Identifying wrong assemblies in *de novo* short read primary sequence assembly contigs

VANDNA CHAWLA[1,2], RAJNISH KUMAR[1] and RAVI SHANKAR[1,*]

[1]*Studio of Computational Biology & Bioinformatics, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Palampur, Himachal Pradesh, India*

[2]*Department of Biotechnology, Guru Nanak Dev University, Amritsar, Punjab, India*

**Corresponding author (Email, ravish@ihbt.res.in)*

With the advent of short-reads-based genome sequencing approaches, large number of organisms are being sequenced all over the world. Most of these assemblies are done using some *de novo* short read assemblers and other related approaches. However, the contigs produced this way are prone to wrong assembly. So far, there is a conspicuous dearth of reliable tools to identify mis-assembled contigs. Mis-assemblies could result from incorrectly deleted or wrongly arranged genomic sequences. In the present work various factors related to sequence, sequencing and assembling have been assessed for their role in causing mis-assembly by using different genome sequencing data. Finally, some mis-assembly detecting tools have been evaluated for their ability to detect the wrongly assembled primary contigs, suggesting a lot of scope for improvement in this area. The present work also proposes a simple unsupervised learning-based novel approach to identify mis-assemblies in the contigs which was found performing reasonably well when compared to the already existing tools to report mis-assembled contigs. It was observed that the proposed methodology may work as a complementary system to the existing tools to enhance their accuracy.

[Chawla V, Kumar R and Shankar R 2016 Identifying wrong assemblies in *de novo* short read primary sequence assembly contigs. *J. Biosci.* **41** 455–474]

1. Introduction

The Sanger sequencing method (Sanger *et al*. 1977) and subsequent improvements in sequencing technologies have revolutionized the world of genome science, deeply impacting the field of modern biology including areas like genetics, crop improvement, disease understanding and solutions, evolution, systems biology, and general understanding of molecular systems of life (Poon *et al.* 2010; Berkman *et al.* 2012; Henry 2012; Consortium TEP 2012b; Fu *et al.* 2013). The advancements in sequencing technologies have impacted the sequence assembling methods. The process of sequencing involves decoding of every base in sequential manner, which runs usually for shorter length of DNA, producing reads. Such short, single-shot products of sequencing are brought together based on their relative position through the process of assembling, giving rise to larger continuous stretches of stitched reads, called contigs. Contigs are further joined together using distant and common paired reads, forming scaffolds. With advent of much cheaper and faster massively parallel sequencing technologies like Illumina and 454, many species are now being sequenced for their genome, using *de novo* assembling methods (Argout *et al.* 2011; Wang *et al.* 2012; Li *et al.*

Keywords. Assembly validation; clustering; contigs; *de novo* assembly; mis-assembly; next generation sequencing; reads

Abbreviations used: ACC, accuracy; BAC, bacterial artificial chromosome; CE, compression-expansion; FCD, fragment coverage distribution; FN, false negative; FP, false positive; MCC, Matthews correlation coefficient; NCBI, National Center for Biotechnology Information; PDBG, paired *de Bruijn* graphs; PE, paired end; SBS, sequencing-by-synthesis; SE, single end; SRA, sequence read archive; TN, true negative; TP, true positive; WGS, whole genome shotgun

Supplementary materials pertaining to this article are available on the Journal of Biosciences Website.

DOI: 10.1007/s12038-016-9630-0

Published online: 5 August 2016

2010b). However, these next generation sequencing technologies produce very short sequence reads compared to the older sequencing methods like shotgun sequencing methods. Performing assembling using very short reads poses a big challenge in obtaining correctly assembled sequences.

