UNIVERSIDADE FEDERAL DO PARANÁ

JANYNNE STEPHANIE DE OLIVEIRA PALHETA

Análise de sequências de DNA Metagenômico de solos da Floresta Atlântica Paranaense: Implicações ecológicas e biotecnológicas

CURITIBA

2017

JANYNNE STEPHANIE DE OLIVEIRA PALHETA

ANÁLISE DE SEQUÊNCIAS DE DNA METAGENÔMICO DE SOLOS DA FLORESTA ATLÂNTICA PARANAENSE: IMPLICAÇÕES ECOLÓGICAS E BIOTECNOLÓGICAS

Dissertação de Mestrado apresentada como requisito parcial à obtenção do grau de mestre em Bioinformática ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná.

Orientador: Dr. Helisson Faoro

CURITIBA

2017

P161 Palheta, Janynne Stephanie de Oliveira
Análise de sequências de DNA Metagenômico de solos da Floresta Atlântica Paranaense: Implicações ecológicas e biotecnológicas/ Janynne Stephanie de Oliveira Palheta. -- Curitiba, 2017.
83 f.: il.

Orientador: Prof. Dr. Helisson Faoro.
Dissertação (Mestrado) - Universidade Federal do Paraná. Setor de Educação Profissional e Tecnológica. Programa de Pós-Graduação em Bioinformática.

Inclui Referências.

1. Metagenômica. 2. Floresta Atlântica. 3. DNA de solo. 4. Sequenciamento NGS. I. Faoro, Helisson. II. Título. III. Universidade Federal do Paraná.

CDD 581.15

TERMO DE APROVAÇÃO

JANYNNE STEPHANIE DE OLIVEIRA PALHETA

"Análise de sequências de DNA Metagenômico de solos da Floresta Atlântica Paranaense: Implicações ecológicas e biotecnológicas"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Dr. Helisson Faoro
Programa de Pós-graduação em Bioinformática – UFPR
Instituto Carlos Chagas – FIOCRUZ/PR

Dr. Eduardo Balsanelli
Universidade Federal do Paraná - UFPR

Drª Lysangela Ronalte Alves
Instituto Carlos Chagas – FIOCRUZ/PR

Curitiba, 24 de fevereiro de 2017

À minha família, que sempre acreditou.

AGRADECIMENTOS

A Deus que guiou e iluminou os meus passos e me permitiu chegar até aqui.

Ao meu orientador, Dr. Helisson Faoro, quem sugeriu o projeto e me guiou em todos os passos, me dando suporte e apoio, sempre sendo paciente e me ajudando nas dificuldades e dúvidas. Obrigada por toda a atenção e por todos os "não fica nervosa", mesmo que eu continue ficando.

À Profa. Dra. Maria Berenice R. Steffens, que foi uma pessoa importantíssima para que eu chegasse até aqui. Obrigada por me estender a mão quando mais precisei e por me apoiar. Eu realmente lhe sou muito grata e nunca esquecerei o que fez por mim. Ainda lembro da senhora me repreendendo para não sofrer sozinha e pedir ajuda. Estou tentando, professora!

À Suzana Gobetti, que foi a primeira pessoa com quem tive contato no PPGBIOINFO e que me cativou com um único e-mail. Obrigada por sempre ser uma pessoa maravilhosa e me receber e acolher com alegria. Sempre saía da sua sala com o espírito mais leve. Realmente obrigada!

À Dra. Michelle Zibetti Tadra-Sfeir, quem auxiliou na obtenção dos dados de sequenciamento para o desenvolvimento deste trabalho.

À minha avó, Raimunda Sousa, e minha tia, Nazaré Oliveira, que são mais que avó e tia, são mães que me amam e cuidam mesmo que de longe, dando-me todo o suporte e apoio e estando sempre comigo, mesmo que seja só para me ouvir chorar ao telefone. Obrigada por todo o apoio e confiança. Amo vocês!

Ao meu irmão, Apollo Palheta, meu confidente, conselheiro, psicólogo e levantador de ânimo, meu tudo. Aquele que me faz enxergar quem sou sempre que estou duvidando de tudo e de mim mesma. Muito obrigada por estar sempre comigo. Eu amo você!

À minha irmã, Jackline Palheta, o anjo que Deus enviou para mim e que foi a pessoa que mais doeu me despedir antes de vir para Curitiba. A mana te ama, você não imagina o quanto!

Às minhas amigas, Priscila Ferreira e Cintia da Silva, que possuem personalidades tão diferentes, mas que tocam o meu coração da mesma forma. Sempre com amor, amizade e confiança. Obrigada por estarem comigo em todos os momentos, compartilhando as *bads* e os surtos e reclamando da vida. *Saranghae*!

Às minhas amigas, Lorenna Pinheiro, Erika Farias e Tuane Oliveira, minhas BD's!, que me julgam mais inteligente e capaz do sou e me tratam como a irmã mais nova nerd. Obrigada por todo o apoio e amizade que não muda independente do tempo e da distância.

Aos meus amigos, Arthur Masahiro, Brunelli Miranda, Tiago Araújo e Paulo Chagas Júnior, que me acompanham desde a graduação, saudades de terminar os trabalhos com imagens de *Street Fighter* e frases de impacto. Obrigada por sempre acreditarem e me incentivarem. Juntos superamos o impossível e decolamos com o momento!

À família Pereira, tio Ronaldo, tia Geocionice, Rayssa, Ronald e Ronaldo Jr., que reencontrei em Curitiba e que me acolheu com amor e carinho. Vocês sempre me deixam emocionada ao contar a família de vocês com seis membros. Obrigada por todos os momentos de alegria e comidas gostosas. Sempre fico mais gorda após estar com vocês!

Esta é a parte que mais gosto de escrever em um trabalho, porque é quando vários momentos vividos voltam-me à mente e percebo que, por mais que tenha me sentido só em muitos momentos, eu não estava só. Durante estes dois anos de mestrado, conheci muitas pessoas, pessoas que me ensinaram, que aprenderam junto comigo e compartilharam os desesperos, que me fizeram rir (até quando não devia), me permitiram provar um pouco mais da vida e ajudaram direta e indiretamente na conclusão deste trabalho. E quero lhes dizer "Obrigada": a todos os professores e funcionários do Programa de Pós-Graduação em Bioinformática por me darem esta oportunidade e por todo o conhecimento compartilhado; aos colegas de laboratório e turma que são preciosos a sua maneira; ao Rhyan Carvalho, que mesmo morando agora em outro estado, continua sendo o "vizinho"; à Gleice Moraes por toda a ajuda no início desta etapa; aos meus tios por todo o apoio; à CAPES, CNPQ e INCT por todo o suporte financeiro.

E por fim, ao EXO, que é uma presença constante na minha vida nos últimos quatro anos. Obrigada por ser a minha luz nos momentos de escuridão, por me fazer sorrir quando quero chorar, por me dar um sopro de incentivo quando quero desistir, por me incentivar a sempre tentar o meu melhor e nunca desistir, por ser meu orgulho e exemplo e por me permitir experimentar um amor incondicional no qual só preciso saber que vocês estão bem e vê-los sorrindo para me sentir feliz. Obrigada por tudo, meus doze anjinhos. Sinto que me tornei alguém melhor e mais forte após conhecê-los. EXO *saranghaja!*

Although time passes, there is a word I cannot express,
sinking down in my heart.
'I'm sorry', 'I love you'
asking you to believe in me like this time
I will hug you and hold your hands.
If we can stay together endlessly,
I will devote myself to you.
I promise you.
Trecho de EXO - Promise (EXO 2014)

RESUMO

A Floresta Atlântica Paranaense é um dos 25 *hotspots* de biodiversidade do mundo, sendo conhecida por sua grande diversidade de fauna e flora. Mas pouco se sabe sobre a diversidade microbiana de seu solo. A Metagenômica é o campo de estudo de DNA de amostras ambientais e por meio de suas técnicas é possível acessar o DNA de comunidades microbianas e inferir informações filogenéticas e funcionais, além de obter genomas de microrganismos não cultiváveis em laboratório. Neste trabalho, foram combinadas técnicas de Metagenômica e tecnologias de sequenciamento de nova geração (NGS) para estudar amostras de DNA do solo da Mata Atlântica Paranaense. As amostras de solo foram coletadas em Julho de 2004 e Janeiro de 2007 em baixa (161 m), média (604 m) e alta altitudes (900 m). O DNA total purificado foi sequenciado pelas plataformas *Illumina MiSeq* e *Ion Proton*. O gene 16S rDNA foi amplificado para todas as amostras e sequenciado na plataforma *Illumina MiSeq*. Foram geradas 371.561 leituras de 16S rDNA e 66.936.100 leituras de DNA total, que foram submetidas a análises de diversidade e funcionais com os programas QIIME e MG-RAST. A montagem de genomas parciais foi realizada com os programas *CLC Genomics Workbench* e MEGAHIT. Durante as análises de diversidade, tanto baseada no DNA total quanto no gene 16S rDNA, foi identificada a predominância dos filos *Acidobacteria* e *Proteobacteria*. Entretanto, houve uma diferença na proporção dos filos identificados comparando-se somente sequências amplificadas de 16S rDNA e DNA total. Já a aplicação de tecnologias diferentes de sequenciamento de DNA não teve influência sobre a distribuição dos filos. As análises funcionais revelaram que a maioria das sequências de DNA total estão relacionadas a funções de metabolismo, principalmente metabolismo de carboidratos e aminoácido e derivados. A melhor montagem de genoma foi obtida pelo programa *CLC Genomics Workbench* para a amostra MAF1 usando dados *Illumina MiSeq* e gerou 16.426 *contigs* acima de 1.000 pb e maior *contig* com 35.630 pb. Os 16.426 *contigs* foram analisados com o *blastn* no banco de dados nr do NCBI e obtiveram similaridade com 3.010 sequências depositadas no banco de dados referentes aos genomas de organismos dos filos *Acidobacteria*, *Proteobacteria* e outros. O uso de sequenciamento de DNA NGS para exploração da diversidade microbiana nos solos da Floresta Atlântica Paranaense revelou a existência de uma grande microdiversidade com a presença de muitos filos bacterianos e que os resultados independem da tecnologia NGS, mas são influenciados pela época que foi coletado o DNA usado para sequenciamento.

Palavras-chave: Metagenômica, Floresta Atlântica, DNA de solo, Sequenciamento NGS, Análise de diversidade, Análise funcional.

ABSTRACT

The Paraná Atlantic Forest is one of the 25 biodiversity hotspots in the world, known for its great diversity of fauna and flora. But little is known about the microbial diversity of its soil. Metagenomics is the field that studies DNA from environmental samples allowing the access to the microbial communities DNA retrieving phylogenetic and functional information, in addition to obtaining genomes from non-cultivable microorganisms. In this work, Metagenomics techniques and new generation sequencing technologies (NGS) were combined to study DNA samples from the soil of the Atlantic Forest of Parana. Soil samples were collected in July 2004 and January 2007 at low (161 m), mean (604 m) and high altitudes (900 m). Total purified DNA was sequenced by the Illumina MiSeq and Ion Proton platforms. The 16S rDNA gene was amplified for all samples and sequenced on the Illumina MiSeq platform. A total of 371,561 reads of 16S rRNA and 66,936,100 reads of total DNA were generated, which were submitted to diversity and functional analysis with the QIIME and MG-RAST programs. The assembly of partial genomes was performed with the CLC Genomics Workbench and MEGAHIT programs. During the analysis of diversity, both based on total DNA and 16S rDNA gene, the predominance of Acidobacteria and Proteobacteria was identified. However, there was a difference in the proportion of the identified phyla comparing only amplified 16S rDNA and total DNA sequences. On the other hand, the application of different technologies of DNA sequencing had no influence on the distribution of phyla. Functional analyzes revealed that most of the total DNA sequences are related to metabolism functions, main carbohydrates and amino acids and derivatives metabolism. The best genome assembly was obtained by the CLC Genomics Workbench program for the MAF1 sample using Illumina MiSeq data and generated 16,426 contigs above 1,000 bp and higher contig with 35,630 bp. The 16,426 contigs were compared to NCBI nr database using blastn and obtained similarity with 3,010 sequences referring to the genomes of organisms from Acidobacteria, Proteobacteria and other phyla. The use of NGS DNA sequencing for the exploration of microbial diversity in the Atlantic Forest soils revealed the existence of a large micro-diversity with the presence of many bacterial phyla and that the results are independent of the NGS technology but are influenced by the time that was collected the DNA used for sequencing.

Keywords: Metagenomics, Atlantic Forest, Soil DNA, NGS sequencing, Diversity analysis, Functional analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

BLAT	-	*Blast-Like-Alignment*
BLAST	-	*Basic Local Alignment Search Tool*
COG	-	*Clusters of Orthologous Groups*
DNA	-	Ácido desoxirribonucleico
dNTP	-	Desoxirribonucleotídeo trifosfato
eggNOG	-	*evolutionary genealogy of genes: Non-supervised Orthologous Groups*
GenBank	-	*Genetic Sequence Database*
IMG	-	*Integrated Microbial Genomes*
KEGG	-	*Kyoto Encyclopedia of Genes and Genomes*
M5NR	-	Banco de dados de proteínas não-redundante
M5RNA	-	*MG-RAST RNA database*
MG-RAST	-	*MetaGenome Rapid Annotation using Subsystems Technology*
NCBI	-	*National Center for Biotechnology Information*
NGS	-	*Next Generation Sequencing*
NR	-	Banco de dados de proteínas não redundante do NCBI
NTP	-	Nucleósido trifosfato
OTU	-	Unidade taxonômica operacional
PATRIC	-	*Pathosystems Resource Integration Center*
PCR	-	Reação em cadeia da polimerase
PCoA	-	*Principal Coordinates Analysis*
PGM	-	*Personal Genome Machine*
pH	-	Potencial hidrogeniônico
rDNA	-	DNA ribossomal
RefSeq	-	*NCBI Reference Sequence*
RNA	-	Ácido ribonucleico
rRNA	-	RNA ribossomal
QIIME	-	*Quantitative Insights Into Microbial Ecology*
SMRT	-	*Single Molecule Real Time*

LISTA DE SÍMBOLOS

Gb	-	Giga bases
H^+	-	Íon
m	-	Metro
pb	-	Pares de bases
®	-	Marca registrada

SUMÁRIO

1 INTRODUÇÃO

O solo é uma parte importante do ecossistema. Participa de vários processos biológicos além de fornecer nutrientes para as plantas e microrganismos. Provavelmente possui a maior diversidade microbiana de todos os ambientes da Terra (DANIEL, 2005).

As bactérias são altamente adaptáveis e capazes de decompor todos os produtos químicos produzidos por organismos vivos e, no solo, agem como principais agentes de decomposição e desintoxicação de contaminantes ambientais (ØVREÅS, 2000). Uma grama de solo pode abrigar até 10 bilhões de microrganismos de milhões de espécies diferentes, com diferentes graus de abundância e funções biológicas. O solo agrega uma enorme quantidade de microrganismos de espécies ainda desconhecidas (cerca de 99,9%), uma vez que não crescem em culturas de laboratório (TORSVIK, GOKSØYR e DAAE, 1990; TORSVIK e ØVREÅS, 2002) e acredita-se que essas populações microbianas nunca cultivadas possuem grande riqueza de diversidade genética e biotecnológica. Diversas técnicas foram desenvolvidas para extração e purificação de DNA de amostras ambientais como forma de ter acesso a esses microrganismos não cultiváveis em laboratório, dando origem ao campo da Metagenômica que pode ser definida como o estudo do DNA de amostras ambientais (HANDELSMAN et al., 1998).

Atualmente com apenas 12% do seu tamanho original, a Floresta Atlântica brasileira está entre os 25 *hotspots*[1] de biodiversidade do mundo (MYERS et al., 2000; RIBEIRO et al., 2011). Mesmo após a perda de território para a industrialização e o crescimento populacional, a Floresta Atlântica estende-se por toda a costa leste do Brasil, de nordeste a sul, e norte da Argentina e sul do Paraguai, abrigando mais de 8.000 espécies de plantas e 567 espécies de vertebrados, correspondendo a 2,7% e 2,1%, respectivamente, do total global (RIBEIRO et al., 2011). Conhecida como uma área de preservação, a Floresta Atlântica é o alvo deste trabalho. Especificamente, o solo desta.

Nos anos de 2004 e 2007 foram coletadas amostras de solo da Floresta Atlântica no estado do Paraná em 3 regiões, que correspondem a baixa, média e alta altitudes (FAORO et al., 2010; FAORO et al., 2012). O DNA destas amostras foi purificado e sequenciado usando as tecnologias de nova geração *Illumina MiSeq* e *Ion Proton,* sendo o objetivo deste trabalho analisar os dados de sequenciamento de nova geração das amostras de solo da Floresta Atlântica Paranaense.

[1] *Hotspots* são áreas de grande concentração de diversidade de flora e fauna e, por isso, de prioridade global para conservação da biodiversidade (RIBEIRO et al., 2011)

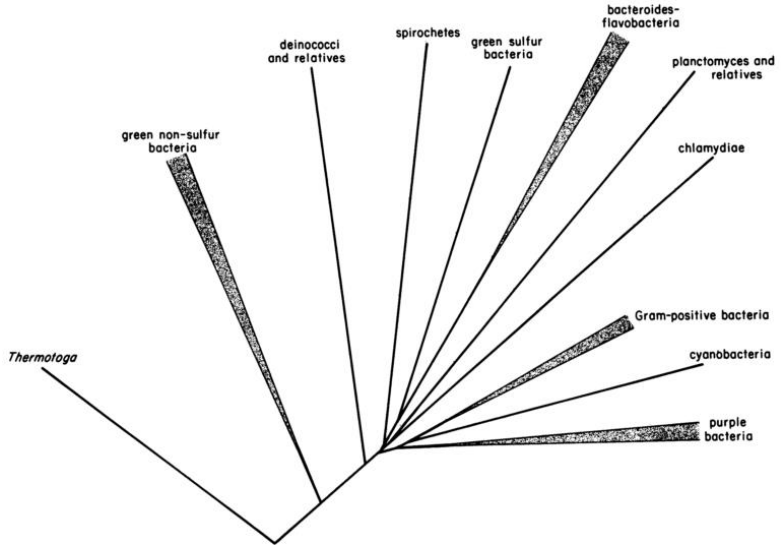
2 REVISÃO DE LITERATURA

2.1 METAGENÔMICA

Em 1978, Torsvik e Goksøyr descreveram dois métodos para determinação de DNA em amostras de solo e identificaram um grande número de moléculas de DNA (TORSVIK e GOKSØYR, 1978). A partir disso, consideraram a possibilidade de isolar DNA puro de amostras de solo, publicando em 1980 um protocolo de isolamento e purificação de DNA de amostras de solo (TORSVIK, 1980) que viabilizou estudos do DNA de amostras ambientais e tornou possível a identificação de 4.000 genomas diferentes de amostras de solo (TORSVIK et al., 1990). Desde então foram publicados vários protocolos de extração de DNA, bem como foram produzidos vários kits comerciais com a mesma finalidade.

Apesar das diferenças entre os protocolos já conhecidos, os métodos de extração de DNA podem ser divididos em dois: indireto e direto. No método indireto, é feita a separação das células da matriz do solo que depois passarão pela lise celular. Já no método direto, é feita a lise celular com toda a matriz do solo que ajuda a separar o DNA da matriz e dos detritos celulares encontrados no solo. Com o método direto é possível recuperar uma maior quantidade de DNA e ter acesso a maior diversidade bacteriana, embora este seja mais fragmentado e não tão puro quanto o DNA obtido pelo método indireto (DANIEL, 2005).

FIGURA 1 – ÁRVORE FILOGENÉTICA DO DOMÍNIO *BACTERIA* PROPOSTA POR WOESE

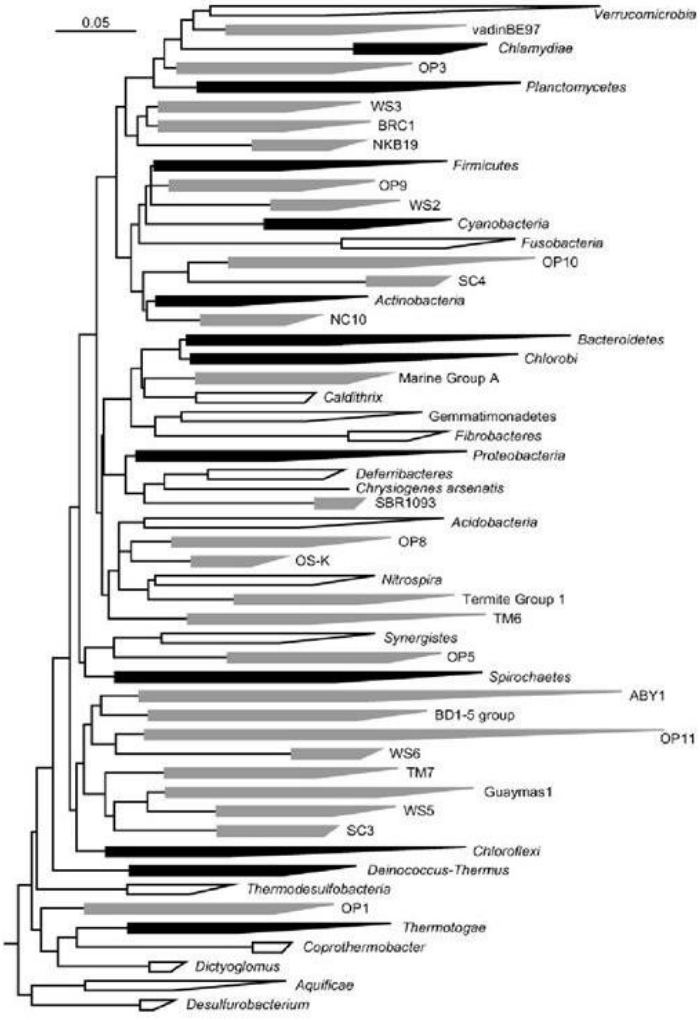


FONTE: WOESE (1987).
LEGENDA: Árvore filogenética do domínio *Bacteria* proposta por Woese com 11 filos taxonômicos.

Em 1987, a árvore filogenética do domínio *Bacteria* era composta apenas por 11 filos

descritos com base nos perfis de restrição da molécula de 16S rRNA (WOESE, 1987) (FIGURA 1). Usando a purificação de DNA ambiental, Pace e colaboradores (1986) desenvolveram uma abordagem que envolvia a amplificação do gene que codificava o 16S rRNA (16S rDNA) e sua clonagem como forma de explorar a biodiversidade bacteriana. A partir disso e conforme foram descobertos novos organismos, a árvore filogenética foi expandida e modificada com novos filos sendo adicionados. Rappé e Giovannoni (2003) expandiram a árvore original para 52 filos e, dentre esses, 26 eram filos candidatos os quais não possuíam representantes isolados e cultivados em laboratório (FIGURA 2).

FIGURA 2 – ÁRVORE FILOGENÉTICA DO DOMÍNIO *BACTERIA* PROPOSTA POR WOESE MODIFICADA POR RAPPÉ E GIOVANNONI

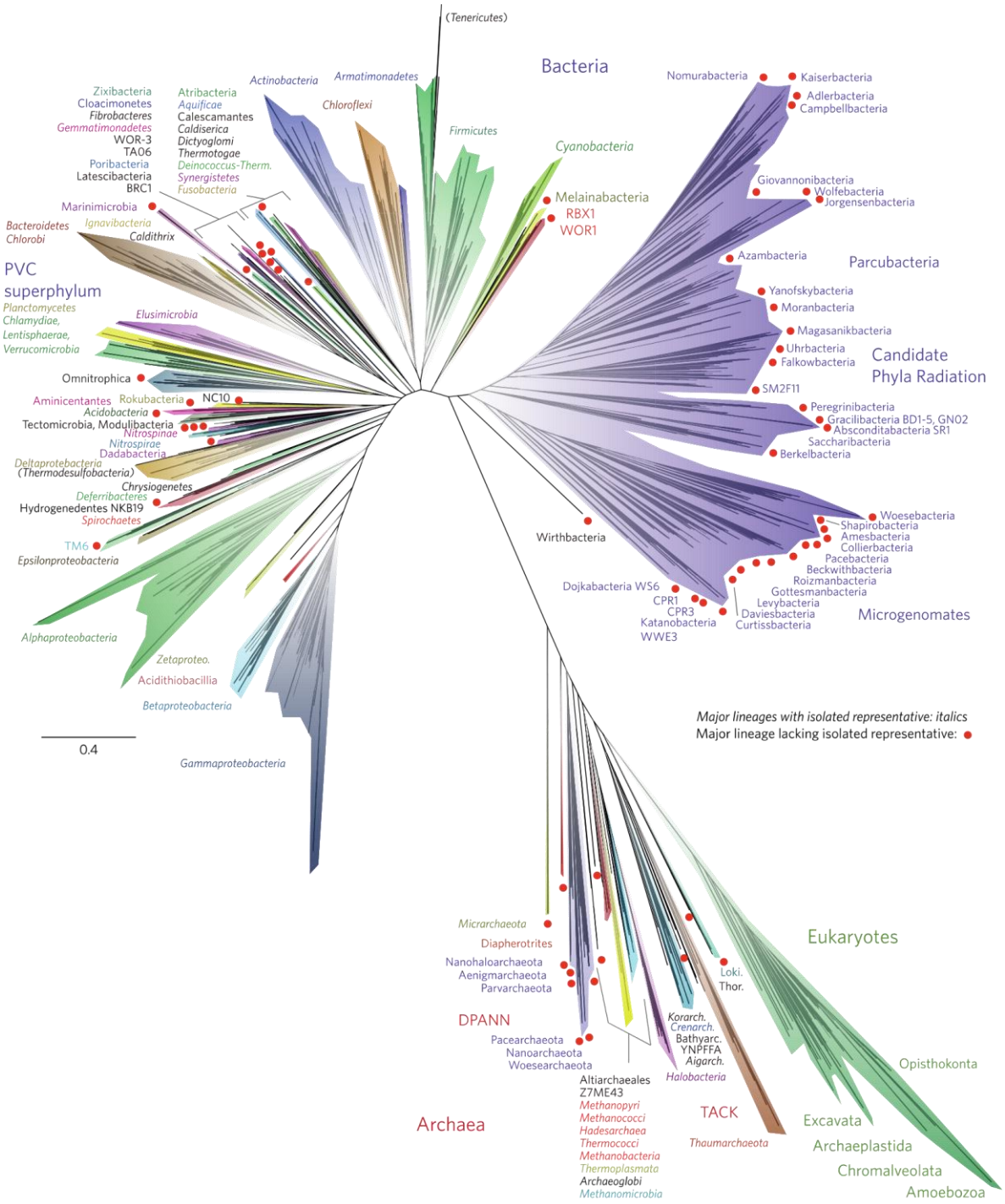


FONTE: RAPPÉ e GIOVANNONI (2003).

LEGENDA: Árvore filogenética do domínio *Bacteria* expandida com 52 filos taxonômicos. Setas pretas representam os filos originais da árvore de Woese (com o filo *Gram-positive bacteria* dividido nos filos *Actinobacteria* e *Firmicutes*), setas brancas representam os filos cultiváveis em laboratório e, as cinzas, os filos candidatos.

Atualmente, as técnicas de sequenciamento de DNA de nova geração (NGS) aliadas à Metagenômica permitiram uma grande reformulação da árvore filogenética (HUG et al., 2016) (FIGURA 3).

FIGURA 3 – ÁRVORE FILOGENÉTICA DOS 3 DOMÍNIOS DA VIDA REFORMULADA POR HUG



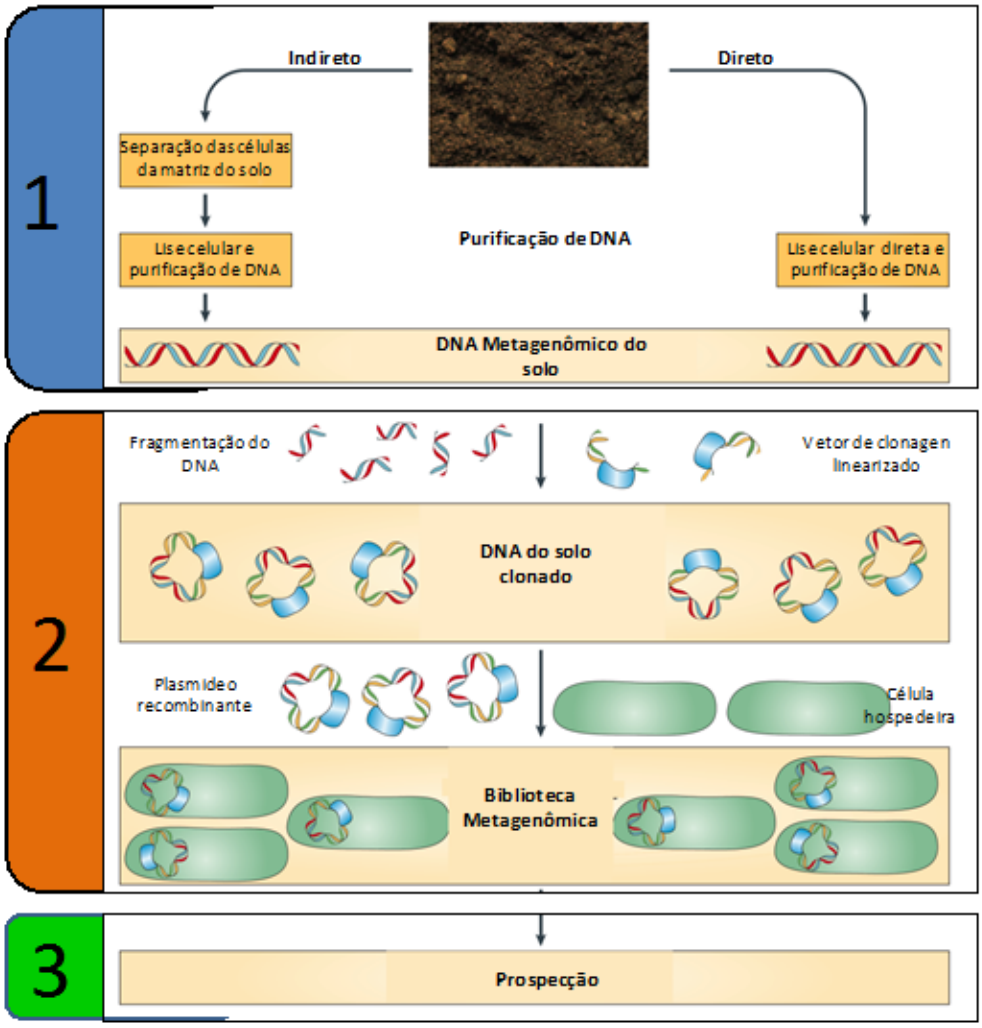
FONTE: HUG et al. (2016).

