


2

METAGENÔMICA E AMPLICON: PERGUNTAS FREQUENTES E RESPOSTAS ESSENCIAIS

Autores 2.1

Sávio de Souza Costa 

Revisão: Diego Mariano 

Cite este artigo 2.1

Costa, SS. **Metagenômica e Amplicon: perguntas frequentes e respostas essenciais.**
BIOINFO. ISSN: 2764-8273. Vol. 3. p.02 (2023). doi: 10.51780/bioinfo-03-02

Resumo 2.1

Neste artigo, você irá aprender sobre Metagenômica e Amplicon.

2.1 O que é Metagenômica?

OBTER informações genéticas sobre comunidades microbianas presentes nos mais variados ambientes com as técnicas clássicas de biologia molecular possui um grande obstáculo, que é como acessar o genoma desses organismos se apenas 1% das bactérias conhecidas podem ser cultivadas utilizando metodologias atuais? [1]. Dessa forma, uma alternativa foi a clonagem de genes específicos ao invés da clonagem do genoma completo [2].

Na década de 1980, trabalhos moleculares utilizaram a técnica de PCR para explorar a diversidade de sequências de RNA ribossômico, levando a ideia de clonar DNA diretamente de amostras ambientais já em 1985. [3]. Após este ponto de partida, foi publicado o primeiro trabalho a utilizar o termo metagenômica. Este trabalho analisou as bases para clonagem para a análise funcional de microrganismos do solo [2] e, a partir desta técnica, também surgiu a chamada análise por Amplicon. Atualmente a metagenômica e a Amplicon são técnicas que garantem acesso a diversidade e caracterização microbiana nos mais variados ambientes sem a necessidade de cultura dos microrganismos [4].

*"Dessa forma, a **metagenômica** é uma valiosa ferramenta para a descoberta de novos genes, vias metabólicas e enzimas que são de extrema importância biotecnológica e para saúde [5]."*

Antes de se encontrar as respostas pra pergunta fundamental sobre metagenômica, ecologia microbiana e tudo mais. Primeiro deve-se saber a pergunta, e as principais perguntas que são respondidas em um estudo de metagenômica são "Quem são os microrganismos presentes em um determinado ambiente", "O que eles estão fazendo" e "Como estão fazendo isto?".

2.2 Metagenômica e amplicon? São as mesmas coisas?

A resposta rápida é: não! Contudo precisamos entender que no início ambos eram sinônimos para análises de genes de microrganismos amplificados/clonados direto de um determinado ambiente. Atualmente essas análises metagenômicas são diferenciadas e chamadas de metagenômica e amplicon.

*"A terminologia **amplicon** vem da amplificação de um determinado gene marcador presente no ambiente que será analisado."*

Tal gene marcador é geralmente caracterizado como *house-keeping*, ou seja, essencial para certo grupo de microrganismos. Além disso, ele precisa ter outras características, como ser altamente conservado em uma espécie, mas ter certas diferenças em outras espécies. Isso permite que tais genes possam ser usados como uma ferramenta para distinguir diferentes tipos de microrganismos [6].

A utilização do gene **rRNA 16S** como marcador filogenético está consolidada principalmente por apresentar características como: baixa taxa de transferência horizontal, baixa taxa de recombinação gênica, sequência extremamente conservada, mas, que possui regiões hipervariadas que são espécie-específicas [6]. Portanto, a partir da análise dessa região hipervariada, é possível se identificar a nível taxonômico e distinguir diferentes espécies bacterianas [7]. Outros exemplos de genes marcadores de microrganismos incluem o gene **recA** em bactérias, o gene **ITS** em fungos e o gene **SSU rRNA** em eucariotos unicelulares. A importância de utilizar genes bem descritos na literatura permite a realização de comparações entre diferentes comunidades, ressalvadas as diferenças metodológicas empregadas [7][8]. A metodologia padrão desse tipo de análise busca formar clusters com as sequências dos genes marcadores semelhantes, esses clusters são denominados de "Unidade taxonômica operacional" (OTU's) [6].

Já o termo metagenômica é normalmente utilizado para tratar da metagenômica *shotgun* que é uma técnica que permite analisar o material genético de uma amostra complexa contendo todos os genes de muitas espécies diferentes. Ou seja, além de amplificar e obter genes, como o 16S bacteriano, no método *shotgun*, o DNA da amostra é fragmentado aleatoriamente em pequenos pedaços sendo, em seguida, sequenciado em massa para assim se obter uma grande variedade de genes presentes neste ambiente [9][10]. A análise metagenômica *shotgun* tem a uma ampla variedade de aplicações, incluindo a investigação de ecossistemas microbianos em ambientes naturais, estudos de microbiomas humanos para entender a relação entre a microbiota intestinal e a saúde humana, e até mesmo para descobrir novas enzimas e produtos naturais produzidos por microrganismos [11].

