Deterring the Manipulation of N50 Assembly Quality Scoring

Hershil Devnani
Queen's University
Kingston, Ontario, Canada

Andrew Ma
Queen's University
Kingston, Ontario, Canada

Rayan Shaikli
Queen's University
Kingston, Ontario, Canada

ABSTRACT

With genome assembly becoming a routine procedure along with the advances in next-generation sequencing, it is important to evaluate the methods used to assess the quality of the genome assembly. Significant problems associated with N50, a common assembly quality measure, is the susceptibility of misrepresentation in data sets containing misassemblies and overlapping contigs. In response to the problems, we propose a novel metric, UA50, that determines the U50 score for a set of aligned contig blocks rather than the initial set of contigs. The algorithm for UA50 performs sequence alignment of the initial contigs to the provided reference genome. Misassemblies are identified and the erroneous joints are split into contig blocks which are aligned to the reference genome. UA50 score is computed by calculating the U50 score with the set of aligned blocks. Our results highlight the advantage of UA50 over N50 where a large proportion of the contig set is contains misassemblies. Further exploration opportunities on UA50 include its efficacy in real, published data sets, scoring of largely overlapping contig sets, and a percentage based UA50 metric.

CCS CONCEPTS

• **Applied computing** → **Computational genomics**; **Computational genomics**; **Computational biology**; • **Hardware** → *Biology-related information processing*; • **Theory of computation** → *Bio-inspired optimization*; **Genetic programming**; • **Computing methodologies** → **Genetic algorithms**; **Genetic programming**.

KEYWORDS

genome assembly quality, genomics, computational biology

ACM Reference Format:
Hershil Devnani, Andrew Ma, and Rayan Shaikli. 2021. Deterring the Manipulation of N50 Assembly Quality Scoring. In . ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

N50 statistics is frequently used for assessing genome assembly quality. It is defined as the shortest contig length at the 50% of the total genome length which is equivalent to half of the total length of the contigs. Although N50 is a useful indicator of assembly quality, it can be easily manipulated to output a higher value. In this study, we attempt to address the problems associated with using the N50 statistics. One of the problems using N50 is the misassembly problem. In the genome assembly of the reference genome (Figure 1), three correctly aligning contigs can result from the assembly (Figure 2), each having an equal unit length of $1u$. In this set of three contigs, the N50 is $1u$. However, the three contigs can be joined incorrectly in respect to the reference genome to manipulate the N50 score, resulting in a longer contig with a length of $3u$ (Figure 3). These incorrect joints of a contig is a misassembly which can be a result of poor genome assembly or the use of an incorrect algorithm. This contig set containing multiple misassemblies results in an overinflated N50 score since the contig is three times the size of the correct contigs. The misassembly problem highlights the limitation of N50 to evaluate the correctness of the assembly and its susceptibility to overinflation given a contig set containing high proportions of incorrectly assembled contigs.

Figure 1: Hypothetical reference genome.

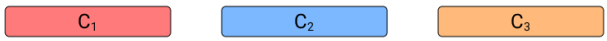

Figure 2: Correctly assembled contigs (C_1, C_2, C_3) with a length of 1u each. The N50 for this contig set is $1u$.

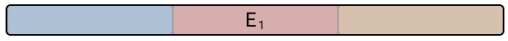

Figure 3: Incorrectly assembled contig (E_1), resulting from erroneous joints between C_1, C_2, C_3. Total length of E_1 is $3u$.

The overlapping contigs problem is the inherent limitation of N50 to assign a high score to a set of largely overlapping contigs. For example, given a reference genome (Figure 4), a possible set of non-overlapping contigs is shown in (Figure 5), and another set of contigs with overlapping regions (Figure 6). The contig set with high proportion of contigs with large overlaps have an incomplete coverage of the reference genome. Another set of shorter contigs containing less overlaps and more coverage results in a lower N50 value compared to the contig set with many overlaps between contigs. The overlapping contigs problem highlights the failure of N50 to correctly evaluate the coverage of the assembly and the susceptibility to overinflation given a contig set containing high proportions of overlapping contigs. The problems associated with

Queen's University - CISC471 2021, April 23, 2021, Kingston, ON, Canada

N50 statistics highlight the need for an alternative assembly quality metric that is able to evaluate the correctness and the coverage of the assembly.

Figure 4: Hypothetical reference genome.

Figure 5: Non-overlapping contigs (C_1, C_2, C_3, C_4), with lengths of $1u$ with complete coverage of reference genome.

Figure 6: Overlapping Contigs (O_1, O_2) with a length of $2u$ with incomplete coverage of reference genome.

