Bioinformatics, 2024
Advance Access Publication Date: Maio, 2024

Universidade Federal de São Carlos

Rarefação vs TSS: Impactos na Detecção de Espécies Diferencialmente Abundantes

[1,*]Fernando Quintiliano Faria Junior

[1]Department of Ecology and Evolutive Biology

Associate Editor: Célio Dias Santos Júnior

Resumo

Motivation: A análise de dados de sequenciamento, como os obtidos em estudos de metagenômica e microbioma, enfrenta um grande desafio: normalizar contagens de sequências (reads) para permitir comparações justas entre amostras. Dois métodos de normalização de dados amplamente discutidos são o Total Sum Scaling (TSS) e a rarefação. Devido ao seu significado metodológico e interpretativo, ambos têm sido objeto de intenso debate na literatura científica. Compreender as implicações e a eficácia de cada método é crucial para garantir a precisão e a robustez das análises de diversidade e abundância em estudos de microbioma. Este estudo se propõe a explorar e comparar esses métodos, contribuindo para a discussão sobre a melhor abordagem para a normalização de dados de sequenciamento.

Results: A curva do coletor obtida pelo método de Total Sum Scaling (TSS) se aproxima mais da curva dos dados brutos em comparação com a rarefação. Observou-se que algumas amostras na curva da rarefação demoram mais para atingir o platô, implicando que espécies com baixas abundâncias podem ser perdidas no processo. A análise de dados brutos ajuda a elucidar a aplicabilidade dos dois tratamentos, uma vez que a geração de dados aleatórios seguiu uma distribuição uniforme. Este estudo contribui para a compreensão das vantagens e limitações de TSS e rarefação na normalização de dados de sequenciamento.

Contact: fernando.faria@estudante.ufscar.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introdução

A análise de dados de sequenciamento, como os obtidos em estudos de metagenômica e microbioma, enfrenta um grande desafio: normalizar contagens de sequências (reads) para permitir comparações justas entre amostras. Dois métodos de normalização de dados amplamente discutidos são escalonamento de soma (TSS) e rarefação. Devido ao seu significado metodológico e interpretativo, ambos têm sido objeto de intenso debate na literatura científica.

O Total Sum Scaling (TSS) envolve dividir o número de leituras de cada amostra pelo número total de leituras daquela amostra, seguido de uma multiplicação por uma constante. Este método é simples e permite comparações diretas entre amostras, assumindo que a distribuição de sequências é proporcional à abundância real das espécies presentes (McMurdie & Holmes, 2014). No entanto, TSS pode ser influenciado por variações na profundidade de sequenciamento e pode não corrigir adequadamente para a presença de leituras raras ou extremas, potencialmente introduzindo viés (Weiss et al., 2017).

Por outro lado, a rarefação envolve a subamostragem aleatória de leituras para igualar a profundidade de sequenciamento entre amostras. Este método visa eliminar a influência da variação no número total de reads, facilitando comparações entre amostras com diferentes profundidades de sequenciamento (Sanders, 1968). Embora a rarefação possa ser útil para padronizar dados, ela também pode resultar na perda de informações, especialmente em amostras com baixa abundância de espécies (Gotelli & Colwell, 2001). Críticas à rarefação incluem a perda de poder estatístico e a introdução de vieses devido à exclusão de leituras (McMurdie & Holmes, 2014).

Portanto, Este estudo se propõe a explorar e comparar esses métodos, contribuindo para a discussão sobre a melhor abordagem para a normalização de dados de sequenciamento

2 Metodologia

Geração de Dados

Para comparar os métodos de rarefação e Total Sum Scaling (TSS), foi gerada uma tabela de Operational Taxonomic Units (OTU) com 26 colunas, representando as amostras, e 100 linhas, representando as espécies. A geração dos dados aleatórios foi realizada utilizando os pacotes pandas e numpy em Python 3. Para aproximar os dados de um sequenciamento real, zeros foram adicionados aleatoriamente nas colunas.

Funções Desenvolvidas

Foram desenvolvidas funções específicas para criar a tabela, gerar dados aleatórios, zerar células aleatórias, realizar a rarefação e a normalização dos dados. Adicionalmente, foram implementadas funções para calcular a diversidade de Shannon e para gerar a curva do coletor. As funções foram estruturadas da seguinte forma:

1. Criação da Tabela: Função para inicializar a tabela OTU com 26 colunas e 100 linhas.
2. Geração de Dados Aleatórios: Função para preencher a tabela OTU com valores aleatórios.
3. Zerar Células Aleatórias: Função para adicionar zeros aleatoriamente nas colunas, simulando dados de sequenciamento reais.
4. Rarefação: Função que amostra os dados aleatoriamente para padronizar o número de reads entre as amostras.
5. Normalização por TSS: Função que normaliza os dados dividindo cada valor pelo total de reads da amostra, multiplicando por uma constante padrão.
6. Índice de Shannon: Função para calcular a diversidade de Shannon em cada amostra.
7. Curva do Coletor: Função para gerar e plotar a curva do coletor a partir dos dados.

