KAD Profiling

21 August 2019

KAD is valuable for evaluating the accuracy of nucleotide base quality of assemblies. Briefly, abundance of k-mers are quantified for both sequencing reads and assembly sequences. Comparison of both values results in a single value per k-mer, K-mer Abundance Difference (KAD), which indicates how well the assembly matches read data for each k-mer.

Run environment and KAD script

R environment:
x86_64-pc-linux-gnu, x86_64, linux-gnu, x86_64, linux-gnu, , 3, 5.1, 2018, 07, 02, 74947, R, R version 3.5.1 (2018-07-02), Feather Spray

KAD script:
seqKADprofile.pl 0.10

INPUT

1. reads
ID: /homes/liu3zhen/scripts2/KAD/data/Xv1601/XV1601.R1.pair.fq.gz
/homes/liu3zhen/scripts2/KAD/data/Xv1601/XV1601.R2.pair.fq.gz

```
/homes/liu3zhen/scripts2/KAD/data/Xv1601/XV1601.R1.pair.fq.gz
/homes/liu3zhen/scripts2/KAD/data/Xv1601/XV1601.R2.pair.fq.gz
```

2. assemblies

```
canu:   /homes/liu3zhen/scripts2/KAD/data/Xv1601/XV1601.v01.canu.fasta
final:  /homes/liu3zhen/scripts2/KAD/data/Xv1601/XV1601Ref1.fasta
```

Parameters for k-mer analysis

1. jellyfish 2.3.0 was used to generate k-mers from sequences
2. length of k-mer:
3. minimum counts of k-mers: 15