In response to the problem, we developed a novel metric called UA50. The UA50 performs sequence alignment bewteen the initial set of contigs and the provided reference genome. When erroneous joints are identified within a contig, it is separated into alignment contig blocks where each block is aligned to the reference genome. This process corrects any misassemblies present in the initial contig set and produces a corrected set of contigs. Only the aligned contigs are stored and sorted by their lengths from the longest to the shortest. From the aligned contig set, unique regions of the contigs are identified using their coordinates in respect to the reference genome. Only the unique segments of the contigs are stored, removing any existing overlaps between the contigs. The unique contig set is sorted in the same manner as the aligned contig set. The cutoff for UA50 is the summation of contig lengths multiplied by the threshold percentage, for UA50 the threshold percentage is 50%. The UA50 score of the assembly is the length of the shortest unique contig at the first instance where the running sum becomes greater than or equal to the UA50 cutoff.

The two problems associated with N50 are the overinflated scores for misassemblies and the large overlappings between the contigs. Our UA50 metric measures the length of the contig which covers 50% of a set of aligned, unique lengths of the contigs. The sequence alignment of the initial contigs to the reference genome identifies the misassemblies present in the contigs and splits any erroneous joints into distinct contig blocks, producing an aligned set of contigs. The alignment processing of initial contigs removes the possibility of overinflated score on misassembled contigs. The aligned contig set is mapped to the reference genome to generate a unique set of contigs where the overlapping portions of each contig is removed. Using a unique contig set removes the possibility of a overinflated score given a contig set with high proportion of large overlaps.

2 RELATED WORK

N50 was first introduced as an alternative measure for contiguity by Lander et. al [8] in the paper regarding the initial sequencing and analysis of the human genome as a part of the Human Genome Project. N50 was defined as the largest length L such that 50% of all nucleotides are contained in contigs of size at least L. The N50 statistic was developed in response to the problems of using the average length of a contig and the length-weighted average length. Average length metric was prone to deflation by a small proportion of the genome consisting of multiple short contigs whereas the length-weighted average length was prone to inflation by large segments of finished sequence. NG50 is a N50 statistics-based measurement that computes the smallest contig length where 50% of the reference genome is contained in contigs of size NG50 or larger [3]. Despite its advantages, N50 is susceptible to inaccurate results [10]. A poor assembly can lead to erroneous joints between unrelated contigs and produce large missassembled contigs, resulting in an inflated N50 score. Another disadvantage of N50 is its inaccuracy when measuring microbial or viral datasets since most of the reads are noise, thus skewing the N50 score [9]. Lastly, N50 does not factor the uniqueness of the contigs which can lead to inflation of N50 given a largely overlapping contig set.

U50 and related metrics were developed to address the problems associated with the N50 statistic [2]. U50 is the length of the smallest contig such that 50% of the sum of all unique, target-specific contigs is contained in contigs of size U50 or larger. U50 first sorts the contigs by their lengths in decreasing order. All contigs are then mapped to the reference genome, preserving the unique portions of each contig and removing non-unique regions of each contig. The unique portions of contigs are sorted in decreasing lengths. From the set of unique contigs, the cutoff is defined as the summation of all unique contigs multiplied by the threshold percentage. A cumulative sum of unique contigs is compared against the threshold percentage. The contig length at the first instance where the cumulative sum is greater or equal to the cutoff is the U50 score. UL50 describes the number of contigs at the first instance where cumulative sum is greater or equal to the cutoff. UG50% is the percentage score of U50. Compared to N50, U50 metric produced a more accurate measure of assembly given sequences containing high noise, such as viral and microbial sequencing. Limitations of U50 include the requirement of a reference genome and lack of consideration for genome coverage due to the removal of overlapping regions.

Similarly, Gnerre et al. [4] conducted research to develop new assembly techniques to baseline the quality of assembly reads against a reference genome of a related species. The use of reference genomes may seem counter-intuitive since you first need a genome to be able to sequence back against that genome. However, if we know about two closely related species and happened to have the genome data available for one, we can use that as reference to baseline the assembly quality given the data set of assembly contigs for the related species.

Research has also been conducted to reduce the impact of error rates in the raw data obtain for a sequencing assembly. Using techniques such as the N50 and L50 metrics in conjunction with average contig-to-chromosome levels can give a better estimation of the

assembly quality in these cases, by using linkage maps to filter out erroneous contigs [6]. Other techniques used to compare the contig data set quality include the quality assessment tool for evaluating and comparing genome assemblies (QUAST) software. The QUAST project aims to use the common scoring metrics, such as N50 and L50 amongst other, in conjunction with algorithms to better summarize these different metrics to both mitigate errors and provide an accurate assessment of quality of a genome assembly against the contig data set [5]. Some of the other metrics include not only assembly quality metric, but making educated observations between patterns of GC ratios, duplication ratios, insertion/deletions, and prediction for the number of coding genes that may match with sequences of contigs [5].