Genome size, duplication, and its complexity are the major decisive factors considered for prior assessment of all genomes to be sequenced. The human genome is a diploid genome of about 3.3 Gb size, consisting of repeats of varying length. The plant genomes are also large and have much higher ploidy levels, higher rates of heterozygosity and complex repeats. Major challenge in sequence assembling is posed by any sort of sequence redundancy and morphism. Therefore, the main challenge for the assembly algorithms is to find the true overlaps to prevent mis-assembling. A related issue is the varying length of complex repeats which may range from few hundred bases to several kilo bases, from short SINEs to huge and prevalent transposons and LTR elements, making it difficult to most of the existing sequencers to pass them. It is easier to sequence the simple repeats regions compared to long complex repeats regions which exist widely and interspersed with regions longer than the reads produced by several existing NGS platforms.

Assessment of *de novo* assembly from short read data is of prime importance as genome sequencing projects for several species have increased rapidly. The rapid increment in NGS-led sequencing projects, relying mainly upon the short reads have achieved good acceptance while solving complex genomes like that of cotton (Wang *et al.* 2012), panda (Li *et al.* 2010a), turkey (Dalloul *et al.* 2010), cucumber (Huang *et al.* 2009), and many more (Ewing and Kazazian 2011; Consortium T 1000 2010, 2012a). The benchmarking projects like Assemblathon 1, Assemblathon 2 and GAGE for comparative study among assemblers, datasets and assembly parameters have already been established (Earl *et al.* 2011; Salzberg *et al.* 2012; Bradnam *et al.* 2013). These projects detailed the performance of various assemblers for *de novo* genome assembly and describe how well the assemblers performed on various genomes. However, no clear evidence was given to select the 'best' assembler, as the assemblers produced different quality of results on different datasets. The entire analysis was dependent upon the availability of reference genome, which turned out to be a limiting point for the up-coming *de novo* genome assembly-based projects. Also, tools like Amosvalidate (Phillippy *et al.* 2008), CGAL (Rahman and Pachter 2013), REAPR (Hunt *et al.* 2013), FRCbam (Vezzi *et al.* 2012) and ALE (Clark *et al.* 2013) have been developed for assembly analysis without available reference. Most of them share similarity in assembly analysis process as they use positions of read pairs within the assembly. The Amosvalidate tool (Phillippy *et al.* 2008) uses the compression-expansion (CE) statistic (Zimin *et al.* 2008) to identify the regions of assembly where paired-end reads deviate from an expected normal distribution for the distance between the partner reads or fragment size. It also calculates statistics based on the overall read coverage, the distribution of *K*-mers, and the presence of fragmented read alignments. CGAL and ALE both produce a summary likelihood score of an assembly, while FRCbam uses a number of metrics to identify features which correspond to erroneous regions in an assembly, used to plot a feature response curve, evaluating the assembler's performance and precision. More recently, the Recognition of Errors in Assemblies using Paired Reads (REAPR) tool (Hunt *et al.* 2013) applied similar metrics of fragment coverage and insert-size distribution to identify mis-assembled regions and introduced the ability to call errors at specific bases in an assembly hypothesis. REAPR is reported to work best using mapped read pairs from a large insert library (at least 1000 bases).

The above-mentioned assembly analysis tools came up with effort to identify regions of mis-assembly within a single assembled sequence, or they try to recognize the best assembly either from several assemblies obtained from different assemblers or done using different combinations of assembly parameters. The existing tools to identify mis-assemblies display lots of scope for improvement. The present study provides an assessment and comparative analysis using data from different sequencing platforms, analyzing various possible factors and parameters for mis-assembly, discuss the existing algorithms and proposes a much simpler and effective approach to identify wrong assembly.

2. Methods

2.1 *Data*

SRA (NCBI) (Shumway *et al.* 2010) is a major freely available resource for short reads data. The genome specific reads were collected for three organisms, namely, *Homo sapiens* (GRCh37), *Drosophila. melanogaster* (BDGP5) and *Staphylococcus aureus* (*S. aureus subsp. aureus USA300_FPR3757*) (Adams *et al.* 2000; Lander *et al.* 2001; MacCallum *et al.* 2009). For these, high quality genome sequence is available along with the re-sequencing data at SRA. Runs selected for study consisted of paired end reads from Illumina GAIIx, Illumina HiSeq and Illumina MiSeq. The details related to each sample were represented in supplementary file 1. The genomic sequence of *H. sapiens* (GRCh37: Release 70), *D. melanogaster* (BDGP5:Release 70) and *S. aureus* were downloaded from Ensembl database (Flicek *et al.* 2013; Kersey *et al.* 2014). Repeat annotations for *H. sapiens* and *D. melanogaster* were downloaded from (*ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/*

chromOut.tar.gz) and (*ftp://hgdownload.cse.ucsc.edu/goldenPath/dm3/bigZips/chromOut.tar.gz*), respectively.

