

METHODOLOGY ARTICLE

Open Access

# GMASS: a novel measure for genome assembly structural similarity



Daehong Kwon<sup>†</sup>, Jongin Lee<sup>†</sup> and Jaebum Kim<sup>\*†</sup>

## Abstract

**Background:** Thanks to the recent advancements in next-generation sequencing (NGS) technologies, large amount of genomic data, which are short DNA sequences known as reads, has been accumulating. Diverse assemblers have been developed to generate high quality de novo assemblies using the NGS reads, but their output is very different because of algorithmic differences. However, there are not properly structured measures to show the similarity or difference in assemblies.

**Results:** We developed a new measure, called the GMASS score, for comparing two genome assemblies in terms of their structure. The GMASS score was developed based on the distribution pattern of the number and coverage of similar regions between a pair of assemblies. The new measure was able to show structural similarity between assemblies when evaluated by simulated assembly datasets. The application of the GMASS score to compare assemblies in recently published benchmark datasets showed the divergent performance of current assemblers as well as its ability to compare assemblies.

**Conclusion:** The GMASS score is a novel measure for representing structural similarity between two assemblies. It will contribute to the understanding of assembly output and developing de novo assemblers.

**Keywords:** Measure, Genome, Assembly, Structural similarity

## Background

Recent advances in next-generation sequencing (NGS) technologies have facilitated a rapid production of large amounts of short sequencing data from genomes, known as reads, at a low cost [1–3]. The length of reads is up to several hundred base pairs in the case of short read sequencing technologies such as Illumina, and up to several hundred kilo base pairs in the case of long read sequencing technologies such as Oxford Nanopore. Since the NGS reads are still much shorter than chromosomes lengthwise, a de novo assembly process is necessary to reconstruct the whole chromosome sequences [4]. The de novo assembly begins with finding and connecting overlapping NGS reads, which leads to the generation of longer sequences, known as contigs. Contigs are then ordered and oriented to construct much longer sequences, known as scaffolds, using paired-end or mate-pair reads generated from long DNA fragments.

Several algorithms have been developed to generate high-quality de novo assemblies. They are mainly classified into three categories: the greedy graph based algorithm, the Overlap-Layout-Consensus based algorithm, and the *de Bruijn* graph based algorithm [4]. The greedy graph based algorithm, such as SSAKE [5], VCAKE [6] and SHARCGS [7], incrementally connects a read to another read or contig with the highest overlap score until there is no more overlapping. The overlap scores are calculated using the number of matching bases in the overlap. Both the Overlap-Layout-Consensus based algorithm and *de Bruijn* graph-based algorithm rely on a graph structure constructed from the NGS reads. They first construct an overlap graph using overlap information obtained from the NGS reads, and then the de novo assembly is constructed by the traversing the graph. However, the Overlap-Layout-Consensus based algorithm, such as CABOG [8], Newbler [9] and Celera assemblers [10], constructs an overlap graph using the direct overlap among the NGS reads, whereas the *de Bruijn* graph based algorithm, such as ABySS [11], SOAPdenovo [12], ALLPATHS-LG [13] and Velvet [14],

\* Correspondence: [jbkim@konkuk.ac.kr](mailto:jbkim@konkuk.ac.kr)

<sup>†</sup>Daehong Kwon and Jongin Lee contributed equally to this work. Department of Biomedical Science and Engineering, Konkuk University, Seoul 05029, South Korea



is based on the overlap of all possible subsequences of length  $k$ , known as  $k$ -mer, extracted from the NGS reads.

The performance of assemblers is quite different due to algorithmic differences, which results in difficulty in choosing the most appropriate assembler for different NGS data [15–17]. Many benchmarking studies, such as Genome Assembly Gold-standard Evaluations (GAGE) [18], the GAGE for bacterial organisms (GAGE-B) [19], and Assemblathon 1 and 2 [20, 21], have been performed using synthetic and real NGS read datasets with attempts to overcome this problem. In these studies, N50 is the most widely used measure to assess the level of assembly contiguity. N50 is a length-weighted median of ordered scaffolds based on their lengths. Here, the weight of a scaffold is defined as the length of the scaffold divided by the total length of an assembly. NG50 was introduced in the Assemblathon project to consider the genome size. NG50 is similar to N50 except for the fact that in NG50, the weight of a scaffold is not calculated using the length of an assembly but the known or estimated length of a genome. Also, for more comprehensive evaluation of assemblies, Feature-Response Curve was introduced and compared with the standard metrics, such as N50, coverage and contig sizes [22]. The accuracy of an assembly was measured by determining genomic dissimilarity by comparing it to the true assembly and by checking the presence of core eukaryotic genes [23] or single-copy orthologs [24] in the assembly. The mapping pattern of reads was also used for detecting mis-assemblies and measuring accuracy of assembly [25]. By comparing assemblies with these measures, several benchmarking studies have identified that the performance of assemblers varies considerably [18–22].

