Bioinformatics, 2024, 5–31
doi: 10.1093/bioinformatics/12345
Advance Access Publication Date: 31 May 2024
Open Access Article

Methods for eDNA

Diversidade alfa inalterada por rarefação ou *Total Sum Scaling*: evidências de um estudo de simulação

Israel Cassiano-Oliveira[1*]

[1]Departamento de Hidrobiologia, Universidade Federal de São Carlos

*Autor correspondente.

Editor Associado: Célio Santos Dias Júnior

Recebido em 31/05/2024; revisado em 31/05/2024; aceito em 31/05/2024

Resumo

Motivação: Por anos a escolha de métodos para normalização de tabelas de abundância de unidades taxonômicas operacionais (OTUs) tem sido tema de disputa na ecologia. Nesse cenário, dois métodos distintos ganharam destaque: a rarefação e a normalização por *Total Sum Scaling* (TSS). Críticos da rarefação argumentam que tal procedimento prejudica a diversidade das amostras por remover espécies raras e reduz a robustez estatística pela diminuição do n amostral; por outro lado, críticos da normalização por TSS apontam que essa técnica assume que a soma total de sequências por amostra é proporcional à biomassa total da comunidade microbiana, o que nem sempre é verdade, além de ser demasiadamente sensível a *outliers*. Neste trabalho, ambos métodos foram aplicados a um conjunto de dados simulado e diferenças entre os índices de Shannon-Weiner foram avaliadas.

Resultados: Os resultados indicam que a alfa-diversidade provavelmente não é influenciada pelo tipo de normalização adotada, uma vez que tanto a riqueza quanto a uniformidade das amostras se manteve. Entretanto, outras medidas estão sujeitas a distorções. Recomenda-se que, juntamente ao método de normalização, alterações na abundância total, uniformidade e riqueza sejam verificadas e levadas em consideração na análise e interpretação de resultados.

Disponibilidade: Todos os dados, figuras e scripts utilizados na condução deste estudo estão disponíveis no repositório <https://github.com/isracdo/python_biocientistas/tree/main/trabalho_final>.

Contato: israelco@estudante.ufscar.br

Informação suplementar: Dados suplementares disponíveis na *Bioinformatics* online.

1 Introdução

O sequenciamento de *amplicons*, fragmentos de DNA específicos utilizados como marcadores filogenéticos, é uma técnica largamente utilizada para estudar a estrutura de microbiomas, permitindo tanto compreender a composição das comunidades microbianas como obter uma medida das abundâncias de cada táxon (De Filippis *et al*. 2017, Gohl *et al*. 2016, Lundberg *et al*. 2013).

Por meio dessa técnica, obtêm-se tabelas de abundância de sequências (ou *reads*, no inglês) para cada Unidade Taxonômica Operacional (OTUs, nas linhas) em cada amostra (nas colunas), que podem fornecer uma aproximação para o número de indivíduos por táxon em cada amostra e, com isso, podem-se implementar análises de diversidade. No entanto, tais estimativas são limitadas por alguns fatores principais:

- A **ploidia dos organismos** varia de acordo com sua filogenia, e, portanto, o número de sequências marcadoras também. Em geral, um organismo 4N, pela repetição de cromossomos, possui mais cópias da sequência marcadora em comparação com um organismo N, por exemplo. Além disso, a ploidia também pode influenciar a diversidade genética, pois mais cópias do genoma podem levar a uma maior variabilidade genética (Delomas *et al*. 2021, Lavrinienko *et al*. 2021).

- Uma escolha inadequada da **região** ou do **comprimento** da sequência a ser amplificada pode também levar a uma subestimação da riqueza de espécies e a prejuízos no nível de resolução taxonômica (Engelbrektson *et al*. 2010, Lavrinienko *et al*. 2021, Skums *et al*. 2012). Por exemplo, regiões ITS são mais adequadas para detectar fungos do que a região 18s, amplamente usada para eucariotos.
- Uma **profundidade de sequenciamento** insuficiente pode ocasionar perda na riqueza de espécies ao deixar de detectar táxons que não são necessariamente raros, pois a amplificação pode não ser eficaz para todos os táxons (Kelly *et al*. 2019, Majaneva *et al*. 2015).
- **Normalizações** são necessárias para corrigir as assimetrias nas profundidades de sequenciamento (esforço amostral) entre as amostras e assim permitir estabelecer comparações de diversidade

Em especial, os debates acerca de normalização frequentemente dividem opiniões na ecologia microbiana e têm estado longe de atingir um consenso. Dois métodos são mais comumente aplicados: a rarefação, processo de subamostragem aleatória de *reads*, mas que leva em conta as proporções originais de cada táxon em cada amostra, de modo a igualar a profundidade de sequenciamento de todas as amostras ao nível da amostra com menor total de *reads*; e o *Total Sum Scaling* (TSS), que consiste em tomar as abundâncias relativas de cada táxon e multiplicá-las por um valor alto, usualmente o valor de soma da amostra com maior total de *reads*.