With the growing database of genomes, it has now become possible to sequence entire genomes solely given reference to existing genomes data and given large contig reads. Specifically, microbial genomes can be automatically assembled using long contig read data, pieced together using a jigsaw puzzle like approach. Even further, these approaches can almost guarantee greater that 35% of assemblies are 99.99% accurate from 2007 and 2011 [1]. Using PacBio, a long-read sequencer, will further improve these numbers over time. With this in mind, reference genomes can be leveraged to develop new metrics that can assess assembly quality with a much greater degree of accuracy given underlying references that can be used.

Taking scoring analysis to the next level, Naccache et al. worked on an unbiased, next-generation sequencing methods that perform pathogen detection by using high performance, low latency cloud servers. With this compute power, and access to open-source databases as a foundation, such as BLAST, cloud computing was able to accelerate the analysis of next-generation sequencing data of more than 1.1 billion sequences in as quickly as eleven minutes [9]. This breakthrough is significant as it unlocks access to increased compute power to labs and researchers with limited budgets, to conduct analysis and further the collective development of contigs analysis.

In addition to advances in cloud computing, machine learning algorithms have become instrumental in increasing the available throughput to conduct genome assemblies. These enhancements to alignment and assembly scoring algorithms allowed the uncovering and rejection of the overuse in N50 score that are easily manipulated [12]. Two new metrics were developed with new machine learning capabilities in mind. The first of which being a metric to compute the distance from an assembly to a given reference genome. Secondly, a method to conduct model-based learning and retraining against existing data models of genome assembly was introduced.

Johnson et al. have developed an automated pipeline that would programmatically annotate short-read data from species that have no existing reference genomes. These short reads would then be assembled using different automated pipelines that differed algorithmically, and compared against pipelines that were developed by other research groups, specifically with the Nation Center for Genome Research [7]. Their research concluded that there is currently no single best approach when assembling contigs or quantifying the quality of a data set of contigs. The analysis of dissembling and reassembling data is something that can do to improve upon the computation conduced by automated pipeline in counting analysis. This approach is something we aim to further extend with our experiments on simulated disassembly of a reference genome and scoring that assembly set back to in an attempt to train machine learning models to enhance the overall predictability of contig assembly quality given raw data.

3 APPROACH

The approach taken to develop our Unique Assembly metric involved taking a look at the limitation of the current NXX scoring metric and UXX scoring metric.

We found that the NXX algorithm is one of the most widely used and regarded algorithm used to base contig assembly quality against. A greater NXX score relative to the length of the genome generally indicates a better assembly and returns the length of the contig at the XX percent of the total genome assembly length. However, some limitations we observed with NXX assembly scoring algorithm included the misleading nature of the metric in terms of how it can be very easily, and is often times, manipulated to inflate assembly scores by means of filtering out short contigs and with the introduction of misassembly biases.

The most prominent factor that causes the NXX scoring algorithm to be rendered misleading is the use of filters. Many filtering techniques can be used to filter out small length contigs in the assembly data set, which inherently skews the NXX results by increasing the likelihood of smallest contig size making up 50% of the genome assembly to be larger in length than would be without the filtering. Secondly, using a scoring algorithm that incorrectly joins contigs together during assembly, by unjustifiably algorithmically joining certain contigs together to increase the average length of contigs in the data set, which inadvertently construes a greater NXX score. This leads to a positive feedback generated by manipulating the underlying data sets and the resulting inflation in scores due to skewness in proportionality of contig lengths.

These issues regarding inflated assembly quality scores defeat the purpose of the metric in the first place, to assess the quality of the genome assembly given a set of contigs. If the NXX metric can be manipulated easily, it takes away from the efficacy of using the metric as a basis to conduct further analysis and quantify the integrity of an underlying data set. In order to combat these issues, we decided to take a look at these limitations. First, we took a look at properly assembly, since NXX lines up contigs sequentially, not considering unique contig matches to the genome. Second, we wanted to develop a method that discourages filters or manipulation of joining contigs together to inflate assembly quality scores. While the U50 metric addresses the uniqueness of contig assembly against a reference genome, it still does not deter from manipulating the data set to inflate the score since filtering and joining contigs will still inflate this score.

