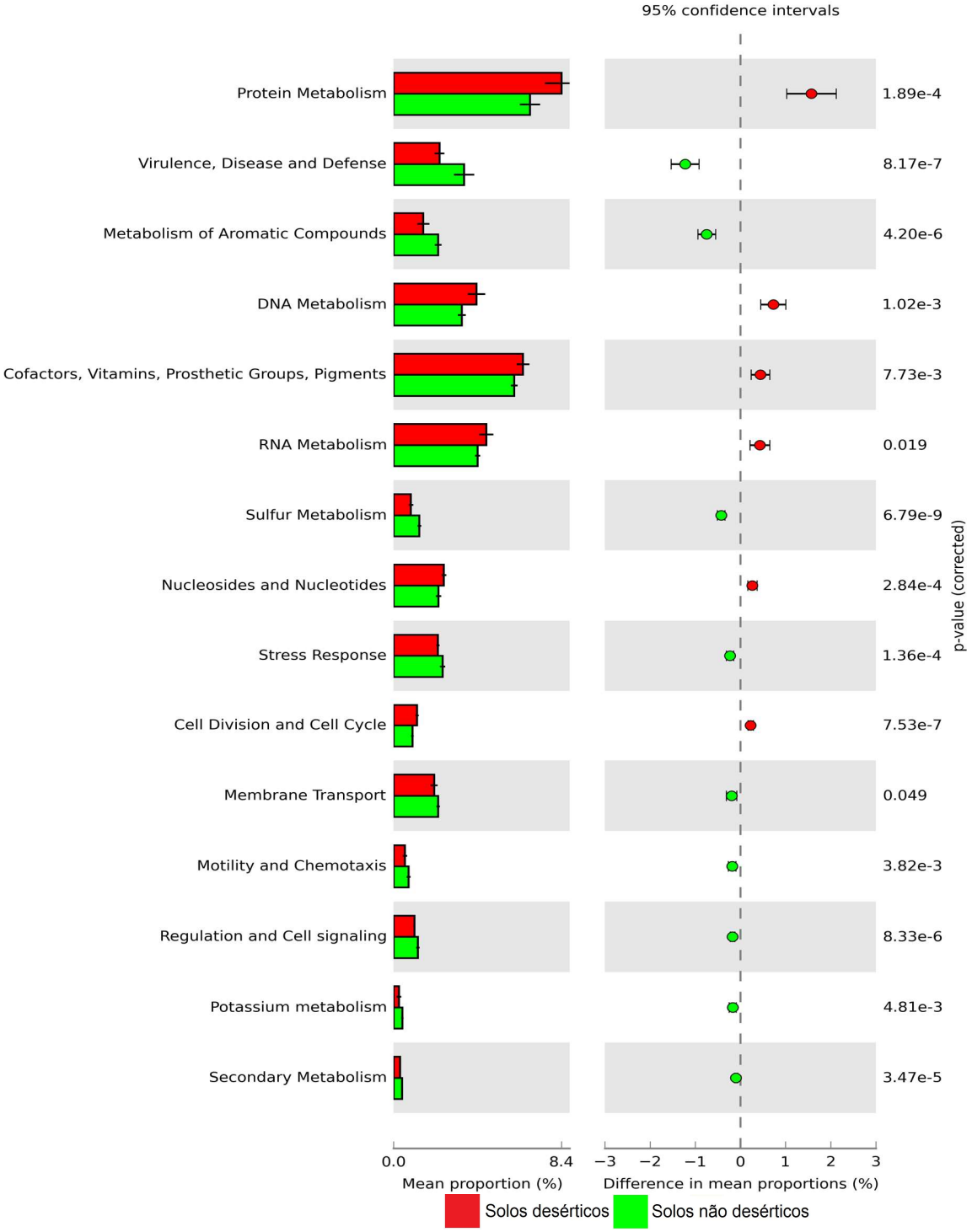


Figura 13. Metabolismos do nível 1 do banco de dados SEED contendo com abundâncias estatisticamente diferentes entre os grupos de biomas desérticos (desertos e semi-áridos) e não desérticos.

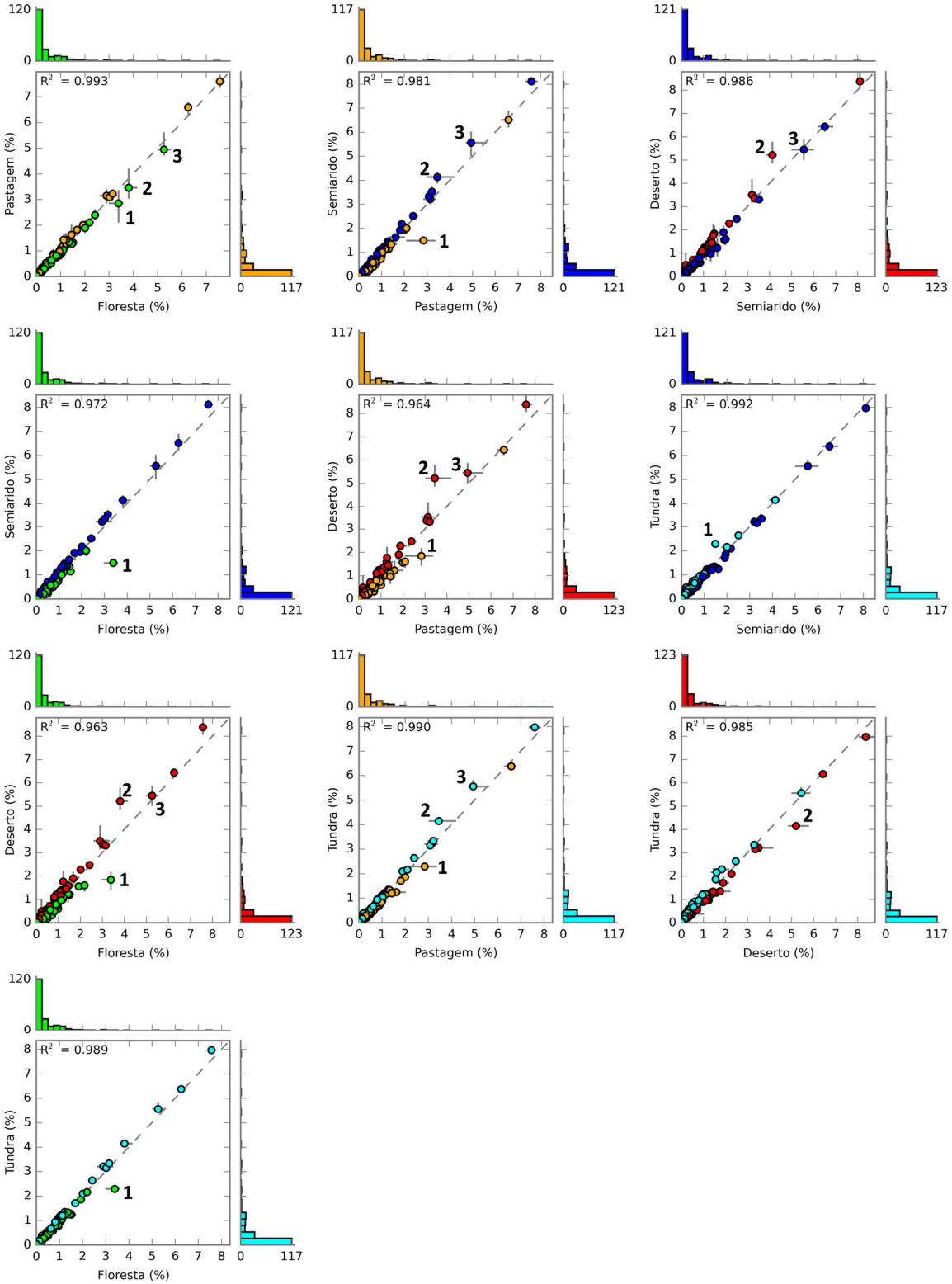


Figura 14. Gráfico de dispersão da comparação par a par dos grupos de biomas baseado no subsistema do SEED (nível 2). 1 - Resistência a antibióticos e compostos tóxicos, 2 - Biosíntese de proteínas e 3 - Metabolismo central de carboidratos.

A fim de identificar possíveis correlações entre grupos filogenéticos específicos (nível de classe) e subsistemas (nível 1) anotados funcionalmente pelo banco de dados do SEED, geramos um rede de co-ocorrência bipartida correlacionando ambas as anotações (Figura 15), sendo que os nós em azul representam os grupos filogenéticos e em vermelho os subsistemas do SEED. Podemos observar dois agrupamentos formados na rede: o primeiro envolvendo as classes Alfaproteobacteria, Betaproteobacteria e Solibacteres e os subsistemas metabolismo de compostos aromáticos e metabolismo de enxofre. Nesse grupo foram encontradas correlações de grupos taxonômicos e metabolismos mais abundantes em solos não desérticos (florestas e pastagens). Já o segundo agrupamento foi formado por grupos taxonômicos e metabolismos mais abundantes em solos de regiões semi-áridas e, principalmente, de solos desérticos. Portanto, essa rede corrobora, correlaciona e resume de forma visual os resultados encontrados anteriormente.

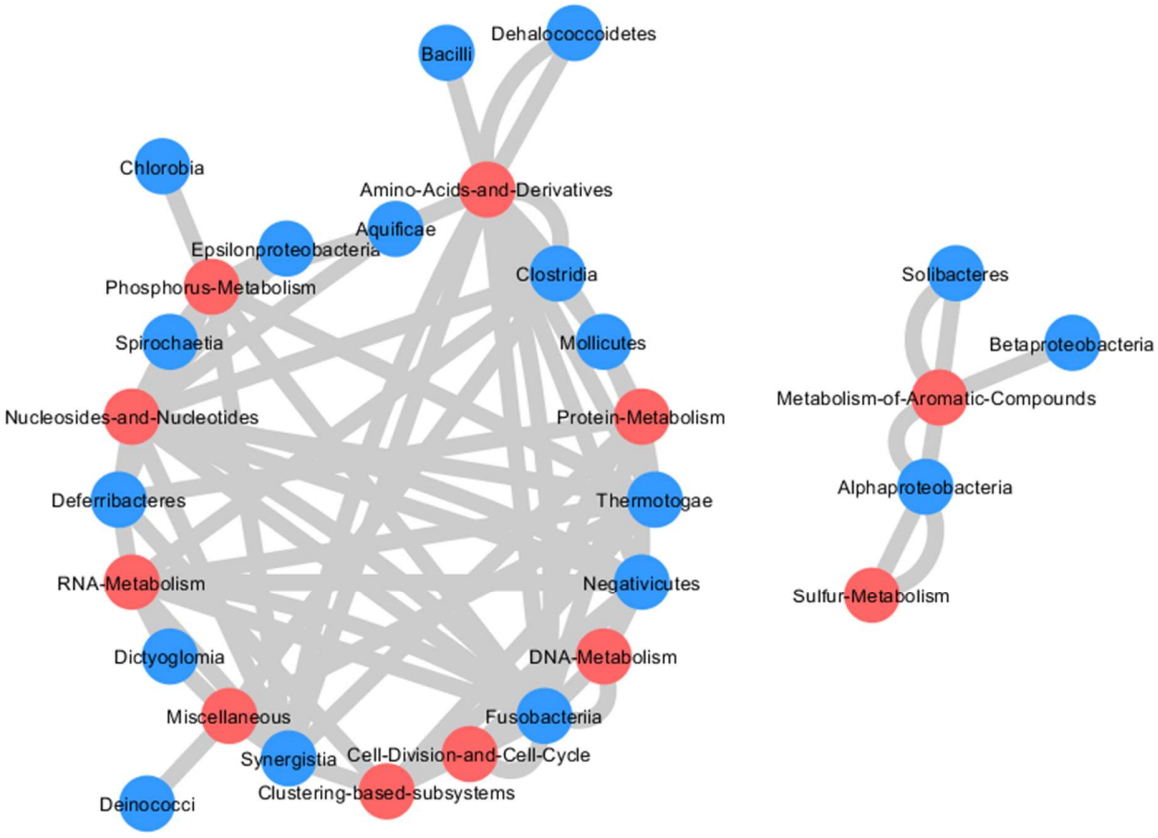


Figura 15. Rede bipartida de co-ocorrência entre a anotação taxômica e funcional de todas as amostras (nível 1 dos subsistemas SEED e nível taxonômico de classe).

Para melhor detalhar os subsistemas que se destacaram na diferenciação dos grupos de biomas (Figura 14), as sequências dos metagenomas foram anotadas em bancos de dados específicos. Na análise do subsistema de resistência aos antibióticos e os compostos tóxicos (1), utilizamos o banco de dados de genes de resistência a antibióticos ARDB (*Antibiotic Resistance gene DataBase*). Para a análise do subsistema de metabolismo central de carboidratos (3), utilizamos o banco de dados do CAZy (Carbohydrate-Active enZYmes database). Adicionalmente, utilizamos o banco de dados HSPIR (*Heat Shock Protein Information Resource*) para avaliar o perfil de proteínas envolvidas em diferentes estresses, conhecidas como proteínas de choque térmico.

5.3. Perfil funcional de enzimas relacionadas à degradação de carboidratos entre os grupos de biomas

Um total de 302.874 sequências dos metagenomas dos biomas, após normalização, alinharam contra 315 famílias do CAZy + celulossomas (*dbCAN*). A abundância relativa das classes de enzimas relacionadas à degradação de carboidratos segue um padrão similar entre os metagenomas (Figura 16).

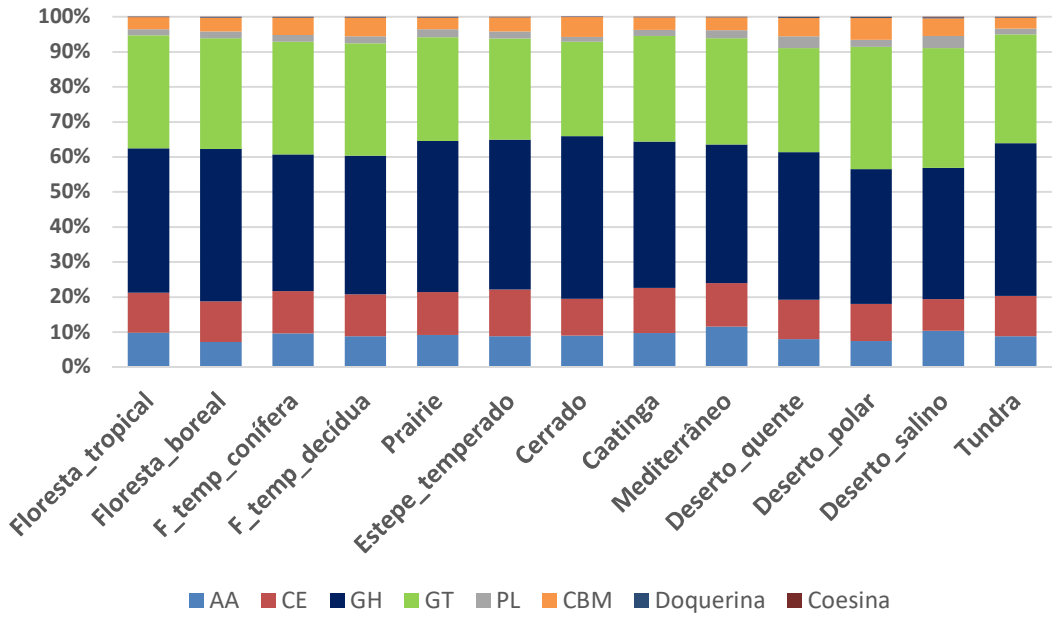


Figura 16. Abundância relativa das classes de CAZymes e celulossomas nos dados de metagenoma normalizados dos biomas.

Exceto para os biomas da tundra e do deserto salino, a GH13 foi a CAZyme mais abundante dentre as glicosil hidrolases nas amostras de todos os biomas (Figura 17). Esse resultado corrobora dados prévios de literatura, já que essa hidrolase foi encontrada como mais abundante em vários trabalhos envolvendo metagenômica de solos (Cardenas *et al.*, 2015; Manoharan *et al.*, 2015). Entretanto, essa hidrolase foi mais abundante em solos de biomas desérticos e semi-áridos em comparação com solos de florestas e de pastagens. Além disso, a GT4 foi a glicosiltransferase mais abundante no solo na maioria dos biomas, exceto na floresta tropical, e também foi encontrada mais abundantemente em solos de biomas de desertos e semi-áridos (Figura 18). Essa glicosiltransferase representa aproximadamente 29% das GTs anotadas em desertos e regiões semi-áridas e 16% em biomas de floresta e pastagens.

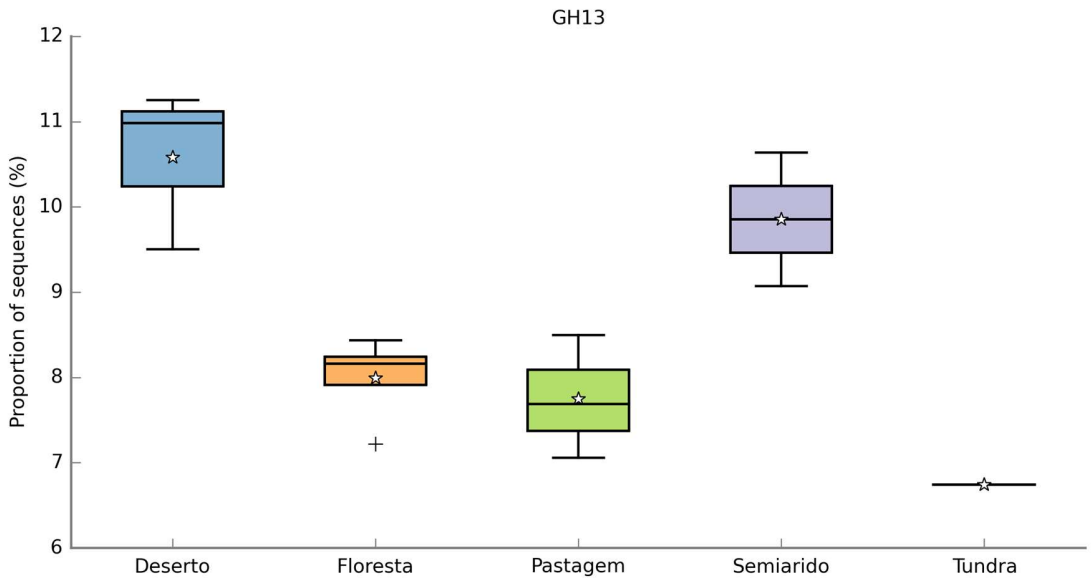


Figura 17. Box plot mostrando a maior abundância da GH13 em solos de deserto e semi-árido em relação as amostras de outros biomas.