2.2 *Data cleaning*

Before using the raw sequences generated by the sequencing machines, the reads were checked for quality and cleaned for any adapter contamination and low quality base calls. Here, filteR (Gahlan *et al.* 2012) was used for quality assessment and filtering of Illumina generated reads (supplementary file 1). The low quality reads were discarded (quality score less than 25 for 70% of bases with all other parameters set to default).

2.3 *De novo sequence assembly*

De novo assembler namely SOAPdenovo2 was used for Illumina generated reads and is an open source assembler (Luo *et al.* 2012). Multiple assemblies were obtained using different parameters (supplementary file 2). Besides SOAPdenovo2, five other different assemblers namely, ABySS (Simpson *et al.* 2009), JR-Assembler (Chu *et al.* 2013), SGA (Simpson and Durbin 2012), Ray (Boisvert *et al.* 2010) and Velvet (Zerbino and Birney 2008), were used to illustrate the effect of assembly assessment approach on the output from different assemblers. The reads were mapped back to the assembled sequences using Bowtie (Langmead *et al.* 2009) while allowing mapping to all mappable regions. The sequences with at least five pairs of reads mapping were used further to get confidence for the assembled regions.

2.4 *Sequence-similarity-based classification*

The resulting assembled contigs (long or short) were similarity searched with the reference sequence. The BLASTN (Camacho *et al.* 2009) based similarity search was carried out at an E-value of 1e−05 with low complexity filter being off while all other parameters were set to default. The assembled sequences were searched for their closest regions across the reference genome. The correctly assembled sequences were primarily assembled contigs which aligned to the reference genome sequence for the entire continuous length without any sort of break, fragmentation or with different arrangement. The sequences, although aligned to reference sequence but with observed fragmentation or with different arrangement, were considered as mis-assembled sequences. Also, the assembled sequences with continuous homology along the length but lacking homology at the trailing ends (≥30 bases) were classified as mis-assembled contig. The fragmented sequences were further classified for orientation and overlapping/non-overlapping regions. Figure 1 illustrates the classification scheme of correctly assembled (positive set) and incorrectly assembled sequences (negative set).

The positive and negative sets were also checked for their association with repeat content. RepeatMasker was used to predict repeats in *de novo* assembled sequences (*http://www.repeatmasker.org*). The default parameters of RepeatMasker (version open-4.0.5) were used for the identification of repeats with RMBLASTN (version 2.2.27+) as the search engine, and RepBase as the data source (*http://www.repeatmasker.org; Jurka et al.* 2005).

2.5 *Features calculation*

The assessment of any assembly considers frequently used traditional parameters like coverage, N50 value, total number of assembled transcripts, average length, percent of transcripts greater than 1 kb length, total number of bases covered and many other parameters. Most of these primarily emphasize on only the assembly size which may be misleading on several occasions. Among them, coverage is the most fundamental influencing factor, emerging directly from reads arrangement. In the present study, the combination of three coverage derived features: normalized coverage, coverage ratio between subsequent positions, and local maxima of coverage, were used to distinguish between the correctly assembled and mis-assembled sequences. It was assumed that for a correctly assembled contig, the coverage value of each point should display certain degree of uniformity, while mis-assembled sequences would display a higher degree of fluctuation for these features across the point of mis-assembly. The reads with good quality score and without sequencing contaminants were considered to calculate the depth of coverage per base for the *de novo* assembled sequences. These short reads were utilized by the assemblers to form a continuous stretch of sequence. The calculation of features was done on length wise classified sets of contigs. The coverage value for each base position was normalized with average contig's coverage and length of the contig to get value in the form of normalized coverage. The values for features; coverage ratio and local maximas, were calculated using these normalized coverage values. The expressions for these features are represented below.

