XVII SIMPÓSIO CIENTÍFICO
DOS PÓS-GRADUANDOS NO CENA
60 anos de pesquisa: passado, presente e futuro

Avaliação do efeito da profundidade dos dados para detecção de *k-mers* exclusivos em cana-de-açúcar

Yuri Dantas Cancelliero[1]*; Gabriely Santos de Oliveira[1]; Prof. Dr. Diego Mauricio Riaño-Pachón[1 2 3]

[1] Laboratório de Biologia Computacional, Evolutiva e de Sistemas (LabBCES-CENA/USP); [2] Instituto Nacional de Ciência e Tecnologia do Bioetanol; [3] Research Centre for Greenhouse Gas Innovation;; *e-mail yuridantas@usp.br

Introdução

A cana-de-açúcar (*Saccharum spp.*) representa uma das culturas de maior importância do planeta, já que é responsável por 80% da produção de açúcar e pioneira na produção de biocombustíveis. Por sua vez, *k-mers* são subsequências curtas de tamanho *k* contidas em sequências maiores, cuja utilidadereside no acesso a propriedades genômicas de forma rápida e eficiente. Para este fim, surgem as abordagens baseadas em *sketches*, subamostras dos conjuntos definidas pelo parâmetro *scaled* (*s*), em que apenas uma fração 1/*s* dos *k-mers* é mantida para análise. Dessa forma, o presente trabalho objetiva investigar e avaliar o efeito dos valores de *k* e *scaled*, bem como da profundidade dos dados de sequenciamento na identificação de *k-mers* exclusivos em 4 espécies de cana-de-açúcar (*S. barberi, S. officinarum, S. robustum,* e *S. spontaneum*).

Materiais e Métodos

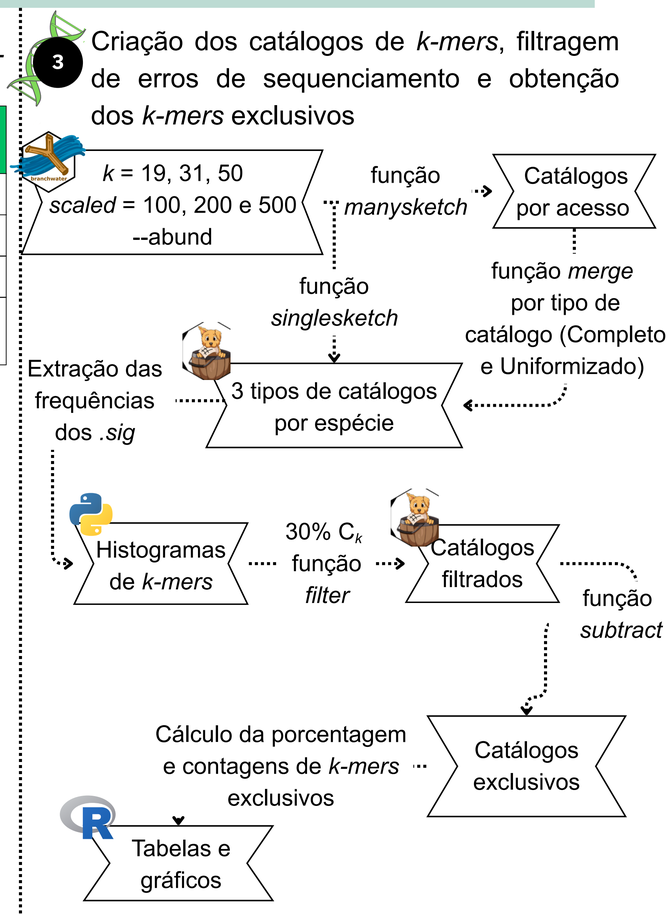
1 Coleta de dados de sequenciamento *paired-end* Illumina

Espécie	Número de acessos	Tamanho (Gbp)	Cobertura (10 Gbp)
S. barberi	26	2287,89	228X
S. officinarum	21	2342,58	234X
S. robustum	23	1807,41	180X
S. spontaneum	8	767,23	76X

2 Definição dos conjuntos

- **Completos** → Todos os acessos de cada espécie
- **Uniformizados** → Quantidade de acessos padronizada para o tamanho em Gpb do menor catálogo (*S. spontaneum*)
- **Normalizados** → Normalização digital dos conjuntos uniformizados:
  - *Software* khmer;
  - *k* = 19;
  - Cobertura de 20X.