Even though these measures are good for evaluating the quality of a single assembly, they do not directly compare two different assemblies. For example, when two assemblies are compared using N50, researchers can identify how dissimilar they are in terms of the level of assembly fragmentation but cannot measure the fraction of highly similar assembly regions between them. This kind of comparison is especially useful when a de novo assembly of a certain species needs to be compared to a reference species, or when multiple assemblies of the same species generated by different assembly programs need to be compared to check how similar or different they are. Comparison of assemblies typically begins with their alignment. If there are large-scale assembly rearrangements, a large number of alignment fragments are normally obtained. Using alignment results, several measures including the number of alignment fragments, and alignment coverage have been separately used to measure assembly similarity. In general, the more similar two assemblies are, it is expected that the smaller number of alignment fragments are constructed between them and

alignment coverage of assemblies are higher. However, the number of alignment fragments can also be reduced by generating low coverage alignments, and the alignment coverage can also be increased by generating large number of small alignment fragments. Therefore, a single quantitative measure combining those two is necessary. In addition, the alignment fragments from assemblies are generated based on a resolution parameter, which determines the minimum length of the alignment, and the choice of the resolution parameter is arbitrary.

In this perspective, a new measure, called GMASS, was developed for comparing the structures of two genome assemblies. The GMASS score was defined based on the distribution pattern of the number and coverage of similar regions between a pair of assemblies. Specifically, the alignments of two assemblies are first generated, and then the number of alignment fragments and their coverage are obtained. This process is repeated using a certain range of resolutions, and those two values from different resolutions are combined to generate a single quantitative score. The GMASS score was related to the structural similarity between assemblies when it was evaluated using simulated assembly datasets. The GMASS score was also applied to compare pairs of assemblies obtained from previous benchmarking studies [18, 20], and the results showed that the current assemblers generate very different assemblies even though the same sequencing read datasets are used.

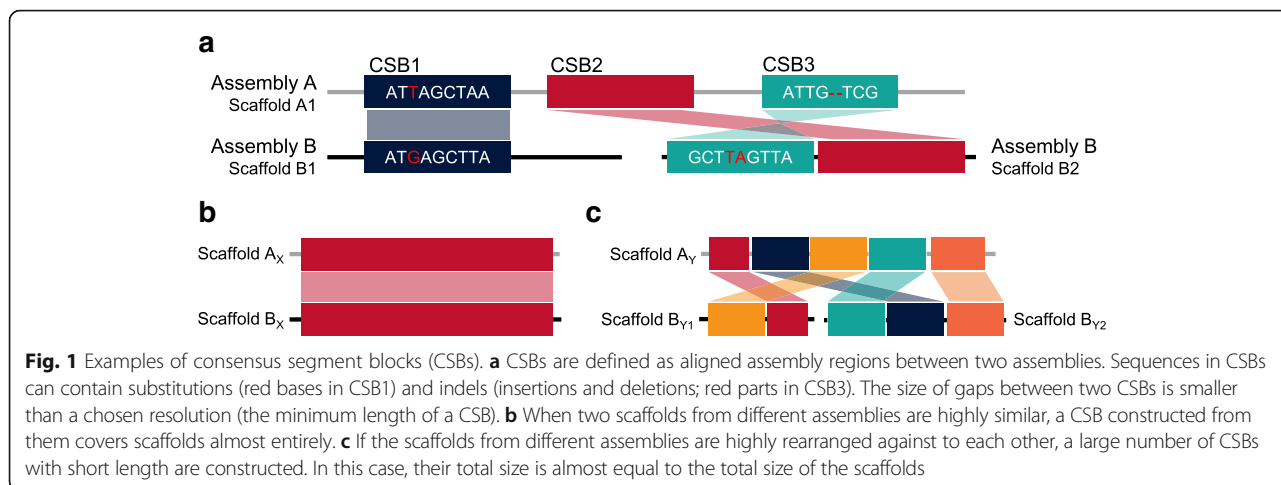
## Methods

### The GMASS score

The GMASS score represents the structural similarity of a pair of genome assemblies based on the length and number of similar genomic regions defined as consensus segment blocks (CSBs) in the assemblies. The CSBs are aligned assembly regions between two assemblies which can contain small number of substitutions and small gaps shorter than a chosen resolution (Fig. 1a). The way of constructing CSBs is described in the following subsection. The GMASS score is defined as Eq. 1 based on the assumption that the number and length of CSBs between two assemblies is equivalent to the number and length of sequence fragments (scaffolds or chromosomes) of the two assemblies as the two assemblies become more similar.

$$GMASS = \frac{\sum_{i=r}^R S_i}{n} = \frac{\sum_{i=r}^R \{L_i(A_1, A_2) \times C_i(A_1, A_2)\}}{n} \quad (1)$$

The GMASS score is the average of structural similarity scores ( $S_i$ ) between two assemblies  $A_1$  and  $A_2$  across a certain range of CSB resolutions. In Eq. 1,  $S_i$  denotes the structural similarity score at the CSB resolution  $i$  belonging to a set of  $n$  different resolutions with a minimum value  $r$  and a maximum value  $R$ .  $S_i$  is calculated by



the product of two terms related to the total length of the CSBs ( $L_i(A_1, A_2)$ ) and the number of CSBs ( $C_i(A_1, A_2)$ ).

$L_i(A_1, A_2)$  is the fraction of the total length of the CSBs against the total length of assemblies used to construct the CSBs with the CSB resolution  $i$  defined as Eq. 2. In Eq. 2,  $l(A_x)$  is the total length of sequence fragments larger than resolution  $i$  in the assembly  $A_x$ , and  $l(CSB_{S_x})$  is the total length of blocks in the CSBs belonging to  $A_x$  defined as  $CSB_{S_x}$ .

$$L_i(A_1, A_2) = \frac{l(CSB_{S_1}) + l(CSB_{S_2})}{l(A_1) + l(A_2)} \tag{2}$$

For example, suppose CSBs are constructed for the two scaffold assemblies  $A_1$  and  $A_2$  using 10 Kbp resolution. If total 2 Mbp and 2.5 Mbp scaffold sequences of  $A_1$  and  $A_2$  respectively are used in this construction, and total 1.8 Mbp and 2.2 Mbp scaffold sequences of  $A_1$  and  $A_2$  respectively are actually included in the CSBs, then  $l(A_1) = 2$  Mbp,  $l(A_2) = 2.5$  Mbp,  $l(CSB_{S_1}) = 1.8$  Mbp, and  $l(CSB_{S_2}) = 2.2$  Mbp. Therefore,  $L_{10Kbp}(A_1, A_2) = (1.8 + 2.2)/(2 + 2.5) = 0.89$ .

