Determine the relationship of k-mer size and sequence length in whole genome shotgun Sequencing

Huilin Zhang&Yuanhao Ruan
CSCI5481

Abstract

Sequencing quality is an essential part in biological research and is highly emphasized by scientists. In genome sequencing, k-mer based analysis plays an important role due to its efficiency, accuracy and effectiveness.[1] Selecting a proper k-mer has remained a debatable problem but scientists recognized k-mer choice affect the correctness of sequencing result. Here we show that k size influence the assembly result at each sequence length and we compared correct assembled fractions and running time at each k value. We selected a range of k sizes, broken the sequence and implemented De Brujin graph to reassemble the sequence and compare with the original sequence to find a k size that satisfy complete match for first time. The equation is k=0.017072*sequence length+6.542938. When a 1/100 sequencing error was added, the equation is k=0.013278*sequence length+6.82868 for a 98% match and k size did not oscillated a lot in this situation. Our results demonstrate runtime has no trend and k size is positively correlated with sequence length. However, the result was based on chromosome of length around 1000 and further analysis need to be performed on other lengths of sequences to get a more accurate relationship of k and sequence length and running time.

Summary of previous finding

The purpose of genome sequencing is to determine the nucleotide orders. Genome sequencing is indispensable due to its significance in biological research, medical diagnosis, forensic science and many other fields, while traditional sequencing methods can only be able to determine a relatively short length in one reaction. Therefore, multiple sequencing methods have emerged and allow researchers to explore the full genome. Whole genome sequencing focuses on analyzing the entire genomes of organisms. Whole genome shotgun sequencing is one approach that enables scientists to obtain the information of the whole genome by breaking DNA into small fragments of different lengths called k-mers and then assemble those fragments into intact genomes. Currently, Eulerian algorithm DNA assembly with De Bruijn graph method has made a huge breakthrough in solving problems of assembly of DNA with repetitive regions[2]. However, the choices of k-mer size have a huge influence on the accuracy of the sequencing result. A smaller size of k-mer complicated the graph while a larger size of k-mer may result in disconnected graph[3]. Scientists have made some progress in developing software in finding a proper value of k such as HyDA-Vista[4], IDBA[5],

KmerGenie[6] and so on. Selecting an appropriate size of k-mer remains as an important problem that needs to be further evaluated.

Results

K size	Fraction correctly assembled	Running time(s)
2	0.26	0.002423048
3	0.29	0.002934933
4	0.264	0.002017021
5	0.31	0.002139807
6	0.33	0.003452063
7	0.334	0.002425909
8	0.426	0.002528906
9	0.438	0.003072023
10	0.476	0.003647804
11	0.584	0.003773212
12	0.576	0.003821135
13	0.58	0.002602816
14	0.968	0.005392075
15	0.98	0.004127026
16	0.968	0.002893925
17	1.0	0.003538132

Table 1: At sequence length of 500bases, fraction correctly assembled and running time

Sequence Length	K size
25	4
50	6
100	7
200	10
300	11
400	11
500	17
600	21
700	23
800	23
900	23
1428	25

Table 2: K size for each sequence length when the assembled sequence completely match with original sequence

Sequence length	K size
100	7
200	10
300	9
400	11
500	16
600	17
700	17
800	17
900	18
1428	25

Table 3: K size for sequence length>100 bases for first 98% match considering sequence error (1 random error per 100 bases)

In order to determine the k value for first complete match, several k sizes were chosen by running the De Brujin algorithm Python program for sequence chrUn_KI270338v1.fa. The k size ranged from 2 to twenty percent of the selected length. In Figure 1, the assembled result was compared with the original sequence at each position when k equal to 2, 3, 4 and 5 at sequence length of 25 bases. When k were set to 2 and 3, we could observe that there were several mismatches at certain positions. When k reached to 4, all the positions were matched. When repeated this process with other sequence lengths, we noticed that when k size was small, mismatch was more likely to occur and complete match appeared when k was at a moderate size.As shown in Table 1, when sequence length was set to 500, k size was 17 to reach complete match. In Figure 2, fraction correctly assembled had a growth trend though there were some fluctuations when k size increases. While the running time varied a lot when k size grew as indicated in Figure 3.