2.3 Como se analisa dados de metagenômica

Shotgun?

Esta pergunta é bastante ampla, uma vez que a metagenômica gera uma grande quantidade de dados a serem analisados. Dentro do “tiroteio de *shotgun*”, existem todos os genes presentes na amostra, como genes marcadores como 16S, genes relacionados às funções metabólicas dos microrganismos, dentre outros. Assim, os objetivos fundamentais da metagenômica podem ser a análise do perfil das comunidades bacterianas, do perfil funcional dos genes e até mesmo a montagem metagenômica para tentativa de obter genomas a partir do metagenoma [10][12]. As análises destes dados ocorre através de ferramentas de **Bioinformática** onde após o tratamento das leituras brutas oriundas do sequenciamento metagenômico, as metodologias computacionais podem prosseguir para diversos caminhos dependendo da pergunta fundamental que será respondida.

Com os dados do metagenoma tratados, é possível responder à pergunta “quem são os microrganismos presentes”. Isso pode ser feito comparando as sequências com bancos de dados de sequências conhecidas, como o banco de dados NCBI, SILVA [13] e RDP [14], usando ferramentas como BLAST [15]

ou DIAMOND [16]. Contudo, essa metodologia pode ser considerada bastante trabalhosa. Dessa forma, surgiram softwares que fazem essas análises utilizando seus próprios métodos, como Kraken2 [17], MetaPhlan2 [18], MEGAN [19] e o MGRAST [20].

A utilização destas ferramentas propicia a obtenção da diversidade e abundância dos microrganismos presentes, sendo que cada uma utiliza sua própria metodologia computacional. Por exemplo, uma maior diversidade bacterianas foi encontrada pela metodologia shotgun do que usando o método 16S rRNA, o que permitiu ainda prever a classificação de táxons de maneira mais eficaz em nível de filo e, em menor grau, em nível de gênero [21].

Outra abordagem é a obtenção de *contigs* por meio de softwares conhecidos como montadores. Os montadores convencionais utilizam diversas abordagens, uma delas é a de grafo *De Bruijn*, o qual divide as leituras em k-mers e reduz a demanda de memória do computador. Alguns exemplos de montadores baseados em grafo *De Bruijn* incluem o MetaVelvet [22], IDBA-UD [23], MEGAHIT [24] e o metaSPAdes [25]. A escolha do montador irá depender do dado e variar para cada amostra, por isso sempre importante comparar as montagens utilizando o **METAQUAST**.

Após a obtenção de *contigs*, é possível aplicar diversas metodologias para análises dos dados dos metagenomas. Uma dessas técnicas consiste em construir genomas a partir de metagenomas (MAGs – *Metagenome-assembled genomes*). Nesse caso, o primeiro passo consiste no *binning* das leituras, que irá agrupar as contigs com base em suas características, como base nas frequências de tetranucleotídeos (TNFs), abundância de genes marcadores e uso de códons [26]. Os softwares mais utilizados com intuito de obter os genomas são o MetaBAT [27], CONCOCT [28] e MaxBin2 [29]. O grau de contaminação de um MAG pode ser analisado através do software CheckM [30]. Assim, a taxa de contaminação depende do método de obtenção do MAG e da diversidade dos microrganismos na amostra do metagenoma. Os MAGs que apresentam alta completude e baixos níveis de contaminação são então selecionados para posterior anotação taxonômica e predição de genes.

As análises apresentadas a seguir podem ser utilizadas tanto para contigs montados quanto para os MAGs preditos. Além disso, eles respondem a duas perguntas funcionais que são: “o que estão fazendo” e “como estão fazendo?”. Dessa forma, para fazer a análise das funções metabólicas dos genes encontrados, pode-se utilizar ferramentas baseadas em homologia, como BLAST [15], para comparar as sequências de genes previstos com as de genes conhecidos. No entanto, métodos modernos, como eggNOG-mapper [31], GhostKOALA [32], MG-RAST [33] e PANNZER2 [34], empregam estratégias de alinhamento otimizadas que permitem alinhamentos rápidos de sequências de genes com bancos de dados. O MG-RAST [33] fornece uma interface de análise metagenômica *online* que inclui *upload* de dados, controle de qualidade e alinhamento com bancos de dados de referência. Ele permite a análise tanto funcional dos genes quanto a predição das comunidades microbianas ali presentes. Dessa forma, percebe-se que a abordagem de quem está presente, o que estão fazendo e como estão fazendo possui vários caminhos e, tudo isso, depende da metodologia escolhida para obter a resposta mais completa. Entretanto, estas são só algumas das ferramentas mais utilizadas para obter estas respostas.