$C_i(A_1, A_2)$  denotes the similarity between two assemblies in terms of the total number of CSBs constructed with the CSB resolution  $i$  defined as Eq. 3.

$$C_i(A_1, A_2) = \begin{cases} \frac{c(CSB_{S_1}) + c(CSB_{S_2})}{c(A_1) + c(A_2)}, & \text{if } c(A_1) + c(A_2) \geq c(CSB_{S_1}) + c(CSB_{S_2}) \\ 1 - \frac{c(CSB_{S_1}) + c(CSB_{S_2}) - c(A_1) - c(A_2)}{\max(A_1) + \max(A_2) - c(A_1) - c(A_2)}, & \text{otherwise} \end{cases} \tag{3}$$

Here  $c(X)$  is the number of sequence fragments larger than given resolution  $i$  (if  $X$  is an assembly) or the number of blocks constructed using the CSB resolution  $i$  (if  $X$  are CSBs). We note that if  $X$  are CSBs, then  $c(CSB_{S_1})$

and  $c(CSB_{S_2})$  are the same because of the definition of a CSB.  $\max(X)$  is the maximum number of CSBs that can be constructed from an assembly  $X$  using the CSB resolution  $i$ . If the sum of the number of CSBs from two assemblies is less than or equal to the sum of sequence fragments larger than resolution  $i$  in the two assemblies,  $C_i(A_1, A_2)$  is defined as the fraction obtained from those two numbers (the upper part of Eq. 3). Otherwise, the positive difference between those two numbers is calculated, and its fraction against a similar difference using the sum of the maximum number of CSBs that can be constructed using the two assemblies is obtained. The result is subtracted from 1 to make a positive correlation between  $C_i(A_1, A_2)$  and assembly similarity (the lower part of Eq. 3). The maximum number of CSBs  $\max(A_x)$  of an assembly  $A_x$  is calculated by Eq. 4.

$$\max(A_x) = \frac{l(A_x)}{i} \tag{4}$$

Here  $i$  is the resolution used to construct CSBs. Eq. 4 implies that the minimum length of a CSB is  $i$ , and therefore at most  $l(A_x)/i$  CSBs can be constructed. This was empirically validated using the GAGE and Assemblathon 1 dataset.

For example, suppose again that CSBs are constructed for the two scaffold assemblies  $A_1$  and  $A_2$  with  $l(A_1) = 2$  Mbp and  $l(A_2) = 2.5$  Mbp using 10 Kbp resolution. If total 20 and 24 scaffolds of  $A_1$  and  $A_2$  respectively are used in this construction, and total 20 CSBs are constructed, then  $c(A_1) = 20$ ,  $c(A_2) = 24$ , and  $c(CSB_{S_1}) = c(CSB_{S_2}) = 20$ . In this case,  $c(A_1) + c(A_2) \geq c(CSB_{S_1}) + c(CSB_{S_2})$  holds, and therefore,  $C_{10Kbp}(A_1, A_2) = (20 + 20)/(20 + 24) = 0.91$  by the upper part of Eq. 3. On the other hand, if  $c(CSB_{S_1}) = c(CSB_{S_2}) = 25$ , the upper part of Eq. 3 cannot be used because  $C_{10Kbp}(A_1, A_2) = (25 + 25)/(20 + 24) > 1$  based on the

upper part of Eq. 3 and the  $C_i(A_1, A_2)$  score between 0 and 1 is needed. In this case,  $C_i(A_1, A_2)$  is calculated by the lower part of Eq. 3 using  $c(A_1) = 20$ ,  $c(A_2) = 24$ ,  $c(CSB_{S_1}) = c(CSB_{S_2}) = 25$ ,  $maxc(A_1) = 2000/10 = 200$ , and  $maxc(A_2) = 2500/10 = 250$ . Therefore,  $C_{10Kbp}(A_1, A_2) = 1 - (25 + 25 - 20 - 24)/(200 + 250 - 20 - 24) = 0.99$ .

#### Calculation of the GMASS score for simulated and benchmark datasets

In this study, CSBs were constructed from three datasets: simulated genome assemblies, GAGE [18], and Assemblathon 1 [20]. In the case of the simulated genome assemblies, the pairwise alignments of simulated genome assemblies (reference against 11 other assemblies with different divergence to the reference) were obtained from previous study [26]. This dataset simulated the evolution of two human chromosomes 21 and 22 from a common ancestor of 12 species with varying divergences. The synteny block generation program in inferCars [27] was then used to construct CSBs between each pair of assemblies with a given resolution.