Nesse sentido, o presente trabalho propõe-se a contribuir com o debate testando ambos métodos de normalização com um conjunto de dados simulado e examinando potenciais divergências em medidas de diversidade alfa.

2 Métodos

Todos os procedimentos descritos nessa seção foram realizados por algoritmos na linguagem Python (versão 3.10.12), configurando uma *random.seed* para repodutibilidade e utilizando os módulos *numpy*, *pandas*, *scipy*, *seaborn*, *matplotlib* e *collections*.

Primeiro, produziu-se uma tabela de abundâncias absolutas aleatórias para 26 amostras (nas colunas), cada uma com 100 OTUs e uma profundidade de sequenciamento de no máximo 100.000 *reads*. Neste procedimento, zeros foram introduzidos aleatoriamente numa proporção de 40-75% das OTUs por amostra.

Em seguida a rarefação foi aplicada à tabela gerada, de acordo com os seguintes passos:

(1) Obter o valor de soma de *reads* da amostra com menor profundidade de sequenciamento

(2) Gerar uma tabela de abundâncias relativas e uma tabela de zeros com as mesmas dimensões da tabela original

(3) Para cada coluna da tabela de zeros, escolher aleatoriamente uma célula, de acordo com as probabilidades definidas pelas abundâncias relativas na coluna correspondente, e adicionar "1"

(4) Repetir o passo 3 até que a soma de *reads* cada uma das colunas se igualasse ao valor obtido no passo 1

Produzida a tabela rarefeita, implementou-se o TSS sobre a tabela original, seguindo os passos:

(1) Gerar uma tabela de abundâncias relativas

(2) Multiplicar todos os seus valores por 100.000 (profundidade de sequenciamento definida ao gerar a tabela original) e transformar os valores resultantes em números íntegros (int)

Com as três tabelas (original, rarefeita e normalizada por TSS), calculou-se a média e o desvio-padrão do total de *reads* por amostra. Além disso, a sufuciência do esforço amostral foi avaliada por curvas de acumulação de espécies (curva do coletor) para cada uma das amostras das três tabelas. Para tal, amostraram-se aleatoriamente OTUs de acordo com suas respectivas abundâncias relativas em cada amostra. O tamanho de cada amostragem foi definido por um intervalo de 0 até o valor total de *reads* da coluna, de 500 em 500, e cada uma dessas amostragens foi repetida 10 vezes para obter um valor médio de riqueza.

Por fim, a riqueza (S), a uniformidade de Pielou (J) e a alfa-diversidade de Shannon-Wiener (H') foram calculadas para as amostras das três tabelas. Os valores de H' foram comparados entre a tabela original e a tabela rarefeita por teste T pareado, após verificação de normalidade pelo teste de Shapiro-Wilk. O mesmo procedimento de comparação foi implementado para os valores de H' das tabelas original e normalizada por TSS.

Todos os dados, análises e gráficos gerados pelo método descrito nessa seção, bem como os códigos executados, estão disponíveis no repositório:
https://github.com/isracdo/python_biocientistas/tree/main/trabalho_final.

3 Resultados

A tabela de OTUs foi gerada com sucesso, com total de *reads* variando de 57.424 (valor da amostra de menor soma) a 99.920 (valor da amostrade maior soma), e cada amostra com 50-66% de OTUs com abundância zero. As normalizações foram bem-sucedidas e o resultado das análises do total de *reads* por amostra são mostrados na figura 1 e na tabela 1.

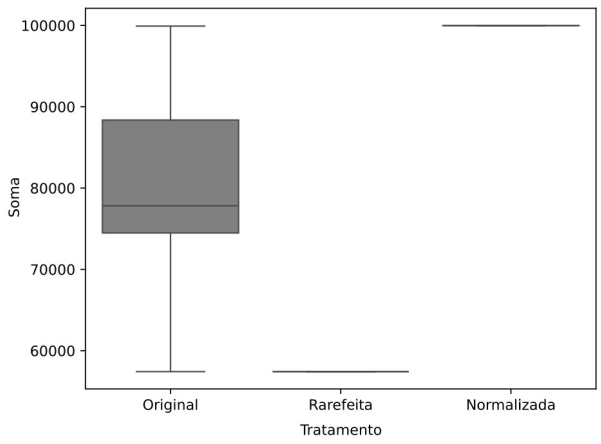

Fig. 1. Boxplot das profundidades de sequenciamento das amostras nas três tabelas (original, rarefeita e normalizada por TSS).

O boxplot da figura 1 e os valores de média na tabela 1 evindenciam o efeito de cada tipo de normalização sobre a profundidade de sequenciamento: na rarefação, a referência para padronização é o total de *reads* mais baixo; no TSS, o mais alto. Em adição, os valores de desvio-

padrão na tabela 1 reiteram a efeitvidade das normalizações em atingir seu propósito: reduzir a disparidade entre amostras.

Tabela 1. Média e desvio-padrão das profundidades de sequenciamento das amostras nas três tabelas (original, rarefeita e normalizada por TSS).

