UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

José Rafael Pilan

Desenvolvimento de um método de classificação taxonômica de dados de metagenomas

Botucatu, Janeiro de 2017.

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

José Rafael Pilan

Desenvolvimento de um método de classificação taxonômica de dados de metagenomas

Monografia apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, em preenchimento dos requisitos para a obtenção do título de Mestre no Programa de Pós-Graduação em Biotecnologia.

Área de Concentração: Biotecnologia
Orientador: Prof. Dr. José Luiz Rybarczyk Filho
Co-Orientadora: Prof.ª Dr.ª Agnes Alessandra Sekijima Takeda

Botucatu, Janeiro de 2017.

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Pilan, José Rafael.
Desenvolvimento de um método de classificação taxonômica de dados de metagenomas / José Rafael Pilan. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Luiz Rybarczyk Filho
Coorientador: Agnes Alessandra Sekijima Takeda
Capes: 90400003

1. Metagenômica. 2. Taxonomia numérica. 3. Metagenoma. 4. Micro-organismos. 5. Código genético.

Palavras-chave: Metagenômica; NGS; Taxonomia.

Agradecimentos

- Ao CNPq por utilizarmos os recursos computacionais referentes aos processos 458810/2013-4 e 473789/2013-2
- Ao Professor Dr. José Luiz Rybarczyk Filho pela orientação e auxilio;
- À Professora Dr.ª Agnes Alessandra Sekijima Takeda pela co-orientação e dedicação;
- Ao Professor Dr. César Martins por disponibilizar o acesso ao *cluster* para este trabalho;
- À toda minha família, principalmente meu pai José Carlos, minha mãe Maria Aparecida e meu irmão César pelo apoio e incentivo que foram fundamentais para a conclusão desse trabalho;
- Aos meus amigos Alex Augusto Biazotti, André Luiz Molan e Carlos Alberto de Oliveira Biagi Junior por toda a ajuda e motivação;
- À minha namorada Ingrid Vidotto de Oliveira por toda paciência e companheirismo;
- Aos demais colegas do departamento de Física e Biofísica e do Instituto de Biotecnologia, professores e funcionários que de alguma forma tenham contribuído para a realização deste trabalho.

Resumo

Na análise de dados metagenômicos temos duas perguntas básicas que podemos fazer: “Quem são?” e “O que estão fazendo?” os microorganismos de uma determinada amostra. Para responder a primeira pergunta utiliza-se a análise taxonômica de microorganismos. Existem diversos *software* que utilizam diferentes metodologias para atingir essa finalidade. Esses métodos são divididos em duas categorias principais: composicional e alinhamento por similaridade. O que diferencia os métodos são principalmente o tempo para ser realizada a análise, poder computacional e eficiência na identificação dos *reads*. Nesse trabalho propomos um novo método por composição que utiliza cinco assinaturas genômicas e suas combinações para identificação dos *reads*: concentração de GC (Guanina/Citocina), entropia de dipletes, entropia de tripletes, entropia de tetrapletes e abundância total de dinucleotídeos. Utilizamos um conjunto de dados referente a 3055 genomas completos de bactérias provenientes do NCBI (*National Center for Biotechnology Information*)que foram fragmentados em dois grupos: teste e controle. Os grupos foram fragmentados em tamanhos de 50-1000pb com partições de tamanho 50pb, buscando se aproximar dos tamanhos de *reads* normalmente gerados pelos equipamentos de sequenciamento de nova geração. O desempenho da metologia foi avaliado por medidas de sensibilidade, especificidade, precisão e média harmônica em comparação aos resultados do grupo teste com o grupo controle. Dentre as combinações analisadas, a concentração de GC apresentou melhor desempenho na identificação dos organismos. Para a comparação do método com os *software* já existentes, prospectamos 233 amostras no EBI (*European Bioinformatics Institute*) do projeto *“A human gut microbial gene catalog established by deep metagenomic sequencing”*, realizamos a análise das amostras com os programas Phymm, Phymmbl e Raiphy e comparamos com os resultados de nossa metodologia. Na comparação, a medida de concentração de GC em conjunto com a medida entropia de dipletes mostrou-se eficiente em comparação as demais atingindo em média 89,5% de identificação dos *reads*.

Abstract

In analyzing metagenomic data we have two basic questions that we can ask: “Who are they?” and “What are doing the microorganisms of a given sample?” . To answer the first question we use the taxonomic analysis of microorganisms. There are several software that use different methodologies to achieve this purpose. These methods are divided into two main categories: compositional and alignment by similarity. What differentiates the methods are mainly the time to perform the analysis, computational power and efficiency in the identification of reads. In this work we propose a new compositional method that uses five genomic signatures and their combinations to identify reads: GC concentration, diplet entropy, triplet entropy, tetraplet entropy and total abundance of dinucleotides. We used a data set of 3055 complete bacterial genomes from the NCBI (National Center for Biotechnology Information) that were fragmented into two groups: test and control. The groups were fragmented in sizes of 50-1000bp with partitions of size 50bp, seeking to approximate the sizes of reads normally generated by the new generation sequencing equipment. The performance of the metology was evaluated by measures of sensitivity, specificity, precision and harmonic mean in comparison to the results of the test group with the control group. Among the combinations analyzed, the GC concentration presented better performance in the identification of organisms. For the comparison of the method with existing software, we prospected 233 samples in the EBI (European Bioinformatics Institute) of the project “A human gut microbial gene established by deep metagenomic sequencing”, we performed the analysis of the samples with the programs Phymm, Phymmbl and Raiphy and compared with the results of our methodology. In the comparison, the GC concentration measure in conjunction with the entropy measurement of diplets proved to be efficient in comparison to the others reaching a mean of 89.5% of the identification of the reads.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

1.1 Comunidades microbianas

As comunidades microbianas têm sido estudadas por séculos. Partindo do preceito que tanto animais como plantas vivem associados com milhares de espécies de microrganismos diferentes (ROSENBERG et al., 2010), a teoria dos hologenomas considera que a soma da informação genética do indivíduo com a sua microbiota é um fator importante na evolução. Através de estudos bioquímicos e genéticos é possível o entendimento dos processos biológicos que essas comunidades estão envolvidas. A observação desses processos pode ajudar na compreensão de metabolismos mais complexos como os eucariotos (ZILBER-ROSENBERG; ROSENBERG, 2008; ROSENBERG; ZILBER-ROSENBERG, 2011; MADIGAN et al., 2010).

No início, os microrganismos foram classificados utilizando como base parâmetros como características morfológicas ou a seleção de perfis bioquímicos (BULLOCK; ASLANZADEH, 2013). Segundo pesquisas (SCHLOSS; HANDELSMAN, 2005; FIERER; JACKSON, 2006; MOREIRA, 2015), cerca de 99 % da microbiota existente na biosfera não é cultivável em laboratório utilizando técnicas de isolamento. Devido a falta de conhecimento das necessidades nutricionais destes organismos ou mesmo da dependência entre eles para o seu crescimento (LEWIS et al., 2010). Esses fatores corroboram com a limitação do entendimento da fisiologia, genética e comunidade desses microrganismos.

1.2 Tecnologias de sequenciamento

1.2.1 Sanger (primeira-geração)

Sanger e colaboradores (SANGER; NICKLEN; COULSON, 1977) criaram um método de sequenciamento que utiliza dideoxinucleotídeos como inibidores de terminação de cadeia de DNA, denominado “primeira-geração” (METZKER, 2010) (figura 1.1). Nesse método de sequenciamento manual, ocorre a síntese de uma fita complementar a sequência de DNA de

interesse, utilizando DNA polimerase, um iniciador (primer), além de uma mistura de deoxinucleotídeos (dNTPs) e dideóxinucleotídeos (ddNTPs). Durante a síntese, várias cópias da fita complementar são sintetizadas com a adição de dNTPs. Porém, quando os ddNTPs são incorporados, ocorre a interrupção da inclusão de novos nucleotídeos. Isso pode ocorrer em qualquer etapa da síntese, gerando fragmentos de diferentes tamanhos. Após essa etapa, as amostras são aplicadas em gel de poliacrilamida, onde ocorre a separação dos fragmentos de DNA. A visualização dessas bandas indicará a sequência de nucleotídeos da fita complementar e, por complementaridade de bases é obtida a sequência de interesse. Posteriormente o método passou a ser automático, com a utilização de computadores para leitura do gel e processamento das sequências e a utilização de fluorocromos para os nucleotídeos (HEATHER; CHAIN, 2016).

Este método permitiu novos tipos de análises nas quais o isolamento de materiais genéticos de um único organismo tem sido gradualmente substituído pelo sequenciamento de vários organismos simultaneamente. Utilizando esse método foi possível, por exemplo, sequenciar o genoma humano (METZKER, 2010).

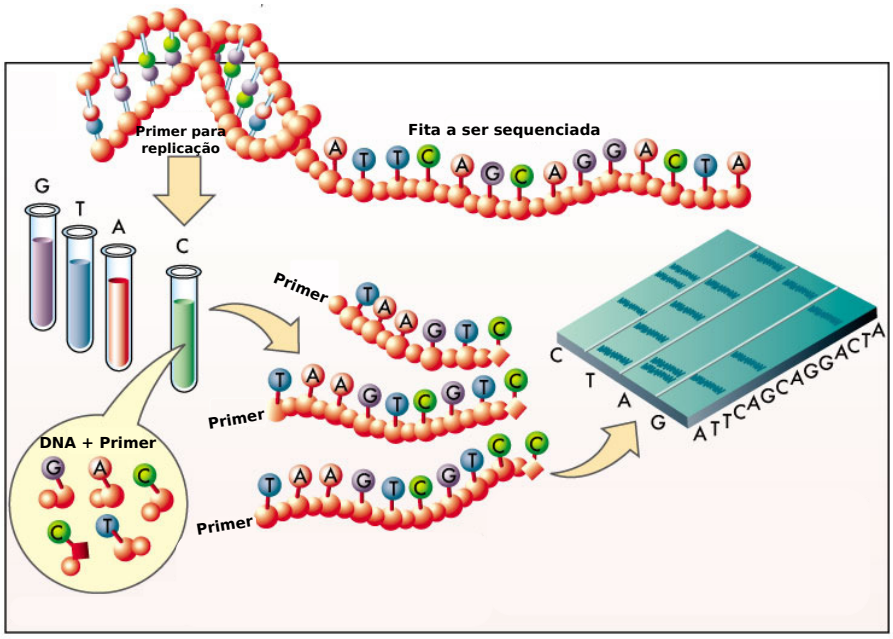

Figura 1.1: Ilustração do funcionamento da tecnologia de sequênciamento de Sanger. À 95°C a fita dupla se desnatura em duas fitas simples. Essa fita simples encontra-se em um tubo contendo *primer*, nucleotídeos e DNA polimerase. A aproximadamente 50°C ocorre o anelamento do *primer* e com 72°C a DNA polimerase entra em ação realização a extensão da fita complementar. Quando os nucleotídeos modificados são incorporados a fita de DNA ocorre o bloqueio da adição de novos nucleotídeos, gerando assim fragmentos de tamanhos diferentes. Ao colocar no gel essas fitas migram por causa da sua massa e carga. No sequenciamento automatizado os ddNTPs (dideoxinucleotídeos) são fluorescentes e utilizando um sistema de laser e detector é possível identificar esses nucleotídeos no capilar. Adaptado de (WINNICK, 2004).

1.2.2 Sequenciamento de nova geração

Os novos métodos de sequenciamento que utilizam estratégias como preparação de fitas molde, sequenciamento e imagem, alinhamento de genoma e métodos de montagem são chamadas de sequenciamentos de nova geração, em inglês *next-generation sequencing (NGS)*. Cada fabricante de equipamentos de NGS utiliza métodos diferentes para a criação dos *templates* (modelos), sequenciamento e captura de imagem. Essas diferenças refletem diretamente na qualidade e no custo dos dados gerados. (METZKER, 2010).

Entre as principais fabricantes de equipamento de sequenciamento temos: Applied Biosystems, Ilumina, Pacific Biosciences, Ion Torrent, Oxford Nanopore e a SOLiD. Para exemplificar essas tecnologias veremos o funcionamento da metodologia aplicada pelas empresas Applied Biosystems, Ilumina, Pacific Biosciences.

1.2.2.1 Applied Biosystems

O método de preparação do *template* onde são fixados *primers* individuais distribuídos espacialmente sobre um suporte sólido é utilizado pela Applied Biosystems (figura 1.2). Nessa técnica os tubos possuem uma mistura de *primers*, *templates*, dNTPs e polimerase. Ocorre uma reação de amplificação por PCR, que é quebrada por emulsão ocorrendo assim a separação do *template*. As esferas são fixadas em uma lamina de vidro.

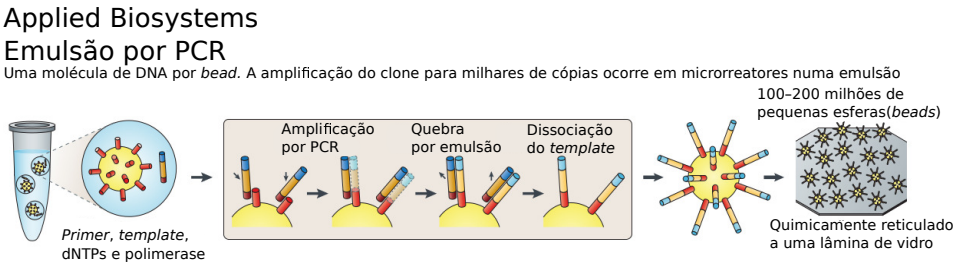

Figura 1.2: Estratégias de imobilização de *templates* utilizada pela Applied Biosystems. Para essa técnica ocorre a adição e ligação dos adaptadores e seleção dos fragmentos com adaptadores. Em uma segunda etapa acontece a ligação da fita simples de DNA às esferas e em seguida a amplificação do DNA por PCR, por último temos a eliminação das gotículas. Com a presença de reagentes (sulfurilase) é gerada luz que é captada pelo equipamento. Adaptado de (METZKER, 2010).

1.2.2.2 Ilumina

A Ilumina utiliza um método de fixação de uma única molécula de DNA por *cluster* (figura 1.3). Nessa técnica são criados agrupamentos utilizando fragmentos ou *mate-paired templates* em uma lâmina de vidro. Cada agrupamento possui *primers* que durante a fase de amplificação criarão pontes entre os segmentos *foward* e *reverse* da fita de DNA.

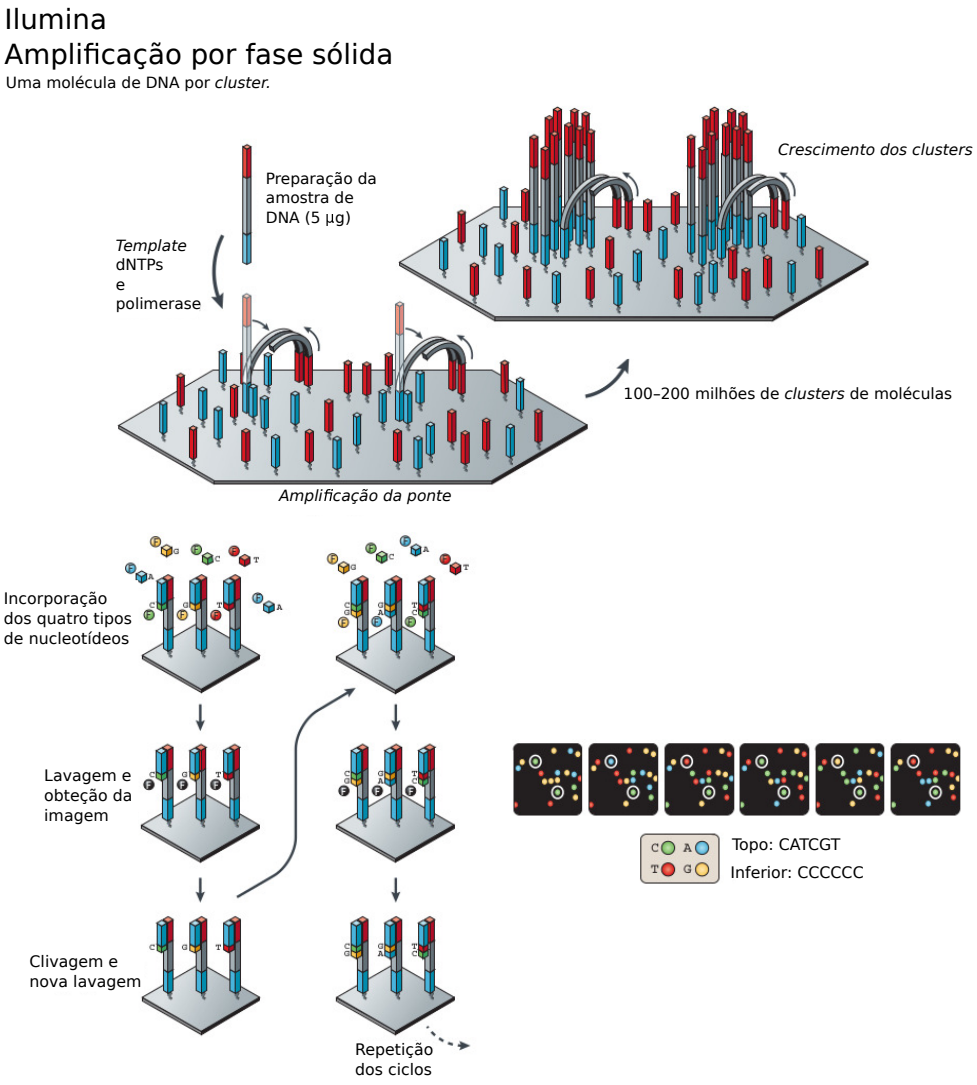

Figura 1.3: Estratégia de amplificação por fase sólida usada pelos equipamentos da Ilumina. Em cada *cluster* temos a ligação entre os segmentos da fita de DNA *foward* e *reverse*. Com isso os nucleotídeos são incorporados para a criação da fita de DNA complementar, ocorre uma lavagem para retirar os nucleotídeos não incorporados, é obtida a imagem pelo equipamento, é realizada a clivagem para retirar os fluorocromos (responsável pela emissão de luz). Esse ciclo é repetido até a obtenção de toda a sequência de nucleotídeos da fita complementar. Adaptado de (METZKER, 2010).

1.2.2.3 Pacific Biosciences

O método utilizado pela Pacific Biosciences realiza a detecção dos fragmentos incorporados em tempo real e é considerado de terceira geração. Nesse método a polimerase é fixada a superfície sólida, a detecção dos nucleotídeos incorporados ocorre em tempo real, com isso é possível conseguir tamanhos de *reads* maiores do que as outras metodologias (figura 1.4)(LIU et al., 2012b; METZKER, 2010).

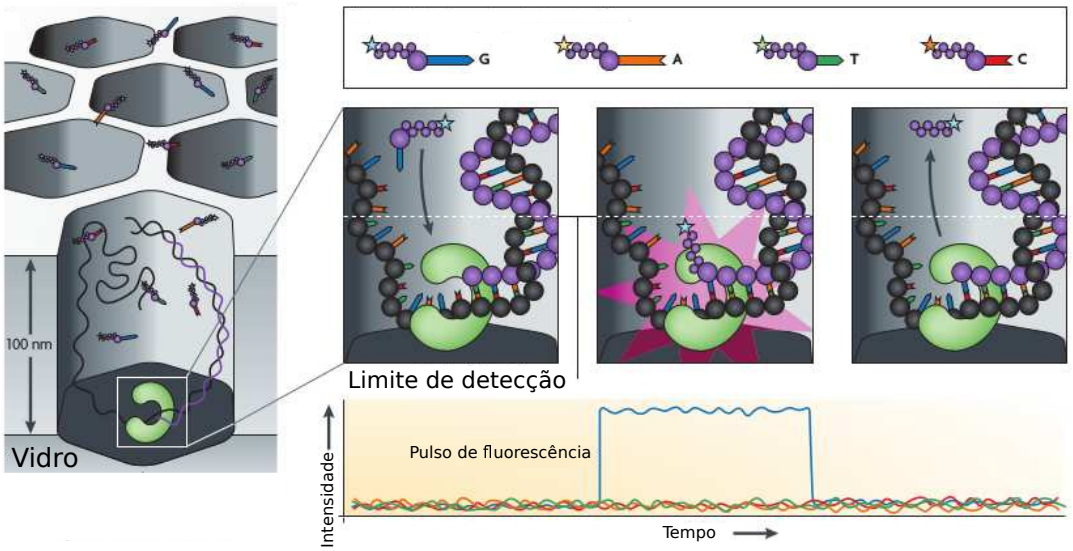

Figura 1.4: Estratégia de fixação da polimerase no suporte utilizada pela Pacific Biosciences. A polimerase é fixada em uma superfície solida. O equipamento possui um limite de detecção para captura da imagem, com isso consegue verificar a intensidade em relação ao tempo do pulso de fluorescência quando um nucleotídeo é incorporado. Após a captura da imagem ocorre a clivagem do fluorocromo. Adaptado de (METZKER, 2010).

1.2.2.4 Captura das imagens, alinhamento e montagem para NGS

Para a fase de captura das imagens é realizado um consenso entre os sinais observados quando os nucleotídeos são ligados a *templates* idênticos. (METZKER, 2010). Os arquivos resultantes da fase de captura de imagem possuem os *reads* que são os fragmentos do DNA obtidos pelos equipamentos de NGS (SHARPTON, 2014).

Alinhamento e montagem são as etapas mais demoradas e custosas computacionalmente que são realizadas após os *reads* terem sido obtidos e dependem do propósito do sequenciamento. Os *reads* podem ser alinhados a um sequência de referência conhecida, com isso conseguimos comparar a genomas já conhecidos. Podemos realizar a montagem *de novo* que não utiliza genomas de referência diminuindo assim a possibilidade de resultado tendencioso.

(PENG et al., 2011).

Com o advento das tecnologias de sequenciamento de nova geração é possível atender a necessidade cada vez maior por informações genômicas rápidas, precisas e de baixo custo. A geração de grande quantidade de dados é uma vantagem em relação aos métodos convencionais. A possibilidade de sequenciar o genoma completo de vários tipos de organismos permitiu a comparação em larga escala assim como realizar estudos evolutivos que antes não eram possíveis (METZKER, 2010; CHUN; RAINEY, 2014).

A tabela 1.1 possui informações comparativas em relação entre as principais tecnologias de sequenciamento em uso atualmente, o tamanho de *reads* gerados por essas tecnologias, o tempo de corrida, rendimento por corrida (Gpb/corrida) e custo por GB (U$$).

Tabela 1.1: Tabela comparativa das principais tecnologias de sequenciamento

Tecnologia de sequenciamento	Tamanho dos reads	Tempo de corrida	Rendimento por corrida (Gpb/corrida)	Custo por Gb (US$)
Applied Biosystems 3730 (capillary)	650	2h	$9{,}6\text{x}10^{-5}$	2.307.692,31
Illumina MiSeq	50 - 600	4h - 56 h	0,30 - 15,00	102,00 -1.866,67
Illumina NextSeq 500	75 - 300	11h - 29 h	19,50 - 120	35,33 -52,82
Illumina HiSeq 2500	50 - 500	10h - 6 dias	15 - 500	28,82 -95,33
Illumina HiSeq 4000	50 - 300	1 dia - 3,5 dias	125 - 750	20,53 -48,08
Illumina HiSeq X - Five	300	$\leq$ 3 dias	900	7,08 -10,63
Ion Torrent - PGM	400	4 hrs - 7h	0,22 - 2,20	397,27 -2.154,55
Ion Torrent - S5	200 - 400	2,5 h - 4h	2 - 16	79,69 -476,50
Oxford Nanopore	10000	Não especificado	6 - 15500	6,14 -150,00
Pacific Biosciences	10000 - 12000	$\leq$ 6h	0,66 - 3,85	181,82 -303,03
SOLiD - 5500	700 - 1410	8 dias	77 - 155,10	67,72 -79,23

Fonte: Adaptado de Glenn, TC. 2011. Field Guide to Next Generation DNA Sequencers. Molecular Ecology. Disponível em: `http://www.molecularecologist.com/next-gen-table-2-2016/`. Acesso em: 29 jul. 2016.

1.3 Descoberta de microorganismos

1.3.1 Gene 16S rRNA

Woese e colaboradores (WOESE; FOX, 1977) apresentaram uma proposta de utilizar uma parte conservada do código genético, os genes ribossomais, para classificação taxonômica de bactérias, sendo que o gene mais utilizado para esse tipo de análise é o 16S rRNA (TORTOLI, 2003) que é ilustrado na figura 1.5.

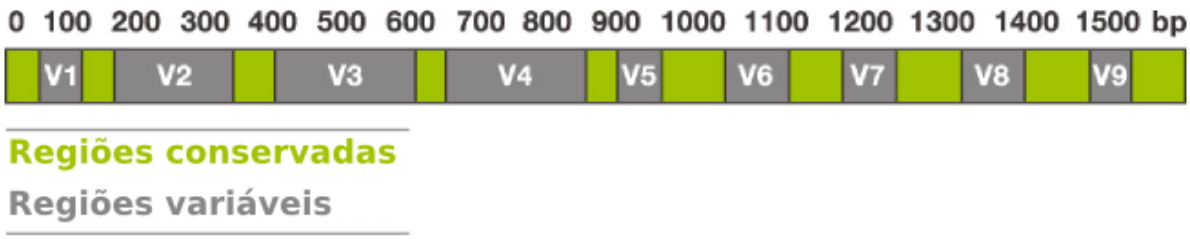

Figura 1.5: Gene 16S rRNA. Em verde é ilustrada as regiões conservadas e em cinza as regiões variáveis do gene. Adaptado de (ALIMETRICS,).

O gene 16S rRNA possui nove regiões variáveis (V1 - V9) que são ladeadas por regiões conservadas. Essas regiões variáveis possuem polimorfismo que são utilizadas para caracterizar os diferentes tipos de microorganismos (CHAKRAVORTY et al., 2007). Quando um microorganismo é comparado às espécies já existentes, e o valor de identidade obtido entre as sequências de 16S rRNA (figura 1.6) é de 99 % o microorganismo é caracterizado como sendo da mesma espécie. Para valores entre 97 % e 99 % é considerado como do mesmo gênero. Quando a identidade é menor do que 97 % o microorganismo é considerado como uma espécie potencialmente nova. (DRANCOURT; BERGER; RAOULT, 2004)

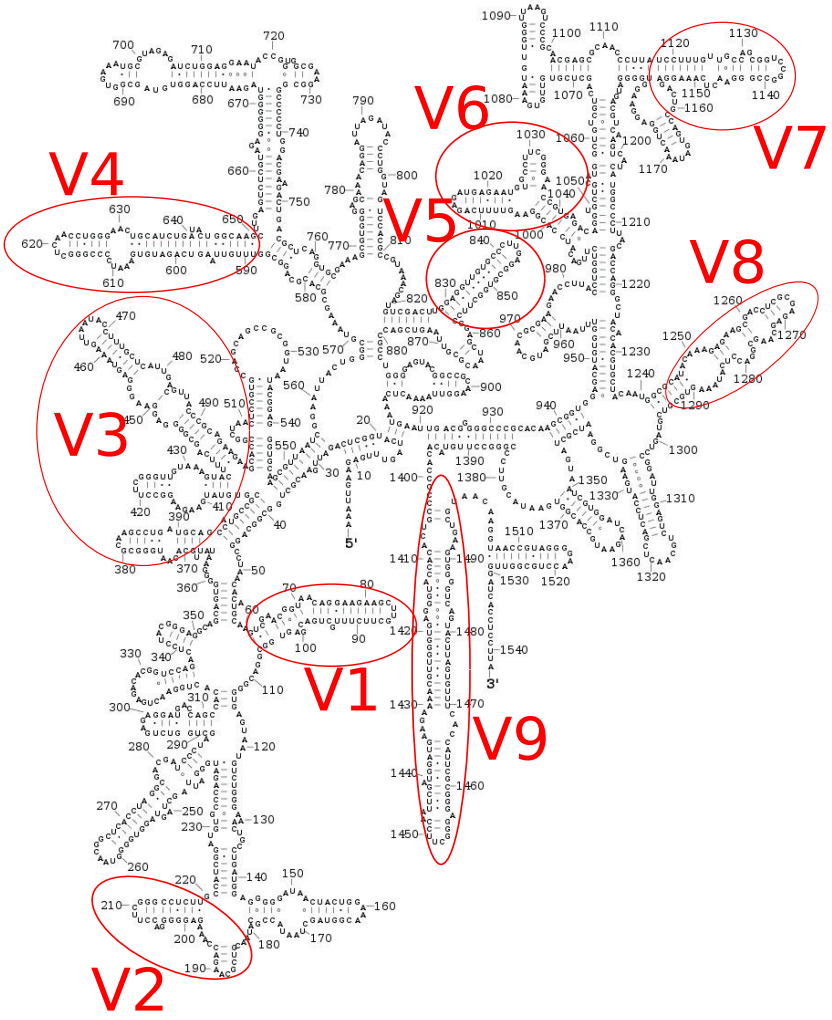

Figura 1.6: Estrutura do 16S rRNA. Na figura estão circuladas as 9 regiões variáveis do 16S rRNA. Essas regiões quando comparadas entre os microorganismos auxiliam na identificação de microorganismos de uma mesma espécie. Adaptado de (CASE et al., 2007)

Entre as razões para a utilização do gene 16S rRNA estão: esse gene pode ser encontrado em praticamente todas as bactérias, como não há alteração na função do gene 16S as análises das sequências nas regiões variáveis podem ser utilizadas como medidas em relação a evolução e o tamanho do gene é suficientemente grande para propósitos relacionados a computação (JANDA; ABBOTT, 2007). Na técnica de aplicação por reação de PCR o gene 16S rRNA é amplificado por reação de PCR utilizando *primers* específicos, clonado e sequenciado (KLINDWORTH et al., 2012). A utilização do método de Sanger em conjunto com a utilização do gene 16S rRNA abriram novas possibilidades revolucionando assim o estudo e classificação dos microrganismos (ESCOBAR-ZEPEDA; LEÓN; SANCHEZ-FLORES, 2015). Com isso, atualmente a diversidade microbiana pode ser caracterizada através de análises dependentes ou independentes de cultivo.

1.4 Metagenômica

Em 1998, Jo Handelsman (HANDELSMAN et al., 1998) utilizou pela primeira vez o termo metagenômica, uma técnica que possibilita a análise de genomas em um nível maior de complexidade sem a necessidade de cultivo prévio. Com a metagenômica, conseguimos obter informação da capacidade metabólica e funcional da comunidade microbiana através da análise do DNA obtido diretamente da amostra ambiental ou orgânica. Essa técnica possibilita responder as perguntas biológicas: “quem são?” e “o que estão fazendo?” os microrganismos de uma determinada amostra (HANDELSMAN, 2004).

