Fundamentos do Sequenciamento Metagenômico de rRNA 16S: Base de estudo e pesquisa destinado ao Projeto YARA

---

1. Introdução à Metagenômica e ao Gene rRNA 16S

1.1. Conceito de Metagenômica e sua Importância

A metagenômica emergiu como um campo de pesquisa transformador, dedicado ao estudo de comunidades microbianas em seus ambientes naturais de forma integrada, transcendendo a análise de organismos isolados cultivados em laboratório.[1] Este campo revolucionou a compreensão das comunidades microbianas, oferecendo percepções sem precedentes sobre sua diversidade genética e funcional em uma vasta gama de ecossistemas terrestres.[2] A importância fundamental da metagenômica reside na constatação de que a esmagadora maioria dos microrganismos, estimada em aproximadamente 99%, presentes em ambientes naturais, não é passível de cultivo utilizando as técnicas laboratoriais convencionais.[1] Esta limitação histórica da microbiologia clássica impedia o acesso ao vasto potencial genético e ecológico da maior parte da biosfera microbiana. A metagenômica contorna esse obstáculo ao permitir a análise direta do material genético (DNA ou RNA) extraído de uma amostra ambiental complexa, desvendando a composição e o potencial funcional de comunidades inteiras, incluindo os membros não cultiváveis.

A capacidade de investigar o "metagenoma" – o conjunto genômico coletivo de uma comunidade microbiana – representa uma mudança de paradigma. Tradicionalmente, a microbiologia focava no estudo de microrganismos em culturas puras, o que, embora fundamental, oferecia uma visão limitada da complexidade das interações ecológicas e do repertório genético total presente em um ambiente. Ao analisar o DNA diretamente do ambiente, a metagenômica permite o estudo da vasta porção "não cultivável" da vida microbiana.[1] Isso significa que os pesquisadores podem agora investigar os papéis ecológicos, a diversidade taxonômica e o potencial funcional de uma fração significativamente maior da biosfera microbiana, que anteriormente permanecia inacessível. Tal abordagem tem implicações profundas para a descoberta de novos genes, vias metabólicas inovadoras e para uma compreensão mais completa da dinâmica de ecossistemas complexos, desde o microbioma humano até os solos e oceanos. As aplicações da metagenômica são, portanto, vastas e importantes, abrangendo desde a ecologia microbiana fundamental, passando pela saúde humana e animal, até a biotecnologia e o monitoramento ambiental.[2]

1.2. O Gene rRNA 16S como Marcador Filogenético

No cerne de muitos estudos metagenômicos focados na identificação taxonômica de bactérias e arqueias está o gene ribossomal 16S (rRNA 16S). Este gene codifica um componente essencial da subunidade menor (30S) do ribossomo procariótico, organela responsável pela síntese proteica.[3] Com um tamanho aproximado de 1500 pares de bases [1], o gene rRNA 16S possui uma estrutura molecular peculiar que o torna um marcador filogenético ideal. Ele é caracterizado pela presença de regiões altamente conservadas ao longo da evolução, que são intercaladas por nove regiões hipervariáveis (designadas de V1 a V9).[1]

A escolha de qual(is) região(ões) hipervariável(eis) sequenciar (por exemplo, V3-V4, V4, V4-V5) é uma consideração importante no desenho experimental, pois diferentes regiões possuem distintos graus de variabilidade e, consequentemente, diferentes capacidades de resolução taxonômica para diversos grupos microbianos.[4] A universalidade do gene rRNA 16S, estando presente em virtualmente todas as bactérias e arqueias conhecidas, combinada com sua taxa de mutação relativamente lenta e aparentemente aleatória ao longo do tempo evolutivo, confere-lhe o status de um "cronômetro molecular".[1] Essa característica permite a reconstrução de relações filogenéticas e a classificação taxonômica dos microrganismos presentes em uma comunidade.

A utilidade do gene rRNA 16S como marcador molecular reside precisamente em sua estrutura em mosaico, onde as regiões conservadas flanqueiam as regiões variáveis. As regiões conservadas são cruciais porque servem como sítios de anelamento para primers "universais" ou de amplo espectro, possibilitando a amplificação por PCR do gene rRNA 16S a partir de uma mistura complexa de DNA de diferentes microrganismos.[1] Sem essas regiões conservadas, seria impraticável utilizar um único conjunto de primers para capturar a diversidade de uma comunidade inteira. Por outro lado, são as sequências das regiões hipervariáveis que fornecem o poder discriminatório necessário para diferenciar os diversos táxons bacterianos e arqueanos.[1] A coexistência dessas duas características no mesmo gene é o que o consagrou como um marcador tão poderoso e amplamente adotado na microbiologia ecológica e metagenômica. Esta arquitetura molecular permite que, através de um único experimento de amplificação e sequenciamento, os pesquisadores obtenham um perfil taxonômico detalhado de comunidades microbianas complexas, o que é a base da abordagem de metagenômica de amplicons.

1.3. Vantagens do rRNA 16S em Estudos de Comunidades Microbianas

O sequenciamento do gene rRNA 16S oferece diversas vantagens que justificam sua ampla utilização em estudos de comunidades microbianas. Uma das principais é o seu custo-benefício. Comparado ao sequenciamento metagenômico shotgun, que analisa todo o DNA presente na amostra, o sequenciamento de amplicons do rRNA 16S é geralmente mais econômico, especialmente para estudos cujo foco principal é a determinação da composição taxonômica e da diversidade da comunidade.[8] Adicionalmente, requer uma profundidade de sequenciamento por amostra consideravelmente menor para atingir uma boa cobertura da diversidade presente.[8]

Outra vantagem crucial é a capacidade de analisar diretamente amostras ambientais sem a necessidade de cultivo prévio dos microrganismos.[1] Isso é particularmente importante dado que a maioria dos microrganismos não cresce sob as condições de laboratório convencionais, permitindo assim o estudo da vasta diversidade microbiana "não cultivável".[1] A técnica é altamente versátil, sendo aplicável a uma ampla gama de tipos de amostras e ecossistemas, desde o microbioma humano até solos, águas e ambientes extremos.[3]

A existência de extensos bancos de dados de referência, como SILVA, Greengenes e RDP (Ribosomal Database Project), que contêm milhões de sequências de rRNA 16S anotadas taxonomicamente, é outra vantagem significativa.[9] Esses bancos de dados são indispensáveis para a etapa de atribuição taxonômica, permitindo que as sequências obtidas no estudo sejam comparadas e classificadas.

O sucesso e a vasta adoção do sequenciamento de rRNA 16S ao longo das últimas décadas fomentaram um ciclo de desenvolvimento contínuo. A popularidade da técnica impulsionou a criação de ferramentas bioinformáticas especializadas (como QIIME2 e Mothur [5]), o refinamento de protocolos e a constante atualização dos bancos de dados. Esse ecossistema robusto de recursos e conhecimento acumulado torna o sequenciamento de rRNA 16S uma escolha metodológica acessível e bem estabelecida para muitos pesquisadores. No entanto, é importante reconhecer que essa "inércia" positiva, resultante da familiaridade e da infraestrutura desenvolvida, pode, em certas situações, levar à escolha do método 16S por conveniência ou limitação de custos, mesmo quando a pergunta de pesquisa poderia ser mais adequadamente respondida por abordagens mais informativas, como o sequenciamento metagenômico shotgun, que oferece insights funcionais diretos e maior resolução taxonômica.[8] Portanto, a seleção do método de sequenciamento deve ser sempre uma decisão criteriosa, ponderando as vantagens e desvantagens de cada abordagem em relação aos

objetivos específicos do estudo, em vez de seguir cegamente o método mais tradicional ou economicamente viável.

2. Fluxo de Trabalho Experimental em Sequenciamento de rRNA 16S

O fluxo de trabalho experimental para o sequenciamento do gene rRNA 16S envolve uma série de etapas críticas, cada uma com potencial para introduzir vieses ou afetar a qualidade dos dados finais. Uma compreensão detalhada dessas etapas é essencial para o planejamento e execução de estudos robustos.

2.1. Coleta e Preparo de Amostras

A etapa inicial de qualquer estudo metagenômico é a coleta de amostras. A representatividade da amostra em relação ao ambiente ou comunidade microbiana de interesse é crucial, assim como a implementação de medidas rigorosas para prevenir a contaminação por microrganismos exógenos ou DNA ambiental durante a coleta, transporte e armazenamento.[5] Diferentes tipos de amostras, como solo, água, fezes humanas ou biópsias de tecido, exigem protocolos de coleta e preservação específicos para manter a integridade do DNA microbiano e a estrutura original da comunidade.[1] Por exemplo, ao coletar amostras de tecido, como as de mama para estudos de microbioma do câncer, é fundamental minimizar a contaminação e a degradação do DNA, sendo preferíveis amostras frescas em detrimento de amostras fixadas em formalina e emblocadas em parafina (FFPE), que podem apresentar DNA degradado e contaminantes.[11]

Dependendo da matriz da amostra, podem ser necessários pré-tratamentos para remover substâncias que possam inibir a reação de PCR subsequente ou interferir na extração de DNA.[5] A etapa de coleta, embora pareça simples, é frequentemente subestimada. Erros cometidos nesta fase inicial, como contaminação cruzada, amostragem não representativa ou degradação do DNA, podem se propagar por todo o fluxo de trabalho e comprometer a validade de toda a análise subsequente, independentemente do quão sofisticadas sejam as técnicas de sequenciamento e bioinformática empregadas. A contaminação é uma preocupação particularmente crítica ao lidar com amostras de baixa biomassa microbiana, onde o DNA contaminante, proveniente de reagentes, kits de extração ou do ambiente do laboratório, pode se tornar uma proporção significativa do DNA total amplificado.[12] A experiência relatada em um estudo sobre a diversidade microbiana de ateromas, onde a contaminação nos controles negativos levou à necessidade de reiniciar completamente o estudo com novas amostras clínicas, sublinha a importância crítica de protocolos de coleta rigorosos e do uso consistente de controles negativos de coleta e extração para garantir a

confiabilidade dos resultados.[12]

2.2. Extração de DNA Microbiano

Após a coleta, o próximo passo é a extração do DNA total da comunidade microbiana presente na amostra. O objetivo principal desta etapa é lisar eficientemente todas as células microbianas, incluindo aquelas com paredes celulares mais resistentes, como as bactérias Gram-positivas, e isolar DNA de alta qualidade, pureza e integridade, livre de inibidores de PCR.[13] Diversos métodos de extração de DNA são comumente empregados, podendo ser categorizados em mecânicos (por exemplo, "bead beating", que utiliza esferas para romper fisicamente as células), químicos (utilizando detergentes, solventes e agentes caotrópicos para lisar membranas e desnaturar proteínas) e enzimáticos (empregando enzimas como a lisozima para degradar as paredes celulares).[5] Frequentemente, uma combinação desses métodos é utilizada para maximizar a eficiência da lise.

No entanto, a escolha do método de extração de DNA não é trivial e pode introduzir vieses significativos. Um dos desafios mais conhecidos é a lise diferencial de microrganismos. Bactérias Gram-positivas, que possuem uma espessa camada de peptidoglicano em suas paredes celulares, são geralmente mais resistentes à lise do que as bactérias Gram-negativas. Se forem utilizados métodos de lise muito brandos, as Gram-positivas podem ser sub-representadas no DNA extraído, levando a uma visão distorcida da composição da comunidade.[14] Por exemplo, foi observado que o protocolo padrão do Human Microbiome Project (HMP) pode sub-representar Firmicutes, um filo predominantemente Gram-positivo.[14] Por outro lado, métodos mecânicos muito agressivos, como o "bead beating" prolongado ou intenso, podem causar o cisalhamento (fragmentação) do DNA, especialmente de microrganismos mais facilmente lisáveis, o que pode ser problemático se a integridade do DNA de alto peso molecular for necessária para técnicas como o sequenciamento de reads longas.[13] Além disso, a variabilidade na aplicação de métodos mecânicos, como o tempo e a intensidade do "bead beating", pode afetar a reprodutibilidade dos resultados entre amostras ou experimentos.[13]

Reconhecendo esses desafios, pesquisadores têm desenvolvido e otimizado protocolos de extração. Um exemplo é o protocolo 'Rapid', que propõe uma abordagem não mecânica e não enzimática, utilizando uma combinação de tratamento alcalino, calor e detergente para lisar as células.[13] Estudos comparativos indicaram que este método pode fornecer uma representação mais consistente de bactérias de difícil e fácil lise, incluindo uma melhor recuperação de Firmicutes, além de ser mais rápido e adequado para o processamento de

múltiplas amostras.[13]

A escolha do método de extração de DNA, portanto, não é apenas um passo técnico, mas uma decisão que pode introduzir um viés sistemático capaz de alterar fundamentalmente a percepção da composição da comunidade microbiana. O perfil microbiano obtido não reflete a realidade de forma absoluta, mas sim uma aproximação filtrada pelas eficiências e vieses inerentes ao método de extração empregado. Se um método sub-representa consistentemente um grupo bacteriano específico, todas as análises subsequentes, incluindo cálculos de diversidade e identificação de táxons diferencialmente abundantes, serão baseadas nessa visão distorcida. Isso tem implicações diretas na interpretação ecológica ou clínica dos resultados. Consequentemente, é crucial que os estudos comparem ou validem métodos de extração, especialmente ao lidar com novos tipos de amostra ou ao investigar comunidades microbianas diversas. O relato detalhado do método de extração utilizado é essencial para permitir a comparação e a reprodutibilidade entre diferentes estudos, e a conscientização sobre esses vieses é fundamental para a robustez e o avanço do campo da metagenômica.