As shown in Table 2, k size of lengths of 25 bases, 50 bases, 100 bases, 200 bases, 300 bases, 400 bases, 500 bases, 600 bases, 700 bases, 800 bases, 900 bases and full length of 1428 bases were compared. The k size for first complete match increased with the increasing sequence length. A linear regression model was depicted in Figure 4. The equation was y=0.017072x+6.542938, where x represented sequence length and y was the k size corresponding to the sequence length when first complete match happened. R^2 was 0.8249. Considering simulation error (1 random error per 100 bases), k sizes were redetermined by implementing random error method for sequence length larger than 100 bases as demonstrated in Table 4 and we used 98% percent as an optimal match. K sizes were more varied when applying sequencing error in the assembly process. The linear regression model in Figure 5 predicted the relationship between sequence length and k was y=0.013278x+6.82868, giving x for sequence length and y for k size.R^2 was 0.9156.

Conclusions

The whole genome shotgun sequencing simulation successfully proved that the choice of k size largely affect the assembly quality. When k was too small, the resulting sequence was more likely to have errors, while k was too large, it may be memory inefficient. Fractions correctly assembled have positive relationship with k at each sequence length. The running time did not show any trend. A equation y=0.017072x+6.542938 could be used to determine the k size for complete match. With regard to sequencing error, y=0.013278x+6.82868 could be applied to find a 98%

match. Because the sequence used was 1428 bases and in order to find a more comprehensive equation to find an appropriate k size, we need to examine longer, for example, sequences that have millions bases.

Methods

Simulation

A chromosome sequence was obtained and was broken up into k-mer fragments with a selection of k. A De Brujin graph was built using the k-mer list as node with a Hamiltonian circuit (Figure 6). Then an Eulerian circuit (Figure 7) was implemented to traverse the k-mers and a valid path was found to be the result of assembly. Several short sections with different length were selected to perform the analysis with different k values. The runtime and fraction correctly assembled were compared for each k value. For each sequence length, a k value was determined when the original sequence and assembled sequence was totally matched.

In order to simulate sequencing error for 1 random error per 100 bases, 15 bases at random position of original sequence were randomly chosen to be replaced by a different base. For each sequence length, a k value was determined when the original sequence and assembled sequence was 98% matched

Statistical analysis

Linear regression model was used to depict the relationship between sequence length and k size.

Figures

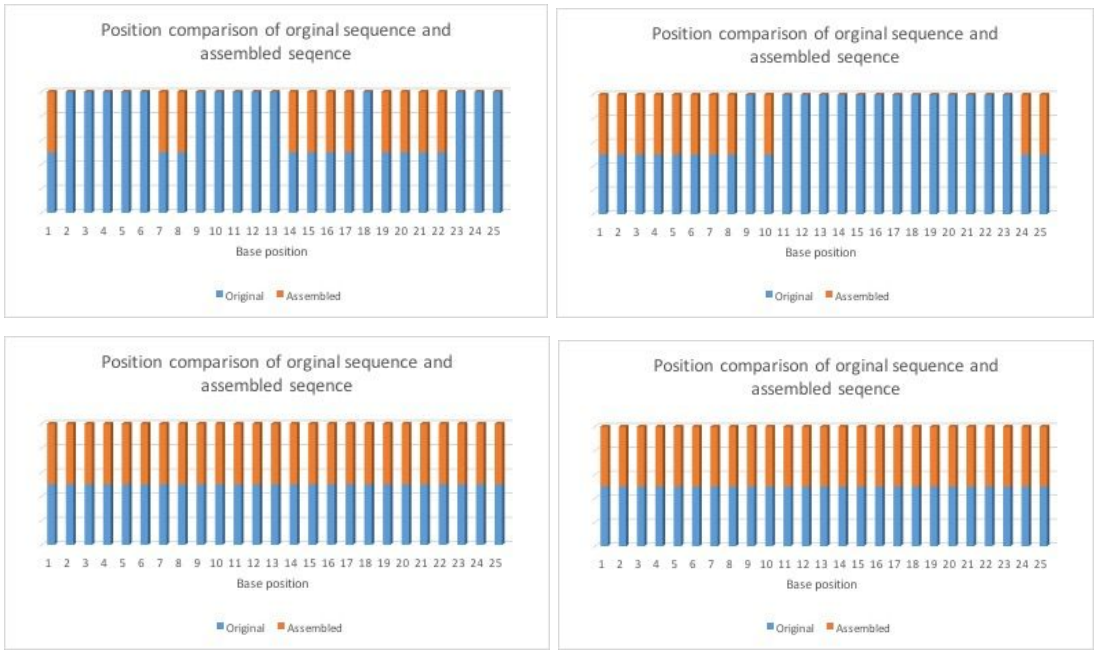

Figure 1: Compare bases for original sequence and assembled sequence at each position(sequence length=25, k=2,3,4,5)