Plotagem dos Gráficos

A plotagem dos gráficos foi realizada utilizando o pacote Matplotlib.pyplot. Este pacote foi utilizado para gerar as curvas do coletor e o boxplot, comparando as tabelas (original, normalizada e rarefeita).

Aplicação dos Tratamentos e Comparação

Após a geração dos dados, foram aplicados os tratamentos de TSS e rarefação. As comparações entre os métodos foram realizadas calculando a diversidade de Shannon e gerando curvas do coletor para ambos os tratamentos. As curvas e os índices resultantes permitiram uma avaliação comparativa entre os métodos.

Disponibilidade dos Dados

Todos os dados e códigos utilizados neste estudo estão disponíveis em nosso repositório GitHub, acessível pelo link: [linkgithub].

3 Resultados e Discussão

Métodos de tratamento de dados são amplamente discutidos na comunidade científica. Antes de analisar qualquer dado, é crucial compreender a natureza dos dados, sua ordenação e o contexto em que estão inseridos. No presente trabalho, foram gerados dados aleatórios para realizar as análises, porém, é importante ponderar que esses dados gerados possuem uma distribuição uniforme, sem vieses, o que representa um contexto ideal, mas não real. Em situações reais, os dados de sequenciamento apresentam variabilidade significativa e podem incluir diversos tipos de vieses, como diferenças no esforço de amostragem, variações técnicas, e a presença de ruído biológico. Apesar dessas limitações, a utilização de dados aleatórios permite uma avaliação controlada das técnicas de normalização e rarefação.

3.1 Índice de Shannon

O Índice de Shannon é uma medida de diversidade que leva em consideração tanto a abundância quanto a equitabilidade das espécies presentes em uma determinada comunidade. Ele é calculado com base na distribuição de frequência das espécies em uma amostra e fornece uma medida quantitativa da complexidade da comunidade. Quanto maior o Índice de Shannon, maior a diversidade da comunidade (Shannon, C. E.,1948).

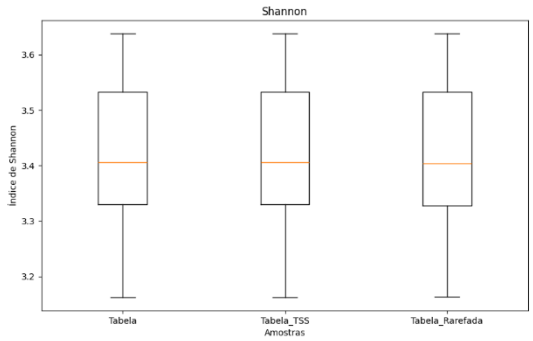

Fig 1. Boxplot dos índices de Shannon no controle e nos dois tratamentos

Na Figura 1, é evidente uma discrepância mínima na diversidade entre os tratamentos. Este fenômeno provavelmente decorre da homogeneidade dos dados, influenciada pela distribuição aleatória dos mesmos.

3.2 Análises dos Tratamentos

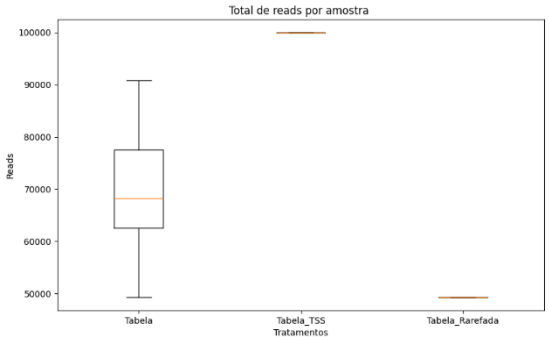

Fig 2. Boxplot do total de reads por amostra, no contole e nos dois tratamentos

Tabelas	Média	Desvio Padrão
Tabela Original	69318.807692	10712.801453
Tabela TSS	99981.192308	3.237520
Tabela Rarefada	49289.000000	0.00000

Tabela 1. Media e desvio padrão do total de reads por amostra no controle e nos dois tratamentos

No processo de tratamento de dados usando a rarefação ou o Total Sum Scaling (TSS), o principal objetivo é equalizar as contagens totais entre as amostras. Na rarefação, como foi determinado que a amostragem ocorreria até atingir o máximo de contagens da amostra com menor número de reads, todas as amostras resultaram com 49.289 reads. Devido à ausência de variância neste caso, a Figura 2 mostra apenas um traço.

Para o TSS, embora o método ajuste as contagens para refletir uma abundância relativa, multiplicando as proporções por um fator comum, os números de reads resultantes tendem a se transformar em valores decimais. Durante o processo de conversão para inteiros, ocorre uma pequena variância, como demonstrado na Tabela 1.