In the case of the other two datasets, assembly sequences as a form of scaffolds were downloaded from a website with their datasets. Then repeats in the assemblies were first masked using RepeatMasker (version 4.0.5) [28], and pairwise alignments between two different assemblies were generated using LASTZ (version 1.02) [29] with parameters “-E=150 -H=2000 -K=4300 -L=2200 -M=254 -O=600 -T=2 -Y=15000”. In evaluation of the GMASS score, self-alignments of each assembly were also constructed using LASTZ parameters “-O=400 -E=30 -X=910 -Y=9400 -K=3000 -L=3000 -H=2000 -T=1 -Z=1”. These parameter values were based on the recommendation found in the UCSC genome browser [30]. Once pairwise alignments of genome assemblies were prepared, the synteny block generation program in inferCars [27] was used to construct CSBs between a pair of assemblies with a given resolution.

In the process of the CSB generation between two assemblies, one of the assemblies needs to be a reference. In the case of the simulated assembly dataset, a reference assembly specified in that study was also used as the reference in this study. Therefore, the CSBs of total 11 pairs (reference against 11 other assemblies) were constructed. In the GAGE and Assemblathon 1 dataset, a total of eight assemblies by eight assembly programs (ABYSS2 [31], ALLPATHS-LG [13], Bambus2 [32], CABOG [8], SGA [33], SOAPdenovo [12], MSR-CA [34] and Velvet [14]) and five assemblies by five assembly programs (ABYSS [11], ALLPATHS-LG [13], Celera [10], Meraculous [35] and SOAPdenovo [12]) exist respectively, and CSBs of all possible assembly pairs were constructed by using one assembly with a higher N50 as the reference. In

terms of the resolution, five different resolutions (100, 200, 300, 400, and 500 Kbp) were used for the simulated assembly dataset. In the case of the GAGE and Assemblathon 1 dataset, N50, N60, N70, N80, and N90 values of a non-reference assembly were used as resolutions.

To assess the relationship between the GMASS score and structural similarity of assembly pairs, the Pearson's correlation coefficient test was conducted between structural similarities and the GMASS scores calculated using CSBs between simulation datasets. The correlation analyses were also conducted for  $L_i(A_1, A_2)$ ,  $C_i(A_1, A_2)$  and  $S_i$  scores in a given resolution  $i$ . The evolutionary distances which were used for constructing simulated assemblies were treated as indicators of sequence similarity between two assemblies. Information about the evolutionary distance, including phylogenetic tree topology and evolutionary distances between a reference assembly and the simulated assemblies were obtained from [26].

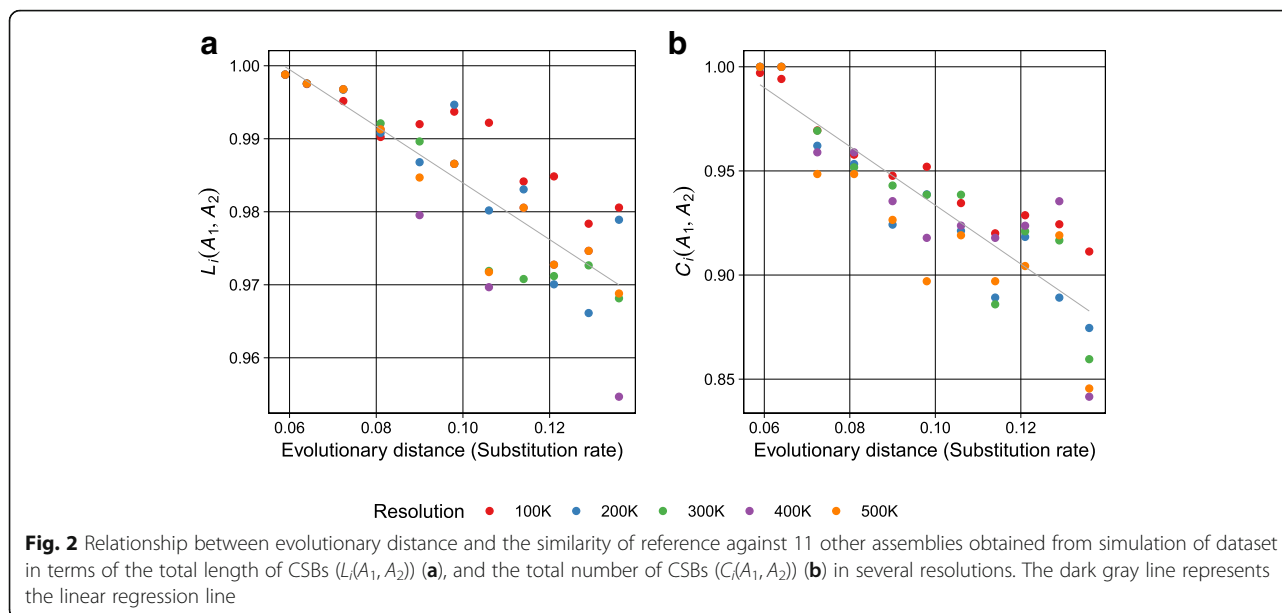
## Results

### Evaluating the GMASS score with simulation data

In this study, it was assumed that if two assemblies are similar, the number and length of CSBs are closer to the number and length of sequence fragments in the assemblies. As shown in Fig. 1b, when two sequences are highly similar, CSBs cover almost the entire sequences, and the total size of CSBs is similar to the one of a smaller sequence (more examples in Additional file 1: Figure S1 and S2). However, CSBs almost covering the entire sequences do not always mean that two assemblies are more similar. If each small region of an assembly is even nearly identical to the counterpart of the other assembly but they are highly rearranged, a large number of small CSBs covering almost the whole of the assemblies can be constructed (Fig. 1c, more examples in Additional file 1: Figure S1 and S2). The GMASS score was developed to provide a single quantitative score measuring the structural similarity of two assemblies by resolving aforementioned issues (Methods).