And so, to combat this, out approach was to mitigate these artificial inflations by introducing a new metric based on the both the NXX and UXX metric, that will enhance the raw estimated assembly score given a set of contigs. The idea was to create a new metric that can be used to compare the assembly of different sets of contigs against a reference genome that is deterrent to inflated scoring caused by filtering, and being a subsequent extension of

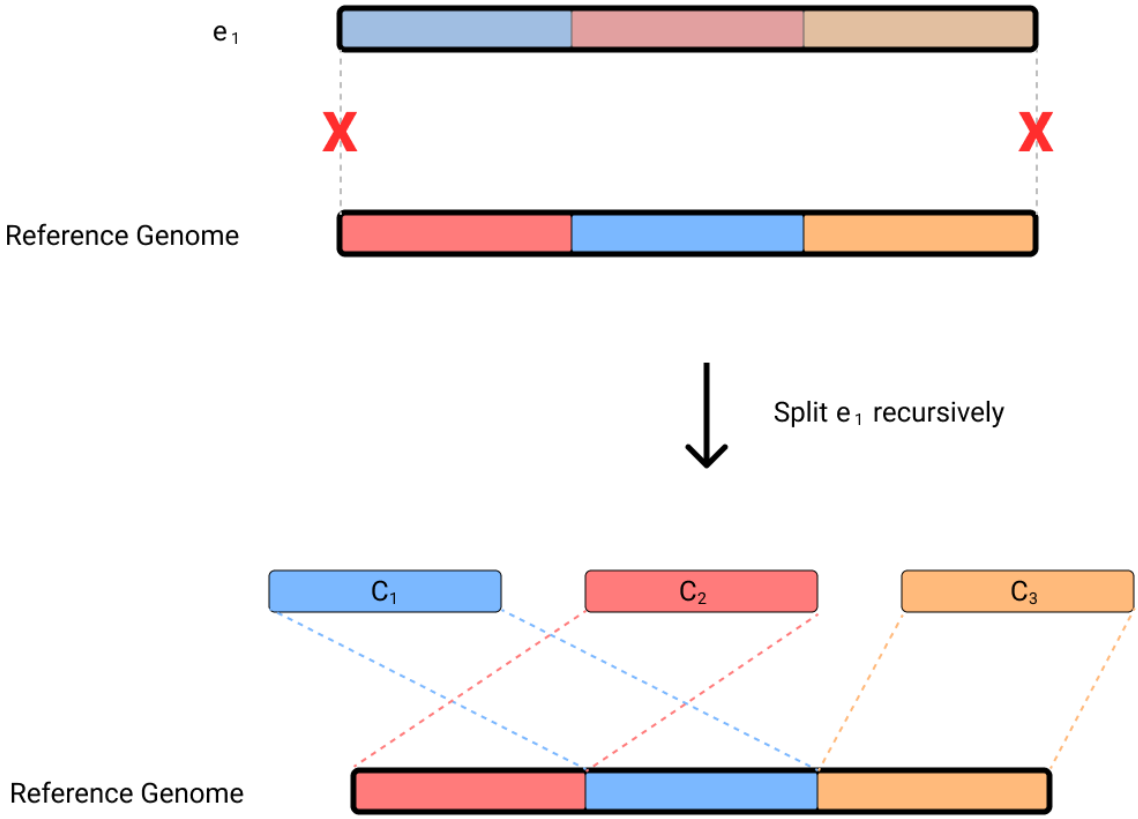

Figure 7: E_1 does not completely align with the reference genome. e_1 is recursively split into three new contigs (C_1, C_2, C_3, C_4) that completely align to the reference genome.

the UXX metric, acts a baseline for sequence assembly for related organisms.

Our new Unique Assembly, UA50, metric provides a metric to understand the assembly quality of a raw set of data. Our approach was to enable an accurate quantifiable score that determines assembly quality based on correcting misassemblies and identifying the unique matches to contigs against a reference genome. Our algorithm matches the unique contigs matches against a genome using a mask array, in which once a contigs matches a portion of the genome, we do not double count the length of subsequent contigs that may overlap in the same portion of the genome. This provides a deterrent to filtering out short contigs or joining contigs together since they will neither promote nor detract from the metrics scoring analysis. We chose to use a reference genome in our scoring metric since we found the NXX score to arbitrary, so a basis of reference genome will provide a better indicator genome assembly of related organisms.

Overall, the key approach to our metric as to develop something that discourages skewing or filtering data for falsely better N50 scores when there is no basis for the score to increase [11].

3.1 Misassembly Correction

The process of identifying and correcting misassemblies in the initial contig set involves the alignment of the contig set to the given reference genome. Here, we assume that all of the correct contigs are completely aligned to the reference genome, meaning the contig is a subsequence of the reference genome. Each contig is checked and added to a list of aligned contigs if it is completely aligned. A misassembly is defined as the incorrect joint between contigs, in other words, the misassembled contig is not a subsequence of the reference genome. When a misassembled contig is identified, it is recursively split into individual contigs until each of the distinct contigs align to the reference genome (Figure 7). The resulting contigs and their coordinates in respect to the reference genome are then added to the list of aligned contigs. Using this process removes all the erroneous joints in the misassembled contigs and produces a set of correct contigs from the misassembled contigs.