As bibliotecas genômicas para utilização na metagenômica são criadas utilizando análise baseada na sequência ou na função (HANDELSMAN, 2004). A técnica consiste na obtenção do DNA de uma amostra ambiental adquirida de forma direta ou indireta. Na extração direta ocorre a lise celular direto do solo. Após a lise, o DNA é extraído e purificado para posterior análise. Na extração indireta as células bacterianas são extraídas da matriz do solo e depois lisadas para a subsequente extração do DNA, purificação e análise. O método direto de extração de DNA é o mais utilizado. Após a extração, os fragmentos de DNA são ligados a vetores, montando vetores de clonagem com capacidade de replicação, denominados DNA recombinante. Por último, essas moléculas de DNA são inseridas em uma bactéria hospedeira cultivável em laboratório, originando uma biblioteca metagenômica, em que os clones recombinantes com parte do DNA presente em um determinado ambiente ou amostra podem ser multiplicados e mantidos por muitos anos (DELMONT et al., 2011; HANDELSMAN, 2004). A análise baseada em sequência pode ser utilizada por ferramentas computacionais para as posteriores etapas de análise (Figura 1.7). A metagenômica possui grande importância no setor de biotecnologia, podendo atender a grande demanda das indústrias e da medicina em busca de enzimas mais eficientes. Isso é possível através da descoberta de novos microrganismos ou do desenvolvimento de processos de base biotecnológica pois ela permite que seja acessada uma fonte de conhecimento ainda não explorada de diversidade biológica e molecular (MOREIRA, 2015).

1.5 Análise taxônomica de dados metagenômicos

Com a diminuição do custo e o aumento da quantidade de dados gerados através da metagenômica as análises foram divididas em etapas, facilitando assim a descoberta de novos organismos, a montagem correta do genoma ou o mapeamento das sequências. Essas etapas podem ser realizadas isoladas ou combinadamente para a análise do metagenoma, e são classificadas em: controle de qualidade, montagem, detecção de genes, anotação de genes, classificação ta-

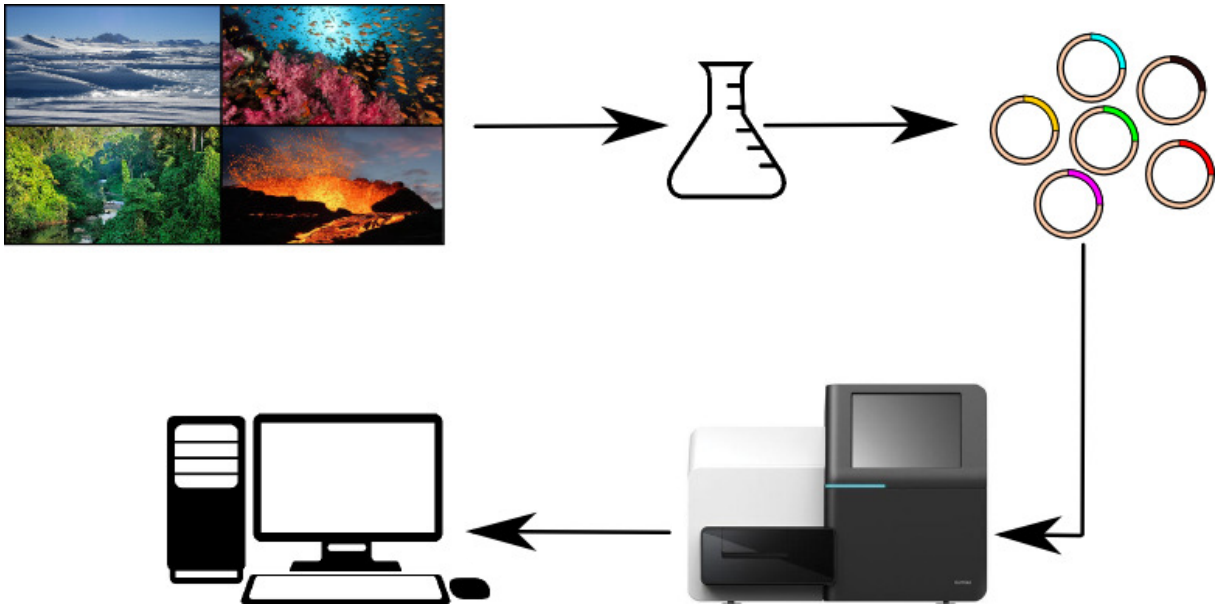

Figura 1.7: Etapas de obtenção de dados de metagenoma: a amostra de interesse é obtida para o estudo, realiza-se a extração do material genético que são inseridas em vetores para a criação de bibliotecas metagenômica. Esse vetores são sequenciados utilizando equipamentos de NGS. A informação adquirida é armazenada para posterior análise.

xonômica (*binning*), análise comparativa integrada e gerenciamento de dados (LADOUKAKIS; KOLISIS; CHATZIIOANNOU, 2015).

O processo denominado *binning* ou classificação taxonômica permite a análise dos dados metagenômicos sem a necessidade de utilização de genes marcadores filogenéticos, como o gene 16S rRNA, utilizado na identificação taxonômica de bactérias que apesar de possuir alto nível de precisão é encontrado somente em cerca de 1 % das sequências obtidas da amostra do metagenoma (MOREIRA, 2015).

Está análise de dados metagenômica pode ser baseada na similaridade entre as sequências do metagenoma com sequências já classificadas e depositadas em banco de dados de referência como em programas baseados em BLAST (*Basic local alignment search tool*), que possuem um alto nível de precisão mas que tornam o processo lento e custoso computacionalmente ou na composição dos nucleotídeos (LIU et al., 2012a) que utiliza cálculos de medidas como: uso de códon, conteúdo GC e frequência de oligonucleotídio (HIGASHI et al., 2012). Existem vários programas híbridos que mesclam os dois métodos.

1.6 Programas de análise taxonômica

Segundo a base de dados OMICtools (HENRY et al., 2014) até agosto de 2016 existiam disponíveis para *download* 66 programas de classificação taxonômica sendo desses: 48 baseados em técnicas de alinhamento, 11 baseados em composição e 7 programas que mesclam as duas técnicas. Como podemos ver na figura 1.8 esses programas foram desenvolvidos entre América do Norte, Europa e Ásia.

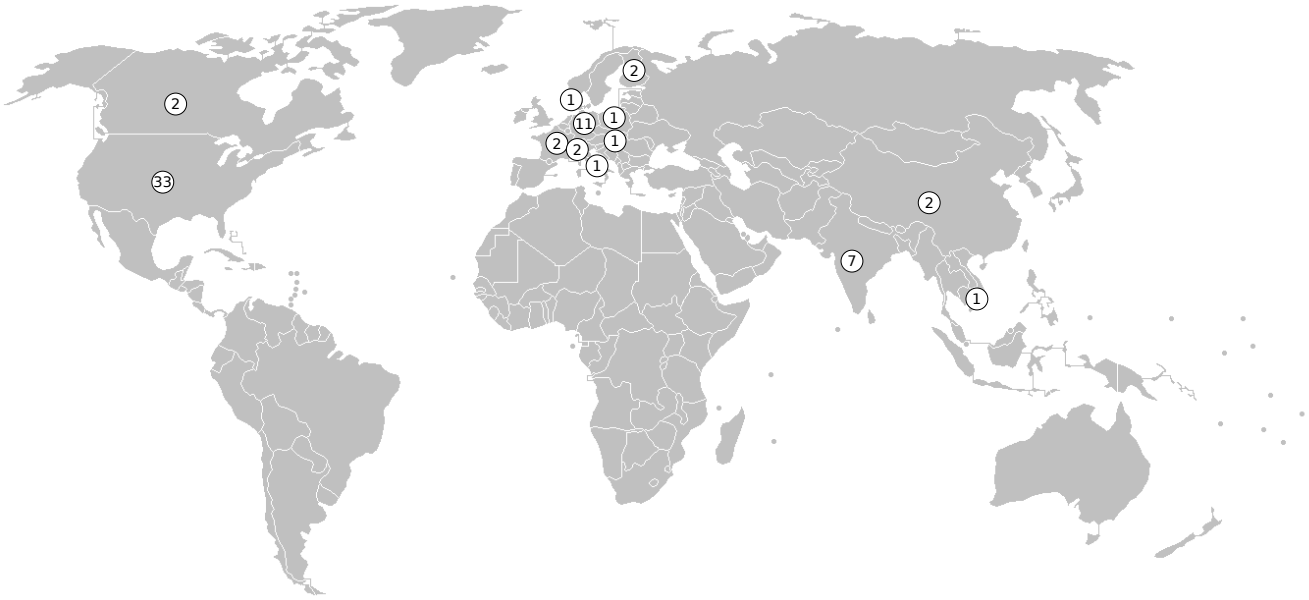

Figura 1.8: Mapa mundi mostrando a distribuição de programas para identificação taxonômica desenvolvida por cada país. Estados Unidos, Europa e Asia são os lugares com maior desenvolvimento de programas de análise taxonômica.

Nas buscas pelo BLAST, as sequências podem ter múltiplas combinações. Programas como o Kraken (WOOD; SALZBERG, 2014), MG-RAST (KEEGAN; GLASS; MEYER, 2015) e MEGAN (HUSON et al., 2007) utilizam uma aproximação chamada "Menor Ancestral Comum" (*Lowest Common Ancestor* (LCA)) (HUSON et al., 2007) que atribui a sequência ao maior nível taxonômico que pode ser classificado como pertencendo ao LCA reduzindo assim o resultado de falsos positivos. Porém essa metodologia possui seu viés. O MEGAN, por exemplo, pode fornecer resultados falsos negativos reduzindo a sensibilidade do *binning* quando descarta sequências que não satisfaçam cortes pré-definidos pelo usuário. Uma vez que o tamanho do genoma está relacionado ao número de leituras em amostras metagenômicas, o Kraken tenta resolver esse problema utilizando um banco de dados com informações de todos os k-mers (frequência de oligonucleotídeos) e dos LCAs de todos os organismos que possuam aquele k-mer exato. Programas como o CARMA (KRAUSE et al., 2008) e o SOrt-ITEMS

(HAQUE et al., 2009) são baseados em BLAST mas não utilizam o LCA. O CARMA realiza a classificação filogenética dos *reads* baseados em domínios conservados do banco de dados de proteínas Pfam. O SOrt-ITEMS adota uma abordagem exaustiva para julgar primeiro a qualidade do alinhamento do *read* com as sequências homólogas na base de dados para encontrar um nível adequado na árvore taxonômica em que o *read* possa ser atribuído. O algoritmo usa então a abordagem de ortologia para identificar sequências homólogas que mostram ortologia significativa, ou seja, semelhança de reciprocidade com o *read* da amostra.

Os programas que utilizam o método de composição de sequência baseiam-se no padrão de disposição dos nucleotídeos, levando em consideração a conservação entre as espécies ao longo da evolução. A ferramenta Phylopythia (MCHARDY et al., 2007) utiliza aprendizado de máquina por *Support Vector Machine* (SVM) junto com marcadores filogenéticos para a identificação dos *reads* de acordo com a composição de oligonucleotídeos de vários tamanhos. O RAIphy (NALBANTOGLU et al., 2011) utiliza um modelo de índice relativo de abundância (RAI), em que um valor de índice é calculado para cada *k-mer*, com base nas estatísticas super e sub-abundância recolhidos a partir do táxon. INDUS (MOHAMMED et al., 2011a) verifica a distância entre as frequências dos vetores de tetranucleotídeos. Sequedex (BERENDZEN et al., 2012) usa um *k-mer* de tamanho 10 para realizar a identificação e localiza *strings* de amino ácidos de tamanho 10 que compartilham ao menos dois genomas de referência de gêneros distintos de bactérias. MetaCV (LIU et al., 2012a) transforma o nucleotídeo em peptídeos de 6 *frames* que é decomposto em oligopeptídeos de tamanho fixo. Estes são utilizados para a classificação taxonômica quando comparado a um banco de dados de proteínas. FOCUS (SILVA et al., 2014) usa a frequência de oligonucleotídeos (k-mers) do genoma completo e calcula o *set* de abundância otimizado do organismos usando o método de mínimos quadrados não negativos. TAC-ELM (RASHEED; RANGWALA, 2012) utiliza oligonucleotídeos e conteúdo GC para criar uma rede neural. *Large scale metagenomics* (VERVIER et al., 2016) usa algoritmos de aprendizagem de máquina para realizar a classificação taxonômica.

Existem programas que utilizam uma combinação da metodologia de alinhamento com o método de composição de sequências, esses métodos são denominados métodos híbridos. O Binpairs (DUTTA et al., 2014) usa uma estratégia da tecnologia do Ilumina para emparelhamento das informações de *reads* pequenos para aumentar a acurácia do *binning* em conjuntos de dados metagenômicos. PhymmBL (BRADY; SALZBERG, 2009) realiza o BLAST junto a modelos interpolados de Markov para caracterizar oligonucleotídeos de tamanho variável típicos de grupos filogenéticos. RITA (MACDONALD; PARKS; BEIKO, 2012) primeiro utiliza um classificador Naive Bayers e depois três algoritmos de BLAST (D-BLASTN, BLASTN e BLASTX). SeMeta (LE; TRAN; TRAN, 2016) separa os *reads* em *clusters* e depois atribui

cada *cluster* ao táxon que melhor se enquadra utilizando como base a similaridade entre os *reads* e a base de dados de referência. SPHINX (MOHAMMED et al., 2011b) possui três etapas: a primeira é composicional e identifica um pequeno *subset* de sequências na base de dados de referência que é similar em composição ao *read* a ser identificado; a segunda utiliza BLAST para verificar o alinhamento do *read* contra o *subset*; a terceira analisa o resultado do BLAST e classifica o *read*. TWARIT (REDDY; MOHAMMED; MANDE, 2012) faz a classificação em duas etapas: a primeira etapa procura sequências iguais às já existentes no banco de dados de referência; a segunda etapa verifica em um banco de dados de composição a distância entre vetores de frequência de tetranucleotídeos. Treephyler (SCHREIBER et al., 2010) baseia-se nas predições do PFAM junto à perfis de modelo oculto de Markov pré calculados da família do PFAM e utiliza uma árvore filogenética de referência para atribuir os táxons aos *reads*.

2 *Objetivos*

Neste trabalho, propomos o desenvolvimento de uma nova metodologia de classificação taxonômica usando medidas de concentração de guanina e citosina (GC %), entropia de dipletes (S2), entropia de tripletes (S3), entropia de tetrapletes (S4) e cálculos de abundância de dinucleotídeos (din). GC % é uma medida utilizada em vários programas de classificação taxonômica, porém as medidas de S2, S3, S4 e din são parâmetros novos a serem avaliados. De posse desses parâmetros verificaremos a eficiência das medidas utilizadas de forma isolada ou combinada para cada nível de táxon (gênero, família, classe e ordem) e quais medidas retornam melhores resultados para classificação taxonômica. A técnica que estamos propondo é inovadora pois utiliza medidas com potencial para melhorar o índice de identificação de microrganismos em amostras de metagenômica.

2.1 Objetivos Específicos

- validação da metologia para os táxons de gênero, família, classe, ordem em bactérias;
- comparação da metodologia com outros *software* de análise composicional de sequências e alinhamento por similaridade em amostras de microbiota humana;

3 *Material e Métodos*

3.1 *Workflow*

O trabalho foi realizado em duas frentes: desenvolvimento da metodologia e comparação com outras metodologias (Figura 3.1). Para o desenvolvimento da metodologia, adquirimos um conjunto de dados referente a genomas completos de bactérias armazenados no NCBI, esses genomas foram filtrados, separados em grupos teste e controle, aplicados cálculos da metodologia e realizado as medições de sensibilidade, especificidade e precisão. Para a comparação com outras metodologias utilizamos amostras da microbiota humana armazenadas no EBI *Metagenomics*, essas amostras foram analisadas pelos programas RAIphy, Phymm e Phymmbl e comparados a metodologia proposta.

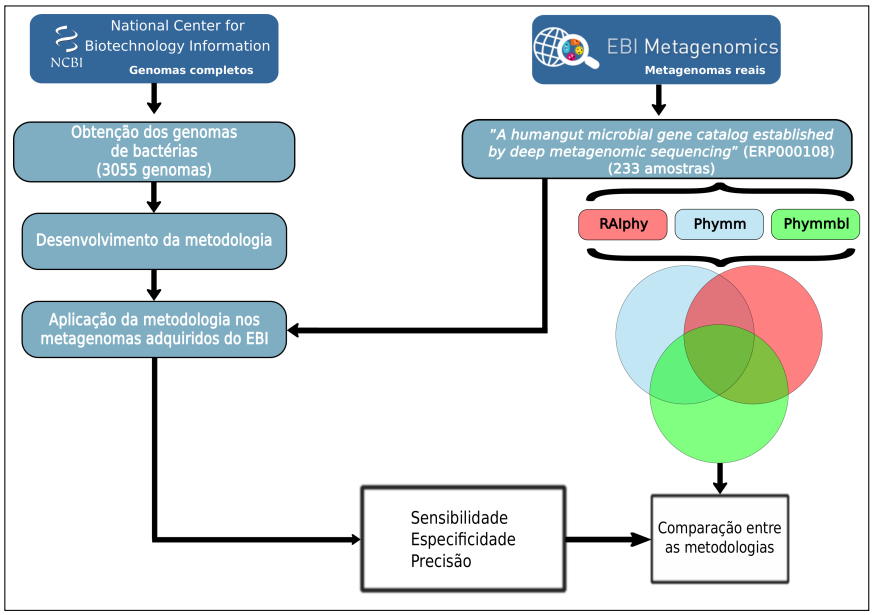

Figura 3.1: *Workflow* com as etapas de desenvolvimento da metodologia e comparação com outras metodologias.

3.2 Aquisição de dados

3.2.1 Conjunto de dados para criação do banco de dados

O conjunto de dados referente a 3055 genomas completos de bactérias, provenientes do NCBI, foi adquirido em 07/07/2015 e usado para a criação do banco de dados com os resultados das medidas, assim como para os arquivos que foram utilizados na comparação com o banco de dados do grupo controle.

3.2.2 Conjunto de dados para comparação com outras ferramentas

O conjunto de dados utilizado para a comparação entre os resultados gerados da nossa metodologia com os programas de análise taxonômica já existentes, foi adquirido com o *download* de 233 amostras disponíveis no EBI genomes em 07/05/2015 do projeto *"A human gut microbial gene catalog established by deep metagenomic sequencing"* (ERP000108). Esse projeto coletou amostras fecais de indivíduos com sobrepeso ou adultos obesos do sexo masculino e feminino, assim como pacientes com doença inflamatória intestinal, de idades entre 18 e 69 anos provenientes de dois países da Europa: Espanha e Dinamarca (QIN et al., 2010).

3.3 *Hardware*

Para as análises foram utilizados computadores do Laboratório de Estudos em Biocomplexidade (BSL) assim como computadores alocados no Instituto de Biotecnologia de Botucatu (IBTEC) e um servidor dedicado no departamento de Laboratório de Genômica Interativa (LGI) (tabela 3.1). Na sessão de resultados é apresentado a quantidade de dados gerados e quantos computadores foram utilizados em cada etapa.

Tabela 3.1: Especificações de memória RAM e quantidades de núcleos dos computadores utilizados nas análises.

	Workstation	Servidor dedicado
Núcleo	8	32
Memória RAM (GB)	16	128

3.4 Medidas

3.4.1 Concentração de GC (GC %)

A concentração de guanina e citosina (GC) possui uma interação mais estável do que a interação adenina e timina (AT) por apresentar 3 ligações de hidrogênio enquanto que AT possui apenas duas com isso quanto maior for a concentração de GC mais estável será a sequência. A concentração de GC tem sido utilizada como medida para classificação taxonômica em programas composicionais (YAKOVCHUK; PROTOZANOVA; FRANK-KAMENETSKII, 2006; GERHARDT; CORSO, 2006).

$$GC\% = \frac{G+C}{A+C+G+T}$$

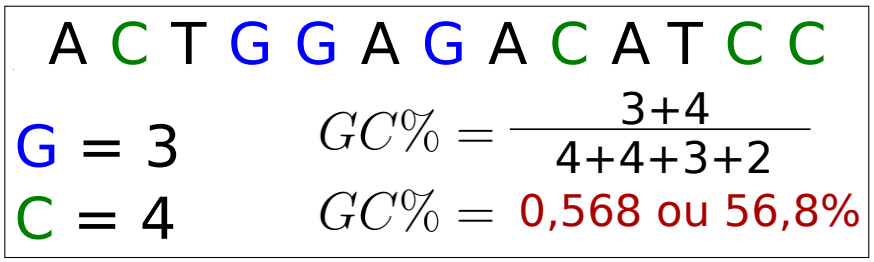

Figura 3.2: Exemplo do cálculo de concentração de GC. A concentração de GC é obtida através da razão entre a soma da totalidade de guanina e citosina pela soma de todos os nucleotídeos que constituem o fragmento.

3.4.2 Entropia (S)

A entropia é a medida da variabilidade de um conjunto de variáveis discretas. A utilização da entropia pode informar o grau de ordem-desordem dentro de um sistema. Neste caso a entropia representa o padrão de distribuição de oligonucleotídios em uma sequência. Utilizamos a entropia de dipletes (S2), entropia de tripletes (S3) e entropia de tetrapletes (S4) (SANTOS; RYBARCZYK-FILHO; GERHARDT, 2011).

O cálculo é realizado estimando a probabilidade de cada oligonucleotídeos em uma janela móvel, dividido pela quantidade total de oligonucleotídeos na sequência (Figura 3.3). Esse cálculo é normalizado tendendo de valores entre 0 a virtualmente 1. Entre as entropia de dipletes, tripletes e tetrapletes temos variações na equação em relação a janela e também ao total de combinações de nucleotídeos existentes, para a entropia de dipletes temos um total de 16 combinações, para entropia de tripletes e tetrapletes temos respectivamente 64 e 256

combinações de nucleotídeos possíveis. Com uma sequência com baixa variação de oligonucleotídeos o resultado se aproxima de 0.

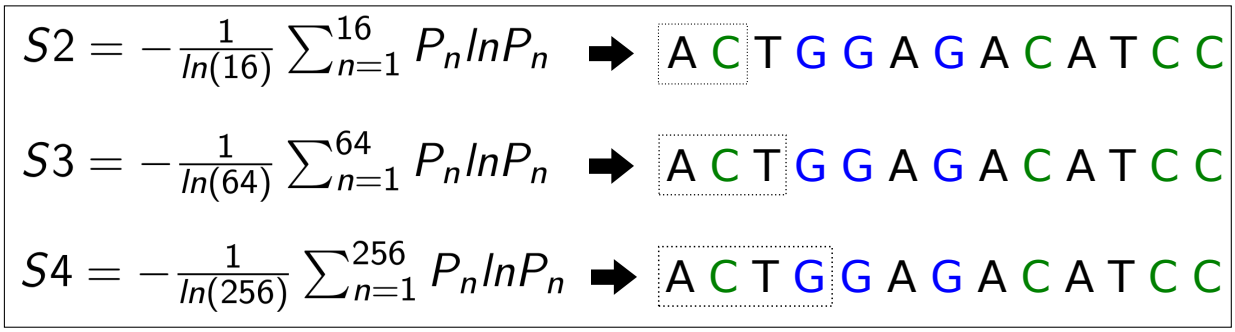

Figura 3.3: Equações utilizadas para entropia.

3.4.3 Abundância total de dinucleotídeos (din)

A abundância total é uma medida estatística para cálculo de frequência de dinucleotídeos. Através da análise dos padrões de frequência podemos inferir a diferença entre as espécies e entre os organismos de diferentes ambientes. A forma como estão distribuídos os dinucleotídeos influenciam diretamente a estrutura da molécula de DNA, podendo influenciar na interação com proteínas, principalmente as que estão relacionadas a replicação e tradução. (KARLIN; BURGE, 1995; KARLIN, 1998). Devido a importância desses processos para as células, os organismos possuem padrões distintos de frequências de dinucleotídeos entre eles (DUTTA; PAUL, 2012) permitindo assim a utilização dessa medida como parâmetro de identificação de organismos (Figura 3.4). Para a equação de abundância total de dinucleotídeos temos a somatória da frequência dos nucleotídeos XY, dividido pela frequência do nucleotídeos X e nucleotídeo Y aonde X e Y pertencem ao conjunto adenina, citosina, timina e guanina. A somatória de todas as frequências são divididas por 16 que são o total de combinações de dinucleotídeos.

$$din = \frac{1}{16} \sum_X \sum_Y \frac{f_{XY}}{f_X f_Y}, X \in \{A, C, T, G\}, Y \in \{A, C, T, G\}$$

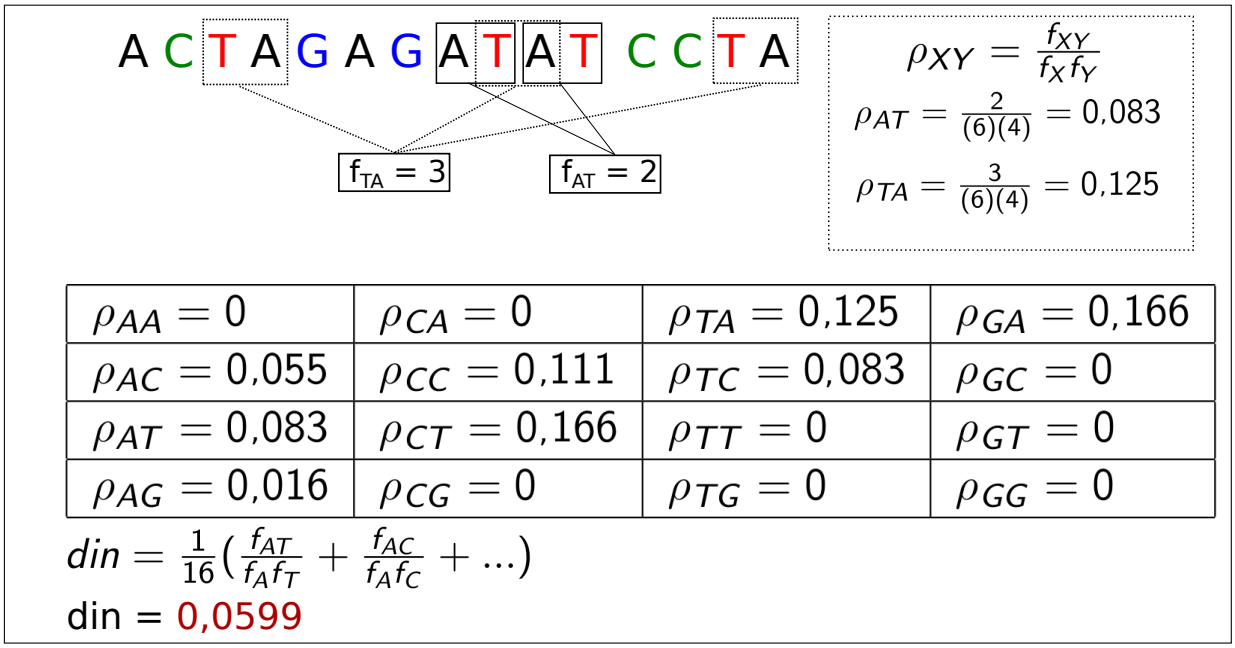

Figura 3.4: Cálculo de abundância total de dinucleotídeos. É a média da somatória da razão de todas as combinações de dinucleotídeos possíveis.

3.5 Filtragem dos dados para criação do banco de dados

Foram selecionados os genomas obtidos do NCBI que possuíssem mais de 70000 pb e que tivessem apenas os nucleotídeos A, C, T e G. Esse tamanho foi escolhido para que pudéssemos separá-los em dois grupos (teste e controle) e tivéssemos uma amostragem de arquivos com baixa possibilidade de *reads* duplicados para cada um dos organismos. Os organismos do grupo controle possuem os primeiros 50 % do total de nucleotídeos do genoma. Para os organismos do grupo teste foram utilizados os outros 50 %. Separamos os genomas por táxons (espécie, gênero, família, ordem, classe e filo) e realizamos cortes na sequência do DNA desses organismos utilizando como parâmetros tamanhos entre 50 a 1000 nucleotídeos (passos de 50 pb). Estes tamanhos foram selecionados visando abranger os diversos tamanhos de *reads* gerados pelas tecnologias de sequenciamento (Figura 3.5).