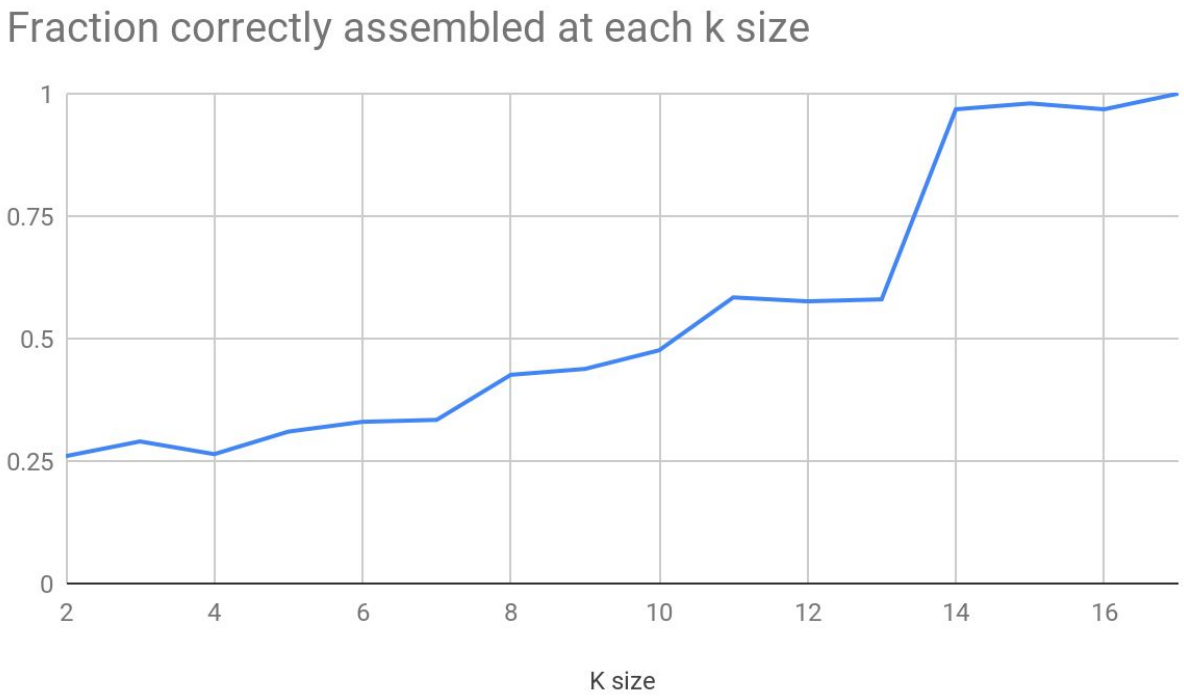

Figure 2: When sequence length=500bases, fraction correctly assembled at each k size