3.2 UA50 Calculation

Step 1: The first step for calculating UA50 involves performing sequence alignment on the initial set of contigs (c) to the reference genome. Contigs that do not completely align with the reference genome are considered misassemblies (e), since misassemblies are composed of one or more incorrect joint between contigs that are aligned to the reference genome. Any contig with misalignment in the initial contig set is recursively split into equal length contig blocks until the individual blocks are aligned to the reference genome. The new contig blocks and the aligned contigs (a) are added to the aligned contig set (A).

$$e = c_1 + ... + c_n \tag{1}$$

$$A = a_1 + a_2 + ... + a_n \tag{2}$$

Step 2: The aligned contigs are sorted by their lengths from the longest (a_1) to the shortest (a_n).

$$sort(A) = a_1 + a_2 + ... + a_n \tag{3}$$

Step 3: A mask array is created which has the same length as the reference genome. Initially, the mask array consists of only "0"s which indicates the unmapped regions of the reference genome. The initial values are replaced by 1 when a region of a contig maps to the reference genome. The mapping of the contigs to the reference genome is accomplished by comparing the contig coordinates in respect to the reference genome.

Step 4: Using the mask array, each of the aligned contigs are mapped to the reference genome. A running sum is used to track the number of unique mapping of the contig. This sum represents the length of the unique segments of each of the aligned contigs. Only the unique segments of each aligned contig are stored (a') in the unique contig set (U), disregarding the overlapping portions of the aligned contigs. The unique contigs are sorted by their lengths from the longest (a'_1) to the shortest (a'_n).

$$sort(U) = a'_1 + a'_2 + ... + a'_n \tag{4}$$

Step 5: The cutoff for UA50 is given by the summation of contig lengths multiplied by the threshold percentage. For UA50, the threshold percentage is 50%.

$$UA_{50}cutoff = (\sum_{k=1}^{n} a'_k) \times 50\% \tag{5}$$

Step 6: The running sum is calculated by adding the lengths in the sorted unique contig set in a sequential order. The UA50 score is the length of the shortest contig at the first instance where the running sum is greater or equal to the UA50 cutoff.

$$UA_{50} = a'_k, where(\sum_{k=1}^{n} a'_k) \geq UA_{50}cutoff \tag{6}$$

3.3 Testing

The first step was to baseline the results by understating how data sets respond to change in N50, L50, U50 and UG50% metrics. Next, we took a look at how simulated filtering and assembly manipulation by joining shorter contigs together would affect the efficacy of the N50 and U50 score.

In order to generate the sample data used to conduct our experiment, we develop a function that would take in the following parameters:

- Proportion of small to medium to large length contigs in the data set
- The size of the data set or the number of contigs present in the data set
- The reference genome to base contigs off in an attempt to simulate the assembly data set for related organisms (introducing 5% error to random contigs)

In order to be able to test our metrics resistance to change when manipulating the data sets, we added an options parameter:

- The amount of simulated filtering and contig joining to a percentage of the entire data set

These control variables in our sample data generator allowed us to build data sets that were used to baseline our understanding of the different scoring algorithms and enable us to test our proposed advantages by introducing the same manipulations usually done to inflate NXX and UXX scores. Our data generator used parameters for the size of contigs based on our experience with de novo assembly the proportions of contig lengths generally considered to be small (36bp-150bp), medium (150bp-250bp) and large (250bp-500bp). These parameters were set as global variables and can be adjusted in the code very easily.

4 RESULTS AND DISCUSSION

We constructed two experiments assessing the performance of the UA50 metric using theoretical contig data sets. One reference genome was used to generate the data sets per experiment. Total of seven theoretical contig data sets were generated. Three were used in the first comparative analysis concerning the scoring metrics given contig sets with varying proportions of short, medium, and large contig lengths, the proportions are indicated in Table 2. Four contig sets were generated for the comparative analysis between N50 and UA50 with different percentages of induced error, the percentages of error are indicated in Table 4.

The first experiment, shown 1, involves the comparative analysis of the UA50 metric to pre existing metrics on varying contig length composition in the theoretical data set, without any misassemblies. The UA50 is consistent in all three data sets. This behaviour is equivalent to the U50 results, having the same score for the different proportions of contig lengths. Similarly, UG50% is consistent in the three theoretical contig sets. Conversely, the N50 appears to be increasing as the proportions of large contigs increase in the data sets. The L50 follows the same suit as the N50 where its value increases with increase of the proportions of large contigs in the data sets.