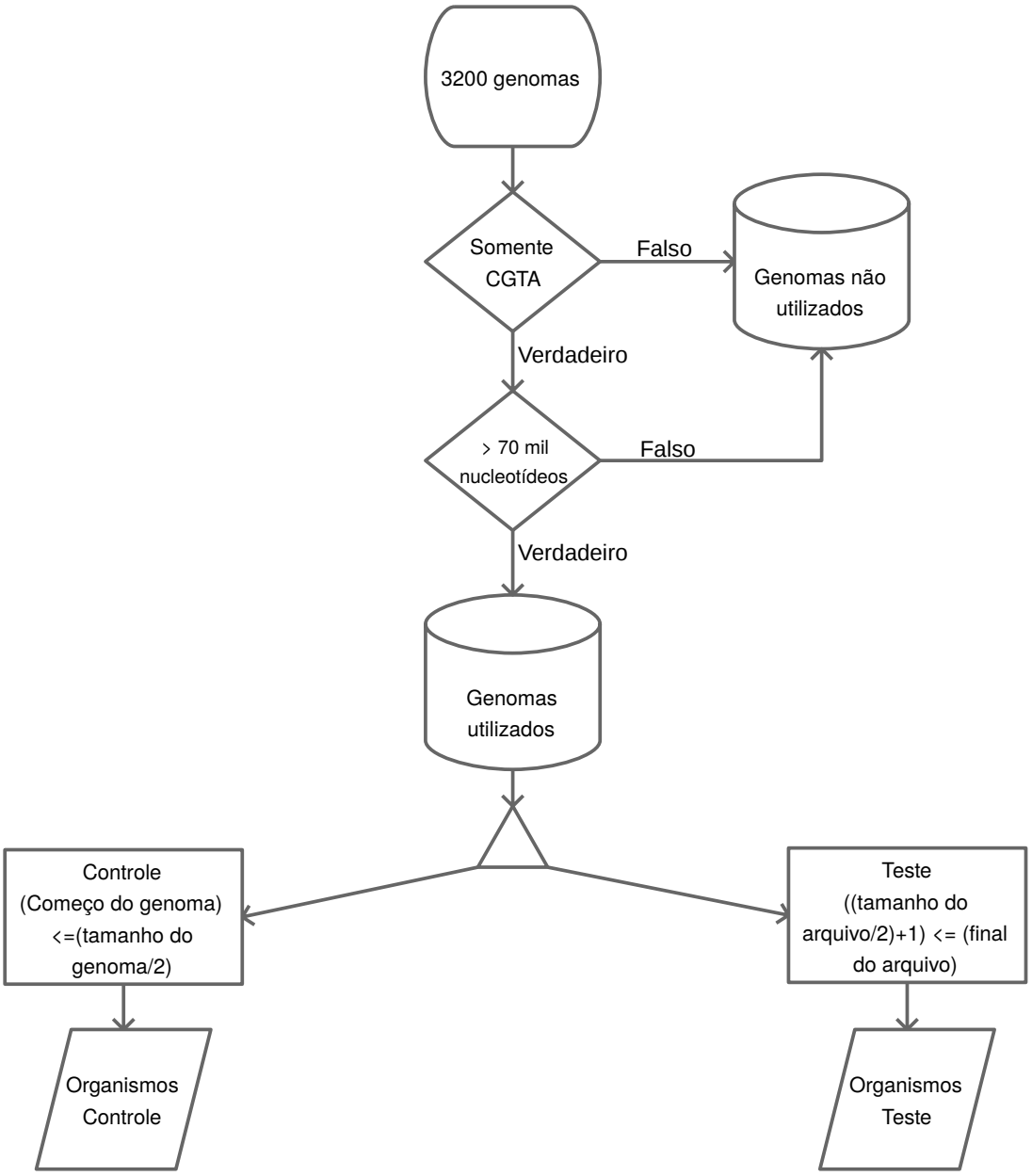

Figura 3.5: Fluxograma dos passos para filtragem e separação dos genoma. A partir dos 3200 genomas completos adquiridos do NCBI *genomes* foi realizado um filtro para selecionar apenas os que possuíam nucleotídeos A,C,G e T. O segundo filtro foi realizado para escolher os genomas maiores que 70 mil nucleotídeos afim de diminuir a chance de fragmentos com repetição para os grupos teste e controle

3.6 Grupo controle

Para o grupo controle, foram criados 1000 arquivos para cada organismo contendo 1000 *reads* do mesmo organismo. Isto foi realizado para os diversos tamanhos de *reads* propostos anteriormente que abrangem tamanhos entre 50 a 1000 nucleotídeos. Estes *reads* foram selecionados a partir da primeira metade do genoma completo do organismo e foram selecionados de

forma aleatória. Ao completar os 1000 *reads* de cada arquivo o algoritmo verifica se já foram criados os 1000 arquivos, caso essa condição não seja verdadeira ele cria o próximo de forma automática, o processo só é finalizado após ter sido criados todos os arquivos para todos os organismos daquele táxon. (Figura 3.6).

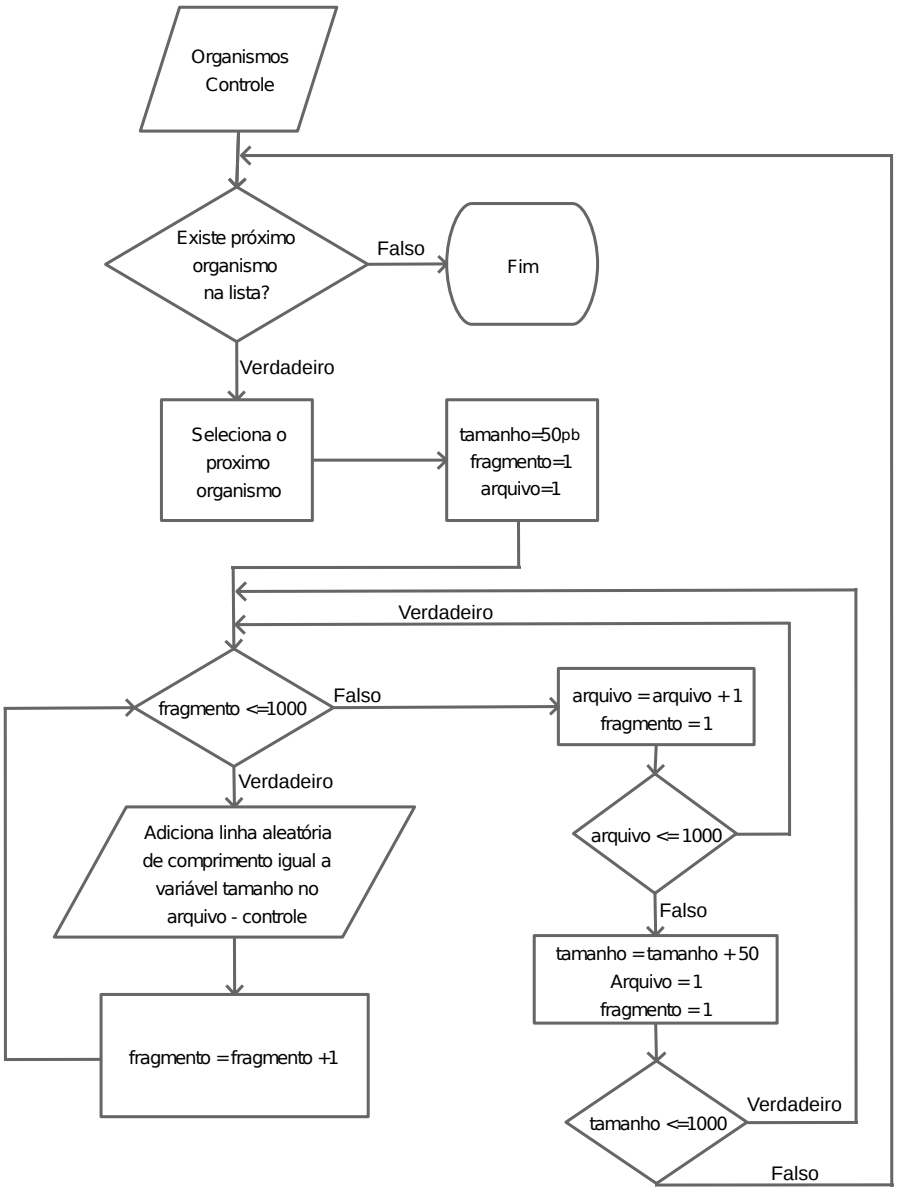

Figura 3.6: Fluxograma para criação do grupo controle. Para o grupo controle foram selecionados todos os organismos e criados 1000 arquivos contendo 1000 fragmentos aleatórios para tamanhos de fragmentos de 50 a 1000 nucleotídeos.

Com os arquivos do grupo controle calculamos a concentração de GC %, entropia de dipletes, entropia de tripletes, entropia de tetrapletes e abundância de dinucleotídeos para cada fragmento do arquivo (Figura 3.7).

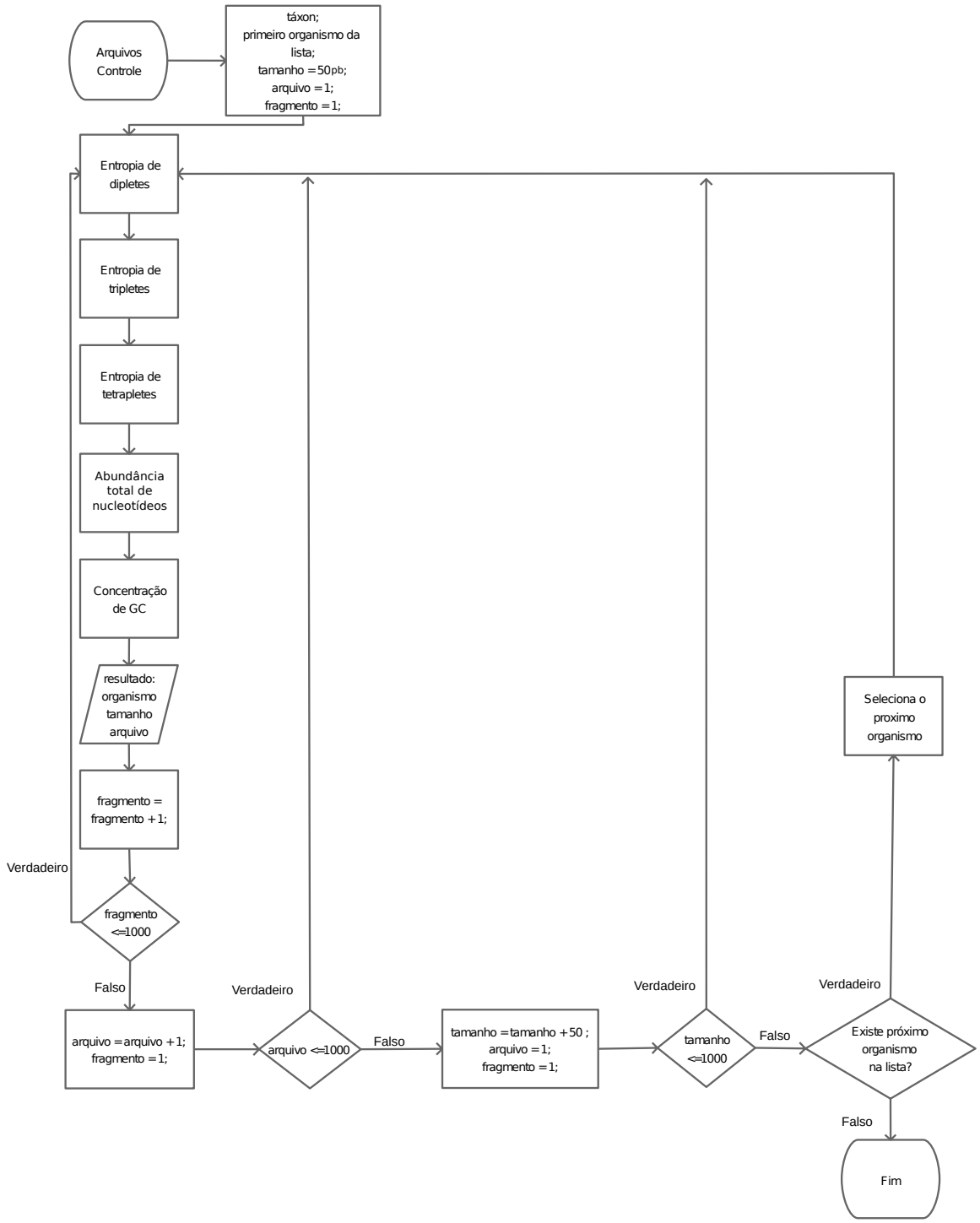

Figura 3.7: Fluxograma para criação do grupo controle. Para todos os arquivos contendo fragmentos do grupo controle foram aplicados os cálculos de entropia de dipletes, entropia de tripletes, entropia de tetrapletes, abundância total de dinucleotídeos e concentração de GC.

Utilizando os arquivos gerados com os resultados das medidas do grupo controle, criamos um banco de dados (Figura 3.8) com as informações de média, variância e desvio padrão para todas as medidas. Esse banco de dados foi criado utilizando Linguagem de Consulta Estruturada (SQL) que fornece alta confiabilidade e velocidade nas pesquisas realizadas dentro das tabelas existentes no banco de dados.

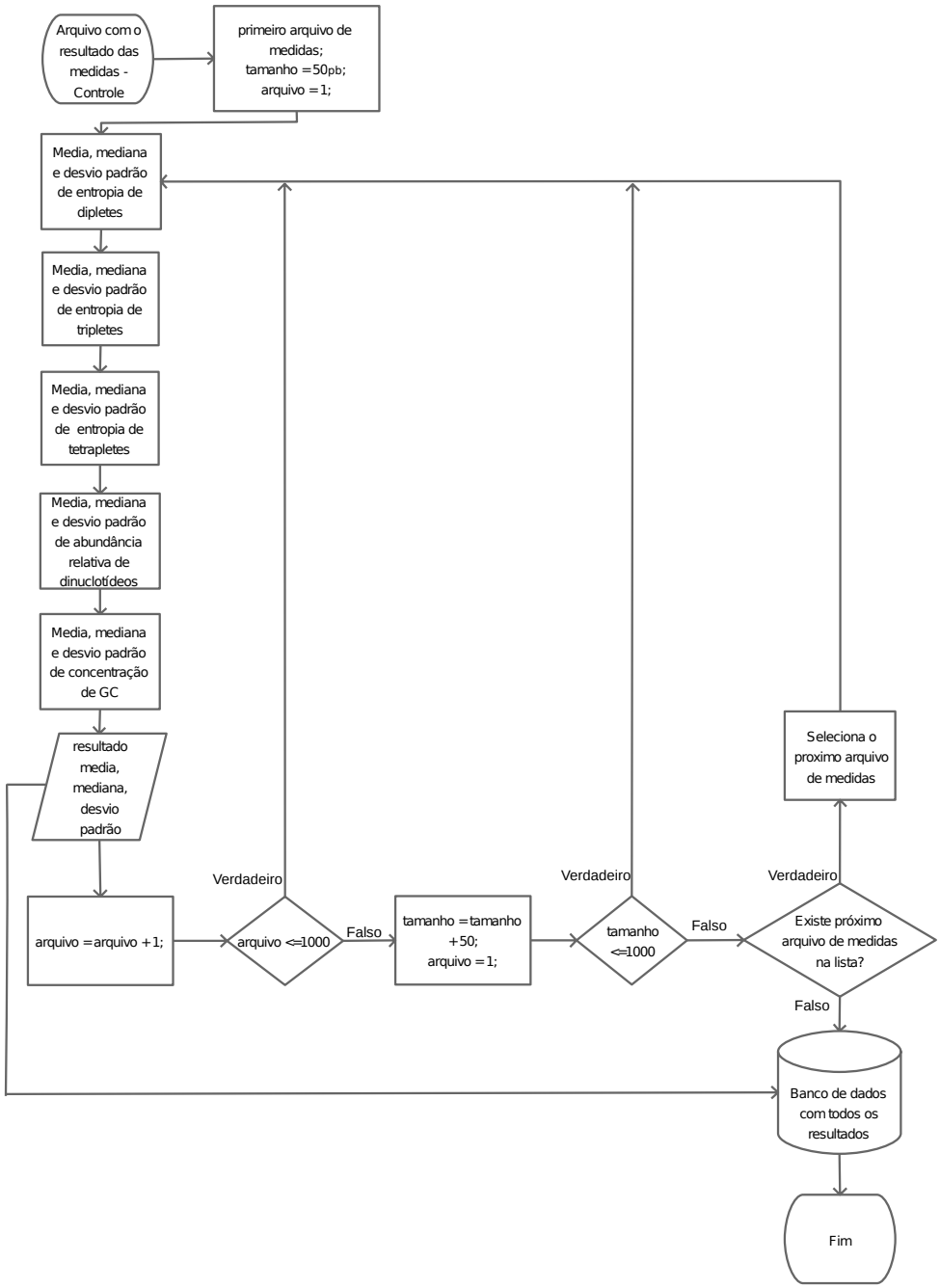

Figura 3.8: Fluxograma da criação do banco de dados com resultados dos cálculos de concentração de GC %, entropia de dipletes, entropia de tripletes, entropia de tetrapletes, abundância total de dinucleotídeos para os arquivos do grupo controle.

3.7 Grupo teste

Foram gerados 100 arquivos, para cada um dos tamanhos que utilizamos como parâmetro (50 a 1000 pb), contendo aleatoriamente 500 *reads* de organismos pertencentes a cada um dos táxons de interesse e 500 *reads* diferentes desse táxon. Esses arquivos de amostragem fazem parte do grupo teste (Figura 3.9).

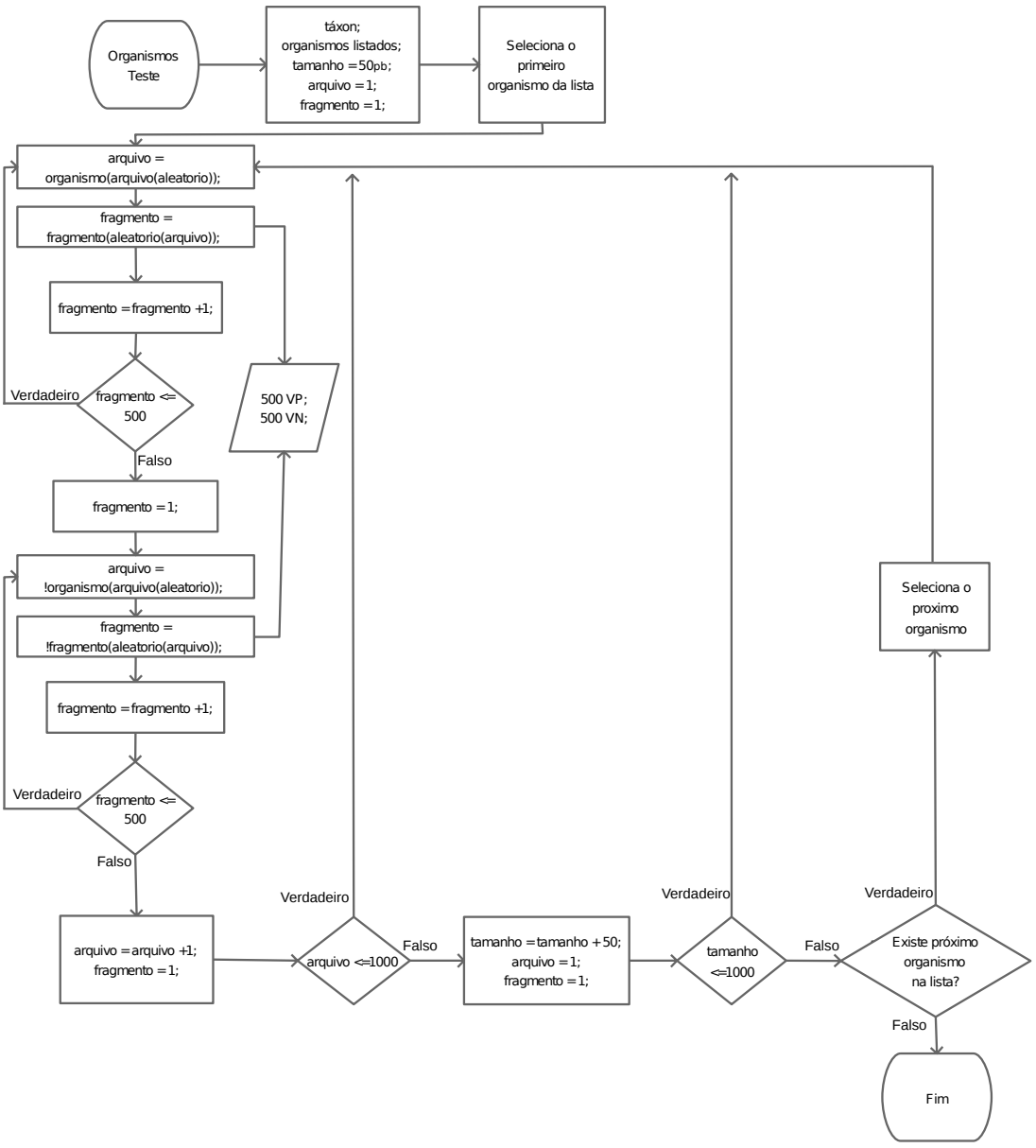

Figura 3.9: Fluxograma para criação do grupo teste. Para o grupo teste foram gerados 1000 arquivos para cada organismo em cada tamanho de fragmento indo de 50 a 1000, cada arquivo possui 500 fragmentos aleatórios do organismo alvo e 500 fragmentos aleatórios de outro organismo.

3.8 Medidas estatísticas

Comparando o grupo teste com o banco de dados podemos aferir as medidas estatísticas de sensibilidade, especificidade, precisão e média harmônica para todos os cálculos isolados e as intersecções entre as combinações, totalizando 29 possibilidades (Tabela 3.2). Os resultados das combinações foram analisados pelo programa RStudio para a criação dos gráficos.

Tabela 3.2: Tabela de identificadores das combinações de medidas.

ID	Medidas	ID	Medidas	ID	Medidas
1	din	11	S3 ∩ S4	21	GC % ∩ S2 ∩ S3 ∩ S4
2	S2	12	S3 ∩ S4 ∩ din	22	GC % ∩ S2 ∩ Entropy3 ∩ S4 ∩ din
3	S2 ∩ din	13	S4	23	GC % ∩ S2 ∩ S4 ∩ din
4	S2 ∩ S3	14	S4 ∩ din	24	GC % ∩ S3
5	S2 ∩ S3 ∩ din	15	GC %	25	GC % ∩ S3 ∩ din
6	S2 ∩ S3 ∩ S4	16	GC % ∩ din	26	GC % ∩ S3 ∩ S4
7	S2 ∩ S3 ∩ S4 ∩ din	17	GC % ∩ S2	27	GC % ∩ S3 ∩ S4 ∩ din
8	S2 ∩ S4	18	GC % ∩ S2 ∩ din	28	GC % ∩ S4
9	S3	19	GC % ∩ S2 ∩ S3	29	GC % ∩ S4 ∩ din
10	S3 ∩ din	20	GC % ∩ S2 ∩ S3 ∩ din		

Fonte: O autor (2016)

3.8.1 Matriz de confusão

A matriz de confusão é um conjunto de medidas estatísticas que permite mensurar, quantificar e categorizar valores de: verdadeiro positivo (VP), falso positivo (FP), falso negativo (FN) e verdadeiro negativo (VN) na comparação entre o grupo teste e o grupo controle. Verdadeiros positivos são os fragmentos que foram registrados como sendo do organismo de referência correto do grupo controle. Os falsos positivos são fragmentos que não pertencem ao organismo de referência mas que obtiveram resultados indicando como sendo desse organismo. Os falsos negativos são fragmentos que pertencem ao organismo de referência mas que obtiveram resultados identificando como não sendo desse organismo. Os verdadeiros negativos são os fragmentos que não pertencem ao organismo de referência e que obtiveram resultados indicando como não sendo desse organismo.

3.8.2 Sensibilidade

Evidencia a quantidade de *reads* VP que assim foram caracterizados. Para exemplificar essa medida, podemos analisar o caso de um conjunto de *reads* do gênero *Escherichia* (teste) que realmente foram caracterizados como sendo do gênero *Escherichia* (controle) (Figura 3.10 A).

$$\text{sens} = \frac{VP}{VP+FN}$$

3.8.3 Especificidade

Representa a fração de *reads* VN que assim foram caracterizados. A especificidade informará, por exemplo, se os *reads* que não são do gênero *Escherichia*(teste) foram caracterizados como tal (Figura 3.10 B).

$$\text{espec} = \frac{VN}{VN+FP}$$

3.8.4 Precisão

Quantifica os *reads* preditos corretamente como VP. A precisão utiliza como parâmetros os FP. Por exemplo, verificará quantos *reads* não são do gênero *Escherichia* (teste) mas foram caracterizados como pertencendo por estarem dentro dos valores desse gênero (Figura 3.10 C).

$$\text{prec} = \frac{VP}{VP+FP}$$

3.8.5 Média harmônica entre sensibilidade e especificidade

Calculamos a média harmônica para saber a média das medidas de sensibilidade e especificidade (Figura 3.10 D).

$$\text{harm} = \frac{2VP}{(2VP+FP+FN)}$$

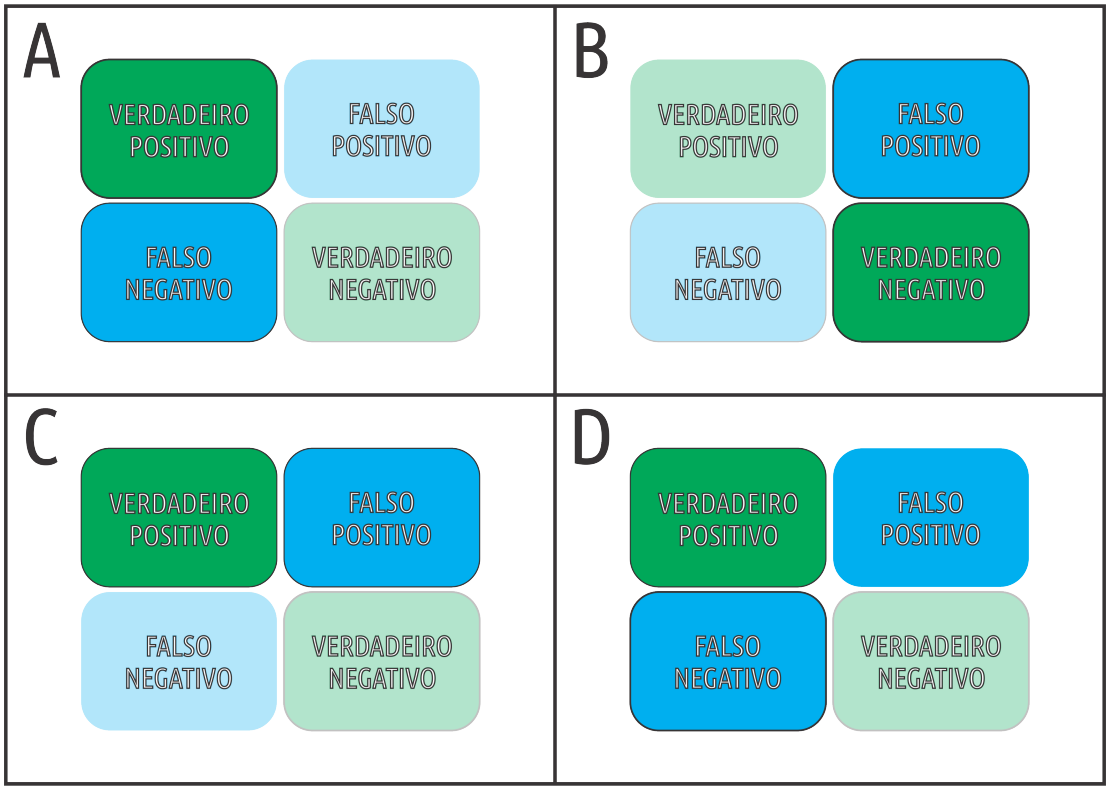

Figura 3.10: Matriz demonstrando como são calculadas as medidas estatísticas de: (a) sensibilidade, (b) especificidade, (c) precisão e (d) média harmônica entre sensibilidade e especificidade. Os retângulos com cores escuras na figura são representam os valores utilizadas para cada medida estatística. A sensibilidade é calculada utilizando os valores de VP e FN. Para o cálculo de especificidade são utilizados os valores de FP e VN. Para a precisão são utilizados VP e FP e para a média harmônica são utilizados VP, FP e FN.

3.9 Comparação com outras metodologias

Dentre os programas testados foram selecionados 4: Raiphy (composicional), Phymm (composicional), Phymmbl (alinhamento) e Kraken (alinhamento). Estes programas foram utilizados para análise do conjunto de microbiota fecal humana (seção 3.2.2) para posterior análise com a nova metodologia proposta.

3.10 Critério de Corte

Para avaliar os resultados da metodologia proposta, definimos o valor mínimo de corte igual ou superior à 70 % para a especificidade e sensibilidade ao analisarmos as combinações de medidas. Esse valor é acima do valor médio dos algoritmos utilizados por programas de análise taxonômica conforme podemos averiguar na figura 3.11 (NALBANTOGLU et al., 2011). Nesse trabalho aplicamos a metodologia nos táxons gênero, família, classe e ordem.

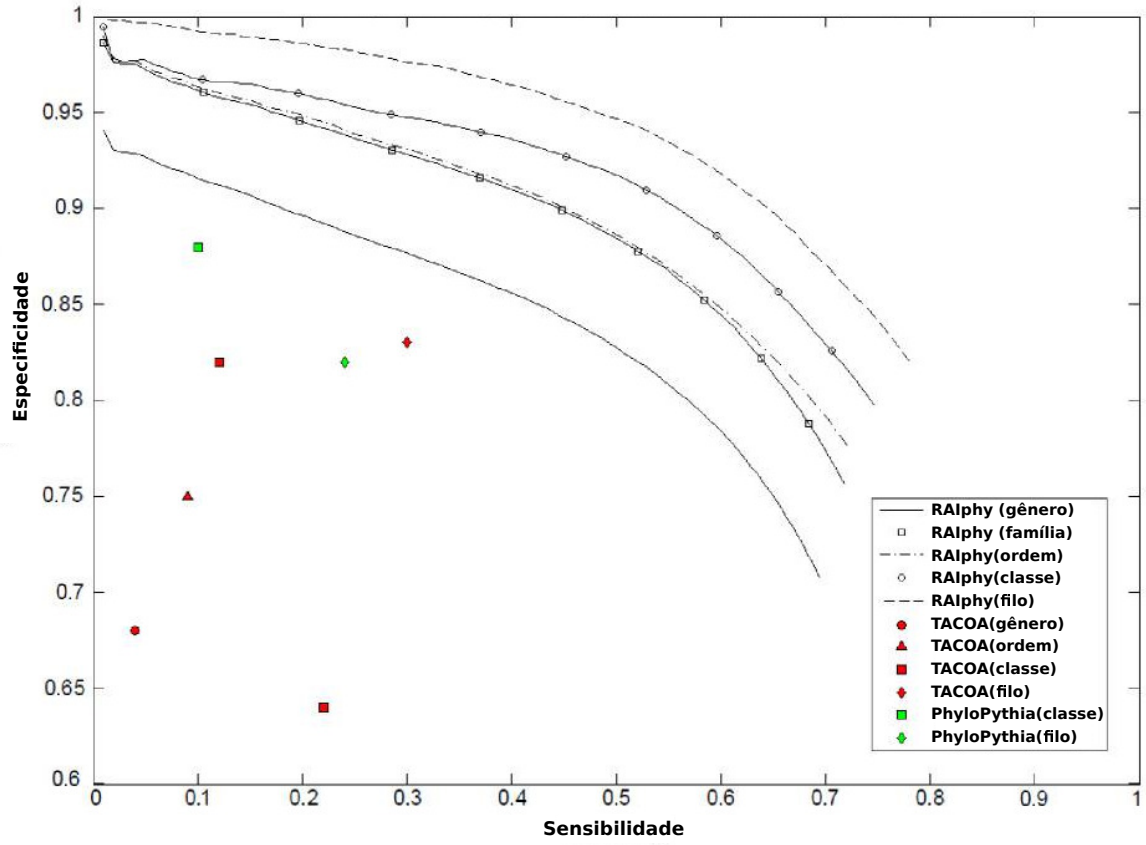

Figura 3.11: Características operantes de sensibilidade e especificidade para classificação taxonômica de fragmentos de 1 kpb das ferramentas RAIphy, TACOA e PhyloPythia. Adaptado de (NALBANTOGLU et al., 2011).