2.4 Como se analisa dados de amplicon?

A tecnologia de amplicon é uma abordagem mais comum para a análise de dados ambientais, pois utiliza a amplificação de regiões específicas do DNA para estudar a diversidade e função de comunidades microbianas em diferentes ambientes. Essa abordagem pode ser usada para estudar comunidades microbianas em diversos tipos de amostras, incluindo solo, água, fezes e mucosas. Uma das principais vantagens da tecnologia de amplicon é a sua alta sensibilidade e especificidade, que permite a detecção de baixas abundâncias de microrganismos e a identificação de espécies específicas dentro de uma comunidade complexa. Além disso, a tecnologia de amplicon é relativamente simples e acessível, o que torna essa abordagem uma das mais populares na análise de dados de metagenômica [35] [36].

A técnica de amplicon é baseada em homologia e predição. As espécies podem ser identificadas com base na sequência lida da região variável dos genes

marcadores. O método requer o alinhamento de uma sequência do gene escolhido com todas as sequências de um banco de dados de referência. Dentre os bancos de dados utilizados, encontram-se alguns similares a metagenômica *shotgun*, como SILVA [13] e RDP [14]. Algumas ferramentas e *pipelines* estão disponíveis para a análise dessas sequências de forma automatizada, como QIIME (*Quantitative Insights Into Microbial Ecology*) [37], MOTHUR [38] e USEARCH [39], bem como opções mais recentes DADA2 [40] e Qiime2-Deblur [41].

Os softwares QIIME, MOTHUR e USEARCH-UPARSE agrupam sequências com 97% de identidade em Unidades Taxonômicas Operacionais (OTUs). Já os Qiime2-Deblur, DADA2 e USEARCH-UNOISE3 tentam reconstruir as sequências biológicas exatas presentes na amostra, chamadas de *Amplicon Sequence Variants* (ASVs). Assim, uma OTU representa um grupo de sequências muito próximas (>97% de identidade), que se separa das demais OTUs pela aplicação de técnicas de agrupamento hierárquico utilizando limites de identidade de sequência independentemente de inferências filogenéticas [7].

ASVs são referidos por outros autores como “*zero noise OTUs*” ou “*sub-OTUs*”. Assim, elas buscam identificar e distinguir sequências de amplicons individuais com base em diferenças nucleotídicas. Em vez de agrupar as sequências em OTUs, o método ASV atribui um número de identificação único para cada sequência de amplicon presente nos dados, representando assim variantes únicas. O uso de ASVs permite uma resolução mais alta na análise da diversidade microbiana. Ao considerar cada sequência individualmente, é possível identificar diferenças sutis entre as variantes, como mutações ou polimorfismos, que poderiam ser agrupados em uma única OTU utilizando uma abordagem baseada em similaridade [42][43].

O sequenciamento de metagenoma também é particularmente útil no estudo de comunidades virais. Como os vírus carecem de um marcador filogenético universal compartilhado, a única maneira de acessar a diversidade genética da comunidade viral de uma amostra ambiental é por meio da metagenômica. A metagenômica tem o potencial de avançar o conhecimento em uma ampla variedade de campos. Também pode ser aplicado para resolver desafios práticos em medicina, engenharia, agricultura, sustentabilidade e ecologia [11][21]. Por

fim, cabe ressaltar que a escolha entre as abordagens de metagenômica por amplicon ou *shotgun* depende principalmente dos objetivos da pesquisa, dos recursos disponíveis e do tipo de amostra a ser analisada.

Saiba mais 2.1

Este artigo está disponível em <https://bioinfo.com.br/a-bioinformatica-na-metagenomica-e-amplicon-perguntas-frequentes-e-respostas-essenciais/>

2.5 Referências

- [1] Hawksworth, D. L. The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological Research*, 105(12), 1422–1432. (2001). doi:10.1017/s0953756201004725.
- [2] Handelsman J, Rondon MR, Brady SE, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* Oct;5(10):R245-9. (1998). doi: 10.1016/s1074-5521(98)90108-9. PMID: 9818143.
- [3] Pace NR, Stahl DA, Lane DJ, Olsen GJ. “The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences”. In Marshall KC (ed.). *Advances in Microbial Ecology*. Vol. 9. Springer US. pp. 1–55. (1986). doi:10.1007/978-1-47570611-6.
- [4] Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Huttenhower, C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9), 814–821. (2013). doi:10.1038/nbt.2676.
- [5] Culligan, E. P., Marchesi, J. R., Hill, C., Sleator, R. D. Combined metagenomic and phenomic approaches identify a novel salt tolerance gene from the human gut microbiome. *Frontiers in Microbiology*, 5. (2014). doi:10.3389/fmicb.2014.00189
- [6] Tikhonov, Mikhail, Robert W. Leach, and Ned S. Wingreen. “Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution.” *The ISME journal* 9.1 (2015): 68-80.
- [7] Yarza, P., Yilmaz, P., Pruesse, E. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 12, 635–645 (2014). <https://doi.org/10.1038/nrmicro3330>
- [8] Harris, J. K., Kelley, S. T. Pace, N. R. New perspective on uncultured bacterial phylogenetic division OP11. *Appl. Environ. Microbiol.* 70, 845–849 (2004).