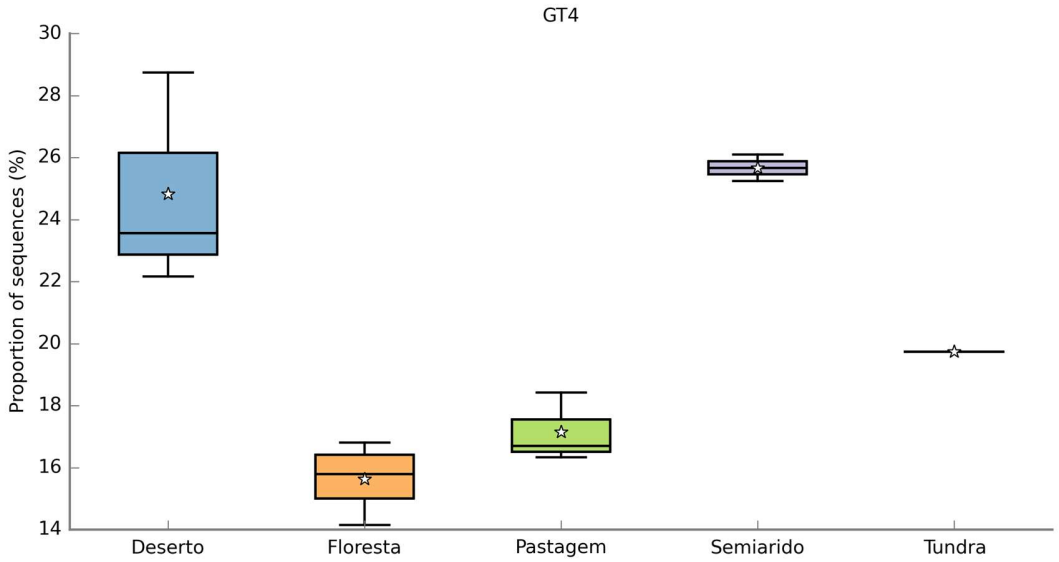


Figura 18. Box plot mostrando a maior abundância da GT4 em solos de deserto e semi-árido em relação as amostras de outros biomas**.**

GH13 e GT4 são famílias que incluem uma variedade de enzimas envolvidas no metabolismo da trealose e de sinalização, incluindo trealose sintase (EC: 3.2.1.93), trealose-6-fosfato-hidrolase (EC 3.2.1.93), trehalohydrolase malto-oligosyltrehalose (EC: 3.2.1.141) e de malto-sintase oligosyltrehalose (EC: 5.4.99.15). A grande abundância destas CAZymes em tais ambientes hostis pode ser explicado pelo fato de que a trealose é um dissacárido não redutor, estável em uma ampla variedade de pH (3,5-10), conhecida por desempenhar um importante papel no armazenamento de energia e a proteção celular sob muitas condições de estresse, incluindo o aquecimento, congelamento, seca e dessecação (Jiang *et al.*, 2013; Walmagh, Zhao e Desmet, 2015).

O *t*-teste Welch foi utilizado para selecionar as CAZymes com abundâncias significativamente diferentes nas comparações entre os grupos de biomas e separadas por: glicosil hidrolases (Figura 19), glicosiltransferases (Figura 20), esterases (Figura 21), liases (Figura 22) e enzimas auxiliares (Figura 23). Posteriormente, uma análise manual baseada em revisões bibliográficas foi realizada com as CAZymes encontradas. Com base nesses resultados, destacamos algumas CAZymes que podem elucidar os potenciais metabolismos dos carboidratos nesses grupos de biomas.

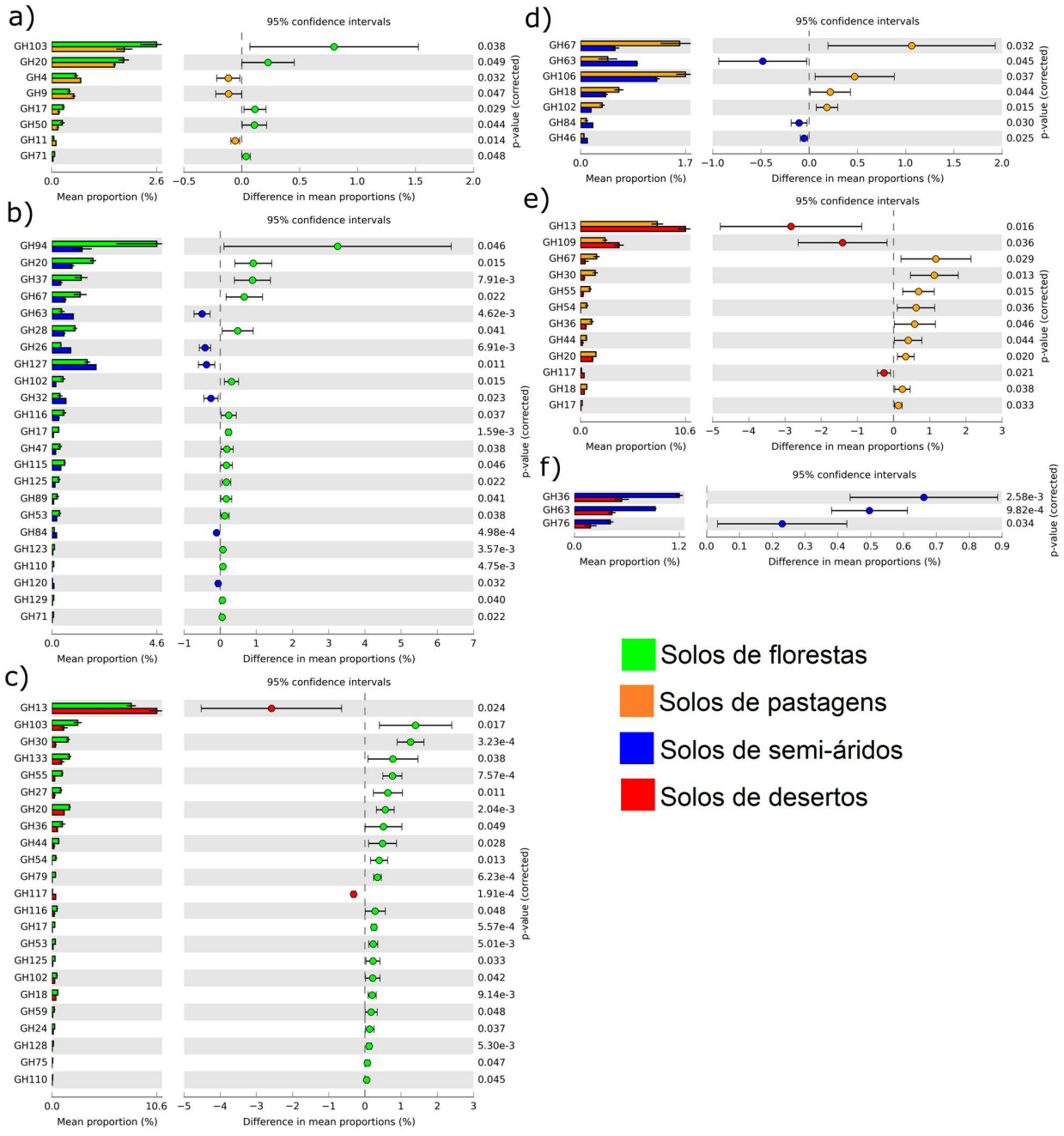


Figura 19. Glicosil hidrolases com abundâncias estatisticamente diferentes entre os grupos de biomas. a) Florestas *vs*. Pastagens; b) Florestas *vs*. semi-áridos; c) Florestas *vs*. Desertos; d) Pastagens *vs*. semi-áridos; e) Pastagens *vs*. desertos; f) Semi-áridos *vs*. Desertos**.**

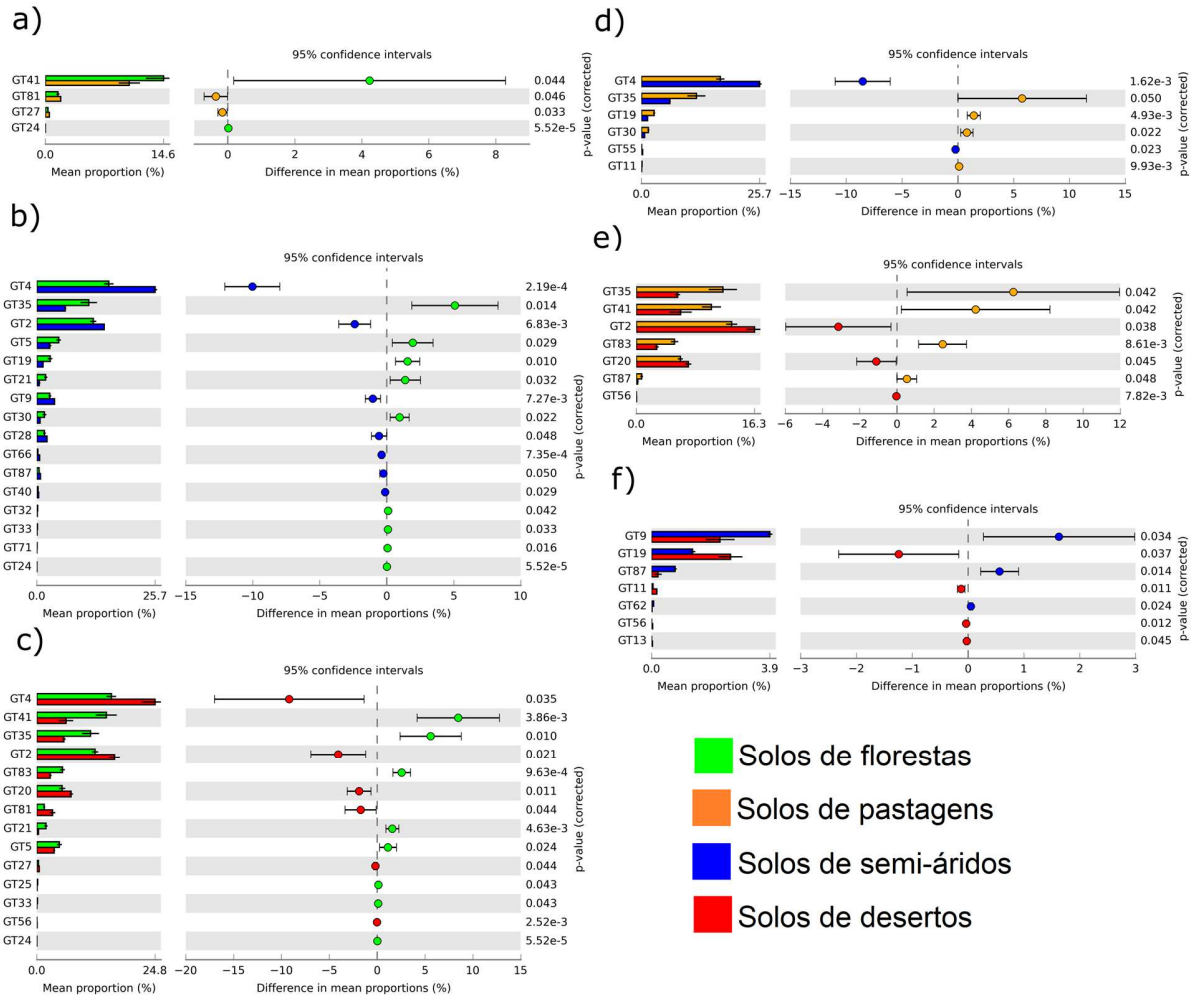


Figura 20. Glicosiltransferases com abundâncias estatisticamente diferentes entre os grupos de biomas. a) Florestas *vs*. Pastagens; b) Florestas *vs*. Semi-áridos; c) Florestas *vs*. Desertos; d) Pastagens *vs*. Semi-áridos; e) Pastagens *vs*. Desertos; e f) Semi-áridos *vs*. Desertos.

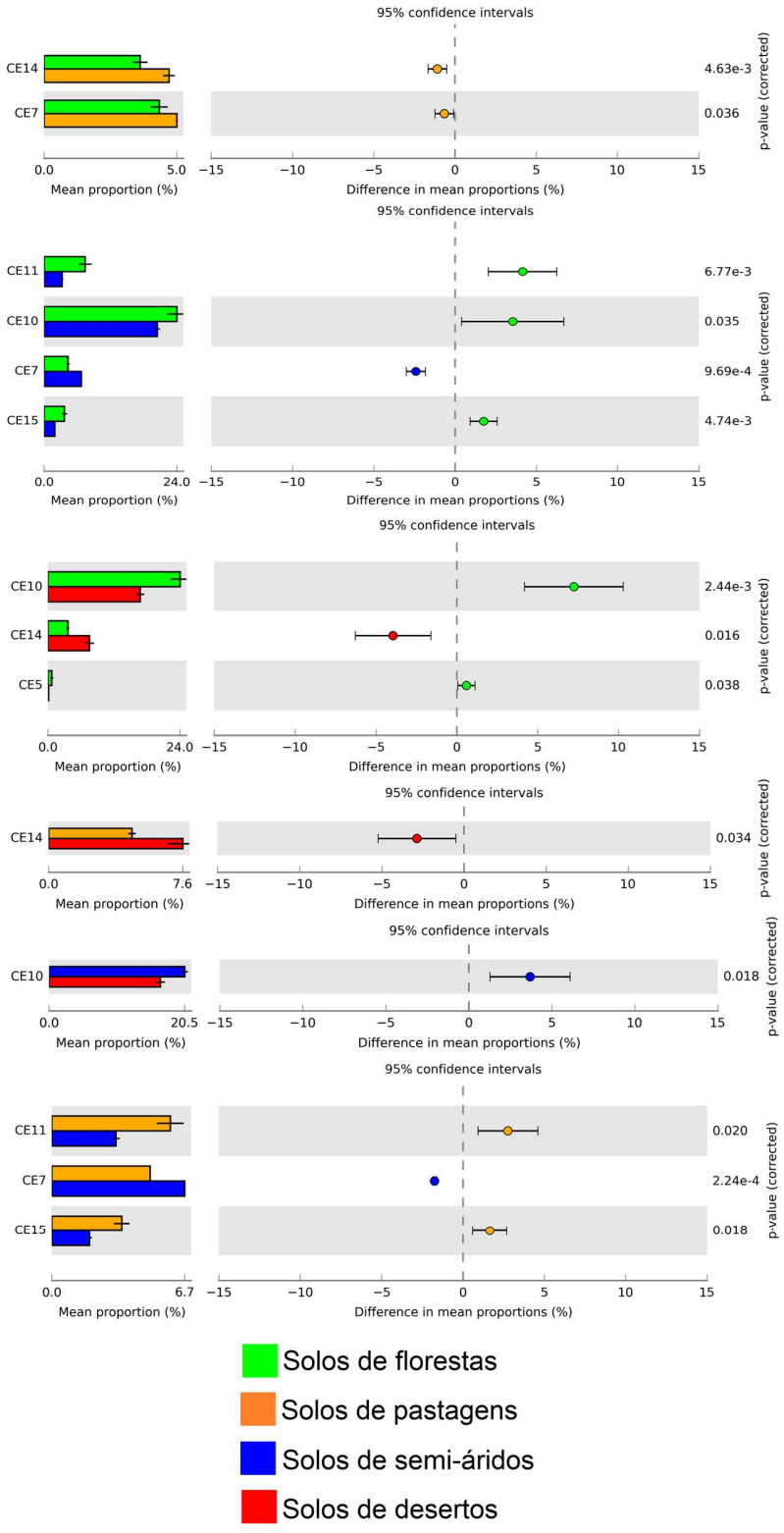


Figura 21. Esterases com abundâncias estatisticamente diferentes entre os grupos de biomas.

