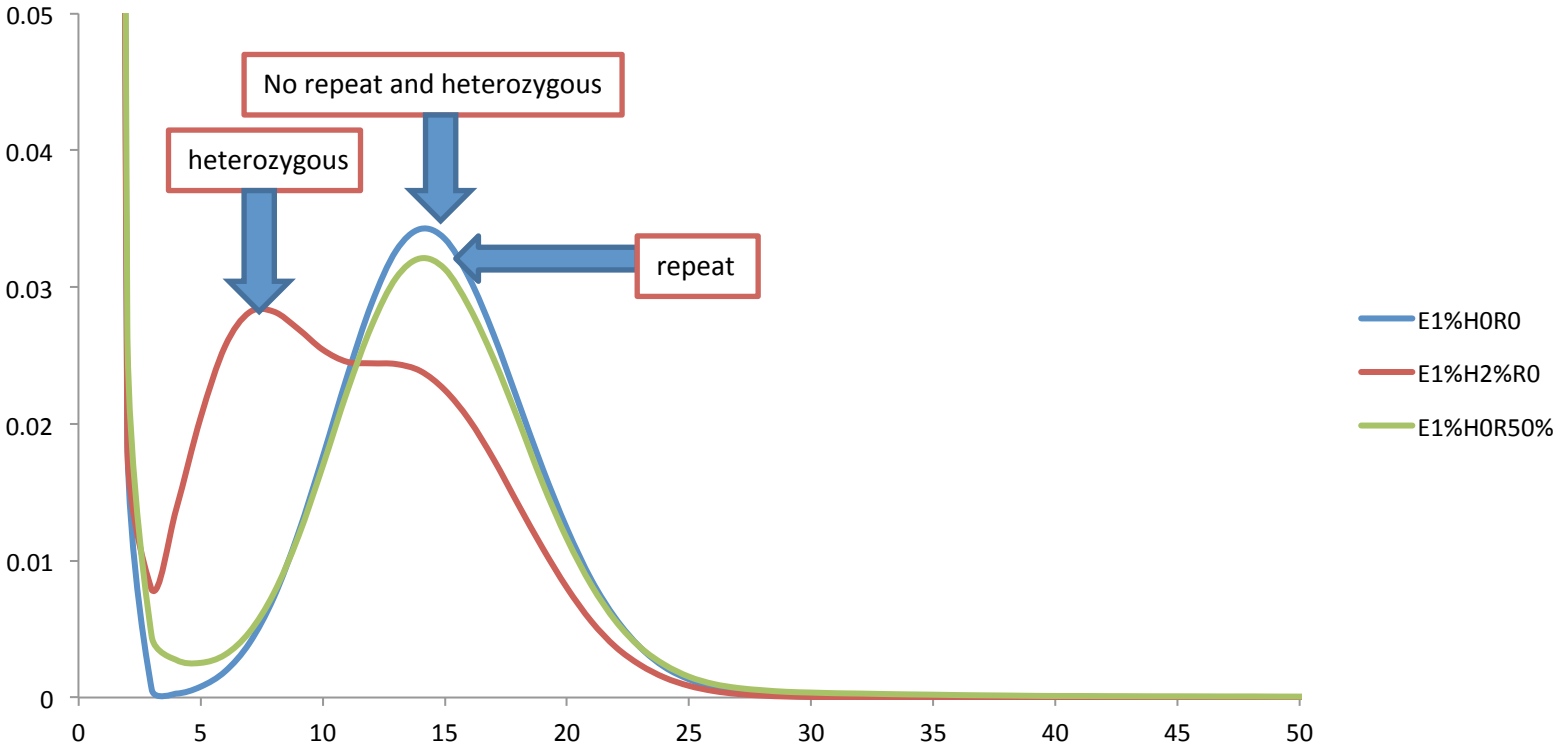
K-mer Analysis in Model

Dongliang Zhan
Bioinformatics Center

Outline

- Importance of k-mer Analysis
- Basic definition
- Algorithm:
  - De-noising from errors
  - Estimate genome size
  - Estimate repeat rate
  - Estimate heterozygosity