LEGENDA: Árvore filogenética expandida e reformulada com os três domínios taxonômicos (*Archaea*, *Eukayotes* e *Bacteria*) construída a partir de informações moleculares. Os pontos vermelhos representam os filos sem representantes isolados e o parâmetro 0.4 corresponde à distância evolutiva calculada pelos autores.

O trabalho desenvolvido por Norman Pace em 1986 foi o primeiro a propor o estudo de comunidades baseado na purificação, amplificação e clonagem do gene 16S rDNA de amostras ambientais (PACE et al., 1986). Entendendo-se por metagenoma, o conjunto de genomas presentes em uma amostra ambiental.

O termo Metagenômica foi cunhado por Handelsman e colaboradores (1998) e é atualmente conhecida como Metagenômica Clássica. Essa técnica consiste no isolamento do DNA metagenômico diretamente do solo, purificação por método direto ou indireto (FIGURA 4-1), fragmentação do DNA e clonagem em vetor artificial, formando as bibliotecas metagenômicas (FIGURA 4-2) que serão analisadas quanto às atividades biológicas e prospecção de novos produtos naturais (FIGURA 4-3).

FIGURA 4 – METAGENÔMICA CLÁSSICA DESCRITA POR HANDESLMAN



FONTE: Modificada de DANIEL (2005).
LEGENDA: Técnica da Metagenômica Clássica.
1 – Purificação de DNA metagenômico;
2 – Fragmentação e clonagem em vetor artificial e construção de bibliotecas metagenômicas; e
3 – Prospecção de novos produtos naturais.

Usando a técnica descrita, Rondon e colaboradores (2000) identificaram a atividade de lipase, amilase, nuclease, hemolítica e compostos antibacterianos em bibliotecas metagenômicas construídas com DNA de solo, sendo esse trabalho considerado o primeiro dentro da Metagenômica.

2.2 FLORESTA ATLÂNTICA

Estendendo-se por toda a costa leste do Brasil, de nordeste a sul, e norte da Argentina e sul do Paraguai, abrigando mais de 8.000 espécies de plantas e 567 espécies de vertebrados, correspondendo a 2,7% e 2,1%, respectivamente, do total global (RIBEIRO et al., 2011), a Floresta Atlântica é uma área de preservação e está entre os 25 *hotspots* de biodiversidade do mundo (MYERS et al., 2000; RIBEIRO et al., 2011).

A fauna e flora da Floresta Atlântica já foi muito estuda e explorada no decorrer dos anos, implicando na redução de espécies nativas e risco de extinção de algumas espécies de animais e vegetais. Sendo necessária a supervisão constante por órgãos de preservação como o Ministério do Meio Ambiente (http://www.mma.gov.br/), que promove programas de recuperação, conservação e sustentabilidade no território nacional, e a implantação de reservas naturais ao longo da Mata Atlântica brasileira.

Assim como a fauna e a flora, o solo da Floresta Atlântica vem sendo estudado ao longo da última década. Em 2010 foi publicado um dos primeiros trabalhos tendo o solo da floresta como foco, que usando técnicas de Metagenômica analisou amostras de solo da Floresta Atlântica no estado do Paraná coletadas nos anos de 2004 e 2007 e, após realizadas as análises de biodiversidade e estatística, mostrou-se que o solo da Floresta Atlântica do Paraná possui grande diversidade bacteriana e a predominância dos filos de bactéria *Acidobacteria* (63%), *Proteobacteria* (25.2%), *Gemmatimonadetes* (1.6%) e *Actinobacteria* (1.2%) (FAORO et al., 2010). Outro trabalho realizado, desta vez usando amostras de solo da Parque Nacional da Serra dos Órgãos (PARNASO), no estado do Rio de Janeiro, mostrou a predominância dos filos *Acidobacteria* (29-54%) e *Proteobacteria* (16-38%), além do filo *Verrucomicrobia* (0.6-14%) (BRUCE et al., 2010), que não foi identificado no solo da Floresta Atlântica paranaense no trabalho de Faoro (2010). Ambos os trabalhos mostraram que o solo da Floresta Atlântica brasileira apresenta uma riqueza de biodiversidade bacteriana e que apresenta diferenças dependendo da região e das características físico-químicas e métodos de análises utilizados.

2.3 SEQUENCIAMENTO NGS

O sequenciamento de DNA consiste em determinar a sequência de nucleotídeos que compõem uma molécula de DNA. O método de terminação de cadeia de Sanger, publicado em 1977, foi o primeiro método popular de sequenciamento de DNA e o mais rápido e simples dentre os métodos disponíveis no momento (SANGER e NICKLEN, 1977). Este método possibilitou o crescimento e aprimoramento da técnica, sendo lançado em 1986 o primeiro sequenciador automático de DNA, o A370, da empresa *Applied Biosystems*. Quatro anos depois, a *Applied Biosystems* lançou o ABI 373, o primeiro sequenciador automático de DNA usando eletroforese. A automação do sequenciamento de DNA foi fundamental para a realização de grandes projetos genômicos, como o Projeto Genoma Humano publicado em 2001 (SPRINGER, 2006).

Durante décadas, o método de sequenciamento de DNA sofreu poucas alterações até que em 2004 foi lançado pela Roche o primeiro sequenciador de nova geração (NGS) chamado de 454. Nos anos seguintes, o 454 foi seguido pelos sequenciadores das empresas *Illumina* e *Life Technologies* o que deu início a uma revolução nas Ciências Biológicas. Os sequenciadores NGS baseiam-se no processamento paralelo de grandes quantidades de moléculas de DNA e permitem o acesso a enormes quantidades de dados de sequenciamento em tempo reduzido e com menor custo (MARDIS, 2008; METZKER, 2010; SHOKRALLA et. al, 2012). As tecnologias de nova geração foram fundamentais para o crescimento da Metagenômica, justamente por permitir o sequenciamento de amostras de DNA ambiental, que podem conter fragmentos genômicos de centenas a milhares de indivíduos em uma única amostra (SHOKRALLA et al., 2012). Posteriormente, em 2011, uma nova tecnologia de sequenciamento de DNA, chamada de sequenciamento por semi-condução, foi apresentada pela empresa *Life Technologies* (ex *Applied Biosystems*) nas plataformas *Ion Torrent* (PGM) e *Ion Proton*. Atualmente a revolução continua com a tecnologia de sequenciamento de molécula única em tempo real (SMRT - *Single Molecule Real Time, PacBio*) e por nanoporo (*Oxford Nanopore*) (MUNROE e HARRIS, 2010).

As grandes revoluções introduzidas pelos sequenciadores NGS foram a eliminação da etapa de clonagem de fragmentos de DNA previamente ao sequenciamento e a miniaturização da reação em cadeia da polimerase (PCR), que passou a ser realizada em micro reatores ou direto na célula de sequenciamento. Ambas as técnicas diminuíram dramaticamente o tempo e o investimento necessário para o sequenciamento de genomas completos (METZKER, 2010). De modo geral, em todas as plataformas NGS, a molécula de DNA a ser sequenciada é quebrada

em fragmentos pequenos (100-500 pb) aos quais são ligados adaptadores, pequenas moléculas de DNA produzidas artificialmente. A sequência de nucleotídeos dos adaptadores é complementar a uma sequência de oligonucleotídeos imobilizada na superfície do aparato usado no sequenciamento (microesferas ou células de sequenciamento). As moléculas de DNA de interesse imobilizadas são amplificadas de modo clonal para aumentar a intensidade do sinal e utilizadas para o sequenciamento.

Neste trabalho foram usadas as plataformas *Illumina MiSeq* e *Ion Proton.* O sequenciador *Illumina MiSeq* foi lançado em 2011 e tem a capacidade de gerar fragmentos de leituras (*paired-end reads*) de até 600 pb e um total de 15 Gb por corrida. Ele utiliza a abordagem de sequenciamento por síntese. Desse modo os fragmentos de DNA ligados a adaptadores são imobilizados em uma lâmina onde é realizada a PCR para amplificação de cada fragmento, formando agrupamentos de sequências clones (*clusters*). Posteriormente, cada *cluster* recebe DNA polimerase e nucleotídeos fluorescentes marcados com a cor referente a cada base nitrogenada e com a terminação 3' quimicamente inativada, permitindo que somente uma base seja adicionada a cada ciclo. A cada vez que uma base é incorporada, ocorre um sinal fluorescente e uma imagem é captada para a identificação da base adicionada a cada *cluster*, depois os terminadores e sinais fluorescentes são removidos e uma nova base pode ser adicionada. Isto se repete até que não haja bases a serem adicionadas (MARDIS, 2008; SHOKRALLA et al., 2012).

Lançado em 2010, o sequenciador *Ion Proton* gera fragmentos de leituras simples (*single reads*) de até 200 pb e até 6 Gb por corrida. Também usa a abordagem de sequenciamento por síntese, porém, diferente do *Illumina MiSeq*, combina a tecnologia de semicondutores com mudanças químicas para gerar dados digitais. Os fragmentos de DNA são ligados aos adaptadores, conectados a microesferas e amplificados por PCR em emulsão. Posteriormente, as microesferas recobertas com fragmentos de DNA são depositadas em poços no chip semicondutor, que receberá um nucleotídeo dNTP e DNA polimerase. A cada NTP incorporado na fita de DNA, um H^+ é liberado alterando o pH do poço, o que permite determinar qual e quantas bases foram adicionadas à sequência ("ION PROTON SYSTEM - THERMO FISHER SCIENTIFIC"; MUNROE e HARRIS, 2010; SHOKRALLA et al., 2012).

2.4 MONTAGEM DE GENOMAS

Quando sequenciado, o DNA é fragmentado em sequências menores chamadas leituras (*reads*), sendo que a montagem de um genoma consiste em organizar essas leituras em uma

sequência que representa o genoma do organismo alvo. Durante a montagem, as leituras são agrupadas e, quando há regiões idênticas, são sobrepostas formando *contigs*, que podem ser ordenados e agrupados em sequências maiores, os *scaffolds*. Também podem ser encontradas lacunas (*gaps*) entre ou dentro dos *scaffolds* que, normalmente, ocorrem quando não é encontrada uma sequência que preencha determinada região (BAKER, 2012).

Há dois tipos de montagem: montagem por referência e montagem *de novo*. A montagem por referência consiste em usar como modelo o genoma de um organismo próximo e já conhecido para reconstruir a sequência genômica do organismo em estudo. Enquanto na montagem *de novo*, são usados somente os dados de sequenciamento para reconstruir a sequência genômica (THOMAS, GILBERT e MEYER, 2012).

A montagem de um genoma é realizada por programas específicos chamados de "montadores". Atualmente há vários programas montadores de genoma publicados e consolidados que usam algoritmos e técnicas de montagem variados (HUSON et al., 2011; LI et al., 2016; PENG et al., 2011; ZERBINO e BIRNEY, 2008). A performance de cada programa depende do tipo, volume e qualidade dos dados, além do equipamento utilizado. Apesar do avanço nas tecnologias de sequenciamento e dos vários programas montadores disponíveis, a montagem de genomas a partir de amostras ambientais apresenta muitos desafios. Uma única amostra ambiental pode conter dados biológicos de centenas ou milhares de organismos, que são todos fragmentados e misturados durante o sequenciamento. Extrair o genoma de um organismo dentre desses dados é extremamente difícil. É necessário um grande volume de dados de sequenciamento de DNA de forma a aumentar a cobertura e a confiabilidade das sequências montadas e diminuir o risco de montagem de quimeras. No entanto, uma vez que a Metagenômica visa a montagem de genomas de organismos não cultiváveis em laboratório, é difícil dizer a exatidão de uma montagem quando não há uma referência para comparar (HOWE e CHAIN, 2015; SHARPTON, 2014; THOMAS, GILBERT, e MEYER, 2012; TRINGE et al., 2005).

2.5 ANÁLISE DE DIVERSIDADE

A análise de diversidade objetiva traçar o perfil taxonômico de uma comunidade microbiana, determinando quais organismos estão presentes na comunidade e qual a sua abundância (SHARPTON, 2014). Para isso são usadas sequências completas ou parciais do gene que codifica o RNA ribossomal 16S que pode ser amplificado diretamente de amostras de DNA ambiental (PACE et al., 1986). Algumas características que tornam o 16S rRNA um bom

marcador molecular são: 1) está presente em todos os procariotos; 2) estrutura e função conservadas entre os diferentes taxa; 3) sequência nucleotídica com alternância entre regiões conservadas e variáveis, sendo essas últimas classificadas de V1 a V9; 4) grande número de sequências depositadas em banco de dados (PACE et al., 1986).

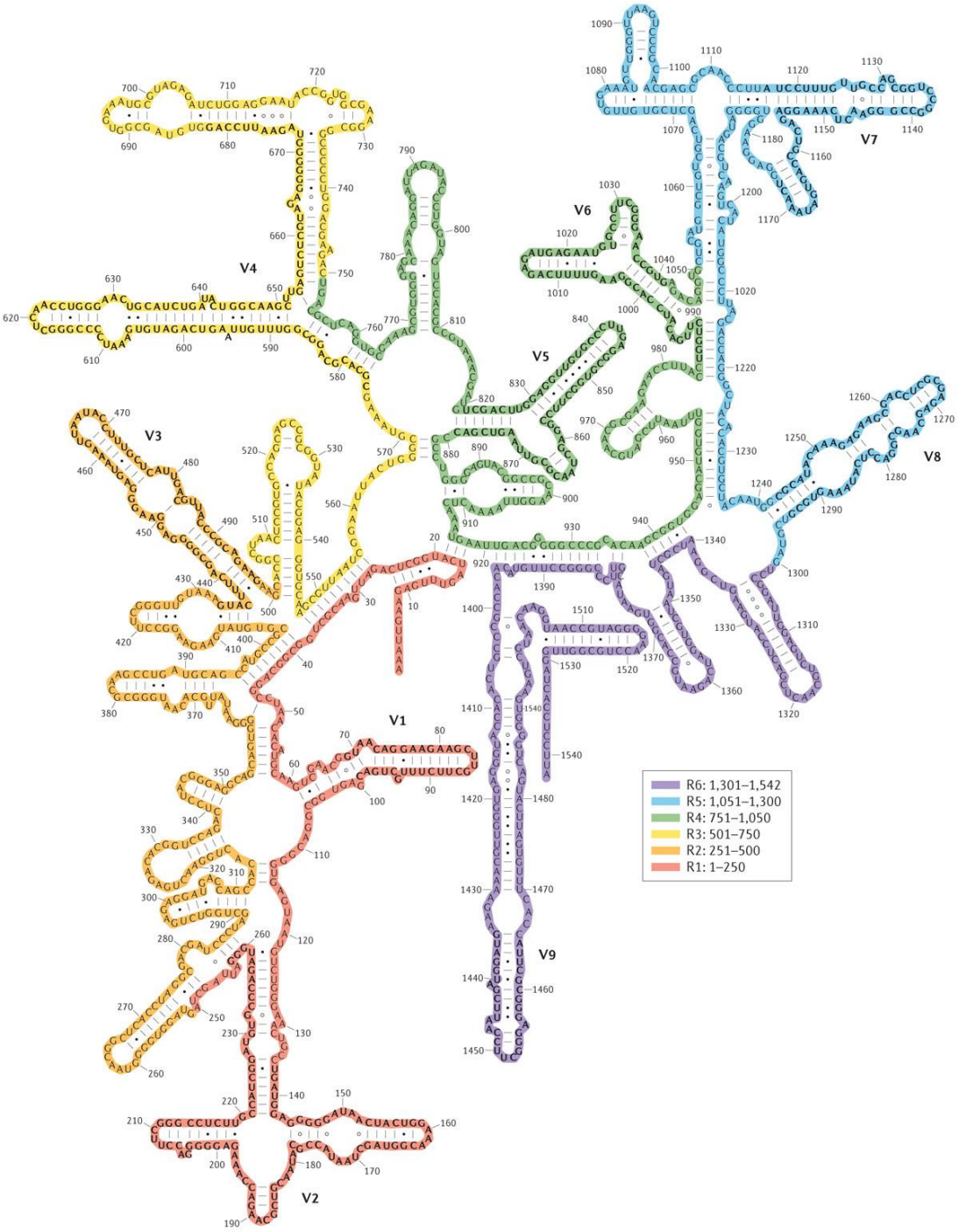
Os algoritmos de classificação de sequências de 16S rRNA para inferência de grupos taxonômicos estão divididos em duas abordagens: dependente da taxonomia e independente da taxonomia. Na abordagem dependente da taxonomia, as sequências de 16S rRNA são comparadas contra um banco de dados usando algoritmos de alinhamento como o BLAST (ou BLAT para o MG-RAST, usado neste trabalho) e são atribuídas aos organismos com a maior taxa de correspondência. Esta abordagem depende que os organismos estejam bem caracterizados nos bancos de dados para que tenha resultados confiáveis (SUN et al., 2012).

Diferentemente da abordagem descrita acima, a abordagem independente de taxonomia não precisa de um banco de dados de referência, o que aumenta a possibilidade de identificação de organismos não cultiváveis em laboratório e que ainda não foram sequenciados. Nesta abordagem, as sequências são comparadas entre si, gerando uma matriz de distância, e são agrupadas em unidades taxonômicas operacionais (OTU's) de acordo com um valor de variabilidade permitida entre as sequências dentro de cada OTU. Embora bastante eficiente, esta abordagem não é totalmente eficaz, uma vez que não há informações de como as OTU's foram agrupadas e, os diferentes algoritmos, usam formas diferentes de calcular as matrizes de distâncias (SUN et al., 2012).

Segundo White e colaboradores (2010), uma pequena alteração nos parâmetros usados no mesmo algoritmo pode resultar em variações significativas na composição das OTU's. O raciocínio pode ser aplicado em relação aos resultados de diferentes abordagens, uma vez que usam algoritmos diferentes, e mesmo que configurados os mesmos parâmetros, os resultados serão diferentes entre si.

Além da abordagem e algoritmos usados, os resultados das análises de diversidade com sequências de 16S rRNA também irão variar conforme a região do gene analisada. A figura 5 mostra a estrutura do gene 16S rRNA com as 9 regiões variáveis separadas em fragmentos de aproximadamente 250 nucleotídeos, sendo as regiões V3, V4 e V9 as maiores com aproximadamente 250 nucleotídeos cada uma.

FIGURA 5 – ESTRUTURA SECUNDÁRIA DO RRNA 16S



FONTE: (YARZA et al., 2014).

LEGENDA: Estrutura secundária do gene 16S rRNA da *Escherichia coli* com suas 9 regiões variáveis divididas em 6 fragmentos de aproximadamente 250 nucleotídeos. Fragmento R1 (vermelho): V1 e V2; fragmento R2 (laranja): V3; fragmento R3 (amarelo): V4; fragmento R4 (verde): V5 e V6; fragmento R5 (azul): V7 e V8; e fragmento R6 (roxo): V9.

Cada região variável do gene 16S rRNA possui um grau de conservação diferente, o que implica na variação dos resultados taxonômicos obtidos. Por exemplo, o fragmento R1

composto pelas regiões V1-V2 é muito curto (apenas 250 nucleotídeos) e altamente variável, o que compromete o acesso a informações taxonômicas de níveis mais altos, como de filo (YARZA et al., 2014). Sendo necessárias sequências maiores de 16S rRNA e alta cobertura para acessar maiores informações de diversidade e taxonomia das comunidades microbianas.

Além do gene 16S rDNA, o gene 5S rDNA também é usado para análises de diversidade mesmo possuindo somente 120 nucleotídeos, como observado por Pace et al. (1986).

No trabalho anterior, de Faoro e colaboradores (2010), foram amplificadas as regiões V1-V2 do gene 16S rRNA para a análise de diversidade da comunidade microbiana da Floresta Atlântica Paranaense usando o método de Sanger. Neste trabalho, amplificamos as regiões V4-V5 formadas por aproximadamente 300 nucleotídeos (500 a 816) usando plataformas de sequenciamento NGS para a caracterização do mesmo solo.

2.6 ANÁLISE FUNCIONAL

Na Metagenômica, a análise funcional objetiva identificar os genes e suas funções em uma determinada comunidade microbiana e pode ser quantificada pelo número de sequências obtidas que são similares a sequências gênicas depositadas em banco de dados. Diferente do 16S rDNA, essas sequências são obtidas do DNA total e codificadoras de proteínas (SHARPTON, 2014). Inferir a função para um metagenoma é mais complicado que inferir taxonomia, uma vez que envolve dois passos: a predição e a anotação do gene. Essas etapas são totalmente dependentes da comparação contra banco de dados biológicos a fim de identificar sequências homólogas de um gene conhecido. Um problema relacionado a essa abordagem é quando o gene está pouco representado no banco de dados ou procura-se identificar genes novos. O tamanho das sequências é outro problema encontrado, pois a maioria das abordagens trabalham com fragmentos longos de DNA ou genomas montados (LINDGREEN et al., 2016; SHARPTON, 2014; THOMAS et al., 2012). Os anotadores *GeneMark* (BESEMER e BORODOVSKY, 2005) e GLIMMER (DELCHER et al., 1999), por exemplo, são baseados em Modelo de Markov, o que dificulta a identificação de um gene novo ou que não esteja bem caracterizado, uma vez que precisam de um conjunto de dados de genes conhecidos. Há também os anotadores que utilizam informações a *downstream* para realizar a anotação, esse método exige um tamanho mínimo para as sequências e uma cobertura superior a 100x, uma vez que os dados de Metagenômica são muito fragmentados (EKBLOM e WOLF, 2014; KOONIN e GALPERIN, 2003).

A análise funcional para dados metagenômicos representa um desafio computacional. Programas como o MG-RAST que realiza anotação de sequências curtas por meio de buscas por similaridade em banco de dados, mesmo fazendo o agrupamento (*binning*) de sequências para reduzir o volume de dados, para amostras ambientais é exigido muito poder computacional e demanda tempo. E estima-se que possam ser anotadas somente 20 a 50% das sequências submetidas (LINDGREEN et al., 2016; THOMAS et al., 2012).

3 OBJETIVOS

3.1 OBJETIVO GERAL

Analisar e comparar dados de sequenciamento de DNA de amostras de solo da Floresta Atlântica do Paraná produzidos nos sequenciadores *Illumina MiSeq* e *Ion Proton*.

3.2 OBJETIVOS ESPECÍFICOS

- Realizar a montagem de genomas parciais;
- Comparar os resultados obtidos para as diferentes amostras de solo e períodos de coleta;
- Comparar os resultados obtidos entre as plataformas *Illumina MiSeq* e *Ion Proton*;
- Analisar a diversidade microbiana baseado no sequenciamento do gene 16S rDNA;
- Realizar a análise funcional baseada no sequenciamento do DNA total.

4 JUSTIFICATIVA

A Floresta Atlântica brasileira é um *hotspot* de biodiversidade de fauna e flora, mas existe pouca informação sobre a diversidade microbiana e seu potencial biotecnológico. Trabalhos prévios usando amostras de solo dessa floresta começaram a revelar um pouco desse potencial baseados em sequenciamento Sanger do gene 16S rDNA e técnicas clássicas de Metagenômica envolvendo clonagem de DNA e prospecção em placas (FAORO et al, 2010; FAORO et al, 2011; FAORO et al, 2012). Apesar de promissores, os resultados obtidos anteriormente foram limitados pela tecnologia aplicada. O desenvolvimento das tecnologias de sequenciamento de DNA de nova geração representou uma nova oportunidade para complementar e aprofundar esses estudos permitindo a identificação de mais organismos, bem como organismos raros, montagem de genomas parciais de organismos desconhecidos e obtenção de sequências de genes de interesse biotecnológico.

5 MATERIAL E MÉTODOS

5.1 AMOSTRAS

As amostras de DNA de solo foram coletadas ao longo da Rodovia PR 410 no estado do Paraná, que cruza 28,5 km de área da Floresta Atlântica, e consistem em dois grupos coletados em julho de 2004 e janeiro de 2007 (nas estações climáticas inverno e verão, respectivamente) em alta, média e baixa altitudes (TABELA 1). O DNA das amostras coletadas em 2004 foi purificado conforme especificações padrão do kit de DNA *UltraClean soil DNA* (*MoBio Laboratories*) (FAORO et al., 2010) enquanto o DNA das amostras coletadas em 2007 foi purificado seguindo as especificações padrão do kit *PowerMax Soil DNA Isolation Kit* (*MoBio Laboratories*) (FAORO et al., 2012).

TABELA 1 – AMOSTRAS DE DNA DE SOLO

Altitude (m)	**Amostras**	
	Jul/2004	**Jan/2007**
900	MA02	MAF1
604	MA05	MAF2
161	MA07	MAF3

FONTE: (FAORO et al., 2010; FAORO et al., 2012)

5.2 SEQUENCIAMENTO *ILLUMINA MISEQ* E *ION PROTON*

As amostras de solo foram divididas em dois grupos. O primeiro, dedicado à análise do DNA total, foi formado pelas amostras MAF1, MAF2 e MAF3. Os dados de sequenciamento foram gerados pelas plataformas *Illumina MiSeq* e *Ion Proton*.

O segundo grupo, dedicado à análise do gene 16S rDNA, foi composto por todas as amostras (MA02, MA05, MA07, MAF1, MAF2 e MAF3) e passou pela amplificação do gene 16S rDNA usando iniciadores modificados segundo Caporaso e colaboradores 2012. Esses iniciadores amplificam a região V4-V5 do gene 16S rDNA e também carregam a sequência dos adaptadores utilizados na construção de bibliotecas genômicas da plataforma *Illumina*. Essas amostras foram sequenciadas na plataforma *Illumina MiSeq*.

Os sequenciadores *Illumina MiSeq* e *Ion Proton* pertencem ao Departamento de Bioquímica e Biologia Molecular da Universidade Federal do Paraná e foram administrados pelos pesquisadores Dr. Helisson Faoro e Dra. Michelle Zibetti Tadra-Sfeir.

5.3 MONTAGEM DE GENOMAS PARCIAIS

Para este trabalho foram usados dois programas montadores de genomas: *CLC Genomics Workbench* 7.5.2 (CLC bio®) e MEGAHIT v1.0 e v1.0.6 (LI et al., 2016). O *CLC Genomics Workbench* é um programa proprietário que agrupa uma série de ferramentas para análises genômicas, dentre elas: montagem *de novo*, mapeamento e *trimming* de *reads*. O MEGAHIT é um programa livre e tem como foco a montagem de dados metagenômicos. Seu algoritmo consiste na construção de um grafo de *Bruijn* e é indicado para ambientes com grande capacidade de processamento, devido ao excessivo uso de memória na construção do grafo.

A montagem dos genomas foi realizada usando as leituras produzidas pelas plataformas *Illumina MiSeq* e *Ion Proton*. Para esta etapa as leituras *Ion Proton* foram tratadas para remoção de sequências dos adaptadores e leituras de baixa qualidade (menor que Phred 20, 1 erro a cada 100 pb). Os dados produzidos na plataforma *Illumina MiSeq* são tratados pelo próprio programa do sistema antes de gerar o arquivo final. Depois os dados de sequenciamento *Illumina MiSeq* e *Ion Proton* foram submetidos à montagem *de novo* pelos montadores *CLC Genomics Workbench* 7.5.2, usando os parâmetros padrões descritos na tabela 2, e o programa MEGAHIT v1.0 e v1.0.6 sem e com o parâmetro "*--presets meta-large*", específico para análises de metagenoma de solo. Foram abordados três tipos de montagem: montagens somente com dados *Illumina MiSeq*, somente com dados *Ion Proton* e montagens híbridas, combinando os dados das duas plataformas de sequenciamento.

TABELA 2 – PARÂMETROS *CLC GENOMICS WORKBENCH* PARA MONTAGEM *DE NOVO*

Parâmetro		**Valor**
Graph parameters	*Automatic word size**	20
	Automatic bubble size*	50
Contig length	Minimum contig length	200
Auto-detect paired distances		yes
Perform scaffolding		yes

FONTE: O autor (2017).

NOTA: *Word size* e *bubble size* são parâmetros referentes à construção do grafo de *Bruijn*. *Word* é o tamanho das sequências de DNA que formarão os nós do grafo e *bubble* é o tamanho das bolhas que poderão ser formadas e resolvidas durante a construção do grafo.

Com os resultados de todas as montagens, a melhor montagem foi selecionada e os *contigs* acima de 1.000 pb foram comparados com o *GenBank* usando o programa *blastn* da ferramenta *BLAST at NCBI* disponível dentro do pacote *CLC Genomics Workbench* com os parâmetros descritos na tabela 3.

TABELA 3 – PARÂMETROS *BLAST AT NCBI*

Parâmetro	**Valor**
Limit by entrez	*All organisms*
Choose filter	*Filter low complexity*
Expect	10.0
Word size	20 *(default 11)*
Match/mismatch	*Match2, Mismatch -3*
Gap costs	*Existence 5, Extension 2*
Number of hit sequences	100

FONTE: O autor (2017).