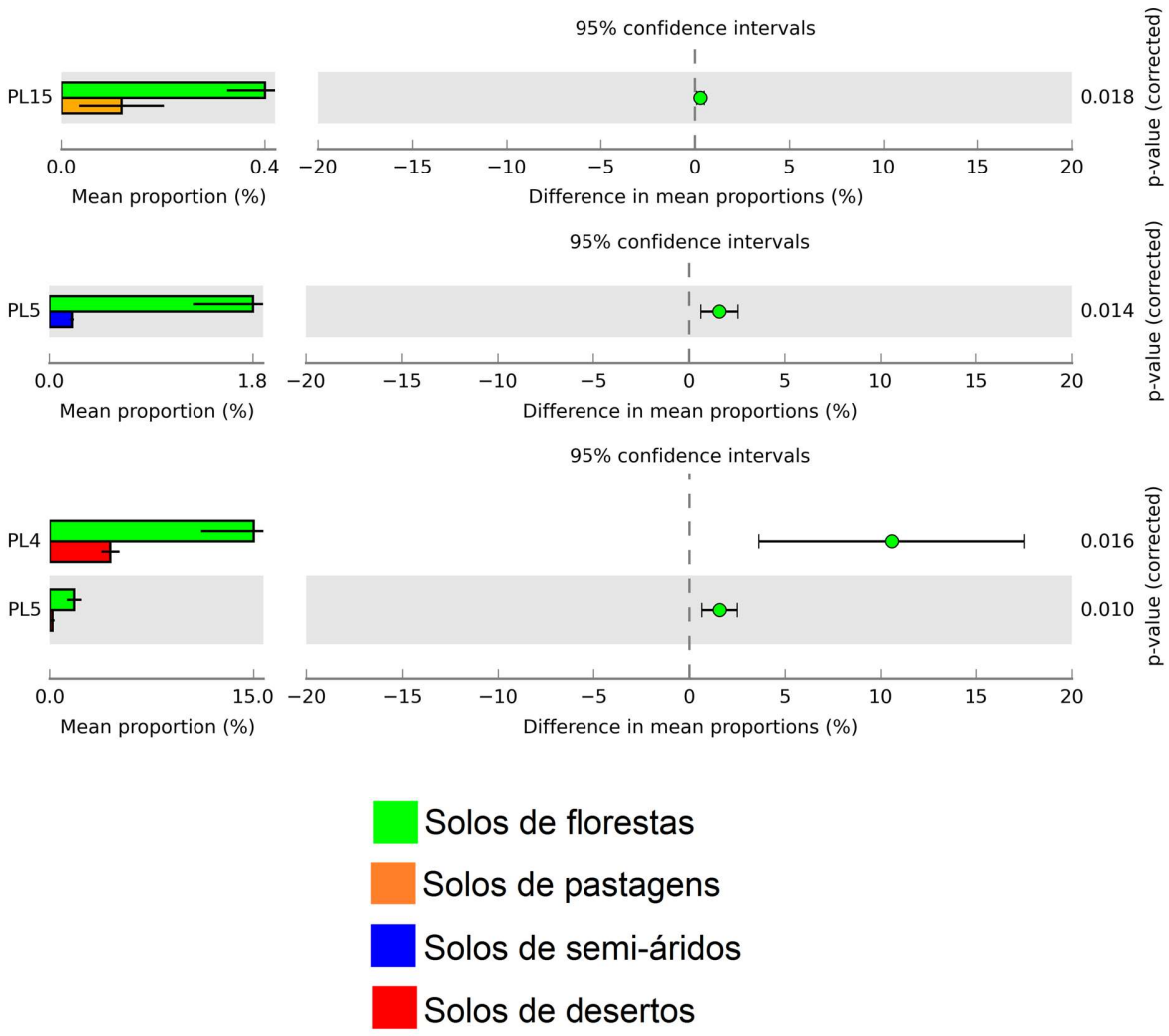
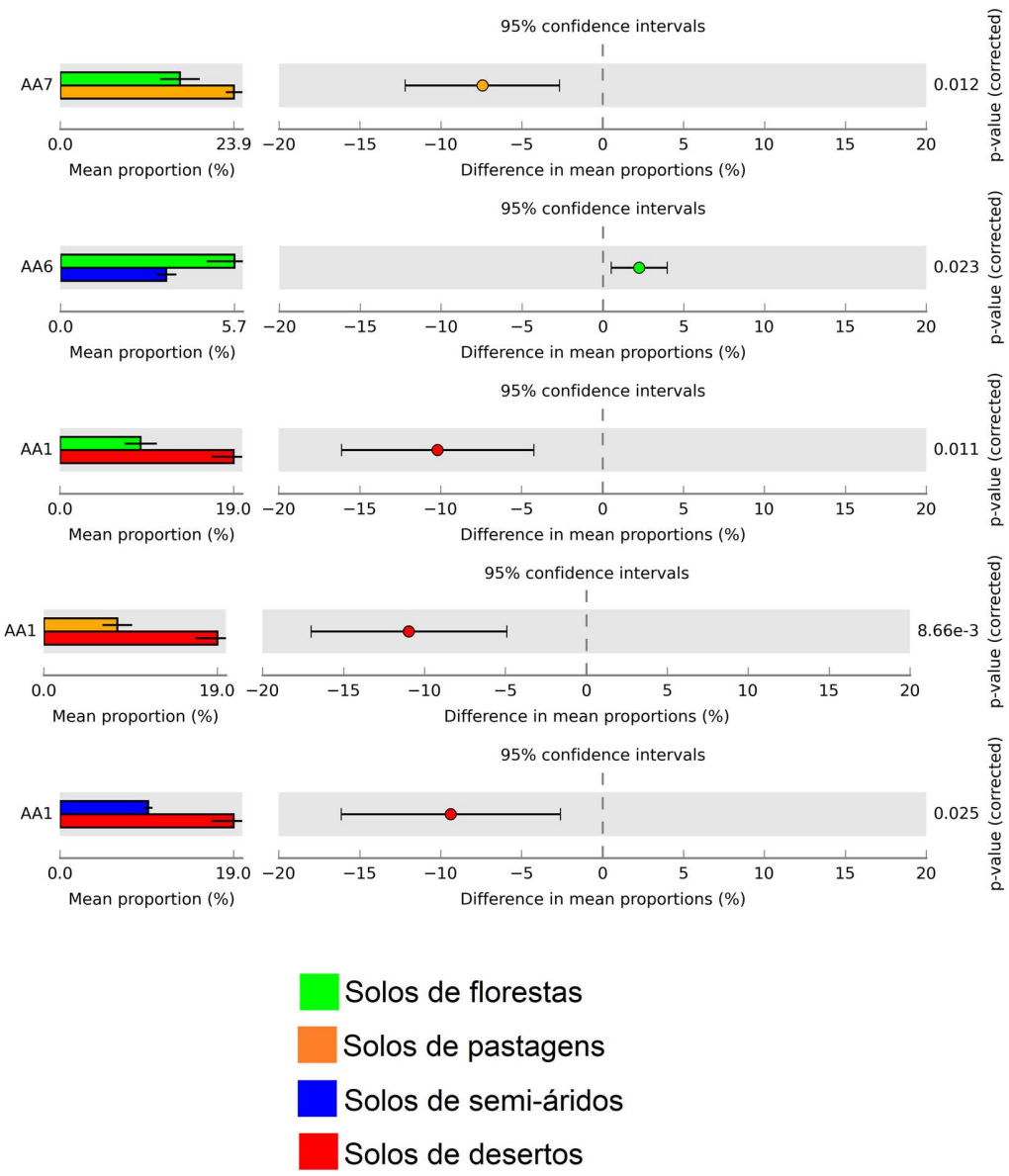


Figura 22. Liases com abundâncias estatisticamente diferentes entre os grupos de biomas.



.

Figura 23. Enzimas auxiliares (AA) com abundâncias estatisticamente diferentes entre os grupos de biomas.

Sequências de duas transglicosilases líticas, a GH102 e GH103, foram observadas em maior abundância em solos de florestas quando comparadas a solos de desertos e semi-áridos (Figura 24). Transglicosilases líticas são enzimas N-acetil-muramidases que clivam a ligação 1,4-glicosídica entre o MurNAc (ácido N-acetilmurâmico) e as unidades de GlcNAc (N-acetilglucosamina) e também catalizam uma reação de transferência de glicosil intramolecular em que uma ligação 1,6-anidro é formada entre o C1 e o C6 do ácido N-acetilmurâmico. Essas enzimas

são componentes essenciais na biossíntese e renovação de heteropolímeros da parede celular (Blackburn e Clarke, 2001). Esse mecanismo de renovação de parede celular pode estar relacionado à resistência de alguns tipos de antibióticos, como vancomicina e teixobactina (Bugg *et al.*, 1992; Ling *et al.*, 2015). Além disso, solos de florestas e pastagens apresentaram uma maior abundância de genes associados à síntese do dissacarídeo lipídio-A (GT19), uma possível Lípido IVA 4-amino-4-desoxi-L-arabinosiltransferase (GT83) e, adicionalmente, uma carboidrato esterase CE11 (UDP-3-0-acil N-acetilglicosamina desacetilase). Estas enzimas estão envolvidas no processo de biossíntese de lipídio A, essencial na síntese da membrana externa de bactérias Gram negativas, proporcionando proteção contra estresses ambientais, como antibióticos. Por isso, essas enzimas são alvos comuns para a produção de antibióticos bacterianos (Jackman *et al.*, 2000; Metzger *et al.*, 2012; Lee e Lee, 2013).

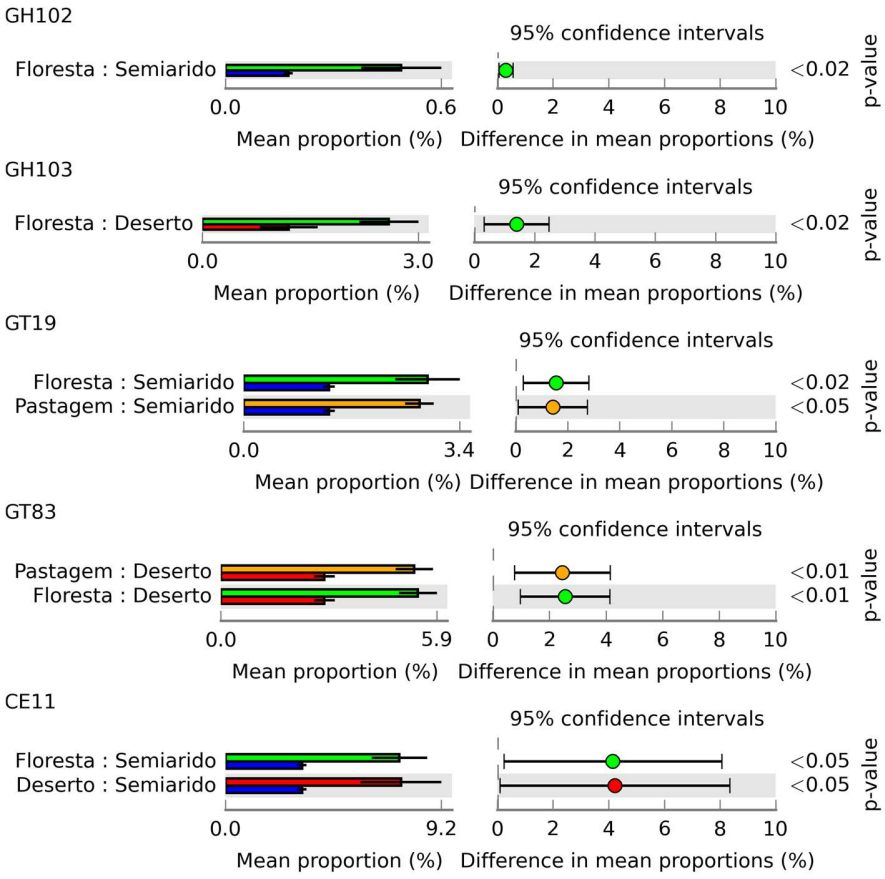


Figura 24. Enzimas envolvidas na biosíntese da parede celular com abundâncias estatisticamente diferentes entre os solos dos diferentes grupos de biomas.

Sete CAZymes exibiram maior abundância em solos de florestas e pastagens em relação aos biomas de desertos e semi-áridos: GH17, GH18, GH20, GH30, GH44, GH53 e GH55. Além disso, solos de pastagens apresentaram maior abundância de GH54 e GH67 em comparação com solos de biomas desérticos (Figura 25). Interessantemente, essas enzimas estão envolvidas no processo de degradação de biomassa (Tabela 4) (Zhou *et al.*, 2014; Couger *et al.*, 2015). Esse resultado corrobora com resultados de trabalhos prévios, como o de Houghton e colaboradores, onde os autores observaram maior densidade de biomassa em solos de florestas, seguido de pastagens e semiáridos e, posteriormente, em menor densidade, solos desérticos (Houghton, Hall e Goetz, 2009). Por outro lado, a neoagarobiose hidrolase GH117 exibiu maior abundância em solos de desertos quando comparada a solos de florestas e pastagens (Figura 25). A função dessa hidrolase é ainda pouco conhecida, entretanto é sabido que está envolvida na conversão de ágar em açúcar fermentável (Lee *et al.*, 2009). Essa fonte de ágar está possivelmente disponível através da parede celular de algas encontradas na crosta de solos desérticos, como relatado em trabalhos anteriores (Cameron, 1960; Lewis e Lewis, 2005; Cardon, Gray e Lewis, 2008). Sequências de solos de desertos anotadas como neogarabiose hidrolase foram selecionadas para anotação taxonômica e, interessantemente, todas as sequências mostraram-se afiliadas ao filo Bacteroidetes.

Tabela 4. Descrição das CAZymes envolvidas na degradação de biomassa e identificadas como significantemente mais abundantes em solos de florestas e pastagens em comparação com desertos e semi-áridos.

CAZyme	**Descrição (Cazy)**
GH17	Glucano endo-1,3-p-glucosidase; Glucano 1,3-p-glucosidase; Liqueninase; B-glucosidase específica de ABA; B-1,3-glucanosiltransglicosilase
GH18	Quitinase; Lisozima; Endo-p-N-acetilglucosaminidase; Hidrolase de peptidoglicano com especificidade de endo-p-N-acetilglucosaminidase; Hidrolase do fator nod; Inibidor de xilanase; Concanavalina B; narbonina
GH20	P-hexosaminidase; Lacto-N-biosidase; B-1,6-N-acetilglucosaminidase; B-6-SO3-N-acetilglucosaminidase
GH30	Endo-p-1,4-xilanase; P-glucosidase; B-glucuronidase; P-xilosidase; P-fucosidase; Glucosilceramidase; B-1,6-glucanase; Glucuronoarabinoxilano endo-p-1,4-xilanase; Endo-p-1,6-galactanase; β-xilosidase
GH44	endoglucanase; xiloglicanase
GH53	Endo-p-1,4-galactanase
GH54	A-L-arabinofuranosidase; B-xilosidase
GH55	Exo-p-1,3-glucanase; Endo-p-1,3-glucanase
GH67	A-glucuronidase; Xilano α-1,2-glucuronidase

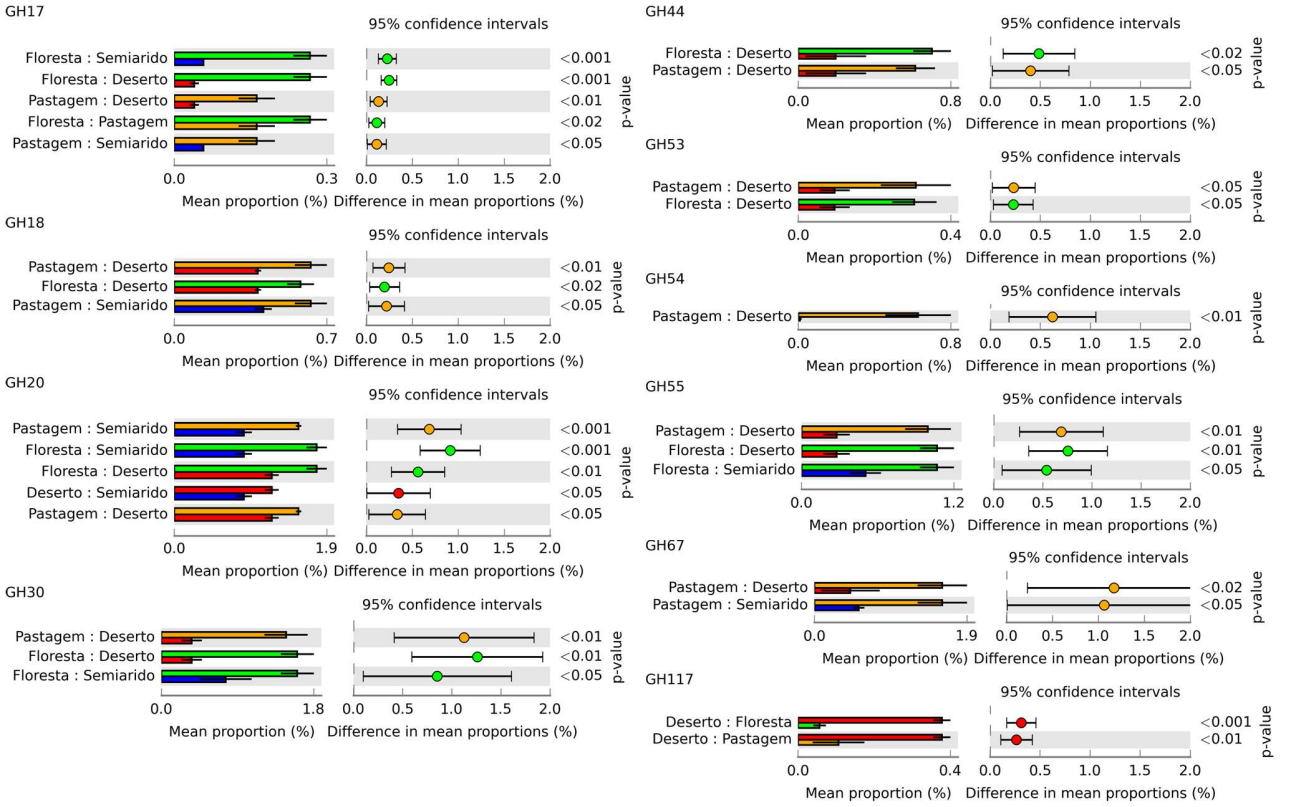


Figura 25. Enzimas identificadas como significantemente mais abundantes em diferentes grupos de biomas relacionadas a degradação de biomassa e neogarabiose (GH 117).

A abundância de enzimas envolvidas no metabolismo de amido e sacarose mostrou-se estatisticamente diferente entre os grupos de biomas. A GH133 (amilo-α-1,6-glucosidase), a GT5 (glucosiltransferase de possível amido) e a GH79 (uma possível β-glucuronidase) foram mais abundantes em solos de florestas em comparação com solos de semi-áridos e desertos. Por outro lado, duas possíveis α-glucosidases (GH4 e GH63) e uma possível endoglucanase (GH6) mostraram maior proporção em solos de semi-áridos que em solos de outros grupos de biomas. Além disso, três grandes famílias de enzimas foram mais abundantes em solos semi-áridos e desérticos, a GH13, GT2, GT4, e GT20, possivelmente relacionadas com o metabolismo de síntese de trealose, como já mencionado acima. A GT81, mais abundante em solos de desertos e semi-áridos, é uma enzima envolvida na biosíntese de glucosilglicerato e, juntamente com a trealose, parece estar envolvida em uma maior resistência a estresses osmóticos e térmicos (Costa, Empadinhas e Da Costa, 2007; Argandoña *et al.*, 2010; Reina-Bueno *et al.*, 2012) (Figura 26).