Importance of k-mer analysis



- K-mer distribution can reveal the characteristic of the genome

Basic definition

For each read:

kmer_num = read_len-kmer_size+1

For the whole genome

total_kmer_num ≈ genome_size

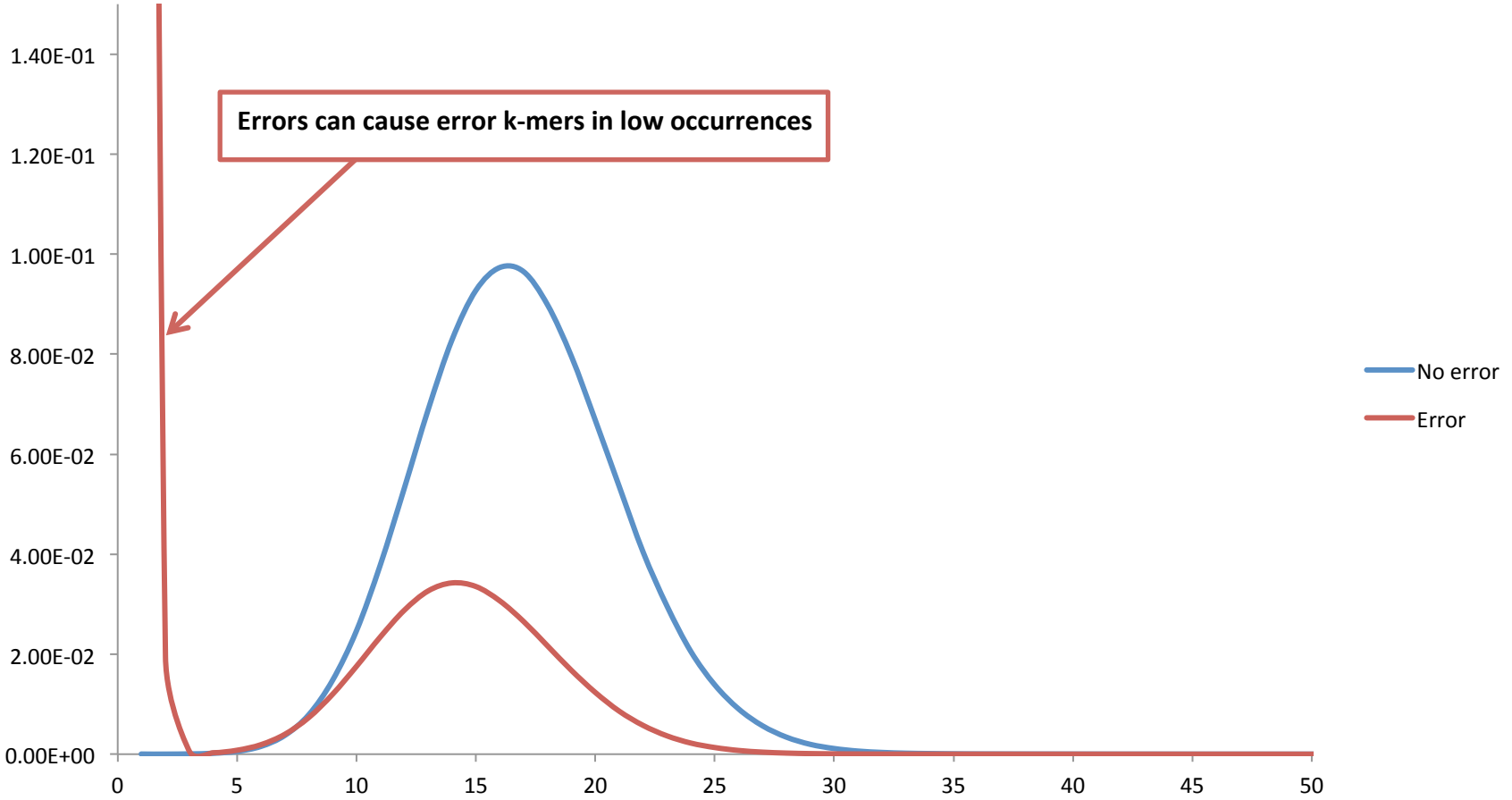
Basic definition

- Repeat is the sequence that appear more than twice in the haploid genome.

- Hetegozygosity is caused by SNPs, insertion and deletions(InDels).