Além da lista com todos os *hits* de cada *contig* analisado, o *BLAST at NCBI* disponibiliza uma tabela com as informações simplificadas da pesquisa no banco de dados. A tabela informa o nome do *contig* e o número de *hits* e os classifica por menor *e-value*, maior % de identidade, maior % de positivos, maior tamanho do *hit* e maior *bit score*, informando para cada um o número de acesso e a descrição das sequências depositadas no banco de dados *nr*.

Baseado nessa tabela, foram filtradas as sequências do banco de dados com maior número de *hit* em todas as categorias. Essas sequências foram usadas como referência para mapeamento das leituras *Illumina MiSeq* e *Ion Proton*. A partir disso, foram extraídas as leituras mapeadas em cada sequência do banco de dados e submetidas a uma nova montagem *de novo* com o *CLC Genomics Workbench*, com os parâmetros padrão (TABELA 2).

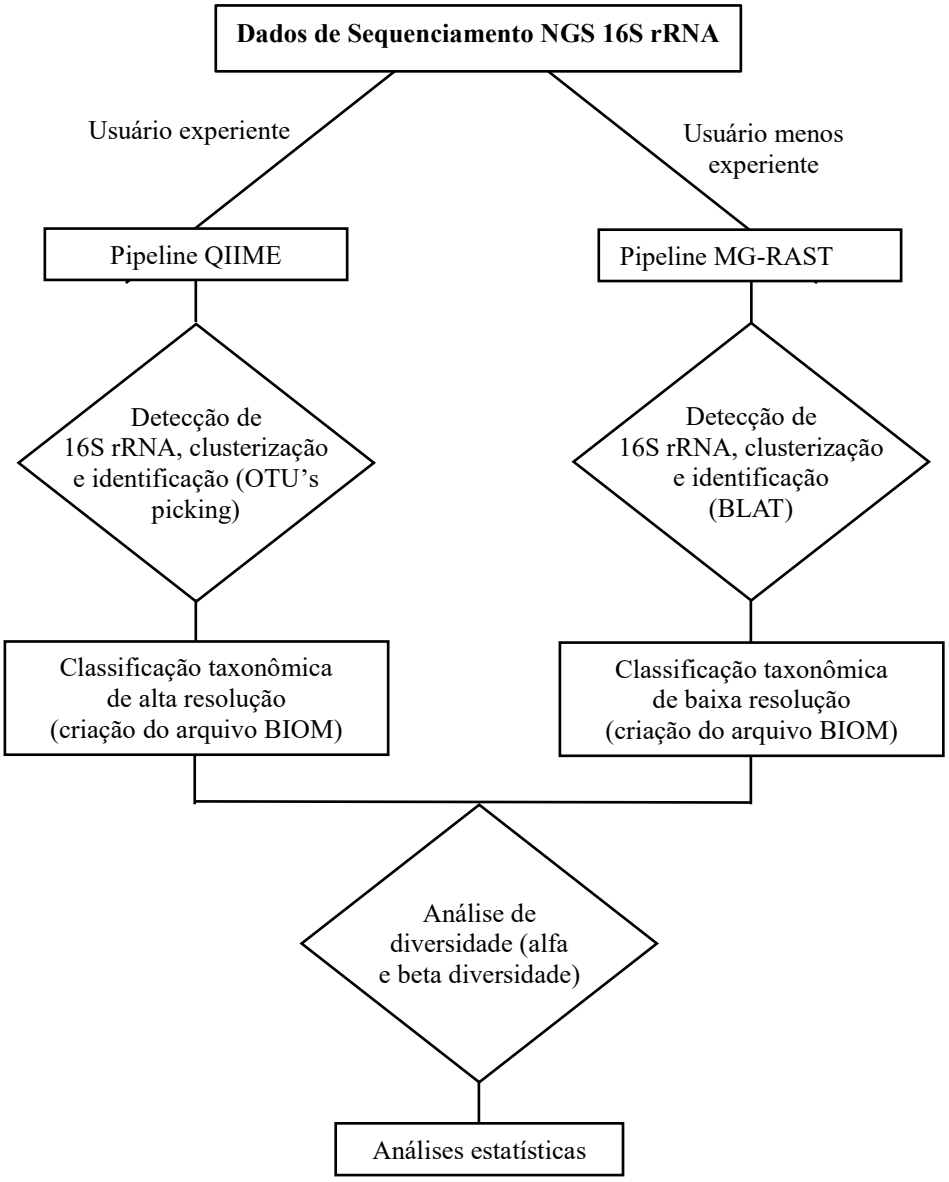
5.4 ANÁLISE DE DIVERSIDADE

Neste trabalho, foram usados o programa QIIME (CAPORASO et al., 2010) e o servidor MG-RAST (MEYER et al., 2008) para a análise de diversidade com dados metagenômicos. O QIIME (*Quantitative Insights Into Microbial Ecology*) é um programa livre (*open-source)* baseado em *scripts Phyton* que permitem a classificação das sequências de 16S rDNA em OTU's e, usando-as como base, construir árvores filogenéticas, plotar gráficos taxonômicos, construir redes de interação, calcular medidas de alfa e beta diversidade, entre outros (CAPORASO et al., 2010).

O servidor MG-RAST (*MetaGenome Rapid Annotation using Subsystems Technology*) é um sistema *web open-source* para análises de dados metagenômicos. Os dados de sequenciamento são enviados para o servidor que irá seguir um *pipeline* interno de anotação e posteriormente é disponibilizado um conjunto de ferramentas para análises pós-anotação, além dos dados de anotação que podem ser baixados ou reanalisados no *workbench* do sistema. Dentre as ferramentas oferecidas, estão: criação de *heatmap*, *boxplot*, curva de rarefação,

árvores filogenéticas, análise de diversidade com busca em bancos de dados de proteínas (*GenBank* e M5NR, por exemplo) e de RNA (*Greengenes* e M5RNA, por exemplo) e análise funcional com bancos de dados (GLASS et al., 2010; MEYER et al., 2008).

FIGURA 6 – FLUXOGRAMA DE ANÁLISES QIIME E MG-RAST



FONTE: Modificada de D'ARGENIO et al. (2014).
LEGENDA: Fluxograma de análises dos programas QIIME e MG-RAST para dados de sequenciamento de nova geração para amostras de 16S rRNA. A principal diferença está na etapa de identificação e clusterização das sequências de 16S rRNA.

A figura 6 mostra o fluxograma comparativo das análises de dados de sequenciamento de nova geração para 16S rRNA realizadas pelos programas QIIME e MG-RAST (D'ARGENIO et al. 2014). Embora sigam basicamente o mesmo pipeline, os programas usam estratégias diferentes para a clusterização e identificação das sequências de 16S rDNA. O MG-

RAST faz a pesquisa das sequências de 16S rRNA usando o algoritmo *Blast-Like-Alignment* (BLAT) contra um banco de dados reduzido, uma versão clusterizada com 90% de identidade do banco de dados SILVA. As sequências são selecionadas e agrupadas com 97% de identidade e as mais longas são selecionadas como representativas para cada cluster e são submetidas a uma nova pesquisa com o BLAT contra os bancos de dados de 16S rRNA. Usando a abordagem OTU-*picking*, o QIIME usa um algoritmo de clusterização (padrão UCLUST) para agrupar as sequências por unidades taxonômicas com 97% de identidade, depois são selecionadas as sequências representativas para cada OTU e são submetidas às análises com banco de dados.

Para a análise de diversidade com o programa QIIME foram usados os *scripts* em *Phyton* (TABELA 4) sobre as amostras de 16S rDNA (MA02, MA05, MA07, MAF1, MAF2 e MAF3) usando o banco de dados *Greengenes* (DESANTIS et al., 2006) com 97% de identidade e o método de busca USEARCH como parâmetros.

TABELA 4 – *SCRIPTS* QIIME E DESCRIÇÕES

Script	**Descrição**
Pick_open_reference_otus.py	Criação das unidades taxonômicas funcionais (OTU's) e árvores filogenéticas.
Summarize_taxa_through_plots.py	Cria tabelas e gráficos taxonômicos baseados em tabelas de OTUs.
Alpha_rarefaction.py	Gera tabelas OTUs rarefeitas, calcula métricas de alfa diversidade para cada tabela OTU rarefeita e compara os resultados, e gera tabelas de rarefação alfa.
Core_diversity_analyses.py	*Workflow* dos *scripts* que geram resultados de alfa e beta diversidade, análise de componentes principais, gráficos de distância e taxonômicos.

FONTE: QIIME *SCRIPTS*

Com o servidor MG-RAST foi feita a análise de diversidade usando os bancos de dados de RNA *Greengenes* e M5RNA, que combina os bancos de dados SILVA (PRUESSE et al., 2007), *Greengenes* e RDP (COLE et al., 2003), (LINDGREEN et al., 2016) com 97% de identidade e os bancos de dados de proteína M5NR (WILKE et al., 2012), *RefSeq* (PRUITT, TATUSOVA e MAGLOTT, 2004), *GenBank* (BENSON et al., 2013), KEGG (KANEHISA e GOTO, 2000), SEED (OVERBEEK et al., 2005), PATRIC (GILLESPIE et al., 2011), IMG (MARKOWITZ et al., 2012), eggNOG (JENSEN et al., 2008), *SwissProt* e *TrEMBL* (BAIROCH e APWEILER, 1999) com 60% de identidade para as amostras de DNA total (MAF1, MAF2 e MAF3) *Illumina MiSeq* e *Ion Proton*. Além das análises com o banco de dados *Greengenes* com 97% de identidade para as amostras de 16S rRNA (MA02, MA05, MA07, MAF1, MAF2 e MAF3).

5.5 ANÁLISE FUNCIONAL

As análises funcionais foram feitas usando o servidor MG-RAST com comparações com o banco de dados COG (TATUSOV et al., 2000) e SEED *Subsystems* com 60% de identidade para as amostras de DNA total (MAF1, MAF2 e MAF3) *Illumina MiSeq* e *Ion Proton*. O banco de dados KEGG foi utilizado para reconstrução de vias metabólicas usando 60% de identidade.

6 RESULTADOS E DISCUSSÃO

6.1 SEQUENCIAMENTO NGS

No trabalho anterior, de Faoro e colaboradores (2010), o gene 16S rDNA de 10 amostras de solo da Floresta Atlântica paranaense foi amplificado, clonado em vetores e transformados na bactéria hospedeira *Escherichia coli* TOP10 originando 10 bibliotecas gênicas que foram sequenciadas pelo método de Sanger. A partir destas foi possível extrair 754 sequências completas das regiões V1-V2 do gene 16S rDNA com tamanhos entre 234-341 pb. Neste trabalho, as amostras MA02, MA05 e MA07 escolhidas dentre as 10 amostras do trabalho anterior como representantes das alta, média e baixa altitudes, respectivamente, junto com as amostras coletadas em 2007 (MAF1, MAF2 e MAF3) tiveram seu gene 16S rDNA amplificado e foram sequenciadas pela plataforma NGS *Illumina MiSeq* gerando um total de 371.561 sequências de até 300 pb que foram usadas para análises de diversidade (TABELA 5).

TABELA 5 – DADOS DE SEQUENCIAMENTO 16S rDNA

	MA02	**MA05**	**MA07**	**MAF1**	**MAF2**	**MAF3**
Total de *reads*	4.912	131.459	160.849	53.441	14.605	6.295
Total pb	1.232.215	32.987.492	40.360.939	13.410.107	3.664.992	1.579.577

FONTE: O autor (2017).

Separadamente, o DNA das amostras ambientais de 2007 foi sequenciado nas plataformas NGS *Illumina MiSeq* e *Ion Proton* gerando um total de 66.938.372 sequências, sendo 66.936.100 sequências após a remoção dos adaptadores das sequências geradas pelo sequenciador *Ion Proton* (TABELA 6). Durante as análises de diversidade, funcional e montagem de genomas parciais, foram usadas as sequências do *Ion Proton* sem os adaptadores.

TABELA 6 – DADOS DE SEQUENCIAMENTO DO DNA TOTAL

		MAF1	**MAF2**	**MAF3**
Proton	Total de *reads* (bruto)	6.629.692	29.481.055	8.150.945
	Total de pb (bruto)	718.927.473	3.160.331.413	850.651.840
	Total de *reads* (*trimming*)	6.629.647	29.478.885	8.150.888
	Total de pb (*trimming*)	717.832.937	3.153.760.570	849.299.269
MiSeq	Total de *reads*	11.952.048	6.349.518	4.375.114
	Total de pb	2.903.972.402	1.499.008.698	1.079.331.503
Total de *reads**		18.581.740	35.830.573	12.526.059
Total de pb*		3.622.899.875	4.659.340.111	1.929.983.343

FONTE: O autor (2017).
NOTA: *Totais dos valores brutos para *Illumina MiSeq* e *Ion Proton*.

6.2 MONTAGEM DE GENOMAS PARCIAIS

Com os dados de sequenciamento *Illumina MiSeq* e *Ion Proton* foram feitas 38 montagens *de novo* com os montadores *CLC Genomics Workbench* e MEGAHIT (com e sem o parâmetro "*--presets meta-large*") (TABELA 7).

TABELA 7 – MONTAGENS *DE NOVO*

	MiSeq	**Proton**	**Híbridas**
CLC	3	3	3
MEGAHIT v0.1	4	3	4
MEGAHIT v1.0.6	8	6	4

FONTE: O autor (2017).

Apesar de o MEGAHIT ser um programa montador para dados metagenômicos, mesmo usando o parâmetro *"--presets meta-large*" que é próprio para a montagem de sequências de DNA de solo, o *CLC Genomics Workbench* conseguiu utilizar maior quantidade de sequências durante as montagens e teve melhor performance em relação à quantidade e tamanho dos *contigs* montados. Levando esses critérios em consideração, a montagem com o *CLC Genomics Workbench* para a amostra MAF1 (*MiSeq*) foi escolhida como a melhor montagem. A tabela 8 mostra os resultados de todas as 38 montagens feitas.

TABELA 8 – MONTAGENS *CLC GENOMICS WORKBENCH* E MEGAHIT

continua

	Illumina MiSeq			
	MAF1	**MAF2 (S2 e S7)***		**MAF3**
MEGAHIT 1.0				
Total de *contigs*	554.484	67.514	130.492	147.007
Min *contig*	200	200	200	200
Max *contig*	6.324	3.727	3.771	4.914
N50	478	477	489	448
Cobertura	491	480	490	455
Total de bases	272.367.510	32.404.554	63.919.331	66.901.735
MEGAHIT 1.0.6				
Total de *contigs*	568.343	70.293	132.931	153.665
Min *contig*	200	200	200	200
Max *contig*	6.239	3.218	3.758	4.914
N50	477	474	486	446
Cobertura	488	474	486	448
Total de bases	277.143.754	33.333.467	64.658.640	68.882.878
MEGAHIT 1.0.6 (--presets meta-large)				
Total de *contigs*	500.321	58.685	121.320	138.838
Min *contig*	200	200	200	200
Max *contig*	7.244	3.080	3.797	4.846
N50	481	472	479	433
Cobertura	478	458	467	425
Total de bases	239.055.328	26.899.936	56.715.896	59.011.820

TABELA 8 – MONTAGEM *CLC GENOMICS WORKBENCH*

continuação

	Illumina MiSeq		
	MAF1	**MAF2 (S2 e S7)**	**MAF3**
CLC Genomics Workbench			
Total de *contigs*	693.915	391.043	260.326
Min *contig*	25	22	27
Max *contig*	35.630	9.997	6.699
N50	460	471	446
Cobertura	414	418	393
Total de bases	287.348.098	163.442.629	102.272.080

	Ion Proton		
	MAF1	**MAF2**	**MAF3**
MEGAHIT 1.0			
Total de *contigs*	276	43.423	969
Min *contig*	210	201	215
Max *contig*	4611	2.897	4.541
N50	345	374	345
Cobertura	378	393	366
Total de bases	104.404	17.057.060	354.282
MEGAHIT 1.0.6			
Total de *contigs*	318	46.099	1.088
Min *contig*	210	201	210
Max *contig*	3.993	3.015	6.291
N50	342	373	345
Cobertura	372	392	364
Total de bases	118.162	18.063.673	396.149
MEGAHIT 1.0.6 (--presets meta-large)			
Total de *contigs*	473	64.824	1.556
Min *contig*	208	200	203
Max *contig*	5.895	5.083	4.984
N50	305	351	310
Cobertura	331	370	330
Total de bases	156.416	23.980.502	513.504
CLC Genomics Workbench			
Total de *contigs*	6.354	138.633	10.209
Min *contig*	141	116	162
Max *contig*	5.401	5.286	4.949
N50	224	308	249
Cobertura	235	316	258
Total de bases	1.494.826	43.786.425	2.634.037

	Híbrida			
	MAF1	**MAF2 (S2 e S7)**		**MAF3**
MEGAHIT 1.0				
Total de *contigs*	703.880	163.086	265.268	214.842
Min *contig*	200	200	200	200
Max *contig*	7.143	5.122	5.972	6.292
N50	471	439	450	446
Cobertura	485	448	459	451
Total de bases	341.635.992	73.030.764	121.732.859	96.835.164
MEGAHIT 1.0.6				
Total de *contigs*	722.563	171.981	275.407	219.780
Min *contig*	200	200	200	200
Max *contig*	7.540	5.121	6.157	4.961
N50	470	435	448	430
Cobertura	483	444	456	419
Total de bases	348.824.963	76.276.320	125.536.020	92.015.150

TABELA 8 – MONTAGEM *CLC GENOMICS WORKBENCH*

conclusão

	Híbrida		
	MAF1	**MAF2 (S2 e S7)**	**MAF3**
CLC Genomics Workbench			
Total de *contigs*	754.457	625.414	65.460
Min *contig*	15	15	15
Max *contig*	35.630	9.516	4.984
N50	445	414	448
Cobertura	406	391	415
Total de bases	306.684.868	244.813.763	27.162.393

FONTE: O autor (2017).

NOTA: *Foram feitas 2 corridas da amostra MAF2 que foram unidas durante as montagens com o *CLC Genomics Workbench*, porém o MEGAHIT não aceitou a junção dos arquivos.

Durante a montagem, foram gerados 16.426 *contigs* acima de 1.000 pb. Estes foram comparados com o banco de dados *GenBank* usando a ferramenta *BLAST at NCBI* com o *blastn*. O *BLAST at NCBI* gerou uma tabela na qual lista todos os *contigs* analisados, a quantidade de *hits* obtidos para cada um e os classifica por menor *e-value*, maior % de identidade, maior % de positividade, maior tamanho do *hit* e maior *bit score* (APÊNDICE 1). A partir desta tabela foram filtrados todos os *contigs* que tiveram ao menos um *hit* com alguma sequência do banco de dados, gerando um total de 8.289 *contigs* e 41.445 sequências classificadas[2]. Essas 41.445 sequências foram filtradas, restando 3.010 sequências únicas que foram organizadas em ordem decrescente conforme o número de ocorrências. Dentre as 3.010 sequências únicas, foram selecionadas 31 sequências com mais de 100 ocorrências (TABELA 9).

A sequência com o maior número de ocorrências, 7,34% de 41.445 ocorrências, é referente ao genoma da bactéria *Solibacter usitatus Ellin6076*, pertencente ao filo *Acidobacteria*, que foi identificada como o melhor *hit* durante as análises de Faoro e colaboradores (2012). A tabela também reflete o que foi identificado durante as análises de diversidade: a predominância dos filos *Proteobacteria* (58%) e *Acidobacteria* (29%) entre as sequências. Além disso, podem ser observadas a presença de sequências de genomas candidatos, que são genomas montados a partir de dados de DNA ambiental e não possuem informações de experimentos em laboratório.

Os arquivos fasta de cada uma das 31 sequências foram baixados do NCBI e submetidos ao *CLC Genomics Workbench* para o mapeamento de *reads* na referência. Foram mapeadas sobre os genomas de referência as 66.936.100 *reads* referentes aos dados de sequenciamento *Illumina MiSeq* e *Ion Proton* (MAF1, MAF2 e MAF3) (TABELA 9).

[2] *Contigs* com 1 ou mais *hits* X categorias da tabela do *blastn*: 8.289 *contigs* x 5 *categorias* = 41.445 sequências classificadas (sequências não únicas).

TABELA 9 – OCORRÊNCIAS *BLAST AT NCBI*

continua

Nº de Acesso	Definição	Filo Taxonômico	Quant. de Ocorrências	% de ocorrência	% de mapeio
CP000473	*Solibacter usitatus Ellin6076, complete genome.*	*Acidobacteria*	3042	7,34	47,28
CP000360	*Candidatus Koribacter versatilis Ellin345, complete genome.*	*Acidobacteria*	3011	7,27	53,52
CP011801	*Nitrospira moscoviensis strain NSP M-1, complete genome.*	*Nitrospirae*	2174	5,25	41,59
FP929003	*Candidatus Nitrospira defluvii chromosome, complete genome.*	*Nitrospirae*	1108	2,67	31,42
CP001472	*Acidobacterium capsulatum ATCC 51196, complete genome*	*Acidobacteria*	1033	2,49	55,95
LN885086	*Nitrospira sp. ENR4 genome assembly NiCh1, chromosome: 1.*	*Nitrospirae*	698	1,68	36,67
CP003130	*Granulicella mallensis MP5ACTX8, complete genome.*	*Acidobacteria*	641	1,55	37,70
CP015136	*Acidobacteria bacterium DSM 100886, complete genome.*	*Acidobacteria*	622	1,50	42,97
CP003379	*Terriglobus roseus DSM 18391, complete genome*	*Acidobacteria*	465	1,12	38,78
CP002480	*Granulicella tundricola MP5ACTX9, complete genome.*	*Acidobacteria*	380	0,92	42,48
CP007128	*Gemmatirosa kalamazoonesis strain KBS708, complete genome.*	*Gemmatimonadetes*	369	0,89	47,67
CP002467	*Terriglobus saanensis SP1PR4, complete genome.*	*Acidobacteria*	365	0,88	36,06
CP007440	*Rhodoplanes sp. Z2-YC6860, complete genome.*	*Proteobacteria*	278	0,67	43,61
CP003969	*Sorangium cellulosum So0157-2, complete genome.*	*Proteobacteria*	163	0,39	23,51
CP011509	*Archangium gephyra strain DSM 2261, complete genome.*	*Proteobacteria*	162	0,39	24,83
CP011806	*Acidobacteria bacterium Mor1 sequence.*	*Acidobacteria*	156	0,38	36,63
CP016428	*Bradyrhizobium icense strain LMTR 13, complete genome.*	*Proteobacteria*	148	0,36	41,30
CP003364	*Singulisphaera acidiphila DSM 18658, complete genome.*	*Planctomycetes*	125	0,30	23,18
CP011125	*Sandaracinus amylolyticus strain DSM 53668, complete genome.*	*Proteobacteria*	124	0,30	28,33
CP001804	*Haliangium ochraceum DSM 14365, complete genome.*	*Proteobacteria*	124	0,30	30,71
LN907826	*Bradyrhizobium sp. G22 genome assembly, chromosome: I.*	*Proteobacteria*	119	0,29	42,40

TABELA 9 – OCORRÊNCIAS *BLAST AT NCBI*

conclusão

Nº de Acesso	Definição	Filo Taxonômico	Quant. de Ocorrências	% de ocorrência	% de mapeio
LT607803	*Variovorax sp. HW608 genome assembly, chromosome: I.*	*Proteobacteria*	117	0,28	40,90
CP012333	*Labilithrix luteola strain DSM 27648, complete genome.*	*Proteobacteria*	110	0,27	26,26
CP000769	*Anaeromyxobacter sp. Fw109-5, complete genome.*	*Proteobacteria*	109	0,26	46,86
CP001032	*Opitutus terrae PB90-1, complete genome.*	*Verrucomicrobia*	109	0,26	36,84
CP001715	*Candidatus Accumulibacter phosphatis clade IIA str. UW-1, complete genome.*	*Proteobacteria*	106	0,26	40,63
CP004025	*Myxococcus stipitatus DSM 14675, complete genome.*	*Proteobacteria*	105	0,25	27,52
CP011971	*Steroidobacter denitrificans strain DSM 18526, complete genome.*	*Proteobacteria*	102	0,25	43,87
CP000251	*Anaeromyxobacter dehalogenans 2CP-C, complete genome.*	*Proteobacteria*	101	0,24	51,19
CP003389	*Corallococcus coralloides DSM 2259, complete genome.*	*Proteobacteria*	101	0,24	28,41
CP006003	*Myxococcus fulvus 124B02, complete genome.*	*Proteobacteria*	100	0,24	26,95

FONTE: O autor (2017).

As leituras mapeadas foram extraídas e submetidas à montagem *de novo* com o *CLC Genomics Workbench* (TABELA 10). Durante a montagem, foram montadas em média 8% das *reads* mapeadas durante o *BLAST at NCBI* e as montagens ficaram muito fragmentadas. Foram obtidos muitos *contigs* em cada uma das montagens, aproximadamente 1.000 *contigs*, com pouquíssimos ou nenhum *contig* acima de 1.000 pb.

Embora o sequenciamento NGS possibilitou o acesso a um grande volume de sequências de DNA em comparação ao sequenciamento com Sanger realizado anteriormente (FAORO et al., 2010), o baixo mapeio de leituras e a fragmentação das montagens obtidas levanta duas hipótese: 1) a estirpe de *S. usitatus* encontrada no solo da Floresta Atlântica paranaense pode ser diferente da disponível no banco de dados; 2) é necessário um volume de dados muito maior para extrair genomas completos de dados metagenômicos de solo. Segundo Luo e colaboradores (2011), para montar genomas individuais a partir de metagenoma de solo é necessária uma cobertura de 20x do genoma referência. Ou seja, para montar o genoma da estirpe *S. usitatus* que tem o tamanho de 9.965.640 pb, seriam necessários aproximadamente 199 Mbp.