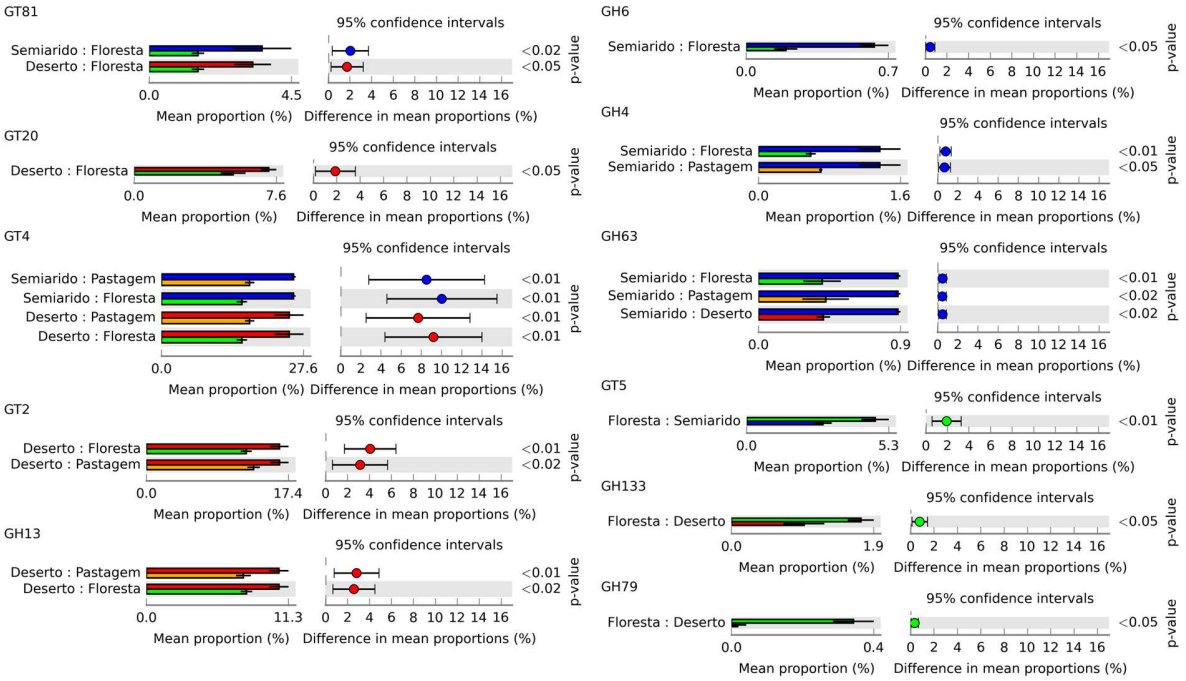


Figura 26. Enzimas relacionadas ao metabolismo de amido e sacarose e identificadas como significantemente mais abundantes em diferentes grupos de biomas.

A glicosilação é uma modificação pós-traducional de proteínas que consiste na adição de sacarídeos à cadeia proteica, sendo um processo primordial para a formação das proteínas de membrana e secretórias. Existem três tipos de glicosilação: N-glicosilação, O-glicosilação e C-glicosilação. Na O-glicosilação, os açúcares são ligados ao radical OH da serina ou da treonina e na N-glicosilação os açúcares são ligados no radical NH_2 de resíduos asparagina. Já na C-glicosilação, o açúcar é ligado a um carbono numa cadeia lateral de triptofano (Faridmoayer *et al.*, 2007). Estes mecanismos são encontrados em todos os domínios de vida, entretanto a N-glicosilação é rara em bactérias. O primeiro relato de N-glicosilação em bactérias foi descrito em *Campylobacter jejuni* e está atualmente descrito como limitado a algumas espécies das classes Deltaproteobacteria e Epsilonproteobacteria, principalmente em bactérias patógenas (Nothaft e Szymanski, 2010). Já a O-glicosilação foi encontrada em algumas linhagens de *Neisseria* spp. e de *Pseudomonas aeruginosa* (Faridmoayer *et al.*, 2007; Nothaft e Szymanski, 2010).

Curiosamente, as análises estatísticas revelaram que os solos do bioma semi-árido tinha uma maior abundância de glicosiltransferases relacionadas à glicosilação, como GT27 e GT39, esta última também abundante em solos de deserto, possivelmente envolvida no mecanismo de O-

glicosilação, e GT26 (WecB) possivelmente envolvida na N-glicosilação (Jarrell *et al.*, 2014). Uma α-N-acetilgalactosaminidase (GH109) foi encontrada como mais abundante em solos de desertos quando comparados com solos de florestas e de pastagens (Figura 27). Essa proteína está associada com a remoção do antígeno A em células vermelhas do sangue, entretanto, apesar de sua ação no ambiente ainda não ser bem estudada, ela parece estar envolvida na degradação de oligossacáridos O-ligados (Blackman *et al.*, 2015). Um trabalho de proteômica mostrou uma maior expressão dessa proteína em temperaturas mais altas em *Leptospirillum* spp. presentes em um biofilme (Mosier *et al.*, 2014).

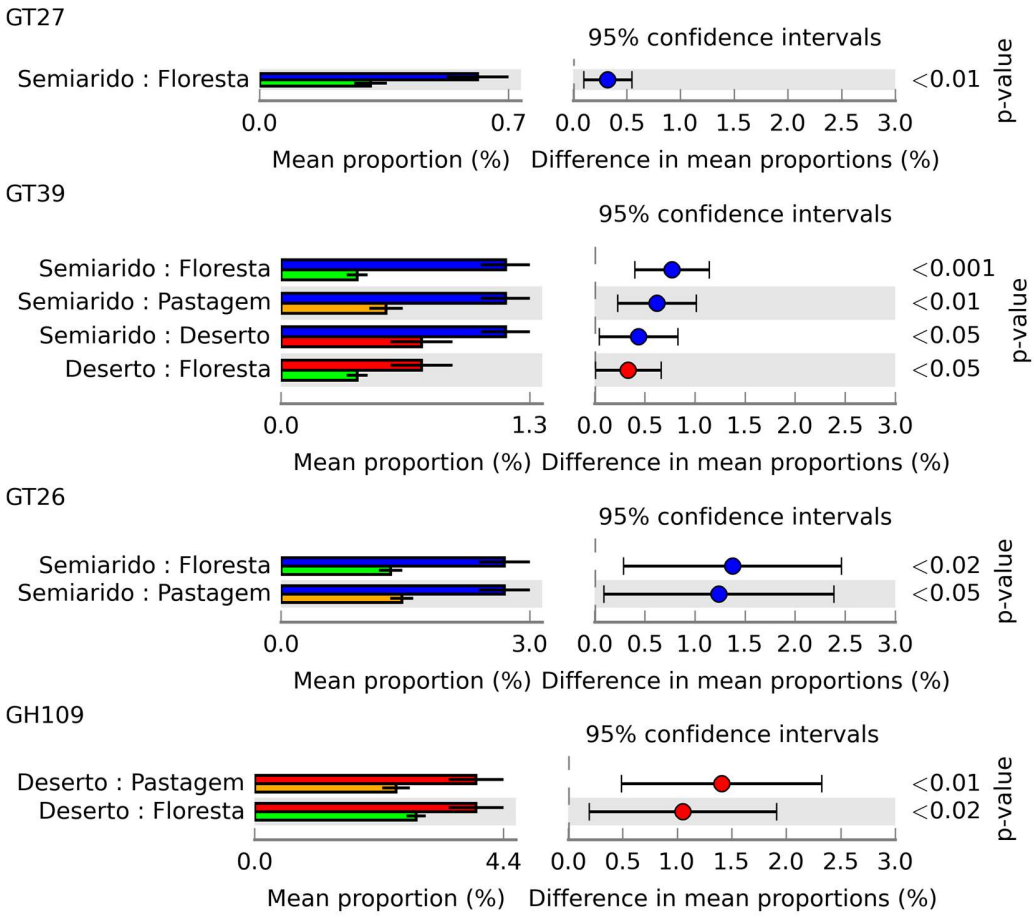


Figura 27. Enzimas identificadas como significantemente mais abundantes em diferentes grupos de biomas relacionadas a glicosilação.

Duas famílias de enzimas ligninolíticas, multicobre oxidase AA1 e um domínio redutase de ferro AA8, apresentaram maior abundância em solos de deserto e de semi-áridos, respectivamente, quando comparado a outros grupos de biomas (Figura 28). Uma família de polissacarídeo liase, uma alginato liase (PL5), foi significativamente mais abundante em solos florestais e de pastagens quando comparado com outros grupos de biomas (Figura 29). Curiosamente, as sequências de alginato liase foram afiliadas à classe Acidobacteria. Esta liase desempenha papel importante na assimilação de alginato, uma fonte de carbono muito abundante na natureza (Kim, Lee e Lee, 2011).

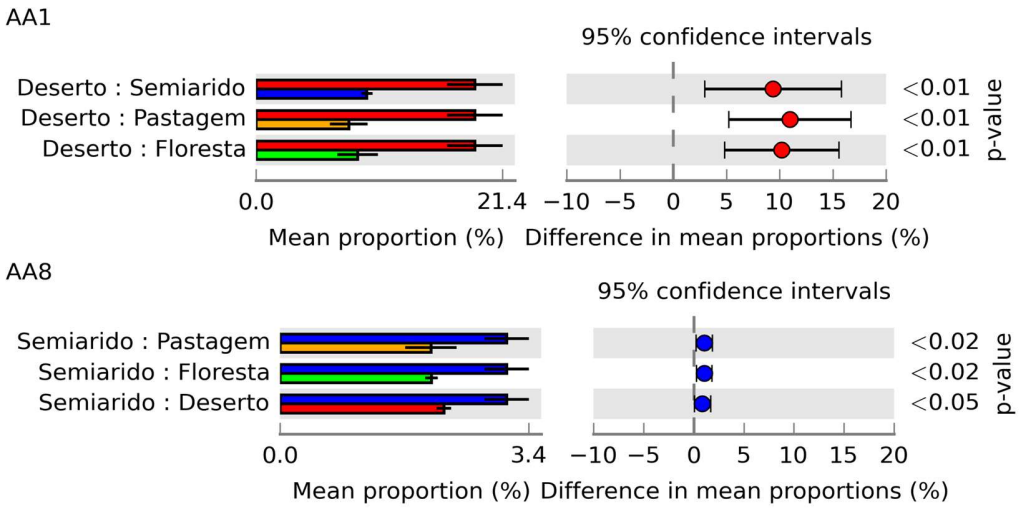


Figura 28. Enzimas ligninolíticas identificadas como mais abundantes em amostras de solos de semi-áridos e desertos.

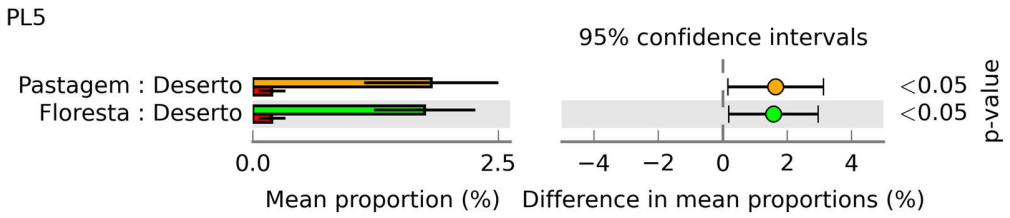


Figura 29. Liase identificada como mais abundante em solos de florestas e pastagens.

5.4. O resistoma nos diferentes biomas

As bactérias do solo têm um extenso repertório de genes de resistência a antibióticos (ARGs) (Nesme *et al.*, 2014; Ling *et al.*, 2015). O conjunto desses em um dado organismo ou ambiente é definido como resistoma. Nesse contexto, a metagenômica vem sendo utilizada para estudar o resistoma em diferentes ambientes. Apesar da maior abundância de ARGs estar mais relacionada a ambientes contaminados, urbanizados e/ou de interação direta com humanos e animais, é conhecida a existência de uma gama de genes de resistência a antibióticos mesmo em regiões remotas e de ambientes naturais, como solos de pastagens (Delmont *et al.*, 2012), geleiras (Segawa *et al.*, 2013), solos de savanas (Riesenfeld, Goodman e Handelsman, 2004) e solos de regiões remotas do Alasca (Allen *et al.*, 2009). Nesse sentido, avaliamos a abundância de ARGs e os grupos taxonômicos que carregam estes genes nas diferentes amostras de solos dos biomas em estudo.

Um total de 2.775 sequências distribuídas em 52 classes de ARG foi identificado entre todas as amostras de solos dos diferentes biomas. Dentre estas, 16% mostraram um percentual de similaridade contra as sequências do banco de dados maior ou igual a 90%. Cerca de 0,002-0,00002% do total de *reads* sequenciados foram identificados como ARGs, percentual este semelhante àqueles encontrados em solos de plantação de arroz (0,0007-0,0010%) (Xiao *et al.*, 2016) e, como esperado, menor que em amostras de lama proveniente de estações de tratamento de água (0,008-0,01%) (Wang *et al.*, 2013) e de sedimentos de rio contaminados por antibióticos (0,02-1,71%).

Amostras de solos dos biomas de floresta e pastagens exibiram uma maior proporção de possíveis ARGs em comparação aos outros grupos de biomas, corroborando com dados de um trabalho anterior (Noah Fierer *et al.*, 2012). Estes resultados são também coerentes com aqueles apontados pela análise das CAZymes, onde foi demonstrada uma maior abundância de enzimas relacionadas à resistência aos antibióticos nas amostras de florestas e de pastagens.

O mecanismo de bomba de efluxo para resistência a multidrogas (MDR) foi a classe de resistência mais predominante entre todas as amostras. Dentre estas, os mecanismos MDR que conferem resistência à cloranfenicol e fluoroquinolona da classe MexEF (22%), a macrolídeos das classes MacAB (14%) e a cloranfenicol da classe CEO (8%) foram as três classes de ARGs mais abundantes. Cerca de 6% das sequências anotadas como possíveis ARGs mostraram-se envolvidas

na resistência à tetraciclina, 6% à bacitracina e 1,4% à vancomicina (Tabela 5). As classes de bombas de efluxo que conferem multirresistência aos antibióticos (MexHI, MexVW, MexEF e MacAB), resistência à tetraciclina (tet_RPP) e à bacitracina (Baca e BCR) foram amplamente distribuídas entre todos os biomas. Bomba de efluxo MDR é um mecanismo comum a todos os microorganismos e funciona de forma eficiente para reduzir a concentração de antibiótico intracelular, além de estar também envolvido na desintoxicação de metabólitos intracelulares, à virulência e ao tráfico de sinal (Martínez, 2008; Xiao *et al.*, 2016).

Para melhor visualizar a correlação entre as classes de ARGs anotadas e os biomas em que estão inseridos, geramos uma rede bipartida onde os biomas são representados por *nós* diferenciados pela cor amarela, em contraste com as classes de ARGs que são os *nós* identificados na cor branca. As arestas ligam os possíveis ARGs a cada bioma em que foi identificado, sendo que a espessura da aresta é equivalente à abundância na qual o ARG foi encontrado e a cor da aresta foi definida de acordo com o grupo de bioma: verde para florestas, laranja para pastagens, azul escuro para semi-áridos, vermelho para deserto e azul ciano para Tundra (Figura 30). Nessa rede podemos observar a maior abundância de ARGs nos grupos de biomas de florestas e pastagens. A Caatinga e deserto quente apresentaram alta abundância dos genes de resistência *MacAB* e *MexEF*. Já o bioma da tundra mostrou elevado número de sequências identificadas pertencentes as classes *tet_RPP* e *MacAB* (resistência à tetraciclina e macrolídeos – MDR bomba de efluxo). Alguns ARGs foram identificados em apenas um único bioma de acordo com os parâmetros que utilizamos, como *KsgA* (resistência à casugamicina) na floresta tropical, *Ble* (resistência à bleomicina) em floresta temperada de coníferas, *CATB* e *cml* (resistência ao cloranfenicol) para desertos salino e quente, respectivamente, *fos* (resistência à fosfomicina) na Caatinga e *TSNR* (resistência à tiostreptona) em amostras de solo do bioma Mediterrâneo. Tioestreptona é um antibiótico derivado de cepas de estreptomicetos e potente agente antibacteriano contra patógenos Gram positivos, tais como bactérias resistentes à meticilina (*Staphylococcus aureus*) e bactérias resistentes à vancomicina (enterococos) (Kelly, Pan e Li, 2009).

Genes relacionados a resistência à polimixina foram significativamente mais abundantes em solos dos biomas de florestas, pastagens e tundra. A polimixina é produzida por bactérias Gram-positivas e é seletivamente tóxica para muitas bactérias Gram-negativas (Velkov *et al.*, 2010). No entanto, algumas bactérias Gram-negativas carregam ARGs que codificam

proteína de resistência à polimixina (*Arna*), envolvidos na biossíntese de lipídio A, os quais se mostraram mais abundantes em amostras de solos de florestas (anotação dada pelo CAZy). Um número elevado de genes que conferem resistência à tetraciclina foi encontrado no bioma tundra (17% do total de sequências identificadas como possíveis ARGs em Tundra). Apesar de mecanismos de resistência à tetraciclina estarem normalmente associados com fezes humanas e animais (Zhu *et al.*, 2013; W. Wang *et al.*, 2014), alguns estudos têm identificado a ocorrência de genes de resistência a antibióticos em regiões pristinas, como na camada ativa de solos de regiões remotas extremamente frias, principalmente no norte do Canadá e em solos do Alasca (Allen *et al.*, 2009; Schloss *et al.*, 2010; Perron *et al.*, 2015).

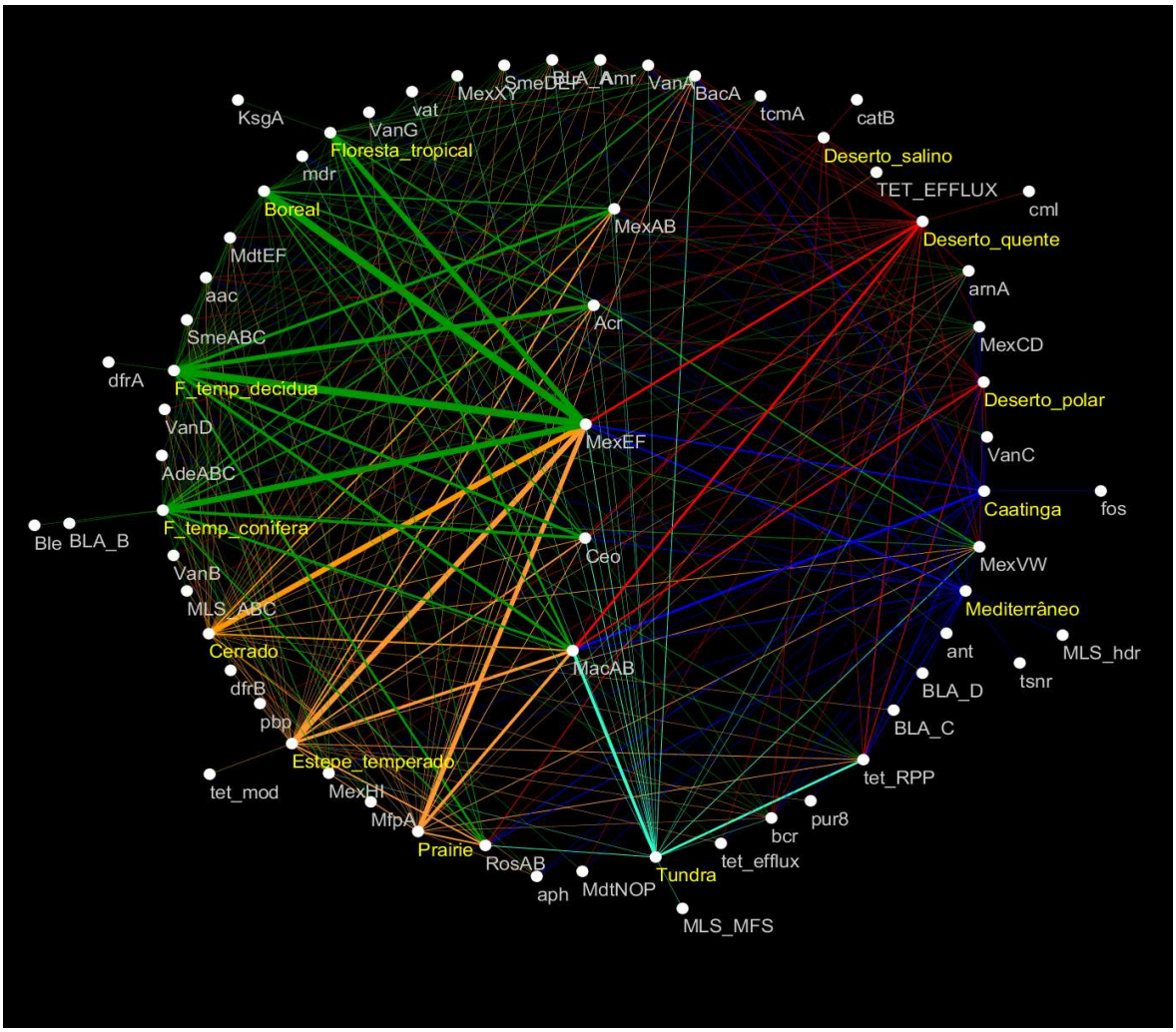


Figura 30. Rede correlacionando os genes de resistência a antibióticos ao bioma e grupo de bioma ao qual estão relacionados. Os nós na cor amarela representam os biomas e os nós em cor branca representam os ARGs.