A. **Normalized coverage:** The normalized coverage at each base position i for each contig of length l, was calculated using the following formula:

$$NorCov(i) = \frac{c_i}{AvgCov \times l \times Cov_{exp}}$$

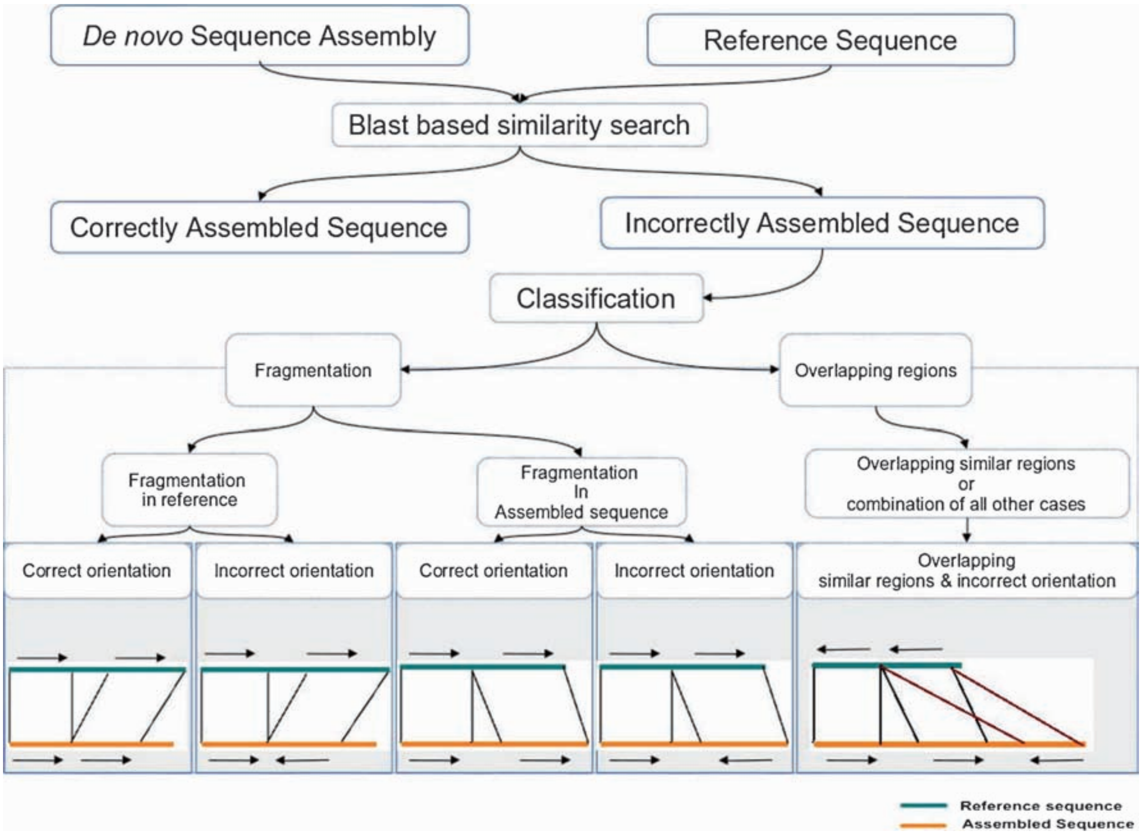

De novo Sequence Assembly
Reference Sequence
Blast based similarity search
Correctly Assembled Sequence
Incorrectly Assembled Sequence
Classification
Fragmentation
Overlapping regions
Fragmentation in reference
Fragmentation In Assembled sequence
Overlapping similar regions or combination of all other cases
Correct orientation
Incorrect orientation
Correct orientation
Incorrect orientation
Overlapping similar regions & incorrect orientation
Reference sequence
Assembled Sequence

