Assembly of long, error-pront reads using repeat graphs

Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner

05.07.2021

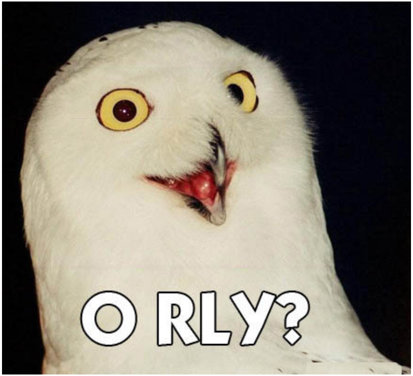
Long reads and de Bruijn graphs?

- de Bruijn graphs need correct bases

Long reads and de Bruijn graphs?

- de Bruijn graphs need correct bases
- otherwise tangled graph

Long reads and de Bruijn graphs?



Repeat graphs

- generalization of de bruijn graphs

```
>- from disjointigs = random walk of reads on the repeat gr
>- means the repeat graph hasn't to be known
```

Repeat graphs

- generalization of de bruijn graphs
- structure

```
>- from disjointigs = random walk of reads on the repeat gr
>- means the repeat graph hasn't to be known
```

Repeat graphs

- generalization of de bruijn graphs
- structure
- creation

```
>- from disjointigs = random walk of reads on the repeat gr
>- means the repeat graph hasn't to be known
```

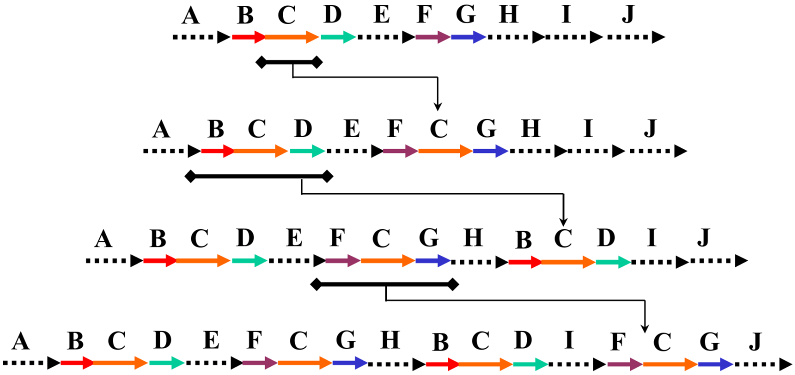
Difference repeat graph de Bruijn graph

Repeat resolution

Results

- human dataset

Segmental duplications


A B C D E F G H I J
A B C D E F C G H I J
A B C D E F C G H B C D I J
A B C D E F C G H B C D I F C G J

- Segmental duplications are duplicated blocks of genomic DNA typically ranging in size from 1-200 kb (IHGSC 2001)

- Segmental duplications are duplicated blocks of genomic DNA typically ranging in size from 1-200 kb (IHGSC 2001)
- They often contain sequence features such as high-copy repeats and gene sequences with intron-exon structure.