Tabela 5. Classes de ARGs definidas pelo banco de dados ARDB e seus respectivos mecanismos e antibióticos alvos.

Classes das ARGs	Mecanismo	Antibiótico
mdtef, mdtnop, acr, amr, ceo, smeabc, smedfe, mexab, mexcd, mexef, mexhi, mexvw, mexxy, adeABC	sistema de transporte de resistência-nodulação-divisão celular	multidrogas
Macab	sistema de transporte de resistência-nodulação-divisão celular	multidrogas: macrólido
tet_rpp, tet_efflux, tet_mod	proteção da proteína ribossômica, bomba de efluxo, desconhecido	tetraciclina
cml, catB	classe de transportadores da superfamília dos principais facilitadores	cloranfenicol
anrA	modificação do lipídio A	polimixina
ant, aac, aph	aminoglicosídeo O-nucleotidiltransferase/ N-acetiltransferase	resistência a multidrogas
baca, bcr	bomba de efluxo	bacitracina
mls_mfs	classe de transportadores da superfamília dos principais facilitadores	multidrogas: macrólido
mls_abc	sistema transportador ABC	multidrogas: macrólido
Fos	glutationa transferase, metalo-glutationa transferase	fosfomicina
vanA, vanB, vanC, vanD, vanG	*operon* de genes de resistência a vancomicina	vancomicina
bla_A, bla_B, bla_C, bla_D	classe A de beta-lactamase	β-lactamases
dfra, dfrb	dihidrofolato redutase	trimetoprim
Rosab	sistema antiporter de potássio	multidrogas
Tcma	transportadores da *s*uperfamília dos principais facilitadores	multidrogas
Mfpa	proteção do DNA girase a partir da inibição de quinolonas	fluoroquinolona
Vat	acetiltransferases	estreptogramina
pur8	-	puromicina
Pbp	proteína de ligação a penicilina	β-lactamase
Tsnr	Metiltransferase de RNA	tiostreptona
Ksga	dimetiladenosina transferase	casugamicina
Ble	Prevenção da quebra de DNA induzida por bleomicina	bleomicina

Para entender melhor quais são os grupos taxonômicos que carregam esses genes de resistência a antibióticos em cada bioma (Figura 31) e a qual tipo de antibiótico esses genes conferem resistência (Figura 32), as sequências identificadas como possíveis ARGs foram anotadas taxonomicamente. De um total de 2.775 de ARGs identificados, apenas 1.094 seqüências foram anotadas taxonomicamente com sucesso utilizando o algoritmo LCA (*Lower Common Ancestor*). Gamaproteobacteria (34%), Alfaproteobacteria (27%) e Betaproteobacteria (19%) foram as classes

taxonômicas mais abundantes relacionadas aos possíveis ARGs. ARGs de amostras de solos de floresta e de pastagem foram principalmente afiliados a Gamaproteobacteria, Alfaproteobacteria, Betaproteobacteria e Actinobacteria. Por outro lado, sequências de ARGs atribuídas a Actinobacteria foram mais predominantes em amostras de solo dos desertos quente e frio e também em amostras de solo de regiões semi-áridas. Sequências identificadas como pertencentes a Planctomycetia e Acidobacteria só foram identificadas em amostras de solo de floresta boreal (resistência a multidrogas por bomba de efluxo e resistência à bacitracina). Os solos de desertos salinos apresentaram ARGs afiliados apenas à classe Gamaproteobacteria. Apenas os solos da Caatinga exibiram sequências de ARGs atribuídos à classe Bacilli, que é um grupo taxonômico associado a uma vasta gama de genes que conferem resistência a diversos tipos de antibióticos, incluindo vancomicina, bacitracina, fosfomicina, bleomicina, β-lactamase e multidrogas (bomba de efluxo). Outro resultado interessante foi a ocorrência de ARGs relacionados a apenas um grupo taxonômico, como a resistência à bleomicina, mdr e fosfomicina em membros da classe Bacilli; a resistência à fluoroquinolona, puromicina e tiostreptona em Actinobacteria; e a resistência à casugamicina e estreptogramina em Gamaproteobacteria (Figura 31 e Figura 32).

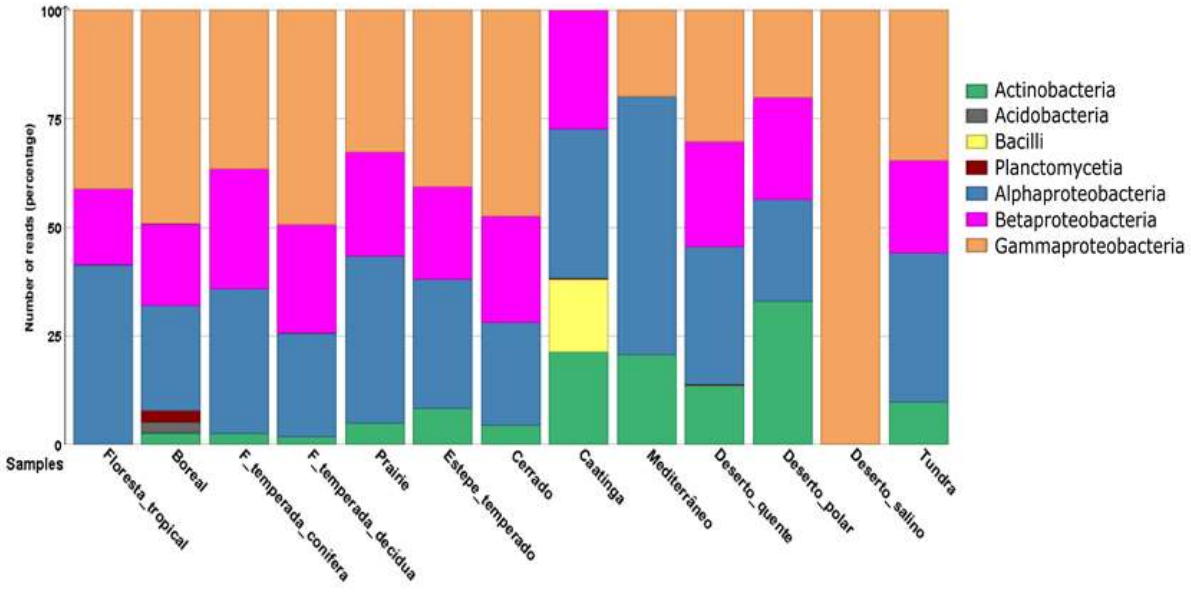


Figura 31. Afiliação taxonômicas das ARGs em cada bioma em nível de classe.

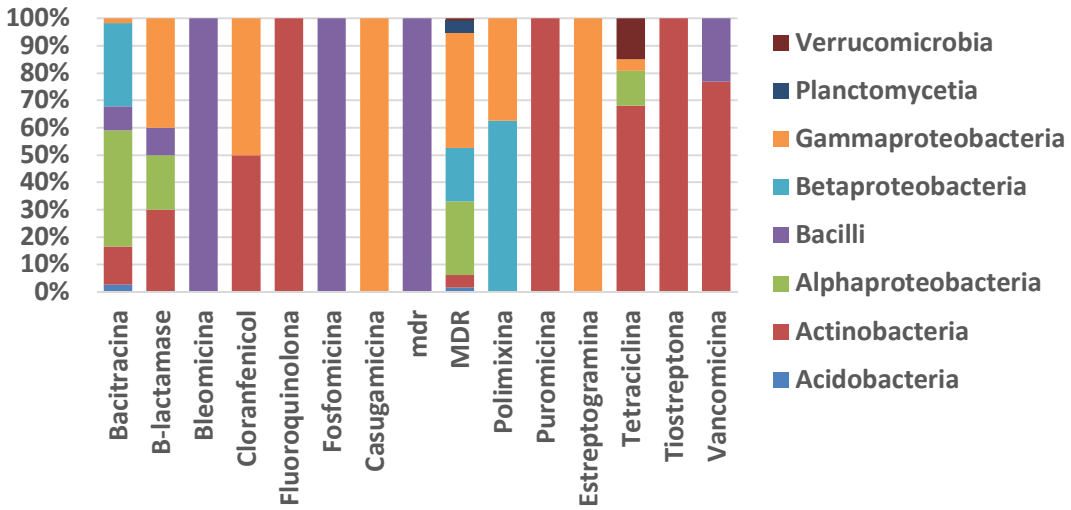


Figura 32. Afiliação taxonômicas das ARGs em cada tipo de resistência aos diferentes antibióticos.

Além disso, para analisar os padrões de co-presença e exclusão mútua entre os grupos taxonômicos e as classes de ARGs identificadas, geramos uma rede bipartida de co-ocorrência que revelou dois agrupamentos (Figura 33). A co-presença é definida pela correlação de presença entre dois atributos e a exclusão mútua é uma correlação na qual a presença de um atributo regula a ausência de outro atributo. Em um primeiro *cluster*, um padrão de exclusão mútua foi observado entre a classe de resistência a multidrogas (MDR) *MexVW* e a classe bacteriana Actinobacteria. Uma maior abundância dessa classe de antibióticos foi encontrada em solos de florestas e pastagens em comparação a amostras de regiões semi-áridas. Por outro lado, houve uma maior abundância de Actinobacteria em solos de regiões semi-áridas em comparação aos outros grupos de biomas. No outro *cluster*, grupos taxonômicos mais abundantes em solos de desertos, como Deferribacteres, Aquificae, Thermotogae, entre outros, apresentaram um padrão de co-presença com a classe de resistência a cloranfenicol (*catB*). Ainda, membros dos filos Chloroflexi (Dehalococcoidetes, Chroflexi (classe) e Thermomicrobia), e das classes Clostridia e Deinococci, apresentaram um padrão de exclusão mútua com classes de MDR. Essas classes estão envolvidas na resistência à acriflavina, aminoglicosideo, beta-lactamases, glicilciclina e macrolídeos (*Acr)*, resistência à oxorrubicina e eritromicina (*MdtEF*), MDR (*MexHI)* e cloranfenicol (*Ceo)*. Interessantemente, esses grupos taxonômicos são mais abundantes em solos de semi-áridos e desérticos, em contraste com a maior abundância dessas classes de genes de resistência a antibióticos em solos de florestas

e pastagens. Alfaproteobacteria, Betaproteobacteria e Verrucomicrobia, mais abundantes em solos de florestas e pastagens, apresentaram um padrão de co-presença com as classes de MDR *MexAB* (resistência a aminoglicósido, beta-lactamase, fluoroquinolona, tetraciclina e tigeciclina) e *Amr* (acriflavina, aminoglicósido e macrólido), resistência a fluoroquinolona (*MfpA*), beta-lactamase (*BLA-A*), polimixina (*arnA*), além da *Acr* e *MdtEF*, o que corrobora com a maior abundância dessas classes de antibióticos em amostras de florestas e pastagens (Figura 33). Além da correlação entre a maior abundância de classes de genes de resistência a antibióticos com a abundância de grupos taxonômicos, essas correlações podem refletir também o resultado do possível enriquecimento de genes de resistência em algumas cepas em diferentes tipos de solos.

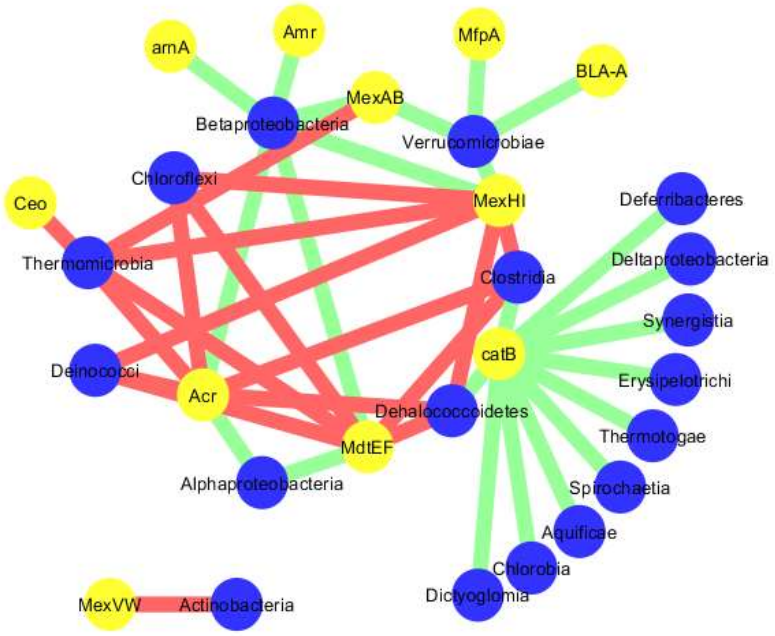


Figura 33. Rede de co-ocorrrência entre as classes de ARGs e os grupos taxonômicos nos quais as ARGs foram anotadas. Os nós em azul representam os grupos taxonômicos (classes) e os nós em amarelo representam as classes de antibióticos definidas pelo ARDB. As arestas em verde representam padrões de co-presença e as arestas em vermelho representam exclusão mútua.

Para obtermos uma perspectiva integrada destes resultados, juntamente com as informações dos grupos de biomas, agrupamos os tipos de resistência em cada bioma e as classes taxonômicas atribuídas às ARGs por grupos de biomas em um gráfico ternário. Primeiramente, realizamos um agrupamento hierárquico (UPGMA) a partir dos dados de abundância de ARGs em

cada conjunto de dados dos biomas (Figura 34). As amostras agruparam-se claramente em 2 clusters, um englobando os grupos de biomas de florestas e pastagens (~74% similaridade), e o outro englobando os grupos de biomas semi-áridos, tundra e desertos (~68% de similaridade), sendo que o deserto salino se comportou como um *outgroup*. Embora a anotação funcional possa sugerir um padrão entre eles, as amostras de desertos diferem uns dos outros com relação à temperatura média, salinidade e muitos outros parâmetros, e que podem influenciar os tipos de ARG associados a estes ecossistemas. Portanto, para o gráfico ternário, os biomas foram agrupados em florestas, pastagens e desertos /regiões semi-áridas, excluindo deserto salino e a tundra (Figura 34).

O gráfico ternário então foi gerado a partir de dois resultados distintos dos anteriores, a anotação taxonômica (*best hit*) dos ARGs e a abundância de ARGs, independentemente de sua afiliação taxonômica. Os ARGs estão classificados no gráfico pelo tipo de resistência aos quais estão relacionados. É possível observar que há um grande grupo de genes de resistência a antibióticos ubíquo aos grupos de biomas. Alguns poucos ARGs parecem bioma-específicos, como a resistência à bleomicina e à casugamicina em solos de florestas, e a resistência à fosfomicina, mdr e tiostreptona em solos desertos e/ou semi-áridos. Interessantemente, a bleomicina e a casugamicina foram descobertas pelo mesmo grupo e isoladas a partir de cepas de *Streptomyces.* Bleomicina é um antibiótico conhecido por ser bastante utilizado no tratamento do câncer e também por causar má formação de células em fungos, tendo assim ação fungicida (Moore *et al.*, 2003). Genes de resistência a esse antibiótico foram encontrados a partir de bibliotecas metagenômicas, não apenas em Actinobacteria, mas principalmente em Proteobacteria e Firmicutes (Mori *et al.*, 2008). A casugamicina também é um eficiente antibacteriano e fungicida (Yoshii, Moriyama e Fukuhara, 2012). Uma possível explicação para essas resistências só terem sido identificadas em solos de florestas é a maior abundância de fungos identificados neste ambiente, portanto, a produção desse antibiótico pode estar relacionada à ação fungicida, mesmo que em pequena escala. Portanto, alguns grupos bacterianos poderiam conter genes de resistência a esse antibiótico neste ambiente. Além disso, uma maior abundância de Proteobacteria foi identificada nesse ambiente, o que explicaria a maior abundância de genes de resistência à *bleomicina* identificados no trabalho de Mori e colaboradores (Mori *et al.*, 2008)(Figura 34).