Analyis 1: read k-mer analysis

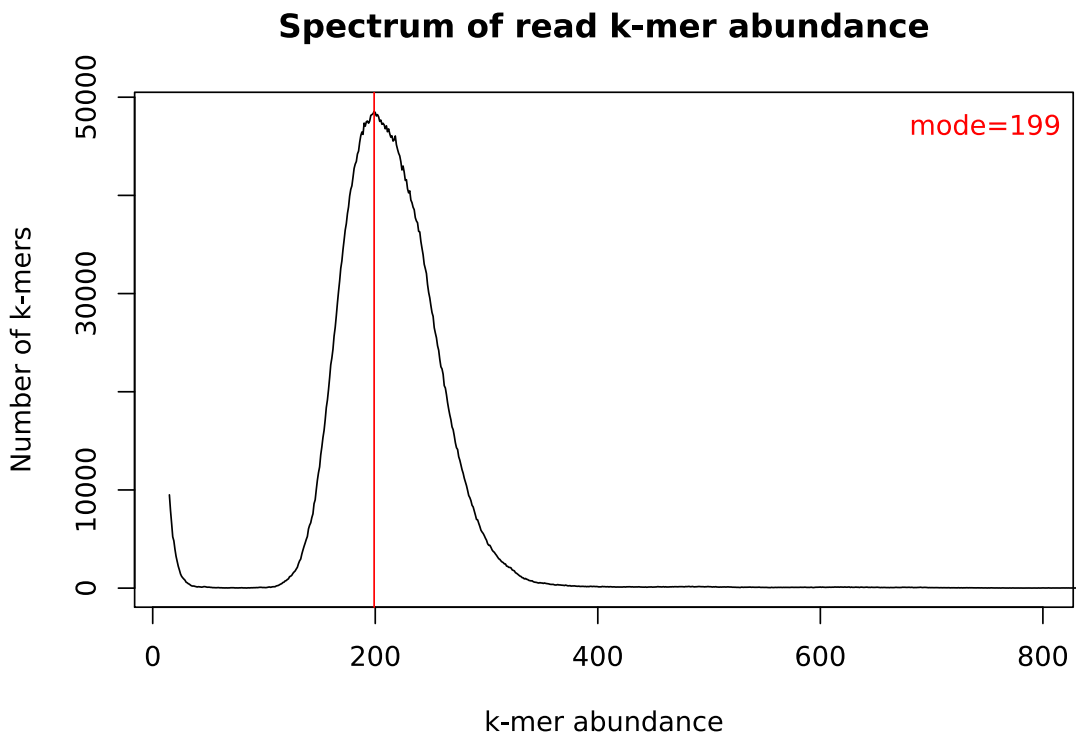

Fig 1. Spectrum of counts of read k-mers

From most genomes, single-copy k-mers each of which is present once in the genome are dominant among all non-redundant k-mers derived from the genome. The mode of sequencing depths of single-copy kmers, representing sequencing depth of read data, can be estimated from the spectrum of k-mer abundance from reads (Fig 1).

The mode of counts of read k-mers: 199. The mode is highlighted in red in Fig1.

Analysis 2: KAD summary for each assembly

Table 1. Summary of KADs

Data	Total	Good	SingleError	MultiError	LowMiss	HighMiss
canu.KAD	4925092	4899849	2649	0	3335	5408
final.KAD	4922443	4906706	0	0	747	5408

Here are criteria used to define k-mer categories:
1. Good: k-mers basically containing no errors; KADs in [-0.5, 0.5]. Note that some k-mers with low counts from reads but absent in the assembly are in this category.
2. SingleError: k-mers showing a single copy in the assembly but with no reads supported; KADs=-1.
3. MultiError: k-mers showing multiple locations in the assembly but read depths indicate lower copies; h KADs <= -2.
4. LowMiss: k-mers with less copies in the assembly as compared to copies indicated by read depths; KADs in [0.75, 2].
5. HighMiss: k-mers showing less copies at a high degree in the assembly as compared to copies indicated by read depths; KADs >=2.

Analysis 3: KAD profiles

Here are tips to understand these KAD profiling figures.
1. “good” k-mers are those with KADs close to 0, representing matching copies indicated by reads and by the assembly.
2. “error” k-mers are located at -1 or smaller. The k-mer count at -1 represents a single-copy error k-mer in the assembly.
3. “missing” k-mers have positive KAD values. The higher values, the higher level of missing in the genome.
4. Left and right figures have different y-axes. Left figure has original count values and right figure has cube root transformation of count values.