Somente a tabela original apresentou variância, o que era esperado, pois estes são os dados brutos, sem ajustes para igualar ou aproximar os totais

de reads por amostra. Após o tratamento dos dados pela rarefação e pelo TSS, percebe-se que o total de reads por amostra se iguala ou se aproxima, tornando as amostras mais comparáveis.

3.2.1 Curvas do coletor

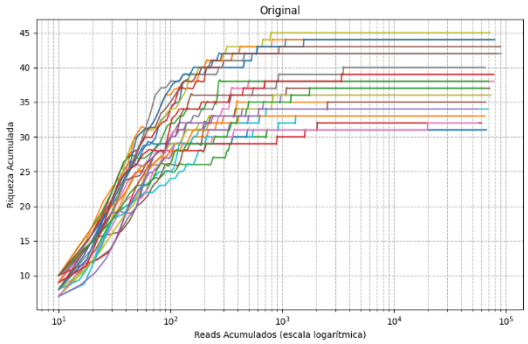

Fig 3. Curva do coletor dos dados originais

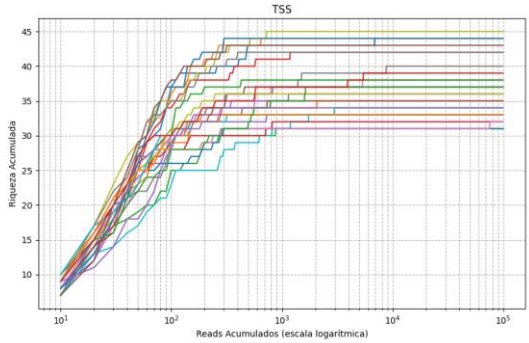

Fig 4. Curva do coletor dos dados normalizados

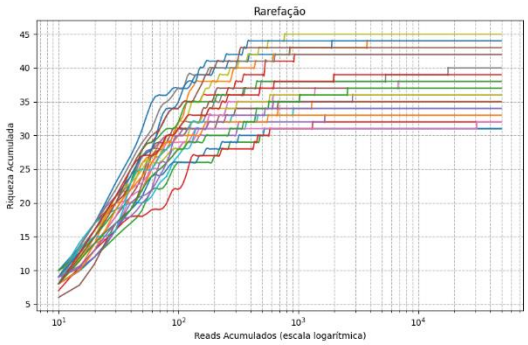

Fig 5. Curva do coletor dos dados rarefados

Na produção da curva do coletor, alguns ajustes foram feitos para melhorar a visualização dos resultados. Por exemplo, o eixo X foi ajustado para uma escala logarítmica de base 10, considerando que a curva mostra um rápido acúmulo das espécies mais abundantes em cada amostra. Enquanto a maioria das espécies foi "observada" nas primeiras subamostragens, algumas espécies raras ou de menor abundância demoraram mais para aparecer.

Comparando as três curvas (Figuras 3, 4 e 5), nota-se que a curva do TSS se aproxima mais dos dados brutos do que dos rarefeitos. Algumas amostras dos dados rarefeitos demoram para atingir o platô, o que implica que a rarefação está sujeita a perdas de espécies raras ou de espécies encontradas em baixas abundâncias.

Segundo Sanders (1968), o método de rarefação para medir a diversidade deve ser utilizado com cautela, especialmente na análise de dados de sequenciamento. É crucial observar que, em casos onde a fauna é distribuída aleatoriamente ou uniformemente, mas não de forma agregada, a medição da diversidade por meio da rarefação permanece válida. No entanto, em situações de agregação, embora o método possa revelar a diversidade inerente, ele apresenta limitações na avaliação precisa dos níveis de diversidade. Portanto, ao utilizar a rarefação na análise de dados de sequenciamento, é essencial considerar os padrões de distribuição das espécies e os possíveis efeitos da agregação nas medições de diversidade.

Diante dessas considerações, a escolha entre TSS e rarefação deve ser informada pelo contexto específico do estudo e pelos objetivos analíticos. Estudos recentes têm explorado alternativas e abordagens híbridas que buscam mitigar as limitações de cada método, sugerindo que uma solução universal pode não ser aplicável a todos os tipos de dados e perguntas de pesquisa (Weiss et al., 2017). Assim, compreender as implicações de cada método é crucial para a realização de análises precisas e significativas em pesquisas de microbiomas e outros campos de biologia de sistemas.

Funding

This work has been supported by the Capes

Conflict of Interest: none declared.

References

GOTELLI, N. J.; COLWELL, R. K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters, v. 4, n. 4, p. 379-391, 2001.

McMURDIE, P. J.; HOLMES, S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Computational Biology, v. 10, n. 4, e1003531, 2014.

SANDERS, H. L. Marine benthic diversity: a comparative study. The American Naturalist, v. 102, n. 925, p. 243-282, 1968.

SHANNON, C. E. A mathematical theory of communication. Bell System Technical Journal, v. 27, n. 3, p. 379-423, 1948.

WEISS, S. et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome, v. 5, n. 1, p. 1-18, 2017.