The GMASS score was first evaluated using simulated genome assembly dataset (Methods). Starting from an ancestral form of an assembly, descendent assemblies in the dataset were simulated with different evolutionary divergence (substitution rate) which determined the amount of perturbation in the assembly simulation process. Therefore, in this study, the evolutionary divergence was treated as the surrogate of structural difference of two assemblies. As shown in Fig. 2, both  $L_i(A_1, A_2)$  and  $C_i(A_1, A_2)$ , which represent total length and number of CSBs respectively, showed strong negative correlation with the evolutionary distances in all five resolutions. The Pearson's correlation coefficients between  $L_i(A_1, A_2)$  and evolutionary distances were -0.925 for





100Kbp resolution ( $p$ -value  $4.52E-5$ ),  $-0.883$  for 200Kbp resolution ( $p$ -value  $3.18E-4$ ),  $-0.955$  for 300Kbp resolution ( $p$ -value  $4.75E-6$ ),  $-0.910$  for 400Kbp resolution ( $p$ -value  $9.96E-5$ ), and  $-0.949$  for 500Kbp resolution ( $p$ -value  $8.67E-6$ ). Similar strong negative correlation was observed in the case of  $C_i(A_1, A_2)$ :  $-0.962$  for 100Kbp resolution ( $p$ -value  $2.30E-6$ ),  $-0.952$  for 200Kbp resolution ( $p$ -value  $6.54E-6$ ),  $-0.924$  for 300Kbp resolution ( $p$ -value  $4.89E-5$ ),  $-0.861$  for 400Kbp resolution ( $p$ -value  $6.60E-4$ ), and  $-0.889$  for 500Kbp resolution ( $p$ -value  $2.49E-4$ ). However, as evolutionary distances between assemblies increased, the variance of the scores among different resolutions also increased. Therefore, the decreased pattern of scores were varied by given resolutions.

Next, the  $S_i$  score combining  $L_i(A_1, A_2)$  and  $C_i(A_1, A_2)$  (Methods) was examined at each different resolution. As shown in Fig. 3a, strong negative correlation was observed from all the five resolutions: Pearson's correlation coefficient  $-0.970$  for 100Kbp resolution ( $p$ -value  $7.40E-7$ ),  $-0.969$  for 200Kbp resolution ( $p$ -value  $9.65E-7$ ),  $-0.951$  for 300Kbp resolution ( $p$ -value  $7.23E-6$ ),  $-0.891$  for 400Kbp resolution ( $p$ -value  $2.30E-4$ ), and  $-0.924$  for 500Kbp resolution ( $p$ -value  $4.71E-5$ ). However, slightly different patterns were observed from different resolutions. To deeply understand the effect of the resolution, the  $S_i$  scores calculated from 11 different assemblies were drawn as a function of resolutions (Fig. 3b). More obvious variations of scores among different resolutions were found especially in the case where the evolutionary distance became larger.

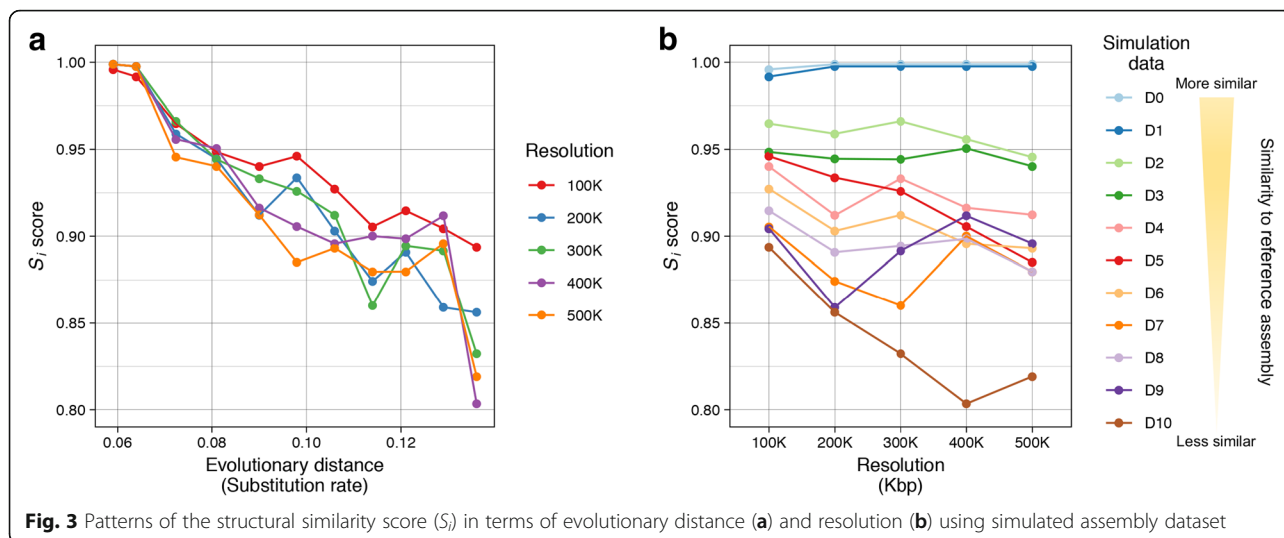
Finally, the GMASS scores were calculated by averaging the  $S_i$  scores from all five resolutions, and comparison made in terms of evolutionary divergences

(Fig. 4). Strong negative correlation between the GMASS score and evolutionary distance was observed (Pearson's correlation coefficient  $-0.962$ ,  $p$ -value  $2.15E-6$ ). There was one exception of the score of D7 which was lower than the ones of D8 and D9, despite being more divergent than D7. This was because larger number of CSBs were constructed from D7 than D8 and D9, although the coverages of CSBs were similar in all three assemblies (Additional file 2: Table S1).