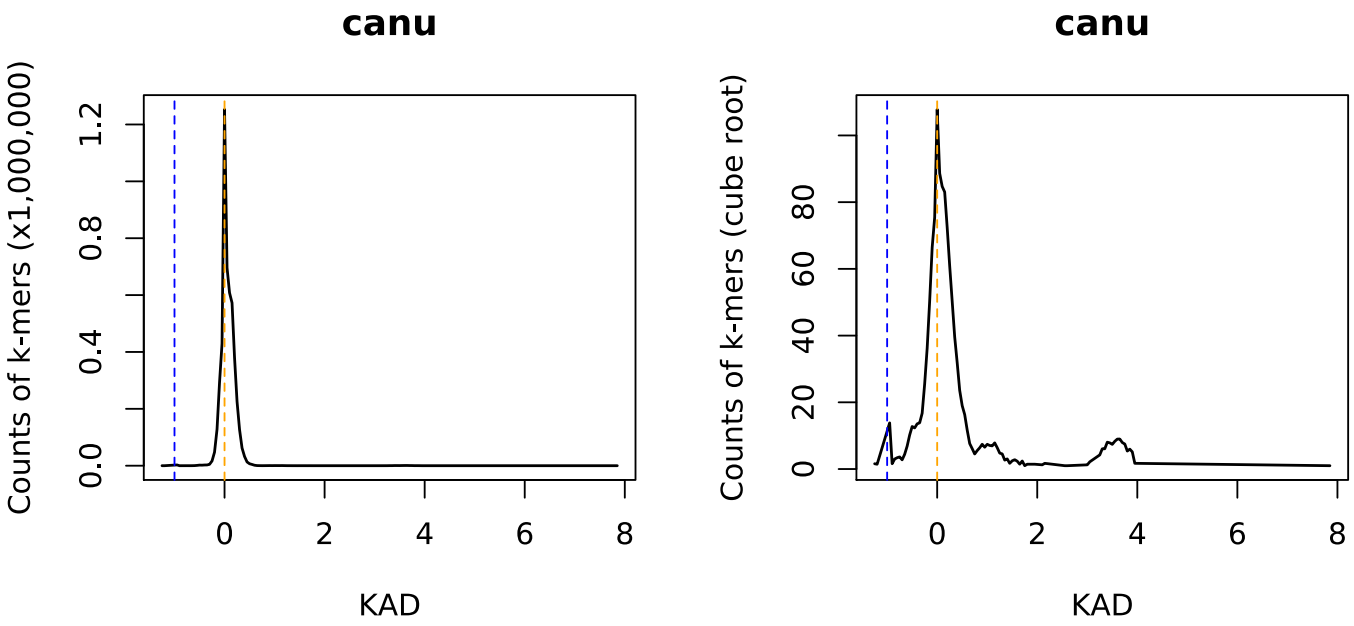

Fig 2. KAD profiles

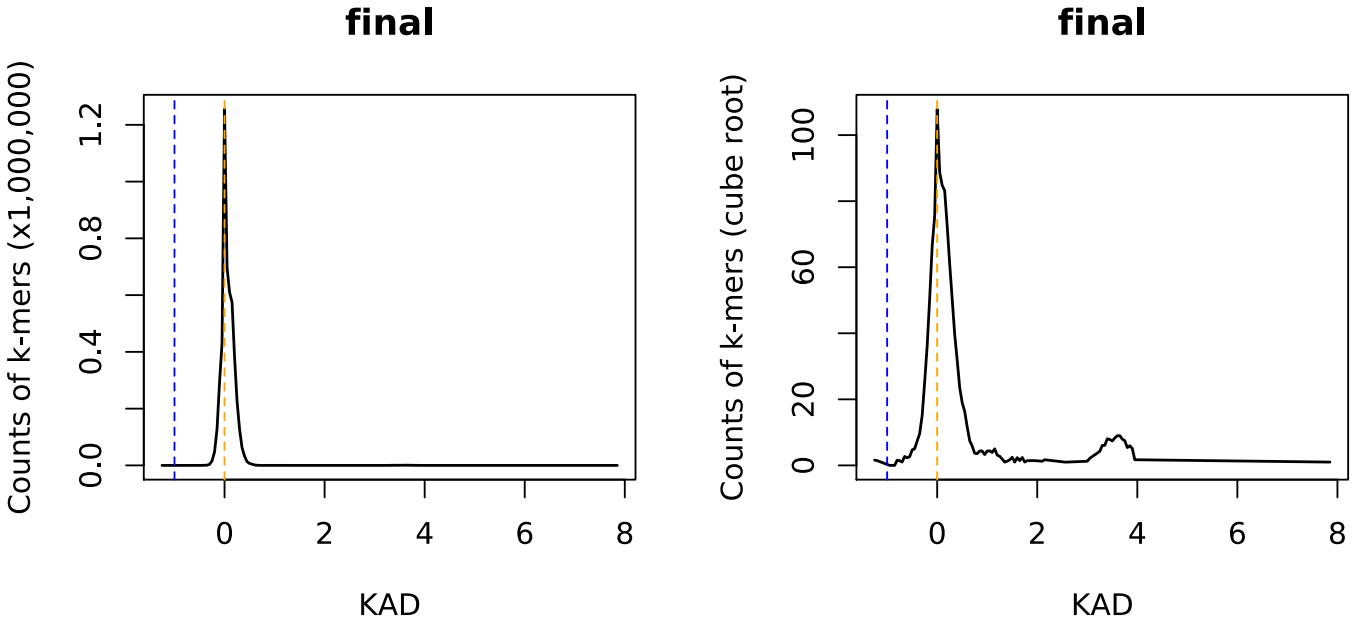

Fig 2. KAD profiles