TABELA 10 – MONTAGENS *DE NOVO* PARA SEQUÊNCIAS DE GENOMA REFERÊNCIA

continua

Nº de Acesso	**Definição**	**Tamanho do genoma**	***Reads* mapeadas**	***Reads* montadas**	**Total de *contigs***	***Contigs* > 1000 pb**	**Min *contig***	**Max *contig***	**N50**	**Cobertura**	**Total pb**
CP000473	*Solibacter usitatus Ellin6076, complete genome.*	9.965.640	1.884.568	292.781	11.510	39	22	2.180	357	341	3.929.523
CP000360	*Candidatus Koribacter versatilis Ellin345, complete genome.*	5.650.368	1.209.546	148.633	4.737	3	43	1.490	310	308	1.460.981
CP011801	*Nitrospira moscoviensis strain NSP M-1, complete genome.*	4.589.485	763.564	70.476	2.923	25	38	1.880	372	364	1.062.582
FP929003	*Candidatus Nitrospira defluvii chromosome, complete genome.*	4.317.083	542.541	49.208	2.168	17	18	1.767	364	352	763.856
CP001472	*Acidobacterium capsulatum ATCC 51196, complete genome*	4.127.356	923.720	100.841	2.829	0	71	909	296	298	842.955
LN885086	*Nitrospira sp. ENR4 genome assembly NiCh1, chromosome: 1*	3.295.117	483.291	46.752	1.902	11	21	1.908	365	354	673.171
CP003130	*Granulicella mallensis MP5ACTX8, complete genome.*	6.237.577	940.651	90.200	2.701	2	46	1.222	303	304	819.818
CP015136	*Acidobacteria bacterium DSM 100886, complete genome.*	7.480.314	1.285.808	96.423	2.586	3	101	1.388	298	303	782.352
CP003379	*Terriglobus roseus DSM 18391, complete genome*	5.227.858	810.935	72.725	1.991	2	33	1.259	289	293	584.026
CP002480	*Granulicella tundricola MP5ACTX9, complete genome.*	4.309.153	732.277	81.068	2.307	1	81	1.086	291	296	681.973
CP007128	*Gemmatirosa kalamazoonesis strain KBS708, complete genome.*	5.311.527	1.012.721	77.057	2.629	0	92	978	324	320	841.299
CP002467	*Terriglobus saanensis SP1PR4, complete genome.*	5.095.226	734.987	56.898	1.523	2	123	1.104	300	300	457.503
CP007440	*Rhodoplanes sp. Z2-YC6860, complete genome.*	8.193.889	1.429.479	169.336	5.619	3	37	1.362	342	331	1.857.755
CP003969	*Sorangium cellulosum So0157-2, complete genome.*	14.782.125	1.390.188	83.984	2.407	0	21	993	309	307	739.328
CP011509	*Archangium gephyra strain DSM 2261, complete genome.*	12.489.432	1.240.606	70.521	1.966	0	90	991	299	303	595.415
CP011806	*Acidobacteria bacterium Mor1 sequence.*	5.989.545	877.543	41.479	834	0	18	805	295	294	245.123

TABELA 10 – MONTAGENS *DE NOVO* PARA SEQUÊNCIAS DE GENOMA REFERÊNCIA

conclusão

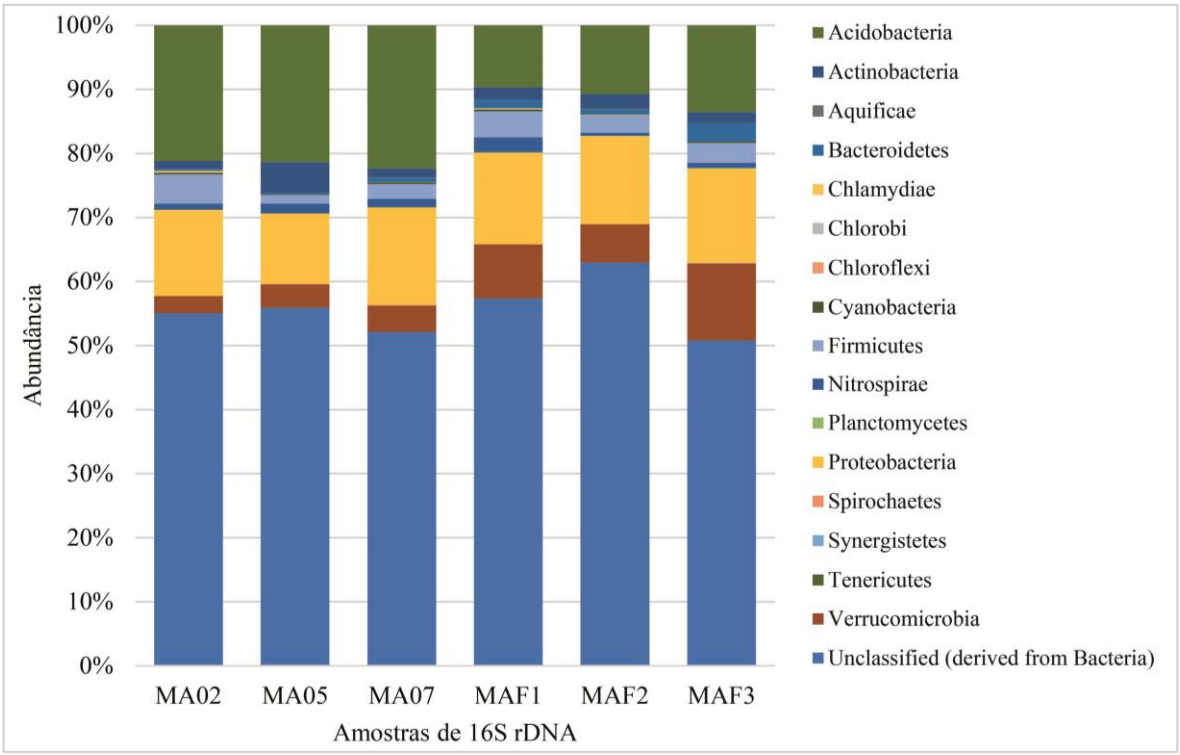
Nº de Acesso	**Definição**	**Tamanho do genoma**	***Reads* mapeadas**	***Reads* montadas**	**Total de *contigs***	***Contigs* > 1000 pb**	**Min *contig***	**Max *contig***	**N50**	**Cobertura**	**Total pb**
CP016428	*Bradyrhizobium icense strain LMTR 13, complete genome.*	8.322.773	1.374.952	165.448	6.052	38	18	2.646	363	348	2.104.757
CP003364	*Singulisphaera acidiphila DSM 18658, complete genome.*	9.629.675	893.031	30.793	1.016	0	108	883	294	303	307.692
CP011125	*Sandaracinus amylolyticus strain DSM 53668, complete genome.*	10.327.335	1.170.418	67.702	1.618	0	117	874	299	300	486.153
CP001804	*Haliangium ochraceum DSM 14365, complete genome.*	9.446.314	1.160.485	60.914	1.441	1	24	1.034	299	299	430.858
LN907826	*Bradyrhizobium sp. G22 genome assembly, chromosome: I*	9.022.917	1.530.231	191.021	7.092	23	62	2.127	360	344	2.437.939
LT607803	*Variovorax sp. HW608 genome assembly, chromosome: I.*	7.733.665	1.265.167	96.162	3.122	7	116	1.450	323	321	1.000.645
CP012333	*Labilithrix luteola strain DSM 27648, complete genome.*	12.191.466	1.280.801	79.003	2.275	2	18	1.149	320	316	717.909
CP000769	*Anaeromyxobacter sp. Fw109-5, complete genome.*	5.277.990	989.335	69.578	1.804	2	7	1.157	295	299	540.265
CP001032	*Opitutus terrae PB90-1, complete genome.*	5.957.605	877.852	34.260	901	0	97	728	287	291	261.914
CP001715	*Candidatus Accumulibacter phosphatis clade IIA str. UW-1, complete genome.*	5.058.518	822.065	51.870	1.357	1	61	1.116	313	312	423.745
CP004025	*Myxococcus stipitatus DSM 14675, complete genome.*	10.350.586	1.139.236	64.040	1.775	0	86	864	296	300	532.175
CP011971	*Steroidobacter denitrificans strain DSM 18526, complete genome.*	3.467.246	608.485	44.435	988	4	18	1.571	324	317	313.639
CP000251	*Anaeromyxobacter dehalogenans 2CP-C, complete genome.*	5.013.479	1.026.505	71.645	1.869	0	74	867	296	300	559.847
CP003389	*Corallococcus coralloides DSM 2259, complete genome.*	10.080.619	1.145.720	67.582	1.806	0	51	960	302	304	548.626
CP006003	*Myxococcus fulvus 124B02, complete genome.*	11.048.835	1.190.899	70.151	1.866	0	96	768	295	299	557.985

FONTE: O autor (2017).

6.3 ANÁLISE DE DIVERSIDADE

Os resultados das análises de diversidade, utilizando somente as sequências de 16S rDNA, com os programas QIIME e MG-RAST baseadas em comparações com o banco de dados *Greengenes* (97% de identidade) apresentaram diferenças entre si. A análise com o MG-RAST identificou 16 filos do domínio *Bacteria* (FIGURA 7), 14 deles em comum com a análise feita com o programa QIIME, que identificou 33 filos do domínio *Bacteria* e 3 filos do domínio *Archaea* (*Crenarchaeota*, *Euryarchaeota* e *Parvarchaeota*) (FIGURA 8). No entanto, apesar dos filos a mais identificados pelo QIIME, quando analisada a abundância dos 14 filos em comum encontrados com os dois programas, é possível notar maior abundância deles sobre os demais filos que apresentaram uma média de 3% de abundância. Dentre os filos identificados em comum, os filos *Acidobacteria* e *Proteobacteria* foram predominantes com 16,5% e 13,7%, respectivamente, dentre os 45% das sequências classificadas na análise com o MG-RAST; e 49,3% e 24,6%, respectivamente, para a análise com o QIIME.

FIGURA 7 – ANÁLISE DE DIVERSIDADE COM MG-RAST E *GREENGENES* PARA 16S rDNA

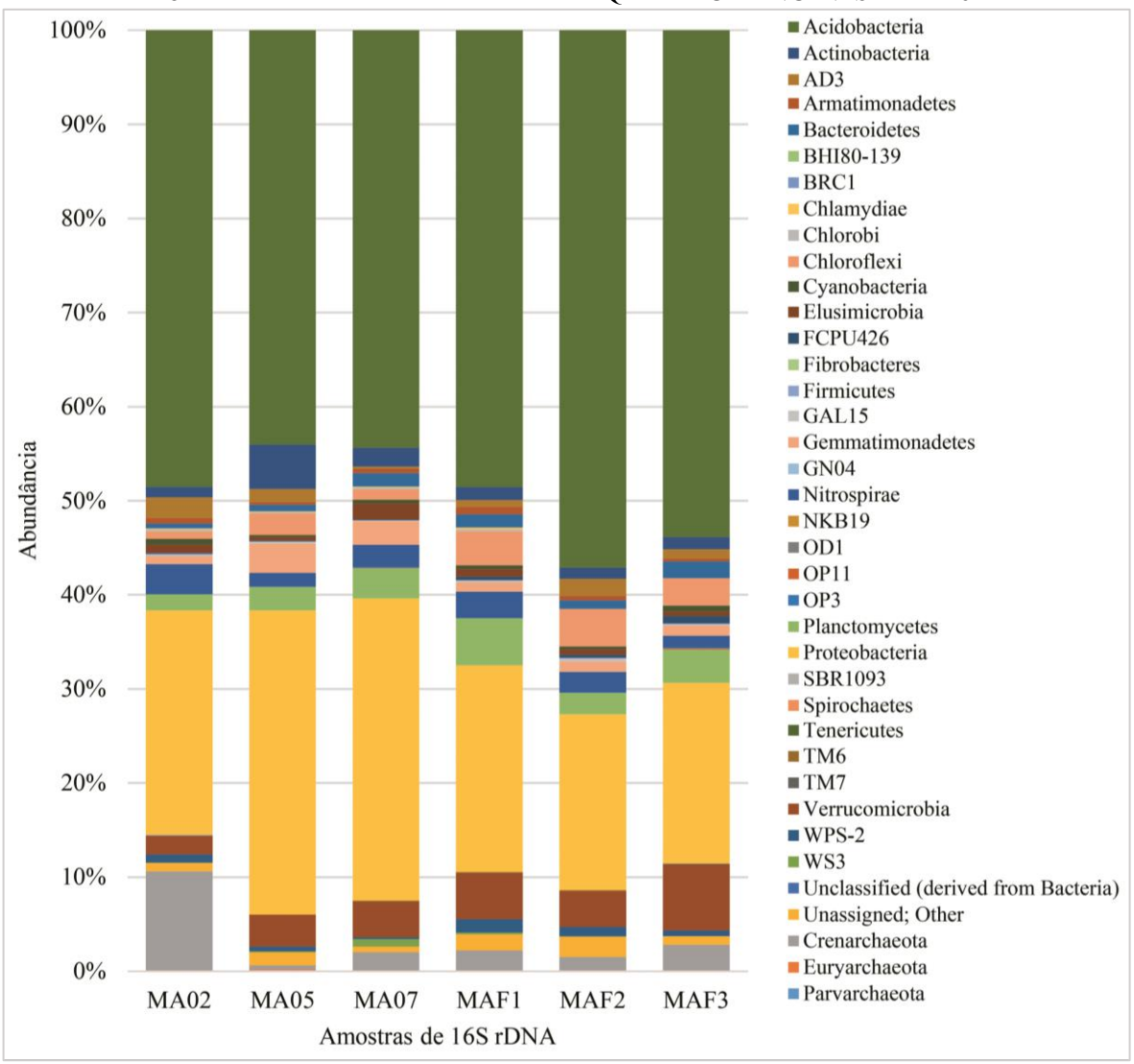


FONTE: O autor (2017).

LEGENDA: Análise de diversidade para amostras de 16S rDNA coletas em 2004 (MAs) e 2007 (MAFs) com o MG-RAST e *Greengenes* com 97% de identidade. Identificados 16 filos taxonômicos com predominância dos filos *Acidobacteria* (16,5%) e *Proteobacteria* (13,7%), em média. E aproximadamente 55,7% de sequências não classificadas.

Outro ponto a ser observado, é a quantidade de sequências não classificadas ("*unclassified (derived from Bacteria)*") e não atribuídas ("*Unassigned; Other*"). A análise com o QIIME teve uma quantidade ínfima de sequências não classificadas (quase nula) e não atribuídas (em média 1,2%), enquanto a análise pelo MG-RAST teve em média 55,7% das sequências "não classificadas".

FIGURA 8 – ANÁLISE DE DIVERSIDADE COM QIIME E *GREENGENES* PARA 16S RDNA



FONTE: O autor (2017).

LEGENDA: Análise de diversidade para amostras de 16S rDNA coletas em 2004 (MAs) e 2007 (MAFs) com o QIIME e *Greengenes* com 97% de identidade. Identificados 33 filos taxonômicos do domínio *Bacteria* e 3 filos do domínio *Archaea* (*Creanarchaeota*, *Euryarchaeota* e *Parvarchaeota*). Os filos mais abundantes foram *Acidobacteria* (49,3%) e *Proteobacteria* (24,6%), em média. E apresentou pouquíssimas sequências não classificadas derivadas do domínio *Bacteria* e, média, 1,2% de sequências não atribuídas.

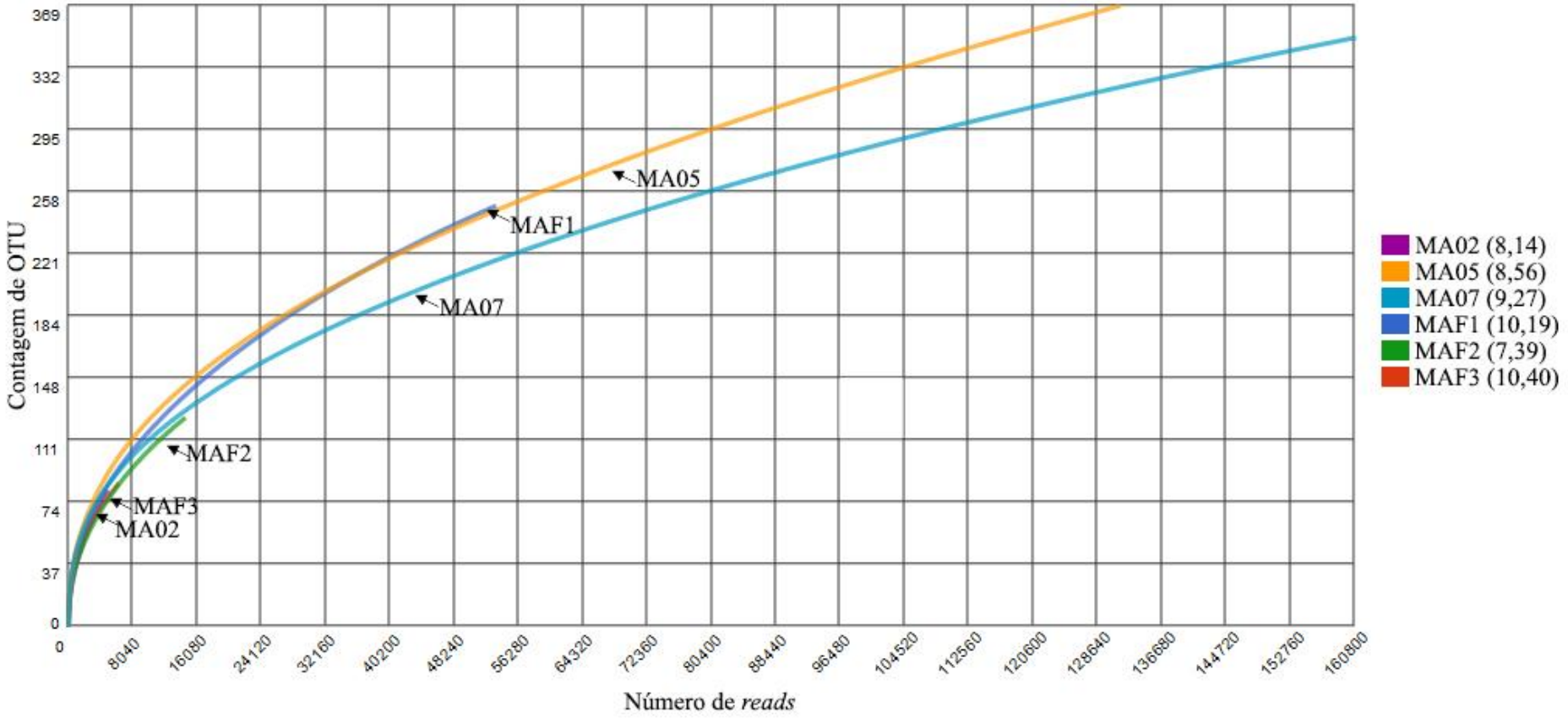
A diferença entre as abordagens usadas pelo MG-RAST e QIIME durante a identificação e clusterização das sequências de 16S rDNA, dependente da taxonomia e independente de taxonomia, respectivamente, pode ter contribuído para que o QIIME identificasse maior quantidade de filos, dentre eles filos candidatos ou com baixa representatividade nos bancos de dados (SUN et al., 2012).

Quando analisadas as amostras em relação ao período de coleta, não houve diferença entre os filos identificados entre as amostras nas análises de cada programa. No entanto há pequenas diferenças nas quantidades de sequências identificadas. Na análise com o MG-RAST (FIGURA 7), foram identificadas mais sequências do filo *Acidobacteria* nas amostras de 2004 (21,6%) que nas amostras de 2007 (11,3%), que apresentaram aumento para os filos *Bacteroidetes* (0,43% e 1,6%), *Proteobacteria* (13,2% e 14,2%) e *Verrucomicrobia* (3,4% e 8,8%) e maior quantidade de sequências não classificadas (54,3% e 57%).

A análise com o QIIME identificou o aumento de sequências do filo *Acidobacteria* para as amostras coletadas em 2007, em contradição com o resultado do MG-RAST. Foram identificados 45,6% para as amostras de 2004 contra 53% para as amostras de 2007. E identificadas 29,4% das sequências das amostras de 2004 para o filo *Proteobacteria* contra 20% para as amostras de 2007.

Com o MG-RAST foi construída a curva de rarefação baseada nos valores de alfa diversidade, que é uma forma de simular a riqueza e abundância de espécies em uma amostra. Na figura 9, pode-se observar as curvas de rarefação calculadas para as 6 amostras de 16S rDNA e que a amostra MA07 apresentar menor diversidade que MA05.

FIGURA 9 – CURVA DE RAREFAÇÃO COM MG-RAST E *GREENGENES* PARA 16S RDNA

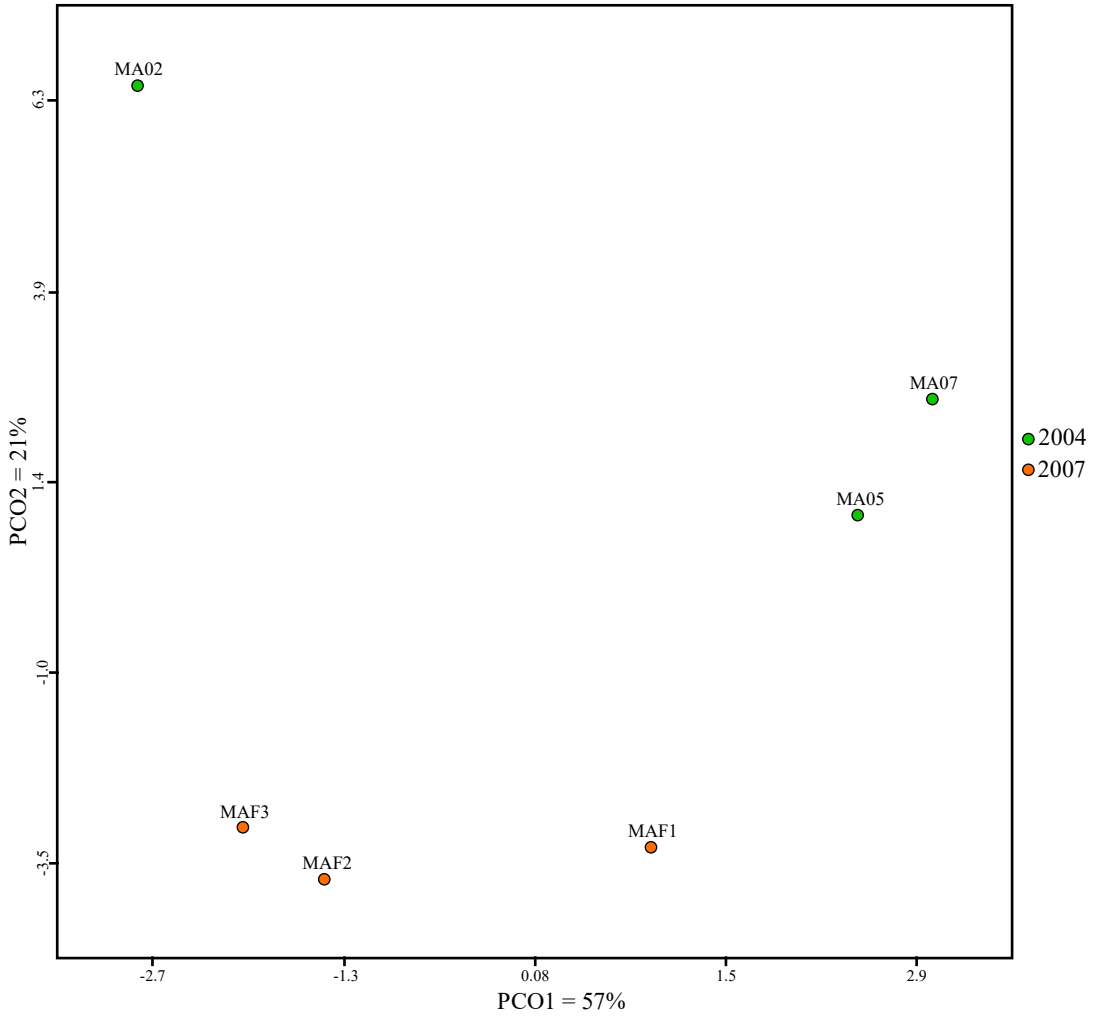


FONTE: O autor (2017).

LEGENDA: Curva de rarefação com alfa diversidade calculada pelo MG-RAST com Greengenes e 97% de identidade para amostras de 16S rDNA.

Análise de componentes principais (PCoA) realizada pelo MG-RAST para as amostras de 16S rDNA com o banco de dados *Greengenes* e 97% de identidade para dados normalizados e distância bray-curtis, revelou que quando observadas pelo eixo PC1, que explica 57% da relação entre as amostras, as amostras não apresentam agrupamento total por período de coleta, uma vez que as amostras MA02 e MAF1 (correspondentes à alta altitude) apresentam um comportamento diferente das amostras de baixa e média altitudes. Porém, se observado pelo eixo PC2, as amostras separam-se por período de coleta em -1.0 (FIGURA 10).

FIGURA 10 – PCOA COM MG-RAST E *GREENGENES* PARA 16S RRNA



FONTE: O autor (2017).

LEGENDA: Gráfico PCoA calculado para amostras de 16S rRNA pelo MG-RAST contra o banco de dados *Greengenes* com 97% de identidade para dados normalizados e método de distância bray-curtis. O eixo PCO1 explica 57% da relação e o eixo PCO2 (21%).

Durante as etapas de processamento e alinhamento das sequências de DNA total (MAF1, MAF2 e MAF3) das plataformas *Illumina MiSeq* e *Ion Proton* pelo MG-RAST, foram

preditas e identificadas sequências de 16S rDNA (TABELA 11). As sequências preditas são referentes às sequências identificadas pelo MG-RAST nos primeiros passos de processamento dos dados. Após a predição dessas sequências, o MG-RAST realiza pesquisas de similaridade entre as sequências de proteínas preditas e bancos de dados de proteínas e pesquisas de similaridade de ácidos nucleicos para sequências preditas de rRNA. Posteriormente, gera uma tabela informando dentre as sequências de proteínas e rRNA preditos inicialmente, quais foram realmente identificados após as buscas de similaridade.

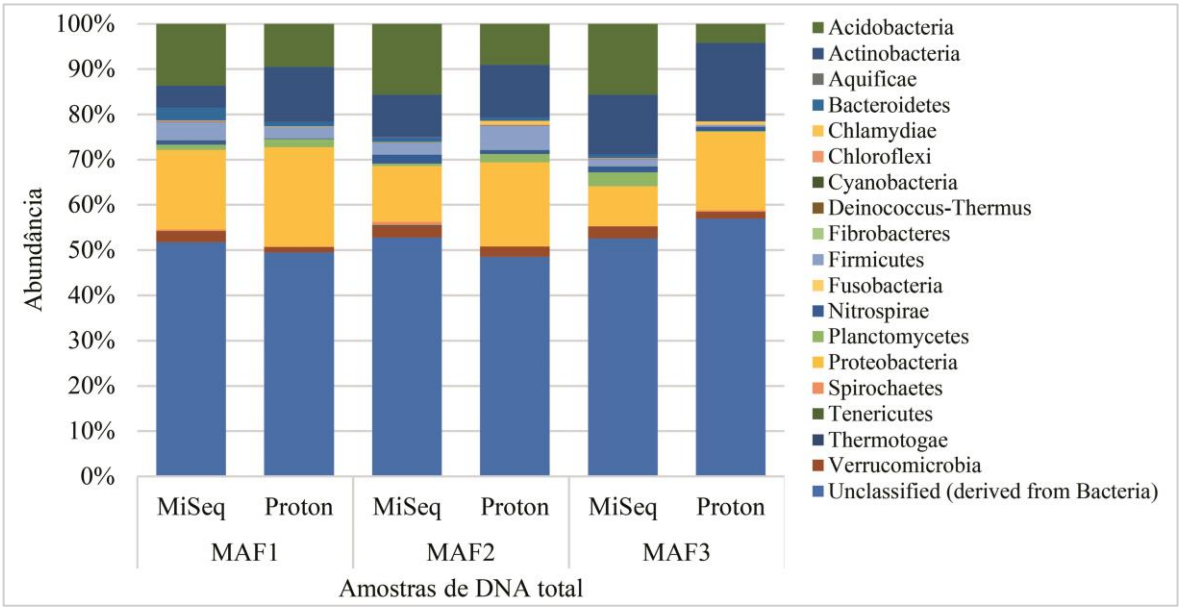
TABELA 11 – ANOTAÇÃO DE SEQUÊNCIAS DE 16S RDNA EM AMOSTRAS DE DNA TOTAL PELO MG-RAST

	MAF1		**MAF2**		**MAF3**	
	MiSeq	**Proton**	**MiSeq**	**Proton**	**MiSeq**	**Proton**
rRNAs preditos	903.028	329.332	475.571	1.582.201	340.425	617.076
rRNAs identificados	2.427	965	1.557	2.375	1.189	1.326

FONTE: O autor (2017).

Usando as sequências de rRNA identificadas nas amostras de DNA total, foram feitas análises de diversidade usando as amostras de DNA total para os bancos de dados de 16S rRNA, *Greengenes* e M5RNA, com 97% de identidade.

FIGURA 11 – ANÁLISE DE DIVERSIDADE COM MG-RAST E *GREENGENES* PARA DNA TOTAL



FONTE: O autor (2017).

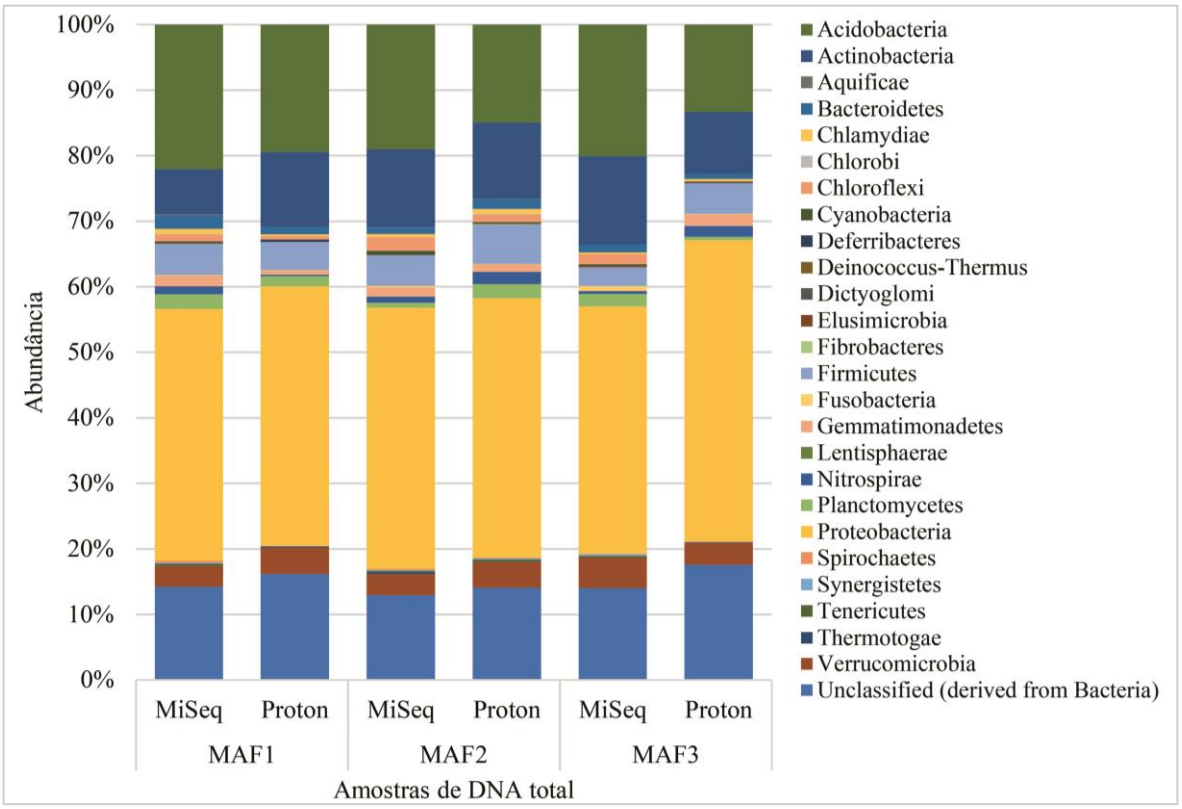
LEGENDA: Análise de diversidade para amostras de DNA total com o MG-RAST e *Greengenes* com 97% de identidade. Identificados 18 filos taxonômicos com predominância dos filos *Proteobacteria* (11,45%) e *Acidobacteria* (11,32%), em média. E aproximadamente 52% de sequências não classificadas.

A seleção dos bancos de dados de rRNA automaticamente limita a análise a sequências de 16S rDNA que foram sequenciadas ao acaso junto com o sequenciamento do DNA total.

O resultado da análise com o *Greengenes* identificou 18 filos do domínio *Bacteria*, dentre eles 14 em comum com a análise anterior. Observando a figura 11, é possível notar o aumento de sequências do filo *Proteobacteria* que se igualou em abundância com o filo *Acidobacteria*, em média 11,45% e 11,32%, respectivamente. Também houve uma pequena diminuição das sequências não classificadas (em média 52%).

Por ser uma combinação dos bancos de dados *Greengenes*, RDP e SILVA, a análise com o banco de dados M5RNA identificou todos os filos identificados pelas análises com o *Greengenes* para as amostras de 16S rDNA e DNA total e mais 5 filos adicionais do domínio *Bacteria* (FIGURA 12). Nesse caso houve uma inversão nos dois filos dominantes com *Proteobacteria* em primeiro e *Acidobacteria* em segundo. Além disso, o número de sequências não classificadas reduziu para a média de 14,8%.