4 Resultados e Discussão

4.1 Processamento das informações

Para analisar os dados prospectados dos repositórios do NCBI e EBI metagenoma foram utilizados até sete *workstation* para as etapas dos fluxogramas apresentados no capítulo 3. Alguns metagenomas requisitaram um tempo superior a seis semanas para serem analisados por *software* de alinhamento. Para as análises do Phymm/Phymmbl foi utilizado um servidor dedicado com alto poder de processamento. Conforme a tabela 4.1, as etapas de análise dos dados prospectados do NCBI resultou em 14,25 TB de informação, enquanto que os *software* análise de metagenomas resultou em 463,6 GB de informação, salientado que os *software* Phymm e Phymmbl analisaram somente 87 % dos dados. Restam 21 % que ainda estão sendo processados.

Tabela 4.1: Tabela com informações relacionadas a geração de dados durante a análise dos táxons.

Programa/Etapa	Status concluído	*Workstation* utilizadas	Dados gerados programas/validação
Método para gênero (202)	100 %	4	2,5 TB
Método para família (147)	100 %	1	1,9 TB
Método para ordem (87)	100 %	1	1,7 TB
Método para classe (35)	100 %	1	1,4 TB
Grupo teste para todos os táxons	100 %	7	6,75 TB
Raiphy	100 %	3	463,6 GB
Phymmbl	87 %	7 + 1 servidor dedicado	
Phymm	87 %	7 + 1 servidor dedicado	

4.2 Avaliação Global

4.2.1 Táxon Gênero

Uma análise geral do táxon gênero mostra que conforme o tamanho do fragmento aumenta, as combinações tendem a melhorar a identificação, como podemos verificar no exemplo das combinações de medidas S2 ∩ S3 ∩ S4 para o gênero *Bartonella* (Figura 4.1). Esse gráfico do tipo *boxplot* é dividido em quatro partes, cada um deles referente a uma das medidas estatísticas que definimos na seção 3.8.1. São eles: sensibilidade, especificidade, precisão e média harmônica entre a sensibilidade e especificidade.

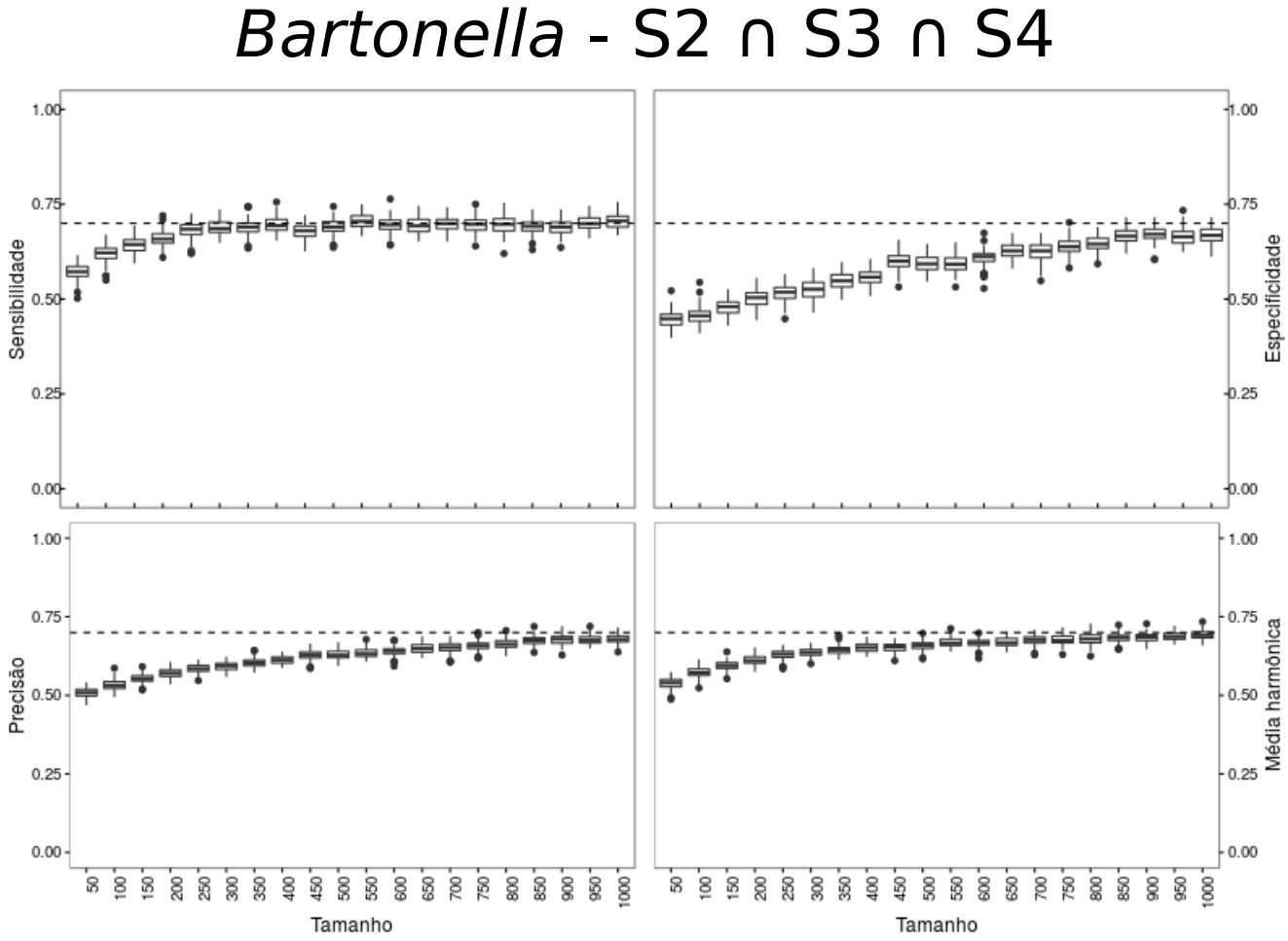

Figura 4.1: Relação de tamanho do fragmento com as combinações de medidasS2 ∩ S3 ∩ S4 para identificação do gênero *Bartonella.*. A sensibilidade e a precisão começam com um percentual de identificação maior que 50 % para tamanhos de fragmentos 50 pb, a especificidade começa com valor próximo a 40 %. Temos uma tendência de aumento do percentual de identificação conforme os tamanhos dos fragmentos aumentam.

O valor de sensibilidade é maior que 50 % para tamanhos de 50 pb e aumenta até atingir o seu platô em *reads* de tamanho 250 pb com o valor aproximadamente em 70 %. O valor de especificidade para *reads* de tamanho 50 pb não atinge os 50 %, mas conforme o tamanho dos fragmentos aumentam a especificidade também aumenta chegando próximo da identificação em 70 % a partir de *reads* com 850 pb. Neste exemplo, a precisão também começou com valores próximos a 50 % e houve um aumento parecido com o da especificidade, também atingindo valores próximos a 70 % para *reads* com tamanho acima de 850 pb. Com esses resultados, podemos afirmar que esta combinação só poderá ser utilizada em *reads* de tamanho minímo de 850 pb para o gênero *Bartonella*.

4.2.2 Táxon Família

Para o táxon família verificamos que há um aumento na porcentagem de identificação de *reads* de tamanho 50-200 pb e para *reads* maiores temos a formação de um platô. Como exemplo, temos a combinação de GC % e S2 para a família *Acetobacteraceae* (Figura 4.2).

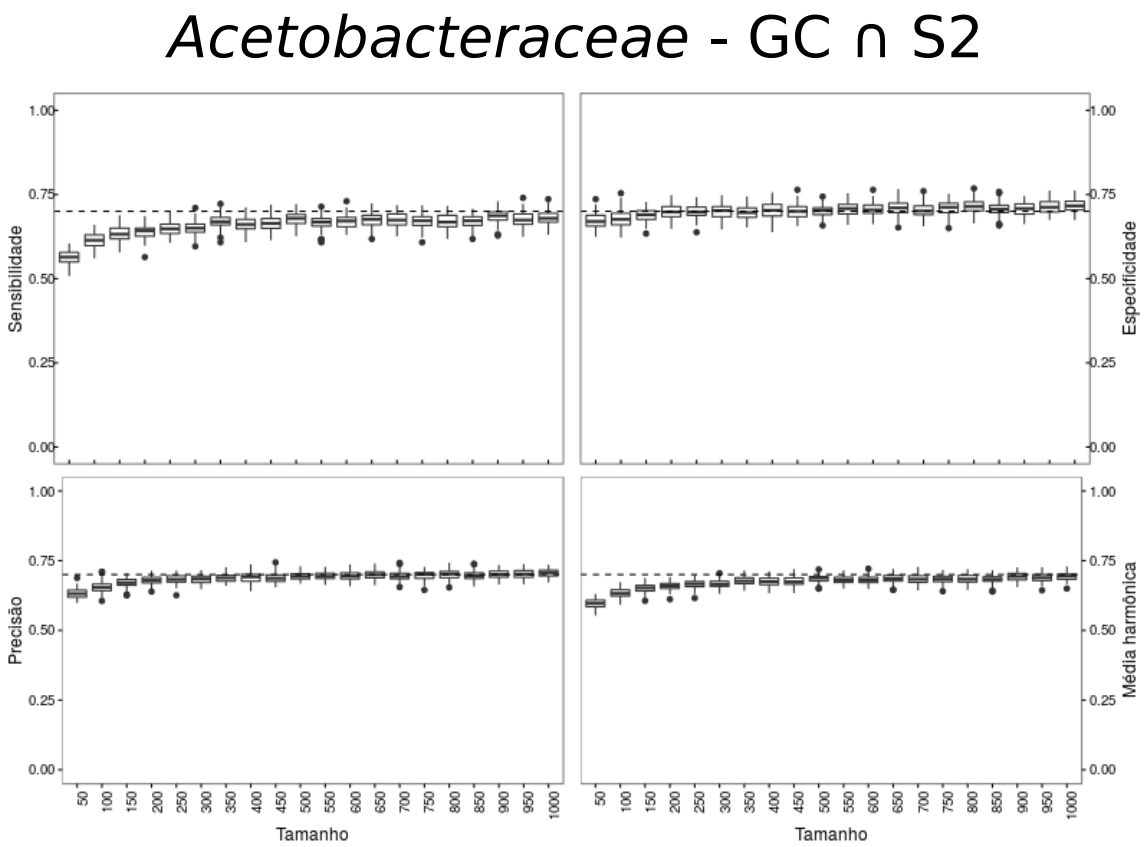

Figura 4.2: Relação de tamanho do fragmento com as combinações de medidas GC % e S2 para identificação da família *Acetobacteraceae*. Na família *Acetobacteraceae* verificamos o surgimento de platô no percentual de identificação a partir dos fragmentos de tamanho 200 pb.

4.2.3 Táxon Ordem

Obtivemos melhores resultados no táxon ordem utilizando as medidas individuais, ou seja, sem estarem combinadas com outras medidas. No exemplo da ordem *Flavobacteriales* (Figura 4.3) a medida GC % identificou mais de 70 % dos organismos para 19 dos 20 tamanhos de fragmentos analisados, com isso somente o tamanho 50 pb ficou fora do corte estabelecido. Em uma análise geral, no táxon ordem também houve melhora na identificação dos organismos com o aumento do tamanho dos fragmentos.

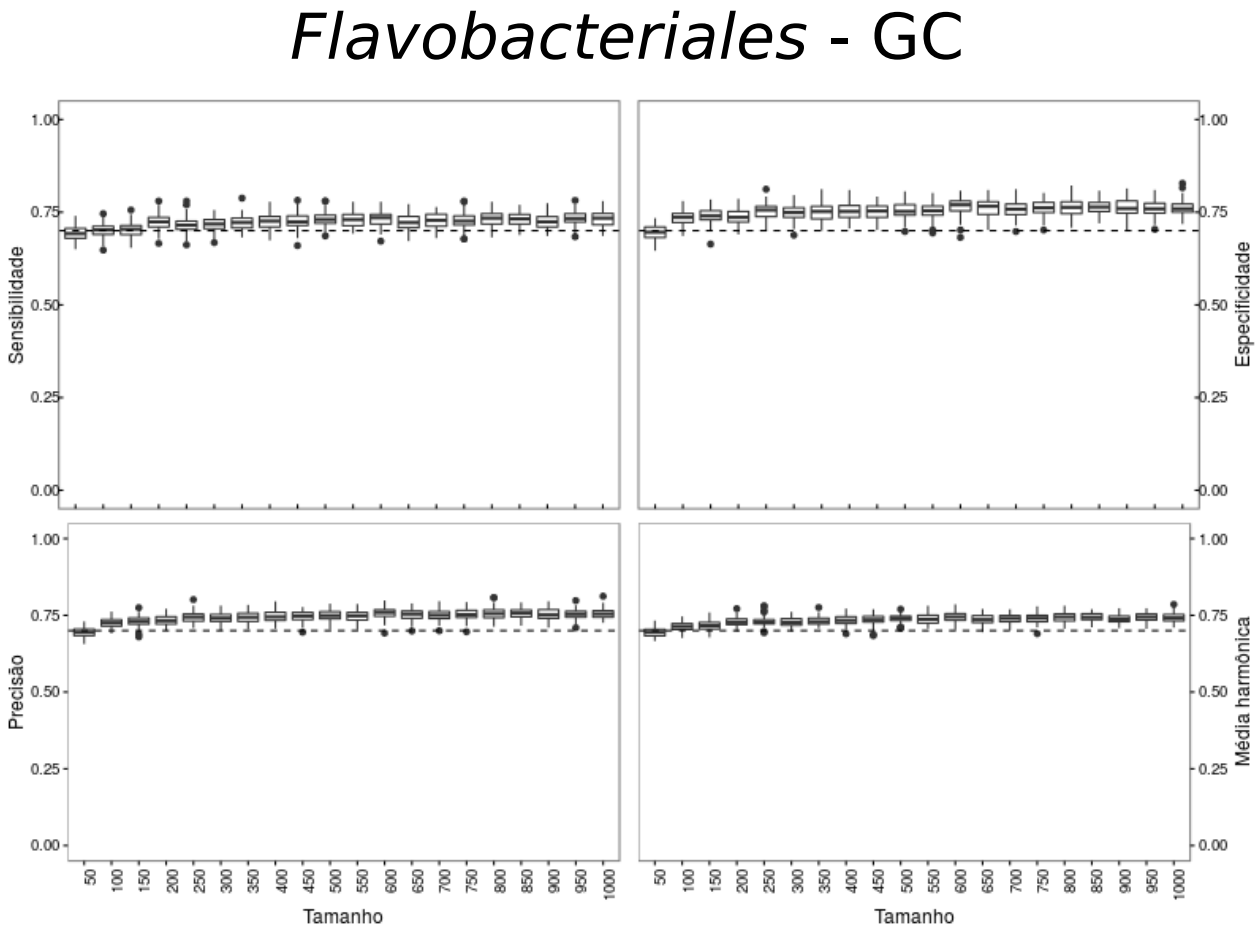

Figura 4.3: Relação de tamanho do fragmento utilizando a medida GC % para identificação da ordem *Flavobacteriales*. Para a ordem *Flavobacteriales* somente o tamanho de fragmento 50 pb não conseguiu atingir os 70 % de identificação para sensibilidade, precisão e especificidade.

4.2.4 Táxon Classe

Uma das medidas que se mostrou interessante para o táxon classe foi a combinação entre S2 e din, conforme podemos observar na classe *Dehalococcoidia* (Figura 4.4) na qual o percentual de identificação começa em mais de 50 % para os tamanhos de fragmentos 50 pb e atinge o percentual de 70 % para tamanhos maiores que 250 pb. Nesse táxon a medida GC % também foi predominante.

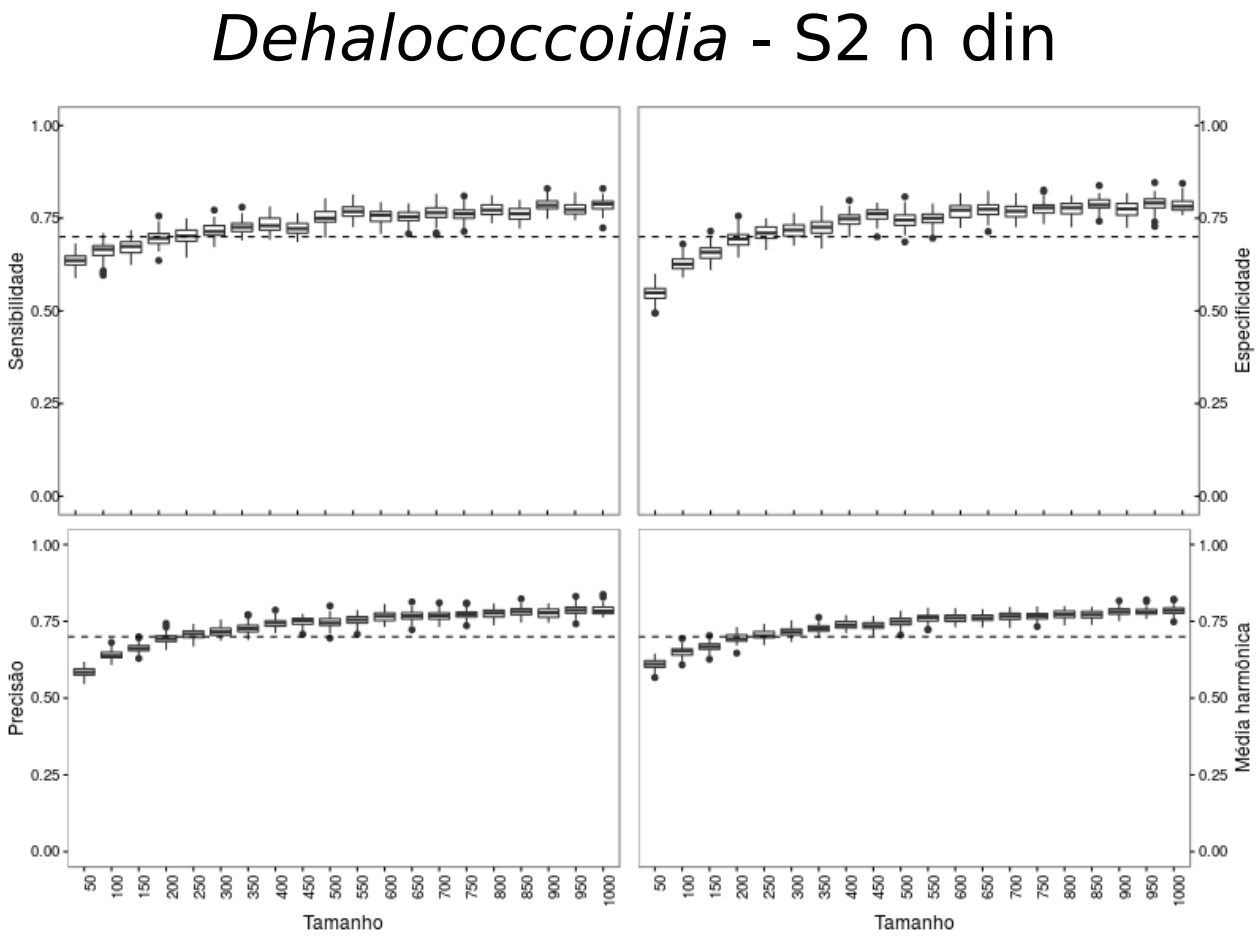

Figura 4.4: Relação de tamanho do fragmento utilizando as combinações de medidas S2 e din para identificação da classe *Dehalococcoidia*. A classe *Dehalococcoidia* obtive percentual de identificação acima de 70 % para fragmentos maiores que 250 pb.

4.3 Aplicação da metodologia no táxon gênero

Ao analisarmos as 29 possíveis combinações para os 202 conjuntos de organismos do táxon gênero, com tamanhos entre 50 pb e 1000 pb, obtivemos 5480 resultados acima do valor definido de 70 % e identificamos 173 gêneros (Tabela 4.2). Dessas combinações, a que obteve maior êxito na identificação do gênero dos organismos foi a 15-GC % que sozinha representa 42,75 % de todas as identificações, o que corresponde a 167 gêneros.

Tabela 4.2: Tabela de combinações para o táxon gênero com valores de especificidade, sensibilidade, precisão e média harmônica com valores iguais ou acima de 70 %.

ID	Medida	Fragmentos Identificados	ID	Medida	Fragmentos Identificados
1	din	184	16	GC % ∩ din	118
2	S2	379	17	GC % ∩ S2	187
3	S2 ∩ din	170	18	GC % ∩ S2 ∩ din	39
4	S2 ∩ S3	203	19	GC % ∩ S2 ∩ S3	68
5	S2 ∩ S3 ∩ din	47	20	GC % ∩ S2 ∩ S3 ∩ din	16
6	S2 ∩ S3 ∩ S4	69	21	GC % ∩ S2 ∩ S3 ∩ S4	25
7	S2 ∩ S3 ∩ S4 ∩ din	25	22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	8
8	S2 ∩ S4	108	23	GC % ∩ S2 ∩ S4 ∩ din	9
9	S3	474	24	GC % ∩ S3	87
10	S3 ∩ din	83	25	GC % ∩ S3 ∩ din	19
11	S3 ∩ S4	188	26	GC % ∩ S3 ∩ S4	38
12	S3 ∩ S4 ∩ din	25	27	GC % ∩ S3 ∩ S4 ∩ din	9
13	S4	449	28	GC % ∩ S4	54
14	S4 ∩ din	43	29	GC % ∩ S4 ∩ din	13
15	GC %	2343			

Fonte: O autor (2016)

A única combinação que conseguiu realizar as identificações em tamanhos de fragmento de 50 pb foi a 15-GC %, que identificou 11.38 % do total de organismos como nos exemplos dos gêneros *Amycolatopsis* (Fig 4.5) e *Halanaerobium* (Fig 4.6).

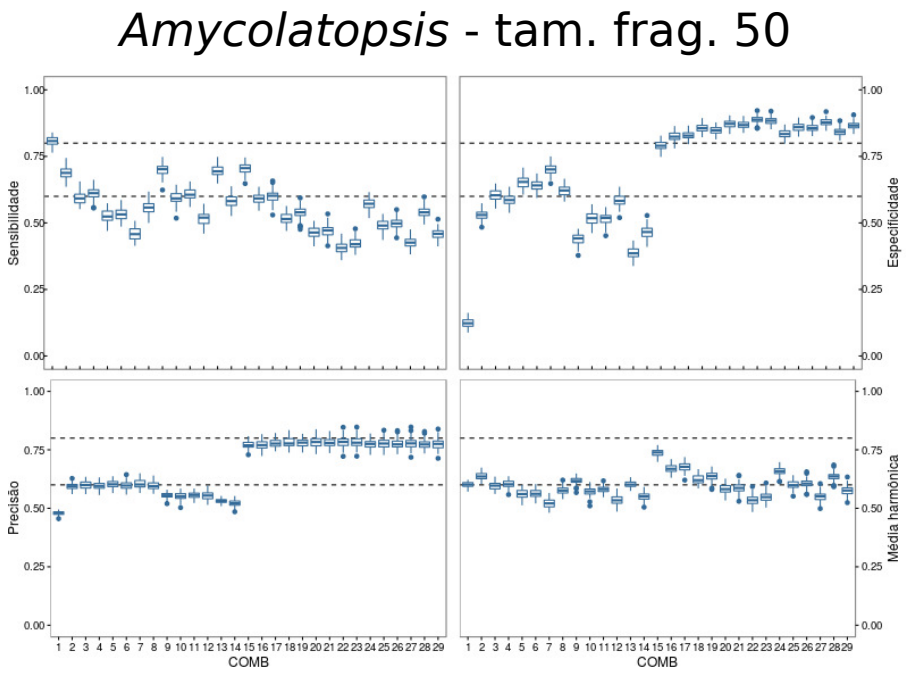

Figura 4.5: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para o gênero *Amycolatopsis*. A medida GC foi a única que conseguiu identificação acima de 70 % para sensibilidade, precisão e especificidade.

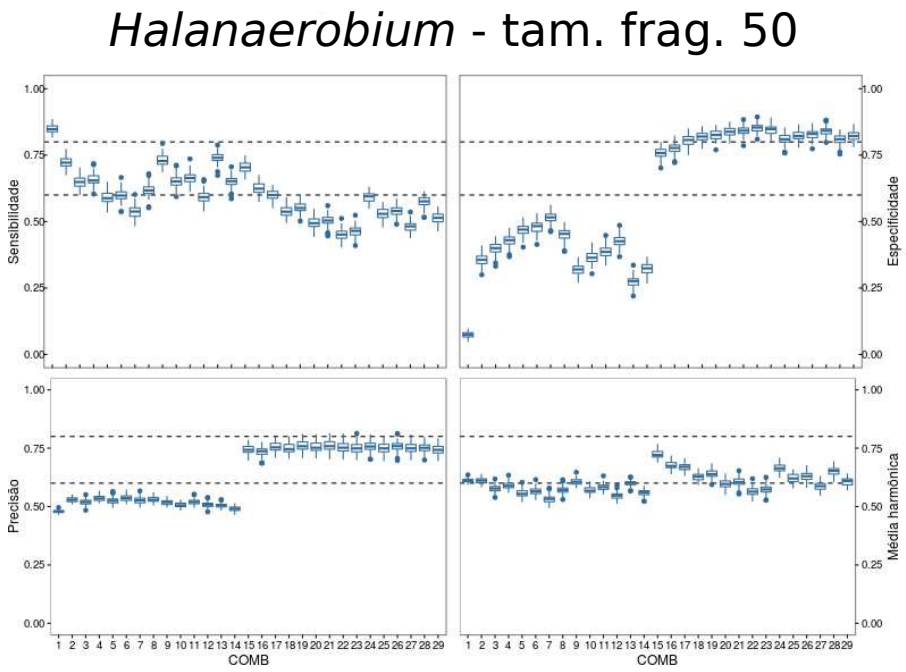

Figura 4.6: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para o gênero *Halanaerobium*. A medida GC foi a única que conseguiu identificação acima de 70 % para sensibilidade, precisão e especificidade.

Com tamanhos maiores de *reads* esse percentual também aumenta, para *reads* de tamanho 600 pb. A combinação 15-GC % consegue identificar 63,36 % dos organismos (Tabela 4.3). As tecnologias de sequenciamento de nova geração geram em média tamanhos de *reads* entre 200-400 pb. A combinação 15-GC % mostrou-se satisfatória para essa faixa de tamanhos de *reads* com identificação média de 115 organismos (56,93 %)

Tabela 4.3: Tabela de gêneros identificados pela metodologia para *reads* de tamanho 50-600 no táxon gênero.

ID	Medida	50	100	150	200	250	300	350	400	450	500	550	600
1	din	-	1	2	3	3	6	9	9	12	12	13	13
2	S2	-	-	2	2	6	7	9	13	13	14	18	22
3	S2 ∩ din	-	-	-	-	1	4	5	5	6	8	9	12
4	S2 ∩ S3	-	-	-	2	1	2	4	4	4	9	9	12
5	S2 ∩ S3 ∩ din	-	-	-	-	1	-	1	1	-	-	3	2
6	S2 ∩ S3 ∩ S4	-	-	-	-	-	-	1	-	-	-	2	1
7	S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	1	1	-
8	S2 ∩ S4	-	-		-	2	1	1	-	2	1	5	3
9	S3	-	-	4	4	10	13	12	18	19	20	28	31
10	S3 ∩ din	-	-	-	-	-	-	1	1	3	2	5	3
11	S3 ∩ S4	-	-	-	-	1	1	3	4	3	4	6	8
12	S3 ∩ S4 ∩ din	-	-		-	-	1	1	1	-	1	1	-
13	S4	-	-	2	5	7	12	12	16	18	21	25	23
14	S4 ∩ din	-	-	-	-	-	-	-	-	-	1	3	1
15	GC %	23	63	88	103	112	120	121	115	121	125	132	128
16	GC % ∩ din	-	-	2	2	5	4	4	4	7	8	10	8
17	GC % ∩ S2	-	-	-	2	4	4	7	4	6	10	9	8
18	GC % ∩ S2 ∩ din	-	-	-	-	1	3	3	1	1	2	3	2
19	GC % ∩ S2 ∩ S3	-	-	-	-	-	1	2	2	3	3	3	4
20	GC % ∩ S2 ∩ S3 ∩ din	-	-	-	-	-	-	1	-	1	1	1	1
21	GC % ∩ S2 ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	1	1	2	-
22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
23	GC % ∩ S2 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
24	GC % ∩ S3	-	-	-	-	-	1	2	3	4	3	5	5
25	GC % ∩ S3 ∩ din	-	-	-	-	-	-	1	1	1	1	1	1
26	GC % ∩ S3 ∩ S4	-	-	-	-	-	-	1	-	1	1	2	2
27	GC % ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	1	-	-	-	-	-
28	GC % ∩ S4	-	-	-	-	1	-	2	1	1	2	3	2
29	GC % ∩ S4 ∩ din	-	-	-	-	-	-	1	-	-	-	-	-

Fonte: O autor (2016)

Com os resultado obtidos, foi observado que a medida 15-GC % pode ser utilizada para identificação de microorganismos em nível de gênero. Estudos indicam que a concentração de GC % pode variar de aproximadamente 16,6 % (*Carsonella ruddii strain Pv*) até 74,9 % (*Anaeromyxobacter dehalogenans Strain 2CP-C*) (BOHLIN et al., 2010). Pesquisadores (BOHLIN et al., 2010; MUTO; OSAWA, 1987) apresentaram estudos sobre a análise do concentração de GC % para classificação de bactérias no táxon filo. A nossa análise indica que é possível identificar gêneros de microorganismos através da concentração de GC %.

Alguns microorganismos não foram identificados pela medida 15-GC % mantendo a taxa de acerto em 70 %. Como exemplo, temos o gênero *Anaeromyxobacter* onde a medida 15-GC % encontra-se no valor mínimo de 70 % para sensibilidade somente no *read* de tamanho 200 pb (Figura 4.7), enquanto que a medida 13-S4 fica acima do valor para *reads* de 150-550 pb (Figura 4.8) (Tabela 4.4). Com esse resultado a medida 13-S4 obteve melhores resultados para tamanhos de fragmentos mais comuns entre as tecnologias de sequenciamento de nova geração.