2.3. Amplificação por PCR do Gene rRNA 16S

Uma vez que o DNA total da comunidade microbiana tenha sido extraído e purificado, a próxima etapa no sequenciamento de amplicons do rRNA 16S é a amplificação seletiva deste gene por meio da Reação em Cadeia da Polimerase (PCR). Este processo utiliza oligonucleotídeos sintéticos, conhecidos como primers (iniciadores), que são desenhados para se anelarem a regiões conservadas do gene rRNA 16S que flanqueiam uma ou mais das suas regiões hipervariáveis.[3] A escolha dos primers e, consequentemente, da(s) região(ões) hipervariável(eis) a ser(em) amplificada(s) é uma das decisões mais críticas no delineamento de um estudo de rRNA 16S, pois impacta diretamente os resultados obtidos.

Diferentes regiões hipervariáveis (V1 a V9) do gene rRNA 16S possuem diferentes graus de variabilidade e, portanto, diferentes capacidades de resolução taxonômica para distintos grupos de microrganismos.[4] A seleção da região alvo depende dos objetivos específicos do estudo, da natureza da comunidade microbiana investigada e das capacidades da plataforma de sequenciamento (por exemplo, o comprimento da read). Primers que amplificam as regiões V3-V4, V4 isoladamente, ou V4-V5 são comumente utilizados em muitos estudos, especialmente com plataformas de sequenciamento de reads curtas como a Illumina.[4] Por exemplo, os primers 515F-806R (que amplificam a região V4) e 515F-926R (que amplificam as regiões V4-V5) são recomendados e amplamente utilizados no âmbito do Earth Microbiome Project (EMP).[15] No entanto, é crucial reconhecer que a escolha do par de primers influencia significativamente a

composição da comunidade microbiana que é detectada e quantificada.[8] Pesquisas comparativas, como a que avaliou os primers 515F-806R e 515F-926R em amostras de solo e os comparou com dados de metagenoma shotgun, demonstraram que diferentes pares de primers podem introduzir vieses taxonômicos específicos, super ou sub-representando determinados grupos bacterianos.[15]

A PCR, embora seja uma técnica poderosa e essencial, é suscetível a vários tipos de vieses que podem distorcer o perfil da comunidade microbiana original. Um dos principais é o viés de amplificação relacionado à eficiência de anelamento dos primers. Pequenas variações na sequência do sítio de ligação do primer no DNA molde de diferentes microrganismos, ou mesmo diferenças na sequência dos próprios primers (no caso de primers degenerados), podem levar a uma amplificação preferencial de certos táxons em detrimento de outros.[8] A concentração do DNA molde na reação de PCR também é um fator crítico. Concentrações muito baixas podem aumentar a variabilidade dos resultados devido a flutuações estocásticas (conhecidas como "drift" da PCR) e tornar a amplificação mais suscetível a contaminantes. Por outro lado, concentrações muito altas de DNA molde podem inibir a reação de PCR ou aumentar a probabilidade de formação de produtos não específicos, como as quimeras.[16] Um estudo sistemático demonstrou que a otimização da concentração do DNA molde é um dos fatores mais importantes para minimizar a variabilidade nos perfis de comunidades microbianas obtidos por sequenciamento de rRNA 16S.[16] O número de ciclos de PCR também deve ser cuidadosamente otimizado; um número excessivo de ciclos pode levar à saturação da reação, à sobrerrepresentação de amplicons que são amplificados mais eficientemente e a um aumento na taxa de erros de polimerização e formação de quimeras.[16] As quimeras são moléculas de DNA híbridas formadas durante a PCR quando uma fita de DNA parcialmente estendida de um molde se anela a um molde diferente e continua a extensão. Essas moléculas artefatuais podem ser erroneamente interpretadas como novos táxons e devem ser identificadas e removidas durante a análise bioinformática. Uma estratégia que tem sido proposta para minimizar o efeito do "drift" da PCR é o pooling (mistura) de várias replicatas de reações de PCR realizadas a partir da mesma amostra de DNA antes do sequenciamento. No entanto, a pesquisa de Berry et al. (2011) sugere que, embora o pooling possa ter algum benefício, seu impacto na redução do viés global é consideravelmente menor do que o da otimização da concentração de DNA molde.[16]

A busca por primers "universais" para o gene rRNA 16S representa um compromisso inerente e complexo. Idealmente, um par de primers universal deveria amplificar o gene 16S de todos os membros de uma comunidade

microbiana com igual eficiência. No entanto, a realidade é que nenhum conjunto de primers é verdadeiramente universal ou completamente livre de viés. A "universalidade" de um primer refere-se à sua capacidade de se anelar a sequências conservadas presentes em uma ampla gama de bactérias e arqueias. Contudo, mesmo pequenas variações nessas regiões consideradas conservadas, ou nas próprias sequências dos primers (especialmente se contiverem bases degeneradas para abranger variações conhecidas), podem afetar drasticamente a afinidade e a eficiência de anelamento a diferentes moldes de DNA. Isso significa que cada estudo de rRNA 16S oferece, na melhor das hipóteses, uma "janela" para a comunidade microbiana, e essa janela é potencialmente enviesada pela escolha dos primers. Alguns primers podem ser mais eficientes para certos grupos taxonômicos ou tipos de ambiente, enquanto sub-representam outros. Isso tem consequências significativas para a meta-análise e a comparação de dados entre estudos que utilizaram diferentes conjuntos de primers. As diferenças observadas nas comunidades microbianas entre tais estudos podem ser, em parte ou totalmente, artefatos metodológicos em vez de reflexos de diferenças biológicas reais. Embora a avaliação *in silico* da cobertura e especificidade dos primers seja uma etapa importante no planejamento, o desempenho *in vitro* em amostras ambientais complexas pode divergir das previsões teóricas, como demonstrado por estudos que comparam resultados de amplicons com dados de metagenoma shotgun.[15] Portanto, a interpretação dos dados de rRNA 16S deve sempre levar em consideração os potenciais vieses introduzidos pela etapa de PCR.

2.4. Preparo de Bibliotecas e Plataformas de Sequenciamento

Após a amplificação por PCR do gene rRNA 16S, os produtos da PCR (amplicons) são purificados para remover primers não incorporados, dNTPs, enzima polimerase e outros componentes da reação. Em seguida, os amplicons purificados são quantificados para garantir que uma quantidade adequada de DNA seja levada para as etapas subsequentes de preparo da biblioteca de sequenciamento.

O preparo de bibliotecas para Sequenciamento de Nova Geração (NGS) é um passo crucial que adapta os amplicons para a plataforma de sequenciamento específica a ser utilizada. Este processo geralmente envolve a adição de sequências adaptadoras nas extremidades dos amplicons. Essas sequências adaptadoras são necessárias para a ligação dos fragmentos de DNA à superfície da célula de fluxo (flow cell) da plataforma de sequenciamento e para a iniciação da reação de sequenciamento. Além dos adaptadores, sequências curtas de DNA conhecidas como índices (ou barcodes) são frequentemente incorporadas.[5] Cada amostra recebe um índice único (ou uma combinação única de índices), o que

permite que múltiplas amostras sejam misturadas e sequenciadas simultaneamente em uma única corrida da plataforma (um processo chamado multiplexing). Após o sequenciamento, os dados brutos são desmultiplexados, ou seja, as reads são separadas e atribuídas à sua amostra de origem com base na sequência do índice. Empresas como a Illumina fornecem kits e protocolos demonstrados especificamente para o preparo de bibliotecas de amplicons do gene rRNA 16S, visando otimizar este processo.[6]

A tecnologia de sequenciamento predominante para estudos de rRNA 16S tem sido, por muitos anos, a da Illumina, baseada no princípio de "sequenciamento por síntese" (SBS). Neste método, os fragmentos de DNA da biblioteca são imobilizados em uma flow cell e amplificados localmente para formar clusters clonais. O sequenciamento ocorre através de ciclos repetidos de incorporação de nucleotídeos marcados com fluoróforos distintos. A cada ciclo de incorporação, a flow cell é imageada, e a identidade do nucleotídeo incorporado em cada cluster é determinada pela cor da fluorescência emitida. Este processo gera milhões de sequências de DNA curtas, tipicamente com comprimentos de 150 a 300 pares de bases (bp) quando operado em modo paired-end (onde ambas as extremidades do fragmento são sequenciadas).[5] Plataformas da Illumina como o MiSeq são frequentemente escolhidas para estudos de rRNA 16S devido à sua relativa rapidez, custo por corrida mais baixo (adequado para um número moderado de amostras) e capacidade de gerar reads com comprimento suficiente para cobrir regiões hipervariáveis como a V4 ou V3-V4.[6] Sistemas de maior capacidade, como o NextSeq, também podem ser utilizados para projetos com um número muito grande de amostras.

Nos últimos anos, as tecnologias de sequenciamento de reads longas, principalmente da Pacific Biosciences (PacBio) e Oxford Nanopore Technologies (ONT), têm emergido como alternativas promissoras para o sequenciamento do gene rRNA 16S.[4] A principal vantagem dessas tecnologias é a capacidade de sequenciar o gene rRNA 16S completo, que tem aproximadamente 1500 bp, ou pelo menos fragmentos muito longos dele. Isso supera uma limitação fundamental das plataformas de reads curtas, que geralmente conseguem sequenciar apenas uma ou duas regiões hipervariáveis adjacentes. Ao obter a sequência do gene inteiro, espera-se uma melhoria significativa na resolução taxonômica, potencialmente permitindo a identificação de microrganismos em nível de espécie ou até mesmo de cepa com maior confiança e precisão.[4]

A tecnologia da Oxford Nanopore (ONT) baseia-se na passagem de moléculas de DNA de fita simples através de nanoporos proteicos embebidos em uma membrana. À medida que o DNA atravessa o poro, ele causa uma alteração

característica na corrente iônica que flui através do poro. Essa alteração é medida em tempo real e decodificada para determinar a sequência de bases. A ONT é conhecida por gerar reads extremamente longas (dezenas a centenas de milhares de bases), pela portabilidade de seus dispositivos (como o MinION) e pela capacidade de análise de dados em tempo real.[17] Tem se mostrado particularmente promissora para aplicações em diagnóstico clínico rápido, onde a velocidade e a capacidade de identificar patógenos em amostras complexas são cruciais.[17] A tecnologia SMRT (Single Molecule, Real-Time) da PacBio também produz reads longas, e suas mais recentes iterações (HiFi reads) oferecem alta acurácia. No entanto, tradicionalmente, o custo do sequenciamento PacBio tem sido mais elevado, embora isso esteja mudando com as novas plataformas.[4]

Apesar das vantagens, as tecnologias de reads longas também enfrentam desafios. Historicamente, elas apresentavam taxas de erro intrínsecas mais altas por base individual em comparação com a Illumina, especialmente para a ONT, embora a acurácia tenha melhorado drasticamente com o desenvolvimento de novos poros, químicas e algoritmos de processamento de sinal.[4] O custo por base sequenciada também pode ser um fator, e os pipelines bioinformáticos para análise de dados de reads longas de 16S ainda estão em um estágio de desenvolvimento e padronização menos maduro em comparação com os fluxos de trabalho bem estabelecidos para dados da Illumina. No entanto, trabalhos recentes destacam que o custo e a praticidade da tecnologia ONT estão se tornando cada vez mais viáveis, inclusive para laboratórios clínicos de rotina.[17]

A transição para o sequenciamento de reads longas para a análise do gene rRNA 16S completo pode ser vista como uma potencial "revolução dentro da revolução" da metagenômica de amplicons. Ao superar a limitação fundamental da cobertura parcial do gene imposta pelas reads curtas, essa abordagem promete revitalizar o gene 16S como uma ferramenta taxonômica ainda mais poderosa. A capacidade de analisar a sequência completa do gene, que contém múltiplas regiões variáveis e, portanto, mais informação filogenética, deve, teoricamente, fornecer uma resolução taxonômica significativamente melhorada, aproximando o poder discriminatório do 16S ao do sequenciamento metagenômico shotgun para fins de identificação taxonômica, mas com um fluxo de trabalho potencialmente mais simples e focado no marcador. Isso é de particular relevância para aplicações onde a identificação precisa em nível de espécie ou cepa é crítica, como no diagnóstico de infecções ou na ecologia microbiana fina.[17] Se as tecnologias de reads longas continuarem a evoluir em termos de acurácia, custo-benefício e robustez dos pipelines bioinformáticos, elas poderão oferecer o melhor de dois mundos: a simplicidade e o foco do marcador 16S com uma capacidade de resolução taxonômica que rivaliza com abordagens mais complexas.