#### Application of the GMASS score to assembly benchmark data

The GMASS score was used to compare each pair of assemblies constructed in the GAGE [18] and Assemblathon 1 [20] dataset (Methods). The statistics of the assemblies were quite varied even though they were constructed using the same sequencing datasets (Additional file 2: Table S2).

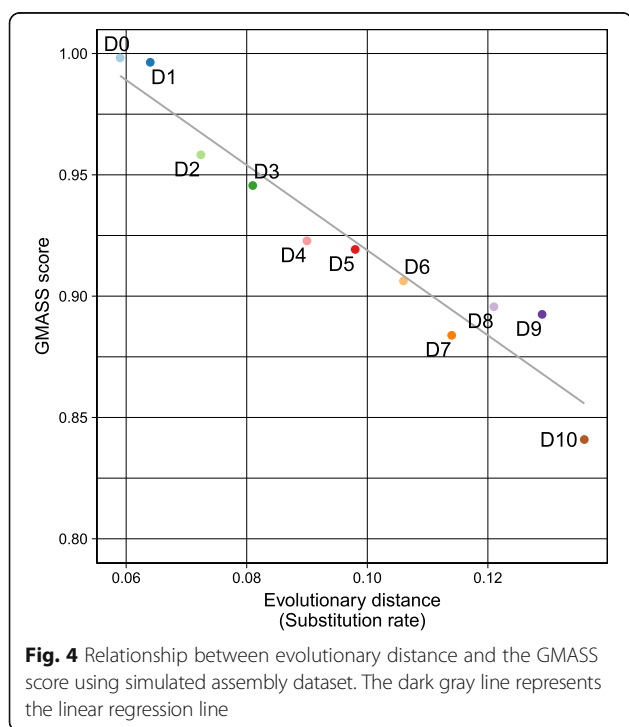
In the case of the GAGE dataset, the GMASS score between the same assembly was approximately 1, and the average of the GMASS scores between different assemblies was 0.537 with standard deviation 0.188 (Table 1). The resolution specific score  $S_i$  of resolution  $i$  calculated from two different assemblies was also varying (Additional file 1: Figure S3, Additional file 2: Table S3). For example, the GMASS score was 0.087 between the MSR-CA assembly and Velvet assembly. The maximum  $S_i$  score between them was 0.4126 when N90 of a non-reference assembly was used as a resolution. However, the minimum score was only 0.0003 when N50 of a non-reference assembly was used as a resolution (see Methods for the selection of resolutions). In the case of the N90 resolution, 253 CSBs were constructed, and



their coverages of compared scaffolds greater than the resolutions of the MSR-CA assembly and Velvet assembly were 0.5023 and 0.4188, respectively (Additional file 2: Table S4). When syntenic relationship was analyzed using the Circos plots drawn by mySyntenyPortal [36], several short CSBs were constructed between one scaffold of one assembly, especially large one, and multiple scaffolds of the other assembly (Additional file 1: Figure S4). However, in the case of the N50 resolution, although 29 scaffolds of the MSR-CA assembly and 51 scaffolds of the Velvet assembly were used to make CSBs, only one CSB was constructed between them. Its

coverage of scaffolds used for comparison was 0.0249 for the MSR-CA assembly and 0.0037 for the Velvet assembly.

In the case of Assemblathon 1 dataset, the average GMASS score was lower than the score from the GAGE dataset (Table 2, average = 0.47, standard deviation = 0.158). The minimum GMASS score of 0.274 was obtained between assemblies from WTSI-S and CSHL. The highest GMASS score was only 0.647 obtained between assemblies from WTSI-S and Broad. The features of constructed CSBs,  $C(A_1, A_2)$ ,  $L(A_1, A_2)$  and  $S_i$  scores were also different when difference resolutions were used (Additional file 1: Figure S5, Additional file 2: Table S5 and S6). The GMASS scores calculated using the self-alignments of assemblies showed almost 1 in both cases of GAGE and Assemblathon 1.



### Discussion

The GMASS score was designed to represent the degree of structural similarity between a pair of assemblies. This new measure was developed based on the distribution pattern of CSBs between a pair of assemblies, which were identified by applying the concept of syntenic block in comparative genomics. With pairwise alignment results, CSBs are constructed with algorithms for constructing syntenic blocks, the GMASS scores are then calculated by considering both the length and number of CSBs.

Based on evaluation with simulated dataset, we found that the GMASS score was highly related to the structural similarity. However, the GMASS score of the D7 assembly in the dataset was lower than the scores of the D8 and D9 assembly, which were more divergent than D7 from a reference assembly. In this case, larger numbers of CSBs were constructed from the D7 assembly than the D8 and D9 assembly, but the CSB coverage of

**Table 1** The GMASS scores of assembly pairs obtained from the GAGE dataset

ABySS2	0.992							
ALLPATHS-LG	0.671	1.000						
Bambus2	0.660	0.523	1.000					
CABOG	0.689	0.620	0.446	1.000				
MSR-CA	0.662	0.447	0.376	0.419	1.000			
SGA	0.687	0.706	0.740	0.759	0.701	0.984		
SOAPdenovo	0.665	0.731	0.542	0.595	0.589	0.734	0.993	
Velvet	0.457	0.134	0.250	0.259	0.087	0.499	0.384	1.000
	ABySS2	ALLPATHS-LG	Bambus2	CABOG	MSR-CA	SGA	SOAPdenovo	Velvet

D7 was higher than D8 and D9. It was as a result of the high rearrangement of small regions in the D7 assembly which are nearly identical to the counterpart of the reference assembly. This kind of rearrangements in small regions were mostly observed in D7 than D8 and D9 (Additional file 1: Figure S6). The CSBs pattern in other resolutions can be identified from our supplementary website (<http://bioinfo.konkuk.ac.kr/GMASS/>).