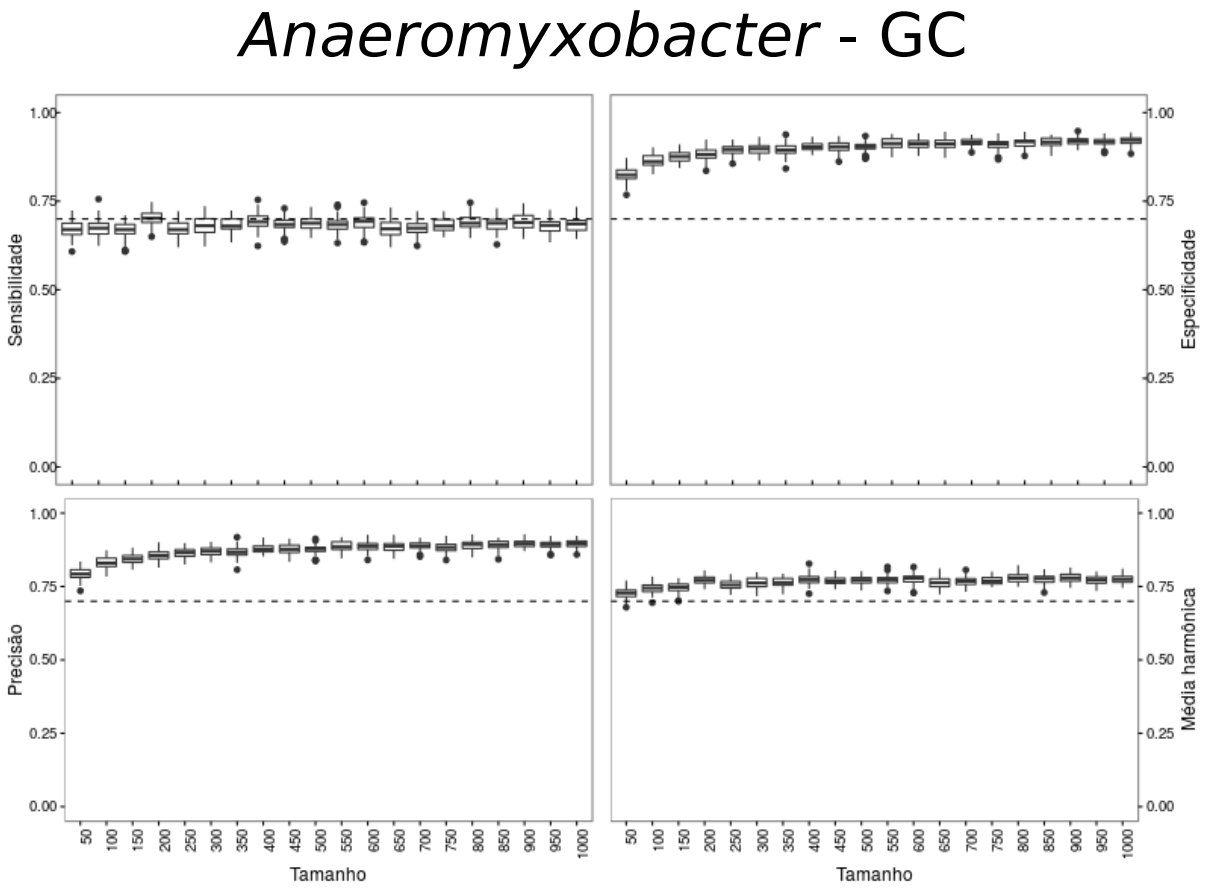

Figura 4.7: Identificação do gênero *Anaeromyxobacter* por tamanho de *read* para a medida 15-GC %. Apesar da sensibilidade não aumentar em decorrência do aumento do tamanho do fragmento, a precisão e a especificidade melhoram atingindo um platô em fragmentos de tamanho 250 pb.

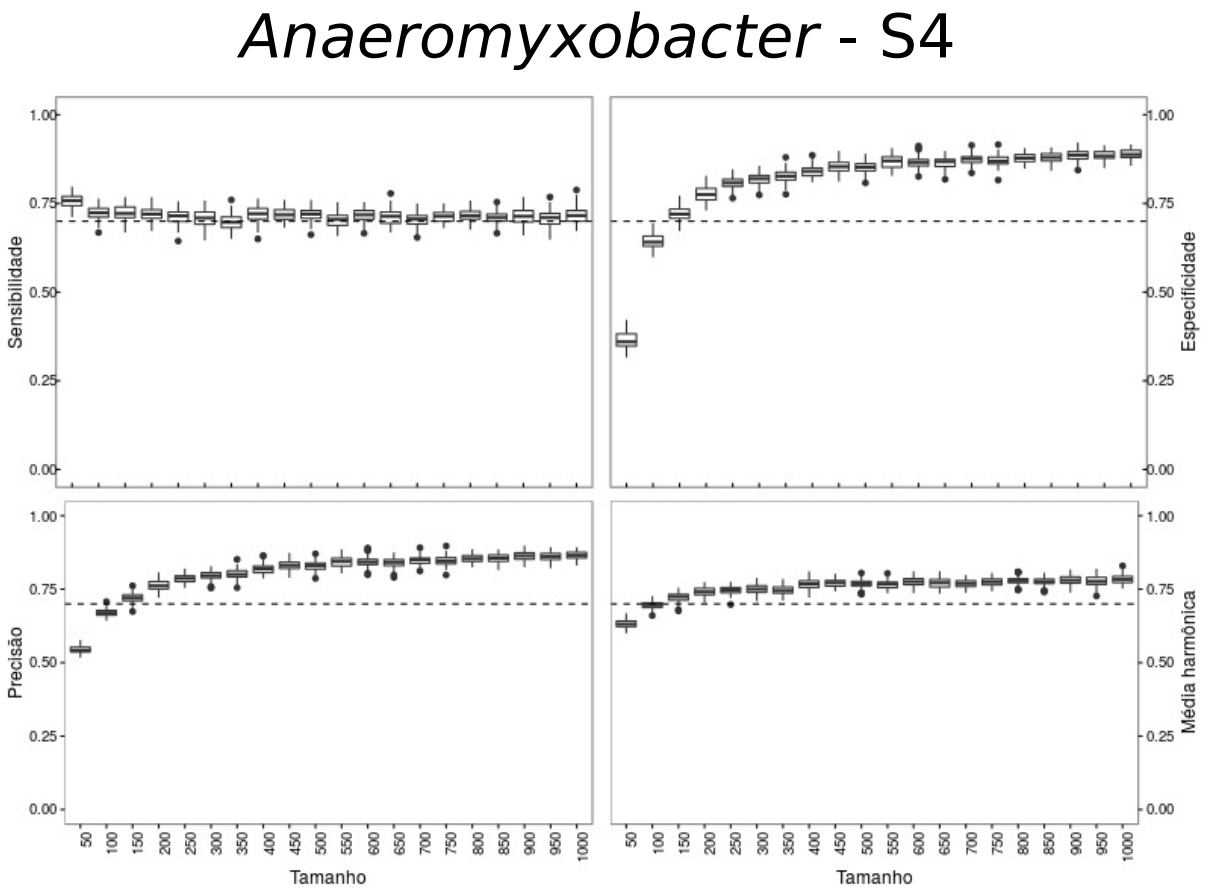

Figura 4.8: Identificação do gênero *Anaeromyxobacter* por tamanho de *read* para a medida 13-S4. Para o tamanho de fragmento 50 pb sensibilidade conseguiu identificar acima de 75 % dos gêneros, contudo a precisão e especificidade tiveram valores de identificação abaixo dos 60 % para esse tamanho de fragmento. A partir dos tamanhos de fragmentos 150 pb a identificação ficou acima de 70 % para a sensibilidade, precisão e especificidade.

Tabela 4.4: Tabela de medidas que identificaram organismos do gênero *Anaeromyxobacter* em relação ao tamanho do *read*.

Anaeromyxobacter	150	200	250	300	350	400	450	500	550	600
1-din							*	*	*	*
9-S3	*	*	*	*	*	*	*		*	*
13-S4	*	*	*	*	*	*	*	*	*	
15-GC %		*								

Fonte: O autor (2016)

O mesmo ocorreu com a identificação do organismo *Sorangium*, onde a medida 15-GC % consegue estar dentro do valor mínimo de 70 % somente em *reads* de tamanho 200 e 500 pb (Figura 4.9), enquanto que a medida 9-S3 fica acima do valor de 70 % para todos os *reads* com exceção de 50, 100, 150, 200 e 300 pb (Figura 4.10) (Tabela 4.5).

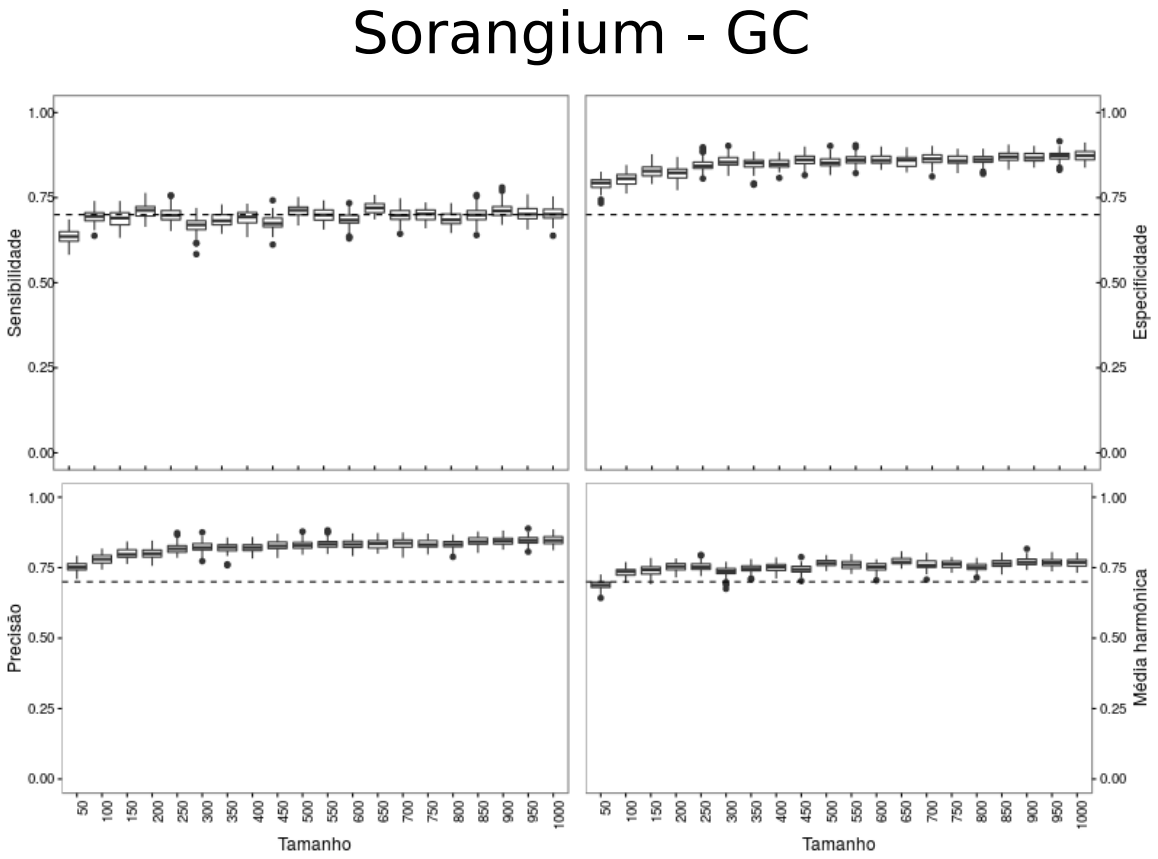

Figura 4.9: Identificação do gênero *Sorangium* por tamanho de *read* para a medida 15-GC %. Apesar dos resultados próximos, somente os tamanhos de fragmentos 200 pb e 500 pb obtiveram acima de 70 % de identificação para sensibilidade.

Tabela 4.5: Tabela de medidas que identificaram organismos do gênero *Sorangium* em relação ao tamanho do *read*.

Sorangium	150	200	250	300	350	400	450	500	550	600
9-S3			*		*	*	*	*	*	*
15-GC %		*						*		

Fonte: O autor (2016)

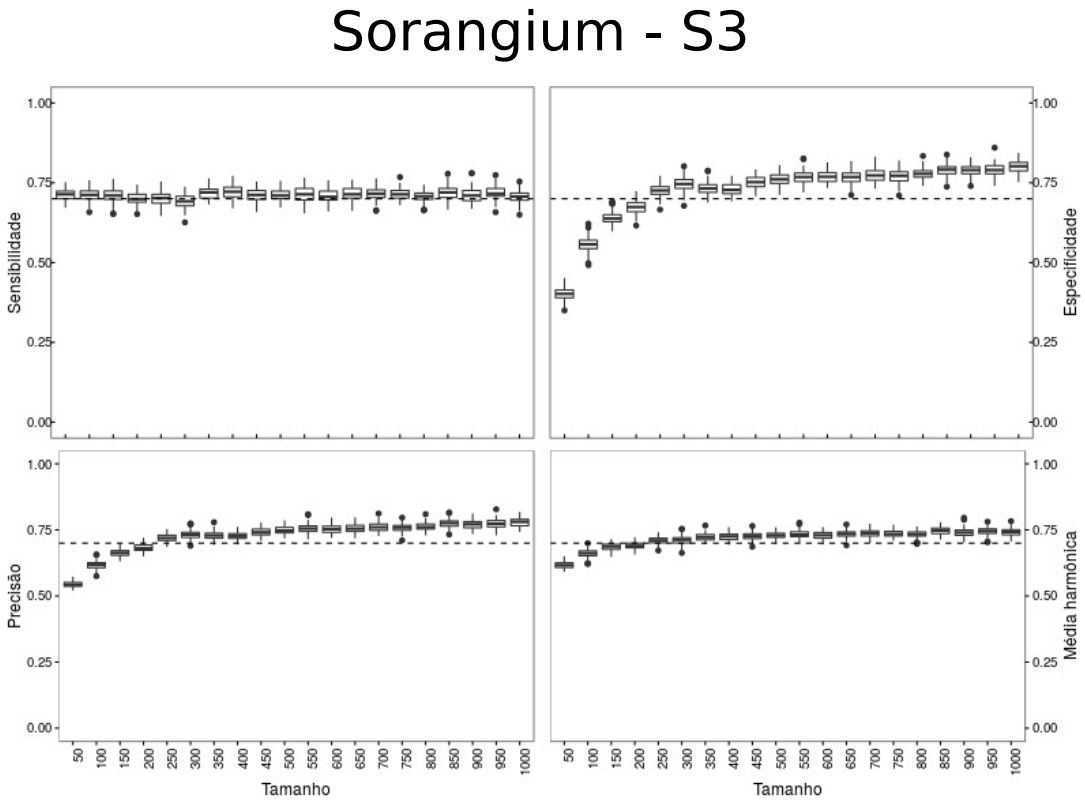

Figura 4.10: Identificação do gênero *Sorangium* por tamanho de *read* para a medida 9-S3. Essa medida conseguiu identificar com sensibilidade acima de 70 % os fragmentos de tamanho 250 pb e fragmentos acima do valor 350 pb.

4.4 Aplicação da metodologia no táxon família

Analisamos as 29 possíveis combinações para os 147 conjuntos de organismos do táxon família com *reads* entre 50-1000 pb, obtivemos 2706 resultados de identificações acima do valor definido de 70 % e identificamos 100 famílias (Tabela 4.6). Assim como nos resultados do táxon gênero a combinação que obteve maior êxito na identificação de família dos organismos foi a 15-GC % com um percentual de 44,16 % de todas as identificações e conseguiu identificar 95 famílias. O percentual de identificações realizadas pela medida 15-GC %, para o tamanho de fragmento 50 pb foi de 15,64 % do total de organismos enquanto que para tamanhos maiores como 600 pb o percentual de identificação atinge 42,17 % (Tabela 4.7), a medida 15-GC % obteve o ápice de identificação de famílias em fragmentos de tamanho 450 atingindo 66 famílias identificadas. Podemos verificar esse resultado nos *reads* de tamanho 50 pb nas famílias *Nocardiaceae* (Fig 4.11) e *Chlamydiaceae* (Fig 4.12). Nos dois exemplos a combinação 1-din possui os melhores resultados para sensibilidade, atingindo valores acima de 80 %, e assim sendo considerada uma importante medida para verificação de verdadeiros positivos. As medidas 16

a 29 também obtiveram resultados melhores que a 15-GC % para a especificidade e precisão podendo assim identificar melhor os verdadeiros negativos e falsos positivos.

Tabela 4.6: Tabela de combinações e organismos identificados no táxon família para valores de especificidade, sensibilidade, precisão e média harmônica acima de 70 %.

ID	Medida	Fragmentos Identificados	ID	Medida	Fragmentos Identificados
1	din	182	16	GC % ∩ din	33
2	S2	186	17	GC % ∩ S2	66
3	S2 ∩ din	97	18	GC % ∩ S2 ∩ din	12
4	S2 ∩ S3	76	19	GC % ∩ S2 ∩ S3	26
5	S2 ∩ S3 ∩ din	11	20	GC % ∩ S2 ∩ S3 ∩ din	4
6	S2 ∩ S3 ∩ S4	16	21	GC % ∩ S2 ∩ S3 ∩ S4	13
7	S2 ∩ S3 ∩ S4 ∩ din	1	22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	3
8	S2 ∩ S4	33	23	GC % ∩ S2 ∩ S4 ∩ din	4
9	S3	253	24	GC % ∩ S3	34
10	S3 ∩ din	31	25	GC % ∩ S3 ∩ din	6
11	S3 ∩ S4	76	26	GC % ∩ S3 ∩ S4	23
12	S3 ∩ S4 ∩ din	5	27	GC % ∩ S3 ∩ S4 ∩ din	4
13	S4	253	28	GC % ∩ S4	31
14	S4 ∩ din	24	29	GC % ∩ S4 ∩ din	8
15	GC %	1195			

Fonte: O autor (2016)

Tabela 4.7: Tabela de organismos identificados pela metodologia para *reads* de tamanho 50-600 pb no táxon família.

ID	Medida	50	100	150	200	250	300	350	400	450	500	550	600
1	din	-	1	3	4	4	6	10	10	12	12	12	12
2	S2	-	1	-	1	5	2	4	7	9	9	10	11
3	S2 ∩ din	-	-	-	-	1	2	2	2	6	5	5	6
4	S2 ∩ S3	-	-	-	-	-	1	1	1	-	2	2	3
5	S2 ∩ S3 ∩ din	-	-	-	-	1	1	1	-	-	-	1	-
6	S2 ∩ S3 ∩ S4	-	-	-	-	-	-	1	-	-	-	-	-
7	S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	1	-	-	-	-	-
8	S2 ∩ S4	-	-		-	1	1	1	-	-	-	3	-
9	S3	-	-	1	2	4	8	6	11	12	13	13	14
10	S3 ∩ din	-	-	-	-	-	1	-	-	-	1	2	1
11	S3 ∩ S4	-	-	-	-	-	-	1	1	1	1	2	4
12	S3 ∩ S4 ∩ din	-	-		-	1	-	1	-	-	-	-	-
13	S4	-	1	2	3	3	7	8	11	12	14	15	13
14	S4 ∩ din	-	-	-	-	-	1	-	-	-	1	2	-
15	GC %	16	33	48	57	50	60	61	60	66	62	64	62
16	GC % ∩ din	-	-	-	-	3	-	1	-	3	2	3	3
17	GC % ∩ S2	-	-	-	-	-	-	2	-	7	2	3	3
18	GC % ∩ S2 ∩ din	-	-	-	-	-	-	1	-	-	-	-	-
19	GC % ∩ S2 ∩ S3	-	-	-	-	-	-	-	-	1	2	2	2
20	GC % ∩ S2 ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
21	GC % ∩ S2 ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	1	1	1	-
22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
23	GC % ∩ S2 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
24	GC % ∩ S3	-	-	-	-	-	-	1	1	1	2	2	2
25	GC % ∩ S3 ∩ din	-	-	-	-	-	-	1	-	-	-	-	-
26	GC % ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	2	2	2
27	GC % ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
28	GC % ∩ S4	-	-	-	-	-	-	2	1	1	2	2	2
29	GC % ∩ S4 ∩ din	-	-	-	-	-	-	1	-	-	-	-	-

Fonte: O autor (2016)

Com os resultado obtidos na análise do táxon família, foi constatado que a medida 15-GC % também pode ser utilizada para identificação de microorganismos em nível de família.

Assim como verificamos para gênero, algumas famílias não foram identificadas pela medida 15-GC %. Como exemplo, para família temos a *Microbacteriaceae* onde a medida 15-GC % (Fig 4.13) não conseguiu identificar os *reads* acima de 70 % de sensibilidade para tamanhos de 150 pb a 400 pb, enquanto a medida 1- din (Fig 4.14) obteve êxito na identificação para tamanhos de 250 pb a 600 pb (Tabela 4.8).

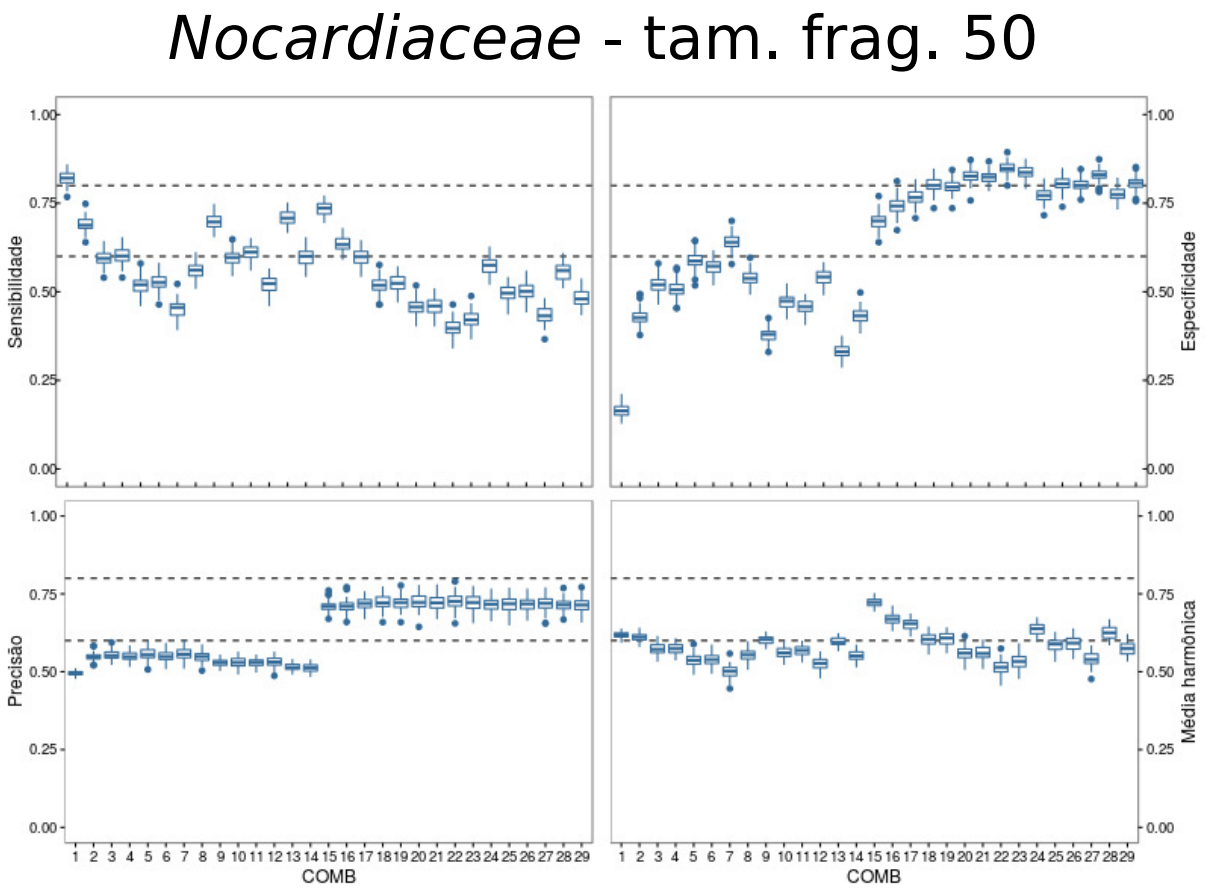

Figura 4.11: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para o família *Nocardiaceae*. A medida GC foi a única a conseguir identificar os fragmentos em táxon família acima de 70 %.

Tabela 4.8: Tabela de medidas que identificaram organismos da família *Microbacteriaceae* em relação ao tamanho do *read*.

Microbacteriaceae	150	200	250	300	350	400	450	500	550	600
1-din			*	*	*	*	*	*	*	*
15-GC %							*	*	*	*

Fonte: O autor (2016)

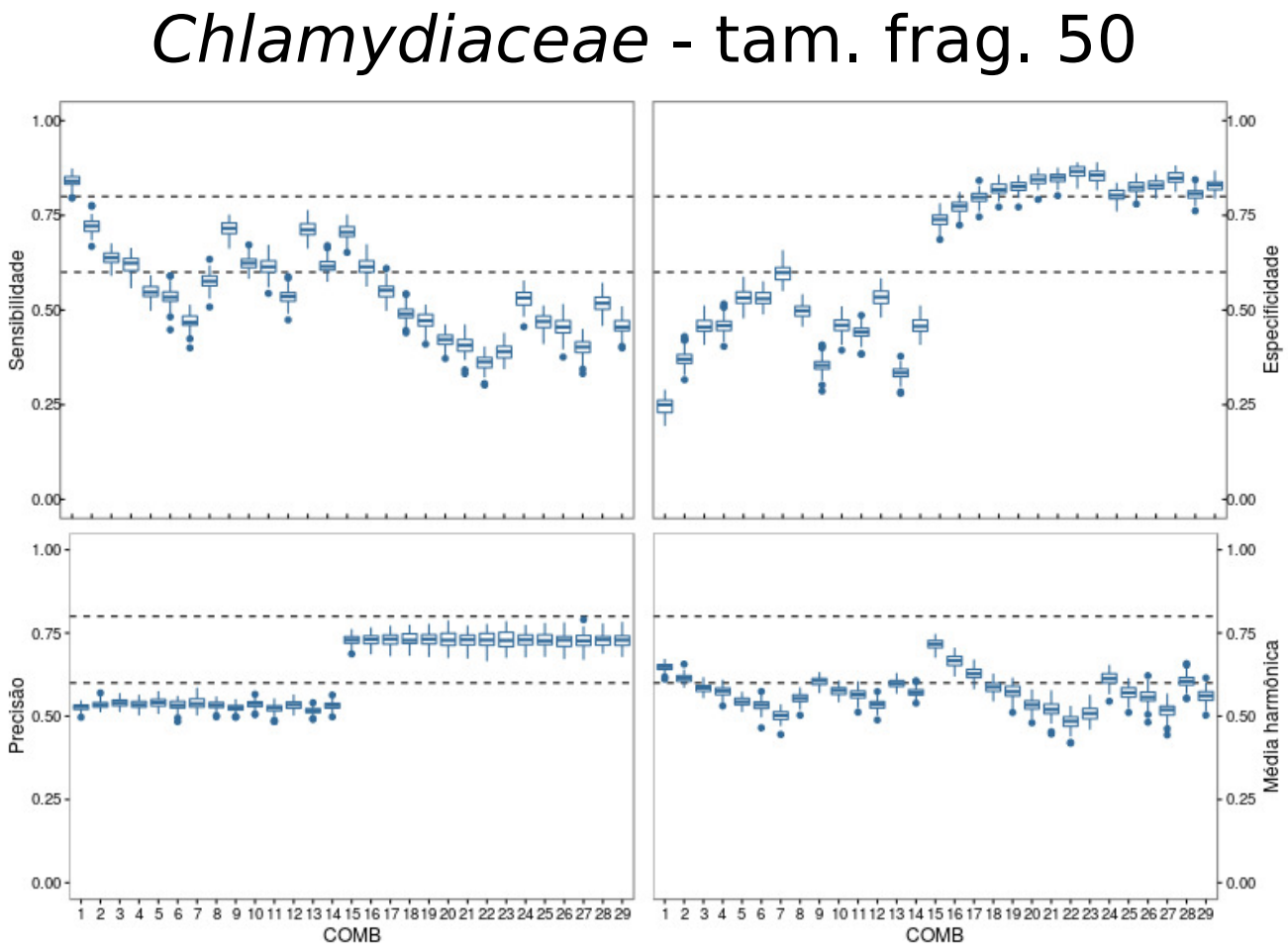

Figura 4.12: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a família *Chlamydiaceae*. Para a família *Chlamydiaceae* a única medida que conseguiu atingir o corte de 70 % de identificação para sensibilidade, precisão e especificidade foi a medida 15-GC %. Medidas como 1-din conseguiu valores acima de 80 % de identificação para sensibilidade, mas obteve valores baixo para precisão e especificidade.