A tabela a seguir resume e compara as principais plataformas de sequenciamento utilizadas para a análise do gene rRNA 16S:

Tabela 1: Comparativo de Plataformas de Sequenciamento para rRNA 16S

Característica	Illumina MiSeq/NextSeq	PacBio Sequel/Revio	Oxford Nanopore MinION/GridION
Princípio de Sequenciamento	Sequenciamento por Síntese (SBS)	Sequenciamento de Molécula Única em Tempo Real (SMRT)	Detecção de alteração de corrente iônica por nanoporo
Comprimento Típico da Read	150-300 bp (paired-end)	10-25 kb (HiFi reads >99.9% acurácia)	>10 kb (pode chegar a Mb), acurácia variável (melhorando)
Capacidade de Sequenciar Gene 16S Completo	Não (geralmente regiões parciais, e.g., V4, V3-V4)	Sim	Sim
Taxa de Erro Típica (por base individual)	Muito baixa (~0.1%)	Muito baixa para HiFi reads (<0.1%), maior para CLR	Moderada a alta, mas melhorando (R10.4.1 flow cells >Q20)
Custo Relativo (por amostra para 16S)	Baixo a moderado	Moderado a alto	Baixo a moderado (especialmente MinION para menor escala)
Vantagens Principais para 16S	Alta acurácia, pipelines estabelecidos, alto throughput	Sequenciamento do gene completo, alta acurácia (HiFi)	Sequenciamento do gene completo, portabilidade, tempo real
Desvantagens Principais para 16S	Reads curtas limitam resolução taxonômica	Custo historicamente mais alto, menor throughput por corrida	Taxa de erro historicamente mais alta, pipelines em evolução

3. Análise Bioinformática de Dados de rRNA 16S

Após a etapa de sequenciamento, os dados brutos, geralmente na forma de arquivos FASTQ, passam por um extenso processo de análise bioinformática. Este processo é fundamental para transformar as sequências de nucleotídeos em informações biológicas significativas sobre a composição e a diversidade das comunidades microbianas.

3.1. Pré-processamento e Controle de Qualidade dos Dados Brutos

O objetivo inicial da análise bioinformática é limpar os dados brutos, removendo sequências de baixa qualidade, artefatos de sequenciamento e sequências não biológicas (como adaptadores e primers), a fim de preparar os dados para as análises taxonômicas e de diversidade subsequentes.

As principais etapas incluem:

- **Avaliação da Qualidade das Reads:** Ferramentas como FastQC são utilizadas para gerar estatísticas sobre a qualidade das sequências brutas, incluindo a distribuição dos scores de qualidade por base ao longo das reads, conteúdo de GC, presença de sequências super-representadas e adaptadores residuais.[20] Essa avaliação inicial ajuda a identificar potenciais problemas com a corrida de sequenciamento e a guiar as etapas de filtragem.
- **Remoção de Adaptadores e Primers:** Sequências de adaptadores e primers que podem ter sido sequenciadas junto com o inserto biológico devem ser removidas. Ferramentas como Cutadapt [20] ou Trimmomatic são comumente usadas para esta finalidade. Um exemplo prático usando qiime cutadapt trim-paired dentro do ambiente QIIME2 é fornecido em.[21]
- **Filtragem por Qualidade (Quality Trimming/Filtering):** Bases com scores de qualidade Phred baixos, que indicam uma maior probabilidade de erro na chamada da base, são tipicamente removidas das extremidades das reads (trimming) ou reads inteiras de baixa qualidade podem ser descartadas (filtering).[5] A escolha dos limiares de truncamento (onde cortar as reads) é uma etapa crucial e pode depender da qualidade geral dos dados e da plataforma de sequenciamento. Um truncamento muito agressivo pode levar à perda de informação biológica, enquanto um truncamento insuficiente pode reter erros que afetam as análises subsequentes.[20]
- **União de Reads Pareadas (Paired-End Read Merging):** Para dados de sequenciamento paired-end, como os gerados pela Illumina, as reads forward (R1) e reverse (R2) que se originam do mesmo fragmento de DNA e se sobrepõem são unidas para reconstruir uma sequência consenso mais longa que cobre a região hipervariável de interesse de forma mais completa.[20]

Ferramentas como PEAR, VSEARCH, o script fastq_mergepairs do pacote USEARCH, ou funcionalidades integradas em pipelines como DADA2, são usadas para esta etapa. A eficiência e a acurácia da união dependem da qualidade das reads e do comprimento e qualidade da região de sobreposição.

Um pré-processamento rigoroso e cuidadoso é absolutamente fundamental. A propagação de erros de sequenciamento, sequências quiméricas, ou a presença residual de adaptadores e primers podem levar a uma inflação artificial da diversidade observada e à identificação de unidades taxonômicas espúrias (OTUs ou ASVs que não existem biologicamente). A qualidade dos dados que entram nos algoritmos de inferência taxonômica é diretamente proporcional à confiabilidade dos resultados finais; um princípio bem resumido pela máxima "garbage in, garbage out". Investir tempo na otimização e na verificação minuciosa das etapas de pré-processamento, incluindo o uso de controles positivos (comunidades mock) e negativos, é crucial para garantir a validade das conclusões biológicas extraídas do estudo. A remoção incorreta de adaptadores ou um truncamento inadequado das reads, por exemplo, pode levar à perda de sequências válidas ou influenciar negativamente a performance dos algoritmos de clustering ou denoising.[20]

3.2. Geração de Unidades Taxonômicas

Após o pré-processamento, o próximo passo é agrupar as sequências limpas em unidades que representem, da forma mais acurada possível, os diferentes tipos de microrganismos presentes nas amostras. O objetivo é distinguir as variações biológicas reais das variações introduzidas por erros de sequenciamento e PCR. Duas abordagens principais são utilizadas: o agrupamento em Unidades Taxonômicas Operacionais (OTUs) e a geração de Variantes de Sequência de Amplicons (ASVs).

- Agrupamento em Unidades Taxonômicas Operacionais (OTUs):
  Este é o método tradicional, que consiste em agrupar sequências de DNA com base em um limiar de similaridade de sequência predefinido, comumente 97% para o gene rRNA 16S.20 A ideia é que sequências dentro deste limiar provavelmente pertencem à mesma "espécie" ou a um grupo taxonômico proximamente relacionado. Diversos algoritmos foram desenvolvidos para o clustering de OTUs, como UCLUST (frequentemente usado no pipeline QIIME1 20), UPARSE 24 e CD-HIT.

  No entanto, a abordagem de OTUs possui várias limitações. O limiar de 97% de similaridade é, em grande parte, arbitrário e pode não corresponder

consistentemente a espécies biológicas reais em todos os grupos taxonômicos; algumas espécies podem ter genes 16S mais similares que 97%, enquanto outras, mais distantes, podem ser agrupadas indevidamente.

Além disso, os métodos de clustering de OTUs tendem a agrupar erros de sequenciamento junto com as sequências biológicas corretas, o que pode inflar artificialmente as estimativas de diversidade. A escolha da sequência representativa (centroide) para cada cluster de OTU também pode ser problemática e afetar a atribuição taxonômica subsequente. Finalmente, os clusters de OTUs são definidos dentro do contexto de um único estudo ou conjunto de dados, o que dificulta a comparação direta e a reprodutibilidade entre diferentes estudos.

- Geração de Variantes de Sequência de Amplicons (ASVs):
  Uma abordagem mais recente e cada vez mais adotada é a geração de ASVs, também conhecidas como "zero-radius OTUs" (zOTUs) ou "Exact Sequence Variants" (ESVs). Em vez de agrupar sequências com base em um limiar de similaridade, os métodos de ASV visam resolver sequências únicas com precisão de um único nucleotídeo, após modelar e remover estatisticamente os erros de sequenciamento e PCR.20 Os algoritmos mais proeminentes para a geração de ASVs incluem DADA2 20, Deblur 22 e UNOISE3 (parte do pacote USEARCH).

  As ASVs oferecem várias vantagens sobre as OTUs. Elas proporcionam uma resolução taxonômica potencialmente maior, permitindo a distinção entre variantes de sequências que diferem por apenas algumas bases. São mais eficazes na diferenciação de sequências raras e na remoção de contaminantes e sequências espúrias. Crucialmente, as ASVs são consideradas rótulos biológicos mais estáveis e consistentes, o que aumenta significativamente a reprodutibilidade e a comparabilidade dos resultados entre diferentes estudos e laboratórios, pois a mesma sequência biológica resultará na mesma ASV, independentemente do conjunto de dados em que é encontrada. O algoritmo DADA2, em particular, tem se mostrado eficaz no processamento de reads de baixa abundância e na inferência de sequências biológicas exatas. Estudos comparativos, como o apresentado em 22 e 22, indicam que, embora o número absoluto de ASVs ou OTUs e alguns índices de diversidade alfa possam variar entre os métodos, os perfis taxonômicos gerais e as conclusões biológicas (por exemplo, em estudos de diagnóstico de doenças) podem ser semelhantes. No entanto, DADA2 demonstrou oferecer a melhor sensibilidade na detecção de variantes reais em algumas avaliações.

A transição da abordagem de OTUs para a de ASVs representa um avanço metodológico significativo na busca por uma representação mais fiel e precisa da diversidade microbiana. Essa mudança reflete um movimento de uma heurística (o limiar de similaridade de 97%) para um modelo estatístico mais sofisticado que leva em consideração os padrões de erro específicos da tecnologia de sequenciamento utilizada. Isso não apenas melhora a acurácia e a resolução dos estudos de microbioma, mas também tem implicações profundas para a capacidade do campo de realizar meta-análises robustas e construir conhecimento cumulativo. Como as ASVs são, em teoria, as sequências biológicas exatas inferidas que estavam presentes na amostra original, elas funcionam como unidades biológicas mais estáveis e universalmente comparáveis. Isso facilita a criação de bancos de dados de ASVs com anotações taxonômicas consistentes e permite o rastreamento da distribuição e da dinâmica de "espécies genéticas" específicas através de diferentes ambientes, hospedeiros ou condições de doença com uma precisão sem precedentes.

A tabela abaixo resume as principais diferenças entre as abordagens de OTU e ASV:

Tabela 2: Comparativo entre Abordagens de OTU e ASV

Característica	Agrupamento em OTUs	Geração de ASVs
Método Principal	Clustering baseado em limiar de similaridade (e.g., 97%)	Denoising baseado em modelo de erro de sequenciamento
Unidade Gerada	Unidade Taxonômica Operacional (OTU)	Variante de Sequência de Amplicon (ASV) / Exact Sequence Variant (ESV)
Resolução	Limitada pelo limiar; agrupa variantes próximas	Precisão de um único nucleotídeo; distingue variantes sutis
Tratamento de Erros de Seq.	Agrupa erros com sequências reais; pode inflar diversidade	Modela e remove erros; visa inferir sequências biológicas exatas
Reprodutibilidade	Menor; OTUs são	Maior; ASVs são rótulos

Inter-estudos	dependentes do conjunto de dados	biológicos consistentes
Vantagens Principais	Método tradicional, ferramentas estabelecidas	Maior resolução, melhor tratamento de erros, maior reprodutibilidade
Desvantagens Principais	Limiar arbitrário, menor resolução, inflação de diversidade	Computacionalmente mais intensivo (alguns métodos), sensível a parâmetros
Exemplos de Ferramentas	UCLUST, UPARSE, CD-HIT	DADA2, Deblur, UNOISE3

3.3. Atribuição Taxonômica

Uma vez que as sequências foram agrupadas em OTUs ou processadas em ASVs, o próximo passo fundamental é atribuir uma identidade taxonômica a cada uma dessas unidades. Este processo visa classificar cada OTU/ASV dentro da hierarquia taxonômica padrão (Reino, Filo, Classe, Ordem, Família, Gênero e, idealmente, Espécie), comparando suas sequências com as de microrganismos conhecidos e classificados presentes em bancos de dados de referência.