From the GMASS scores calculated using the assembly benchmark datasets (GAGE and Assemblathon 1), we found that the current assembly tools generated different assemblies even though the same sequencing read datasets are used. For example, the GMASS score between Velvet and ALLPATHS-LG assembly in the GAGE dataset, which were both developed based on the *de Bruijn* graph, was very low. This may indicate that the quality of assembly is not only affected by the graph and traversal algorithm, but also by additional pre- and post-processing algorithms and parameters used in the assembler. In the case of the GAGE dataset, the GMASS scores calculated using the self-alignments of the ABySS2, SGA and SOAPdenovo assembly were not 1 (Table 1). The reason is that there were short scaffolds with high fraction of repetitive sequences, which were masked out in a repeat masking step, and eventually ignored during sequence alignment for the CSB construction.

The GMASS scores can be used in structural-level similarity analysis between diverse assemblies as well as genomes of multiple species or individuals. Especially, because the score was significantly related to the

evolutionary distance, it can be also used as alternative evolutionary distances among species or individuals by complementing traditional nucleotide-level similarity measures, such as ANI [37] and GBDP [38]. This measure can also be used to infer the accuracy of a novel assembly by calculating the score between the assembly and true assembly for target species. In addition, the relative improvement of a new assembler against existing assemblers can be approximately inferred by comparing the GMASS scores which are calculated from true assembly against assemblies created by the new assembler and existing assemblers.

The CSBs of different assemblers can be used to improve or develop the assemblers. For example, researchers can identify regions with structural differences, which are not covered by CSBs among assemblies built by different assemblers. These regions might be too complicated to assemble using existing assemblers or assembly algorithms, therefore new assemblers need to be developed to assemble such regions more accurately.

Resolution is a critical factor to consider when calculating the GMASS score. The resolution makes it difficult to compare scaffolds shorter than a given resolution by ignoring them in the construction of CSBs. In addition, the contiguity of an assembly was quite variable according to the result displayed by GAGE and Assemblathon datasets, it is not possible to use a fixed set of resolutions. Therefore, we used a dynamic set of resolutions for different assembly pairs based on the statistics of scaffold lengths (N50, N60, N70, N80, and N90) of a non-reference assembly for the GAGE and

**Table 2** The GMASS scores of assembly pairs obtained from the Assemblathon 1 dataset. The teams' names constructed each assembly are written in parentheses

ALLPATHS-LG (Broad)	1.000				
ABySS (BGI)	0.642	1.000			
SOAPdenovo (DOEJGI)	0.370	0.632	1.000		
Meraculous (WTSI-S)	0.647	0.607	0.557	1.000	
Celera (CSHL)	0.333	0.313	0.327	0.274	1.000
	ALLPATHS-LG (Broad)	ABySS (BGI)	SOAPdenovo (DOEJGI)	Meraculous (WTSI-S)	Celera (CSHL)

Assemblathon 1 datasets. However, a large fraction of sequences is ignored if a large value is used as a resolution. Further studies to define appropriate resolutions are required.

## Conclusions

The GMASS score is a novel measure for representing structural similarity between two assemblies. The usability of the GMASS score was successfully evaluated using various assembly datasets. The novel measure will contribute to the understanding of assembly output and developing de novo assemblers.

## Additional files

**Additional file 1: Figure S1.** Examples of different patterns of CSBs constructed among human (GRCh38/hg38 assembly) and chimpanzee (Pan\_tro 3.0/panTro5 assembly) chromosomes. Linear plots were drawn by mySyntenyPortal (<https://github.com/jkimlab/mySyntenyPortal/>).

**Figure S2.** Examples of different patterns of CSBs constructed among human (GRCh38/hg38 assembly) and mouse (GRCm38/mm10 assembly) chromosomes. Linear plots were drawn by mySyntenyPortal (<https://github.com/jkimlab/mySyntenyPortal/>).

**Figure S3.** Patterns of  $L(A_1, A_2)$ ,  $C(A_1, A_2)$  and  $S_i$  scores of assembly pairs obtained from the GAGE dataset in different resolutions.

**Figure S4.** Circos plots representing CSBs constructed between the MSR-CA and Velvet assembly in five different resolutions. N50 (843,765 bp), N60 (669,249 bp), N70 (484,156 bp), N80 (298,277 bp), and N90 values (89,649 bp) of a more fragmented Velvet assembly are used as the resolution. Colored and white boxes represent the scaffolds of the MSR-CA and Velvet assembly respectively.

**Figure S5.** Patterns of  $L(A_1, A_2)$ ,  $C(A_1, A_2)$  and  $S_i$  scores of assembly pairs obtained from the Assemblathon 1 dataset in different resolutions.