Algorithm in k-mer analysis

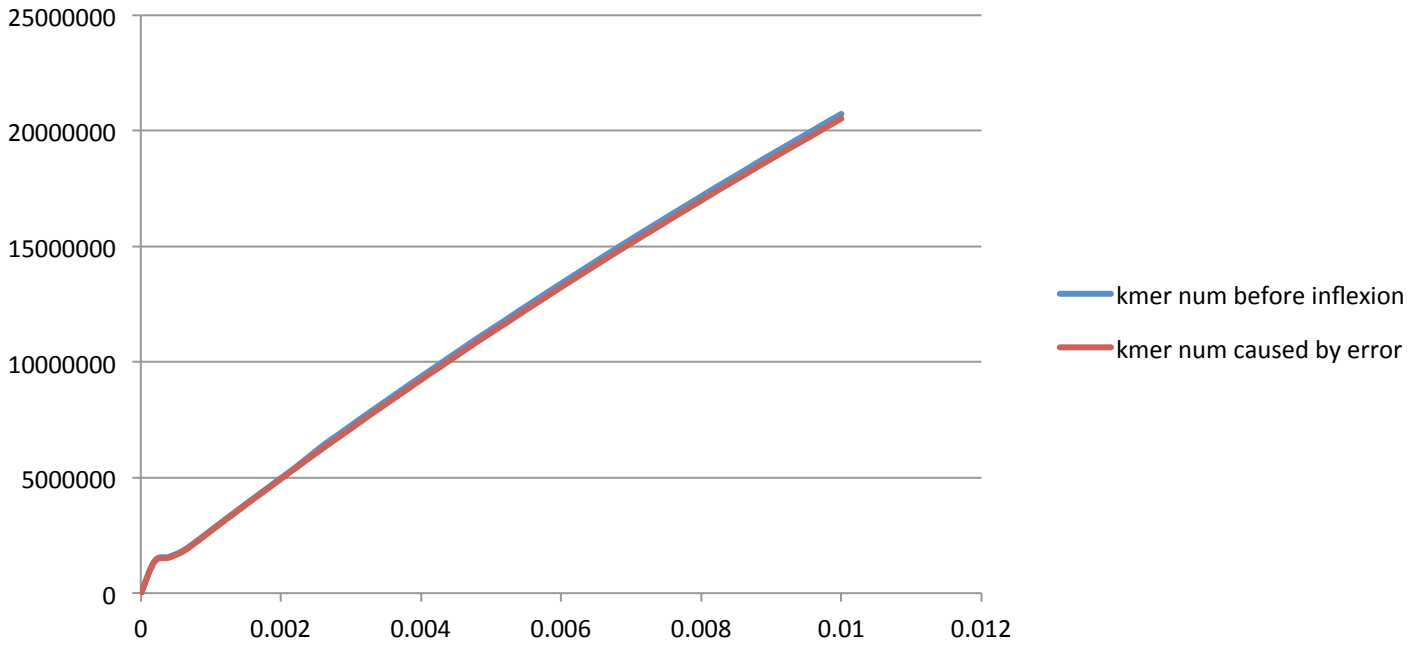
De-noising from error



De-noising from error

- The total k-mers can be divided into two parts: error k-mers and correct k-mers.
- The error k-mer’s occurrences are very small. I guess that almost all error k-mers is before the inflexion.
- To verify the hypothesis, I simulate reads with error rate from 0.001 to 0.01 and observe the number of k-mer before the inflexion and the real error number of k-mer.

De-noising from error



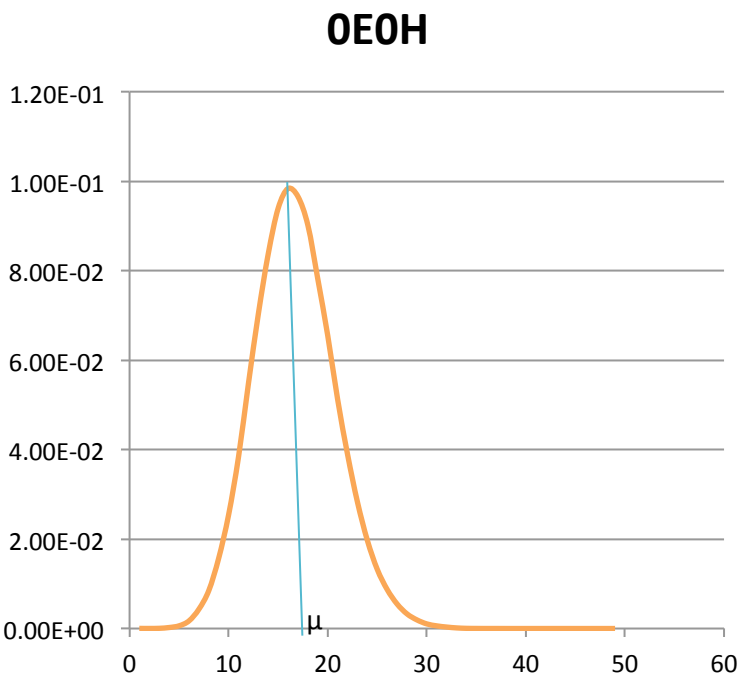
The tendency of k-mer num before inflexion is consensus with the real's error number of k-mer.