- [9] Wang, T. et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* 6, 320–329 (2012).
- [10] Quince, C., Walker, A., Simpson, J. et al. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 35, 833–844 (2017). <https://doi.org/10.1038/nbt.3935>
- [11] Madhavan, A., Sindhu, R., Parameswaran, B. et al. Metagenome Analysis: a Powerful Tool for Enzyme Bioprospecting. *Appl Biochem Biotechnol* 183, 636–651 (2017). <https://doi.org/10.1007/s12010-017-2568-3>
- [12] Scholz, M. et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13, 435–438 (2016).
- [13] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Opens external link in new windowNucl. Acids Res.* 41 (D1): D590-D596. (2013).
- [14] Cole, J. R. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research*, 31(1), 442–443. (2003). doi:10.1093/nar/gkg039
- [15] Altschul S, Gish W, Miller W, Myers E, Lipman D: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403–410.
- [16] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015 Jan;12(1):59-60. doi: 10.1038/nmeth.3176. Epub 2014 Nov 17. PMID: 25402007.
- [17] Wood, D.E., Lu, J. Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20, 257 (2019). <https://doi.org/10.1186/s13059-019-1891-0>
- [18] Truong, D., Franzosa, E., Tickle, T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 12, 902–903 (2015). <https://doi.org/10.1038/nmeth.3589>
- [19] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res.* 2007 Mar;17(3):377-86. doi: 10.1101/gr.5969107. Epub 2007 Jan 25. PMID: 17255551; PMCID: PMC1800929.
- [20] Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol.* 2016;1399:207-33. doi: 10.1007/978-1-4939-3369-3_3. *PMID* : 26791506.
- [21] Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun.* 2016 Jan 22;469(4):967-77. doi: 10.1016/j.bbrc.2015.12.083. Epub 2015 Dec 22. PMID: 26718401; PMCID: PMC4830092.

- [22] Namiki, T., Hachiya, T., Tanaka, H., Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20), e155. DOI: 10.1093/nar/gks678
- [23] Peng, Y., Leung, H. C., Yiu, S. M., Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11), 1420-1428. DOI: 10.1093/bioinformatics/bts174
- [24] Li, D., Liu, C. M., Luo, R., Sadakane, K., Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10), 1674-1676. DOI: 10.1093/bioinformatics/btv033
- [25] Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27(5), 824-834 DOI: 10.1101/gr.213959.116
- [26] Setubal JC. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys Rev*. 2021 Nov 4;13(6):905-909. doi: 10.1007/s12551-021-00865-y. PMID: 35059016; PMCID: PMC8724365.
- [27] Kang, D. D., Froula, J., Egan, R., Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3, e1165. DOI: 10.7717/peerj.1165
- [28] Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11), 1144-1146. DOI: 10.1038/nmeth.3103
- [29] Wu, Y. W., Simmons, B. A., Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), 605-607. DOI: 10.1093/bioinformatics/btv638
- [30] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14. PMID: 25977477; PMCID: PMC4484387.
- [31] Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., ... Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, 47(D1), D309-D314 DOI: 10.1093/nar/gky1085
- [32] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1), D457-D462. DOI: 10.1093/nar/gkv1070
- [33] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., ... Edwards, R. A. (2008). The metagenomics RAST server—a public resource for the automatic

phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1), 386.
DOI: 10.1186/1471-2105-9-386

[34] Tonini, M., Ureta-Vidal, A., Bateman, A. (2021). PANNZER2: a rapid functional annotation web server. *Nucleic acids research*, 49(W1), W542-W546. DOI: 10.1093/nar/gkab408

[35] Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell*. (2021) May;12(5):315-330. doi: 10.1007/s13238-020-00724-8. Epub 2020 May 11. PMID: 32394199; PMCID: PMC8106563.

[36] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. (2010) ;7:335–336

[37] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5), 335-336. DOI: 10.1038/nmeth.f.303

[38] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23), 7537-7541. DOI: 10.1128/AEM.01541-09

[39] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460-2461. DOI: 10.1093/bioinformatics/btq461

[40] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7), 581-583. DOI: 10.1038/nmeth.3869

[41] Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z. Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, 2(2), e00191-16. DOI: 10.1128/mSystems.00191-16.

[42] Callahan, B. J., McMurdie, P. J., Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12), 2639-2643. DOI: 10.1038/ismej.2017.119

[45] Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, 34(14), 2371-2375. DOI: 10.1093/bioinformatics/bty113