FIGURA 12 – ANÁLISE DE DIVERSIDADE COM MG-RAST E M5RNA PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade para amostras de DNA total com o MG-RAST e M5RNA com 97% de identidade. Identificados 25 filos taxonômicos com predominância dos filos *Proteobacteria* (40,18%) e *Acidobacteria* (18,16%), em média. E aproximadamente 14,8% de sequências não classificadas.

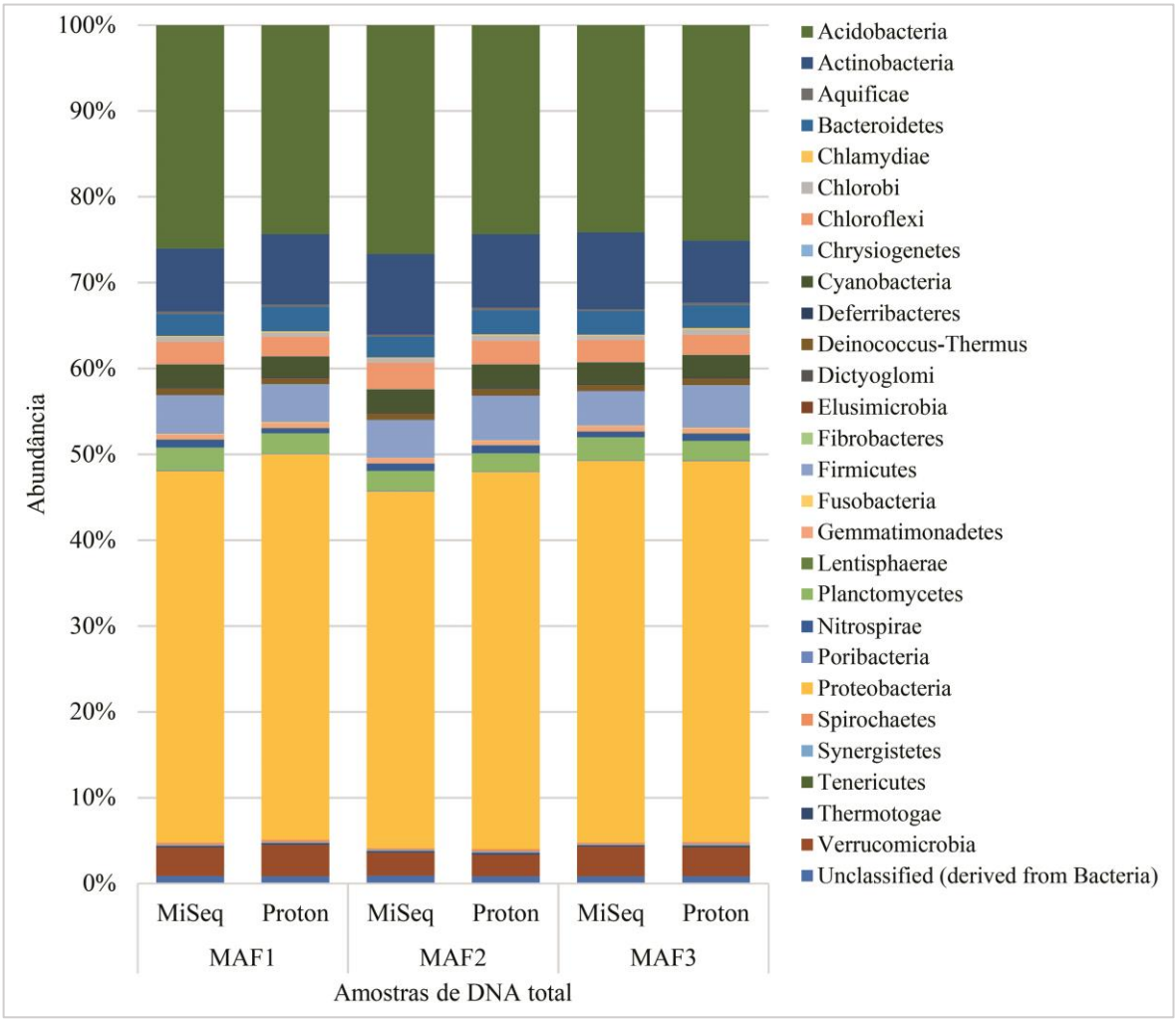
Durante a análise de diversidade com bancos de dados de proteínas, o MG-RAST usa as informações de anotação das sequências geradas no pré-processamento dos dados. A partir dessas informações, reconstrói o perfil filogenético das sequências baseando-se nas funções de cada proteína e realiza buscas por similaridade nos bancos de dados de proteínas.

Levando em consideração as diferenças entre as análises de diversidade com os bancos de dados de 16S rRNA e entre as técnicas de classificação de sequências dos programas, foram feitas análises de diversidade com as amostras de DNA total com comparações com os bancos de dados de proteínas (60% de identidade) disponibilizados pelo servidor MG-RAST. Ao todo, foram feitas 10 análises (APÊNDICE 2) com os bancos de dados M5NR, *GenBank*, SEED, IMG, KEGG, TrEMBL, PATRIC, *SwissProt*, eggNOG e *RefSeq*. O resultado mais representativo foi obtido na comparação com o banco de dados M5NR, tendo sido identificados 27 filos bacterianos (FIGURA 13).

Uma característica interessante dessa análise foi a pouca discrepância encontrada entre os resultados das duas tecnologias de sequenciamento, sendo que a maioria dos filos apresentaram uma distribuição muito semelhante, tanto para o banco M5NR quanto para os outros bancos. A maior discrepância encontrada entre os bancos foi em relação à distribuição dos filos *Acidobacteria* e *Proteobacteria*.

Quando considerado apenas sequências de proteínas, a abundância de *Acidobacteria* diminui quando comparada com as análises de sequências de 16S rDNA. Esse efeito pode ser explicado pela quantidade de sequências de proteínas de *Acidobacteria* e *Proteobacteria* depositadas nos bancos de dados. O filo *Proteobacteria* é um grupo importante com muitos organismos de importância médica, agrícola, farmacêutica, biotecnológica e de cultivo relativamente fácil em laboratório. Por esses motivos é mais estudado e tem maior depósito de sequências. O filo *Acidobacteria* é muito mais recente, de cultivo difícil em laboratório e com pouca importância para a sociedade. Assim, apresenta menos depósitos de sequências de proteínas nos bancos de dados.

FIGURA 13 – ANÁLISE DE DIVERSIDADE COM MG-RAST E M5NR PARA DNA TOTAL

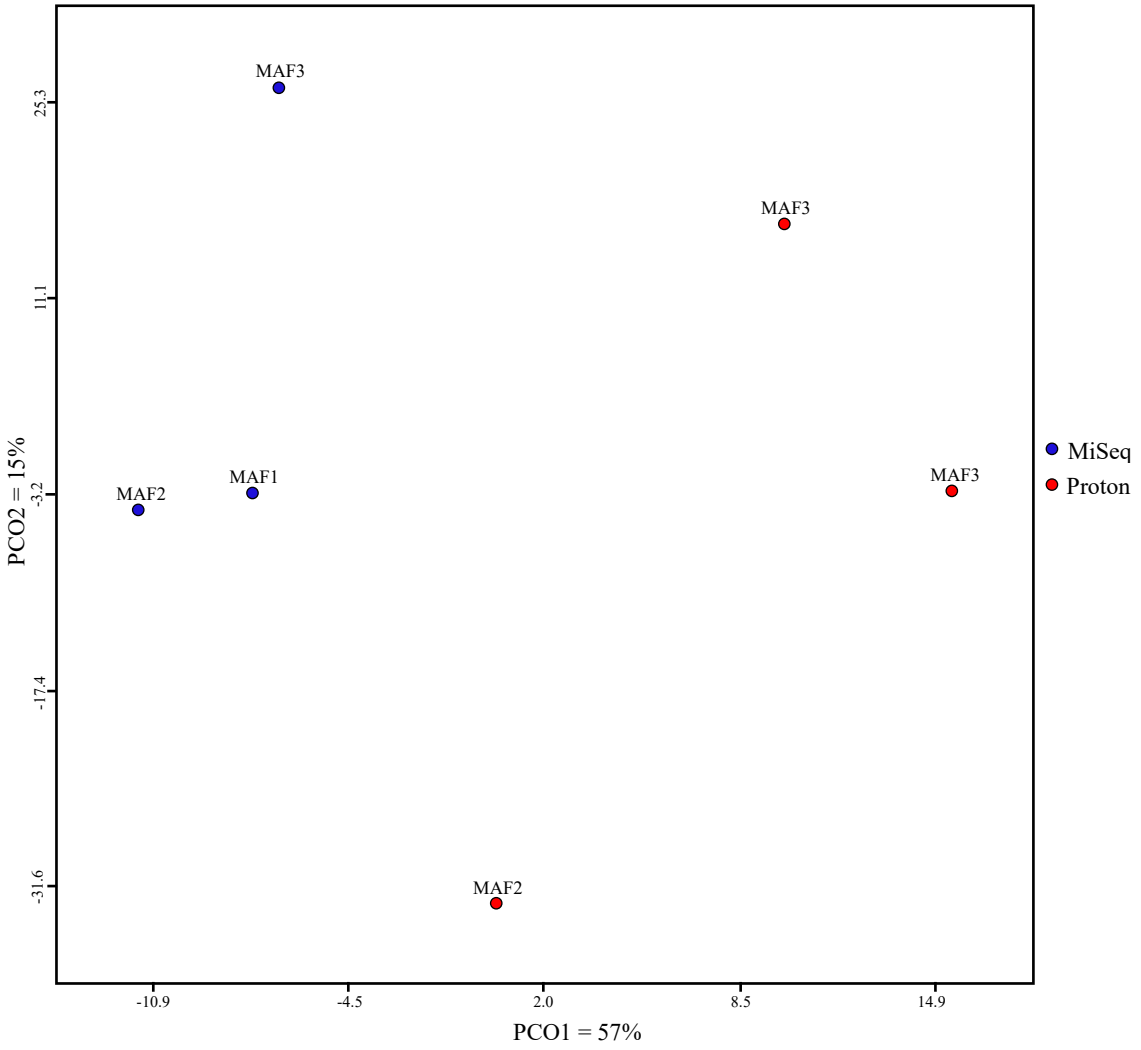


FONTE: O autor (2017).

LEGENDA: Análise de diversidade para amostras de DNA total com o MG-RAST e banco de dados M5NR com 60% identidade e escolhida como a representativa dentre as 10 análises com os bancos de dados de proteínas. 26 filos identificados e predominância dos filos *Proteobacteria* (43,74%) e *Acidobacteria* (25,11%). A quantidade de sequências não classificadas foi mínima, em média 0,8%.

A figura 14 mostra o PCoA gerado pelo MG-RAST para as amostras de DNA total com o banco de dados M5NR com 60% identidade para dados normalizados e método de distância bray-curtis. Assim como no PCoA gerado com o *Greengenes* (97% de identidade), as amostras estão agrupadas por plataforma de sequenciamento e apresentam comportamento parecido no plano do gráfico.

FIGURA 14 – PCOA COM MG-RAST E M5NR PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Gráfico PCoA calculado para amostras de DNA total pelo MG-RAST contra o banco de dados M5NR com 60% de identidade para dados normalizados e método de distância bray-curtis. As amostras estão agrupadas por plataforma de sequenciamento. O eixo PCO1 explica 57% da relação e o eixo PCO2 15%.

Os filos taxonômicos *Proteobacteria*, *Acidobacteria*, *Actinobacteria*, *Firmicutes* e *Verrucomicrobia* foram os filhos em maior abundância na maioria das análises, seguidos pelo filo *Cyanobacteria* que apresentou maior abundância que o filo *Verrucomicrobia* em algumas das análises (APÊNDICE 2). A quantidade de sequências não classificadas foi de aproximadamente 1%. Na tabela 12, está discriminada a ocorrência de cada filo nas 10 análises com os bancos de dados de proteínas.

TABELA 12 – REPRESENTAÇÃO DOS FILOS TAXONÔMICOS NAS ANÁLISES DE DNA TOTAL SEGUNDO DISTRIBUIÇÃO PARA CADA BANCO DE DADOS DE PROTEÍNAS

Filo	M5NR	SwissProt	RefSeq	GenBank	KEGG	TrEMBL	SEED	IMG	eggNOG	PATRIC
Acidobacteria	x	x	x	x	x	x	x	x	x	x
Actinobacteria	x	x	x	x	x	x	x	x	x	x
Aquificae	x	x	x	x	x	x	x	x	x	x
Bacteroidetes	x	x	x	x	x	x	x	x	x	x
Chlamydiae	x	x	x	x	x	x	x	x	x	x
Chlorobi	x	x	x	x	x	x	x	x	x	x
Chloroflexi	x	x	x	x	x	x	x	x	x	x
Chrysiogenetes	x	-	x	x	x	x	-	x	-	-
Cyanobacteria	x	x	x	x	x	x	x	x	x	x
Deferribacteres	x	-	x	x	x	x	x	x	-	x
Deinococcus-Thermus	x	x	x	x	x	x	x	x	x	x
Dictyoglomi	x	x	x	x	x	x	x	x	-	x
Elusimicrobia	x	x	x	x	x	x	x	x	-	x
Fibrobacteres	x	x	x	x	x	x	-	x	-	x
Firmicutes	x	x	x	x	x	x	x	x	x	x
Fusobacteria	x	x	x	x	x	x	x	x	x	x
Gemmatimonadetes	x	x	x	x	x	x	-	x	-	x
Lentisphaerae	x	-	x	x	-	-	-	x	-	x
Nitrospirae	x	x	x	x	x	x	-	x	-	x
Planctomycetes	x	x	x	x	x	x	x	x	x	x
Poribacteria	x	-	x	x	-	x	-	-	-	-
Proteobacteria	x	x	x	x	x	x	x	x	x	x
Spirochaetes	x	x	x	x	x	x	x	x	x	x
Synergistetes	x	-	x	x	x	x	x	x	x	x
Tenericutes	x	x	x	x	x	x	x	x	x	x
Thermotogae	x	x	x	x	x	x	x	x	x	x
Verrucomicrobia	x	x	x	x	x	x	x	x	x	x
unclassified	x	x	x	x	x	x	x	x	x	x

FONTE: O autor (2017).
NOTA: Os “x” representam os filos presentes em cada análise e os “-”, os filos ausentes.

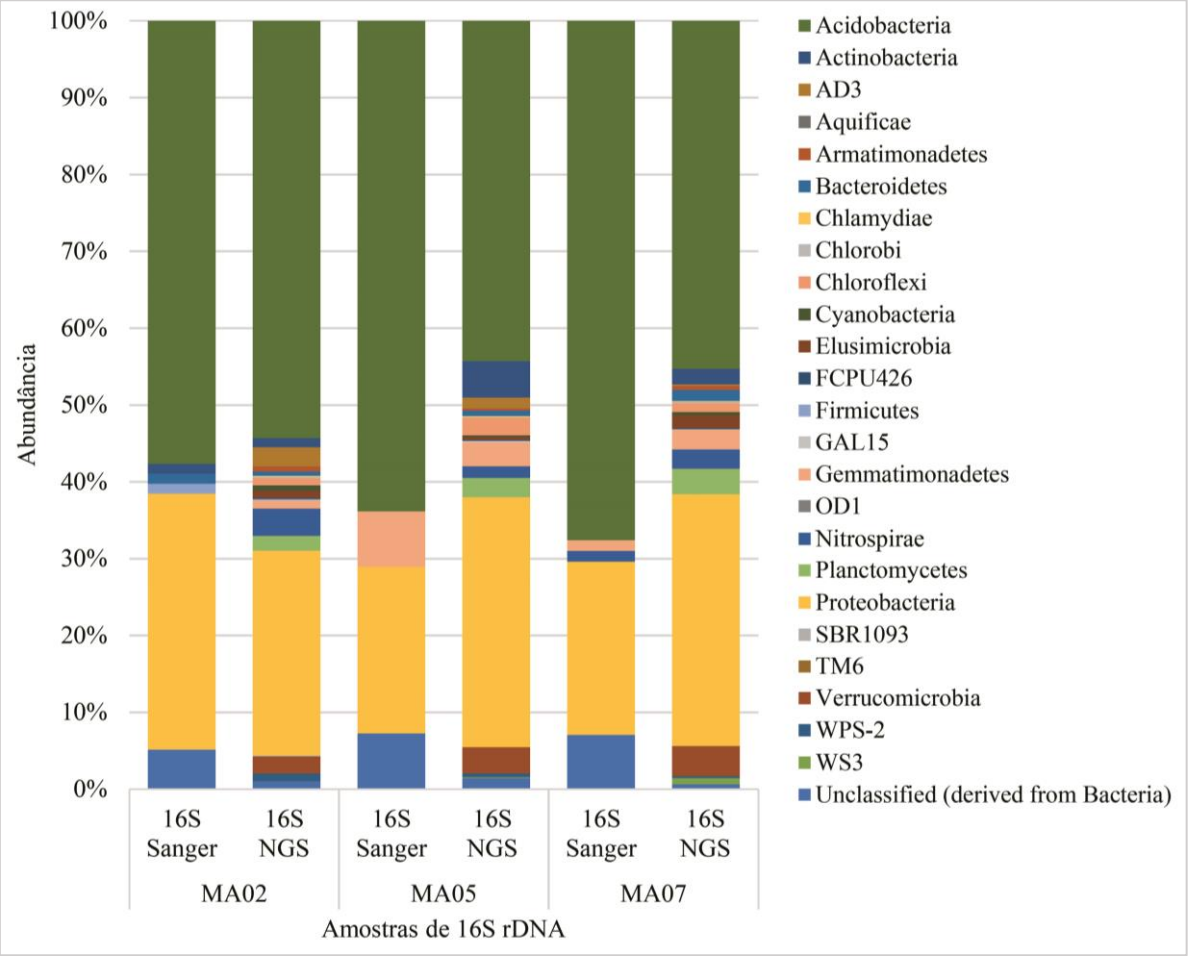
As diferenças entre as análises realizadas com os programas QIIME e MG-RAST mostram alguns pontos a serem observados. O primeiro deles é o método de classificação das sequências utilizado. Como já mencionado, o MG-RAST usa o método de classificação dependente da taxonomia no qual primeiramente faz uma pesquisa com o algoritmo BLAT contra um cluster de 90% de identidade do banco de dados SILVA, que é um banco de dados curado de sequências de rRNA. É de conhecimento geral que os bancos de dados curados possuem menos sequências depositadas do que os bancos que não possuem uma curadoria, devido ao tempo e esforço demandados por este processo. A presença de filos candidatos nesses bancos de dados é menor ainda, uma vez que é difícil curar o genoma de um organismo que não possui dados de experimentos laboratoriais, e essas sequências acabam não sendo classificadas. Levando isso em consideração e o fato de as amostras ambientais possuírem muitas sequências de organismos não cultiváveis em laboratório e de filos candidatos, é justificável o percentual de sequências dadas como “não classificadas” durante as análises com o MG-RAST. Quando observadas as análises feitas com o MG-RAST e o QIIME para as

amostras de 16S rDNA com o banco de dados *Greengenes*, outro banco de dados curados, entretanto com mais sequências depositadas comparado ao SILVA, incluindo filos candidatos, (YOUSSEF et al., 2015), a análise com o MG-RAST identificou 55,7% das sequências como "não classificadas" contra um valor de menos de 1% para a análise realizada com o QIIME. A discrepância entre esses valores pode estar relacionada a uma quantidade maior de filos candidatos identificados pelos QIIME, mesmo com pouca abundância. Por outro lado, quando observados os filos mais conhecidos e com várias sequências depositadas nos bancos de dados de 16S rDNA, como os filos *Acidobacteria*, *Actinobacteria*, *Proteobacteria* e *Verrucomicrobia* comumente encontrados em amostras de solo, pôde-se perceber que não houve diferenças quanto à dominância nas análises realizadas com os dois programas, embora haja diferenças entre as abundâncias.

Nesse ponto surgiu uma nova questão: e se os programas reportassem a mesma quantidade de sequências não classificadas, será que a abundância entre os filos identificados seria diferente? Quando observadas isoladamente as análises realizadas somente com o MG-RAST para as amostras de DNA total contra 10 banco de dados de proteínas, podemos observar que essa diferença não foi muito grande considerando dados normalizados sendo que, com poucas exceções, foram identificados os mesmos filos e com praticamente a mesma abundância para todos eles.

Os resultados obtidos com o sequenciamento do gene 16S rDNA apresentados no trabalho de Faoro e colaboradores (2010) foram comparados com os resultados obtidos para a mesma amostra utilizando tecnologia de sequenciamento NGS apresentados neste trabalho. Analisando-se essa comparação (FIGURA 15), é possível evidenciar a grande inclusão de novos filos nos dados de sequenciamento NGS em relação aos dados de sequenciamento Sanger. Esse resultado pode estar diretamente relacionado ao grande número de sequências adicionais produzidas pela tecnologia NGS que permitiu atingir filos menos abundantes nas amostras. Entretanto, quando observamos apenas os filos dominantes, as poucas sequências de 16S rDNA obtidas com sequenciamento Sanger foram suficientes para obter o mesmo resultado, mesmo considerando as regiões diferentes do 16S rDNA utilizadas em cada trabalho. No trabalho anterior foram usadas as regiões V1-V2 do gene 16S rDNA, enquanto neste foram usadas as regiões V4-V5.

FIGURA 15 – COMPARAÇÃO ENTRE AS ESTRATÉGIAS DE SEQUENCIAMENTO SANGER E NGS PARA O 16S RDNA



FONTE: O autor (2017).

LEGENDA: A coluna "16S Sanger" se refere à análise de diversidade original das amostras MA02, 05 e 07 (FAORO et al., 2010). A coluna 16S NGS se referem à análise das leituras produzidas a partir da mesma amostra de DNA na plataforma MiSeq após amplificação do gene 16S rDNA segundo CAPORASO et al. (2012).

As diversas análises realizadas nos permitiram observar que embora não se agrupem por período diferentes de coleta (amostras de 16S rDNA), as amostras se agruparam pelas diferentes plataformas de sequenciamento (amostras de DNA total) pela análise de componentes principais. Também apresentaram os mesmos perfis de distribuição de grupos taxonômicos entre si. Nas amostras de 16S rDNA, as amostras de média e baixa altitude estão mais próximas entre si tanto para as amostras coletadas em 2004, quanto para as amostras coletadas em 2007. Também foi possível afirmar a presença dos principais filos comumente identificados no solo, uma vez que foram identificados pelas análises dos dois programas com os diferentes conjuntos de dados contra os vários bancos de dados, seja de RNA e proteínas.

6.4 ANÁLISE FUNCIONAL

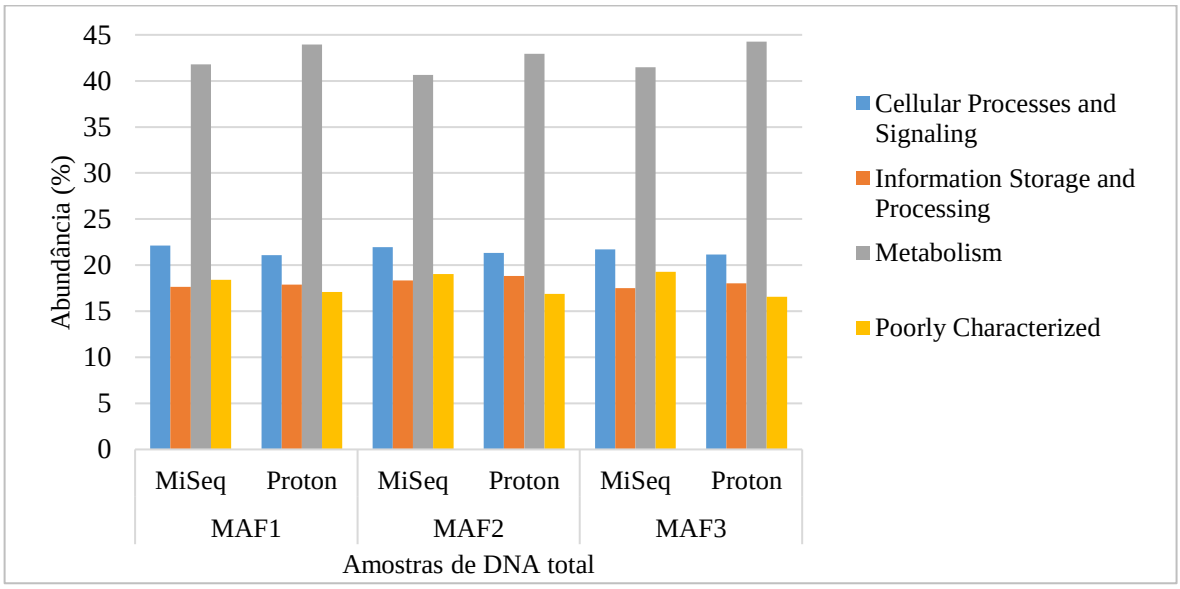
As leituras de DNA total das bibliotecas MAF1, 2 e 3, obtidas com os sequenciadores *Illumina MiSeq* e *Ion Proton*, foram analisadas quanto à função no servidor MG-RAST (TABELA 13). Todas as sequências foram analisadas quanto à função com os bancos de dados COG e SEED *subsystems*. As análises com o COG identificaram que em média 42,5% das sequências estão relacionadas à atividade de Metabolismo, seguidas por Processos celulares e de sinalização com 21,5% (FIGURA 16).

TABELA 13 – ESTATÍSTICA GERAL DE ANOTAÇÃO PARA DNA TOTAL COM MG-RAST

	MAF1		**MAF2**		**MAF3**	
	MiSeq	**Proton**	**MiSeq**	**Proton**	**MiSeq**	**Proton**
Proteínas preditas	8.299.367	2.386.512	4.926.498	16.238.402	3.013.603	4.253.853
Proteínas identificadas	2.614.613	425.882	1.716.210	2.264.541	1.084.013	748.7
Categorias funcionais identificadas	2.027.692	325.609	1.324.844	1.694.881	841.421	569.512

FONTE: O autor (2017).

FIGURA 16 – ANÁLISE FUNCIONAL COM COG



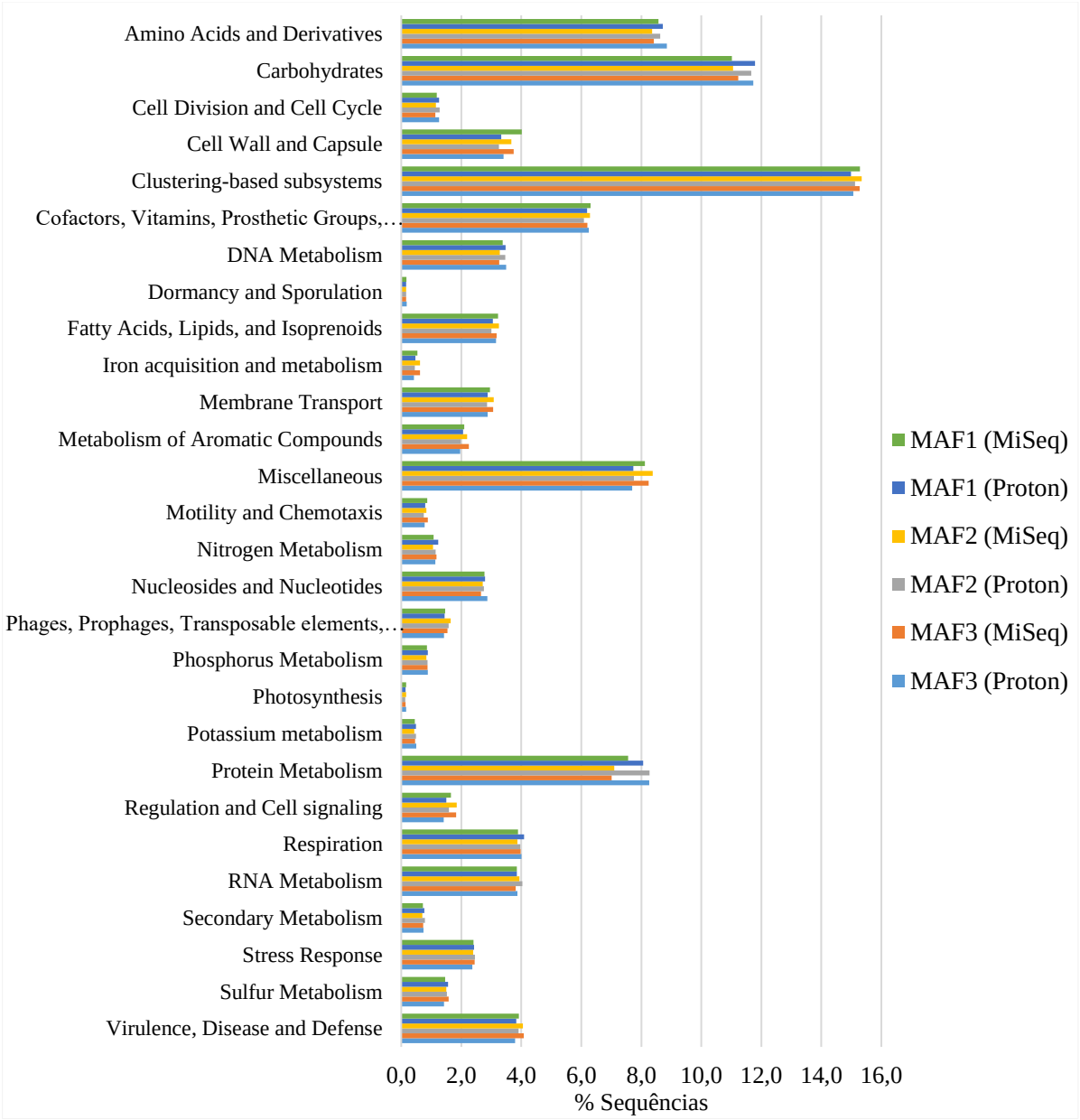
FONTE: O autor (2017).