- Uso de Bancos de Dados de Referência:
  A acurácia da atribuição taxonômica depende criticamente da qualidade, abrangência e atualização do banco de dados de referência utilizado. Os principais bancos de dados para o gene rRNA 16S incluem:
  - **SILVA:** É um banco de dados abrangente, cuidadosamente curado e regularmente atualizado, que inclui sequências de rRNA de subunidade pequena (SSU) e grande (LSU) de todos os três domínios da vida (Bacteria, Archaea, Eukarya). É amplamente considerado um dos bancos de dados mais acurados e robustos para a classificação taxonômica baseada no 16S rRNA.[9]
  - **Greengenes (GG):** Foi um banco de dados historicamente popular, especialmente nos primórdios do QIIME. No entanto, sua última atualização ocorreu em 2013 [20], o que levanta sérias preocupações sobre sua relevância e utilidade atuais, dado o rápido crescimento do conhecimento sobre a diversidade microbiana. Estudos comparativos mais recentes indicam que o Greengenes tende a ter um desempenho inferior a outros bancos de dados, com uma menor proporção de sequências anotadas em nível de espécie e falhas na classificação de diversos

gêneros.[10]

- **RDP (Ribosomal Database Project):** É outro banco de dados bem estabelecido e conceituado, focado especificamente em sequências de rRNA. Oferece ferramentas para alinhamento e classificação, e sua acurácia é geralmente comparável à do SILVA.[9]
- **Outros bancos de dados:** Incluem o banco de dados de nucleotídeos do NCBI (GenBank), que é vasto, mas menos curado especificamente para taxonomia de rRNA; o LTP (All-Species Living Tree Project) [20]; e bancos de dados integrados como o 16S-ITGDB, que visa combinar informações de SILVA, RDP e Greengenes.[9] É fundamental ressaltar que a escolha do banco de dados e da sua versão específica pode impactar significativamente os resultados da classificação taxonômica. Diferentes bancos de dados podem usar taxonomias ligeiramente diferentes ou ter coberturas distintas de certos grupos microbianos.[20]

- Ferramentas e Algoritmos de Classificação:
  Diversos algoritmos e ferramentas são empregados para realizar a atribuição taxonômica. Os mais comuns incluem:
  - **Classificadores Naive Bayes:** Estes são algoritmos de aprendizado de máquina que calculam a probabilidade de uma sequência pertencer a cada táxon no banco de dados de referência, com base na frequência de "palavras" de k-mers (subsequências de comprimento k) na sequência query e nas sequências de referência. O classificador Naive Bayes implementado no RDP e o q2-feature-classifier classify-sklearn do QIIME2 são exemplos populares.[21] Esses classificadores geralmente precisam ser treinados em sequências de referência que correspondem à região hipervariável específica do 16S rRNA que foi sequenciada no estudo, para otimizar a acurácia.[20]
  - **Métodos Baseados em Alinhamento:** Ferramentas como BLAST (Basic Local Alignment Search Tool), VSEARCH ou USEARCH podem ser usadas para alinhar as sequências query (OTUs/ASVs) contra as sequências do banco de dados de referência. A taxonomia é então atribuída com base na melhor correspondência (hit) que atende a certos critérios de similaridade e cobertura do alinhamento. Pipelines integrados como QIIME2 [21] e Mothur [5] oferecem módulos e fluxos de trabalho que facilitam a atribuição taxonômica usando diferentes algoritmos e bancos de dados.
- Desafios na Atribuição Taxonômica:
  Um desafio comum é a ocorrência de sequências que não podem ser atribuídas a nenhum táxon conhecido em um determinado nível hierárquico (por exemplo, gênero ou espécie), ou mesmo em níveis mais altos como filo. Essas são frequentemente rotuladas como "unassigned", "unclassified" ou

"Bacteria;Other". As razões para isso podem ser várias 26:

- A sequência pode pertencer a um microrganismo novo ou ainda não caracterizado, cuja sequência de 16S rRNA não está presente (ou não está corretamente anotada) no banco de dados de referência utilizado.
- A qualidade da sequência query pode ser baixa, ou o fragmento sequenciado pode ser muito curto ou pouco informativo para permitir uma classificação confiável.
- A sequência pode ser uma quimera ou outro tipo de artefato de PCR ou sequenciamento. Uma pequena proporção de sequências não atribuídas é geralmente considerada normal, especialmente ao investigar ambientes microbianos pouco explorados ou diversos.[26]

A atribuição taxonômica, portanto, não deve ser vista como uma verdade absoluta, mas sim como uma hipótese científica baseada na similaridade da sequência observada com as sequências de referência conhecidas. A qualidade dessa hipótese é intrinsecamente limitada pela completude, acurácia e atualização do banco de dados de referência escolhido. Um rótulo como "Unassigned Bacteria" ou "Bacteria;Firmicutes;Clostridia;Clostridiales;[Eubacterium] coprostanoligenes group;g__uncultured" não significa necessariamente que o microrganismo correspondente não exista ou não seja biologicamente importante. Pelo contrário, pode representar uma linhagem completamente nova, uma variação de uma sequência conhecida que não atinge o limiar de confiança para classificação em níveis taxonômicos mais baixos com o banco de dados e o classificador atuais, ou simplesmente uma lacuna no nosso conhecimento taxonômico. É importante não descartar automaticamente as sequências não atribuídas, especialmente se forem abundantes ou mostrarem padrões de abundância diferencial interessantes entre os grupos de amostras. Ferramentas como o BLAST, quando usadas para comparar essas sequências contra bancos de dados mais amplos e abrangentes (como o NCBI nt/nr), podem, por vezes, fornecer pistas adicionais sobre sua possível identidade ou afiliação filogenética.[26] A curadoria e a atualização contínua dos bancos de dados de referência, com a incorporação de novas sequências de genomas e de microrganismos não cultiváveis, são cruciais para melhorar a precisão e a profundidade da classificação taxonômica.

A tabela a seguir oferece uma visão geral dos principais bancos de dados de referência para o gene rRNA 16S:

Tabela 3: Visão Geral dos Principais Bancos de Dados de Referência para rRNA 16S

Nome do Banco de Dados	Foco Principal/Características	Frequência de Atualização	Vantagens	Desvantagens/Limitações	URL/Acesso (Exemplo)
SILVA	rRNA (SSU/LSU) de alta qualidade, curado, abrangente (todos domínios)	Regular (semestral/anual)	Alta acurácia, boa cobertura, taxonomia consistente, ferramentas online	Pode ser computacionalmente intensivo para algumas tarefas	www.arb-silva.de
Greengenes (GG)	rRNA 16S/18S, historicamente usado com QIIME	Desatualizado (desde 2013)	Legado em estudos mais antigos	Desatualizado, menor cobertura de espécies, menor acurácia [20]	greengenes.secondgenome.com
RDP	rRNA, focado em classificação e análise filogenética	Regular	Boas ferramentas de classificação, taxonomia bem estabelecida	Cobertura pode ser menor que SILVA para alguns grupos	rdp.cme.msu.edu
NCBI GenBank	Repositório geral de sequências de nucleotídeos	Contínua (diária)	Maior banco de dados, inclui sequências de projetos genômicos diversos	Menos curado para taxonomia de 16S, pode conter erros/redundâncias	www.ncbi.nlm.nih.gov/genbank/

3.4. Análise de Diversidade Microbiana

Após a geração das unidades taxonômicas (OTUs/ASVs) e sua respectiva atribuição taxonômica, a análise de diversidade microbiana é realizada para

quantificar e comparar a riqueza (número de tipos diferentes de microrganismos), a uniformidade (distribuição das abundâncias relativas dos diferentes tipos) e a composição geral das comunidades microbianas, tanto dentro de cada amostra (diversidade alfa) quanto entre diferentes amostras ou grupos de amostras (diversidade beta).

- Diversidade Alfa (within-sample diversity):
  A diversidade alfa descreve a complexidade de uma comunidade microbiana dentro de uma única amostra. Diversas métricas são usadas para capturar diferentes aspectos da diversidade alfa:
  - **Métricas de Riqueza:** Estas métricas quantificam o número de diferentes táxons (OTUs ou ASVs) presentes em uma amostra.
    - *Observed OTUs/ASVs (Riqueza Observada):* Simplesmente a contagem do número de OTUs/ASVs únicos detectados na amostra.[28]
    - *Chao1 e ACE (Abundance-based Coverage Estimator):* São estimadores não paramétricos que tentam inferir a riqueza total de táxons na comunidade original, incluindo aqueles que podem não ter sido detectados devido à profundidade de sequenciamento insuficiente. Eles fazem isso dando um peso maior aos táxons raros, especialmente aqueles observados apenas uma vez (singletons) ou duas vezes (doubletons) na amostra.[27] É importante notar que ferramentas de denoising como DADA2 removem singletons por padrão, o que pode tornar métricas como Chao1 inadequadas ou exigir ajustes na interpretação se os singletons não forem considerados.[28] Em contraste, DEBLUR retém singletons, permitindo o cálculo dessas métricas.[28]
  - **Métricas de Uniformidade (Evenness):** Estas métricas avaliam quão equitativamente as abundâncias dos diferentes táxons estão distribuídas dentro da amostra. Uma alta uniformidade significa que muitos táxons estão presentes em abundâncias semelhantes, enquanto uma baixa uniformidade indica que alguns poucos táxons dominam a comunidade.
    - *Índice de Simpson (ou seu inverso/complemento, como Gini-Simpson):* Mede a probabilidade de que dois indivíduos retirados aleatoriamente da amostra pertençam ao mesmo táxon. Um valor mais alto do índice de Simpson original indica menor diversidade (maior dominância).[27]
    - *Índice de Pielou (Pielou's Evenness):* Compara a diversidade observada com a diversidade máxima possível, dado o número de táxons observados.
  - **Métricas Combinadas (Riqueza e Uniformidade):**
    - *Índice de Shannon (ou Entropia de Shannon):* É uma métrica popular que leva em consideração tanto o número de táxons (riqueza) quanto a

equitabilidade de suas abundâncias (uniformidade). Valores mais altos geralmente indicam maior diversidade.[27]

- **Métricas Filogenéticas:**
  - *Faith's Phylogenetic Diversity (PD):* Esta métrica incorpora as relações evolutivas (filogenéticas) entre os táxons presentes na amostra. É calculada somando os comprimentos dos ramos da árvore filogenética que conectam todos os táxons observados na amostra. Uma maior PD indica que a comunidade é composta por táxons mais distantemente relacionados filogeneticamente.[27] A prática de rarefação, que envolve a subamostragem aleatória das sequências de cada amostra para uma profundidade de sequenciamento igual antes de calcular a diversidade alfa, tem sido tradicionalmente usada para permitir comparações justas entre amostras com diferentes números totais de reads. No entanto, com o advento das ASVs e métodos de denoising que lidam com a profundidade de sequenciamento de maneiras diferentes, o uso e a necessidade da rarefação são temas de debate na comunidade científica.[29]

- Diversidade Beta (between-sample diversity):
  A diversidade beta mede o grau de dissimilaridade ou similaridade na composição de espécies (ou OTUs/ASVs) entre diferentes amostras ou grupos de amostras. Ela ajuda a responder perguntas como: "Quão diferentes são as comunidades microbianas entre o grupo de tratamento e o grupo controle?".
  - **Métricas Baseadas em Presença/Ausência (Qualitativas):**
    - *Índice de Jaccard (ou Distância de Jaccard):* Mede a dissimilaridade com base apenas na presença ou ausência de táxons compartilhados e únicos entre duas amostras, sem considerar suas abundâncias relativas.[27]
  - **Métricas Baseadas em Abundância (Quantitativas):**
    - *Dissimilaridade de Bray-Curtis:* É uma das métricas mais comuns e mede a dissimilaridade com base nas abundâncias relativas dos táxons compartilhados e não compartilhados entre duas amostras. É geralmente mais sensível a mudanças na abundância dos táxons dominantes.[27]
  - **Métricas Filogenéticas:**
    - *Distâncias UniFrac (Ponderada e Não Ponderada):* Estas métricas incorporam informações filogenéticas ao calcular a dissimilaridade entre comunidades.
      - *UniFrac Não Ponderado:* Considera a fração do comprimento total dos ramos da árvore filogenética que leva a descendentes presentes em apenas uma das duas amostras comparadas (ou

seja, linhagens únicas para cada amostra). É mais sensível a mudanças na composição de linhagens raras ou filogeneticamente distintas.[27]

- *UniFrac Ponderado:* Similar ao não ponderado, mas também leva em conta as abundâncias relativas dos táxons ao ponderar os comprimentos dos ramos. É mais influenciado por mudanças na abundância das linhagens filogenéticas dominantes.[27] Após o cálculo das matrizes de dissimilaridade/distância entre todas as amostras, métodos de ordenação multivariada são usados para visualizar essas relações em um espaço de baixa dimensão (geralmente 2D ou 3D). Os mais comuns incluem:

- *Análise de Coordenadas Principais (PCoA):* É um método de ordenação baseado em qualquer matriz de dissimilaridade/distância.
- *Escalonamento Multidimensional Não Métrico (NMDS):* Tenta preservar as ordenações de dissimilaridade entre as amostras, em vez das distâncias exatas.
- A Análise de Componentes Principais (PCA) é menos adequada para dados de microbioma, que são frequentemente esparsos e não seguem uma distribuição normal, em comparação com a PCoA.[30] Para testar estatisticamente se existem diferenças significativas na composição global da comunidade microbiana entre grupos predefinidos de amostras (por exemplo, doentes vs. saudáveis), são utilizados testes baseados em permutação, como PERMANOVA (Permutational Multivariate Analysis of Variance) e ANOSIM (Analysis of Similarities).[27]

A escolha da(s) métrica(s) de diversidade a ser(em) utilizada(s) em um estudo não é uma decisão trivial e pode ter um impacto substancial nas conclusões biológicas alcançadas. Diferentes métricas são projetadas para capturar aspectos distintos da estrutura da comunidade – algumas enfatizam a riqueza, outras a uniformidade, algumas incorporam relações filogenéticas, e outras focam em presença/ausência versus abundância. Portanto, a "sensibilidade" de uma métrica para detectar uma diferença real entre comunidades pode depender da natureza exata dessa diferença.[27] Por exemplo, se a principal diferença entre duas comunidades reside na presença ou ausência de muitos táxons raros, uma métrica como a distância de Jaccard (para diversidade beta) ou o número de ASVs observadas (para diversidade alfa) pode revelar uma grande diferença. Contudo, se a diferença é primariamente nas abundâncias relativas dos mesmos táxons dominantes, a dissimilaridade de Bray-Curtis (beta) ou os índices de Shannon/Simpson (alfa) podem ser mais informativos. Não existe uma "melhor" métrica universal; a escolha deve ser guiada pela pergunta de

pesquisa específica e pela natureza esperada das diferenças biológicas. Utilizar e relatar múltiplas métricas de diversidade, ou justificar cuidadosamente a escolha de uma métrica específica, é considerado uma boa prática científica. Além disso, como sugerido em [27] e [27], a realização de cálculos de poder estatístico para diferentes métricas antes do início do estudo pode ajudar a planejar experimentos com tamanho amostral adequado e a evitar a prática questionável de "p-hacking" (testar múltiplas métricas até que uma delas produza um resultado estatisticamente significativo).