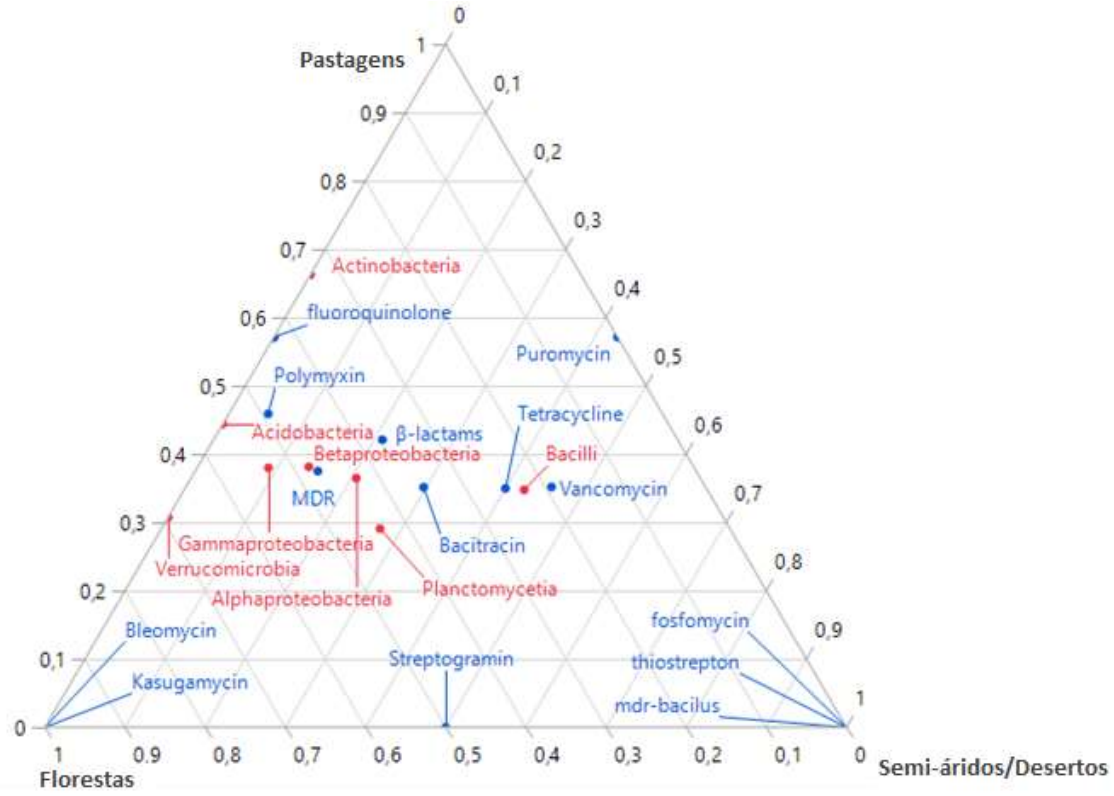


Figura 34. Gráfico ternário mostrando a relação entre os ARGs e os grupos taxonômicos inferidos a partir das ARGs nos três grupos de biomas: floresta, pastagem e deserto/semi-árido.

5.5. Avaliação dos perfis de proteínas de choque térmico (*heat shock protein*) encontrados em cada bioma

As proteínas de choque térmico (HSPs) são uma família altamente conservada de chaperonas induzidas por uma grande variedade de estresses, incluindo a exposição ao frio, luz UV ou estresses bióticos (Park e Seo, 2015). Essas proteínas foram divididas em famílias nomeadas de acordo com o seu peso molecular expresso em kilodaltons, como HSP20, HSP40, HSP60, HSP70, HSP90 e HSP100. Um total de 39.012 HSPs foi identificado em todas as amostras, sendo a amostra de solo de deserto salino a que apresentou a maior abundância de possíveis proteínas de choque térmico. As amostras dos biomas da tundra, semi-árido e deserto mostraram uma maior abundância de HSPs em relação às amostras de solos dos biomas de florestas e pastagens (Figura 36).

As famílias HSP100 e HSP70 foram as mais abundantes em todas as amostras. A HSP100 é uma família de proteínas de ligação a ATP, que atua sobre a agregação da proteína,

permitindo que as proteínas solubilizadas possam ser enoveladas com o auxílio do sistema da família HSP70. Já as proteínas HSP70 são conhecidas como concentradores da rede celular de chaperonas moleculares que possuem um papel fundamental na função celular normal (Parsell *et al.*, 1994; K. Wang *et al.*, 2014). Estas descobertas podem explicar a alta abundância dessas famílias de HSP amplamente distribuídas em todos os biomas. Além disso, a HSP90 foi estatisticamente mais abundante em solos de florestas e pastagens em comparação com solos desérticos e semi-áridos. Essa HSP é conhecida por estar envolvidas na transdução de sinal e manutenção de cromossomos.

O bioma de deserto salino apresentou maior abundância de HSP20 e HSP60 em comparação com os outros biomas. As proteínas da família HSP20 impedem a desnaturação de proteínas, mantendo de forma competente o estado enovelado, de modo que elas podem ser envolvidos em desagregação dependentes de ATP, através do sistema de chaperona HSP70 / 90 (Park e Seo, 2015). HSP60 é uma proteína de choque térmico mitocondrial envolvida no processo de enovelamento adequado de proteínas. Em amostras de solos do deserto salino, 36% das sequências de HSP20 e 70% das sequências de HSP60 foram afiliadas a arqueias, que são conhecidas principalmente por suas propriedades extremofílicas. Gammaproteobacteria, Deltaproteobacteria e Alphaproteobacteria foram as principais classes de bactérias relacionadas às seqüências de HSP20 e HSP60. Essas famílias de HSPs são muito importantes para manter o metabolismo energético bacteriano vivo em ambientes altamente estressantes. Taxas de sobrevivência mais baixas foram detectadas em mutantes de HSP60 em *Caulobacter crescentus* durante estresse oxidativo, salino e osmótico (Susin *et al.*, 2006). A superexpressão de genes de proteínas de choque térmico, tal como HSP20, foi observada em *Halobacterium* durante o tratamento em alta temperatura (Shukla, 2006).

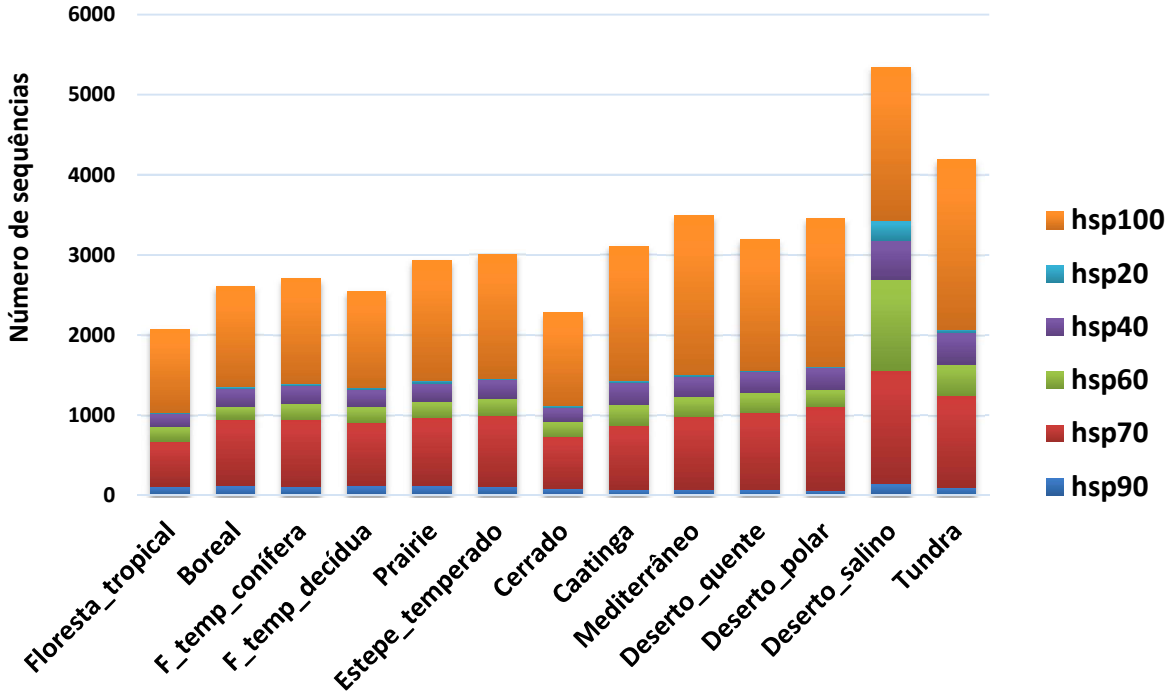


Figura 35. Distribuição de proteínas de choque térmico nas amostras de solos dos diferentes biomas.

6. CONCLUSÕES

Para compreender os perfis taxonômicos e funcionais microbianos dos solos de biomas globais, conjuntos de dados de 11 biomas diferentes, classificados em 5 grupos (florestas, pastagens, semi-áridos, desertos e tundra), foram selecionados a partir de estudos publicamente disponíveis. A estrutura da comunidade microbiana foi claramente diferenciada entre os biomas de florestas, pastagens, semi-áridos e, apesar de mais heterogêneas, as amostras desérticas. Com base na anotação taxonômica, as amostras de tundra agruparam-se juntamente com o grupo de bioma de pastagens, o que pode ser explicado pelo fato destas terem sido coletadas no verão, e nesta estação a vegetação deste bioma é basicamente rasteira, se assemelhando aos biomas de pastagens. Funcionalmente, os grupos de biomas foram mais significativamente diferenciados pela abundância de genes associados com o metabolismo de carboidratos, biosíntese de proteínas e resistência aos antibióticos.

O perfil de enzimas ativas de carboidratos revelou principalmente uma maior abundância de genes relacionados à degradação de biomassa e na biosíntese de lipídio A, envolvido também na resistência a alguns tipos de antibióticos, em amostras de solos de florestas e pastagens. Por outro lado, genes envolvidos na biosíntese da trealose e de compostos coadjuvantes, glicosilação e outros processos relacionados a estresses ambientais mostraram-se mais abundantes em amostras de solos desérticos e de regiões semi-áridas. A análise de genes de resistência aos antibióticos (ARG) revelou uma maior abundância de ARGs em solos de florestas e de pastagens, corroborando com resultados encontrados no perfil de enzimas de enzimas envolvidas com a degradação de carboidratos. As classes de ARGS que confere resistência a multidrogas, à bacitracina e à tetraciclina foram amplamente distribuidas entre todos os biomas, entretanto, um grande percentual de ARGs que conferem resistência à tetraciclina foi encontrado no total de ARGs identificadas no bioma da tundra. As proteínas de choque térmico foram, em geral, mais abundantes em amostras de solo dos biomas desertos, semi-áridos e tundra, sendo o deserto salino com maior número de proteínas identificadas.

Este trabalho mostra que há uma estreita ligação entre o microbioma e o ecossistema ou bioma no qual está inserido. Com base em um abrangente conjunto de dados representando os principais biomas do globo, mostramos com sucesso que o potencial metabólico dos microrganismos está enriquecido de genes que melhor conseguem explorar as fontes energéticas disponíveis no ambiente, e de genes que os tornam mais aptos a sobreviver naquele tipo de

ambiente. Por exemplo, o grande número de ARGs em solos mais competitivos, como florestas e pastagens, e o grande número de genes relacionados a estresses em solos de ambientes extremos. Este estudo forneceu uma percepção ampla da diversidade taxonômica e funcional do microbioma de solos em uma escala global.

REFERÊNCIAS

Aborn, K. e Berbine, J. (2010) *Tropical grassland and savannas*.

Allen, H. K., Moe, L. a, Rodbumrer, J., Gaarder, A. e Handelsman, J. (2009) "Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil.", *The ISME journal*, 3(2), p. 243–251. doi: 10.1038/ismej.2008.86.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. e Lipman, D. J. (1997) "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", *Nucleic Acids Research*, 25(17), p. 3389–3402. doi: 10.1093/nar/25.17.3389.

Argandoña, M., Nieto, J. J., Iglesias-Guerra, F., Calderón, M. I., García-Estepa, R. e Vargas, C. (2010) "Interplay between iron homeostasis and the osmotic stress response in the halophilic bacterium Chromohalobacter salexigens", *Applied and Environmental Microbiology*, 76(11), p. 3575–3589. doi: 10.1128/AEM.03136-09.

Basu, B. e Apte, S. K. (2012) "Gamma radiation-induced proteome of Deinococcus radiodurans primarily targets DNA repair and oxidative stress alleviation.", *Molecular & cellular proteomics : MCP*, 11(1), p. M111.011734. doi: 10.1074/mcp.M111.011734.

Belezas Naturais (2015) *Fotos incríveis da Amazônia*. Disponível em: http://belezasnaturais.com.br/fotos-incriveis-da-amazonia (Acessado: 19 de setembro de 2016).

Beneduzi, A., Moreira, F., Costa, P. B., Vargas, L. K., Lisboa, B. B., Favreto, R., Baldani, J. I. e Passaglia, L. M. P. (2013) "Diversity and plant growth promoting evaluation abilities of bacteria isolated from sugarcane cultivated in the South of Brazil", *Applied Soil Ecology*. Elsevier B.V., 63, p. 94–104. doi: 10.1016/j.apsoil.2012.08.010.

Bhattacharya, T., Ghosh, T. S. e Mande, S. S. (2015) "Global profiling of carbohydrate active enzymes in human gut microbiome", *PLoS ONE*, 10(11), p. 1–20. doi: 10.1371/journal.pone.0142038.

Bissett, A., Fitzgerald, A., Meintjes, T., Mele, P. M., Reith, F., Dennis, P. G., Breed, M. F., Brown, B., Brown, M. V, Brugger, J., Byrne, M., Caddy-retalic, S., Carmody, B., Coates, D. J., Correa, C., Ferrari, B. C. e Gupta, V. V. S. R. (2016) "Introducing BASE : the Biomes of Australian Soil Environments soil microbial diversity database", *GigaScience*. GigaScience, 5(21). doi: 10.1186/s13742-016-0126-5.

Blackburn, N. T. e Clarke, A. J. (2001) "Identification of four families of peptidoglycan lytic transglycosylases.", *Journal of molecular evolution*, 52(1), p. 78–84. doi: 10.1007/s002390010136.

Blackman, L. M., Cullerne, D. P., Torreña, P., Taylor, J. e Hardham, A. R. (2015) "RNA-Seq analysis of the expression of genes encoding cell wall degrading enzymes during infection of lupin (Lupinus angustifolius) by Phytophthora parasitica", *PLoS ONE*, 10(9), p. 1–30. doi: 10.1371/journal.pone.0136899.

Boit, A., Sakschewski, B., Boysen, L., Cano-Crespo, A., Clement, J., Alaniz, N. G., Kok, K., Kolb, M., Langerwisch, F., Rammig, A., Sachse, R., van Eupen, M., von Bloh, W., Zemp, D. e Thonicke, K. (2016) "Large-scale impact of climate change versus land-use change on future biome shifts in Latin America", *Global Change Biology*, p. 1–13. doi: 10.1111/gcb.13355.

Brune, A. (2014) "Symbiotic digestion of lignocellulose in termite guts", *Nature Reviews*

Microbiology. Nature Publishing Group, 12(3), p. 168–180. doi: 10.1038/nrmicro3182.

Bugg, T. D. H., Walsh, C. T., Ligase, D. e Enterococci, V. (1992) "lntracellular Steps of Bacterial Cell Wall Peptidoglycan Biosynthesis : Enzymology, Antibiotics, and Antibiotic Resistance", *Natural Product Reports*, (3), p. 199–215.

Cameron, R. E. (1960) "Arizona-Nevada Academy of Science Communities of Soil Algae Occurring in the Sonoran Desert in Arizona", *Journal of the Arizona Academy of Science*, 1(3), p. 85–88.

Cardenas, E., Kranabetter, J. M., Hope, G., Maas, K. R., Hallam, S. e Mohn, W. W. (2015) "Forest harvesting reduces the soil metagenomic potential for biomass decomposition", *The ISME Journal*. Nature Publishing Group, 9(11), p. 1–12. doi: 10.1038/ismej.2015.57.

Cardon, Z. G., Gray, D. W. e Lewis, L. A. (2008) "The Green Algal Underground: Evolutionary Secrets of Desert Cells", *BioScience*, 58(2), p. 114. doi: 10.1641/B580206.

Chinwuba (2014) *Temperate Deciduous Forest*. Disponível em: http://apbiologysses.blogspot.com.br/2014/08/primary-secondary-succession.html (Acessado: 19 de setembro de 2016).

Cline, L. C. e Zak, D. R. (2015) "Soil Microbial Communities are Shaped by Plant-Driven Changes in Resource Availability During Secondary Succession", *Ecological society of America*, 96(12), p. 3374–3385. doi: 10.1017/CBO9781107415324.004.

Costa, J., Empadinhas, N. e Da Costa, M. S. (2007) "Glucosylglycerate biosynthesis in the deepest lineage of the Bacteria: Characterization of the thermophilic proteins GpgS and GpgP from Persephonella marina", *Journal of Bacteriology*, 189(5), p. 1648–1654. doi: 10.1128/JB.00841-06.

Couger, M. B., Youssef, N. H., Struchtemeyer, C. G., Liggenstoffer, A. S. e Elshahed, M. S. (2015) "Transcriptomic analysis of lignocellulosic biomass degradation by the anaerobic fungal isolate Orpinomyces sp. strain C1A.", *Biotechnology for biofuels*. BioMed Central, 8, p. 208. doi: 10.1186/s13068-015-0390-0.

Cox, M. P., Peterson, D. A. e Biggs, P. J. (2010) "SolexaQA : At-a-glance quality assessment of Illumina second-generation sequencing data", *BMC bioinformatics*.

Cruz, M. (2015) *Biome*. Disponível em: https://www.haikudeck.com/biomes-education-presentation-DF2R6yxU1E (Acessado: 22 de setembro de 2016).

Delmont, T. O., Prestat, E., Keegan, K. P., Faubladier, M., Robe, P., Clark, I. M., Pelletier, E., Hirsch, P. R., Meyer, F., Gilbert, J. A., Le Paslier, D., Simonet, P. e Vogel, T. M. (2012) "Structure, fluctuation and magnitude of a natural grassland soil metagenome.", *The ISME journal*. Nature Publishing Group, 6(9), p. 1677–87. doi: 10.1038/ismej.2011.197.

Dinsdale, E. a, Edwards, R. a, Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. a e Rohwer, F. (2008) "Functional metagenomic profiling of nine biomes.", *Nature*, 452(7187), p. 629–632. doi: 10.1038/nature07346.

Edgar, R. C. (2010) "Search and clustering orders of magnitude faster than BLAST", *Bioinformatics*, 26(19), p. 2460–2461. doi: 10.1093/bioinformatics/btq461.

Eisner, T. T. (2011) *Fall Travel Update: Leaf-Peeping in Vermont*. Disponível em: http://wandermelon.com/tag/green-mountain-national-forest/ (Acessado: 22 de setembro de 2016).

Faridmoayer, A., Fentabil, M. A., Mills, D. C., Klassen, J. S. e Feldman, M. F. (2007) "Functional characterization of bacterial oligosaccharyltransferases involved in O-linked protein glycosylation", *Journal of Bacteriology*, 189(22), p. 8088–8098. doi: 10.1128/JB.01318-07.