LEGENDA: Classificação funcional das sequências nos grupos COG. Em média 42,5% das sequências estão relacionadas a processos de metabolismo e 21,5% a processos celulares e de sinalização.

Quanto à classificação dos *clusters SEED Subsystems*, as sequências foram classificadas em 27 *clusters* mais o *cluster* "*clustering-based subsystems*", que agrupa *clusters* menores que não se encaixam em nenhum dos outros *clusters*. Os *clusters* que mais tiveram sequências agrupadas, são os *clusters* de metabolismo de carboidratos (11,4%) e aminoácidos

e derivados (8,5%) (FIGURA 17).

FIGURA 17 – ANÁLISE FUNCIONAL COM *SEED SUBSYSTEMS*



FONTE: O autor (2017).

LEGENDA: Classificação das sequências nos clusters do SEED *Subsystems*. Predominância dos clusters de metabolismo de carboidratos (11,4%) e aminoácidos e derivados (8,5%), em média.

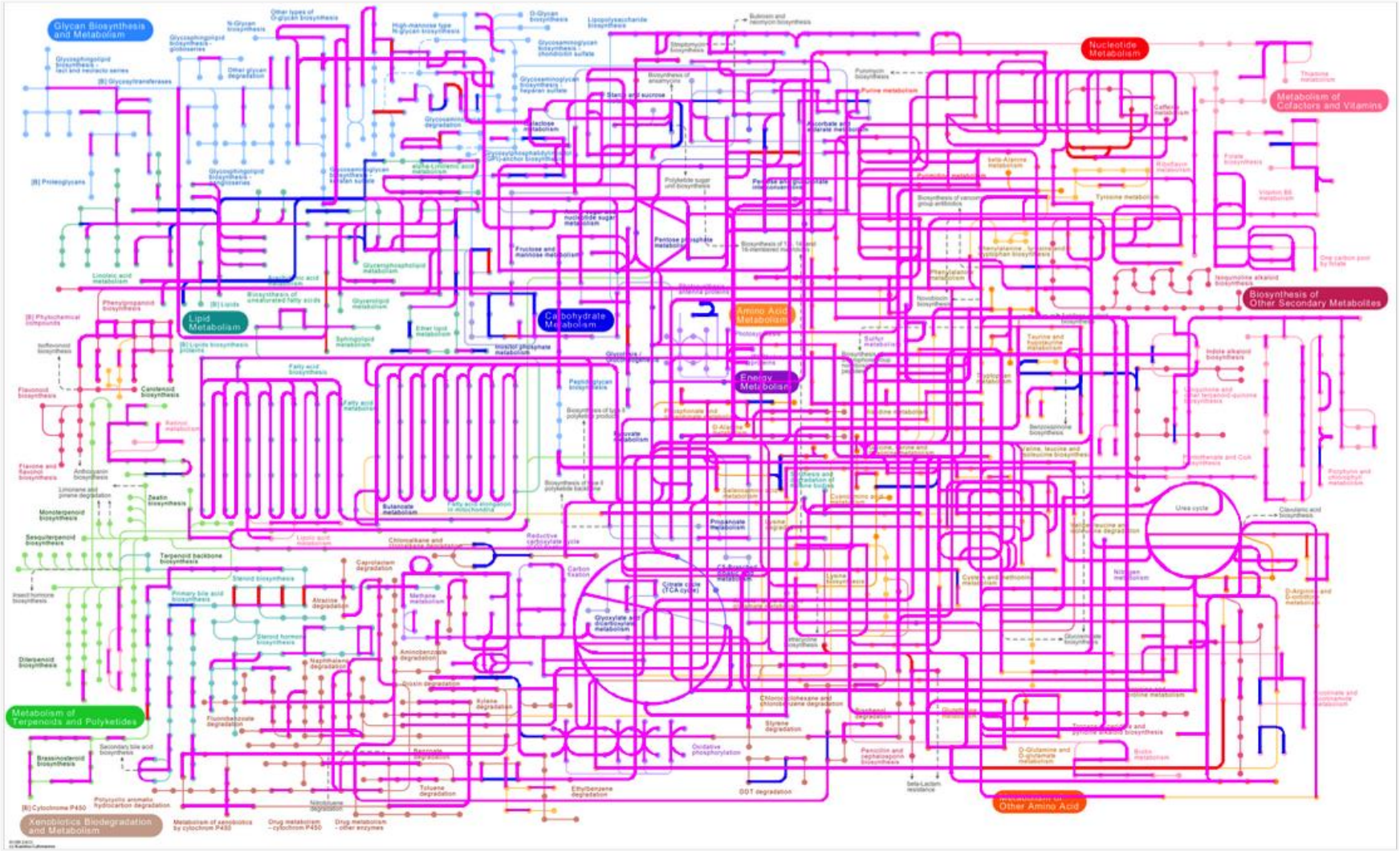
Ao todo foram identificadas 7.101 enzimas e proteínas com funções diversas, como: a fixação de nitrogênio, reparo do DNA e síntese de ATP.

Separando as amostras de DNA total em grupo A (dados *MiSeq*) “azul” e grupo B (dados *Proton*) “vermelho”, com sobreposições em “rosa”, foi construída uma via metabólica com o KEGG com 60% de identidade. Na figura 18, é visível que as plataformas de

sequenciamento apresentaram, de forma geral, resultados semelhantes e não houve tendências ou falta de cobertura. Algumas poucas vias foram identificadas apenas com dados *Illumina MiSeq*. Também percebe-se que os ciclos com maior representatividade na classificação do SEED *Subsystems*, como o metabolismo de carboidratos, apresentaram boa cobertura na via metabólica por ambas as plataformas de sequenciamento.

A análise das vias metabólicas relacionadas ao metabolismo de nitrogênio sugere que em todas as amostras há organismos capazes de realizar a fixação de nitrogênio atmosférico através do complexo da nitrogenase (FIGURAS 19, 20 e 21). A mesma análise para as vias metabólicas relacionadas ao metabolismo de carboidratos sugere que em todas as amostras há organismos capazes de degradar celulose até glucose (FIGURAS 22, 23 e 24). Comparando-se as duas vias para as três amostras é possível verificar que na amostra MAF3 há uma diminuição de rotas metabólicas em comparação com as amostras MAF1 e MAF2, evidenciada pela ausência de algumas enzimas. Essa redução pode estar relacionada com a menor diversidade de espécie dessa amostra (FAORO et al., 2012).

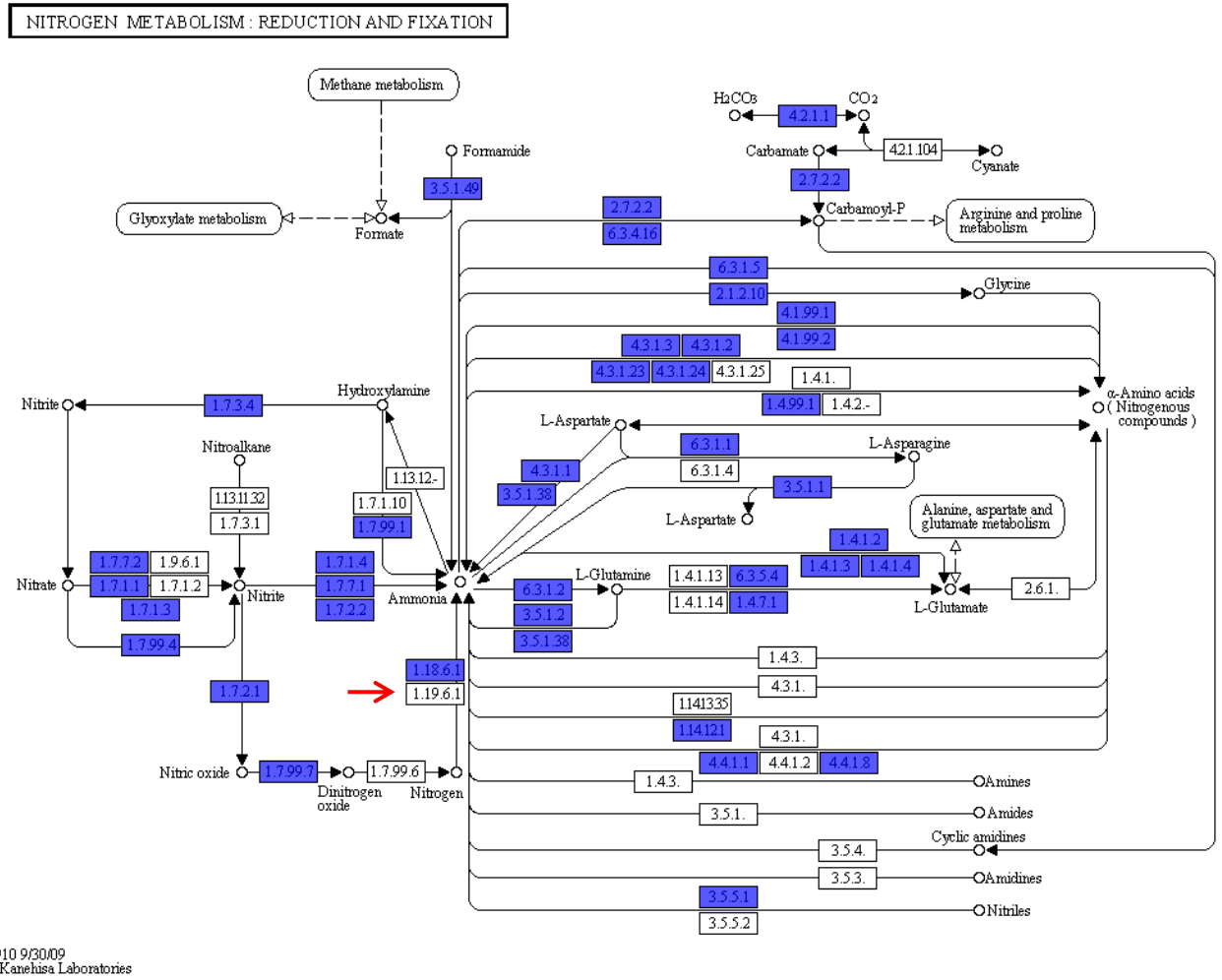
FIGURA 18 – VIA METABÓLICA KEGG PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Via metabólica construída com os dados de DNA total com o KEGG (60% de identidade). Dados MiSeq em azul, dados Proton em vermelho e sobreposições em rosa.

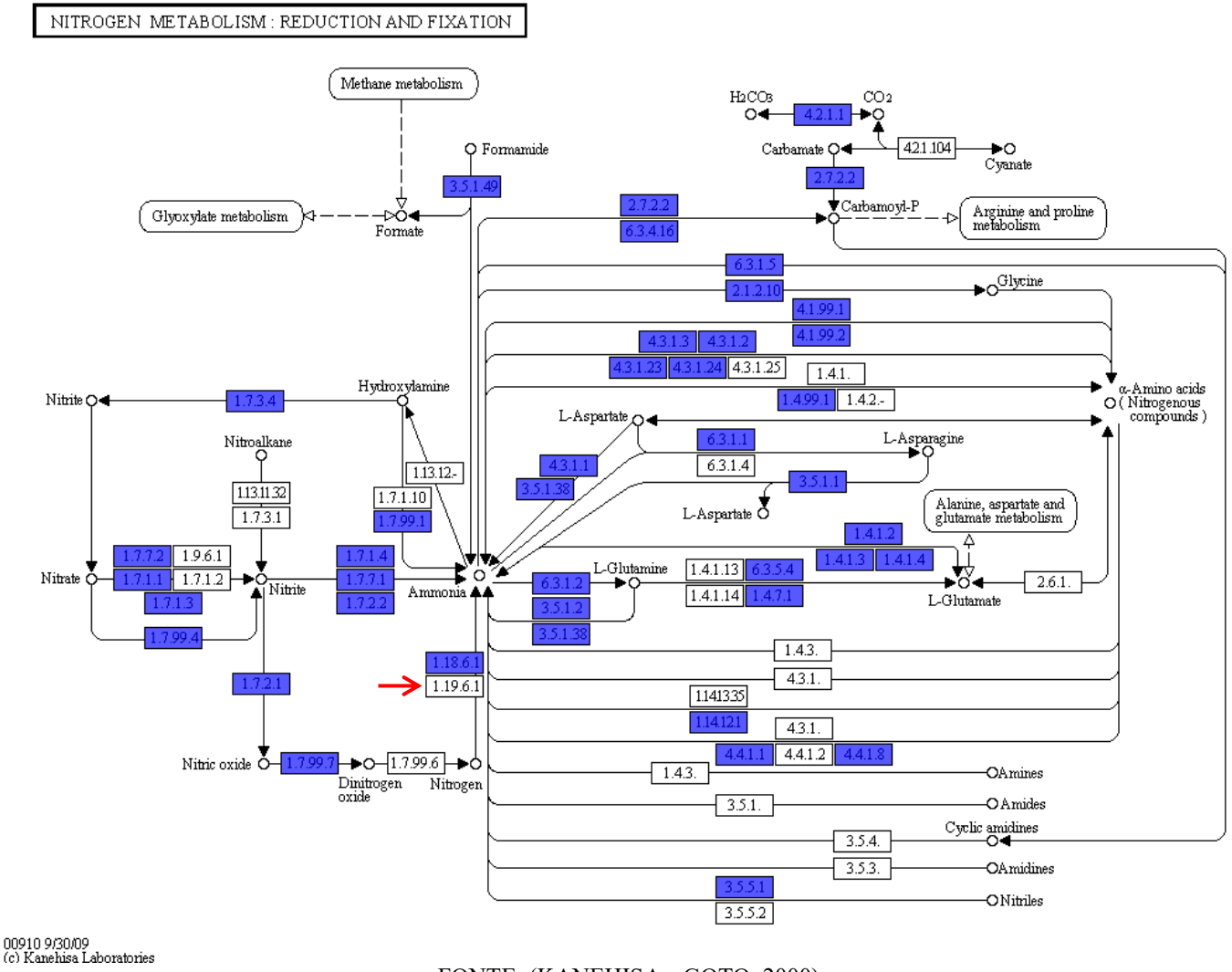
FIGURA 19 – VIA METABÓLICA DE METABOLISMO DE NITROGÊNIO – MAF1



FONTE: (KANEHISA e GOTO, 2000).

LEGENDA: Via metabólica do metabolismo de nitrogênio para a amostra de DNA total MAF1. A seta indica o complexo enzimático da nitrogenase que faz a redução do nitrogênio atmosférico para amônia.

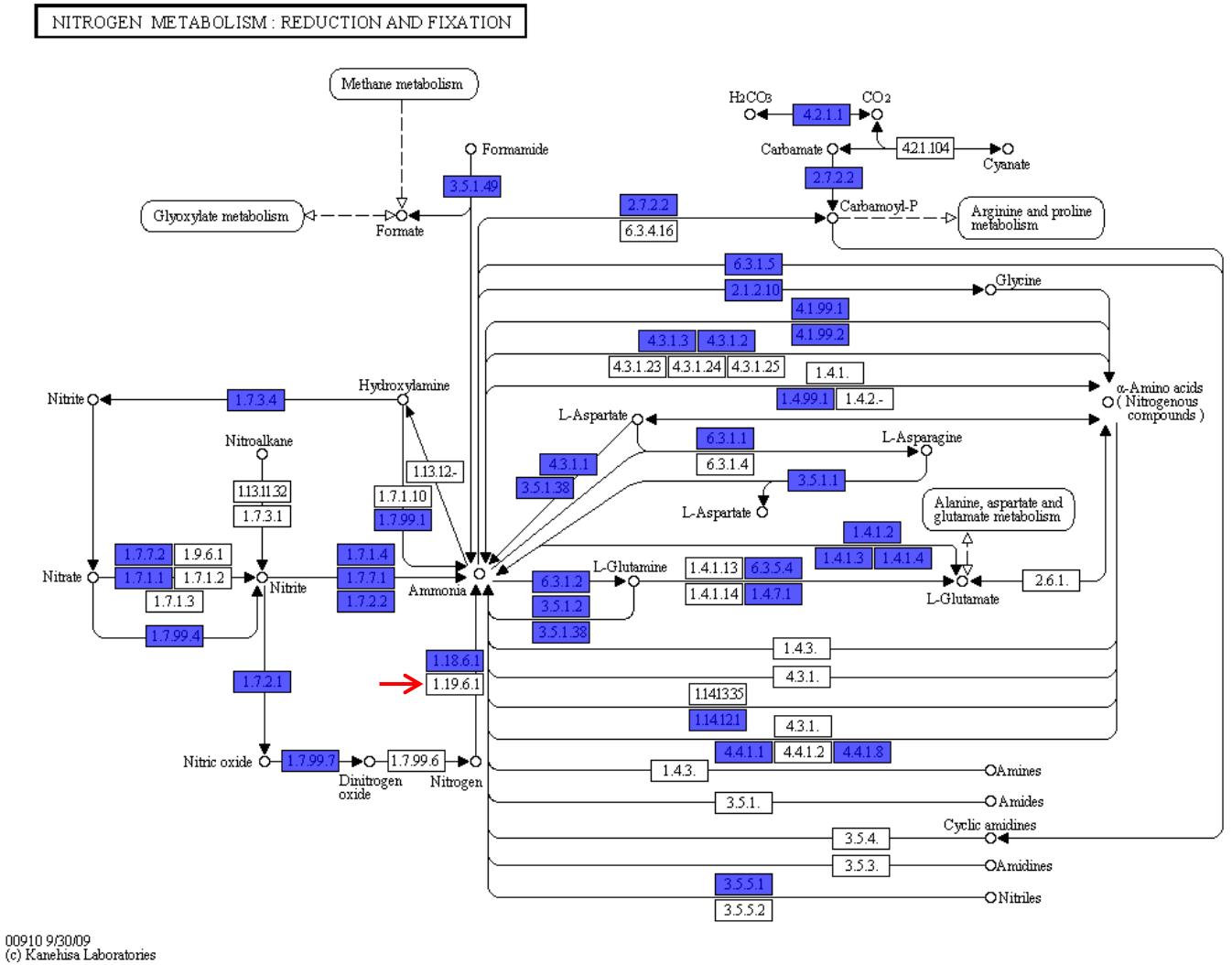
FIGURA 20 – VIA METABÓLICA DE METABOLISMO DE NITROGÊNIO – MAF2



FONTE: (KANEHISA e GOTO, 2000).

LEGENDA: Via metabólica do metabolismo de nitrogênio para a amostra de DNA total MAF2. A seta indica o complexo enzimático da nitrogenase que faz a redução do nitrogênio atmosférico para amônia.

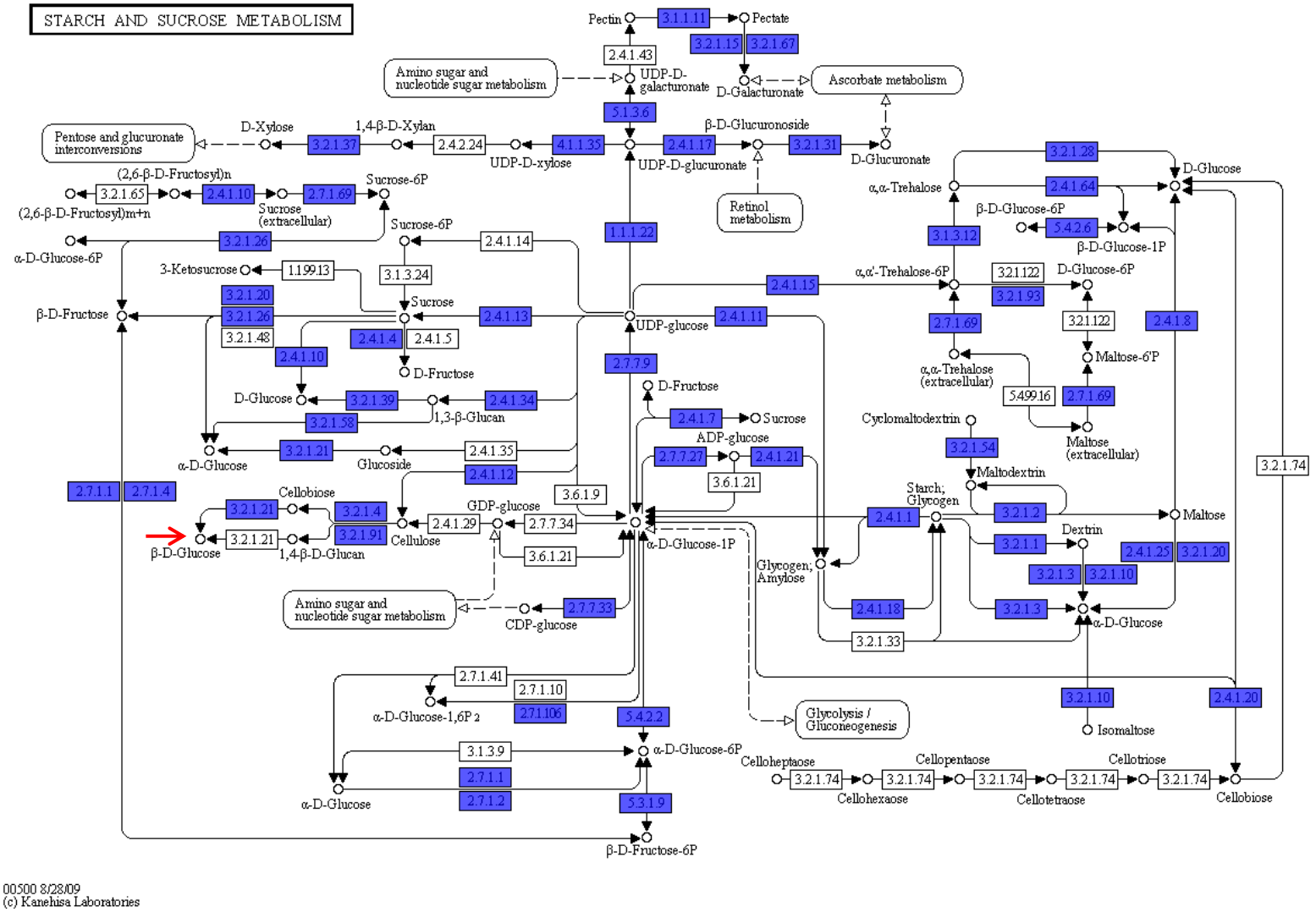
FIGURA 21 – VIA METABÓLICA DE METABOLISMO DE NITROGÊNIO – MAF3



FONTE: (KANEHISA e GOTO, 2000).

LEGENDA: Via metabólica do metabolismo de nitrogênio para a amostra de DNA total MAF3. A seta indica o complexo enzimático da nitrogenase que faz a redução do nitrogênio atmosférico para amônia.

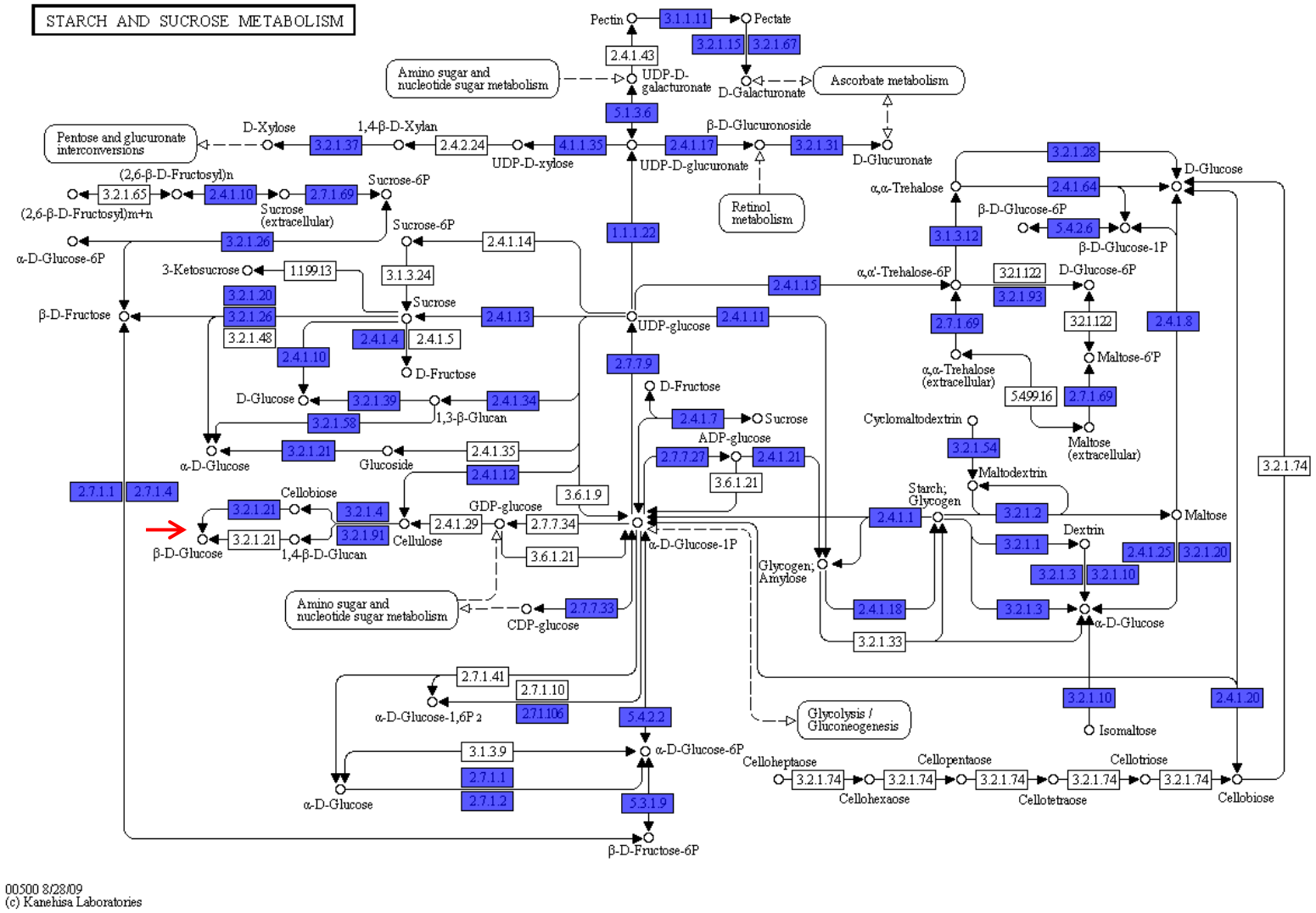
FIGURA 22 – VIA METABÓLICA DE METABOLISMO DE CARBOIDRATO – MAF1



FONTE: (KANEHISA e GOTO, 2000).

LEGENDA: Via metabólica do metabolismo de carboidratos para a amostra de DNA total MAF1. A seta indica a via metabólica que faz degradação da celulose em glucose.

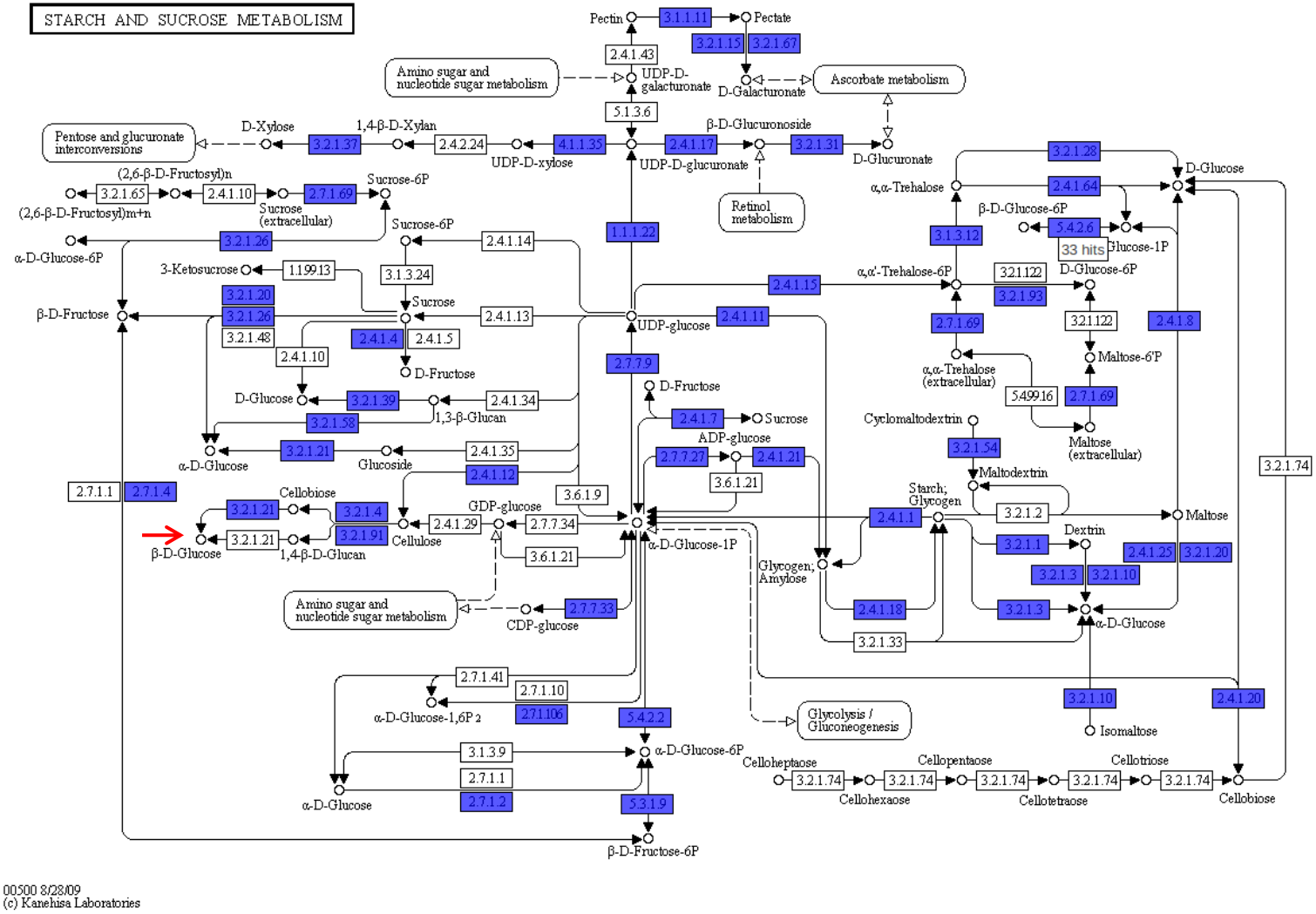
FIGURA 23 – VIA METABÓLICA DE METABOLISMO DE CARBOIDRATO – MAF2



FONTE: (KANEHISA e GOTO, 2000).