The second comparative analysis, shown in Table 3, involves observing the behaviours of N50 and UA50 when misassemblies are induced as the erroneous joints of contigs. The N50 score increased as the percentage of induced error increased whereas UA50 remained resilient to any amount of induced error and computed a consistent score for all four contig data sets. Figure **??** displays the results as the change in the assembly metric score to the percent

Table 1: Comparative Analysis of Assembly Scoring Metrics based on Contig Data Set Features (Average of 5 Trials)

Metric	Control	SmallSkew	MediumSkew	LargeSkew
N50	285.0	154.0	217.0	372.0
UA50	498.0	499.0	499.0	499.0
U50	498.0	499.0	499.0	499.0
UG50pct	99.6	99.8	99.8	99.8
L50	318.0	277.0	421.0	407.0

Table 2: Comparative Analysis of Composition of Contigs for Trial Data

ContigParam	Control	SmallSkew	MediumSkew	LargeSkew
NUM	11250.0	11250.0	11250.0	11250.0
SMPCT	33.618	75.147	13.307	12.604
MDPCT	33.191	12.462	74.24	12.658
LGPCT	33.191	12.391	12.453	74.738

Table 3: Comparative Analysis of N50 vs UA50 Given Percentage of Induced Error

Metric	N50	UA50
5.0	269.0	426.0
10.0	289.0	426.0
25.0	328.0	426.0
50.0	357.0	426.0

Table 4: Comparative Analysis of Composition of Contigs for N50 vs UA50 Experiment Given Percentage of Induced Error

ErrorPCT	NUM	SmallPCT	MediumPCT	LargePCT
5.0	11811.0	48.167	25.933	25.933
10.0	12374.0	46.307	26.75	26.75
25.0	14061.0	41.299	29.052	29.052
50.0	16873.0	35.566	32.17	32.17

induced error. This graph reinforces the resiliency of UA50 in contrast to the increasing N50 for the proportions of misassembly in the contig data set.

Given the results, the UA50 metric proves to produce a consistent score regardless of the number of misassemblies present in the contig data set. The UA50 metric's misassembly correction process removes the influence of the erroneous joints in the contig sets to the UA50 score. The resiliency of the UA50 metric hints the usefulness of the UA50 metric as an alternative assembly measurement in a data set containing misassemblies due to the nature of N50 to produce an overinflated score that does not accurately represent the quality of the assembly. The UA50 metric also harbours the advantages of the U50 metric given a contig set without any misassembly since it behaves identically. The advantages of U50 include eliminating the influence of overlapping contigs , eliminating the influence of an abundance of small contigs and its usefulness for viral and microbial sequencing of samples with high background noise.

Despite the advantages, the UA50 bears limitations that prevent the usage of the metric to be used for every situation. Firstly, the alignment correction process assigns the start and end coordinates of the split contigs when the first instance of the contigs are found in the reference genome. This may overrepresent the beginning portion of the reference genome, resulting in a highly overlapped set of aligned contigs. Although the overlaps are removed in the unique contig set generation, the resulting unique contig set may have decreased coverage of the reference genome. Secondly, the requirement for a reference genome and the coordinates of the contigs poses a limitation for the UA50 to be used under all circumstances. This requirement restricts use of UA50 to be used in assembling

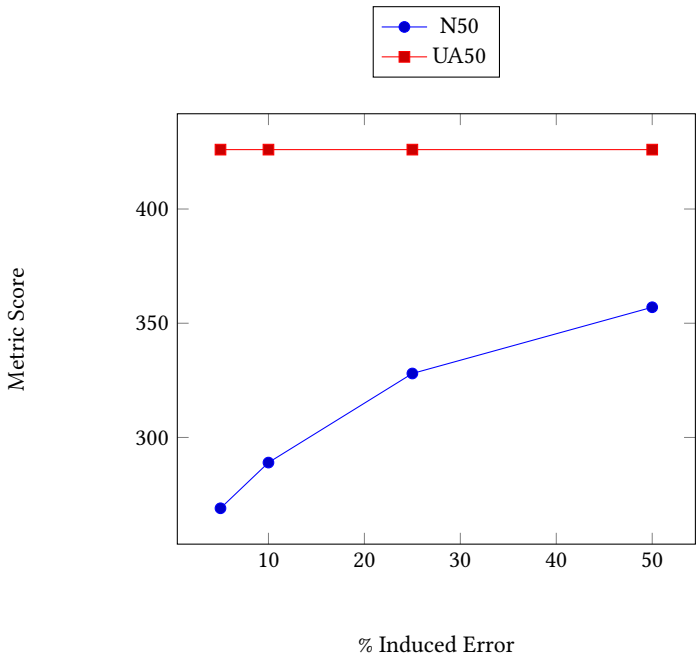

Figure 8: Comparative Analysis of Composition of Contigs for N50 vs UA50 Experiment Given Percentage of Induced Error

novel genomes. Lastly, the UA50 metric shares the limitation of the U50 metric of the possibility that the unique contig set does not sum to the desired percentage threshold. In these situations, a lower percentage threshold should be used such as UA25 or U10. This limits the UA50 to be used in any set of contigs and the reference genome.

