KAD Profiling

25 February 2020

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, 5.2, 2018, 12, 20, 75870, R, R version 3.5.2 (2018-12-20), Eggshell Igloo

KAD script:
KADprofile.pl 0.15

INPUT

1. reads
ID: reads

```
MG1655_1.fq.gz
MG1655_2.fq.gz
```

2. assemblies

```
U00096.1:   U00096.1.fasta
U00096.3:   U00096.3.fasta
```

Parameters for k-mer analysis

1. jellyfish 2.3.0 was used to generate k-mers from sequences
2. length of k-mer: 25
3. minimum counts of k-mers: 5

Analyis 1: read k-mer analysis

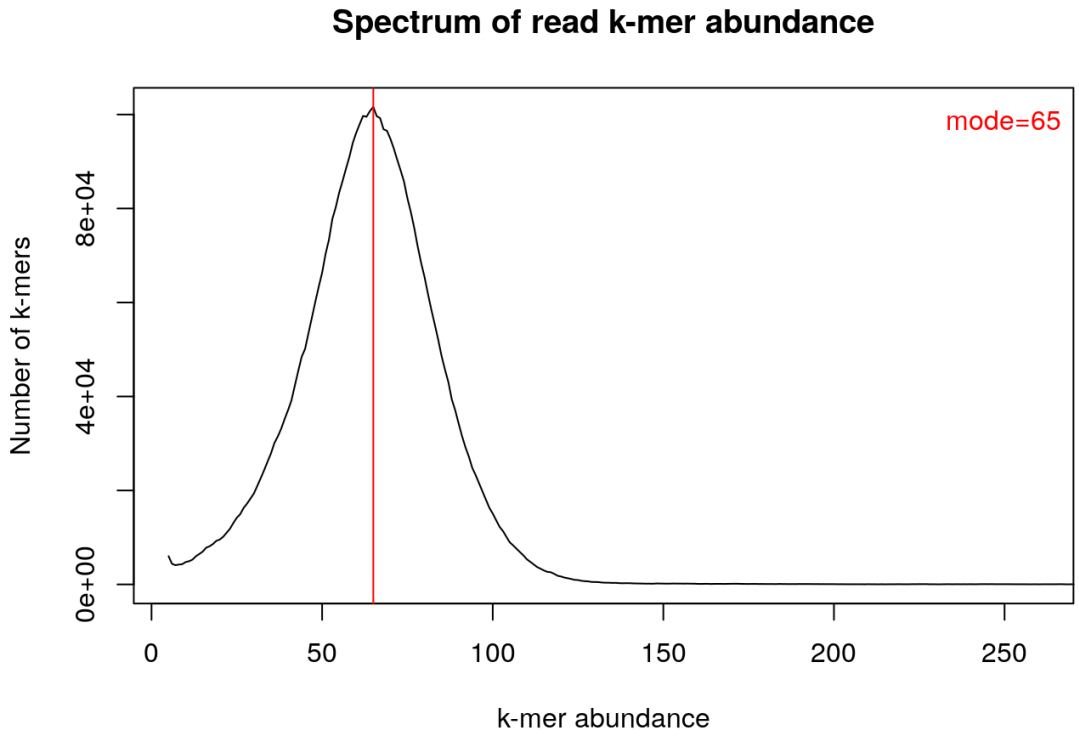

Fig 1. Spectrum of counts of read k-mers

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: 65. The mode is highlighted in red in Fig1.

Analysis 2: KAD summary for each assembly

Table 1. Summary of KADs

Data	Total	Good	Error	OverRep	LowUnderRep	HighUnderRep
U00096.1.KAD	4562071	4354798	15512	3355	4348	56
U00096.3.KAD	4557190	4359867	10627	3348	142	56

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. Error: k-mers showing a single copy in the assembly but with no reads supported; KADs=-1.
3. OverRep: k-mers showing multiple locations in the assembly but read depths indicate lower copies; KADs <= -1 and do not equal -1.
4. LowUnderRep: k-mers with less copies in the assembly as compared to copies indicated by read depths; KADs in [0.75, 2].
5. HighUnderRep: k-mers showing less copies at a high degree in the assembly as compared to copies indicated by read depths; KADs >=2.

Table 2. Estimation of base errors

Data	BaseErrorNum_24	BaseErrorNum_14
U00096.1	646	1108
U00096.3	443	759

Note: Of the header, “BaseErrorNum_xx” represents the number of base errors detected through dividing the number of error k-mers by the conversion rate of “xx”. Because the read depth is >40, the error estimation should be accurate as long as no strong biases in generating read data.

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, representing a single-copy error k-mer in the assembly.
3. “Overepresented” k-mers have negative KAD values smaller than -2, representing some levels of redundancy or systematic sequencing errors. 4. “Underrepresented” k-mers have positive KAD values. The higher values, the higher level of missing in the genome. DNA contamination in Illumina sequencing data and organell DNA sequences are in this category. 5. Left and right figures have different y-axes. Left figure has original count values and right figure has cube root transformed count values.

