A new method of evaluating genome assemblies based on kmers frequencies

Kirill V. Romanenkov
Samsung R&D Institute, Dvintsev 12, Moscow, Russia,
kromanenkov2@yandex.ru

Andrey V. Alexeevski
Belozersky Institute, Lomonosov Moscow State University, Leninskie Gori 1,
Moscow, Russia, *aba@belozersky.msu.ru*

Modern automatic sequencing methods allow to produce hundreds of billions short sequences obtained by reading the input fragments DNA samples of one or more organisms at a rather short time (several days) and low rate (several thousands dollars). However there are serious technological, algorithmic and computational problems on the way from the tens of billions of reads with length ranging from 100 to 250 bp to several sequences representing the complete genome.

Numerous approaches for *de novo* assembly have been proposed. Most of them are based on the de Bruijn graph conception which briefly consists in the following. All kinds of substrings length k (called kmers) are formed from the reads obtained after the sequencing step. Therefore a single string length L forms $L-k+1$ kmers. Each kmer forms a node in the de Bruijn graph and two nodes are connected by a directed edge if $k-1$ length suffix of the first node is the $k-1$ length prefix of the second. Consequently, k value is one of the most important parameters to affect the assembly result.

Running different genome assemblers or one genome assembler with different k values on the same input data commonly leads to a great variety of results. However, there is no generally recognized method for choosing the best assembly. Moreover, standard metrics do not take into account correspondence between resulted assembly and a set of reads and in some cases could not provide a full picture of the assembly quality.

This article introduces a new reference-free method for evaluating genome assembly by kmers frequencies analysis. The proposed method sets up a correspondence between short reads obtained from the sequencer and assembled genome, which allows a more accurate genome assembly assessing. It is based on the assumption that most of the kmers for a reasonable large k are found in small and medium-size genomes in a single copy.

First of all frequency histogram of kmers occurrence in reads is constructed as well as the set of all unique kmers of assembled genome. This histogram has two peaks in general: first corresponds to the sequencing errors and second corresponds to the read coverage. Then a neighborhood of the histogram's second peak is chosen in such a way that eliminates error kmers

and kmers came from repeat regions. After that the proportion of kmers from selected neighborhood among the unique kmers of genome assembly is calculated as follows:

$$Q = \frac{\sum_{i=1}^{|K_{uniq\_reads}|} \left[k^r(i) \in K_1^g\right]}{\left|K_{uniq\_reads}\right|},$$

where $K_1^g = \left\{k^g : k^g \in K^g, abundance\left(k^g\right) = 1\right\}$, $K_i^r = \left\{k^r : k^r \in K^r, abundance\left(k^r\right) = i\right\}$

$K_{uniq\_reads} = \bigcup_{i=a_{p\_l}}^{a_{p\_r}} K_i^r$, $\left[a_{p\_l}; a_{p\_r}\right]$ - selected neighborhood of the second peak on frequency histogram of kmers occurrence in reads, $abundance(k)$ - occurrence of kmer k, K^r - set of all kmers in sequencing reads, K^g - set of all kmers of assembled genome. The higher the Q value, the better the resulting assembly approximates sequencing reads.

The proposed method was validated on different assemblies of Encephalitozoon cuniculi fungus organism [1] performed by four assemblers: Abyss, Ray, SOAPdenovo and Velvet. Assemblers were run with five different k values: 21, 33, 43, 51, 55. Existence of the reference genome for this organism [2] allows validating of the proposed method by comparing it with the reference-dependent metrics calculated by the Quast [3] software. The following tables contains calculated Q values for the different assemblies of the Encephalitozoon cuniculi fungus genome as well as standard metrics. All of them were evaluated using contigs ≥500 bp.

Table 1

Comparison of assemblies of a Encephalitozoon cuniculi fungus by Velvet

k	N50	Number of contigs	NGA50	Total length(Mb)	Q
25	1175	15927	11316	17.7	0.934086
33	928	14977	21452	14.5	0.938702
43	937	8093	18836	8.3	0.937946
51	4091	2610	9237	3.7	0.92775
59	3339	1124	3189	2.4	0.905697

Table 2

Comparison of assemblies of a Encephalitozoon cuniculi fungus by SOAPdenovo

k	N50	Number of contigs	NGA50	Total length(Mb)	Q
25	1101	16920	644	17.4	0.639442
33	932	12249	2578	11.5	0.898704
43	932	3816	14839	4.7	0.936651
51	7480	1245	9248	2.8	0.92729
59	3412	1086	3183	2.3	0.901179

Table 3

Comparison of assemblies of a Encephalitozoon cuniculi fungus by Abyss

k	N50	Number of contigs	NGA50	Total length(Mb)	Q
25	1495	14981	28200	19.7	0.932986
33	1110	12913	55107	14.3	0.940028
43	1082	4939	63324	5.8	0.940482
51	41991	1241	55270	3.1	0.934639
59	20966	488	20966	2.5	0.926928

Table 4

Comparison of assemblies of a Encephalitozoon cuniculi fungus by Ray

k	N50	Number of contigs	NGA50	Total length(Mb)	Q
25	950	8820	78910	9.3	0.895447
33	5780	3274	42981	4.6	0.927833
43	14182	758	15087	2.7	0.92385
51	6909	339	4018	1.6	0.603678
59	2827	419	-	0.9	0.364621

It was found that in most cases *Q* value correlates with the reference-dependent NGAx metrics and could correctly identify the best assembly. Furthermore, an interconnection between assembly quality and standard reference-free metrics was not observed.

1. European Nucleotide Archive
   URL: http://www.ebi.ac.uk/ena/data/view/SRR122309 (accessed 1.02.2017).
2. Encephalitozoon cuniculi GB-M1
   URL: http://www.ncbi.nlm.nih.gov/genome/39?genome_assembly_id=22671
   (accessed: 1.02.2017).
3. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. // Bioinformatics. — 2013. — V. 29, No. 8. — P. 1072–1075.
   URL: http://dx.doi.org/10.1093/bioinformatics/btt086