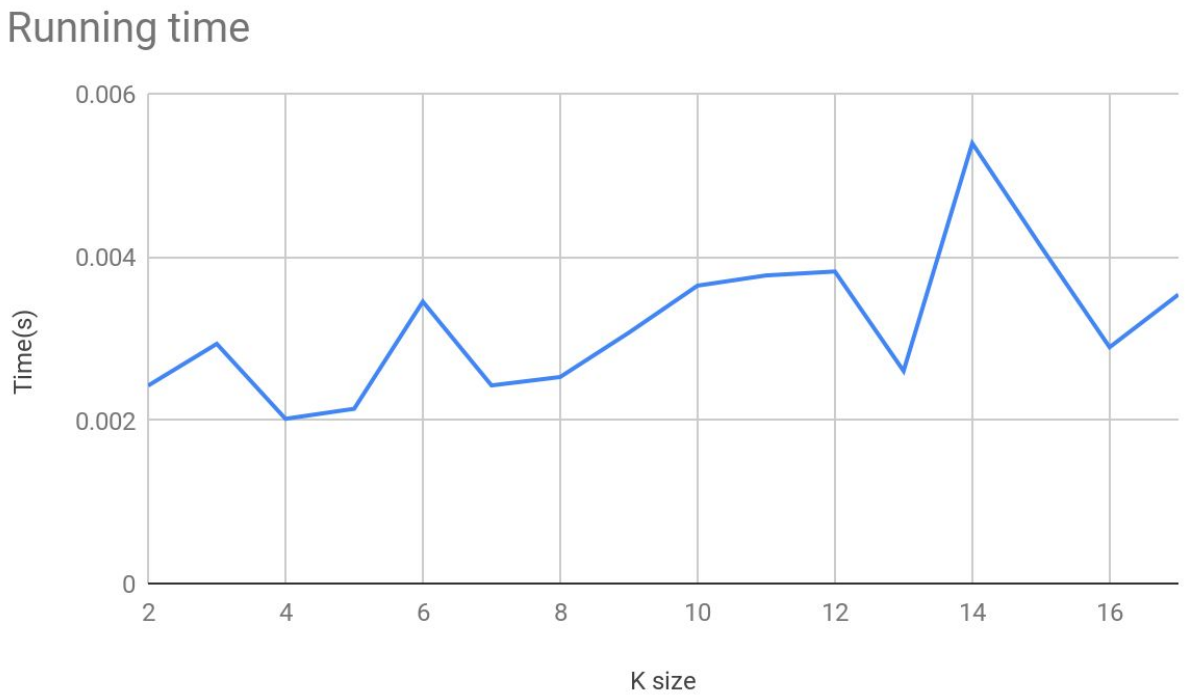

Figure 3: When sequence length=500bases, fraction correctly assembled at each k size

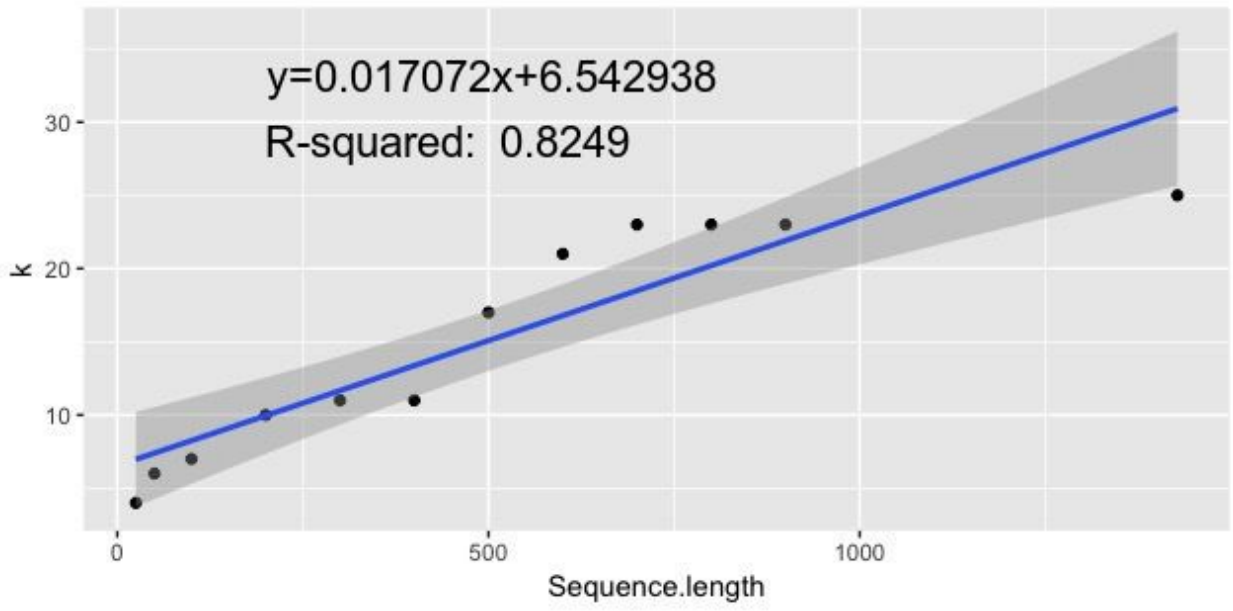

Figure 4: linear regression relationship of sequence length with k size (complete match)