A tabela a seguir resume as principais métricas de diversidade alfa e beta:

Tabela 4: Resumo das Principais Métricas de Diversidade Alfa e Beta

Categoria	Métrica	O que Mede	Principal Interpretação/Sensibilidade
Alfa	Chao1	Riqueza estimada (nº de táxons, incluindo não observados)	Sensível a táxons raros (singletons, doubletons); afetada pela remoção de singletons (e.g., por DADA2)
	Shannon	Riqueza e Uniformidade combinadas	Aumenta com maior nº de táxons e distribuição mais uniforme de suas abundâncias
	Simpson (Índice de Gini-Simpson)	Uniformidade/Dominância (probabilidade de 2 indivíduos serem de táxons ≠)	Aumenta com maior uniformidade (menor dominância por poucos táxons)
	Faith's PD	Diversidade Filogenética (soma dos comprimentos dos ramos da árvore)	Reflete a extensão da história evolutiva representada na amostra
Beta	Bray-Curtis	Dissimilaridade baseada em	Sensível a mudanças na abundância dos

	Dissimilarity	abundância	táxons, especialmente os mais comuns
	Jaccard Distance	Dissimilaridade baseada em presença/ausência	Sensível a diferenças na composição (quais táxons estão presentes), independentemente da abundância
	UniFrac Ponderado	Dissimilaridade filogenética baseada em abundância	Sensível a mudanças na abundância de linhagens filogenéticas
	UniFrac Não Ponderado	Dissimilaridade filogenética baseada em presença/ausência	Sensível a mudanças na presença/ausência de linhagens filogenéticas, mesmo que raras

3.5. Análises Estatísticas Avançadas

Além das análises de diversidade alfa e beta, que fornecem uma visão global da estrutura da comunidade, diversas análises estatísticas mais avançadas podem ser aplicadas para responder a perguntas de pesquisa mais específicas.

- Identificação de Táxons Diferencialmente Abundantes:
  Um objetivo comum em muitos estudos de microbioma é identificar quais táxons específicos (sejam ASVs, gêneros, famílias, etc.) apresentam diferenças estatisticamente significativas em sua abundância relativa entre diferentes grupos de amostras (por exemplo, indivíduos doentes versus saudáveis, diferentes condições de tratamento, ou amostras de ambientes distintos). Dada a natureza particular dos dados de sequenciamento de microbioma – que são composicionais (as abundâncias são relativas e somam 100% para cada amostra), esparsos (contêm muitos zeros, especialmente para táxons raros) e frequentemente sobredispersos (a variância é maior que a média) – o uso de testes estatísticos padrão como o teste t ou ANOVA não é apropriado. Em vez disso, métodos estatísticos especializados foram desenvolvidos ou adaptados para lidar com essas características.
  Algumas das ferramentas e abordagens mais utilizadas incluem:

- **DESeq2** e **edgeR:** Originalmente desenvolvidos para análise de expressão diferencial em dados de RNA-Seq (contagens de reads), esses pacotes do R foram adaptados e são frequentemente usados para dados de microbioma, modelando as contagens com distribuições binomial negativas.[31]
- **ANCOM (Analysis of Compositions of Microbiomes)** e sua versão mais recente **ANCOM-BC (Bias Correction):** São métodos projetados especificamente para dados composicionais, que utilizam transformações log-ratio para identificar táxons diferencialmente abundantes em relação à média geométrica dos demais táxons.[31]
- **LEfSe (Linear discriminant analysis Effect Size):** Combina testes estatísticos padrão (como Kruskal-Wallis e Wilcoxon) com análise de discriminante linear (LDA) para identificar táxons e funções que são diferencialmente abundantes e biologicamente consistentes entre grupos.[31]
- **ALDEx2 (ANOVA-Like Differential Expression tool for compositional data):** Outra ferramenta que lida com a composicionalidade através de transformações log-ratio e inferência estatística baseada em amostragens de Monte Carlo.[31]
- Outras ferramentas mencionadas incluem **MaAsLin2 (Microbiome Multivariable Association with Linear Models)**, **SIAMCAT (Statistical Inference of Associations between Microbial Communities And host phenoTypes)** e **LinDA (Linear Discriminant Analysis for Differential Abundance)**.[31] É importante notar que diferentes ferramentas de análise de abundância diferencial podem produzir resultados variados quando aplicadas ao mesmo conjunto de dados, devido a diferentes suposições estatísticas e métodos de normalização. Portanto, alguns pesquisadores recomendam o uso de múltiplas ferramentas e a busca por um consenso nos resultados, ou a escolha da ferramenta mais apropriada com base nas características específicas dos dados e da pergunta de pesquisa.[31]

- Análise de Correlação com Variáveis Ambientais ou Clínicas:
  Frequentemente, os pesquisadores estão interessados em investigar como a composição da comunidade microbiana como um todo, ou a abundância de táxons específicos, se correlaciona com variáveis ambientais contínuas ou categóricas (como pH, temperatura, concentração de nutrientes, níveis de poluentes) ou com variáveis do hospedeiro (como idade, índice de massa corporal, dieta, parâmetros clínicos, ou mesmo outros dados ômicos).
  Diversos métodos estatísticos podem ser empregados para este fim:
  - **Testes de Correlação Simples:** Coeficientes de correlação de Spearman (não paramétrico, baseado em ranks) ou Pearson (paramétrico, assume

linearidade e normalidade) podem ser usados para avaliar a associação entre a abundância de um táxon e uma variável contínua. No entanto, devem ser usados com cautela devido à natureza composicional dos dados de abundância relativa; transformações como a log-ratio centrada (CLR) podem ser aplicadas antes do cálculo da correlação.

- **Testes de Mantel:** Usados para testar a correlação entre duas matrizes de distância/dissimilaridade. Por exemplo, pode-se testar se a dissimilaridade na composição da comunidade microbiana (calculada com Bray-Curtis ou UniFrac) está correlacionada com a dissimilaridade em um conjunto de variáveis ambientais entre as amostras.[33]
- **Análises de Ordenação Restrita (Constrained Ordination):** Métodos como a Análise de Redundância (RDA) – para dados que podem ser aproximados por um modelo linear – ou a Análise de Correspondência Canônica (CCA) – para dados com uma resposta unimodal às variáveis ambientais – são poderosos para investigar como um conjunto de variáveis ambientais explica a variação na composição da comunidade microbiana.
- **Modelos Lineares Generalizados (GLMs) e Modelos Lineares Mistos (GLMMs):** Podem ser usados para modelar a abundância de táxons específicos em função de múltiplas variáveis preditoras, controlando por fatores de confusão e levando em conta a estrutura dos dados (por exemplo, amostras repetidas do mesmo indivíduo).
- **Análise de Redes de Coocorrência:** Embora não seja estritamente uma análise de correlação com variáveis externas, a construção de redes de coocorrência entre táxons pode revelar associações (positivas ou negativas) entre eles, e essas redes podem então ser analisadas em relação a diferentes condições ambientais ou grupos.[34]

A identificação de táxons que são diferencialmente abundantes entre grupos ou que se correlacionam com variáveis de interesse é, frequentemente, apenas o primeiro passo na investigação biológica. É crucial lembrar que correlação não implica causalidade. Uma associação estatística observada pode ser devida a diversos fatores, incluindo causalidade direta, causalidade reversa, ou a influência de uma terceira variável de confusão não medida ou não considerada na análise. Por exemplo, um determinado micróbio pode estar aumentado em uma condição de doença não porque ele causa a doença, mas porque ele se beneficia do ambiente fisiológico alterado pela própria doença, ou sua abundância pode estar correlacionada com um fator de estilo de vida (como a dieta) que é o verdadeiro motor da mudança observada tanto no microbioma quanto no estado de saúde. Portanto, os resultados dessas análises estatísticas avançadas devem ser interpretados com cautela e, idealmente, vistos como geradores de hipóteses que

requerem validação experimental adicional ou integração com outros tipos de dados (por exemplo, dados funcionais de metagenômica shotgun, metatranscriptômica, ou estudos longitudinais) para se aproximar de uma compreensão mecanística ou inferir relações causais. A coleta e a integração de metadados detalhados e padronizados sobre as amostras e os sujeitos do estudo são, portanto, de importância crítica para permitir uma interpretação mais robusta e contextualizada dos achados estatísticos.[7]

4. Aplicações do Sequenciamento de rRNA 16S

O sequenciamento do gene rRNA 16S tornou-se uma ferramenta onipresente na microbiologia, com aplicações que se estendem por uma vasta gama de disciplinas científicas e setores práticos. Sua capacidade de fornecer um perfil taxonômico de comunidades microbianas complexas de forma relativamente rápida e custo-efetiva impulsionou inúmeras descobertas.

4.1. Ecologia Microbiana

Em seu cerne, o sequenciamento de rRNA 16S é uma ferramenta fundamental para a ecologia microbiana. Ele permite o estudo da diversidade (quem está lá?), da estrutura (como os membros estão organizados e em que proporções?) e da dinâmica (como a comunidade muda ao longo do tempo ou em resposta a perturbações?) de comunidades microbianas em uma miríade de ecossistemas.[1] Isso inclui ambientes terrestres como solos e rizosferas, ambientes aquáticos de água doce (lagos, rios) e marinha (oceanos, estuários, fontes hidrotermais), sedimentos, ambientes extremos (desertos quentes e frios, fontes termais ácidas, salinas), e até mesmo a atmosfera, através da análise de bioaerossóis.[33]

A técnica é amplamente utilizada para investigar como fatores ambientais, sejam eles naturais (como tipo de solo, pH, salinidade, temperatura, disponibilidade de nutrientes) ou antropogênicos (como poluição, mudanças no uso da terra, práticas agrícolas), moldam a composição e a função das comunidades microbianas.[15] Ao comparar perfis de rRNA 16S de diferentes locais ou sob diferentes condições, os ecologistas microbianos podem identificar táxons indicadores, entender nichos ecológicos e desvendar os processos que governam a montagem e a estabilidade das comunidades. Além disso, o sequenciamento de rRNA 16S continua a ser uma ferramenta valiosa para a descoberta de novos táxons bacterianos e arqueanos e para a caracterização de sua distribuição geográfica e ecológica.[5] Exemplos específicos incluem estudos sobre microrganismos em leitos de carvão, que podem estar envolvidos na geração de metano biogênico, e a caracterização de comunidades bacterianas em ecossistemas de zonas úmidas para entender sua estrutura e função ambiental.[35]

4.2. Saúde Humana

A aplicação do sequenciamento de rRNA 16S na área da saúde humana tem sido revolucionária, particularmente no estudo do microbioma humano. Esta técnica permitiu a caracterização detalhada das comunidades microbianas que habitam diferentes sítios do corpo humano, como o trato gastrointestinal, a pele, a cavidade oral, o trato vaginal e, mais recentemente, órgãos antes considerados estéreis, como os pulmões.[3] Esses estudos revelaram a imensa diversidade do microbioma humano e sua considerável variação entre indivíduos, bem como as mudanças que ocorrem ao longo da vida de um indivíduo.

Uma área de intensa pesquisa é a associação entre disbioses – alterações na composição ou função do microbioma normal – e uma ampla gama de doenças humanas. O sequenciamento de rRNA 16S tem sido fundamental para identificar padrões de disbiose associados a doenças inflamatórias intestinais (como a doença de Crohn e a colite ulcerativa), obesidade, diabetes tipo 2, doenças cardiovasculares, distúrbios neurológicos e psiquiátricos, alergias, doenças autoimunes e diversos tipos de câncer, como o câncer colorretal e o câncer de mama.[3] Embora muitos desses estudos sejam correlacionais, eles fornecem pistas importantes sobre o papel potencial do microbioma na patogênese dessas doenças e abrem caminhos para novas estratégias diagnósticas e terapêuticas.