- We can just easily remove the k-mers before the inflexion to de-noising from errors.

Estimate the genome size

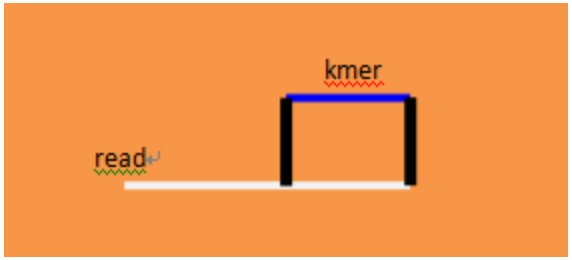
Relationship of average occurrence of k-mer and coverage of base

X	L	K	μ
20	100	17	16
25	100	17	21
40	100	17	33
60	100	17	50
20	40	17	12
25	40	17	15
40	40	17	24
60	40	17	36



$$\mu = \frac{L - k + 1}{L} X$$

Estimate the genome size



- A k-mer can be covered by a read

$$p = \frac{L - k + 1}{G}$$

- Expect of the k-mer is

$$\mu = p \times read\_num = p \times \frac{GX}{L} = \frac{L - k + 1}{L} X$$

Estimate the genome size

- The total num of k-mer T from reads :

$$T = read\_num * (L - k + 1) = \frac{GX}{L}(L - k + 1) = G\mu$$

- So the genome size G is

$$G = \frac{T}{\mu}$$

Estimate the repeat

- If we know the num of repeat k-mers(T_r) and the average k-mer occurrence(μ), we can derive the repeat length(R).