Overall, the study shows the potential of the UA50 metric as a measurement for assembly quality given a theoretical contig set with misassemblies and a reference genome. Further research includes comparative analysis using contig sets with different proportions of overlaps. Furthermore, testing the usefulness of UA50 with published data sets containing misassemblies would be a crucial step in having the UA50 as a standard metric in assessing assembly quality.

Further research extension to our newly develop metric could include building machine learning models based on our scoring metric. Since we are taking a simulated reference genome and dissembling it into a raw data set, we can take a close look at programmatically comparing the disassembly and reconstruction quality of simulated data sets. Changing the parameters of data set, we can understand better the resilience of our metric to erroneous data, and potentially come up with algorithmic models that can score assembly sets by removing errors in the data. We can do this by using a reference genome to compute a most probably assembly that may not necessarily have to use all of the data contained in a raw data set. Lastly, we can look a machine learning model that can be used to predict assembly quality of new data sets by using simulated and real data sets based off of a real genome, to strengthen the validity of our new metric. By using re-constructed assemblies, our new metric can train machine learning models to identify how well a given data sets containing errors will score against a reference genome. The models can be developed over time and automate both data collection and machine learning model training to continually improve our quantitative understanding of raw contig data sets.

5 CONCLUSION

In this study, we explore the UA50 metric as an alternative tool to evaluate assembly quality, aiming to improve on the inherent limitations of the N50 metric. The problems associated with the N50 metric include its tendency to produce overinflated scores when the contig set contains high proportion of large overlaps and when the contig set contains high proportion of misassemblies, defined as erroneous joints between contigs. The UA50 is the proposed metric for assembly quality to address the N50 limitations regarding overlaps and misassemblies within the contig set. Given a reference genome, the UA50 performs sequence alignment to identify the erroneous joints within a contig and splits the contig into smaller contigs that align to the reference genome. Following the misassembly correction, the overlaps are removed in the aligned contig set by mapping the contigs to the reference genome and only storing the unique regions of each contig, resulting in a unique contig set. The UA50 score is defined as the length of the smallest contig such that 50% of the sum of all aligned, unique contigs is contained in contigs of size UA50 or greater.

The initial step in the UA50 algorithm is misassembly correction. The misassembly correction involves comparing individual contigs to the reference genome to determine the correctness of the contig. This is accomplished by checking if a contig is a substring to the reference genome. If the contig is not a substring to the reference genome, the contig is split into equal length contigs and each of the contigs are recursively checked against the reference genome until the distinct contigs are substrings. These correct contigs originating

from the misassembly are stored as an aligned contig set. The overlapping regions of the aligned contigs are removed by mapping the aligned contigs to the reference genome. Using a mask array, only the unique segments of the contigs are stored as a unique contig set. The UA50 cutoff is defined as the summation of unique contig lengths multiplied by 50%, the threshold percentage. The UA50 score is calculated as the length of the contig where the running sum of the contig lengths starting from the longest unique contig is greater or equal to the cutoff value. Thus, the UA50 measures the performance of an assembly using an aligned, unique set of contigs.

Two sets of comparative analysis were conducted. The first analysis tested the behaviour of the UA50 metric given error-free contig sets with varying percentages of short, medium-length, and long contigs. As expected, the UA50 metric behaved similarly to the U50 metric since no misassemblies were present in the contig sets. The second set of analysis observed the UA50 score given a data set containing varying proportions of misassemblies. Regardless of the proportions of the misassemblies, the UA50 metric was consistent whereas the N50 metric showed its inherent limitations to overinflate its scores when the miassemblies are introduced.

The UA50 metric has advantages measuring the assembly quality with contig sets containing misassemblies and overlapping contigs. Despite the advantages, the UA50 cannot be used under all circumstances due to the requirements for a reference genome and relative contig coordinates. Additionally, it has the tendency to overrepresent the beginning region of the reference genome.