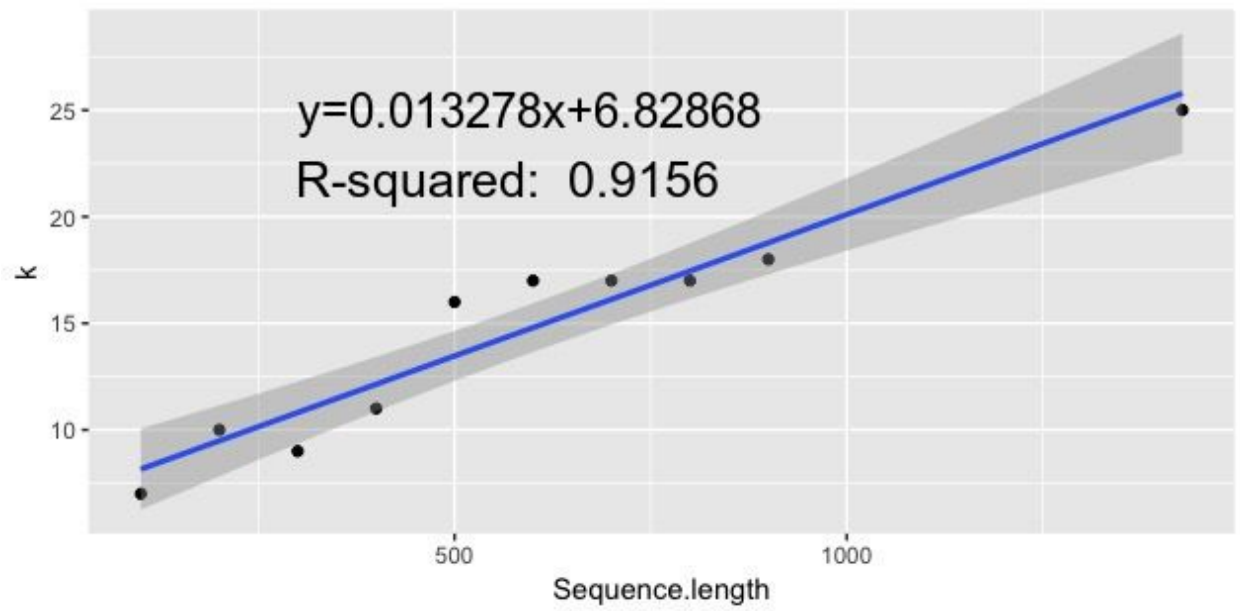

Figure 5: linear regression relationship of sequence length with k size applying sequence error (98% match)