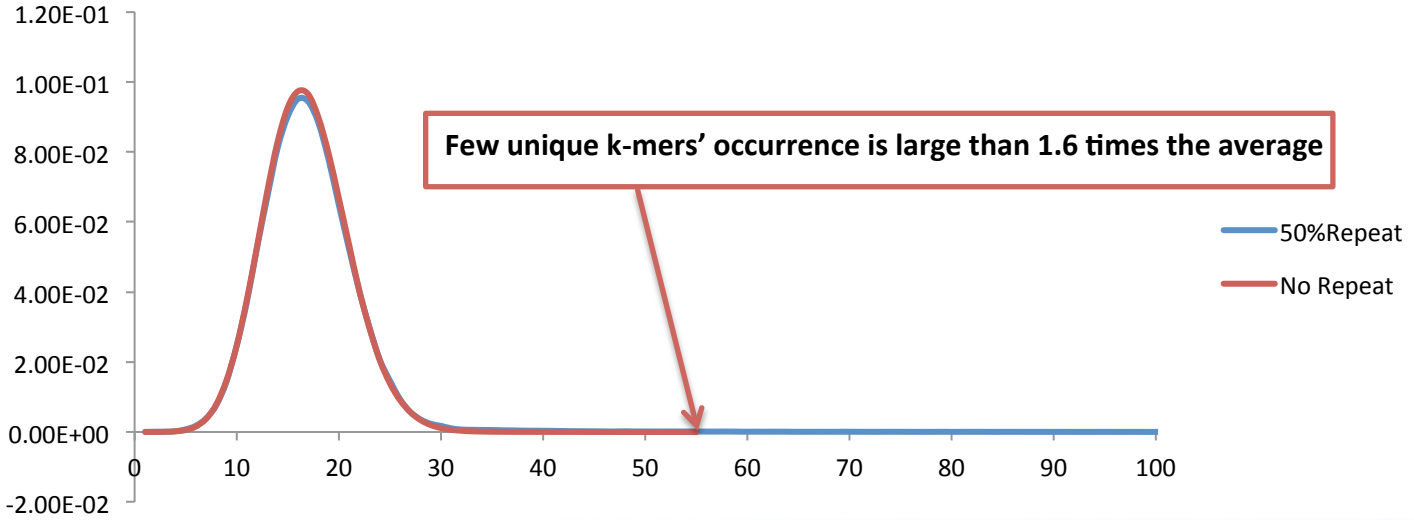
$$R = \frac{T_r}{\mu}$$

- The repeat rate R_p is

$$R_p = \frac{R}{G}$$

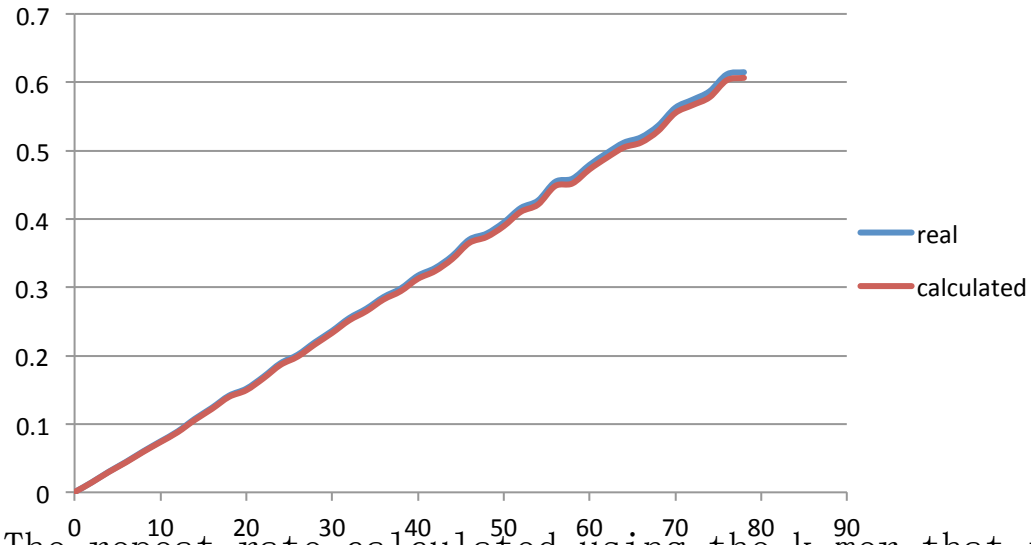
Estimate the repeat

- Since repeat k-mer appear more than twice in the haploid genome, so the occurrence of repeat k-mer is large than twice the unique k-mer’s.
- In the k-mer distribution, few unique k-mer’s occurrence is large than 1.6 times the average.



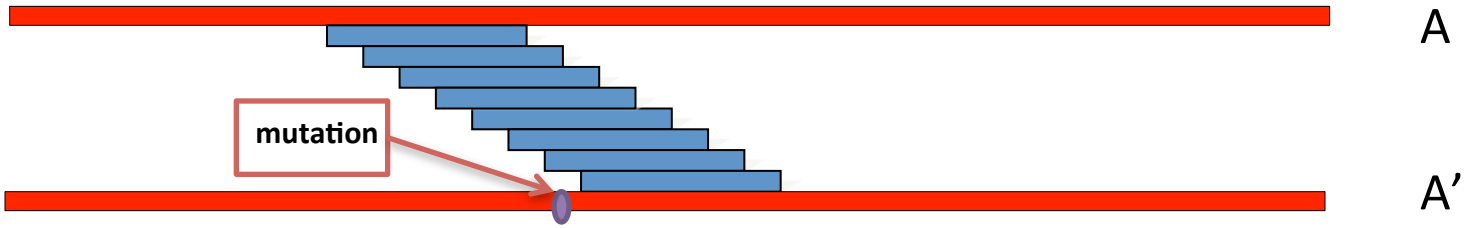
Estimate the repeat

- I simulated DNA sequences with repeat rate from 0 to 0.7 and generate 20X reads for the sequences and count k-mer occurrences.



The repeat rate calculated using the k-mer that the occurrence is >1.6 times consensus with the real's.

Estimate heterozygosity



- The mutation(or heterozygous) rate is h, then a k-mer without mutation is $(1-h)^k$
- So mutation can generate another $1-(1-h)^k$ percent unique k-mers
- The number of unique k-mer of the diploid genome(U) can be counted from k-mer distribution between the inflexion(caused by error) and 1.6 times the average occurrence(repeat k-mer).
- The number of unique k-mer of the haploid genome(U') can be calculated:

  U' =G(1-repeat_rate)

So the unique k-mer caused by mutation is $U-U'=[1-(1-h)^k]U$