LEGENDA: Via metabólica do metabolismo de carboidratos para a amostra de DNA total MAF2. A seta indica a via metabólica que faz degradação da celulose em glucose.

FIGURA 24 – VIA METABÓLICA DE METABOLISMO DE CARBOIDRATO – MAF3



FONTE: (KANEHISA e GOTO, 2000).

LEGENDA: Via metabólica do metabolismo de carboidratos para a amostra de DNA total MAF1. A seta indica a via metabólica que faz degradação da celulose em glucose.

7 CONCLUSÕES

- O sequenciamento das amostras de solo com as tecnologias NGS *Illumina MiSeq* e *Ion Proton* geraram um grande número de sequências, em comparação ao método de Sanger usado no trabalho anterior, tornando possível expandir a análise da diversidade microbiana;
- As análises de diversidade para amostras de 16S rDNA com os programas QIIME e MG-RAST com o banco de dados *Greengenes* apresentaram diferenças na quantidade de filos identificados e sequências não classificadas;
- A análise de diversidade para amostras de 16S rRNA com o programa MG-RAST e banco de dados M5RNA foi muito semelhante à análise com o *Greengenes* para o mesmo programa;
- As análises de diversidade para amostras de DNA total para os 10 bancos de dados de proteínas com o MG-RAST apresentaram perfis taxonômicos semelhantes;
- Em todas as análises de diversidade realizadas, os filos *Acidobacteria* e *Proteobacteria* foram predominantes;
- A análise funcional para amostras de DNA total com o MG-RAST e banco de dados COG apresentaram a predominância de sequências relacionadas a funções de metabolismo;
- A análise funcional para amostras de DNA total baseada no *SEED Subsystems* identificou a predominância de sequências relacionada ao metabolismo de carboidratos e aminoácidos e derivados;
- A via metabólica do KEGG para as amostras de DNA total apresentou resultados semelhantes para os dados *Illumina MiSeq* e *Ion Proton*;
- O genoma com maior cobertura foi *Acidobacterium capsulatum ATCC 51196* (número de acesso CP001472).
- Não foi possível a montagem de genomas devido ao pequeno volume de dados metagenômicos e baixa cobertura das sequências.

REFERÊNCIAS

BAIROCH, A.; e APWEILER, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. **Nucleic Acids Research,** v. 27, n. 1, p. 49-54, PMCID: PCM9847139, 1999.

BAKER, M. De novo genome assembly: what every biologist should know. **Nature Methods**. v. 9, p. 333-337, DOI: 10.1038/nmeth.1935, 2012.

BENSON, D. A.; CAVANAUGH, M.; CLARK, K.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; SAYERS, E. W. GenBank. **Nucleic Acids Research**, v. 41 (Database issue), p. D36-42, DOI: 10.1093/nar/gks1195, 2013.

BESEMER, J.; BORODOVSKY, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. **Nucleic Acids Research**, v. 33 (Web Server issue), p. W451-4, DOI: 10.1093/nar/gki487, 2005.

BRAY, J. R.; CURTIS, J. T. An ordination of upland forest communities of southern Wisconsin. **Ecological Monographs**, v. 7, p. 325-349, DOI: 10.2307/1942268, 1957.

BRUCE, T.; MARTINEZ, I. B.; MAIA NETO, O. VICENTE, A. C. P.; KRUGER, R. H.; THOMPSON, F. L. Bacterial community diversity in the Brazilian Atlantic Forest soils. **Microbial Ecology**, v. 60, n. 4, p. 840-849, DOI: 10.1007/s00248-010-9750-2, 2010.

CAPORASO, J. G.; KUCZYNSKI, J.; STOMBAUGH, J.; BITTINGER, K.; BUSHMAN, F. D.; COSTELLO, E. K.; FIERER, N.; PEÑA, A. G.; GOODRICH, J. K.; GORDON, J. I.; HUTTLEY, G. A.; KELLEY, S. T.; KNIGHTS, D.; KOENIG, J. E.; LEY, R. E.; LOZUPONE, C. A.; MCDONALD, D.; MUEGGE, B. D.; PIRRUNG, M.; REEDER, J.; SEVINSKY, J. R.; TURNBAUGH, P. J.; WALTERS, W. A.; WIDMANN, J.; YATSUNENKO, T.; ZANEVELD, J.; KNIGHT, R. QIIME allows analysis of high-throughput community sequencing data. **Nature Methods**, v. 7, n. 5, p. 335-336, DOI: 10.1038/nmeth.f.303, 2010.

COLE, J. R.; CHAI, B.; MARSH, T. L.; FARRIS, R. J.; WANG, Q.; KULAM, S. A.; CHANDRA, S.; MCGARRELL, D. M.; SCHMIDT, T. M.; GARRITY, G. M.; TIEDJE, J. M. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. **Nucleic Acids Research**, v. 31, n. 1, p. 442-443, PMCID: PMC165486, 2003.

DANIEL, R. The metagenomics of soil. **Nature Reviews Microbiology**, v. 3, n. 6, p. 470-478, DOI: 10.1038/nrmicro1160, 2005.

DELCHER, A. L.; HARMON, D.; KASIF, S.; WHITE, O.; SALZBERG, S. L. Improved microbial gene identification with GLIMMER. **Nucleic Acids Research**, v. 27, n. 23, p. 4636-4641, DOI: 10.1093/nar/27.23.4636, 1999.

DESANTIS, T. Z.; HUGENHOLTZ, P.; LARSEN, N.; ROJAS, M.; BRODIE, E. L.; KELLER, K.; HUBER, T.; DALEVI, D.; HU, P.; ANDERSEN, G. L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Applied and Environmental Microbiology**, v. 72, n. 7, p. 5069-5072, DOI: 10.1128/AEM.03006-05,

2006.

EKBLOM, R.; WOLF, J. B. W. A field guide to whole-genome sequencing, assembly and annotation. **Evolutionary Applications**, v. 7, n. 9, p. 1026-1042. DOI: 10.1111/eva.12178, 2014.

FAORO, H.; ALVES, A. C.; SOUZA, E. M.; RIGO, L. U.; CRUZ, L. M.; AL-JANABI, S. M.; MONTEIRO, R. A.; BAURA, V. A.; PEDROSA, F. O. Influence of soil characteristics on the diversity of bacteria in the Southern Brazilian Atlantic Forest. **Applied and Environmental Microbiology**, v. 76, n. 14, p. 4744-4749. DOI: 10.1128/AEM.03025-09, 2010.

FAORO, H.; GLOGAUER, A.; COUTO, G. H.; SOUZA, E. M.; RIGO, L. U.; CRUZ, L. M.; MONTEIRO, R. A.; PEDROSA, F. O. Characterization of a new Acidobacteria-derived moderately thermostable lipase from a Brazilian Atlantic Forest soil metagenome. **FEMS Microbiology Ecology**, v. 81, n. 2, p. 386–394. DOI: 10.1111/j.1574-6941.2012.01361.x, 2012.

GILLESPIE, J. J.; WATTAM, A. R.; CAMMER, S. A.; GABBARD, J. L.; SHUKLA, M. P.; DALAY, O.; DRISCOLL, T; HIX, D; MANE, S. P.; MAO, C.; NORDBERG, E. K.; SCOTT, M.; SCHULMAN, J. R.; SNYDER, E. E.; SULLIVAN, D. E.; WANG, C.; WARREN, A.; WILLIAMS, K. P.; XUE, T.; YOO, H. S.; ZHANG, C.; ZHANG, Y.; WILL, R.; KENYON, R. W.; SOBRAL, B. W. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. **Infection and Immunity**, v. 79, n. 11, p. 4286-4298. DOI: 10.1128/IAI.00207-11, 2011.

GLASS, E. M.; WILKENING, J.; WILKE, A.; ANTONOPOULOS, D.; MEYER, F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. **Cold Spring Harbor Protocols**, v. 5, n. 1. DOI: 10.1101/pdb.prot5368, 2010.

HANDELSMAN, J.; RONDON, M. R.; BRADY, S. F.; CLARDY, J.; GOODMAN, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. **Chemistry & Biology**, v. 5, n. 10, p. R245–R249. DOI: 10.1016/S1074-5521(98)90108-9, 1998.

HOWE, A.; CHAIN, P. S. G. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). **Frontiers in Microbiology**, v. 6, p. 678. DOI: 10.3389/fmicb.2015.00678, 2015.

HUG, L. A.; BAKER, B. J.; ANANTHARAMAN, K.; BROWN, C. T.; PROBST, A. J.; CASTELLE, C. J.; BUTTERFIELD, C. N.; HERNSDORF, A. W.; AMANO, Y.; ISE, K.; SUZUKI, Y.; DUDEK, N.; RELMAN, D. A.; FINSTAD, K. M.; AMUNDSON, R.; THOMAS, B. C.; BANFIELD, J. F. A new view of the tree of life. **Nature Microbiology**, v. 1, n. 5, p. 16048. DOI: 10.1038/nmicrobiol.2016.48, 2016.

HUSON, D. H.; MITRA, S.; RUSCHEWEYH, H.-J.; WEBER, N.; SCHUSTER, S. C. Integrative analysis of environmental sequences using MEGAN4. **Genome Research**, v. 21, n. 9, p. 1552-1560. DOI: 10.1101/gr.120618.111, 2011.

Ion Proton System - Thermo Fisher Scientific. Disponível em: https://www.thermofisher.com/order/catalog/product/4476610. Acesso: 23 Jan. 2017.

JENSEN, L. J.; JULIEN, P.; KUHN, M.; VON MERING, C.; MULLER, J.; DOERKS, T.; BORK, P. eggNOG: automated construction and annotation of orthologous groups of genes. **Nucleic Acids Research**, v. 36 (Database issue), p. D250-254. DOI: 10.1093/nar/gkm796, 2007.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v. 28, n. 1, p. 27-30. PMCID: PMC10592173, 2000.

KOONIN, E. V.; GALPERIN, M. Y. Principles and Methods of Sequence Analysis. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic, c. 4, 2003.

LI, D.; LUO, R.; LIU, C.-M.; LEUNG, C.-M.; TING, H.-F.; SADAKANE, K.; YAMASHITA, H.; LAM, T.-W. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. **Methods Elsevier**, v. 102, p. 3-11. DOI: 10.1016/j.ymeth.2016.02.020, 2016.

LINDGREEN, S.; ADAIR, K. L.; GARDNER, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. **Scientific Reports**, v. 6, p. 19233. DOI: 10.1038/srep19233, 2016.

MARDIS, E. R. Next-Generation DNA Sequencing Methods. **Annual Review of Genomics and Human Genetics**, v. 9, n. 1, p. 387–402. DOI: 10.1146/annurev.genom.9.081307.164359, 2008.

MARKOWITZ, V. M.; CHEN, I.-M. A.; PALANIAPPAN, K.; CHU, K.; SZETO, E.; GRECHKIN, Y.; RATNER, A.; JACOB, B.; HUANG, J.; WILLIAMS, P.; HUNTEMANN, M.; ANDERSON, I.; MAVROMATIS, K.; IVANOVA, N. N.; KYRPIDES, N. C. IMG: the Integrated Microbial Genomes database and comparative analysis system. **Nucleic Acids Research**, v. 40 (Database issue), p. D115-122. DOI: 10.1093/nar/gkr1044, 2012.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46. DOI: 10.1038/nrg2626, 2010.

MEYER, F.; PAARMANN, D.; D'SOUZA, M.; OLSON, R.; GLASS, E. M.; KUBAL, M.; PACZIAN, T.; RODRIGUEZ, A.; STEVENS, R.; WILKE, A.; WILKENING, J.; EDWARDS, R. A. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. **BMC Bioinformatics**, v. 9, n. 1, p. 386. DOI: 10.1186/1471-2105-9-386, 2008.

MUNROE, D. J.; HARRIS, T. J. R. Third-generation sequencing fireworks at Marco Island. **Nature Biotechnology**, v. 28, n. 5, p. 426-428. DOI: 10.1038/nbt0510-426, 2010.

MYERS, N.; MITTERMEIER, R. A.; MITTERMEIER, C. G.; DA FONSECA, G. A. B.; KENT, J. Biodiversity hotspots for conservation priorities. **Nature**, v. 403, n. 6772, p. 853–858. DOI: 10.1038/35002501, 2000.

OVERBEEK, R.; BEGLEY, T.; BUTLER, R. M.; CHOUDHURI, J. V, CHUANG, H.-Y.; COHOON, M.; 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.; RÜCKERT, C.; STEINER, J.; STEVENS, R.; THIELE, I.; VASSIEVA, O.; YE, Y.; ZAGNITKO, O.; VONSTEIN, V. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. **Nucleic Acids Research**, v. 33, n. 17, p. 5691–5702. DOI: 10.1093/nar/gki866, 2005.

ØVREÅS, L. Population and community level approaches for analysing microbial diversity in natural environments. **Ecology Letters**, v. 3, n. 3, p. 236–251. DOI: 10.1046/j.1461-0248.2000.00148.x, 2000.

PACE, N. R.; STAHL, D. A.; LANE, D. J.; OLSEN, G. J. The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. **Springer US**, p. 1-55. DOI: 10.1007/978-1-4757-0611-6_1, 1986.

PENG, Y.; LEUNG, H. C. M.; YIU, S. M.; CHIN, F. Y. L. Meta-IDBA: a de novo assembler for metagenomic data. **Bioinformatics (Oxford, England)**, v. 27, n. 13, p. i94-101. DOI: 10.1093/bioinformatics/btr216, 2011.

PRUESSE, E.; QUAST, C.; KNITTEL, K.; FUCHS, B. M.; LUDWIG, W.; PEPLIES, J.; GLOCKNER, F. O. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. **Nucleic Acids Research**, v. 35, n. 21, p. 7188-7196. DOI: 10.1093/nar/gkm864, 2007.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Research**, v. 33 (Database issue), p. D501-504. DOI: 10.1093/nar/gki025, 2004.

QIIME Scripts. Disponível em: http://qiime.org/scripts/index.html. Acessado: 13 Jan. 2017.

RAPPÉ, M. S.; GIOVANNONI, S. J. The uncultured microbial majority. **Annual Review of Microbiology**, v. 57, n. 1, p. 369-394. DOI: 10.1146/annurev.micro.57.030502.090759, 2003.

RIBEIRO, M. C.; MARTENSEN, A. C.; METZGER, J. P.; TABARELLI, M.; SCARANO, F.; FORTIN, M.-J. The Brazilian Atlantic Forest: A Shrinking Biodiversity Hotspot. **In Biodiversity Hotspots**, p. 405–434. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-20992-5_21, 2011.

RONDON, M. R.; AUGUST, P. R.; BETTERMANN, A. D.; BRADY, S. F.; GROSSMAN, T. H.; LILES, M. R.; LOIACONO, K. A.; LYNCH, B. A.; MACNEIL, I. A.; MINOR, C.; TIONG, C. L.; GILMAN, M.; OSBURNE, M. S.; CLARDY, J.; HANDELSMAN, J.; GOODMAN, R. M. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. **Applied and Environmental Microbiology**, v. 66, n. 6, p. 2541-2547. PMCID: PMC10831436, 2000.

SANGER, F.; NICKLEN, S. DNA sequencing with chain-terminating. **PNAS**, v. 74, p. 5463-5467, 1977.

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. **Frontiers in Plant Science**, v. 5, p. 209. DOI: 10.3389/fpls.2014.00209, 2014.

SHOKRALLA, S.; SPALL, J. L.; GIBSON, J. F.; HAJIBABAEI, M. Next-generation sequencing technologies for environmental DNA research. **Molecular Ecology**, v. 21, n. 8, p. 1794-1805. DOI: 10.1111/j.1365-294X.2012.05538.x, 2012.

SPRINGER, M. Applied biosystems: Celebrating 25 years of advancing science. **American Laboratory**, v. 38, n. 11, p. 4-8, 2006.

SUN, Y.; CAI, Y.; HUSE, S. M.; KNIGHT, R.; FARMERIE, W. G.; WANG, X.; MAI, V. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. **Briefings in Bioinformatics**, v. 13, n. 1, p. 107-121. DOI: 10.1093/bib/bbr009, 2012.

TATUSOV, R. L.; GALPERIN, M. Y.; NATALE, D. A.; KOONIN, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. **Nucleic Acids Research**, v. 28, n. 1, p. 33-36. PMCID: PMC10592175, 2000.

THOMAS, T.; GILBERT, J.; MEYER, F. Metagenomics - a guide from sampling to data analysis. **Microbial Informatics and Experimentation**, v. 2, n. 1, p. 12. DOI: 10.1186/2042-5783-2-3, 2012.

TORSVIK, V.; GOKSØYR, J.; DAAE, F. L. High diversity in DNA of soil bacteria. **Applied and Environmental Microbiology**, v. 56, n. 3, p. 782-787. PMCID: PMC2317046, 1990.

TORSVIK, V. L. Isolation of bacterial DNA from soil. **Soil Biology and Biochemistry**, p. 15-21. DOI: 10.1016/0038-0717(80)90097-8, 1980.

TORSVIK, V. L.; GOKSØYR, J. Determination of bacterial DNA in soil. Soil **Biology and Biochemistry**, v. 10, n. 1, p 7-12. DOI: 10.1016/0038-0717(78)90003-2, 1978.

TORSVIK, V.; ØVREÅS, L. Microbial diversity and function in soil: from genes to ecosystems. **Current Opinion in Microbiology**, v. 5, n. 3, p. 240-5. PMCID: PMC12057676, 2002.

TRINGE, S. G.; VON MERING, C.; KOBAYASHI, A.; SALAMOV, A. A.; CHEN, K.; CHANG, H. W.; PODAR, M.; SHORT, J. M.; MATHUR, E. J.; DETTER, J. C.; BORK, P.; HUGENHOLTZ, P.; RUBIN, E. M. Comparative metagenomics of microbial communities. **Science (New York, N.Y.)**, v. 308, n. 5721, p 554-557. DOI: 10.1126/science.1107851, 2005.

WHITE, J. R.; NAVLAKHA, S.; NAGARAJAN, N.; GHODSI, M.-R.; KINGSFORD, C.; POP, M. Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. **BMC Bioinformatics**, v. 11, p. 152. DOI: 10.1186/1471-2105-11-152, 2010.

WILKE, A.; HARRISON, T.; WILKENING, J.; FIELD, D.; GLASS, E. M.; KYRPIDES, N.; MAVROMMATIS, K.; MEYER, F. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. **BMC Bioinformatics**, v. 13, p. 141. DOI: 10.1186/1471-2105-13-141, 2012.

WOESE, C. R. Bacterial evolution. **Microbiological Reviews**, v. 51, n. 2, p. 221-271. PMCID: PMC2439888, 1987.

YARZA, P.; YILMAZ, P.; PRUESSE, E.; GLÖCKNER, F. O.; LUDWIG, W.; SCHLEIFER, K.-H.; WHITMAN, W. B.; EUZÉBY, J.; AMANN, R.; ROSSELLÓ-MÓRA, R. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. **Nature Reviews Microbiology**, v. 12, n. 9, p. 635-645. DOI: 10.1038/nrmicro3330, 2014.

YOUSSEF, N. H.; COUGER, M. B.; MCCULLY, A. L.; CRIADO, A. E. G.; ELSHAHED, M. S. Assessing the global phylum level diversity within the bacterial domain: A review. **Journal of Advanced Research**, v. 6, n. 3, p. 269-282. DOI: 10.1016/j.jare.2014.10.005, 2015.

ZERBINO, D. R.; BIRNEY, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, n. 5, p. 821-829. DOI: 10.1101/gr.074492.107, 2008

APÊNDICE 1 – TABELA *BLAST AT NCBI*

Na figura 1, trecho da tabela gerada pelo *BLAST AT NCBI* com o nome do *contig* e o número de *hits* e as classificações por menor *e-value*, maior % de identidade, maior % de positivos, maior tamanho do *hit* e maior *bit score*, e o número de acesso e a descrição das sequências depositadas no banco de dados *nr*.

FIGURA 1 – TRECHO DA TABELA GERADA PELO *BLAST AT NCBI*

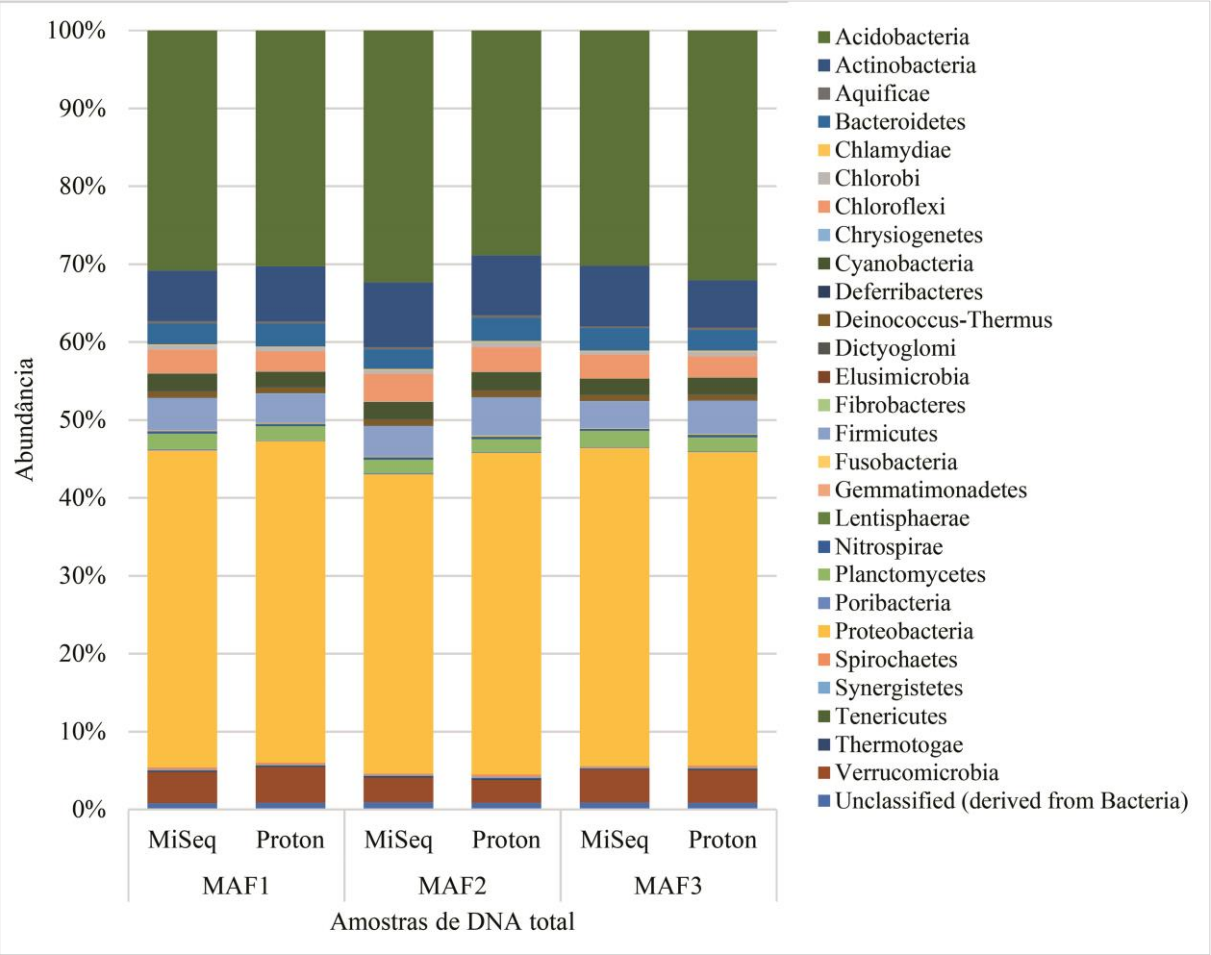
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Query	Number of hits	Lowest E-value	Accession (E-value)	Description (E-value)	Greatest identity %	Accession (identity %)	Description (identity %)	Greatest positive %	Accession (positive %)	Description (positive %)	Greatest hit length	Accession (hit length)	Description (hit length)	Greatest bit score	Accession (bit score)	Description (bit score)
2	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
3	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
4	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
5	MAF1_S1_L001_R1_001	15	6E-117	CP014841	Dyella thioc	74,35897	CP007128	Gemmatirosa	74,35897	CP007128	Gemmatirosa	779	CP000473	Solibacter u	434,28	CP014841	Dyella thioo
6	MAF1_S1_L001_R1_001	2	7,9E-26	CP003969	Sorangium	57,34767	CP003969	Sorangium ce	57,34767	CP003969	Sorangium c	561	CP012159	Chondromyc	131,552	CP003969	Sorangium c
7	MAF1_S1_L001_R1_001	102	0	CP011801	Nitrospira r	80,08256	CP011801	Nitrospira m	80,08256	CP011801	Nitrospira m	994	LN831790	Streptomyce	912,938	CP011801	Nitrospira m
8	MAF1_S1_L001_R1_001	19	0	CP011801	Nitrospira r	69,97792	CP011801	Nitrospira m	69,97792	CP011801	Nitrospira m	1806	CP011801	Nitrospira m	1090,45	CP011801	Nitrospira m
9	MAF1_S1_L001_R1_001	9	3E-173	CP000473	Solibacter u	79,82063	CP015136	Acidobacteri	79,82063	CP015136	Acidobacteri	1173	CP000473	Solibacter u	621,305	CP000473	Solibacter us
10	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
11	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
12	MAF1_S1_L001_R1_001	3	1,5E-18	CP011130	Lysobacter	68,53933	CP011130	Lysobacter ca	68,53933	CP011130	Lysobacter c	261	CP013140	Lysobacter e	106,192	CP011130	Lysobacter c
13	MAF1_S1_L001_R1_001	212	2E-168	CP013457	Burkholderi	81,35593	CP012041	Burkholderia	81,35593	CP012041	Burkholderia	965	CP013467	Burkholderi	603,871	CP013457	Burkholderia
14	MAF1_S1_L001_R1_001	103	0	AB649138	Uncultured	86,2069	CP011801	Nitrospira m	86,2069	CP011801	Nitrospira m	951	LN885086	Nitrospira s	906,599	AB649138	Uncultured b
15	MAF1_S1_L001_R1_001	4	1,1E-15	CP011929	Marinobact	74,59016	CP011929	Marinobacte	74,59016	CP011929	Marinobacte	713	HG794546	Magnetospi	96,6827	CP011929	Marinobacte
16	MAF1_S1_L001_R1_001	110	0	CP000473	Solibacter u	80,51802	CP000473	Solibacter us	80,51802	CP000473	Solibacter us	898	CP000360	Candidatus	860,635	CP000473	Solibacter us
17	MAF1_S1_L001_R1_001	1	3E-120	CP001472	Acidobacter	64,10699	CP001472	Acidobacteri	64,10699	CP001472	Acidobacteri	1153	CP001472	Acidobacter	443,79	CP001472	Acidobacteri
18	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
19	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
20	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
21	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
22	MAF1_S1_L001_R1_001	1	3,8E-25	CP000698	Geobacter u	65,38462	CP000698	Geobacter ur	65,38462	CP000698	Geobacter ur	260	CP000698	Geobacter u	128,382	CP000698	Geobacter ur
23	MAF1_S1_L001_R1_001	3	0	CP011801	Nitrospira r	67,80726	CP011801	Nitrospira m	67,80726	CP011801	Nitrospira m	1417	CP011801	Nitrospira m	748,102	CP011801	Nitrospira m
24	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
25	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
26	MAF1_S1_L001_R1_001	0		not availab	not availab		not available	not available		not available	not availabl		not available	not availab		not availa	not availabl
27	MAF1_S1_L001_R1_001	1	9,6E-87	CP000473	Solibacter u	66,88207	CP000473	Solibacter us	66,88207	CP000473	Solibacter us	619	CP000473	Solibacter u	332,842	CP000473	Solibacter us

FONTE: O autor (2017).

LEGENDA: Tabela gerada pelo *BLAST at NCBI* para os 16.426 *contigs* com mais de 1.000 pb da montagem MAF1 *Illumina MiSeq* com o *CLC Genomics Workbench*. Desses 16.426 *contigs*, 8.289 *contigs* tiveram ao menos 1 *hit* gerando 41.445 ocorrências de sequências, sendo 3.010 de sequências únicas.