3 Criação dos catálogos de *k-mers*, filtragem de erros de sequenciamento e obtenção dos *k-mers* exclusivos



Resultados e Discussão

O valor de *scaled* não altera a detecção de *k-mers* exclusivos, uma vez que as porcentagens foram quase idênticas nos diferentes *k* para os 3 valores testados, o que reafirma sua eficácia em tornar as análises mais rápidas e baratas. Nota-se também uma crescente nas porcentagens de exclusivos em função de *k* em virtude do aumento da espeficidade dos *k-mers*. A exceção reside em *S. officinarum*, cuja sobreposição com *S. robustum* permaneceu estável entre os valores de *k*. Para a profundidade, tem-se influência na detecção de exclusivos, sobretudo em *S. spontaneum*, que teve menos cortes e mostrou aumento em relação ao catálogo completo para todos os valores de *k*.

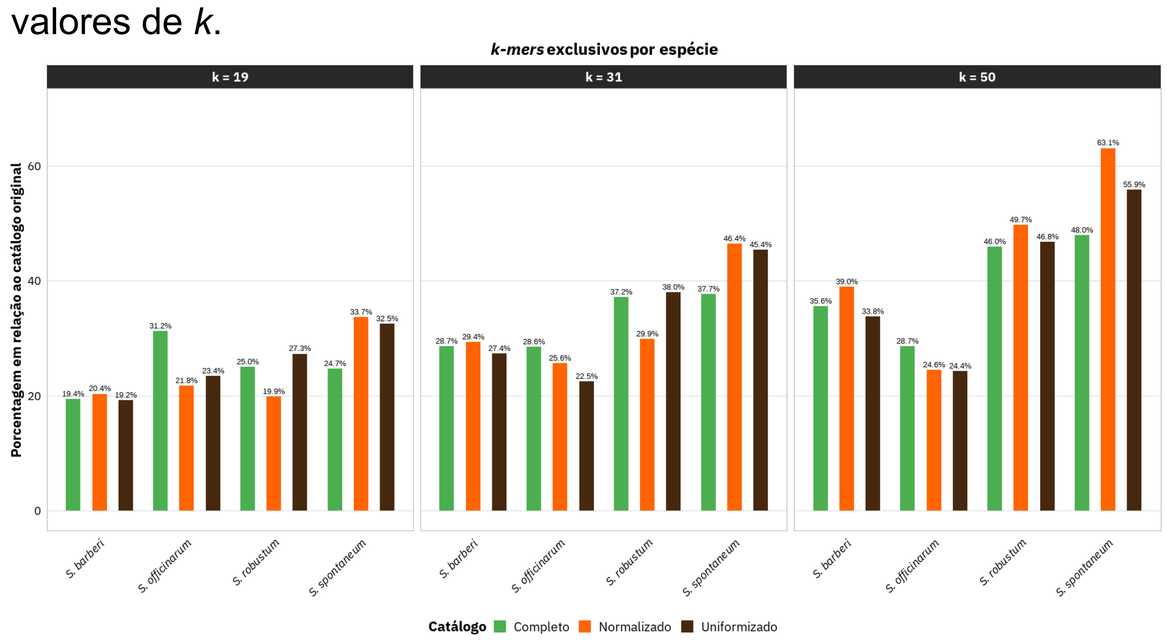


Figura 1. Porcentagens de k-mers exclusivos por espécie e tipo de catálogo para scaled = 500.

Conclusão

Diante do exposto, a ideia da domesticação de *S. officinarum* a partir de *S. robustum* é reforçada pelas subtrações, já que mostraram pouca diferença entre si. A influência da profundidade é vista na redução da cobertura, que aumentou a detecção de exclusivos em *S. spontaneum* pela exclusão de *k-mers* nas etapas de filtragem por que passaram as demais espécies. Portanto, a padronização das coberturas é necessária para análises mais acuradas, de modo a penalizar igualmente todos os conjuntos de dados.

Agradecimentos

Realização | Apoio | Patrocínio