Tratamento	Média	Desvio-padrão
Original	79.375,04	±11.303,35
Rarefeita	57.424,00	±0,00
Normalizada (TSS)	99.979,62	±2,67

As curvas de acumulação de OTUs (curvas do coletor) obtidas para todas as amostras das três tabelas revelam que a profundidade de sequenciamento foi absolutamente adequada para capturar a riqueza em todos os casos . De fato, as curvas atingem a saturação antes de 1000 *reads*. (Por questões de concisão os gráficos das curvas podem ser encontradas no material suplementar). Já em relação às comparações das medidas de S, J e H', nenhuma diferença significativa foi encontrada ($p \geq 0,05$).

4 Discussão

As análises implementadas demonstraram que os procedimentos de normalização testados parecem surtir efeito expressivo sobre a profundidade de sequenciamento, alterando a abundância absoluta de cada OTU e a soma de *reads* por amostra, porém essas mudanças não se refletem nos índices de alfa-diversidade, que não apresentaram diferença significativa.

Para investigar se a estabilidade nos valores de H' poderia ser atribuída a um *tradeoff* entre riqueza e uniformidade, obtiveram-se os valores de S e J para todas as amostras das três tabelas. Percebeu-se então que todos os valores de S eram idênticos, isto é, nenhum dos dois métodos de normalização foi capaz de remover OTUs e diminuir a riqueza. Isso pode ser explicado pela curva do coletor: a perda de espécies só ocorre abaixo de 1000 *reads*, e qualquer profundidade de sequenciamento acima desse valor é suficiente para evitar esse fenômeno – inclusive o valor de 57.424, utilizado na rarefação, o que explica porque esse método não removeu OTUs.

Concomitantemente, os valores de J também foram idênticos em todos os casos, o que pode ser resultado de um aspecto presente em ambos métodos de normalização: a manutenção das proporções originais de OTUs em cada amostra. Assim, se a riqueza não se altera e tampouco a uniformidade, a diversidade alfa se mantém.

Todavia, esses resultados devem ser interpretados com cautela. Primeiro é preciso considerar que, quanto maior o número de OTUs raras e menor a profundidade de sequenciamento mínima, maior a chance de ocorrer perda de riqueza – avaliar essas características é particularmente importante antes de aplicar a rarefação. Segundo, as alterações na abundância absoluta de cada OTU e na soma de *reads* por amostra, decorrentes das normalizações implementadas, podem não ter afetado o Índice de Shannon-Wiener, mas isso de forma alguma assegura que outras medidas comuns, como beta-diversidade, aninhamento, substituição, co-ocorrência de táxons ou até mesmo modelagens de processos ecológicos, por exemplo, não sofrerão distorções relevantes.

Por fim, vale recordar que a profundidade de sequenciamento define o n amostral do conjunto de dados e, portanto, pode afetar a robustez das análises estatísticas executadas. Tendo isso em mente, o TSS, ao normalizar os dados pelo maior total de sequências, pode ajudar a reduzir o viés introduzido pela variação na profundidade de sequenciamento entre amostras, sem prejudicar o n amostral, o que pode proporcionar estimativas mais precisas e comparáveis de medidas ecológicas.

Pesquisadores devem estar cientes dessas limitações e, caso necessário, considerar métodos alternativos (log-transformação, normalização z-score, normalização pela média geométrica, etc) que minimizem os impactos sobre as estimativas ecológicas, sejam adequados às características dos dados utilizados e estejam alinhados aos objetivos e prioridades do estudo.

Agradecimentos

Agradeço ao professor Célio Santos Dias Júnior pela ministração da disciplina "Introdução ao Python para Biocientistas" e aos colegas de turma pelo apoio na realização desse trabalho.

Financiamento

Esse trabalho foi apoiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Conflito de interesses: nada a declarar.

Referências

De Filippis,F. et al. (2017) Different Amplicon Targets for Sequencing-Based Studies of Fungal Diversity. Appl Environ Microbiol, **83**, e00905-17.

Lundberg,D.S. et al. (2013) Practical innovations for high-throughput amplicon sequencing. Nat Methods, **10**, 999–1002.

Gohl,D.M. et al. (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol, **34**, 942–949.

Skums,P. et al. (2012) Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinformatics, **13**, S6.

Delomas,T.A. et al. (2021) Genotyping single nucleotide polymorphisms and inferring ploidy by amplicon sequencing for polyploid, ploidy-variable organisms. Molecular Ecology Resources, **21**, 2288–2298.

Lavrinienko,A. et al. (2021) Does Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial Communities? Trends in Microbiology, **29**, 19–27.

Engelbrektson,A. et al. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. The ISME Journal, **4**, 642–647.

Kelly,R.P. et al. (2019) Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. Sci Rep, **9**, 12133.

Majaneva,M. et al. (2015) Bioinformatic Amplicon Read Processing Strategies Strongly Affect Eukaryotic Diversity and the Taxonomic Composition of Communities. PLoS ONE, **10**, e0130035.