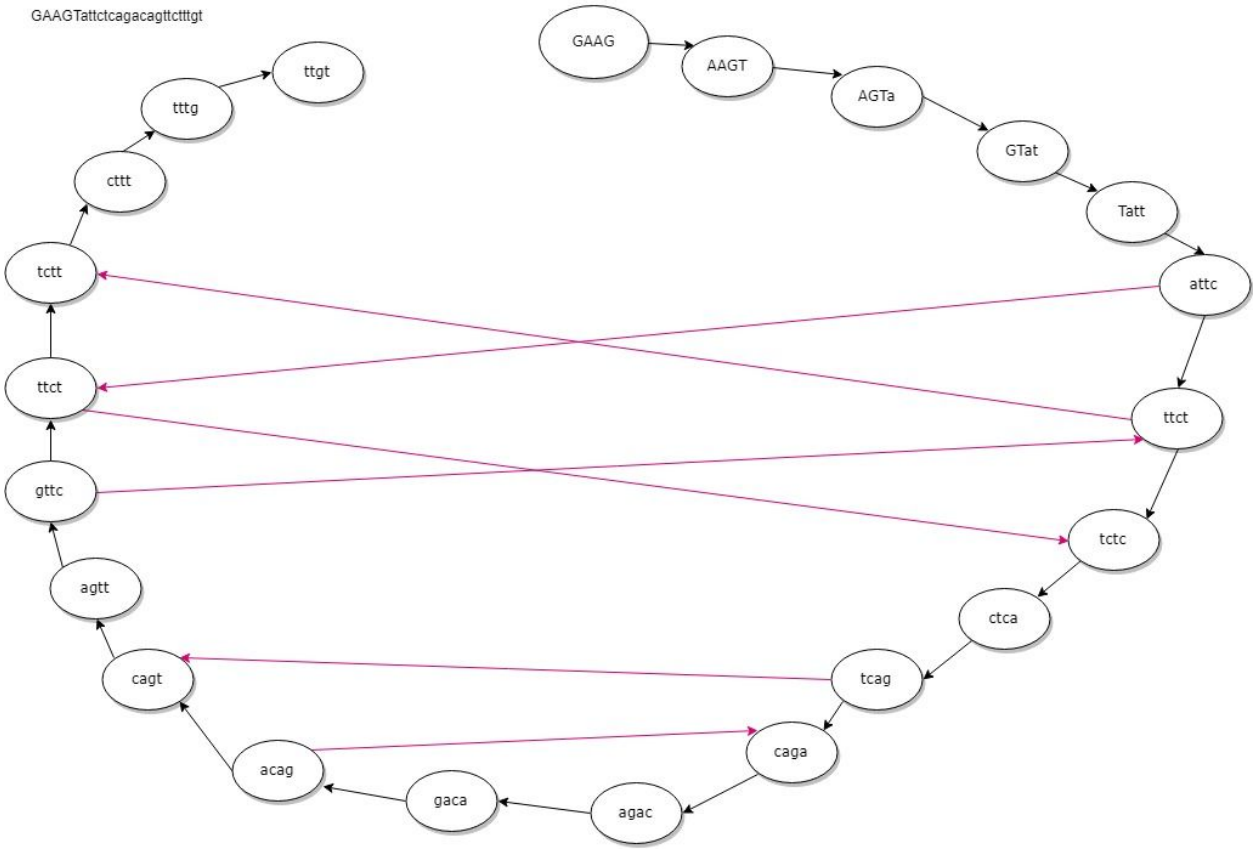

Figure 6: De Brujin graph with Hamiltonian circuit (k dimensional for first 25 bases; k=4)

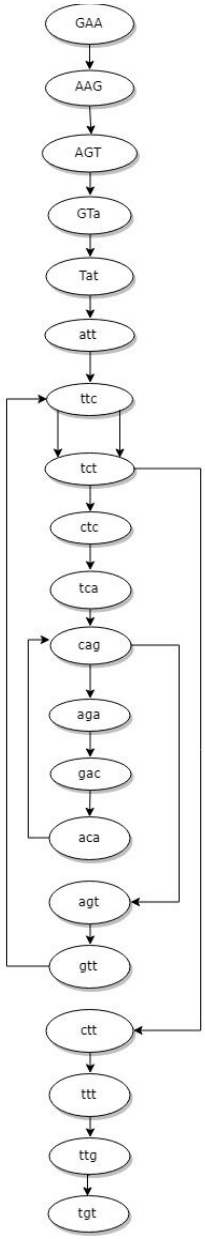

Figure 7: De Brujin graph with Eulerian circuit (k-1 dimensional for first 25 bases; k-1=3)

Acknowledgements

Huilin Zhang: Write half of the code, generated half graph, draw the graph,designed the poster, write paper.

Yuanhao Ruan:Write half of the code, run the simulation,generated half graph,designed the poster, write paper.

References

1.Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology direct*, *8*, 3. doi:10.1186/1745-6150-8-3

2.Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(17), 9748-53.

3.Shariat, B., Movahedi, N. S., Chitsaz, H., & Boucher, C. (2014). HyDA-Vista: towards optimal guided selection of k-mer size for sequence assembly. *BMC genomics*, *15 Suppl 10*(Suppl 10), S9.

4.Shariat, B., Movahedi, N. S., Chitsaz, H., & Boucher, C. (2014). HyDA-Vista: towards optimal guided selection of k-mer size for sequence assembly. *BMC genomics*, *15 Suppl 10*(Suppl 10), S9.

5.Peng, Y., Leung, H. C., Yiu, S. M., & Chin, F. Y. (2010). IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. *Lecture Notes in Computer Science Research in Computational Molecular Biology,* 426-440. doi:10.1007/978-3-642-12683-3_28.

6.Chikhi, R., & Medvedev, P. (2013). Informed and automated k-mer size selection for genome assembly. *Bioinformatics, 30*(1), 31-37. doi:10.1093/bioinformatics/btt310