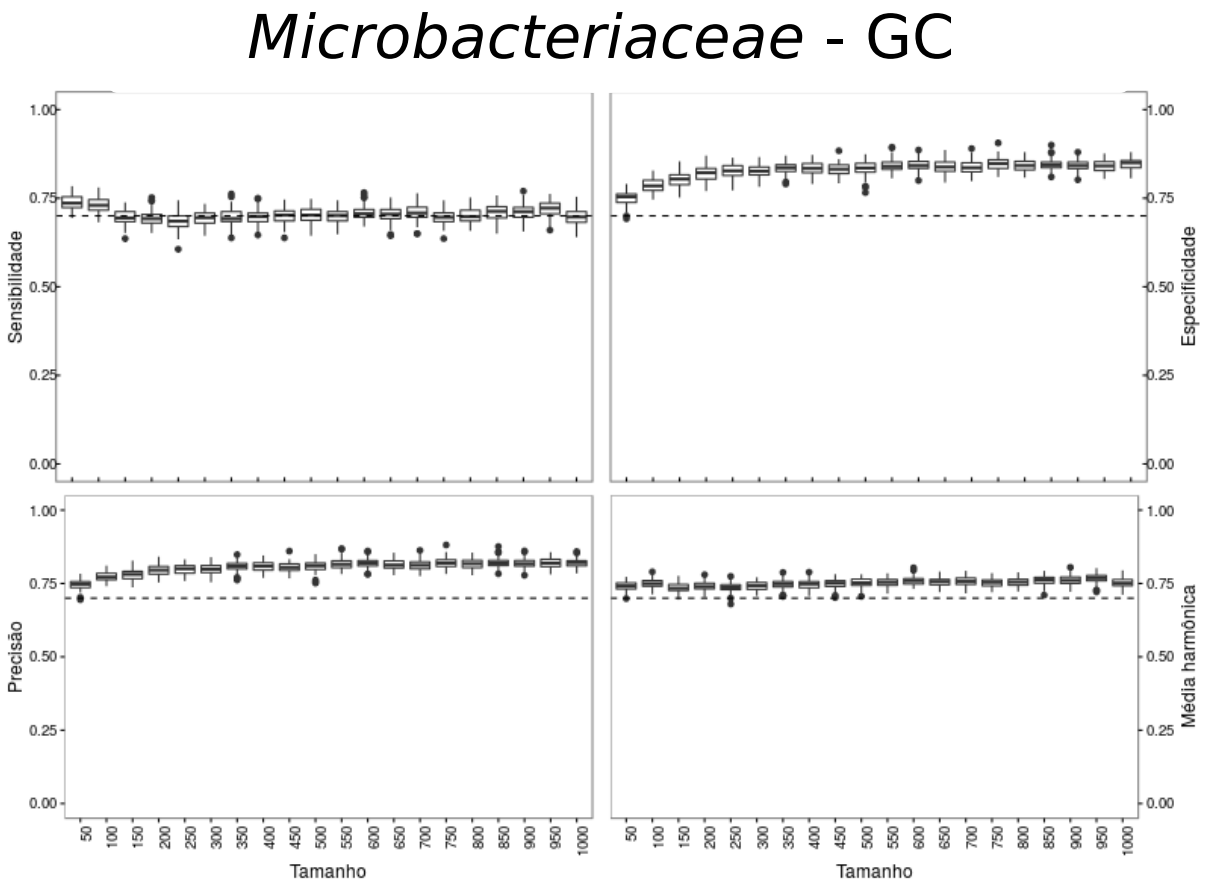

Figura 4.13: Identificação da família *Microbacteriaceae* por tamanho de *read* para a medida 15-GC %. A partir dos fragmentos 100 pb ocorre uma diminuição nos valores de sensibilidade enquanto a precisão e especificidade aumentam.

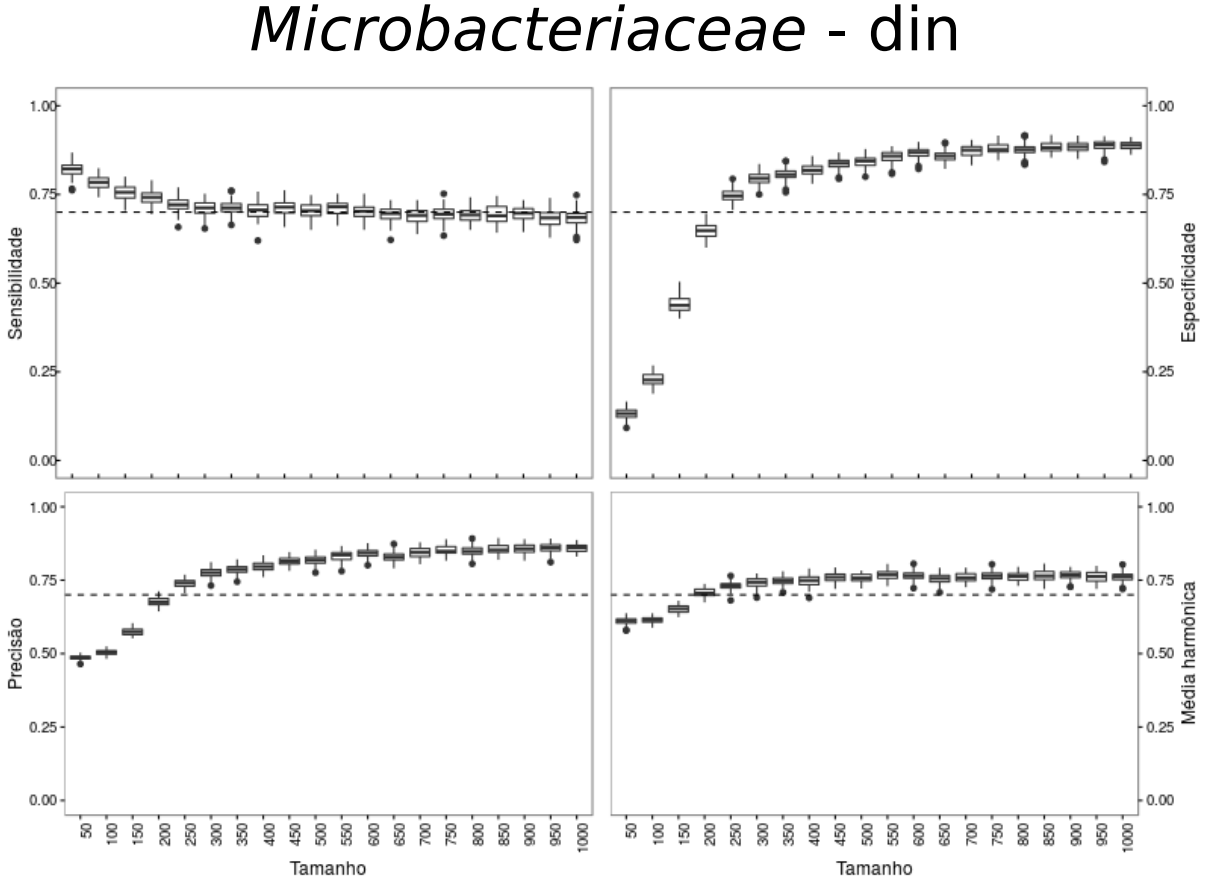

Figura 4.14: Identificação da família *Microbacteriaceae* por tamanho de *read* para a medida 1-din. Enquanto temos uma diminuição na sensibilidade conforme aumenta o tamanho do fragmento, as medidas de precisão e especificidade aumentam. Para a família *Microbacteriaceae* o tamanho de fragmento 250 pb obteve os melhores resultados para as medidas estatísticas.

4.5 Aplicação da metodologia no táxon ordem

De 29 possíveis combinações para os 87 conjuntos de organismos do táxon ordem com *reads* entre 50-1000 pb, obtivemos 911 resultados positivo acima do valor definido de 70 % e encontramos 47 organismos pertencentes a esse táxon (Tabela 4.9). No táxon ordem a combinação que obteve maior êxito foi a 15-GC % com um percentual de 47,96 % o que corresponde a 30 organismos no táxon ordem (Tabela 4.10).

Tabela 4.9: Tabela de combinações para o táxon ordem com valores de especificidade, sensibilidade, precisão e média harmônica com valores iguais ou acima de 70 %.

ID	Medida	Fragmentos Identificados	ID	Medida	Fragmentos Identificados
1	din	92	16	GC % ∩ din	5
2	S2	73	17	GC % ∩ S2	6
3	S2 ∩ din	30	18	GC % ∩ S2 ∩ din	1
4	S2 ∩ S3	28	19	GC % ∩ S2 ∩ S3	0
5	S2 ∩ S3 ∩ din	3	20	GC % ∩ S2 ∩ S3 ∩ din	0
6	S2 ∩ S3 ∩ S4	1	21	GC % ∩ S2 ∩ S3 ∩ S4	0
7	S2 ∩ S3 ∩ S4 ∩ din	0	22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	0
8	S2 ∩ S4	10	23	GC % ∩ S2 ∩ S4 ∩ din	0
9	S3	93	24	GC % ∩ S3	1
10	S3 ∩ din	13	25	GC % ∩ S3 ∩ din	1
11	S3 ∩ S4	18	26	GC % ∩ S3 ∩ S4	0
12	S3 ∩ S4 ∩ din	0	27	GC % ∩ S3 ∩ S4 ∩ din	0
13	S4	87	28	GC % ∩ S4	1
14	S4 ∩ din	9	29	GC % ∩ S4 ∩ din	1
15	GC %	437			

Fonte: O autor (2016)

No táxon ordem a medida 15-GC % também foi a única combinação que conseguiu realizar as identificações em tamanhos de fragmento de 50 pb, ela identificou 9.66 % do total de organismos desse tamanho. Para o tamanho de fragmento 600 pb esse percentual aumenta para 26,43 %. Como exemplo para os fragmentos de tamanho 50 pb temos as ordens *Brachyspirales* (Fig 4.15) e *Chlamydiales* (Fig 4.16). Assim como aconteceu nos táxons gênero e família, a medida 1-din possui os melhores resultados para sensibilidade, atingindo valores acima de 80 %. As combinações 17 a 29 também obtiveram resultados melhores que a 15-GC % para a especificidade e precisão.

Tabela 4.10: Tabela de organismos identificados pela metodologia para *reads* de tamanho 50-600 pb no táxon ordem.

ID	Medida	50	100	150	200	250	300	350	400	450	500	550	600
1	din	-	-	1	1	1	2	5	5	5	7	7	6
2	S2	-	1	-	-	1	2	1	2	2	4	3	2
3	S2 ∩ din	-	-	-	-	1	1	1	1	3	2	2	1
4	S2 ∩ S3	-	-	-	-	-	-	-	-	1	1	1	1
5	S2 ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
6	S2 ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
7	S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
8	S2 ∩ S4	-	-		-	-	-	-	-	-	-	1	-
9	S3	-	-	-	-	1	3	3	3	4	5	6	6
10	S3 ∩ din	-	-	-	-	-	-	-	-	-	1	1	1
11	S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	1	1
12	S3 ∩ S4 ∩ din	-	-		-	-	-	-	-	-	-	-	-
13	S4	-	-	-	1	-	2	3	3	4	5	6	5
14	S4 ∩ din	-	-	-	-	-	1	-	-	-	-	1	1
15	GC %	9	12	18	22	17	21	21	22	20	22	22	23
16	GC % ∩ din	-	-	-	-	1	-	-	-	-	-	-	-
17	GC % ∩ S2	-	-	-	-	-	-	-	-	-	-	-	-
18	GC % ∩ S2 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
19	GC % ∩ S2 ∩ S3	-	-	-	-	-	-	-	-	-	-	-	-
20	GC % ∩ S2 ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
21	GC % ∩ S2 ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
23	GC % ∩ S2 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
24	GC % ∩ S3	-	-	-	-	-	-	-	-	-	-	-	-
25	GC % ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
26	GC % ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
27	GC % ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
28	GC % ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
29	GC % ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-

Fonte: O autor (2016)

Com os resultado obtidos na análise do táxon ordem, foi constatado que a medida 15-GC % também pode ser utilizada para identificação de microorganismos em nível de ordem.

Algumas combinações conseguiram identificar organismos com melhor eficiência do que a medida 15-GC % em determinados tamanhos de fragmento. Como exemplo, para ordem temos a *Frankiales* onde a medida 15-GC % (Fig 4.17) conseguiu apenas identificar os *reads* acima de 70 % de sensibilidade para o tamanho 50 pb, enquanto a medida 1- din (Fig 4.18) obteve êxito na identificação para tamanhos maiores que 450 pb (Tabela 4.11).

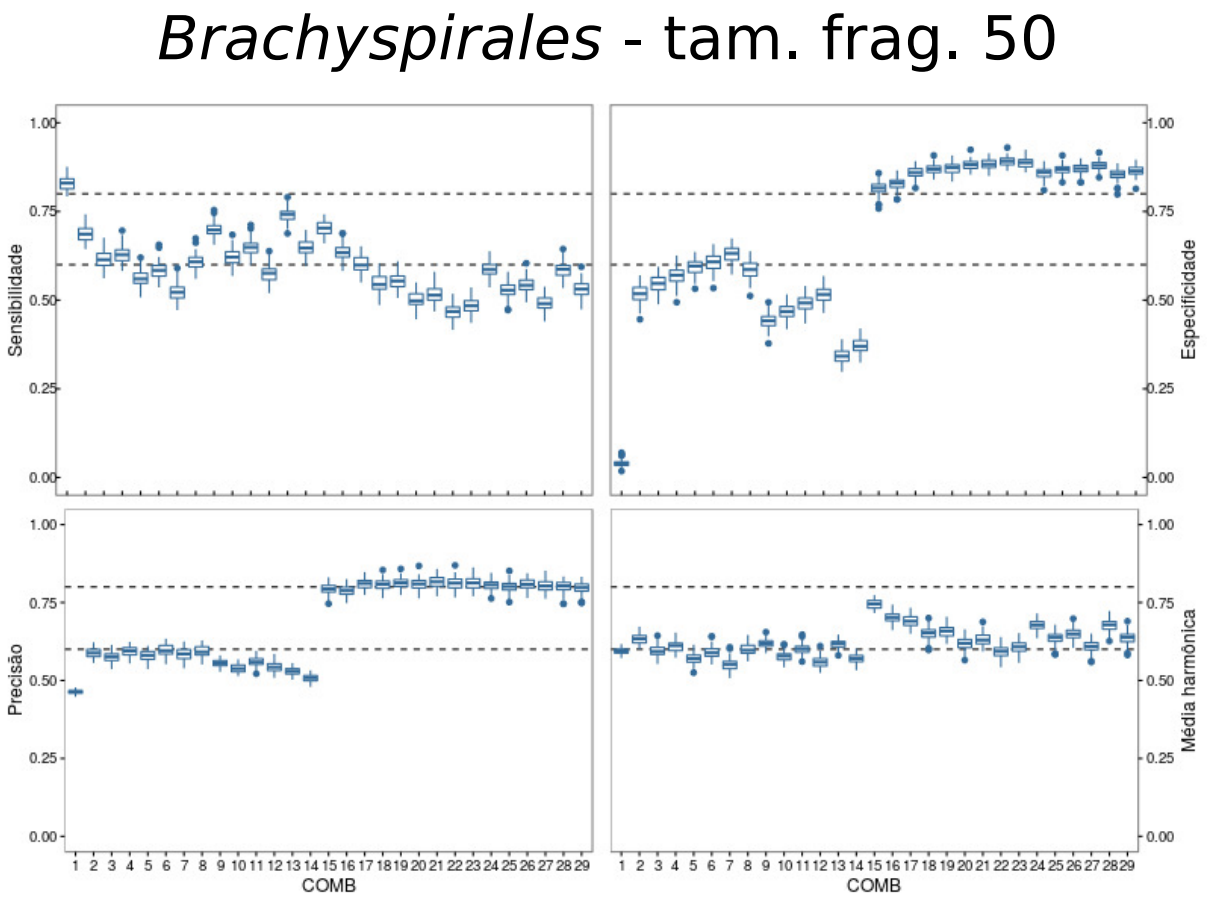

Figura 4.15: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a ordem *Brachyspirales*. Somente as medidas 1-din e 15-GC % conseguiram valores acima de 70 % para sensibilidade sendo consideradas boas medidas para identificação de VP. As medidas de 15 à 29 conseguiram valores acima de 70 % para precisão e especificidade assim sendo consideradas boas medidas para identificação de VN.

Tabela 4.11: Tabela de medidas que identificaram organismos da ordem *Frankiales* em relação ao tamanho do *read*.

Frankiales	150	200	250	300	350	400	450	500	550	600
1-din							*	*	*	*
15-GC %	*									

Fonte: O autor (2016)

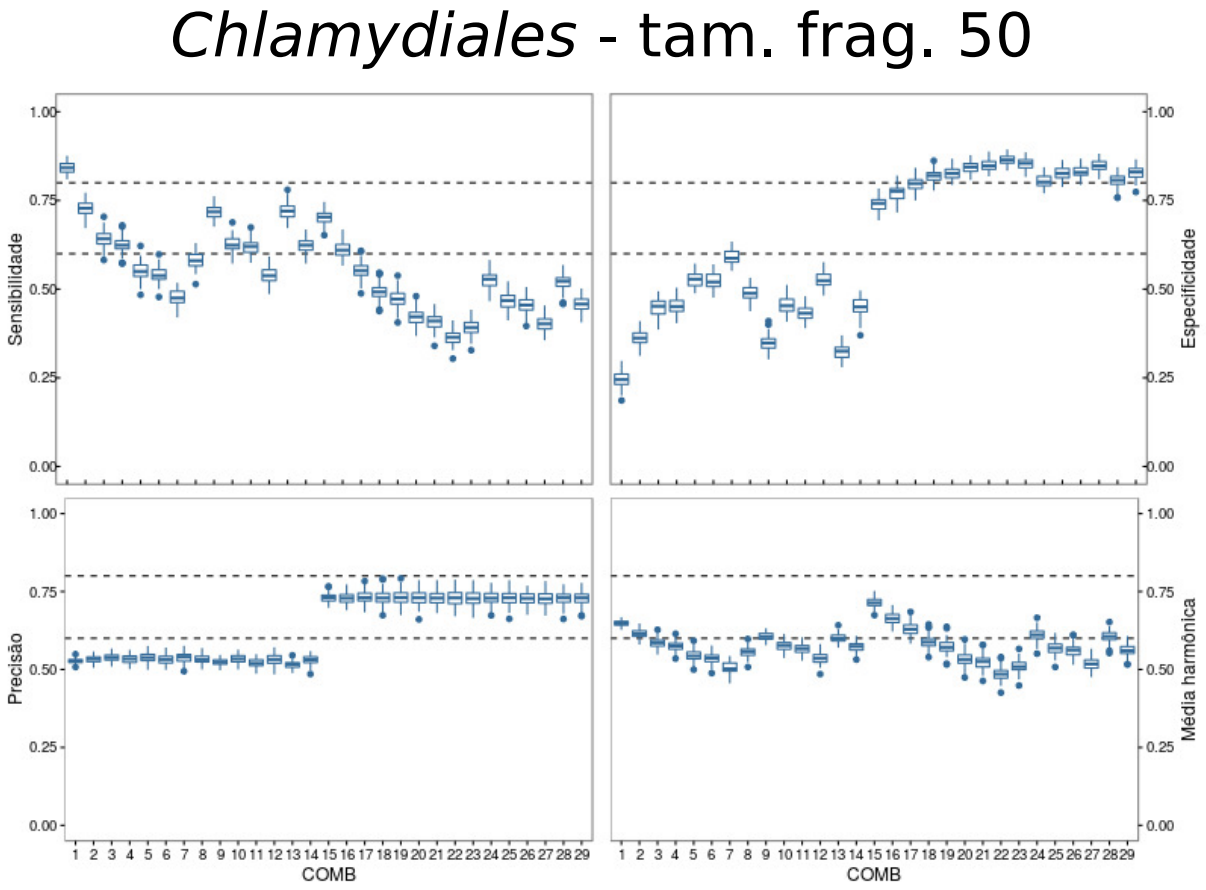

Figura 4.16: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a ordem *Chlamydiales*. Assim como para a ordem *Brachyspirales* as medidas 1-din e 15-GC % tiveram resultados acima de 70 % para identificação de VP enquanto as medidas de 15 à 29 obtiveram resultados acima de 70 % para identificação de VN.

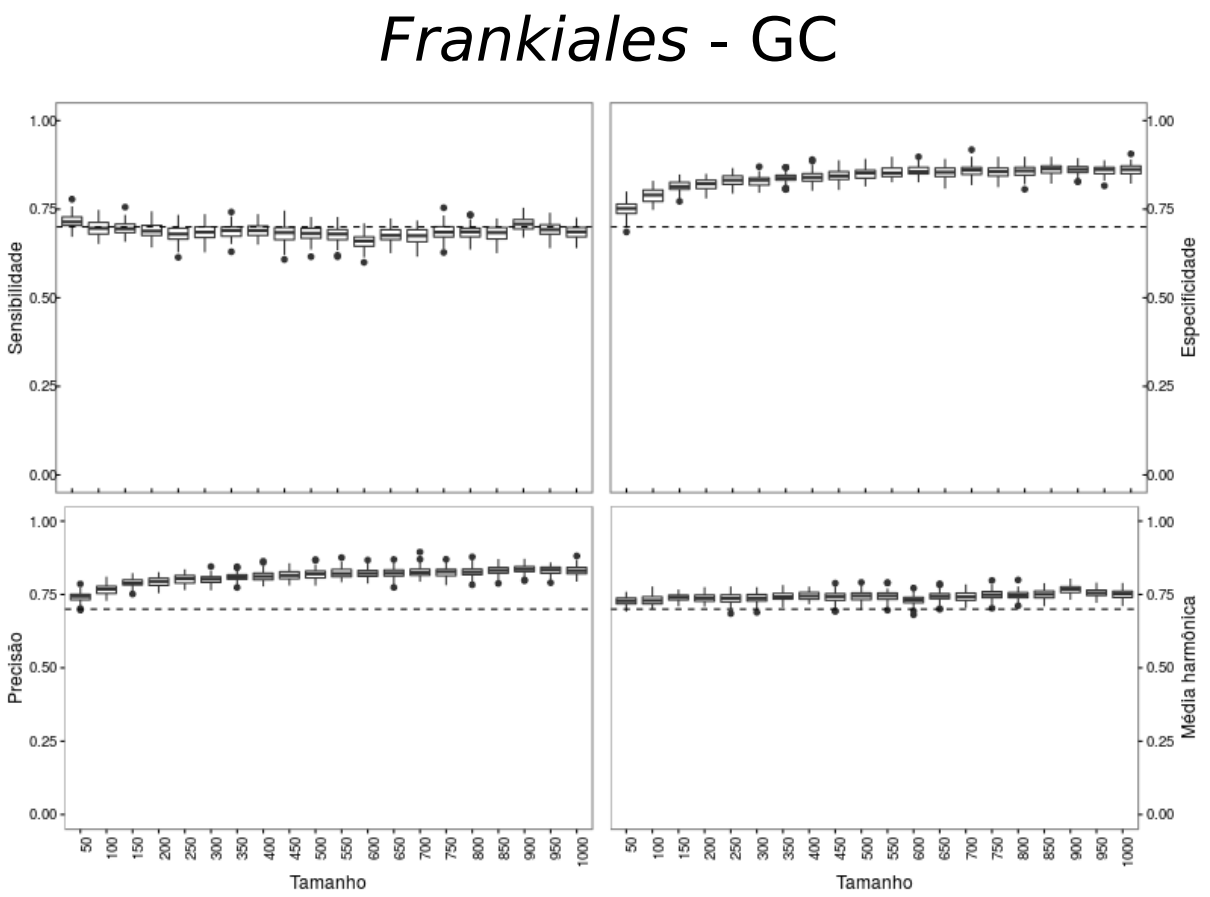

Figura 4.17: Identificação da ordem *Frankiales* por tamanho de *read* para a medida 15-GC %. As medidas estatísticas especificidade e precisão melhoram conforme o tamanho do fragmento aumenta enquanto a sensibilidade diminui com o aumento do fragmento.

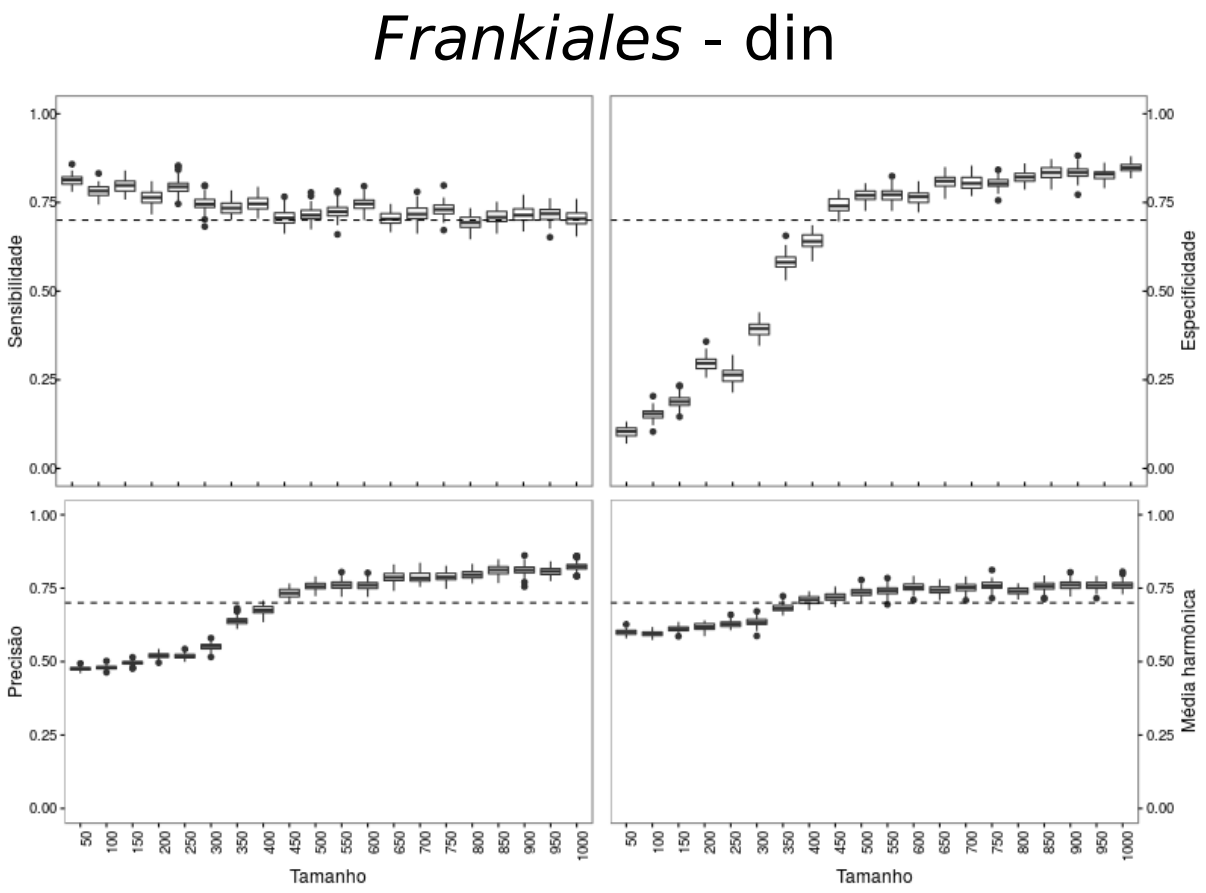

Figura 4.18: Identificação da ordem *Frankiales* por tamanho de *read* para a medida 1-din. As medidas estatísticas de precisão e especificidade começam com valores baixos, com fragmentos de tamanho 450 pb as duas medidas conseguem identificação acima de 70 %. A sensibilidade diminui com o aumento do fragmento porém somente o fragmento 800 pb fica com identificação abaixo de 70 %.

4.6 Aplicação da metodologia no táxon classe

Para o táxon classe obtivemos 293 resultados de identificações acima do valor definido de 70 % para as 29 possíveis combinações dos 35 conjuntos de organismos com *reads* entre 50-1000 pb e identificamos 11 organismos pertencentes a esse táxon (Tabela 4.12). No táxon classe a combinação que obteve maior exito na identificação dos organismos foi a 15-GC % com um percentual de 54,60 % de todas as identificações (Tabela 4.13).

Tabela 4.12: Tabela de combinações para o táxon classe com valores de especificidade, sensibilidade, precisão e média harmônica com valores iguais ou acima de 70 %.

ID	Medida	Fragmentos Identificados	ID	Medida	Fragmentos Identificados
1	din	0	16	GC % ∩ din	11
2	S2	21	17	GC % ∩ S2	4
3	S2 ∩ din	16	18	GC % ∩ S2 ∩ din	0
4	S2 ∩ S3	20	19	GC % ∩ S2 ∩ S3	0
5	S2 ∩ S3 ∩ din	3	20	GC % ∩ S2 ∩ S3 ∩ din	0
6	S2 ∩ S3 ∩ S4	1	21	GC % ∩ S2 ∩ S3 ∩ S4	0
7	S2 ∩ S3 ∩ S4 ∩ din	0	22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	0
8	S2 ∩ S4	9	23	GC % ∩ S2 ∩ S4 ∩ din	0
9	S3	10	24	GC % ∩ S3	0
10	S3 ∩ din	11	25	GC % ∩ S3 ∩ din	0
11	S3 ∩ S4	14	26	GC % ∩ S3 ∩ S4	0
12	S3 ∩ S4 ∩ din	1	27	GC % ∩ S3 ∩ S4 ∩ din	0
13	S4	2	28	GC % ∩ S4	0
14	S4 ∩ din	10	29	GC % ∩ S4 ∩ din	0
15	GC %	160			

Fonte: O autor (2016)

No táxon classe a medida 15-GC % foi a única combinação que conseguiu realizar as identificações em tamanhos de fragmento de 50 pb à 250 pb, no tamanho 50 pb ela identificou 8,57 % do total de organismos desse tamanho (Tabela 4.13). Para o tamanho de fragmento 600 pb esse percentual sobe para 22,85 %. Como exemplo para os fragmentos de tamanho 50 pb temos as classes *Chlamydiia* (Fig 4.19) e *Deinococci* (Fig 4.20). Assim como aconteceu nos táxons gênero, família e ordem, a medida 1-din possui os melhores resultados para sensibilidade, atingindo valores acima de 80 %. As combinações 17 a 29 também obtiveram resultados melhores que a 15-GC % para a especificidade e precisão.

Tabela 4.13: Tabela de organismos identificados pela metodologia para *reads* de tamanho 50-600 pb no táxon classe.

ID	Medida	50	100	150	200	250	300	350	400	450	500	550	600
1	din	-	-	-	-	-	-	-	-	-	-	-	-
2	S2	-	-	-	-	-	-	-	-	1	-	-	1
3	S2 ∩ din	-	-	-	-	1	1	1	1	1	1	1	1
4	S2 ∩ S3	-	-	-	-	-	-	1	1	-	1	1	1
5	S2 ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
6	S2 ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
7	S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
8	S2 ∩ S4	-	-		-	-	-	-	-	-	-	1	-
9	S3	-	-	-	-	-	-	-	-	-	-	-	-
10	S3 ∩ din	-	-	-	-	-	-	-	-	-	1	1	1
11	S3 ∩ S4	-	-	-	-	-	-	-	-	-	1	1	1
12	S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
13	S4	-	-	-	-	-	-	-	-	-	-	1	-
14	S4 ∩ din	-	-	-	-	-	-	-	-	-	-	1	1
15	GC %	3	6	6	7	5	7	7	8	9	11	8	8
16	GC % ∩ din	-	-	-	-	-	-	-	-	-	1	1	1
17	GC % ∩ S2	-	-	-	-	-	-	-	-	-	-	-	-
18	GC % ∩ S2 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
19	GC % ∩ S2 ∩ S3	-	-	-	-	-	-	-	-	-	-	-	-
20	GC % ∩ S2 ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
21	GC % ∩ S2 ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
23	GC % ∩ S2 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
24	GC % ∩ S3	-	-	-	-	-	-	-	-	-	-	-	-
25	GC % ∩ S3 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
26	GC % ∩ S3 ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
27	GC % ∩ S3 ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-
28	GC % ∩ S4	-	-	-	-	-	-	-	-	-	-	-	-
29	GC % ∩ S4 ∩ din	-	-	-	-	-	-	-	-	-	-	-	-

Fonte: O autor (2016)

Verificamos que a medida 15-GC % foi a melhor para identificação de microorganismos em nível de classe. Ao contrário do que ocorreu para os táxons gênero, família e ordem, no táxon classe não houveram combinações que conseguiram identificar organismos que a medida 15-GC % não tenha identificado.