Além disso, o sequenciamento de rRNA 16S é usado para monitorar o efeito de intervenções que visam modular o microbioma, como o uso de probióticos, prebióticos, simbióticos, antibióticos e o transplante de microbiota fecal (TMF). No contexto clínico, a técnica tem um valor diagnóstico importante, especialmente para a identificação de patógenos bacterianos em amostras onde os métodos de cultura tradicionais falham, como em casos de infecções por microrganismos fastidiosos ou não cultiváveis, ou quando o paciente já recebeu tratamento antibiótico empírico, o que pode levar a culturas negativas.[17]

4.3. Biotecnologia

Na biotecnologia, o sequenciamento de rRNA 16S serve como uma ferramenta para a prospecção e caracterização de microrganismos e comunidades microbianas com potencial para aplicações industriais, agrícolas ou ambientais.[3] Ao explorar a diversidade microbiana em ambientes únicos ou extremos, os pesquisadores podem identificar novos microrganismos que produzem enzimas, antibióticos, biossurfactantes ou outros compostos bioativos de interesse.

A técnica é aplicada no estudo de processos de fermentação de alimentos, ajudando a identificar os microrganismos responsáveis pelas transformações

desejáveis e a monitorar a qualidade e segurança dos produtos fermentados.[5] Em processos industriais que dependem de atividades microbianas, como a produção de biogás em digestores anaeróbios ou o tratamento biológico de efluentes em biorreatores, o sequenciamento de rRNA 16S permite o monitoramento da composição da comunidade microbiana, a identificação de microrganismos chave para a eficiência do processo e a otimização das condições operacionais.[35] Na área de biorremediação, a técnica é usada para identificar e caracterizar microrganismos nativos ou consórcios microbianos capazes de degradar poluentes orgânicos persistentes, metais pesados ou outros contaminantes ambientais, auxiliando no desenvolvimento de estratégias de limpeza de sítios contaminados.[3]

4.4. Monitoramento Ambiental

O sequenciamento de rRNA 16S é uma ferramenta valiosa para o monitoramento ambiental, permitindo a avaliação da saúde de ecossistemas e a detecção de impactos de atividades humanas. As comunidades microbianas são altamente sensíveis a mudanças nas condições ambientais e, portanto, sua composição e diversidade podem servir como bioindicadores da qualidade da água, do solo e do ar.[3]

Ao analisar perfis de rRNA 16S de amostras ambientais coletadas ao longo do tempo ou de diferentes locais, é possível detectar o impacto de poluentes (como pesticidas, hidrocarbonetos, metais pesados), de mudanças no uso da terra (como desmatamento ou urbanização) ou de outras perturbações antropogênicas sobre a biodiversidade microbiana e o funcionamento dos ecossistemas.[35] A técnica também pode ser usada para o monitoramento de patógenos de veiculação hídrica ou alimentar em reservatórios ambientais, contribuindo para a saúde pública.

4.5. Ciência Forense

Mais recentemente, o sequenciamento de rRNA 16S, frequentemente combinado com abordagens de aprendizado de máquina (machine learning), tem encontrado aplicações promissoras na ciência forense. O microbioma humano, particularmente o da pele, saliva e intestino, exibe um grau de individualidade que sugere seu potencial como um novo tipo de biomarcador para a identificação de indivíduos ou para associar um indivíduo a um local ou objeto.[36]

Estudos têm explorado a análise do microbioma do solo associado a restos mortais para ajudar a estimar o intervalo post-mortem (o tempo decorrido desde a morte).[36] A combinação de dados de sequenciamento de rRNA 16S com algoritmos de aprendizado de máquina pode melhorar a precisão da classificação de

amostras e da identificação de indivíduos, analisando os padrões complexos de abundância microbiana.[36] Embora ainda seja um campo em desenvolvimento, a metagenômica forense baseada em rRNA 16S oferece novas possibilidades para a investigação criminal.

A notável versatilidade do sequenciamento de rRNA 16S, que o torna aplicável a uma miríade de questões de pesquisa e a uma diversidade impressionante de tipos de amostra, é, paradoxalmente, tanto sua maior força quanto uma fonte potencial de desafios. Essa ampla aplicabilidade, que abrange desde a ecologia de ambientes extremos até o diagnóstico clínico e a investigação forense, significa que os protocolos de amostragem, extração de DNA, escolha de primers, plataformas de sequenciamento e pipelines bioinformáticos podem variar consideravelmente entre diferentes estudos, cada um otimizado para seus objetivos específicos e as características particulares de suas amostras. Embora o princípio central – sequenciar o gene rRNA 16S para obter um perfil taxonômico – permaneça o mesmo, essas variações metodológicas podem dificultar a comparação direta e a integração de dados gerados em contextos tão diversos. Por exemplo, comparar os resultados de um estudo do microbioma do solo, que pode ter usado um método de extração de DNA robusto para lisar esporos e células resistentes, com os de um estudo do microbioma da pele, que pode ter lidado com baixa biomassa e usado primers diferentes, requer extrema cautela. Isso reforça a necessidade crítica de relatórios metodológicos detalhados, transparentes e padronizados em todas as publicações científicas, e o desenvolvimento e adoção de padrões mínimos de reporte (como os propostos por iniciativas como o Genomic Standards Consortium) para facilitar a integração de dados, a realização de meta-análises robustas e a construção de um corpo de conhecimento mais amplo e coeso a partir de estudos individuais.

5. Limitações, Desafios Atuais e Perspectivas Futuras

Apesar de sua ampla utilidade e das inúmeras contribuições para o avanço do conhecimento microbiano, o sequenciamento do gene rRNA 16S não está isento de limitações e desafios. Compreender essas questões é crucial para a interpretação correta dos resultados e para o desenvolvimento futuro da área.

5.1. Limitações Inerentes ao rRNA 16S

Algumas limitações são intrínsecas à natureza do próprio gene rRNA 16S como marcador molecular:

- **Resolução Taxonômica:** Uma das limitações mais reconhecidas é a dificuldade em distinguir espécies bacterianas ou arqueanas muito proximamente relacionadas, e especialmente diferentes cepas dentro de uma

mesma espécie.[5] O gene rRNA 16S, embora contenha regiões hipervariáveis, pode ser altamente conservado entre organismos filogeneticamente próximos. Como resultado, a resolução taxonômica alcançada com o sequenciamento de rRNA 16S é frequentemente limitada ao nível de gênero, e a identificação precisa em nível de espécie pode ser desafiadora para muitos grupos.[8] Essa limitação é exacerbada quando se utilizam tecnologias de sequenciamento de reads curtas que cobrem apenas uma ou algumas regiões hipervariáveis do gene, em vez de sua sequência completa, pois menos informação filogenética está disponível.[10]

- **Vieses de PCR e Primers:** Como discutido anteriormente (Seção 2.3), a etapa de PCR é uma fonte significativa de viés. A escolha dos primers, suas afinidades diferenciais por diferentes moldes de DNA e as condições da reação de PCR podem levar a uma amplificação não representativa da comunidade microbiana original, com alguns táxons sendo super-representados e outros sub-representados ou mesmo não detectados.[8]
- **Heterogeneidade Intragenômica (Variação no Número de Cópias e Sequências):** Muitos procariontes possuem múltiplas cópias do operon rRNA (que inclui o gene 16S) em seus genomas. O número de cópias pode variar consideravelmente entre diferentes espécies, de uma a mais de quinze. Além disso, as sequências dessas múltiplas cópias dentro de um mesmo genoma nem sempre são idênticas; podem existir pequenas variações (microheterogeneidade) entre elas.[35] Essa variação no número de cópias e na sequência intragenômica pode complicar a interpretação dos dados. A variação no número de cópias pode levar a uma superestimação da abundância relativa de organismos com muitas cópias do gene 16S, se as contagens de reads não forem normalizadas por esse fator. A microheterogeneidade intragenômica pode inflar artificialmente as estimativas de diversidade, com diferentes cópias do mesmo organismo sendo erroneamente classificadas como OTUs ou ASVs distintas.
- **Incapacidade de Inferência Funcional Direta:** O gene rRNA 16S é um marcador filogenético; sua sequência informa sobre a identidade taxonômica de um microrganismo, mas não fornece informações diretas sobre o potencial funcional da comunidade microbiana – ou seja, quais genes metabólicos estão presentes ou quais funções os microrganismos são capazes de realizar.[5] Embora existam ferramentas bioinformáticas (como PICRUSt, Tax4Fun, PanFP) que tentam predizer o perfil funcional de uma comunidade com base em seu perfil taxonômico de rRNA 16S e em bancos de dados de genomas de referência, essas predições são inferências indiretas e possuem limitações significativas.[8] Estudos comparativos demonstraram que essas ferramentas de predição funcional baseadas em 16S geralmente não possuem a sensibilidade

necessária para delinear mudanças funcionais sutis ou clinicamente relevantes no microbioma, especialmente quando comparadas com os resultados do sequenciamento metagenômico shotgun, que analisa diretamente todos os genes da comunidade.[37]

- **Desafios na Quantificação de Abundância:** Os dados gerados pelo sequenciamento de amplicons do rRNA 16S são inerentemente relativos; as contagens de reads para cada táxon refletem sua proporção dentro do conjunto total de sequências obtidas para aquela amostra, e não sua abundância absoluta no ambiente original. Converter essas abundâncias relativas em abundâncias celulares absolutas ou biovolumes é um desafio. Os vieses de PCR e, crucialmente, a variação no número de cópias do gene rRNA 16S entre diferentes táxons, complicam significativamente essa quantificação.[8] Organismos com mais cópias do gene 16S produzirão mais amplicons por célula, levando a uma sobrerrepresentação nas contagens de reads, a menos que correções sejam aplicadas.

5.2. Desafios Atuais na Metagenômica de rRNA 16S

Além das limitações inerentes ao marcador, existem desafios metodológicos e interpretativos mais amplos que a comunidade científica enfrenta ao aplicar o sequenciamento de rRNA 16S:

- **Padronização de Protocolos:** Uma das maiores dificuldades na área é a falta de padronização nos protocolos experimentais (desde a coleta da amostra e extração de DNA, até a escolha de primers e condições de PCR) e nos pipelines bioinformáticos (softwares, algoritmos, parâmetros e bancos de dados de referência utilizados). Essa variabilidade metodológica introduz vieses e dificulta enormemente a comparação e a integração de resultados entre diferentes estudos, laboratórios e projetos.[7]
- **Desenho Experimental Robusto:** Um desenho experimental bem planejado é fundamental para a obtenção de resultados confiáveis e biologicamente significativos. Isso inclui:
  - *Tamanho amostral adequado:* É necessário um número suficiente de amostras para garantir poder estatístico para detectar diferenças reais entre os grupos, especialmente quando se esperam efeitos sutis ou quando há alta variabilidade interindividual.[7] Estudos de poder, como os mencionados em [27] e [27], podem ajudar a estimar o tamanho amostral necessário, que pode variar dependendo da métrica de diversidade ou do desfecho escolhido.
  - *Controles apropriados:* O uso de controles negativos (para monitorar contaminação durante a extração e PCR) e controles positivos (comunidades mock de composição conhecida, para validar o pipeline

experimental e bioinformático) é essencial para garantir a qualidade e a validade dos dados.[7]

- *Coleta de metadados:* A coleta de metadados detalhados e padronizados sobre as amostras (origem, características ambientais, informações do hospedeiro, tratamentos, etc.) é crucial para a interpretação contextualizada dos resultados e para a identificação de fatores de confusão.[7]

- **Complexidade da Análise de Dados:** O volume de dados gerado pelo sequenciamento de nova geração é massivo. A multiplicidade de ferramentas bioinformáticas, algoritmos e parâmetros disponíveis para cada etapa da análise (pré-processamento, geração de OTU/ASV, classificação taxonômica, análise de diversidade, testes estatísticos) pode ser esmagadora, especialmente para pesquisadores que não são especialistas em bioinformática. Diferentes escolhas no pipeline de análise podem levar a resultados e conclusões diferentes, tornando a reprodutibilidade um desafio.[7]
- **Limitações dos Bancos de Dados de Referência:** A acurácia da atribuição taxonômica depende diretamente da qualidade dos bancos de dados de referência. Estes podem ser incompletos (especialmente para microrganismos de ambientes pouco explorados ou não cultiváveis), conter erros de anotação ou estar desatualizados (como o caso do Greengenes [20]). Essas limitações podem levar a classificações taxonômicas incorretas, incompletas ou a uma alta proporção de sequências não atribuídas.[7]
- **Interpretação dos Resultados:** Um dos maiores desafios é a interpretação biológica correta dos resultados. É fundamental distinguir o sinal biológico real do ruído técnico ou artefatos metodológicos. Deve-se evitar a superinterpretação de correlações estatísticas como evidência de causalidade, e reconhecer que muitos fatores podem influenciar a composição do microbioma. A inferência de mecanismos a partir de dados de rRNA 16S é particularmente difícil devido à falta de informação funcional direta.[7]

5.3. Perspectivas Futuras

Apesar dos desafios, o campo do sequenciamento de rRNA 16S continua a evoluir, com várias perspectivas promissoras para o futuro:

- Impacto Contínuo do Sequenciamento de Reads Longas:
  Como mencionado anteriormente (Seção 2.4), as tecnologias de sequenciamento de reads longas (PacBio e ONT) estão transformando a análise do gene rRNA 16S ao permitir o sequenciamento de sua sequência completa (full-length).4 Isso tem o potencial de:
  - *Melhorar significativamente a resolução taxonômica*, permitindo uma

identificação mais precisa em nível de espécie e, possivelmente, até de cepa, superando uma das principais limitações das abordagens baseadas em reads curtas.
    - *Facilitar uma ligação mais confiável entre a taxonomia derivada do 16S e os genomas completos de referência*, o que poderia, indiretamente, melhorar a inferência funcional.
    - *Expandir as aplicações em diagnóstico clínico*, onde a identificação rápida e precisa de patógenos é crucial. A portabilidade e a capacidade de análise em tempo real da tecnologia ONT são particularmente promissoras nesse sentido.[17]
- Aprimoramento de Ferramentas Bioinformáticas e Bancos de Dados:
  Espera-se um desenvolvimento contínuo de algoritmos bioinformáticos mais precisos, eficientes e robustos para todas as etapas da análise de dados de rRNA 16S, desde o denoising e a geração de ASVs até a classificação taxonômica e a análise estatística. A curadoria e a expansão dos bancos de dados de referência são cruciais, com a incorporação de sequências de genomas de microrganismos não cultiváveis, de ambientes pouco explorados e de dados de sequenciamento de reads longas do 16S. Métodos para avaliar a qualidade e a precisão das anotações taxonômicas, como o Qscore proposto em 10 para avaliar o desempenho de diferentes amplicons e estratégias de sequenciamento, também contribuirão para o refinamento da área.
- Integração Multi-ômica:
  Uma das direções futuras mais importantes para a pesquisa em microbioma é a integração de dados de rRNA 16S com outras abordagens "ômicas".38 Isso inclui:
    - *Metagenômica shotgun:* Fornece informações diretas sobre o potencial genético funcional da comunidade (todos os genes presentes).
    - *Metatranscriptômica:* Analisa os transcritos de RNA, revelando quais genes estão ativamente expressos pela comunidade em um determinado momento.
    - *Proteômica:* Identifica e quantifica as proteínas expressas, que são os efetores moleculares da célula.
    - *Metabolômica:* Caracteriza o perfil de metabólitos (pequenas moléculas) produzidos ou modificados pela comunidade microbiana e/ou pelo hospedeiro. A combinação desses diferentes tipos de dados permite uma visão muito mais holística e mecanística das comunidades microbianas, de suas funções e de suas interações com o hospedeiro ou o ambiente. No entanto, a integração de dados de naturezas, escalas e níveis de ruído tão diferentes apresenta desafios computacionais e estatísticos significativos,

exigindo o desenvolvimento de métodos sofisticados de análise.[39] Questões como a ambiguidade na atribuição taxonômica (que afeta a ligação entre táxons e funções), a composicionalidade dos dados, a esparsidade, a variabilidade e, muitas vezes, o tamanho amostral limitado em estudos multi-ômicos precisam ser cuidadosamente consideradas.[39]

O futuro da metagenômica de rRNA 16S parece residir não apenas em seu contínuo refinamento técnico, como o proporcionado pelo sequenciamento de reads longas, mas, crucialmente, em sua integração inteligente e estratégica com outras abordagens ômicas. Em vez de ser visto apenas como uma alternativa "inferior" ou menos informativa ao sequenciamento metagenômico shotgun, o sequenciamento de rRNA 16S pode continuar a desempenhar um papel valioso como uma ferramenta de triagem taxonômica robusta, custo-efetiva e de alto rendimento. Os perfis taxonômicos gerados pelo 16S podem ser usados para identificar padrões de interesse (por exemplo, disbiose associada a uma doença, ou mudanças na comunidade em resposta a uma perturbação ambiental) em um grande número de amostras. Com base nesses achados, um subconjunto de amostras pode então ser selecionado para investigações mais profundas e funcionalmente orientadas, utilizando técnicas como metagenômica shotgun, metatranscriptômica, proteômica ou metabolômica, para desvendar os mecanismos subjacentes às diferenças taxonômicas observadas. Essa abordagem em etapas posiciona o sequenciamento de rRNA 16S como um componente valioso e complementar dentro de um pipeline de pesquisa multi-ômica mais amplo. A padronização de protocolos e a melhoria contínua da resolução taxonômica do 16S (especialmente com o advento das reads longas) tornariam essa integração ainda mais poderosa, pois permitiriam uma ligação mais precisa e confiável entre a identidade taxonômica fornecida pelo 16S e os dados funcionais derivados de outras plataformas ômicas.

6. Conclusão

O sequenciamento do gene rRNA 16S consolidou-se como uma ferramenta fundamental e transformadora no estudo das comunidades microbianas. Desde sua concepção, tem permitido aos pesquisadores explorar a vasta diversidade do mundo microbiano em uma escala sem precedentes, revelando a composição de microbiomas em uma miríade de ambientes, desde o corpo humano até os ecossistemas mais remotos da Terra. O fluxo de trabalho, que abrange desde a cuidadosa coleta de amostras, passando pela extração de DNA, amplificação por PCR do gene alvo, sequenciamento de nova geração e culminando em complexas análises bioinformáticas, requer atenção meticulosa a cada etapa para garantir resultados robustos e interpretáveis.

As aplicações do sequenciamento de rRNA 16S são vastas e impactantes, permeando campos como a ecologia microbiana fundamental, a saúde humana (onde desvendou o papel do microbioma em inúmeras doenças e condições), a biotecnologia (na prospecção de novos microrganismos e otimização de processos), o monitoramento ambiental (como bioindicador da saúde de ecossistemas) e, mais recentemente, a ciência forense.

No entanto, é crucial reconhecer e abordar as limitações inerentes a esta técnica. A resolução taxonômica, frequentemente limitada ao nível de gênero, os vieses introduzidos durante a amplificação por PCR, a heterogeneidade intragenômica do gene 16S, a incapacidade de inferir diretamente o potencial funcional da comunidade e os desafios na quantificação precisa da abundância microbiana são fatores que exigem uma interpretação cautelosa dos resultados e escolhas metodológicas informadas. A falta de padronização em protocolos experimentais e bioinformáticos continua a ser um desafio para a comparabilidade e a integração de dados entre estudos.

Apesar dessas limitações, o futuro do sequenciamento de rRNA 16S permanece promissor. Os avanços tecnológicos, notadamente o desenvolvimento e a crescente acessibilidade do sequenciamento de reads longas, prometem superar algumas das limitações históricas, como a resolução taxonômica, ao permitir a análise da sequência completa do gene 16S. Paralelamente, o aprimoramento contínuo de ferramentas bioinformáticas, a expansão e curadoria de bancos de dados de referência e, crucialmente, a integração de dados de rRNA 16S com outras abordagens ômicas (metagenômica, metatranscriptômica, proteômica e metabolômica) estão pavimentando o caminho para uma compreensão mais profunda, holística e mecanística das comunidades microbianas e de suas complexas interações. O sequenciamento de rRNA 16S, portanto, continuará a ser uma peça valiosa no arsenal de ferramentas disponíveis para desvendar os mistérios e o potencial do vasto e dinâmico mundo microbiano.

7. Estado da Arte

O projeto se posiciona na interseção entre bioinformática, inteligência artificial conversacional e processamento de linguagem natural (PLN), com foco na interpretação automática de dados metagenômicos do QIIME 2 — especialmente em análises 16S rRNA. Ele parte de um problema claro: apesar de ferramentas como o QIIME 2 serem amplamente usadas para estudos de microbiomas, a interpretação de seus resultados ainda exige

conhecimento técnico especializado, algo escasso em regiões como a Amazônia.

O estado atual das tecnologias relacionadas mostra que:

- IA e PLN já são usados para análise de dados científicos, mas aplicações específicas para bioinformática metagenômica com interface conversacional ainda são pouco exploradas.
- Agentes conversacionais (ChatGPT, Siri, Alexa) são reconhecidos e utilizados amplamente, mas raramente adaptados para o vocabulário e fluxo de trabalho da bioinformática.
- Frameworks open source como o Rasa oferecem flexibilidade, transparência e controle total do pipeline de diálogo, permitindo integração direta com bibliotecas Python (pandas, seaborn, biom-format, plotly) para análise de dados biológicos.
- No contexto regional, não há ferramentas que conciliem soberania de dados, acessibilidade computacional e interpretação científica automatizada voltada para microbiomas da Amazônia.

O diferencial do YARA está em:

1. Foco local e soberania científica – modelo treinado com vocabulário técnico da microbiologia e adaptado à realidade amazônica.
2. Integração prática com QIIME 2 – leitura direta de arquivos .qzv, .tsv e .biom.
3. Geração automática de relatórios, gráficos e interpretações em linguagem acessível para pesquisadores não especialistas.
4. Acessibilidade via Telegram – interação simples, sem necessidade de infraestrutura pesada.
5. Metodologia validada – uso de testes de usabilidade (SUS) e coleta de feedback de pesquisadores da EMBRAPA e INPA.

7. Referências Bibliográficas Citadas no Relatório

- [1] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews, 68*(4), 669-685. [1]
- [2] Luef, B., & Gueddouri, D. (2024). Metagenomics. *Nature Primers, (Primer atualizado em 2025)*.
- [3] Uncoded. (2024, November 6). *16S Metagenomics: A Powerful Tool for Microbial Diversity Analysis*.
- [4] CD Genomics. (n.d.). *Microbial Species Identification Methods: 16S rRNA and Metagenomic Sequencing*.
- [5] Aryal, S. (2024, December 13). *16S rRNA Gene Sequencing- Principle, Steps, Limitations, Applications*. Microbe Notes.
- [6] Illumina, Inc. (2025). *16S rRNA Sequencing*.
- [8] Zaura, E., et al. (2021). Relative Fecal Microbiome Compositions and Functional Potential Unchanged by Antibiotic Prophylaxis for Dental Procedures. *Frontiers in Microbiology, 12*, 670336.
- [8] Zaura, E., et al. (2021). Relative Fecal Microbiome Compositions and Functional Potential Unchanged by Antibiotic Prophylaxis for Dental Procedures. *Frontiers in Microbiology, 12*, 670336. [8]
- [9] Hsieh, Y. (n.d.). *16S-ITGDB README*. GitHub.
- [10] Liu, Y., et al. (2023). Qscore: a comprehensive method to evaluate the performance of 16S rRNA gene amplicons in microbiome profiling. *Microbiology Spectrum, 11*(6), e00563-23.
- [5] Aryal, S. (2024, December 13). *16S rRNA Gene Sequencing- Principle, Steps, Limitations, Applications*. Microbe Notes. [5]
- [6] Illumina, Inc. (2025). *16S rRNA Sequencing*. [6]
- [11] Valdes-Mas, R., et al. (2024). A Systematic Review and Meta-Analysis of 16S rRNA and Cancer Microbiome Atlas Datasets to Characterize Microbiota Signatures in Normal Breast, Mastitis, and Breast Cancer. *Cancers, 13*(2), 467.
- [12] Relman, D. A. (2012). Learning about who we are. *Nature, 486*(7402), 194-195. [12]
- [13] Minot, S. S., et al. (2023). Improved DNA Extraction and Amplification Strategy for 16S rRNA Gene Amplicon-Based Microbiome Studies. *Research Square (Preprint)*. [13]
- [14] Minot, S. S., et al. (2024). Improved DNA extraction and amplification strategy for 16S rRNA gene amplicon-based microbiome studies. *PLoS ONE, 19*(3), e0298780. [13]
- [14] Minot, S. S., et al. (2024). Improved DNA extraction and amplification strategy for 16S rRNA gene amplicon-based microbiome studies. *PLoS ONE,*

19(3), e0298780. [14]

- [7] Jovel, J., et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology, 7*, 459. [7]
- [15] Fischer, M., et al. (2023). Performance and potential bias of two standard 16S rRNA gene primer pairs in profiling soil prokaryotic communities. *Frontiers in Microbiology, 14*, 1140487.
- [15] Fischer, M., et al. (2023). Performance and potential bias of two standard 16S rRNA gene primer pairs in profiling soil prokaryotic communities. *Frontiers in Microbiology, 14*, 1140487. [15]
- [16] Berry, D., et al. (2011). Continental-scale distributions of nitrogen-fixing bacteria reflect soils and climate. *The ISME Journal, 5*(6), 991-1004. [16]
- [16] Berry, D., et al. (2011). Continental-scale distributions of nitrogen-fixing bacteria reflect soils and climate. *The ISME Journal, 5*(6), 991-1004. [16]
- [17] O'Sullivan, D. M., et al. (2024). Validation of Oxford Nanopore Technology (ONT) for 16S rRNA gene sequencing in clinical diagnostics using characterized reference materials. *Frontiers in Cellular and Infection Microbiology, 15*, 1517208. [17]
- [18] O'Sullivan, D. M., et al. (2024). Validation of Oxford Nanopore Technology (ONT) for 16S rRNA gene sequencing in clinical diagnostics using characterized reference materials. *Frontiers in Cellular and Infection Microbiology, 15*, 1517208. [17]
- [17] O'Sullivan, D. M., et al. (2024). Validation of Oxford Nanopore Technology (ONT) for 16S rRNA gene sequencing in clinical diagnostics using characterized reference materials. *Frontiers in Cellular and Infection Microbiology, 15*, 1517208. [17]
- [18] O'Sullivan, D. M., et al. (2024). Validation of Oxford Nanopore Technology (ONT) for 16S rRNA gene sequencing in clinical diagnostics using characterized reference materials. *Frontiers in Cellular and Infection Microbiology, 15*, 1517208. [17]
- [19] Xu, Y., et al. (2023). Rapid diagnosis of infected body fluids using nanopore 16S rRNA gene sequencing. *Frontiers in Microbiology, 14*, 1324494.
- [20] Pauvert, C., et al. (2019). Systematic comparison of 16S rRNA gene specific primers and test of a new primer pair for Human Gut Microbiome studies. *bioRxiv (Preprint)*. [20]
- [21] Lappan, R. (n.d.). *VL QIIME2 analysis: Pre-processing of sequence reads*. GitHub Pages.
- [21] Lappan, R. (n.d.). *VL QIIME2 analysis: Pre-processing of sequence reads*. GitHub Pages. [21]
- [22] Li, F., et al. (2023). Comparison of OTU clustering and denoising methods for 16S rRNA amplicon sequencing data in colorectal cancer. *Frontiers in*

Microbiology, 14, 1178744.