APÊNDICE 2 – ANÁLISE DE DIVERSIDADE COM MG-RAST PARA BANCOS DE DADOS DE PROTEÍNAS

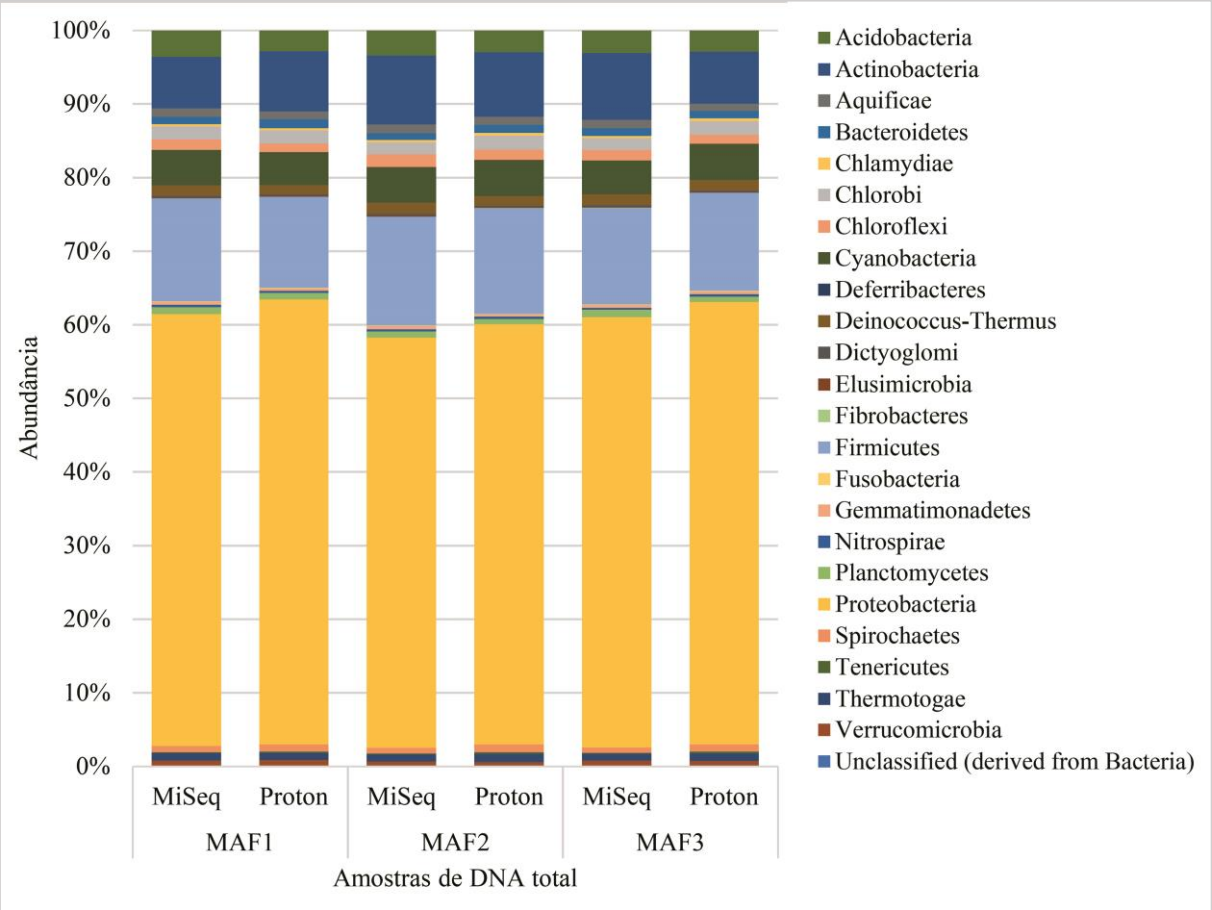
FIGURA 1 – ANÁLISE DE DIVERSIDADE COM MG-RAST E *GENBANK* PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e *Genbank* com 60% de identidade para amostras de DNA total.

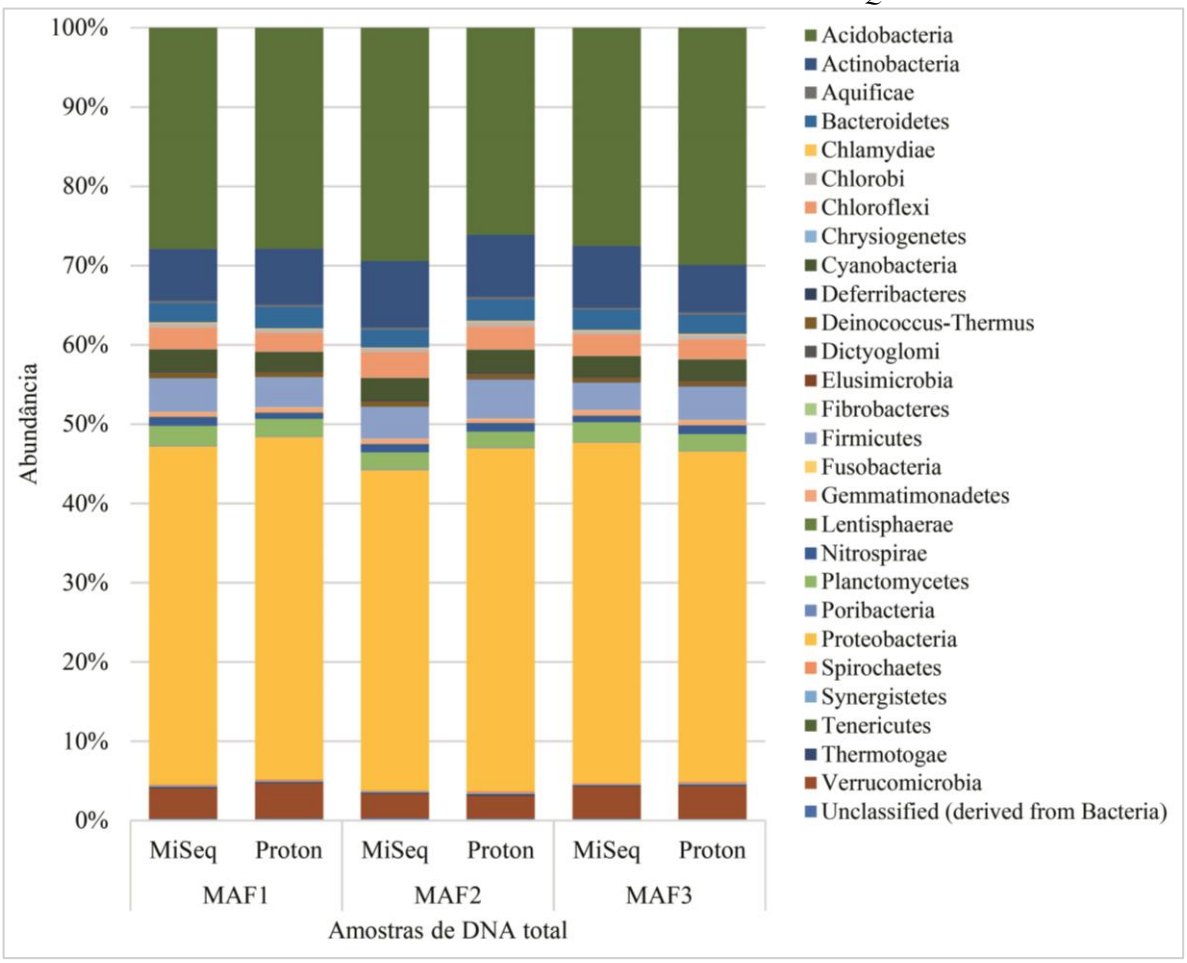
FIGURA 2 – ANÁLISE DE DIVERSIDADE COM MG-RAST E *SWISSPROT* PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e *SwissProt* com 60% de identidade para amostras de DNA total.

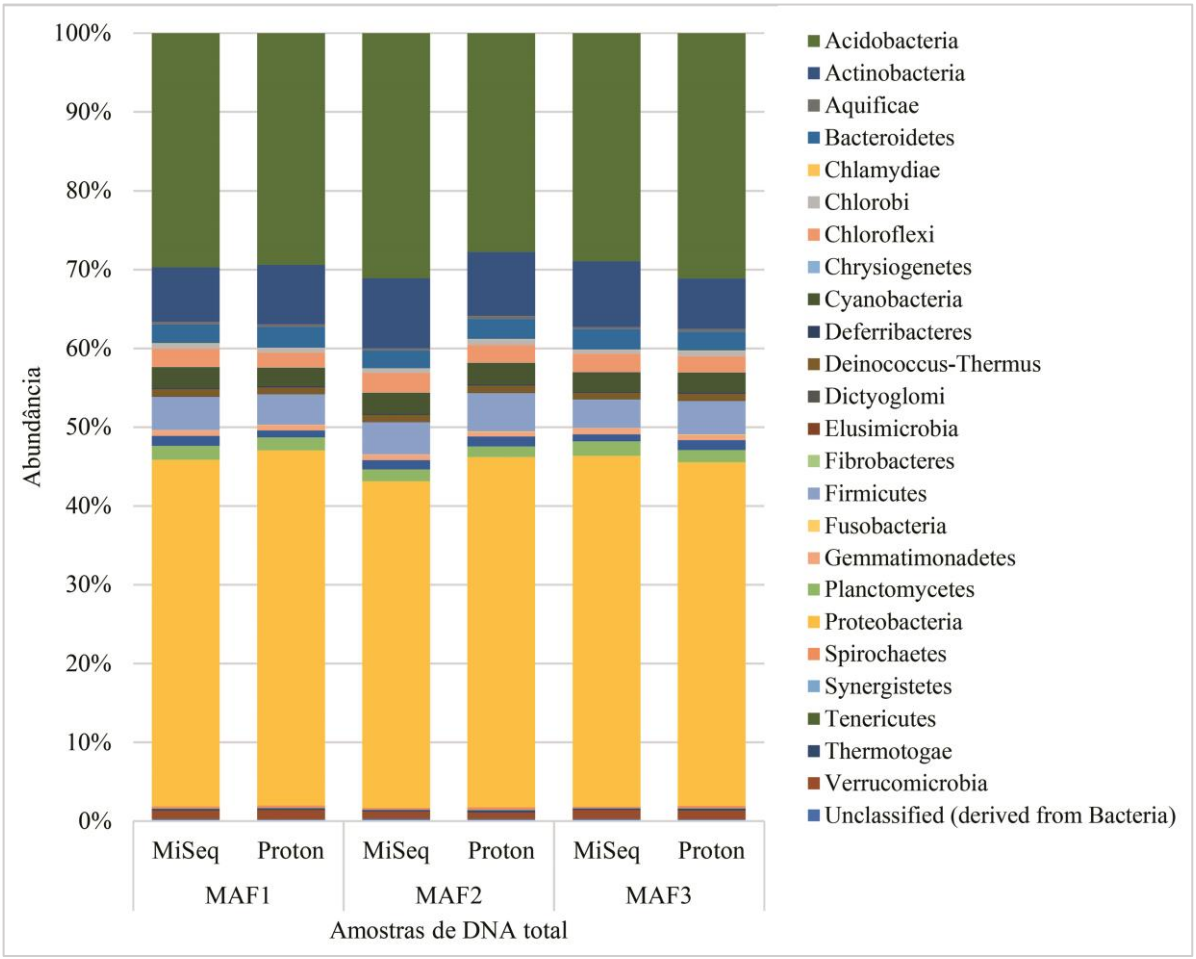
FIGURA 3 – ANÁLISE DE DIVERSIDADE COM MG-RAST E *REFSEQ* PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e *RefSeq* com 60% de identidade para amostras de DNA total.

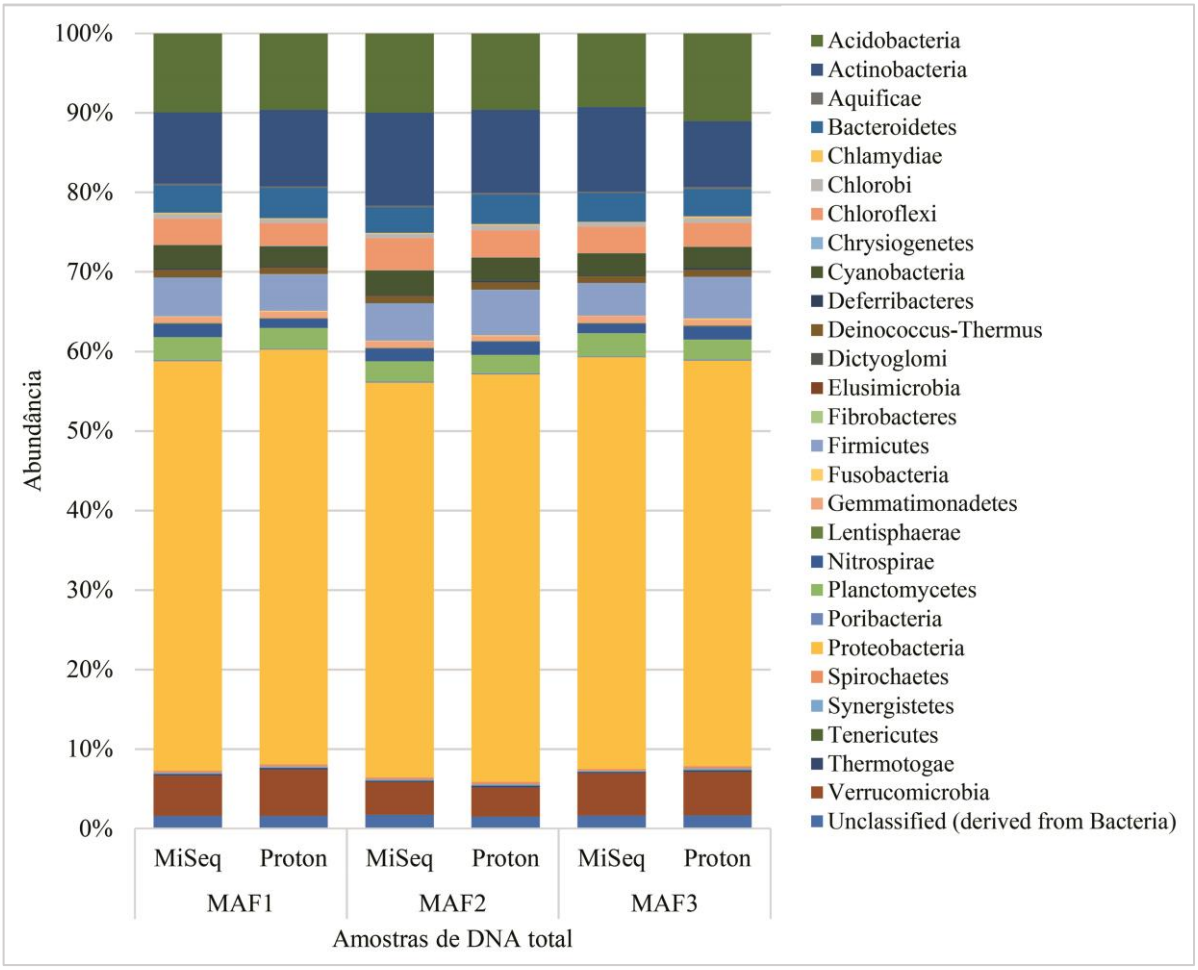
FIGURA 4 – ANÁLISE DE DIVERSIDADE COM MG-RAST E KEGG PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e KEGG com 60% de identidade para amostras de DNA total.

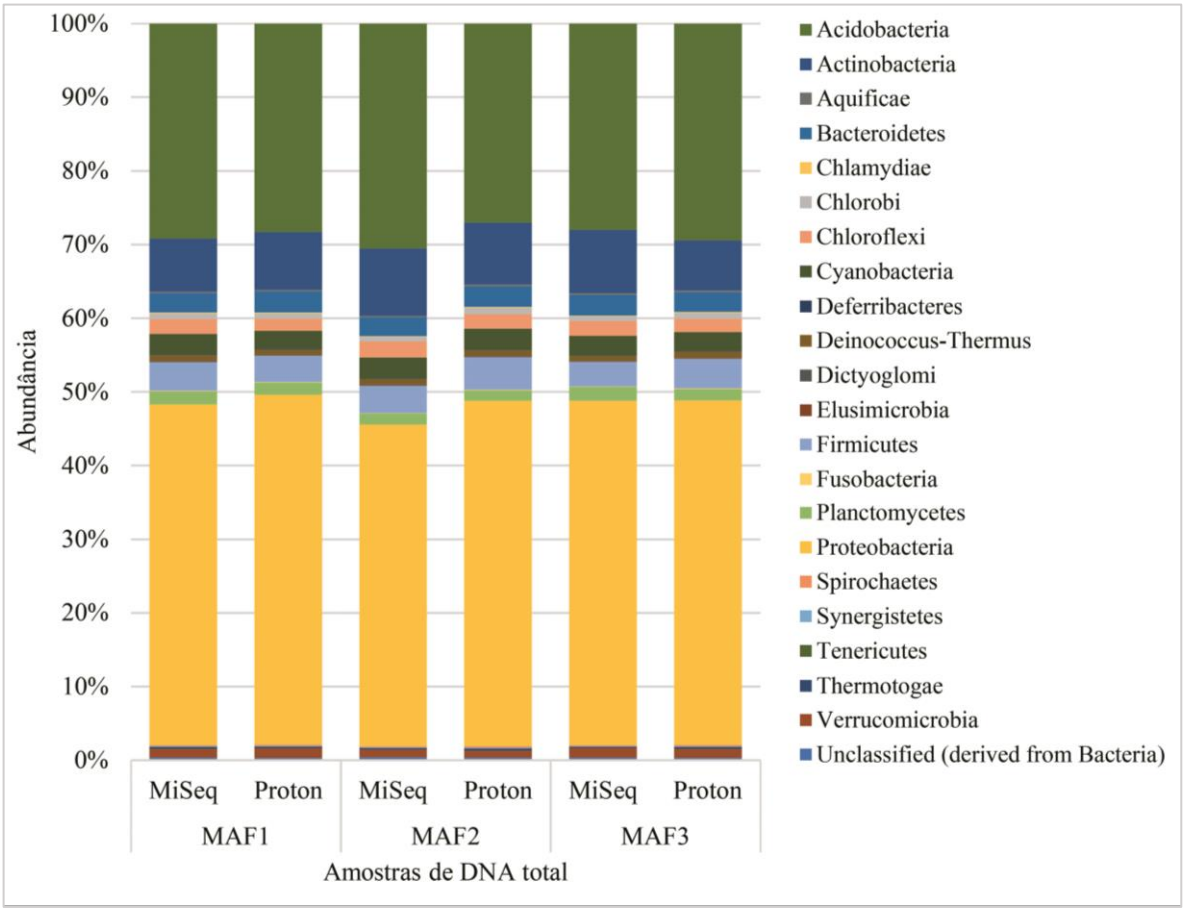
FIGURA 5 – ANÁLISE DE DIVERSIDADE COM MG-RAST E TREMBL PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e TrEMBL com 60% de identidade para amostras de DNA total.

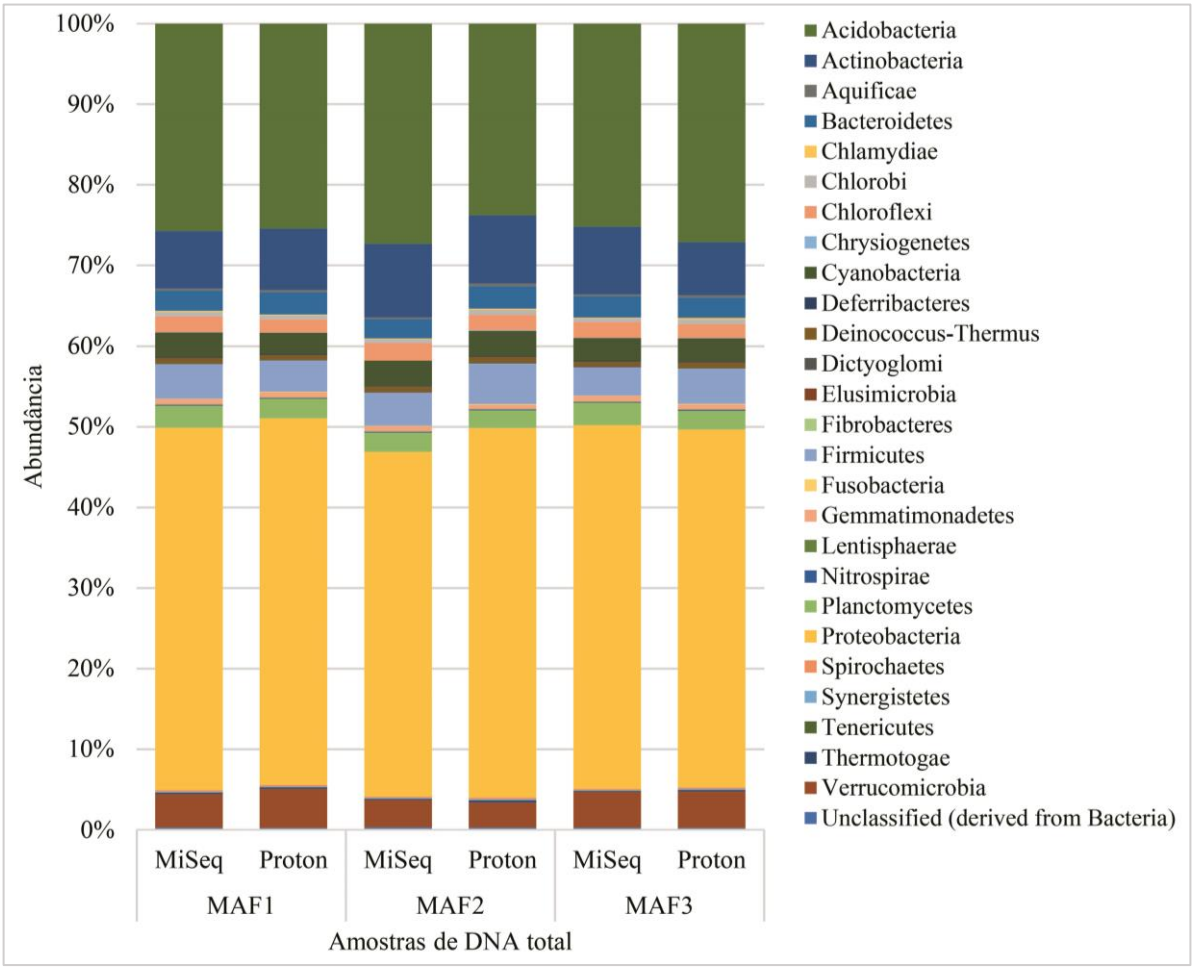
FIGURA 6 – ANÁLISE DE DIVERSIDADE COM MG-RAST E SEED PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e SEED com 60% de identidade para amostras de DNA total.

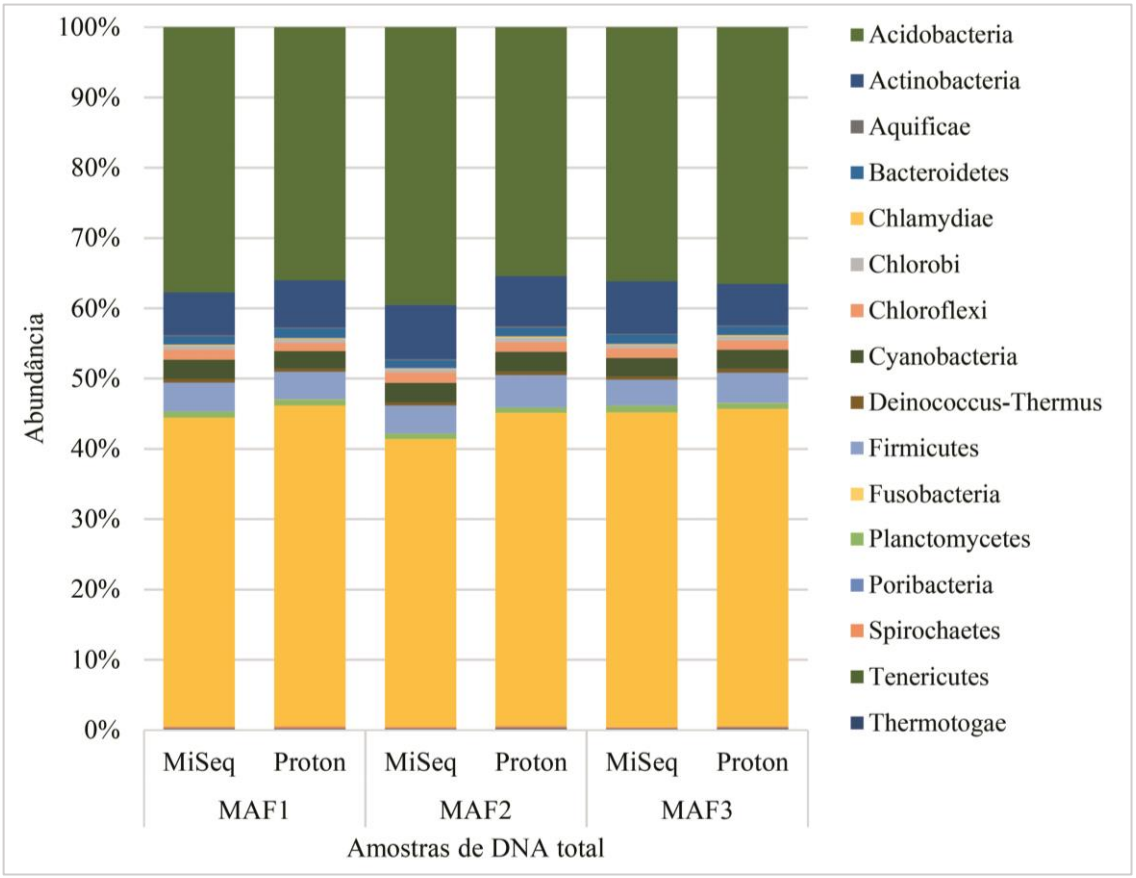
FIGURA 7 – ANÁLISE DE DIVERSIDADE COM MG-RAST E IMG PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e IMG com 60% de identidade para amostras de DNA total.

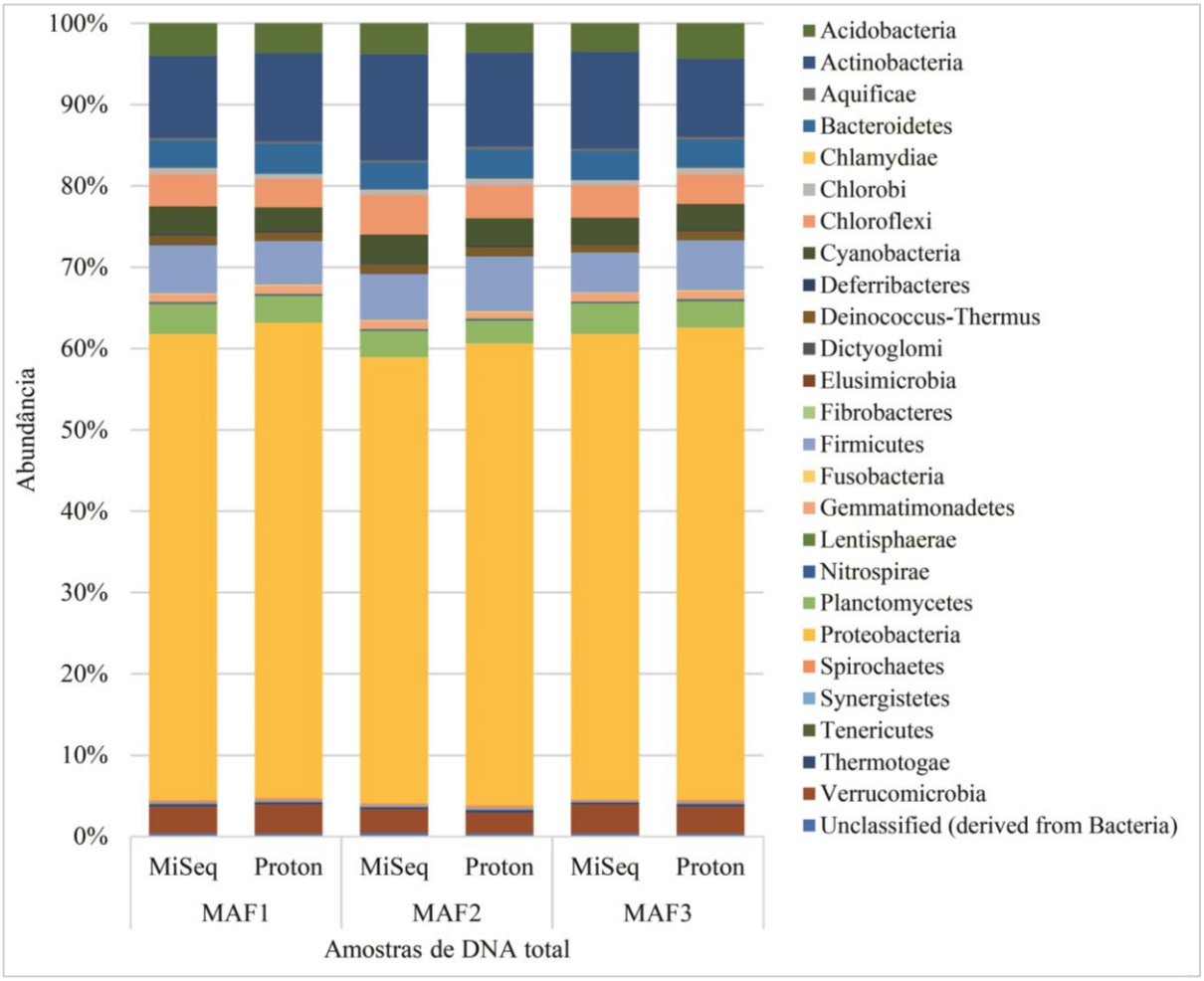
FIGURA 8 – ANÁLISE DE DIVERSIDADE COM MG-RAST E EGGNOG PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e eggNOG com 60% de identidade para amostras de DNA total.

FIGURA 9 – ANÁLISE DE DIVERSIDADE COM MG-RAST E PATRIC PARA DNA TOTAL



FONTE: O autor (2017).

LEGENDA: Análise de diversidade com o MG-RAST e PATRIC com 60% de identidade para amostras de DNA total.