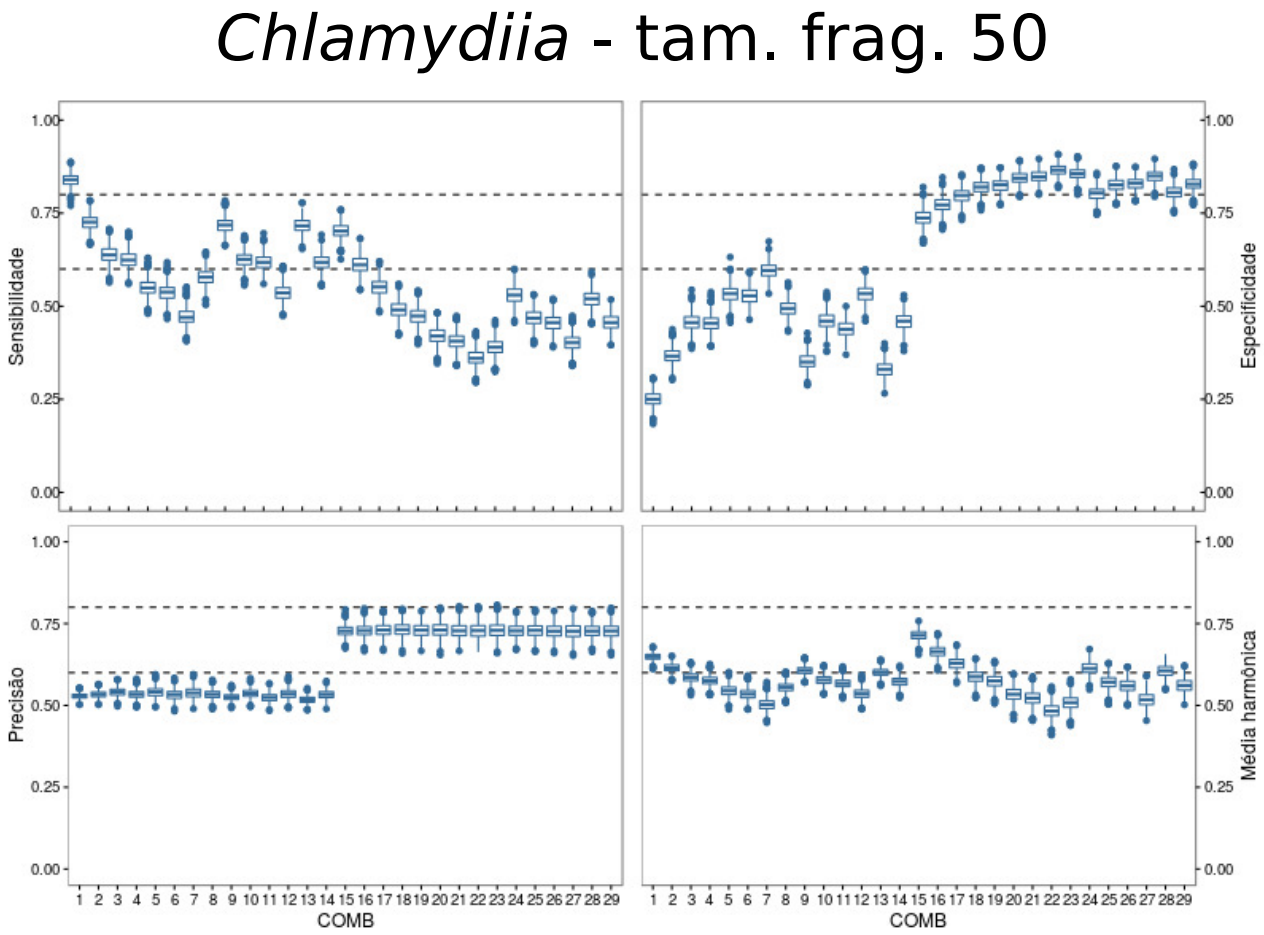

Figura 4.19: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a classe *Chlamydiia*. Para a classe *Chlamydiia* as medidas 1-din, 2-S2, 9-S3, 13-S4 e 15-GC % obtiveram resultados acima de 70 % para a identificação de VP. As medidas 15 à 29 tiveram resultados acima de 70 % para a identificação de VN.

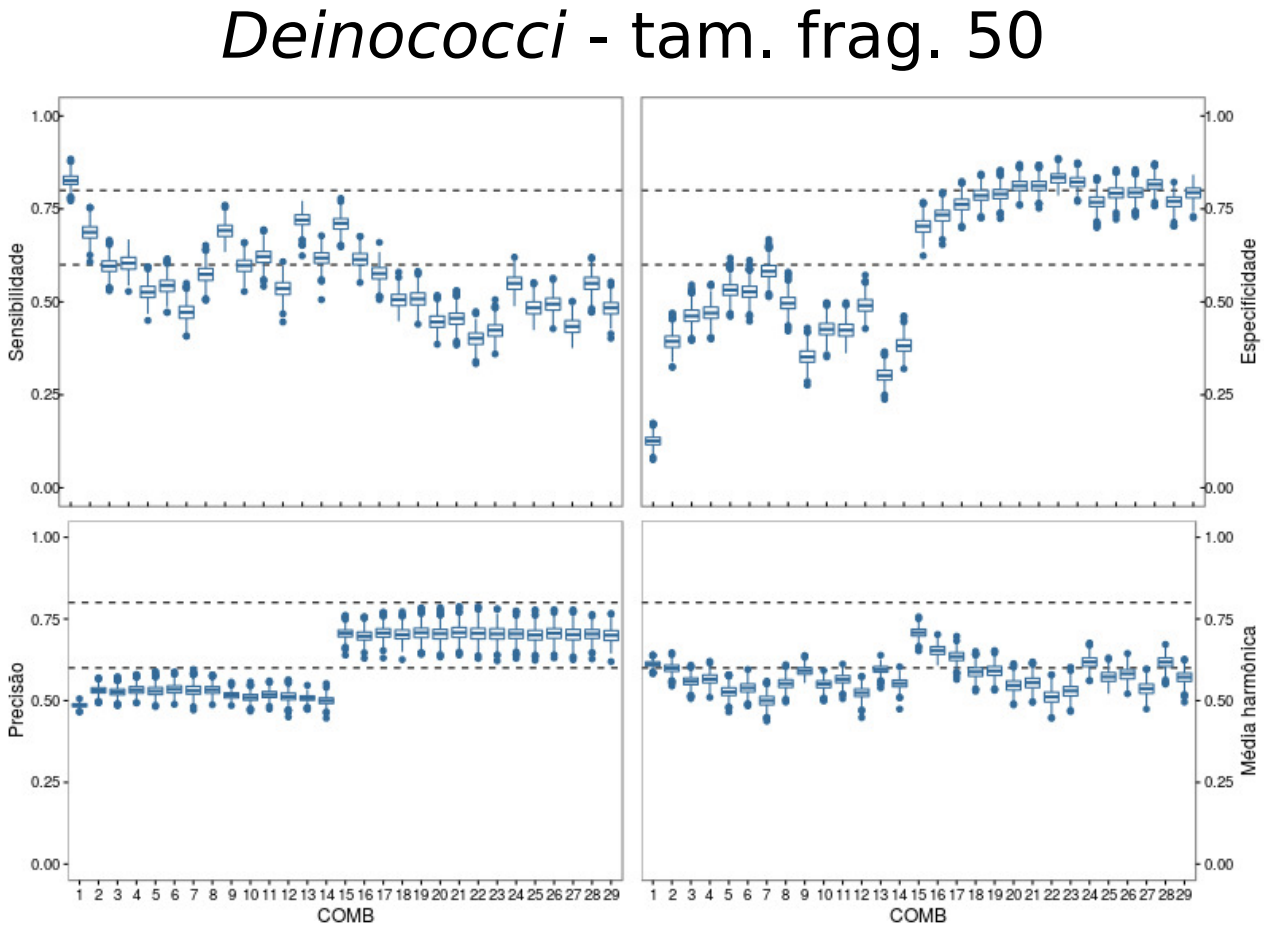

Figura 4.20: *Boxplots* das medidas estatísticas para o *read* de tamanho 50 pb em comparação com todas as 29 combinações para a classe *Deinococci*. Para a a classe *Deinococci* as medidas de 1 até 14 ficaram com um percentual de identificação abaixo dos 60 % em relação a precisão e especificidade. As medidas com valores de identificação acima de 70 % para sensibilidade foram: 1-din, 9-S3, 13-S4 e 15-GC, com isso a única medida que conseguiu identificação acima de 70 % para sensibilidade, precisão e especificidade foi a medida 15-GC %.

4.7 Comparação com programas de análise taxonômica.

Realizamos uma análise comparativa das amostras obtidas do projeto *"A human gut microbial gene catalog established by deep metagenomic sequencing"* com os programas Raiphy, Phymm, Phymmbl e a nossa metodologia. Como exemplo, aqui colocamos os resultados de 8 amostras: ERR011172, ERR011190, ERR011230, ERR011259, ERR011277, ERR011305, ERR011308 e ERR011323. As amostras ERR011190 e ERR011172 são de indivíduos da Dinamarca, do sexo masculino e de idades 54 e 44 anos respectivamente. As amostras ERR011259 e ERR011230 também são de indivíduos da Dinamarca, de sexo feminino e de idades 59 e 44 anos respectivamente. As outras quatro amostras são de indivíduos da Espanha, sendo que as amostras ERR011305 e ERR011308 são pessoas do sexo feminino de idades 51 e 41 anos. As amostras ERR011323 e ERR011277 são de indivíduos do sexo masculino de idades 62 e 43 anos respectivamente (QIN et al., 2010).

Aplicamos todas as combinações de medidas nas amostras para realizar a identificação dos *reads*. Do total de medidas, as medidas que tiveram identificação acima de 70 % são: S2, S2 $\cap$ S3, S3, S4, GC %, GC % $\cap$ S2 e GC % $\cap$ S3. (Tabela 4.14). Como a medida 1-din não conseguiu identificar os *reads*, todas as combinações que utilizam essa medida ficaram com 0 %. Esse fato deve-se ao baixo valor do desvio padrão dessa medida, muitos *reads* das amostras tiveram resultados com valor próximo do aceito nos organismos dessa medida mas não ficavam na faixa aceita pela metodologia.

Para cada *read*, foi encontrado um conjunto de gêneros com valores dentro da faixa de aceitação pela metodologia para a identificação dos organismos. Para essa visualização, criamos histogramas (Fig 4.21, Fig 4.22 e Fig 4.23) mostrando a relação entre o percentual de gêneros encontrados e a quantidade total de gêneros. Quanto maior for o percentual em valores no eixo-y em relação aos valores mais baixos de gêneros encontrados do eixo-x, a medida terá um melhor desempenho na identificação.

A medida GC % apresentou um percentual médio de 96,5 % de identificação dos *reads* nas amostras. Ao analisarmos a quantidade de gêneros que tiveram resultados na faixa de aceitação pela metodologia, verificamos que são encontrados até aproximadamente 50 dos possíveis 202 gêneros (Fig 4.21). Nos histogramas da figura 4.21, nos itens A, B, C, E e G verificamos que a maior concentração de gêneros entre os *reads* foram entre os valores de 30 à 36. Nos itens D, F e H, além de uma concentração alta entre os valores de gêneros 30 à 36 também houveram valores mais altos para gêneros que não foram encontrados, confirmando os dados da tabela 4.14 onde essas três amostras obtiveram menor quantidade de *reads* identificados.

A medida S2 encontrou em média 94,5 % dos *reads*. Entretanto, apesar do alto valor de identificações, para cada *read* foi encontrado muitos gêneros candidatos compatíveis àquele valor. Com isso, alguns *reads* possuíam mais de 150 possíveis gêneros (Fig 4.22). As amostras identificadas utilizando a medida S2 resultaram em três quantidades de valores mais significativos. Em todas as amostras temos um percentual alto de amostras com apenas um gênero, temos uma faixa de gêneros encontrados entre os valores 56 à 61 e um pico no valor 121 demonstrando que muitos gêneros possuem faixas de valores próximas. Os *reads* que tiveram o valor de 121 gêneros encontrados não obtiveram consenso entre os gêneros identificados nos programas Raiphy, Phymm e Phymmbl.

A combinação das medidas GC % com S2 conseguiu identificar em média 89,5 % dos *reads*. Ao utilizar essas duas medidas combinadas foi possível diminuir a quantidade de gêneros encontrados por *read* pela medida GC % quando utilizada sozinha (Fig 4.23). Com essa diminuição significativa de gêneros encontrados é possível realizar um filtro mais adequado para a identificação de cada *read*.

Quando comparamos a metodologia com os programas Raiphy, Phymm e Phymmbl, a combinação que obteve melhores resultados foi a GC % $\cap$ S2. A quantidade de *reads* identificados pelos quatro programas como sendo do mesmo gênero são de 95 (0,0147 %) (ERR011259), 53 (0,0117 %) (ERR011323), 34 (0,0091 %) (ERR011305), 49 (0,0159 %) (ERR011277), 50 (0,0148 %) (ERR011172), 31 (0,0052 %) (ERR011190), 55 (0,0717 %) (ERR011308) e 163 (0,0139 %) (ERR011230) (Fig 4.24).Os gêneros mais abundantes encontrados por todos os programas nas amostras são: *Bacteroides* e *Eubacterium* confirmando os dados publicados por ((QIN et al., 2010)).

Tabela 4.14: Tabela com o percentual de organismos identificados nas amostras de metagenomas por medida da metodologia

ID	Medida	ERR011172	ERR011190	ERR011230	ERR011259	ERR011277	ERR011305	ERR011308	ERR011323
1	din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
2	S2	95 %	95 %	94 %	93 %	95 %	92 %	95 %	92 %
3	S2 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
4	S2 ∩ S3	85 %	85 %	77 %	86 %	79 %	87 %	80 %	83 %
5	S2 ∩ S3 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
6	S2 ∩ S3 ∩ S4	45 %	51 %	30 %	54 %	31 %	52 %	48 %	36 %
7	S2 ∩ S3 ∩ S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
8	S2 ∩ S4	45 %	51 %	30 %	54 %	31 %	53 %	48 %	37 %
9	S3	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
10	S3 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
11	S3 ∩ S4	54 %	58 %	38 %	63 %	37 %	63 %	54 %	48 %
12	S3 ∩ S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
13	S4	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
14	S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
15	GC %	96 %	97 %	97 %	93 %	99 %	92 %	99 %	92 %
16	GC % ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
17	GC % ∩ S2	90 %	91 %	89 %	88 %	91 %	87 %	90 %	87 %
18	GC % ∩ S2 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
19	GC % ∩ S2 ∩ S3	64 %	66 %	58 %	67 %	58 %	67 %	62 %	63 %
20	GC % ∩ S2 ∩ S3 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
21	GC % ∩ S2 ∩ S3 ∩ S4	33 %	39 %	21 %	40 %	22 %	38 %	37 %	25 %
22	GC % ∩ S2 ∩ S3 ∩ S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
23	GC % ∩ S2 ∩ S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
24	GC % ∩ S3	72 %	74 %	66 %	74 %	68 %	74 %	69 %	70 %
25	GC % ∩ S3 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
26	GC % ∩ S3 ∩ S4	30 %	35 %	20 %	37 %	19 %	36 %	34 %	24 %
27	GC % ∩ S3 ∩ S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
28	GC % ∩ S4	37 %	43 %	26 %	44 %	27 %	42 %	41 %	38 %
29	GC % ∩ S4 ∩ din	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %

Fonte: O autor (2016)

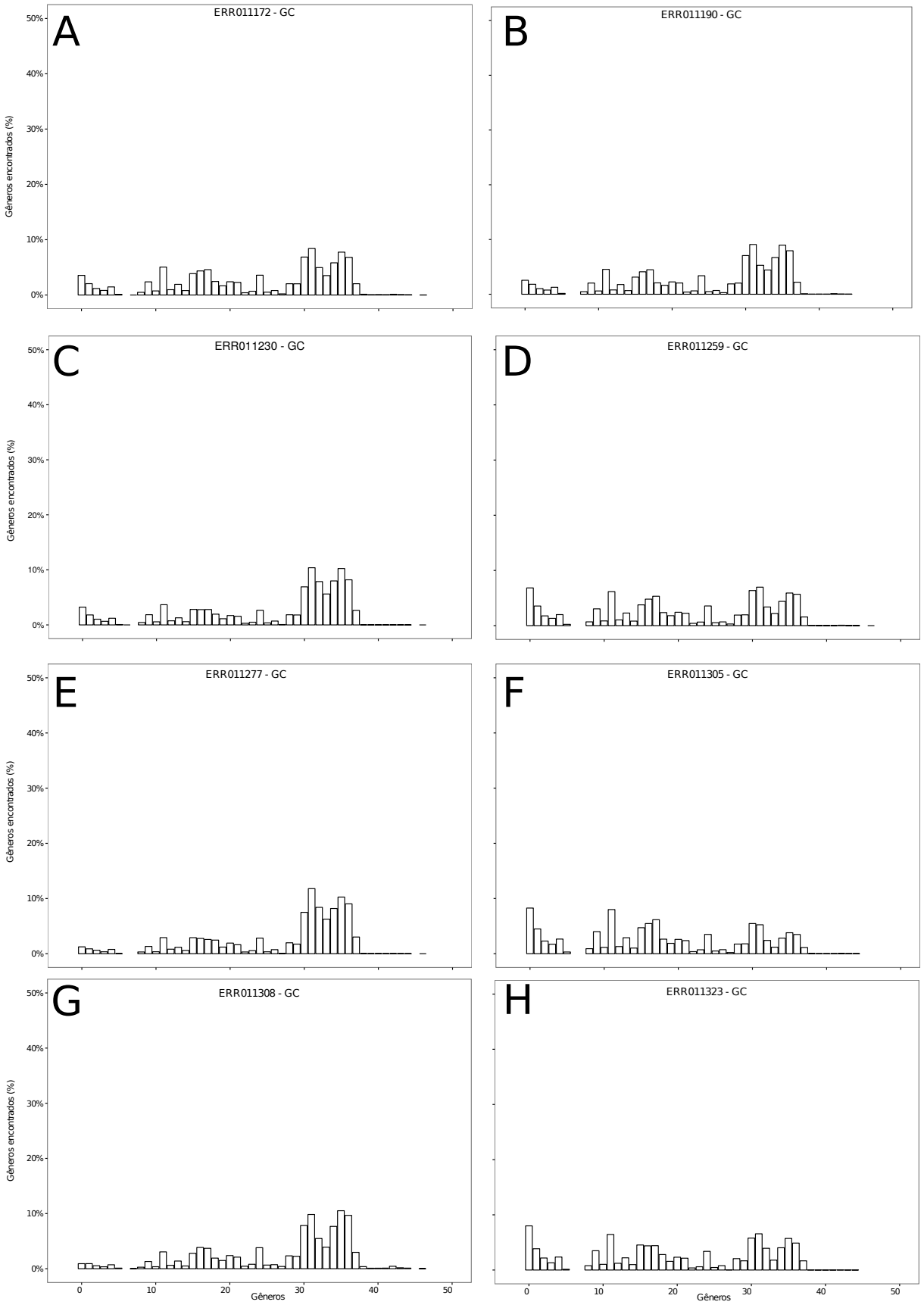

Figura 4.21: Percentual de gêneros encontrados em relação ao total de gêneros da metodologia aplicados nas amostras de metagenomas utilizando a medida GC %. A) amostra ERR011172, B) amostra ERR011190, C) amostra ERR011230, D) amostra ERR011259, E) amostra ERR011277, F) amostra ERR011305, G) amostra ERR011308 e H) amostra ERR011323

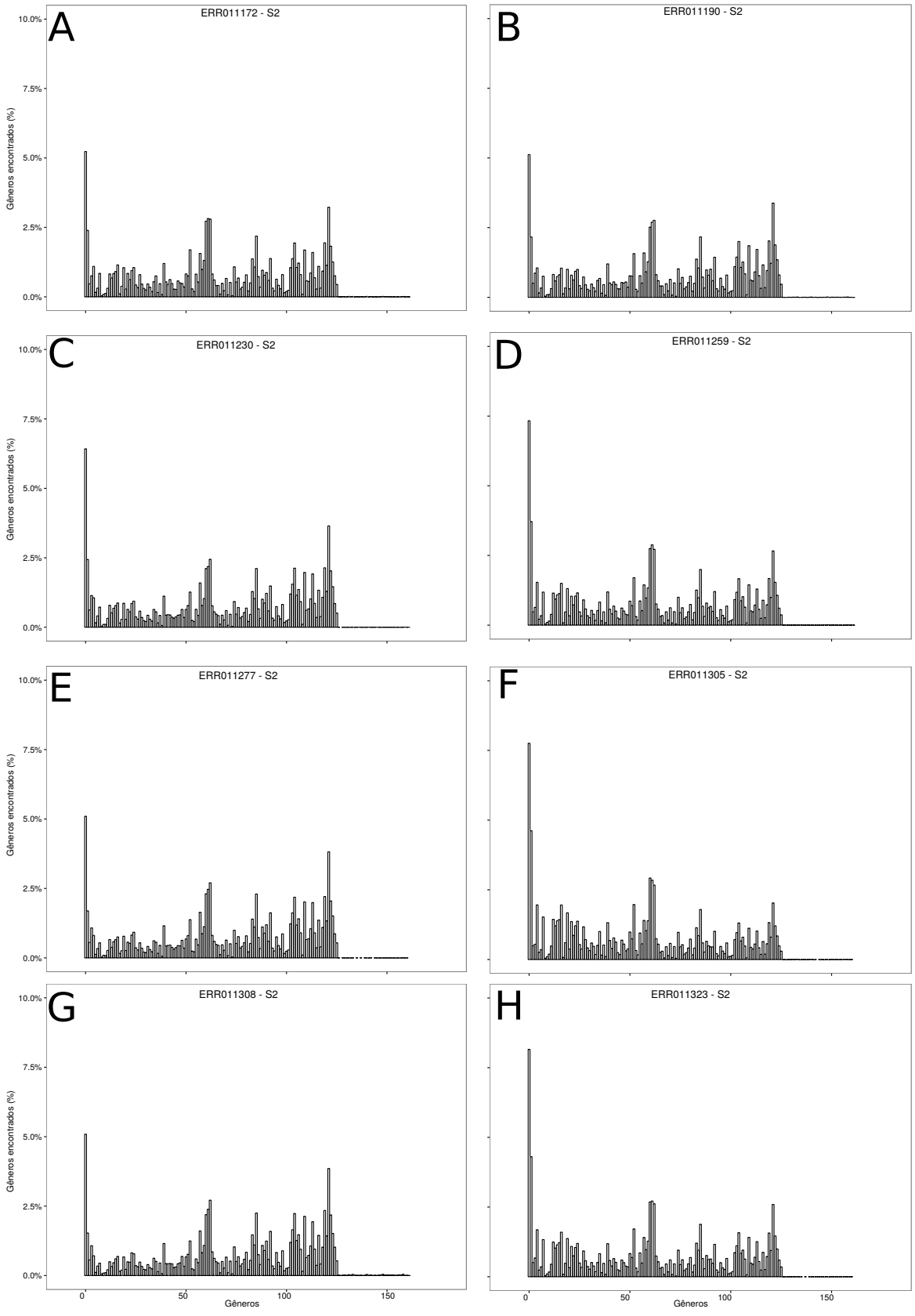

Figura 4.22: Percentual de gêneros encontrados em relação ao total de gêneros da metodologia aplicados nas amostras de metagenomas utilizando a medida S2. A) amostra ERR011172, B) amostra ERR011190, C) amostra ERR011230, D) amostra ERR011259, E) amostra ERR011277, F) amostra ERR011305, G) amostra ERR011308 e H) amostra ERR011323

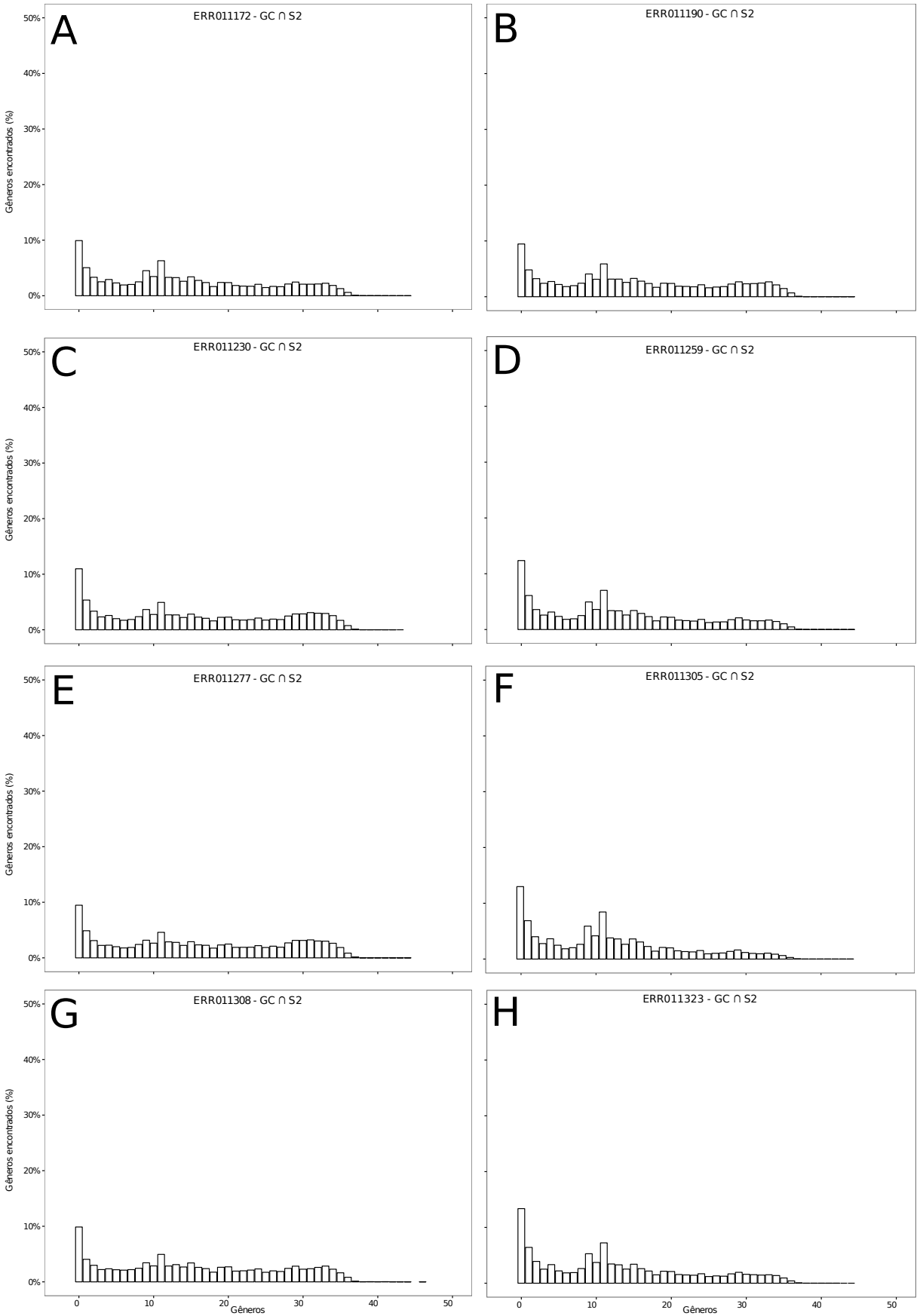

Figura 4.23: Percentual de gêneros encontrados em relação ao total de gêneros da metodologia aplicados nas amostras de metagenomas utilizando a combinação de medidas GC % e S2. A) amostra ERR011172, B) amostra ERR011190, C) amostra ERR011230, D) amostra ERR011259, E) amostra ERR011277, F) amostra ERR011305, G) amostra ERR011308 e H) amostra ERR011323

Figura 4.24: Diagrama de Venn comparando a quantidade de organismos identificados no táxon de gênero por *software* em relação a combinação das medidas GC % ∩ S2 para as amostras ERR011172, ERR011190, ERR011230, ERR011259, ERR011277, ERR011305, ERR011308 e ERR011323.

4.8 Dados finais da metodologia

Com o resultado dos testes foi desenvolvido o executável para a utilização da metodologia. O executável possui o tamanho de 830,9 KB. A soma do tamanho dos bancos de dados para todas os táxons ficou em 7,7 MB. O programa consegue classificar 100 mil *reads* em aproximadamente 4 minutos. Se compararmos aos *softwares* de classificação taxonômica utilizados nesse trabalho vemos uma discrepância entre esses valores, o Phymm/Phymmbl precisa entre 120 GB (mínimo) e 200 GB (recomendado) para a instalação dos scripts e o banco de dados que utiliza. O executável do RAIphy possui 9,5 MB e seu banco de dados 166,7 MB. Para classificar 100 mil sequências o Phymm/Phymmbl demora em média 8 horas enquanto o RAIphy realiza a mesma operação em 9 minutos.

5 *Conclusões*

Com a análise das combinações de medidas para identificação de microorganismos nos táxons gênero, família, ordem e classe, verificamos que a medida predominante foi a concentração de GC. Contudo, identificamos que para alguns tamanhos de fragmentos outras medidas como abundância total de dinucleotídeos, entropia de tripletes e entropia de tetrapletes conseguiram resultados acima de 70% de identificação.

Na aplicação da metodologia e comparação com as metodologias já existentes Raiphy, Phymmbl e Phymm nas amostras de metagenomas obtidas do projeto *"A human gut microbial gene catalog established by deep metagenomic sequencing"* houveram poucos *reads* identificamos pelas três metodologias como sendo do mesmo gênero. Ao aplicarmos as medidas do método que estamos propondo a combinação das medidas GC e S2 obtiveram os melhores resultados com identificações acima de 89 % e conseguindo diminuir a quantidade de possíveis gêneros candidatos por *read*. O tempo gasto para as análises foi menor em comparação com os programas já existentes.