- [23] Pylro, V. S., et al. (2024). OTU clustering or ASV inference: Which is the future of microbial community diversity studies? *PLoS ONE, 19*(10), e0309065.
- [23] Pylro, V. S., et al. (2024). OTU clustering or ASV inference: Which is the future of microbial community diversity studies? *PLoS ONE, 19*(10), e0309065. [23]
- [24] Nearing, J. T., et al. (2022). Denoising the Denoisers: An Independent Evaluation of Microbiome Amplicon Processing Pipelines. *mBio, 13*(1), e02 denoising21.
- [24] Nearing, J. T., et al. (2022). Denoising the Denoisers: An Independent Evaluation of Microbiome Amplicon Processing Pipelines. *mBio, 13*(1), e02 denoising21. [24]
- [22] Li, F., et al. (2023). Comparison of OTU clustering and denoising methods for 16S rRNA amplicon sequencing data in colorectal cancer. *Frontiers in Microbiology, 14*, 1178744. [22]
- [25] Beiting, D. (n.d.). *CHMI QIIME2 SOPs*. GitHub Pages.
- [26] Reddit r/bioinformatics. (2023). *16S rRNA analysis using QIIME2 - unassigned sequences*.
- [20] Schloss, P. D. (2021). Reintroducing Mothur: A Comprehensive Software Package for Microbial Ecology. *Applied and Environmental Microbiology, 87*(11), e02343-20. [20]
- [27] La Rosa, M., et al. (2022). Power calculations for 16S rRNA microbiome data. *Frontiers in Microbiology, 12*, 796025.
- [28] Calgaro, M., et al. (2025). Alpha diversity metrics for microbiome data: a comparative review. *Briefings in Bioinformatics, bbab544*.
- [29] One Codex. (n.d.). *Alpha Diversity*.
- [27] La Rosa, M., et al. (2022). Power calculations for 16S rRNA microbiome data. *Frontiers in Microbiology, 12*, 796025. [27]
- [28] Calgaro, M., et al. (2025). Alpha diversity metrics for microbiome data: a comparative review. *Briefings in Bioinformatics, bbab544*. [28]
- [30] One Codex. (n.d.). *Beta Diversity*.
- [31] EzBioCloud. (n.d.). *Differential Abundance Analysis*.
- [32] Haworth, S. E., et al. (2021). A comparison of differential abundance testing methodologies for microbiome data. *bioRxiv (Preprint)*.
- [31] EzBioCloud. (n.d.). *Differential Abundance Analysis*. [31]
- [33] Ling, F., et al. (2018). Different airborne bacterial communities in outdoor environment as characterized by 16S rRNA and 16S rRNA gene sequencing. *Scientific Reports, 8*(1), 1389.
- [34] CD Genomics. (n.d.). *Microbial Community Study: Correlation Analysis of 16S and ITS Sequencing*.
- [34] CD Genomics. (n.d.). *Microbial Community Study: Correlation Analysis of 16S*

and ITS Sequencing. [34]

- [7] Jovel, J., et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology, 7*, 459. [7]
- [35] Wang, Y., & Qian, P. Y. (2009). The applications of the 16S rRNA gene in microbial ecology: current situation and problems. *Applied Microbiology and Biotechnology, 82*(2), 251-260.
- [35] Wang, Y., & Qian, P. Y. (2009). The applications of the 16S rRNA gene in microbial ecology: current situation and problems. *Applied Microbiology and Biotechnology, 82*(2), 251-260. [35]
- [5] Aryal, S. (2024, December 13). *16S rRNA Gene Sequencing- Principle, Steps, Limitations, Applications*. Microbe Notes. [5]
- [36] Zhang, Y., et al. (2024). Application of 16S rRNA gene sequencing and machine learning in forensic microbiome for individual identification. *Frontiers in Microbiology, 15*, 1360457.
- [36] Zhang, Y., et al. (2024). Application of 16S rRNA gene sequencing and machine learning in forensic microbiome for individual identification. *Frontiers in Microbiology, 15*, 1360457. [36]
- [8] Zaura, E., et al. (2021). Relative Fecal Microbiome Compositions and Functional Potential Unchanged by Antibiotic Prophylaxis for Dental Procedures. *Frontiers in Microbiology, 12*, 670336. [8]
- [8] Zaura, E., et al. (2021). Relative Fecal Microbiome Compositions and Functional Potential Unchanged by Antibiotic Prophylaxis for Dental Procedures. *Frontiers in Microbiology, 12*, 670336. [8]
- [37] Wirbel, J., et al. (2024). 16S rRNA gene-based functional inference tools generally do not have the necessary sensitivity to delineate health-related functional changes in the microbiome and should thus be used with care. *Microbiome, 10*(2), 001203.
- [37] Wirbel, J., et al. (2024). 16S rRNA gene-based functional inference tools generally do not have the necessary sensitivity to delineate health-related functional changes in the microbiome and should thus be used with care. *Microbiome, 10*(2), 001203. [37]
- [38] ResearchGate. (n.d.). *Bioinformatics for Multi-Omics Data Integration (Collection of articles)*.
- [39] Braccia, D., et al. (2024). Statistical methods for multi-omic integration of microbiome data. *Gut Microbes, 16*(1), 2297860.
- [39] Braccia, D., et al. (2024). Statistical methods for multi-omic integration of microbiome data. *Gut Microbes, 16*(1), 2297860. [39]

Referências citadas

1. The 16S rRNA gene in the study of marine microbial communities, acessado em maio 11, 2025, https://www.scielo.org.mx/scielo.php?pid=S0185-38802015000400297&script=sci_arttext&tlng=en
2. Analysis of metagenomic data | Springer Nature Experiments, acessado em maio 11, 2025, https://experiments.springernature.com/nature/primers/10.1038/s43586-024-00376-6
3. 16S Metagenomics: A Powerful Tool for Microbial Diversity Analysis - Uncoded, acessado em maio 11, 2025, https://uncoded.in/16s-metagenomics-a-powerful-tool-for-microbial-diversity-analysis/
4. Microbial Species Identification Methods: 16S rRNA and Metagenomic Sequencing, acessado em maio 11, 2025, https://www.cd-genomics.com/microbioseq/resource-microbial-species-identification-methods-16s-rrna-and-metagenomic-sequencing.html
5. 16S rRNA Gene Sequencing: Principle, Steps, Uses, Diagram, acessado em maio 11, 2025, https://microbenotes.com/16s-rrna-gene-sequencing/
6. 16S and ITS rRNA Sequencing | Identify bacteria & fungi with NGS, acessado em maio 11, 2025, https://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/16s-rrna-sequencing.html
7. Current challenges and best-practice protocols for microbiome ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7820839/
8. Comparative Analysis of 16S rRNA Gene and ... - Frontiers, acessado em maio 11, 2025, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.670336/full
9. 16S-ITGDB/README.md at master · yphsieh/16S-ITGDB · GitHub, acessado em maio 11, 2025, https://github.com/yphsieh/16S-ITGDB/blob/master/README.md
10. Comprehensive Assessment of 16S rRNA Gene Amplicon ..., acessado em maio 11, 2025, https://journals.asm.org/doi/10.1128/spectrum.00563-23
11. A Systematic Review and Meta-Analysis of 16S rRNA and Cancer Microbiome Atlas Datasets to Characterize Microbiota Signatures in Normal Breast, Mastitis, and Breast Cancer - MDPI, acessado em maio 11, 2025, https://www.mdpi.com/2076-2607/13/2/467
12. Conducting a Microbiome Study - PMC, acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5074386/
13. Improved DNA Extraction and Amplification Strategy for 16S rRNA Gene Amplicon-Based Microbiome Studies - ResearchGate, acessado em maio 11, 2025, https://www.researchgate.net/publication/378736037_Improved_DNA_Extraction_and_Amplification_Strategy_for_16S_rRNA_Gene_Amplicon-Based_Microbiome_Studies

14. Improved DNA Extraction and Amplification Strategy for 16S rRNA ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10932036/
15. Fine-scale evaluation of two standard 16S rRNA gene ... - Frontiers, acessado em maio 11, 2025, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1140487/full
16. Evaluating Bias of Illumina-Based Bacterial 16S rRNA Gene Profiles ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4178620/
17. Standardization of 16S rRNA gene sequencing using nanopore long read sequencing technology for clinical diagnosis of culture negative infections - Frontiers, acessado em maio 11, 2025, https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2025.1517208/full
18. Standardization of 16S rRNA gene sequencing using nanopore long ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11922894/
19. The clinical utility of Nanopore 16S rRNA gene sequencing for direct bacterial identification in normally sterile body fluids - Frontiers, acessado em maio 11, 2025, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1324494/full
20. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8544895/
21. Amplicon analysis with QIIME2 - VL microbiome project, acessado em maio 11, 2025, https://rachaellappan.github.io/VL-QIIME2-analysis/pre-processing-of-sequence-reads.html
22. An independent evaluation in a CRC patient cohort of ... - Frontiers, acessado em maio 11, 2025, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1178744/full
23. ASV vs OTUs clustering: Effects on alpha, beta, and gamma ... - PLOS, acessado em maio 11, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0309065
24. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC6964864/
25. QIIME2 workflow | CHMI services, acessado em maio 11, 2025, https://chmi-sops.github.io/mydoc_qiime2.html
26. 16S rRNA analysis using Qiime2 : r/bioinformatics - Reddit, acessado em maio 11, 2025, https://www.reddit.com/r/bioinformatics/comments/1bf9fxm/16s_rrna_analysis_using_qiime2/
27. The Power of Microbiome Studies: Some Considerations on Which ..., acessado em maio 11, 2025, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.7

96025/full
28. Key features and guidelines for the application of microbial alpha ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11698868/
29. Alpha Diversity - One Codex Docs, acessado em maio 11, 2025, https://docs.onecodex.com/en/articles/4136553-alpha-diversity
30. Beta Diversity | One Codex Docs, acessado em maio 11, 2025, https://docs.onecodex.com/en/articles/4150649-beta-diversity
31. Differential Abundance | Knowledge Base, acessado em maio 11, 2025, https://kb.ezbiocloud.net/home/protocols/shotgun-microbiome/analyze-datasets/differential-abundance
32. Comparison study of sixteen differential abundance methods using two large Parkinson disease gut microbiome datasets | bioRxiv, acessado em maio 11, 2025, https://www.biorxiv.org/content/10.1101/2021.02.24.432717v1.full
33. Analysis of Airborne Microbial Communities Using 16S ribosomal RNA: Potential Bias due to Air Sampling Stress, acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5805565/
34. Microbial Community Study: Correlation Analysis of 16S and ITS ..., acessado em maio 11, 2025, https://www.cd-genomics.com/microbioseq/resource-microbial-community-study-correlation-analysis-of-16s-and-its-sequencing.html
35. The applications of the 16S rRNA gene in microbial ecology: current ..., acessado em maio 11, 2025, https://www.researchgate.net/publication/281735309_The_applications_of_the_16S_rRNA_gene_in_microbial_ecology_current_situation_and_problems
36. Research progress on the application of 16S rRNA gene ... - Frontiers, acessado em maio 11, 2025, https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2024.1360457/full
37. On the limits of 16S rRNA gene-based metagenome prediction and ..., acessado em maio 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10926695/
38. Bioinformatics for Multi-Omics Data Integration | Request PDF - ResearchGate, acessado em maio 11, 2025, https://www.researchgate.net/publication/384905024_Bioinformatics_for_Multi-Omics_Data_Integration
39. Full article: Multi-omic approaches for host-microbiome data ..., acessado em maio 11, 2025, https://www.tandfonline.com/doi/full/10.1080/19490976.2023.2297860