Fatih Ozsolak (2013) "Third Generation Sequencing Techniques and Applications to Drug Discovery", *Expert Opin Drug Discov*, 7(3), p. 231–243. doi: 10.1517/17460441.2012.660145.Third.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J. e Huttenhower, C. (2012) "Microbial co-occurrence relationships in the Human Microbiome", *PLoS Computational Biology*, 8(7). doi: 10.1371/journal.pcbi.1002606.

Ferrari, B. (2015) *Microbial life in Antarctic polar desert soils*. Disponível em: http://www.science.unsw.edu.au/events/microbial-life-antarctic-polar-desert-soils (Acessado: 22 de setembro de 2016).

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A. e Knight, R. (2012) "Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients.", *The ISME Journal*. Nature Publishing Group, 6(5), p. 1007–17. doi: 10.1038/ismej.2011.159.

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H. e Caporaso, J. G. (2012) "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes", *Proceedings of the National Academy of Sciences*, 109(52), p. 21390–21395. doi: 10.1073/pnas.1215210110.

Finn, R. D., Clements, J. e Eddy, S. R. (2011) "HMMER web server: Interactive sequence similarity searching", *Nucleic Acids Research*, 39(SUPPL. 2), p. 29–37. doi: 10.1093/nar/gkr367.

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, A. e Bork, P. (2013) "Country-specific antibiotic use practices impact the human gut resistome", p. 1163–1169. doi: 10.1101/gr.155465.113.23.

George, Kathryn e Sarah (sem data) *The Taiga Biome*. Disponível em: http://weldytaigabiome.weebly.com/landforms.html (Acessado: 19 de setembro de 2016).

GoKutch (sem data) *Rann Utsav*. Disponível em: http://gokutch.com/ (Acessado: 22 de setembro de 2016).

Hammer, Ø., Harper, D. A. T. e Ryan, P. D. (2001) "Paleontological statistics software package for education and data analysis", *Palaeontologia Electronica*, 4(1), p. 9–18. doi: 10.1016/j.bcp.2008.05.025.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. e Goodman, R. M. (1998) "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.", *Chemistry & biology*, 5(10), p. R245–R249. doi: 10.1016/S1074-5521(98)90108-9.

Haroon, M. F., Thompson, L. R. e Stingl, U. (2016) "Draft Genome Sequence of Uncultured SAR324 Bacterium lautmerah10 , Binned from a Red Sea Metagenome", *Genome announcements*, 4(1), p. 1–2. doi: 10.1128/genomeA.01711-15.Copyright.

Hess, M., Sczyrba, A., Rob, E., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennachio, L. A., Tring, S. G., Visel, A., Woyle, T., Wang, Z. e Rudin, E. M. (2011) "Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen", *Science*, 463(6016), p. 463–467. doi: 10.1126/science.1200387.

Hongrittipun, P., Youpensuk, S. e Rerkasem, B. (2014) "Screening of Nitrogen Fixing Endophytic Bacteria in Oryza sativa L", *Journal of Agricultural Science*, 6(6), p. 66–74. doi: 10.5539/jas.v6n6p66.

Houghton, R. A., Hall, F. e Goetz, S. J. (2009) "Importance of biomass in the global carbon cycle", *Journal of Geophysical Research: Biogeosciences*, 114(3), p. 1–13. doi: 10.1029/2009JG000935.

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., Harden, J., Turetsky, M. R., McGuire, A. D., Shah, M. B., VerBerkmoes, N. C., Lee, L. H., Mavrommatis, K. e Jansson, J. K. (2015) "Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes", *Nature*. Nature Publishing Group, 521, p. 208–212. doi: 10.1038/nature14238.

Huson, D., Mitra, S. e Ruscheweyh, H. (2011) "Integrative analysis of environmental sequences using MEGAN4", *Genome Research*, 21(9), p. 1552–1560. doi: 10.1101/gr.120618.111.Freely.

Jackman, J. E., Fierke, C. a, Tumey, L. N., Pirrung, M., Uchiyama, T., Tahir, S. H., Hindsgaul, O., Raetz, C. R. H., Of, I. e Deacetylases, D. U.--O. R. N. (2000) "Antibacterial Agents That Target Lipid A Biosynthesis in Gram-negative Bacteria", *The Journal of biological chemistry*, 275(15), p. 11002–11009.

Jarrell, K. F., Ding, Y., Meyer, B. H., Albers, S.-V., Kaminski, L. e Eichler, J. (2014) "N-Linked Glycosylation in Archaea: a Structural, Functional, and Genetic Analysis.", *Microbiology and molecular biology reviews : MMBR*, 78(2), p. 304–341. doi: 10.1128/MMBR.00052-13.

Jiang, L., Lin, M., Zhang, Y., Li, Y., Xu, X., Li, S. e He Huang (2013) "Identification and Characterization of a Novel Trehalose Synthase Gene Derived from Saline-Alkali Soil Metagenomes", *PLoS ONE*, 8(10), p. 1–11. doi: 10.1371/journal.pone.0077437.

John Muir School (sem data) *Chaparral - Native plants of San Diego*. Disponível em: https://sites.google.com/site/nativeplantsofsandiego/biomes/chaparral (Acessado: 21 de outubro de 2016).

Kamilova, F., Kravchenko, L. V., Shaposhnikov, A. I., Makarova, N. e Lugtenberg, B. (2006) "Effects of the Tomato Pathogen Fusarium oxysporum f. sp. radicis-lycopersici and of the Biocontrol Bacterium Pseudomonas fluorescens WCS365 on the Composition of Organic Acids and Sugars in Tomato Root Exudate", *Molecular Plant-Microbe Interactions*, 19(10), p. 1121–1126. doi: 10.1094/MPMI-19-1121.

Kanehisa, M. e Goto, S. (2000) "Yeast Biochemical Pathways. KEGG: Kyoto encyclopedia of genes and genomes", *Nucleic Acids Res*, 28(1), p. 27–30. doi: 10.1093/nar/28.1.27.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. e Tanabe, M. (2016) "KEGG as a reference resource for gene and protein annotation", *Nucleic Acids Research*, 44(D1), p. D457–D462. doi: 10.1093/nar/gkv1070.

Keegan, K. P., Trimble, W. L., Wilkening, J., Wilke, A., Harrison, T., Souza, M. D. e Meyer, F. (2012) "A Platform-Independent Method for Detecting Errors in Metagenomic Sequencing Data : DRISEE", *PLoS Computational Biology*, 8(6). doi: 10.1371/journal.pcbi.1002541.

Kelly, W. L., Pan, L. e Li, C. (2009) "Thiostrepton biosynthesis: Prototype for a new family of

bacteriocins", *Journal of the American Chemical Society*, 131(12), p. 4327–4334. doi: 10.1021/ja807890a.

Kim, H. S., Lee, C. e Lee, E. Y. (2011) "Alginate Lyase : Structure , Property , and Application", *Biotechnology and Bioprocess Engineering*, 851, p. 843–851. doi: 10.1007/s12257-011-0352-8.

Langmead, B., Trapnell, C., Pop, M. e Salzberg, S. L. (2009) "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome Biology*, 10(3). doi: 10.1186/gb-2009-10-3-r25.

LaPointe, A., Gibbons, S. M., Frazier, A., Hampton-Marcell, J. e Gilbert, J. (2015) "Aquarium microbiome response to ninety-percent system water change: Clues to microbiome management", *Zoo Biology*, 34(4), p. 360–367. doi: 10.1002/zoo.21220.

Lauber, C. L., Hamady, M., Knight, R. e Fierer, N. (2009) "Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale", *Applied and Environmental Microbiology*, 75(15), p. 5111–5120. doi: 10.1128/AEM.00335-09.

Lee, C. e Lee, J.-H. (2013) "Lipid A Biosynthesis of Multidrug-Resistant Pathogens-A Novel Drug Target", *Current pharmaceutical design*, p. 1–17. doi: 10.2174/13816128113199990494.

Lee, S., Lee, J. Y., Ha, S. C., Jung, J., Shin, D. H., Kim, K. H. e Choi, I. G. (2009) "Crystallization and preliminary X-ray analysis of neoagarobiose hydrolase from Saccharophagus degradans 2-40", *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 65(12), p. 1299–1301. doi: 10.1107/S174430910904603X.

Leisner, J. J., Jørgensen, N. O. G. e Middelboe, M. (2016) "Predation and selection for antibiotic resistance in natural environments", *Evolutionary Applications*, 9(3), p. 427–434. doi: 10.1111/eva.12353.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. e Henrissat, B. (2013) "Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes.", *Biotechnology for biofuels*, 6(1), p. 41. doi: 10.1186/1754-6834-6-41.

Lewis, L. A. e Lewis, P. O. (2005) "Unearthing the Molecular Phylodiversity of Desert Soil Green Algae (Chlorophyta)", *Syst. Biol*, 54(6), p. 936–947. doi: 10.1080/10635150500354852.

Li, B., Yang, Y., Ma, L., Ju, F., Guo, F., Tiedje, J. M. e Zhang, T. (2015) "Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes", *The ISME Journal*. Nature Publishing Group, 9(11), p. 1–13. doi: 10.1038/ismej.2015.59.

Li, Z. e Srivastava, P. (2004) "Heat-shock proteins.", *Current protocols in immunology / edited by John E. Coligan ... [et al.]*, Appendix 1, p. Appendix 1T. doi: 10.1002/0471142735.ima01ts58.

Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., Mueller, A., Hughes, D. E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. A., Cohen, D. R., Felix, C. R., Fetterman, K. A., Millett, W. P., Nitti, A. G., Zullo, A. M., Chen, C. e Lewis, K. (2015) "A new antibiotic kills pathogens without detectable resistance", *Nature*, 0. doi: 10.1038/nature14098.

Liu, B. e Pop, M. (2009) "ARDB - Antibiotic resistance genes database", *Nucleic Acids Research*, 37(SUPPL. 1), p. 443–447. doi: 10.1093/nar/gkn656.

Llorens-Marès, T., Yooseph, S., Goll, J., Hoffman, J., Vila-Costa, M., Borrego, C. M., Dupont, C. L. e Casamayor, E. O. (2015) "Connecting biodiversity and potential functional role in modern euxinic

environments by microbial metagenomics", *The ISME journal*, 9(7), p. 1648–1661. doi: 10.1038/ismej.2014.254.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. e Henrissat, B. (2014) "The carbohydrate-active enzymes database (CAZy) in 2013", *Nucleic Acids Research*, 42(D1), p. 490–495. doi: 10.1093/nar/gkt1178.

Loo, Melanie (2010). "Biomes" . BIO 7: Lecture 31-32 Preview, California State University (Sacramento State). Disponível em: http://www.csus.edu/indiv/l/loom/lect%2031-32%20s07.htm (Acessado: 16 de setembro de 2016).

Louro, L. (sem data) *Biomas de América*. Disponível em: https://br.pinterest.com/pin/480055641502844420/ (Acessado: 20 de setembro de 2016).

Ma, B., Wang, H., Dsouza, M., Lou, J., He, Y., Dai, Z., Brookes, P. C., Xu, J. e Gilbert, J. A. (2016) "Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China.", *The ISME journal*. Nature Publishing Group, p. 1–11. doi: 10.1038/ismej.2015.261.

Ma, L., Xia, Y., Li, B., Yang, Y., Li, L. G., Tiedje, J. M. e Zhang, T. (2016) "Metagenomic Assembly Reveals Hosts of Antibiotic Resistance Genes and the Shared Resistome in Pig, Chicken, and Human Feces", *Environmental Science and Technology*, 50(1), p. 420–427. doi: 10.1021/acs.est.5b03522.

Makhalanyane, T. P., Valverde, A., Gunnigle, E., Frossard, A., Ramond, J. B. e Cowan, D. A. (2015) "Microbial ecology of hot desert edaphic systems", *FEMS microbiology reviews*, 39(2), p. 203–221. doi: 10.1093/femsre/fuu011.

Manoharan, L., Kushwaha, S. K., Hedlund, K. e Ahrén, D. (2015) "Captured metagenomics: Large-scale targeting of genes based on 'sequence capture' reveals functional diversity in soils", *DNA Research*, 22(6), p. 451–460. doi: 10.1093/dnares/dsv026.

Martínez, J. L. (2008) "Antibiotics and antibiotic resistance genes in natural environments.", *Science*, 321(5887), p. 365–367. doi: 10.1126/science.1159483.

Melo, P. (2014) *Caatinga*. Disponível em: https://www.estudokids.com.br/caatinga/ (Acessado: 23 de setembro de 2016).

Mendes, L. W., Tsai, S. M., Navarrete, A. A., De Hollander, M., Van Veen, J. A. e Kuramae, E. E. (2015) "Soil-Borne Microbiome: Linking Diversity to Function", *Microbial Ecology*. doi: 10.1007/s00248-014-0559-2.

Metzger, L. E., Lee, J. K., Finer-Moore, J. S., Raetz, C. R. H. e Stroud, R. M. (2012) "LpxI structures reveal how a lipid A precursor is synthesized.", *Nature structural & molecular biology*. Nature Publishing Group, 19(11), p. 1132–8. doi: 10.1038/nsmb.2393.

Meyer, F., Paarmann, D., D'Souza, M. e Etal. (2008) "The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes", *BMC bioinformatics*, 9, p. 386. doi: 10.1186/1471-2105-9-386.

Mhuantong, W., Charoensawan, V., Kanokratana, P., Tangphatsornruang, S. e Champreda, V. (2015) "Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass-degrading enzymes among lignocellulolytic microbial communities.", *Biotechnology for biofuels*, 8, p. 16. doi: 10.1186/s13068-015-0200-8.

Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., Hoopen, P. ten, Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., Sterk, P. e Finn*, R. D. (2016) "EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data", *Nucleic Acids Research*, 44(D1), p. D595–D603. doi: 10.1093/nar/gkv1195.

Moore, C. W., McKoy, J., Del Valle, R., Armstrong, D., Bernard, E. M., Katz, N. e Gordon, R. E. (2003) "Fungal cell wall septation and cytokinesis are inhibited by bleomycins", *Antimicrobial Agents and Chemotherapy*, 47(10), p. 3281–3289. doi: 10.1128/AAC.47.10.3281-3289.2003.

Mori, T., Mizuta, S., Suenaga, H. e Miyazaki, K. (2008) "Metagenomic screening for bleomycin resistance genes", *Applied and Environmental Microbiology*, 74(21), p. 6803–6805. doi: 10.1128/AEM.00873-08.

Mosier, A. C., Li, Z., Thomas, B. C., Hettich, R. L., Pan, C. e Banfield, J. F. (2014) "Elevated temperature alters proteomic responses of individual organisms within a biofilm community.", *The ISME journal*. Nature Publishing Group, 9(1), p. 1–15. doi: 10.1038/ismej.2014.113.

Navarrete, A. A., Tsai, S. M., Mendes, L. W., Faust, K., De Hollander, M., Cassman, N. A., Raes, J., Van Veen, J. A. e Kuramae, E. E. (2015) "Soil microbiome responses to the short-term effects of Amazonian deforestation", *Molecular Ecology*, 24(10), p. 2433–2448. doi: 10.1111/mec.13172.

Nesme, J., C??cillon, S., Delmont, T. O., Monier, J. M., Vogel, T. M. e Simonet, P. (2014) "Large-scale metagenomic-based study of antibiotic resistance in the environment", *Current Biology*, 24(10), p. 1096–1100. doi: 10.1016/j.cub.2014.03.036.

Nothaft, H. e Szymanski, C. M. (2010) "Protein glycosylation in bacteria: sweeter than ever", *Nat. Rev. Microbiol.* Nature Publishing Group, 8(11), p. 765–778. doi: 10.1038/nrmicro2383.

O'Brien, S. L., Gibbons, S. M., Owens, S. M., Hampton-Marcell, J., Johnston, E. R., Jastrow, J. D., Gilbert, J. A., Meyer, F. e Antonopoulos, D. A. (2015) "Spatial scale drives patterns in soil bacterial diversity", *Environmental Microbiology*, In review(May). doi: 10.1111/1462-2920.13231.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rül;ckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. e Vonstein, V. (2005) "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes", *Nucleic Acids Research*, 33(17), p. 5691–5702. doi: 10.1093/nar/gki866.

Pandit, A. S., Joshi, M. N., Bhargava, P., Ayachit, G. N., Shaikh, I. M., Saiyed, Z. M., Saxena, A. K. e Bagatharia, S. B. (2014) "Metagenomes from the Saline Desert of Kutch", *Genome Announcements*, 2(3), p. e00439-14-e00439-14. doi: 10.1128/genomeA.00439-14.

Pareek, S., Mathur, N., Singh, A. e Nepalia, A. (2015) "Review Article Antibiotics in the Environment : A Review", 4(11), p. 278–285.

Park, C. e Seo, Y. (2015) "Heat Shock Proteins : A Review of the Molecular Chaperones for Plant Immunity", 31(4), p. 323–333. doi: 10.5423/PPJ.RW.08.2015.0150.

Parks, D. H., Tyson, G. W., Hugenholtz, P. e Beiko, R. G. (2014) "STAMP: Statistical analysis of taxonomic and functional profiles", *Bioinformatics*, 30(21), p. 3123–3124. doi:

10.1093/bioinformatics/btu494.

Parsell, D. A., Kowal, A. S., Singer, M. A. e Lindquist, S. (1994) "Protein disaggregation mediated by heat-shock protein Hsp104.", *Nature*, p. 475–478. doi: 10.1038/372475a0.

Pasternak, Z., Al-Ashhab, A., Gatica, J., Gafny, R., Avraham, S., Minz, D., Gillor, O. e Jurkevitch, E. (2013) "Spatial and Temporal Biogeography of Soil Microbial Communities in Arid and Semiarid Regions", *PLoS ONE*, 8(7). doi: 10.1371/journal.pone.0069705.

Peng, M., Zi, X. e Wang, Q. (2015) "Bacterial community diversity of oil-contaminated soils assessed by high throughput sequencing of 16s rRNA genes", *International Journal of Environmental Research and Public Health*, 12(10), p. 12002–12015. doi: 10.3390/ijerph121012002.

Perron, G. G., Whyte, L., Turnbaugh, P. J., Goordial, J., Hanage, W. P., Dantas, G. e Desai, M. M. (2015) "Functional characterization of bacteria isolated from ancient arctic soil exposes diverse resistance mechanisms to modern antibiotics", *PLoS ONE*, 10(3), p. 1–19. doi: 10.1371/journal.pone.0069533.