Mesmo com a diminuição da quantidade de gêneros candidatos ainda consideramos alta a quantidade de candidatos. Uma forma de solucionar esse problema seria utilizando uma abordagem híbrida entre o método composicional e o método por alinhamento. O método composicional faria um filtro entre os possíveis gêneros e o método por alinhamento selecionaria entre os possíveis resultados de forma mais específica. Com isso haveria uma diminuição do tempo computacional necessário para a análise pois a metodologia que estamos propondo diminuiria consideravelmente a busca pela informação no banco de dados para realização do alinhamento.

Referências Bibliográficas

ALIMETRICS. *DNA Sequence Analysis - The only species-specific approach for novel bacteria*. `http://www.alimetrics.net/en/index.php/dna-sequence-analysis`. [Online; accessed 19-July-2008].

BERENDZEN, J. et al. Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. *BMC research notes*, BioMed Central Ltd, v. 5, n. 1, p. 460, 2012.

BOHLIN, J. et al. Analysis of intra-genomic gc content homogeneity within prokaryotes. *BMC genomics*, v. 11, p. 464, 2010. ISSN 1471-2164.

BRADY, A.; SALZBERG, S. L. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, Nature Publishing Group, v. 6, n. 9, p. 673–676, 2009.

BULLOCK, N. O.; ASLANZADEH, J. Biochemical profile-based microbial identification systems. In: *Advanced Techniques in Diagnostic Microbiology*. [S.l.]: Springer, 2013. p. 87–121.

CASE, R. J. et al. Use of 16s rrna and rpob genes as molecular markers for microbial ecology studies. *Applied and environmental microbiology*, Am Soc Microbiol, v. 73, n. 1, p. 278–288, 2007.

CHAKRAVORTY, S. et al. A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, v. 69, p. 330–339, May 2007. ISSN 0167-7012.

CHUN, J.; RAINEY, F. A. Integrating genomics into the taxonomy and systematics of the bacteria and archaea. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 64, n. 2, p. 316–324, 2014.

DELMONT, T. O. et al. Metagenomic comparison of direct and indirect soil dna extraction approaches. *Journal of microbiological methods*, Elsevier, v. 86, n. 3, p. 397–400, 2011.

DRANCOURT, M.; BERGER, P.; RAOULT, D. Systematic 16s rrna gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. *Journal of clinical microbiology*, v. 42, p. 2197–2202, May 2004. ISSN 0095-1137.

DUTTA, A. et al. Binpairs: Utilization of illumina paired-end information for improving efficiency of taxonomic binning of metagenomic sequences. *PloS one*, Public Library of Science, v. 9, n. 12, p. e114814, 2014.

DUTTA, C.; PAUL, S. Microbial lifestyle and genome signatures. *Current genomics*, Bentham Science Publishers, v. 13, n. 2, p. 153–162, 2012.

ESCOBAR-ZEPEDA, A.; LEÓN, A. V.-P. de; SANCHEZ-FLORES, A. The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. *Frontiers in genetics*, Frontiers Media SA, v. 6, 2015.

FIERER, N.; JACKSON, R. B. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, National Acad Sciences, v. 103, n. 3, p. 626–631, 2006.

GERHARDT, N. L. G. J.; CORSO, G. Network clustering coefficient approach to dna sequence analysis. *Chaos, Solitons and Fractals*, v. 28, p. 1037–1045, 2006.

HANDELSMAN, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, Am Soc Microbiol, v. 68, n. 4, p. 669–685, 2004.

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, Elsevier, v. 5, n. 10, p. R245–R249, 1998.

HAQUE, M. M. et al. Sort-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, Oxford Univ Press, v. 25, n. 14, p. 1722–1730, 2009.

HEATHER, J. M.; CHAIN, B. The sequence of sesequence: The history of sequencing dna. *Genomics*, v. 107, n. 1, p. 1–8, January 2016.

HENRY, V. J. et al. Omictools: an informative directory for multi-omic data analysis. *Database*, Oxford University Press, v. 2014, p. bau069, 2014.

HIGASHI, S. et al. Analysis of composition-based metagenomic classification. *BMC genomics*, BioMed Central, v. 13, n. 5, p. 1, 2012.

HUSON, D. H. et al. Megan analysis of metagenomic data. *Genome research*, Cold Spring Harbor Lab, v. 17, n. 3, p. 377–386, 2007.

JANDA, J. M.; ABBOTT, S. L. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, Am Soc Microbiol, v. 45, n. 9, p. 2761–2764, 2007.

KARLIN, S. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, v. 1, n. 5, p. 598–610, Oct 1998.

KARLIN, S.; BURGE, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, v. 11, n. 7, p. 283–290, Jul 1995.

KEEGAN, K. P.; GLASS, E. M.; MEYER, F. Mg-rast, a metagenomics service for analysis of microbial community structure and function. *Methods in molecular biology (Clifton, NJ)*, Springer, v. 1399, p. 207–233, 2015.

KLINDWORTH, A. et al. Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, Oxford Univ Press, p. gks808, 2012.

KRAUSE, L. et al. Phylogenetic classification of short environmental dna fragments. *Nucleic acids research*, Oxford Univ Press, v. 36, n. 7, p. 2230–2239, 2008.

LADOUKAKIS, E.; KOLISIS, F. N.; CHATZIIOANNOU, A. A. Integrative workflows for metagenomic analysis. *Multi-omic Data Integration*, Frontiers Media SA, p. 115, 2015.

LE, V. V.; TRAN, L. V.; TRAN, H. V. A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. *BMC bioinformatics*, BioMed Central, v. 17, n. 1, p. 1, 2016.

LEWIS, K. et al. Uncultured microorganisms as a source of secondary metabolites. *The Journal of antibiotics*, Nature Publishing Group, v. 63, n. 8, p. 468–476, 2010.

LIU, J. et al. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic acids research*, Oxford Univ Press, p. gks828, 2012.

LIU, L. et al. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, v. 2012, p. 251364, 2012. ISSN 1110-7251.

MACDONALD, N. J.; PARKS, D. H.; BEIKO, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic acids research*, Oxford Univ Press, p. gks335, 2012.

MADIGAN, M. T. et al. *Brock Biology of Microorganisms 13th edition*. [S.l.]: Benjamin Cummings, 2010.

MCHARDY, A. C. et al. Accurate phylogenetic classification of variable-length dna fragments. *Nature methods*, Nature Publishing Group, v. 4, n. 1, p. 63–72, 2007.

METZKER, M. L. Sequencing technologies - the next generation. *Nature reviews genetics*, Nature Publishing Group, v. 11, n. 1, p. 31–46, 2010.

MOHAMMED, M. H. et al. Indus-a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC genomics*, BioMed Central, v. 12, n. 3, p. 1, 2011.

MOHAMMED, M. H. et al. Sphinx - an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, Oxford Univ Press, v. 27, n. 1, p. 22–30, 2011.

MOREIRA, L. M. *Ciencias genomicas : fundamentos e aplicacoes*. Sociedade Brasileira de Genetica, 2015. Disponível em: <https://drive.google.com/file/d-/0Bwb0CldyiHPwV3IxRGdyYklIdGc/view>.

MUTO, A.; OSAWA, S. The guanine and cytosine content of genomic dna and bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, v. 84, p. 166–169, Jan 1987. ISSN 0027-8424.

NALBANTOGLU, O. U. et al. Raiphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 1, 2011.

PENG, Y. et al. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics (Oxford, England)*, v. 27, p. i94–101, Jul 2011. ISSN 1367-4811.

QIN, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, v. 464, n. 7285, p. 59–65, Mar 2010. Disponível em: <http://dx.doi.org/10.1038-/nature08821>.

RASHEED, Z.; RANGWALA, H. Metagenomic taxonomic classification using extreme learning machines. *Journal of bioinformatics and computational biology*, World Scientific, v. 10, n. 05, p. 1250015, 2012.

REDDY, R. M.; MOHAMMED, M. H.; MANDE, S. S. Twarit: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. *Gene*, Elsevier, v. 505, n. 2, p. 259–265, 2012.

ROSENBERG, E. et al. The evolution of animals and plants via symbiosis with microorganisms. *Environmental microbiology reports*, Wiley Online Library, v. 2, n. 4, p. 500–506, 2010.

ROSENBERG, E.; ZILBER-ROSENBERG, I. Symbiosis and development: the hologenome concept. *Birth Defects Research Part C: Embryo Today: Reviews*, Wiley Online Library, v. 93, n. 1, p. 56–66, 2011.

SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 12, p. 5463–5467, 1977.

SANTOS, L. dos; RYBARCZYK-FILHO; GERHARDT, G. J. L. Triplet entropy in h1n1 virus. *TendÃªncias em MatemÃ¡tica Aplicada e Computacional*, v. 12, p. 253–261, 2011.

SCHLOSS, P. D.; HANDELSMAN, J. Metagenomics for studying unculturable microorganisms: cutting the gordian knot. *Genome biology*, BioMed Central, v. 6, n. 8, p. 1, 2005.

SCHREIBER, F. et al. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, Oxford Univ Press, v. 26, n. 7, p. 960–961, 2010.

SHARPTON, T. J. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science*, v. 5, p. 209, 2014. ISSN 1664-462X.

SILVA, G. G. Z. et al. Focus: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, PeerJ Inc., v. 2, p. e425, 2014.

TORTOLI, E. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. *Clinical microbiology reviews*, Am Soc Microbiol, v. 16, n. 2, p. 319–354, 2003.

VERVIER, K. et al. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, Oxford Univ Press, v. 32, n. 7, p. 1023–1032, 2016.

WINNICK, E. Dna sequencing industry sets its sights on the future: new technologies may duke it out with upgrades of the current method. *The Scientist*, Scientist Inc., v. 18, n. 18, p. 44–48, 2004.

WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 11, p. 5088–5090, 1977.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, BioMed Central, v. 15, n. 3, p. 1, 2014.

YAKOVCHUK, P.; PROTOZANOVA, E.; FRANK-KAMENETSKII, M. D. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic Acids Res*, v. 34, n. 2, p. 564–574, 2006.

ZILBER-ROSENBERG, I.; ROSENBERG, E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS microbiology reviews*, The Oxford University Press, v. 32, n. 5, p. 723–735, 2008.

APÊNDICE A – Tabela comparativa das principais programas de análise

Tabela A.1: Tabela comparativa das principais programas de análise

NOME	Tipo	Link	Pais
AmphoraNet	Alinhamento	http://pitgroup.org/amphoranet/	Hungria
Binpairs	Híbrido	http://metagenomics.atc.tcs.com/binning/binpairs/	Índia
Bracken	Alinhamento	http://ccb.jhu.edu/software/bracken/	USA
CARMA	Alinhamento	http://www.cebitec.uni-bielefeld.de/index.php/2-uncategorised/47-carma?highlight=WyJjYXJtYSJd	Alemanha
CensuScope	Alinhamento	https://hive.biochemistry.gwu.edu/dna.cgi?cmd=censuscope	USA
Centrifuge	Alinhamento	http://www.ccb.jhu.edu/software/centrifuge/	USA
ClaMS	Composição	http://clams.jgi-psf.org/	USA
CLARK	Alinhamento	http://clark.cs.ucr.edu/	USA
CoMeta	Alinhamento	https://github.com/jkawulok/cometa	Polônia
CSSSCL	Alinhamento	https://github.com/oicr-ibc/cssscl	Canada
DiScRIBinATE	Alinhamento	http://metagenomics.atc.tcs.com/binning/DiScRIBinATE/	Índia
DUDes	Alinhamento	https://sourceforge.net/projects/dudes/	Suíça
FOCUS	Composição	http://edwards.sdsu.edu/FOCUS2/	USA
Genometa	Alinhamento	http://genomics1.mh-hannover.de/genometa/index.php?Site=Home	Alemanha
GOTTCHA	Alinhamento	https://github.com/LANL-Bioinformatics/GOTTCHA	USA
GSMer	Alinhamento	https://github.com/qichao1984/GSMer	USA
GSTaxClassifier	Alinhamento	OBSOLETO	Flórida
INDUS	Composição	http://metagenomics.atc.tcs.com/INDUS/	Índia
Kaiju	Alinhamento	http://kaiju.binf.ku.dk/	Dinamarca
Kraken	Alinhamento	http://ccb.jhu.edu/software/kraken/	USA
Large scale metagenomics	Composição	http://cbio.mines-paristech.fr/largescalemetagenomics/	França
LMAT	Alinhamento	http://computation.llnl.gov/projects/livermore-metagenomics-analysis-toolkit	Califórnia
MARTA	Alinhamento	OBSOLETO	USA
MEGAN	Alinhamento	http://ab.inf.uni-tuebingen.de/software/megan5/	Alemanha
metaBEETL	Alinhamento	https://github.com/BEETL/BEETL	Alemanha
MetaBinG	Alinhamento	http://cbb.sjtu.edu.cn/∼ccwei/pub/software/MetaBinG/MetaBinG.php	China
MetaCRAM	Alinhamento	http://web.engr.illinois.edu/∼mkim158/metacram.html	USA
MetaCV	Composição	http://metacv.sourceforge.net/	China
MetaFlow	Alinhamento	https://www.cs.helsinki.fi/en/gsa/metaflow	Finlândia
MetaPalette	Alinhamento	https://github.com/dkoslicki/MetaPalette	USA
MetaPhlAn	Alinhamento	https://bitbucket.org/nsegata/metaphlan/wiki/Home	Itália
MetaPhyler	Alinhamento	http://metaphyler.cbcb.umd.edu/	USA
MG-RAST	Alinhamento	http://metagenomics.anl.gov/	USA

Tabela A.2: Tabela comparativa das principais programas de análise (continuação)

NOME	Tipo	Link	Pais
MLTreeMap	Alinhamento	http://mltreemap.org/	Suíça
mOTU	Alinhamento	http://www.bork.embl.de/software/mOTU/	Alemanha
MyTaxa	Alinhamento	http://enve-omics.ce.gatech.edu/MyTaxa/	USA
Naïve Bayes Classification	Alinhamento	http://nbc.ece.drexel.edu/	USA
One Codex	Alinhamento	https://www.onecodex.com/	USA
PhyloPythia	Composição	OBSOLETO	USA
PhyloPythiaS	Composição	http://phylopythias.bifo.helmholtz-hzi.de/index.php?phase=wait	USA
PhyloSift	Alinhamento	https://github.com/gjospin/PhyloSift	USA
PhymmBL	Híbrido	http://metacv.sourceforge.net/	USA
Pplacer	Alinhamento	http://matsen.fhcrc.org/pplacer/	USA
PROTAX	Alinhamento	http://www.helsinki.fi/science/metapop/Software.htm	Finlândia
ProViDE	Alinhamento	http://metagenomics.atc.tcs.com/binning/ProViDE/	Índia
RAIphy	Composição	http://bioinfo.unl.edu/raiphy.php	USA
RIEMS	Alinhamento	https://www.fli.de/en/institutes/institut-fuer-virusdiagnostik/labore -arbeitsgruppen/labor-fuer-ngs-und-microarray-diagnostik/	Alemanha
RITA	Híbrido	http://kiwi.cs.dal.ca/Software/RITA	Canada
SeMeta	Híbrido	http://it.hcmute.edu.vn/bioinfo/metapro/SeMeta.html	Vietnã
Sequedex	Composição	http://sequedex.lanl.gov/	USA
SOrt-ITEMS	Alinhamento	http://metagenomics.atc.tcs.com/binning/SOrt-ITEMS/	Índia
spaced-seeds-for-metagenomics	Alinhamento	https://github.com/gregorykucherov/spaced-seeds-for-metagenomics	França
SPANNER	Alinhamento	http://kiwi.cs.dal.ca/Software/SPANNER	USA
SPHINX	Híbrido	http://metagenomics.atc.tcs.com/SPHINX/	Índia
TAC-ELM	Composição	http://cs.gmu.edu/∼mlbio/TAC-ELM/	USA
TACOA	Composição	http://www.cebitec.uni-bielefeld.de/index.php/2-uncategorised/99-tacoa? highlight=WyJ0YWNvYSJd	Alemanha
TAEC	Alinhamento	http://cals.arizona.edu/∼anling/sbg/software.htm	USA
TAMER	Alinhamento	http://cals.arizona.edu/∼anling/sbg/software.htm	USA
Taxator-tk	Alinhamento	https://github.com/fungs/taxator-tk/	Alemanha
Taxonomer	Alinhamento	http://taxonomer.iobio.io/	USA
TaxSOM	Composição	http://soma.arb-silva.de/	Alemanha
Treephyler	Híbrido	http://www.gobics.de/fabian/treephyler.php	Alemanha
TWARIT	Híbrido	http://metagenomics.atc.tcs.com/Twarit/	Índia
WebCARMA	Alinhamento	http://wwww.cebitec.uni-bielefeld.de/webcarma.cebitec.uni-bielefeld.de/	Alemanha
WebMGA	Alinhamento	http://weizhong-lab.ucsd.edu/metagenomic-analysis/	USA
WGSQuikr	Alinhamento	https://sourceforge.net/projects/wgsquikr/	USA

Fonte: Henry,V.J., Bandrowski,A.E., Pepin,A.-S. et al. OMICtools: an informative directory for multi-omic data analysis. Database (2014) Vol. 2014: article ID bau069; doi:10.1093/database/bau069. Disponível em: `https://omictools.com/taxonomy-dependent2-category`. Acesso em: 05 ago. 2016.

APÊNDICE B – Amostras de Metagenoma

Tabela B.1: Tabela com informações das amostras obtidas no EBI

ID	Sexo	Localização	Idade	Tamanho (MB)	ID	Sexo	Localização	Idade	Tamanho (MB)
ERR011089	Feminino	Dinamarca	59	18M	ERR011177	Masculino	Dinamarca	58	116K
ERR011090	Feminino	Dinamarca	59	262M	ERR011178	Feminino	Dinamarca	67	772M
ERR011091	Feminino	Dinamarca	59	120K	ERR011179	Feminino	Dinamarca	67	596K
ERR011092	Masculino	Dinamarca	69	117M	ERR011180	Masculino	Dinamarca	59	448M
ERR011093	Masculino	Dinamarca	69	211M	ERR011181	Masculino	Dinamarca	59	248K
ERR011094	Masculino	Dinamarca	69	280K	ERR011182	Masculino	Dinamarca	49	74M
ERR011101	Feminino	Dinamarca	59	2,4M	ERR011183	Masculino	Dinamarca	49	1,2M
ERR011102	Feminino	Dinamarca	59	3,0M	ERR011184	Masculino	Dinamarca	69	290M
ERR011103	Feminino	Dinamarca	59	3,3M	ERR011185	Masculino	Dinamarca	69	252K
ERR011104	Feminino	Dinamarca	59	3,3M	ERR011186	Masculino	Dinamarca	64	503M
ERR011109	Masculino	Dinamarca	64	71M	ERR011187	Masculino	Dinamarca	64	296K
ERR011110	Masculino	Dinamarca	64	162M	ERR011188	Masculino	Dinamarca	59	900M
ERR011111	Masculino	Dinamarca	64	136K	ERR011189	Masculino	Dinamarca	59	676K
ERR011114	Feminino	Dinamarca	-	325M	ERR011190	Masculino	Dinamarca	54	99M
ERR011115	Feminino	Dinamarca	-	52M	ERR011191	Masculino	Dinamarca	54	420K
ERR011116	Feminino	Dinamarca	-	788K	ERR011192	Feminino	Dinamarca	69	733M
ERR011117	Feminino	Dinamarca	42	628M	ERR011193	Feminino	Dinamarca	69	384K
ERR011118	Feminino	Dinamarca	42	532M	ERR011194	Feminino	Dinamarca	54	58M
ERR011119	Feminino	Dinamarca	42	936K	ERR011195	Feminino	Dinamarca	54	152K
ERR011120	Feminino	Dinamarca	42	1,3G	ERR011196	Feminino	Dinamarca	44	743M
ERR011121	Feminino	Dinamarca	42	1,2G	ERR011197	Feminino	Dinamarca	44	680K
ERR011122	Feminino	Dinamarca	42	1,4M	ERR011198	Masculino	Dinamarca	49	421M
ERR011123	Feminino	Dinamarca	42	1,4M	ERR011199	Masculino	Dinamarca	49	636K
ERR011126	Feminino	Dinamarca	54	494M	ERR011200	Feminino	Dinamarca	69	100M
ERR011127	Feminino	Dinamarca	54	1018M	ERR011201	Feminino	Dinamarca	69	60K
ERR011128	Feminino	Dinamarca	54	616K	ERR011202	Feminino	Dinamarca	49	333M
ERR011131	Feminino	Dinamarca	49	556M	ERR011203	Feminino	Dinamarca	49	1,8M
ERR011132	Feminino	Dinamarca	49	560M	ERR011204	Feminino	Dinamarca	49	653M
ERR011133	Feminino	Dinamarca	49	756K	ERR011205	Feminino	Dinamarca	49	1,6M
ERR011140	Feminino	Dinamarca	63	24M	ERR011206	Masculino	Dinamarca	49	581M
ERR011141	Feminino	Dinamarca	63	96K	ERR011207	Masculino	Dinamarca	49	964K
ERR011142	Feminino	Dinamarca	49	339M	ERR011208	Masculino	Dinamarca	59	705M
ERR011143	Feminino	Dinamarca	49	232K	ERR011209	Masculino	Dinamarca	59	1,8M
ERR011148	Feminino	Dinamarca	59	807M	ERR011210	Masculino	Dinamarca	54	15M
ERR011149	Feminino	Dinamarca	59	76K	ERR011211	Masculino	Dinamarca	54	540K
ERR011150	Feminino	Dinamarca	49	271M	ERR011212	Feminino	Dinamarca	54	410M
ERR011151	Feminino	Dinamarca	49	44K	ERR011213	Feminino	Dinamarca	54	284K
ERR011152	Feminino	Dinamarca	49	17M	ERR011214	Feminino	Dinamarca	54	221M

Tabela B.2: Tabela com informações das amostras obtidas no EBI (continuação)

ID	Sexo	Localização	Idade	Tamanho (MB)	ID	Sexo	Localização	Idade	Tamanho (MB)
ERR011153	Feminino	Dinamarca	49	508K	ERR011215	Feminino	Dinamarca	54	504K
ERR011156	Feminino	Dinamarca	44	781M	ERR011216	Masculino	Dinamarca	59	214M
ERR011157	Feminino	Dinamarca	44	72K	ERR011217	Masculino	Dinamarca	59	1,1M
ERR011158	Masculino	Dinamarca	59	97M	ERR011218	Masculino	Dinamarca	54	402M
ERR011159	Masculino	Dinamarca	59	28K	ERR011219	Masculino	Dinamarca	54	760K
ERR011160	Masculino	Dinamarca	69	518M	ERR011220	Feminino	Dinamarca	69	504M
ERR011161	Masculino	Dinamarca	69	96K	ERR011221	Feminino	Dinamarca	69	112K
ERR011162	Masculino	Dinamarca	69	234M	ERR011222	Feminino	Dinamarca	49	315M
ERR011163	Masculino	Dinamarca	69	48K	ERR011223	Feminino	Dinamarca	49	784K
ERR011164	Feminino	Dinamarca	59	60M	ERR011224	Masculino	Dinamarca	59	314M
ERR011165	Feminino	Dinamarca	59	240K	ERR011225	Masculino	Dinamarca	59	148K
ERR011166	Masculino	Dinamarca	54	432K	ERR011226	Feminino	Dinamarca	54	277M
ERR011167	Masculino	Dinamarca	54	228K	ERR011227	Feminino	Dinamarca	54	48K
ERR011168	Masculino	Dinamarca	49	675M	ERR011228	Masculino	Dinamarca	59	533M
ERR011169	Masculino	Dinamarca	49	416K	ERR011229	Masculino	Dinamarca	59	172K
ERR011170	Masculino	Dinamarca	64	46M	ERR011230	Feminino	Dinamarca	44	206M
ERR011171	Masculino	Dinamarca	64	524K	ERR011231	Feminino	Dinamarca	44	368K
ERR011172	Masculino	Dinamarca	44	58M	ERR011232	Masculino	Dinamarca	54	277M
ERR011173	Masculino	Dinamarca	44	520K	ERR011233	Masculino	Dinamarca	54	124K
ERR011174	Feminino	Dinamarca	54	17M	ERR011234	Feminino	Dinamarca	54	244M
ERR011175	Feminino	Dinamarca	54	220K	ERR011235	Feminino	Dinamarca	54	224K
ERR011176	Masculino	Dinamarca	58	27M	ERR011236	Feminino	Dinamarca	59	2,7M
ERR011237	Feminino	Dinamarca	59	82M	ERR011295	Masculino	Espanha	44	169M
ERR011238	Feminino	Dinamarca	59	1020K	ERR011296	Masculino	Espanha	44	196K
ERR011239	Masculino	Dinamarca	49	403M	ERR011297	Feminino	Espanha	44	176K
ERR011240	Masculino	Dinamarca	49	368K	ERR011298	Feminino	Espanha	44	414M
ERR011241	Feminino	Dinamarca	44	584M	ERR011299	Feminino	Espanha	55	443M
ERR011242	Feminino	Dinamarca	44	76K	ERR011300	Feminino	Espanha	55	120K
ERR011243	Feminino	Dinamarca	64	8,3M	ERR011301	Feminino	Espanha	57	126M
ERR011244	Feminino	Dinamarca	64	56K	ERR011302	Feminino	Espanha	57	52K
ERR011245	Masculino	Dinamarca	54	428M	ERR011303	Feminino	Espanha	25	24M
ERR011246	Masculino	Dinamarca	54	220K	ERR011304	Feminino	Espanha	25	760K
ERR011247	Feminino	Dinamarca	49	215M	ERR011305	Feminino	Espanha	62	63M
ERR011248	Feminino	Dinamarca	49	564K	ERR011306	Feminino	Espanha	62	944K
ERR011249	Masculino	Dinamarca	64	501M	ERR011307	Feminino	Espanha	41	78M
ERR011250	Masculino	Dinamarca	64	404K	ERR011308	Feminino	Espanha	41	13M
ERR011251	Feminino	Dinamarca	69	857M	ERR011309	Masculino	Espanha	68	983M
ERR011252	Feminino	Dinamarca	69	416K	ERR011310	Masculino	Espanha	68	1,3M
ERR011253	Feminino	Dinamarca	49	705M	ERR011311	Feminino	Espanha	41	211M
ERR011254	Feminino	Dinamarca	49	408K	ERR011312	Feminino	Espanha	41	444K
ERR011255	Feminino	Dinamarca	49	188M	ERR011313	Feminino	Espanha	34	441M
ERR011256	Feminino	Dinamarca	49	68K	ERR011314	Feminino	Espanha	34	7,4M

Tabela B.3: Tabela com informações das amostras obtidas no EBI (continuação)

ID	Sexo	Localização	Idade	Tamanho (MB)	ID	Sexo	Localização	Idade	Tamanho (MB)
ERR011257	Feminino	Dinamarca	64	901M	ERR011315	Masculino	Espanha	49	18M
ERR011258	Feminino	Dinamarca	64	108K	ERR011316	Masculino	Espanha	49	332K
ERR011259	Feminino	Dinamarca	59	108M	ERR011317	Feminino	Espanha	18	59M
ERR011260	Feminino	Dinamarca	59	832K	ERR011318	Feminino	Espanha	18	2,2M
ERR011261	Feminino	Dinamarca	49	882M	ERR011319	Feminino	Espanha	46	128M
ERR011262	Feminino	Dinamarca	49	196K	ERR011320	Feminino	Espanha	46	236K
ERR011263	Feminino	Dinamarca	59	508M	ERR011321	Feminino	Espanha	36	312M
ERR011264	Feminino	Dinamarca	59	300K	ERR011322	Feminino	Espanha	36	756K
ERR011265	Feminino	Dinamarca	54	519M	ERR011323	Masculino	Espanha	51	80M
ERR011266	Feminino	Dinamarca	54	1016K	ERR011324	Masculino	Espanha	51	164K
ERR011267	Masculino	Dinamarca	64	52M	ERR011325	Feminino	Espanha	48	331M
ERR011268	Masculino	Dinamarca	64	188K	ERR011326	Feminino	Espanha	48	468K
ERR011269	Feminino	Dinamarca	59	605M	ERR011327	Masculino	Espanha	45	37M
ERR011270	Feminino	Dinamarca	59	240K	ERR011328	Masculino	Espanha	45	972K
ERR011271	Feminino	Dinamarca	59	60M	ERR011329	Feminino	Espanha	51	248M
ERR011272	Feminino	Dinamarca	59	2,4M	ERR011330	Feminino	Espanha	51	2,5M
ERR011273	Masculino	Espanha	37	266M	ERR011331	Feminino	Espanha	53	98M
ERR011274	Masculino	Espanha	37	2,7M	ERR011332	Feminino	Espanha	53	184K
ERR011275	Feminino	Espanha	34	494M	ERR011333	Feminino	Espanha	25	66M
ERR011276	Feminino	Espanha	34	128K	ERR011334	Feminino	Espanha	25	336K
ERR011277	Masculino	Espanha	43	55M	ERR011335	Feminino	Espanha	41	33M
ERR011278	Masculino	Espanha	43	68K	ERR011336	Feminino	Espanha	41	164K
ERR011279	Feminino	Espanha	68	104M	ERR011337	Feminino	Espanha	63	52M
ERR011280	Feminino	Espanha	68	28K	ERR011338	Feminino	Espanha	63	384K
ERR011281	Masculino	Espanha	31	178M	ERR011339	Feminino	Espanha	37	234M
ERR011282	Feminino	Espanha	31	24K	ERR011340	Feminino	Espanha	37	96K
ERR011283	Masculino	Espanha	47	257M	ERR011341	Masculino	Espanha	62	231M
ERR011284	Masculino	Espanha	47	88K	ERR011342	Masculino	Espanha	62	84K
ERR011285	Masculino	Espanha	56	357M	ERR011343	Feminino	Espanha	38	500M
ERR011286	Masculino	Espanha	56	32K	ERR011344	Feminino	Espanha	38	3,6M
ERR011287	Masculino	Espanha	48	313M	ERR011345	Feminino	Espanha	19	640K
ERR011288	Masculino	Espanha	48	3,5M	ERR011346	Feminino	Espanha	19	571M
ERR011289	Masculino	Espanha	42	438M	ERR011347	Masculino	Espanha	22	360M
ERR011290	Masculino	Espanha	42	124K	ERR011348	Masculino	Espanha	22	2,7M
ERR011291	Feminino	Espanha	51	337M	ERR011349	Masculino	Espanha	32	275M
ERR011292	Feminino	Espanha	51	676K	ERR011350	Masculino	Espanha	32	6,5M
ERR011294	Feminino	Espanha	49	272K					

Fonte: O autor (2016)