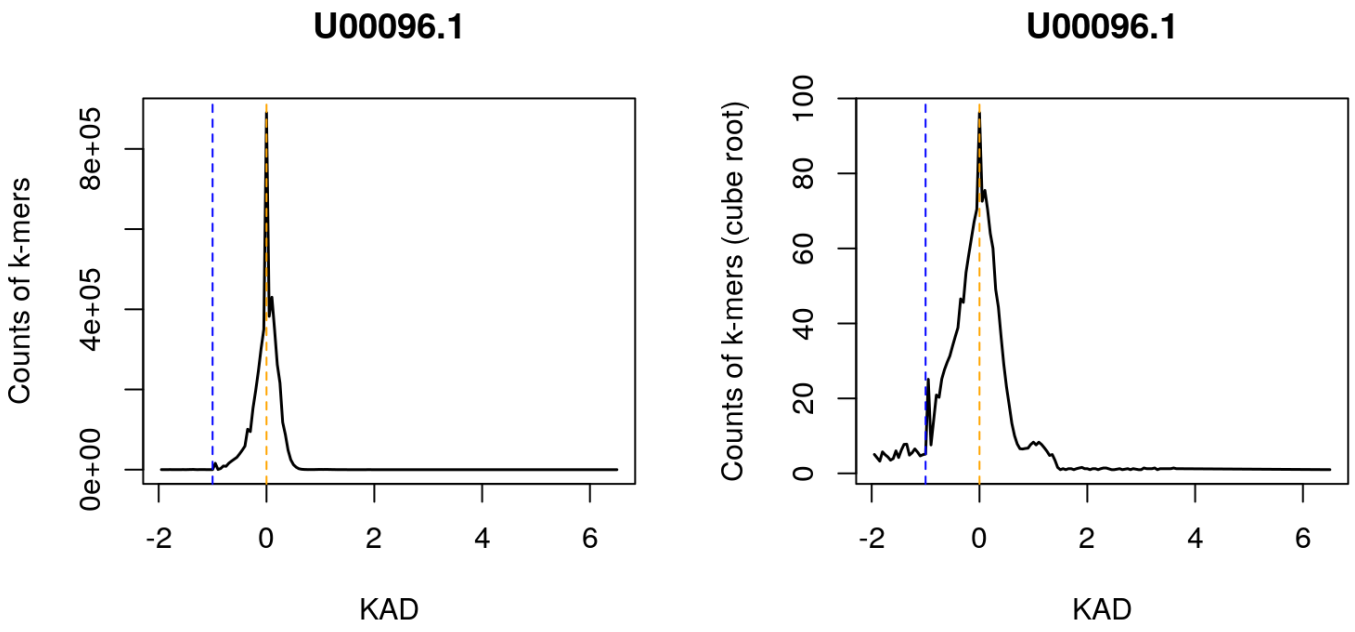

Fig 2. KAD profiles

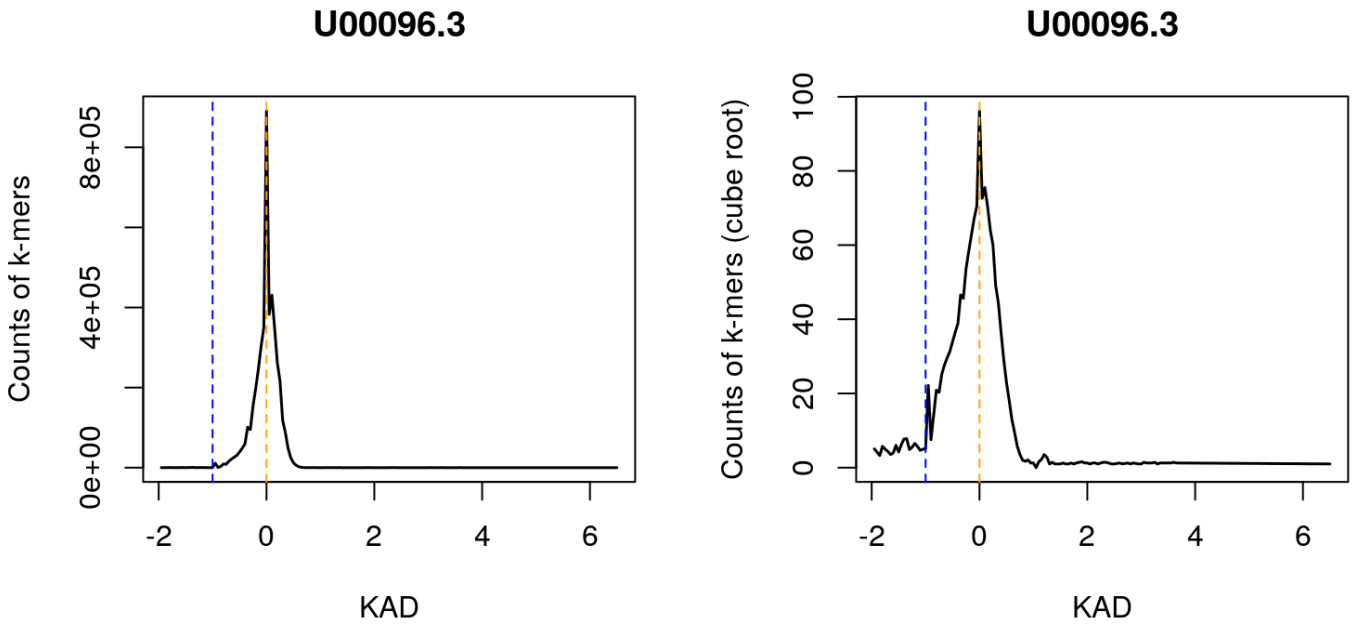

Fig 2. KAD profiles