Plantier, R. D. (2013) *Vegetação do Cerrado: Formação e Características Principais*. Disponível em: http://meioambiente.culturamix.com/natureza/vegetacao-do-cerrado-formacao-e-caracteristicas-principais (Acessado: 22 de setembro de 2016).

Prairie (sem data). Disponível em: https://en.wikipedia.org/wiki/Prairie (Acessado: 22 de setembro de 2016).

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, S., Manichanh, C., Nielsen, T., Pons, N., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J., Hansen, T., Paslier, D. Le, Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N. e Yang, H. (2010) "A human gut microbial gene catalog established by metagenomic sequencing", *Nature*, 464(7285), p. 59–65. doi: 10.1038/nature08821.A.

Quail, M. M., Smith, M. E., Coupland, P., Otto, T. D. T., Harris, S. R. S., Connor, T. R., Bertoni, A., Swerdlow, H. P. H. H. P., Gu, Y., Rothberg, J., Hinz, W., Rearick, T., Schultz, J., Mileski, W., Davey, M., Leamon, J., Johnson, K., Milgrew, M., Edwards, M., Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bentley, D., Balasubramanian, S., Swerdlow, H. P. H. H. P., Smith, G., Milton, J., Brown, C., Hall, K., Evers, D., Barnes, C., Bignell, H., Kozarewa, I., Ning, Z., Quail, M. M., Sanders, M., Berriman, M., Turner, D., Quail, M. M., Otto, T. D. T., Gu, Y., Harris, S. R. S., Skelly, T., McQuillan, J., Swerdlow, H. P. H. H. P., Oyola, S., Syed, F., Grunenwald, H., Caruccio, N., Lam, H., Clark, M., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F., Habegger, L., Carver, T., Harris, S. R. S., Berriman, M., Parkhill, J., McQuillan, J., Ponsting, N., Ning, Z., Otto, T. D. T., Sanders, M., Berriman, M., Newbold, C., Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M., Hirai, A., Takahashi, H., Diep, B., Gill, S., Chang, R., Phan, T., Chen, J., Davidson, M., Lin, F., Lin, J., Carleton, H., Mongodin, E., Achidi, E., Gardner, M., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R., Carlton, J., Pain, A., Nelson, K., Bowman, S., Choi, M., Scholl, U., Ji, W., Liu, T., Tikhonova, I., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Down, T., Rakyan, V., Turner, D., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E., Giresi, P., Kim, J., McDaniell, R., Iyer, V., Lieb, J., Johnson, D., Mortazavi, A., Myers, R., Wold, B., Langridge, G., Phan, M., Turner, D., Perkins, T., Parts, L., Haase, J., Charles, I., Maskell, D., Peters, S., Dougan, G., Licatalosi, D., Mele, A., Fak, J., Ule, J., Kayikci, M., Chi, S., Clark, T., Schweitzer, A., Blume, J., Wang, X., Mamanova, L., Andrews, R., James, K., Sheridan, E., Ellis, P., Langford, C., Ost, T., Collins, J., Turner, D., Myllykangas, S., Buenrostro, J., Natsoulis, G., Bell, J., Ji, H., Shao, N., Hu, H., Yan, Z., Xu, Y., Hu, H.,

Menzel, C., Li, N., Chen, W., Khaitovich, P., Wang, Z., Gerstein, M., Snyder, M., Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F., Burton, J., Walker, B., Sharpe, T., Hall, G., Shea, T., Sykes, S., Levin, J., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D., Friedman, N., Gnirke, A., Regev, A., Adey, A., Asan, N., Xun, X., Kitzman, J., Turner, E., Stackhouse, B., MacKenzie, A., Caruccio, N., Zhang, X., Flusberg, B., Webster, D., Lee, J., Travers, K., Olivares, E., Clark, T., Korlach, J., Turner, S., Holden, T., Lindsay, J., Corton, C., Quail, M. M., Cockfield, J., Pathak, S., Batra, R., Parkhill, J., Bentley, S., Edgeworth, J., Li, H., Durbin, R., Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Angiuoli, S. e Salzberg, S. (2012) "A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers", *BMC Genomics*, 13(1), p. 1. doi: 10.1186/1471-2164-13-341.

Ranjard, L., Dequiedt, S., Chemidlin Prévost-Bouré, N., Thioulouse, J., Saby, N. P. a, Lelievre, M., Maron, P. a, Morin, F. E. R., Bispo, a, Jolivet, C., Arrouays, D. e Lemanceau, P. (2013) "Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity.", *Nature communications*, 4, p. 1434. doi: 10.1038/ncomms2431.

Ratheesh Kumar, R., Nagarajan, N. S., Arunraj, S. P., Sinha, D., Veedin Rajan, V. B., Esthaki, V. K. e D'Silva, P. (2012) "HSPIR: A manually annotated heat shock protein information resource", *Bioinformatics*, 28(21), p. 2853–2855. doi: 10.1093/bioinformatics/bts520.

Raynaud, X. e Nunan, N. (2014) "Spatial Ecology of Bacteria at the Microscale in Soil", *PLoS ONE*, 9(1). doi: 10.1371/journal.pone.0087217.

Reina-Bueno, M., Argandoña, M., Salvador, M., Rodríguez-Moya, J., Iglesias-Guerra, F., Csonka, L. N., Nieto, J. J. e Vargas, C. (2012) "Role of trehalose in salinity and temperature tolerance in the model halophilic bacterium chromohalobacter salexigens", *PLoS ONE*, 7(3). doi: 10.1371/journal.pone.0033587.

Resolute forest products (2014) *Boreal forest facts*. Disponível em: http://borealforestfacts.com/?p=234 (Acessado: 22 de setembro de 2016).

Rho, M., Tang, H. e Ye, Y. (2010) "FragGeneScan: Predicting genes in short and error-prone reads", *Nucleic Acids Research*, 38(20), p. 1–12. doi: 10.1093/nar/gkq747.

Riesenfeld, C. S., Goodman, R. M. e Handelsman, J. (2004) "Uncultured soil bacteria are a reservoir of new antibiotic resistance genes", *Environmental Microbiology*, 6(9), p. 981–989. doi: 10.1111/j.1462-2920.2004.00664.x.

Roesch, L., Fulthorpe, R., Riva, A., Casella, G., Hadwin, A., Kent, A., Daroub, S., Camargo, F., Farmerie, W. e Triplett, E. (2007) "Pyrosequencing enumerates and contrasts soil microbial diversity", *The ISME Journal*, 1(4), p. 283–290. doi: 10.1038/ismej.2007.53.

Sangwan, N., Zarraonaindia, I., Hampton-marcell, J. T., Ssegane, H., Eshoo, T. W., Rijal, G. M., Negri, C. e Gilbert, J. A. (2016) "Differential Functional Constraints Cause Strain-Level Endemism in", *Ecological and Evolutionary Science*, p. 1–11. doi: 10.1128/mSystems.00003-16.Editor.

Sanli, K., Bengtsson-Palme, J., Henrik Nilsson, R., Kristiansson, E., Rosenblad, M. A., Blanck, H. e Eriksson, K. M. (2015) "Metagenomic sequencing of marine periphyton: Taxonomic and functional insights into biofilm communities", *Frontiers in Microbiology*, 6(OCT), p. 1–14. doi: 10.3389/fmicb.2015.01192.

SAS Institute Inc (sem data) "JMP® statistical discovery software, Version 12.2.0".

Schloss, P. D., Allen, H. K., Klimowicz, A. K., Mlot, C., Gross, J. a, Savengsuksa, S., McEllin, J., Clardy, J., Ruess, R. W. e Handelsman, J. (2010) "Psychrotrophic strain of Janthinobacterium lividum from a cold Alaskan soil produces prodigiosin.", *DNA and cell biology*, 29(9), p. 533–541. doi: 10.1089/dna.2010.1020.

Schmidt, T. M., DeLong, E. F. e Pace, N. R. (1991) "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing", *Journal of Bacteriology*, 173(14), p. 4371–4378.

Segawa, T., Takeuchi, N., Rivera, A., Yamada, A., Yoshimura, Y., Barcaza, G., Shinbori, K., Motoyama, H., Kohshima, S. e Ushida, K. (2013) "Distribution of antibiotic resistance genes in glacier environments", *Environmental Microbiology Reports*, 5(1), p. 127–134. doi: 10.1111/1758-2229.12011.

Shannon, P. (2003) "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", *Genome Research*, 13(11), p. 2498–2504. doi: 10.1101/gr.1239303.

Shi, Y., Grogan, P., Sun, H., Xiong, J., Yang, Y., Zhou, J. e Chu, H. (2015) "Multi-scale variability analysis reveals the importance of spatial distance in shaping Arctic soil microbial functional communities", *Soil Biology and Biochemistry*. Elsevier Ltd, 86, p. 126–134. doi: 10.1016/j.soilbio.2015.03.028.

Shi, Y., Xiang, X., Shen, C., Chu, H., Neufeld, J. D., Walker, V. K. e Grogan, P. (2015) "Vegetation-associated impacts on Arctic tundra bacterial and microeukaryotic communities", *Applied and Environmental Microbiology*, 81(2), p. 492–501. doi: 10.1128/AEM.03229-14.

Shields, M. J., Fischer, J. J. e Wieden, H. J. (2009) "Toward understanding the function of the universally conserved GTPase HflX from Escherichia coli: A kinetic approach", *Biochemistry*, 48(45), p. 10793–10802. doi: 10.1021/bi901074h.

Shukla, H. D. (2006) "Proteomic analysis of acidic chaperones, and stress proteins in extreme halophile Halobacterium NRC-1: a comparative proteomic approach to study heat shock response.", *Proteome science*, 4, p. 6. doi: 10.1186/1477-5956-4-6.

Silvey, B. (2009) *Scenery of the American West*. Disponível em: http://www.billiesilvey.com/scenery.html (Acessado: 22 de setembro de 2016).

Snyder, P. K., Delire, C. e Foley, J. A. (2004) "Evaluating the influence of different vegetation biomes on the global climate", *Climate Dynamics*, 23(3–4), p. 279–302. doi: 10.1007/s00382-004-0430-0.

Souza, R. C., Mendes, I. C., Reis-Junior, F. B., Carvalho, F. M., Nogueira, M. A., Vasconcelos, A. T. R., Vicente, V. A. e Hungria, M. (2016) "Shifts in taxonomic and functional microbial diversity with agriculture: How fragile is the Brazilian Cerrado?", *BMC Microbiology*. BMC Microbiology, 16(1), p. 42. doi: 10.1186/s12866-016-0657-z.

Susin, M. F., Baldini, R. L., Gueiros-Filho, F. e Gomes, S. L. (2006) "GroES/GroEL and DnaK/DnaJ have distinct roles in stress responses and during cell cycle progression in Caulobacter crescentus", *Journal of Bacteriology*, 188(23), p. 8044–8053. doi: 10.1128/JB.00824-06.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. e Koonin, E. V (2000) "The COG database: a tool for genome-scale analysis of protein functions and evolution.", *Nucleic acids research*, 28(1), p. 33–36. doi: 10.1093/nar/28.1.33.

The nature conservancy (sem data) *Patagonia's important grasslands*. Disponível em: http://www.nature.org/ourinitiatives/regions/southamerica/argentina/howwework/ (Acessado: 21 de

setembro de 2016).

The Osprey Brand Team (2009) *Beautiful view of the desert*. Disponível em: http://www.ospreypacks.com/stories/rendezvous-in-the-mojave-climbing-camping-and-clinics/ (Acessado: 20 de setembro de 2016).

Thompson, C. E., Beys-da-Silva, W. O., Santi, L., Berger, M., Vainstein, M. H., Guima Rães, J. A. e Vasconcelos, A. T. R. (2013) "A potential source for cellulolytic enzyme discovery and environmental aspects revealed through metagenomics of Brazilian mangroves.", *AMB Express*, 3(1), p. 65. doi: 10.1186/2191-0855-3-65.

Torsvik, V., Goksyr, J. e Daae, F. L. (1990) "High Diversity in DNA of Soil Bacteria", *Applied and Environmental Microbiology*, 56(3), p. 782–787.

Varin, T., Lovejoy, C., Jungblut, A. D., Vincent, W. F. e Corbeil, J. (2012) "Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the high Arctic", *Applied and Environmental Microbiology*, 78(2), p. 549–559. doi: 10.1128/AEM.06354-11.

Velkov, T., Thompson, P. E., Nation, R. L. e Li, J. (2010) "Structure-activity relationships of polymyxins antibiotics", *Journal of Medicinal Chemistry*, 53(5), p. 1898–1916. doi: 10.1021/jm900999h.Structure.

Virtanen, R., Oksanen, L., Oksanen, T., Cohen, J., Forbes, B. C., Johansen, B., K??yhk??, J., Olofsson, J., Pulliainen, J. e T??mmervik, H. (2015) "Where do the treeless tundra areas of northern highlands fit in the global biome system: Toward an ecologically natural subdivision of the tundra biome", *Ecology and Evolution*, p. 143–158. doi: 10.1002/ece3.1837.

Walmagh, M., Zhao, R. e Desmet, T. (2015) "Trehalose analogues: Latest insights in properties and biocatalytic production", *International Journal of Molecular Sciences*, 16(6), p. 13729–13745. doi: 10.3390/ijms160613729.

Wang, K., Zhang, X., Goatley, M. e Ervin, E. (2014) "Heat shock proteins in relation to heat stress tolerance of creeping bentgrass at different N levels", *PLoS ONE*, 9(7), p. 1–10. doi: 10.1371/journal.pone.0102914.

Wang, W., Guo, Q., Xu, X., Sheng, Z. K., Ye, X. e Wang, M. (2014) "High-level tetracycline resistance mediated by efflux pumps Tet(A) and Tet(A)-1 with two start codons", *Journal of Medical Microbiology*, 63(May), p. 1454–1459. doi: 10.1099/jmm.0.078063-0.

Wang, Z., Zhang, X.-X. X., Huang, K. L., Miao, Y., Shi, P., Liu, B., Long, C. e Li, A. M. (2013) "Metagenomic Profiling of Antibiotic Resistance Genes and Mobile Genetic Elements in a Tannery Wastewater Treatment Plant", *PloS one*, 8(10), p. e76079. doi: 10.1371/journal.pone.0076079.

Wenk, M., Ba, Q., Erichsen, V., MacInnes, K., Wiese, H., Warscheid, B. e Koch, H. G. (2012) "A universally conserved ATPase regulates the oxidative stress response in Escherichia coli", *Journal of Biological Chemistry*, 287(52), p. 43585–43598. doi: 10.1074/jbc.M112.413070.

Wilke, A., Glass, E. M., Bischof, J., Braithwaite, D., Harrison, T., Keegan, K., Paczian, T., Trimble, W. L. e Meyer, F. (2015) *MG-RAST Manual for version 3.6*. Disponível em: ftp://ftp.metagenomics.anl.gov/data/manual/mg-rast-manual.pdf.

Woese, C. R. (1987) "Bacterial evolution.", *Microbiological reviews*, 51(2), p. 221–71. Disponível

em: http://www.ncbi.nlm.nih.gov/pubmed/2439888.

Xiao, K.-Q., Li, B., Ma, L., Bao, P., Zhou, X., Zhang, T. e Zhu, Y.-G. (2016) "Metagenomic profiles of antibiotic resistance genes (ARGs) in paddy soils from South China", *FEMS Microbiology Ecology*. doi: 10.1093/femsec/fiw023.

Xu, Z., Hansen, M. A., Hansen, L. H., Jacquiod, S. e Sørensen, S. J. (2014) "Bioinformatic approaches reveal metagenomic characterization of soil microbial community", *PLoS ONE*, 9(4). doi: 10.1371/journal.pone.0093445.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F. e Xu, Y. (2012) "DbCAN: A web resource for automated carbohydrate-active enzyme annotation", *Nucleic Acids Research*, 40(W1), p. 445–451. doi: 10.1093/nar/gks479.

Yoshii, A., Moriyama, H. e Fukuhara, T. (2012) "The novel kasugamycin 2'-N-acetyltransferase gene aac(2')-IIa, carried by the IncP Island, confers kasugamycin resistance to rice-pathogenic bacteria", *Applied and Environmental Microbiology*, 78(16), p. 5555–5564. doi: 10.1128/AEM.01155-12.

Zhang, Y. Z., Cheng, Y. W., Ya, H. Y., Han, J. M. e Zheng, L. (2015) "Identification of heat shock proteins via transcriptome profiling of tree peony leaf exposed to high temperature", *Genetics and Molecular Research*, 14(3), p. 8421–8442. doi: 10.4238/2015.July.28.10.

Zhou, Y., Pope, P. B., Li, S., Wen, B., Tan, F., Cheng, S., Chen, J., Yang, J., Liu, F., Lei, X., Su, Q., Zhou, C., Zhao, J., Dong, X., Jin, T., Zhou, X., Yang, S., Zhang, G., Yang, H., Wang, J., Yang, R., Eijsink, V. G. H. e Wang, J. (2014) "Omics-based interpretation of synergism in a soil-derived cellulose-degrading microbial community", *Scientific Reports*, 4, p. 1–6. doi: 10.1038/srep05288.

Zhu, Y.-G., Johnson, T. A., Su, J.-Q., Qiao, M., Guo, G.-X., Stedtfeld, R. D., Hashsham, S. A. e Tiedje, J. M. (2013) "Diverse and abundant antibiotic resistance genes in Chinese swine farms.", *Proceedings of the National Academy of Sciences of the United States of America*, 110(9), p. 3435–40. doi: 10.1073/pnas.1222743110.

ANEXOS

Profa. Dra. Rachel Meneguello
Presidente
Comissão Central de Pós-Graduação
Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Metagenômica comparativa de comunidades microbianas de solos de biomas globais**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 31/10/2016

Assinatura : Melline Fontes Noronha
Nome do(a) autor(a): **Melline Fontes Noronha**
RG n.° 30209862

Assinatura : Valéria Merzel
Nome do(a) orientador(a): **Valéria Maia Merzel**
RG n.° 12573419

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

DECLARAÇÃO

Em observância ao **§4º do Artigo 1º da Informação CCPG-UNICAMP/002/13**, de 14/08/2013, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado intitulada "***Metagenômica comparativa de comunidades microbianas de solos de biomas globais***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: Melline Fontes Noronha
Nome do(a) aluno(a): Melline Fontes Noronha

Assinatura: Valéria Maia Merzel
Nome do(a) orientador(a): Valéria Maia Merzel

Data: 15/12/2016