**Figure S6.** Linear plots representing CSBs between a reference assembly (R\_seq) and simulated assemblies (D7\_seq, D8\_seq and D9\_seq) in 300 K resolution. (DOCX 4759 kb)

**Additional file 2: Table S1.** Statistics of a reference assembly and 11 descendant assemblies, as well as CSBs constructed between them. AS denotes the scaffolds contained each assembly, and SF represents the CSBs constructed between each assembly and the reference assembly.

**Table S2.** Statistics of analyzed assemblies and reference sequences. In the case of the GAGE dataset, the statistics of the reference sequences were calculated using the human chromosome 14 (HG38/GRCh38 assembly) downloaded from the UCSC genome browser [30]. In the case of the Assemblathon 1 dataset, the teams' names constructed each assembly are written in parentheses, and the statistics of reference are the averages of statistics of two haplotypes used in the study.

**Table S3.**  $L(A_1, A_2)$ ,  $C(A_1, A_2)$  and  $S_i$  scores of assembly pairs obtained from the GAGE dataset in different resolutions. In the "Combination" column, the name of an assembly used as reference appears first.

**Table S4.** Statistics of each assembly and CSBs of assembly pairs obtained from the GAGE dataset. AS denotes the scaffolds in each assembly, and SF represents the CSBs constructed between each assembly pair. In combination column, front assembly is used as reference assembly and rear assembly is used as target assembly.

**Table S5.**  $L(A_1, A_2)$ ,  $C(A_1, A_2)$  and  $S_i$  scores of assembly pairs obtained from the Assemblathon 1 dataset in different resolutions. In the "Combination" column, the name of an assembly used as reference appears first.

**Table S6.** Statistics of each assembly and CSBs of assembly pairs obtained from the Assemblathon1 dataset. AS denotes the scaffolds in each assembly, and SF represents the CSBs constructed between each assembly pair. In combination column, front assembly is used as reference assembly and rear assembly is used as target assembly. Used assemblers are omitted. (XLSX 84 kb)

## Abbreviations

CSBs: Consensus segment blocks; NGS: Next-generation sequencing

## Acknowledgements

Not applicable.

## Funding

This paper was supported by Konkuk University in 2017.

## Availability of data and materials

All assembly datasets used in our study are publicly available from previous studies [18, 20, 26]. The consensus segment blocks generated during our study are available at our supplementary website (<http://bioinfo.konkuk.ac.kr/GMASS>).

## Authors' contributions

JK conceived the idea. DK, JL, and JK developed the study, analyzed the results, and wrote the manuscript. DK and JL implemented the scripts and performed the experiments. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declared that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 26 October 2018 Accepted: 3 March 2019

Published online: 18 March 2019

## References

- Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* 2016;17(6):333–51.
- Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2010;11(1):31–46.
- Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. *Hum Mol Genet.* 2018;27(R2):R234–41.
- Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010;95(6):315–27.
- Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics.* 2007;23(4):500–1.
- Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD. Extending assembly of short DNA sequences to handle error. *Bioinformatics.* 2007;23(21):2942–4.
- Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* 2007;17(11):1697–706.
- Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics.* 2008;24(24):2818–24.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005;437(7057):376–80.
- Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. A whole-genome assembly of drosophila. *Science.* 2000;287(5461):2196–204.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 2009;19(6):1117–23.
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 2010;20(2):265–72.
- Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A.* 2011;108(4):1513–8.
- Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–9.



15. White DJ, Wang J, Hall RJ. Assessing the impact of assemblers on virus detection in a De novo metagenomic analysis pipeline. *J Comput Biol.* 2017;24(9):874–81.
16. Lin Y, Li J, Shen H, Zhang L, Papisian CJ, Deng HW. Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics.* 2011;27(15):2031–7.
17. Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS One.* 2011;6(3):e17915.
18. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 2012;22(3):557–67.
19. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics.* 2013;29(14):1718–25.
20. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* 2011;21(12):2224–41.
21. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience.* 2013;2(1):10.
22. Narzisi G, Mishra B. Comparing de novo genome assembly: the long and short of it. *PLoS One.* 2011;6(4):e19175.
23. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* 2007;23(9):1061–7.
24. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31(19):3210–2.
25. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* 2013;14(5):R47.
26. Kim J, Larkin DM, Cai Q, Asan ZY, Ge RL, Auvil L, Capitanu B, Zhang G, Lewin HA, et al. Reference-assisted chromosome assembly. *Proc Natl Acad Sci U S A.* 2013;110(5):1785–90.
27. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 2006;16(12):1557–65.
28. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. Institute for Systems Biology. 2015. <http://repeatmasker.org>. Accessed 29 Aug 2016.
29. Harris RS: Improved pairwise alignment of genomic DNA: The Pennsylvania State University; 2007.
30. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996–1006.
31. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome Res.* 2017;27(5):768–77.
32. Koren S, Treangen TJ, Pop M. Bambus 2: scaffolding metagenomes. *Bioinformatics.* 2011;27(21):2964–71.
33. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 2012;22(3):549–56.
34. Zimin AV, Marcias G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. *Bioinformatics.* 2013;29(21):2669–77.
35. Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. Meraculous: de novo genome assembly with short paired-end reads. *PLoS One.* 2011;6(8):e23501.
36. Lee J, Lee D, Sim M, Kwon D, Kim J, Ko Y, Kim J. mySyntenyPortal: an application package to construct websites for synteny block analysis. *BMC Bioinformatics.* 2018;19(1):216.
37. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol.* 2007;57(Pt 1):81–91.
38. Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics.* 2013;14:60.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