Further exploration for UA50 as an alternative assembly quality metric include comparative analysis using real, published data sets. For now, we can only tell that our new metric works well for very closely related genomes that can be reference basis for the new assembly and is very resilient to manipulated data sets that may artificially increase N50 scores. Going forward, we can use our new metrics resiliency to changes in raw data sets to train models that can better predict assembly quality, by reverse engineering data sets given a reference genome and observing changes in our UA metric against simulations. These simulations could include introducing artificial genetic distance in the data or using data sets gathered form related species, in an effort to both deter the manipulation of raw data sets that could lead to misrepresented and inflated data score as well as provide a more informative metric to assess assembly quality.

ACKNOWLEDGMENTS

To Dr. Benjamin T. Cecchetto for the guidance throughout the CISC 471 Computational Biology course at Queen's University and laying the framework for this research project.

REFERENCES

[1] 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* 23 (Feb. 2015), 110–120. https://doi.org/10.1016/j.mib.2014.11.014

[2] Christina J. Castro and Terry Fei Fan Ng. 2017. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. *Journal of computational biology : a journal of computational molecular cell biology* 24, 11 (Nov. 2017), 1071–1080. https://doi.org/10.1089/cmb.2017.0013

[3] Dent Earl, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R. Zerbino, Mark Diekhans, Ngan Nguyen, Pramila Nuwantha Ariyaratne, Wing-Kin Sung, Zemin Ning, Matthias Haimel, Jared T. Simpson, Nuno A. Fonseca, İnanç Birol, T. Roderick Docking, Isaac Y. Ho, Daniel S. Rokhsar, Rayan Chikhi, Dominique Lavenier, Guillaume Chapuis, Delphine Naquin, Nicolas Maillet, Michael C. Schatz, David R. Kelley, Adam M. Phillippy, Sergey Koren, Shiaw-Pyng Yang, Wei Wu, Wen-Chi Chou, Anuj Srivastava, Timothy I. Shaw, J. Graham Ruby, Peter Skewes-Cox, Miguel Betegon, Michelle T. Dimon, Victor Solovyev, Igor Seledtsov, Petr Kosarev, Denis Vorobyev, Ricardo Ramirez-Gonzalez, Richard Leggett, Dan MacLean, Fangfang Xia, Ruibang Luo, Zhenyu Li, Yinlong Xie, Binghang Liu, Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Shuangye Yin, Ted Sharpe, Giles Hall, Paul J. Kersey, Richard Durbin, Shaun D. Jackman, Jarrod A. Chapman, Xiaoqiu Huang, Joseph L. DeRisi, Mario Caccamo, Yingrui Li, David B. Jaffe, Richard E. Green, David Haussler, Ian Korf, and Benedict Paten. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research* 21, 12 (Dec. 2011), 2224–2241. https://doi.org/10.1101/gr.126599.111

[4] Sante Gnerre, Eric S Lander, Kerstin Lindblad-Toh, and David B Jaffe. 2009. Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biology* 10, 8 (2009), R88. https://doi.org/10.1186/gb-2009-10-8-r88

[5] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 8 (April 2013), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

[6] Vasanthan Jayakumar and Yasubumi Sakakibara. 2017. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Briefings in Bioinformatics* 20, 3 (Nov. 2017), 866–876. https://doi.org/10.1093/bib/bbx147

[7] Lisa K Johnson, Harriet Alexander, and C Titus Brown. 2019. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience* 8, giy158 (April 2019). https://doi.org/10.1093/gigascience/giy158

[8] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 6822 (Feb. 2001), 860–921. https://doi.org/10.1038/35057062

[9] Samia N. Naccache, Scot Federman, Narayanan Veeraraghavan, Matei Zaharia, Deanna Lee, Erik Samayoa, Jerome Bouquet, Alexander L. Greninger, Ka-Cheung Luk, Barryett Enge, Debra A. Wadford, Sharon L. Messenger, Gillian L. Genrich, Kristen Pellegrino, Gilda Grard, Eric Leroy, Bradley S. Schneider, Joseph N. Fair, Miguel A. Martínez, Pavel Isa, John A. Crump, Joseph L. DeRisi, Taylor Sittler, John Hackett, Steve Miller, and Charles Y. Chiu. 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research* 24, 7 (July 2014), 1180–1192. https://doi.org/10.1101/gr.171934.113

[10] Derrick Scott. 2014. Utilizing Next Generation Sequencing to Generate Bacterial Genomic Sequences for Evolutionary Analysis. *Theses and Dissertations* (Aug. 2014). https://scholarcommons.sc.edu/etd/2887

[11] Adam Thrash, Federico Hoffmann, and Andy Perkins. 2020. Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics* 21, 4 (July 2020), 249. https://doi.org/10.1186/s12859-020-3382-4

[12] Charles Adam Thrash. 2019. A Machine Learning Approach to Genome Assessment. *Mississippi State University* (2019). https://ir.library.msstate.edu/handle/11668/14514