Figure 1. Schematic representation of the assembled sequences identification. The contigs were identified as correctly assembled (positive set) and incorrectly assembled sequences/mis-assembled (negative set). Arrow lines represent the orientation direction. Orange lines represent the *de novo* assembled sequence and green lines represent the reference sequence.

where *AvgCov* is the average coverage of the contig, calculated by the following formula:

$$AvgCov = \frac{\sum_{i=1}^{l} c_i}{l},$$

where *i* is position of each base of the contig, *l* is the length of the contig, c_i is the coverage at each base position of the contig, calculated by following formula:

$$c_i = \sum read\ passed\ through\ the\ i^{th} base\ on\ the\ contig,$$

Cov_{exp} is the experimental coverage, i.e. the average coverage of all the assembled contigs.

B. **Coverage ratio:** For normalized coverage function, *NorCov*, at any position *i*, where *i* = 2 *to l*, Coverage ratio is given by $Cov_{ratio} = \frac{NorCov(i)}{NorCov(i-1)}$

If $NorCov(i) = 0$ and $NorCov(i-1) = 0$, then coverage ratio will be $Cov_{ratio} = 0$.

If $NorCov(i) = 0$ and $NorCov(i-1) \neq 0$, then coverage ratio will be $Cov_{ratio} = 1$,

else $Cov_{ratio} > 0$, means any non- zero positive value.

C. **Local maximas**: For coverage function *NorCov* at any position *i*, if *NorCov*(*i*) is the coverage value at current position under consideration, *NorCov*(*i* - 1) is the coverage value at previous position and *NorCov*(*i* + 1) is the coverage value at next position, then the presence of local maxima $Maxima_{cov}$ is reported if the following conditions (a and b) are satisfied:

a) $NorCov(i) > NorCov(i-1)$

b) $NorCov(i) \gg NorCov(i+1)$

So,

$$Maxima_{cov}(i) = \left\{ \begin{array}{ll} 1 & \text{if the condition above is satisfied} \\ 0 & \text{otherwise} \end{array} \right\}$$

Further, $Maxima_{cov}$ is normalized by experimental coverage Cov_{exp}. Therefore, $NorMaxima_{cov}(i) = \frac{\mathrm{Maxima}_{cov}(i)}{Cov_{exp}}$

2.6 *Performance evaluation*

For each study, the performance of Amosvalidate and REAPR based assembly analysis was evaluated using Sensitivity (Sn), Specificity (Sp), Accuracy (Ac), Matthew's correlation coefficient (MCC).

$$Sn = \frac{TP}{TP + FN} * 100$$
$$Sp = \frac{TN}{TN + FP} * 100$$
$$Ac = \frac{TN + TP}{TN + FP + TP + FN} * 100$$
$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN) * (TP + FP) * (TN + FP) * (TN + FN)}}$$

Here TP and TN are true positive and true negative instances, respectively. Similarly, FP and FN are falsely predicted positive and falsely predicted negative instances, respectively.

Amosvalidate (Phillippy *et al.* 2008) and REAPR (Hunt *et al.* 2013) use their own set of features to determine the erroneous region/contig. The sequences reported with even a single error were selected as mis-assembled one and the remaining were considered as the correctly assembled contigs. TP, TN, FP and FN were categorized while finding the commonality between the similarity based positive/similarity based negative (separately) and Amosvalidate/REAPR-mis-assembled, Amosvalidate/REAPR-correctly assembled contigs (separately). TP is the number of correctly assembled contigs observed for Amosvalidate/REAPR, FP is the number of contigs mis-classified as correctly assembled contigs observed for Amosvaliate/REAPR, TN is the number of mis-assembled contigs observed for Amosvalidate/REAPR and FN is the number of contigs mis-classified as mis-assembled contigs observed for Amosvalidate/REAPR.

2.7 *Unsupervised-learning-based classification and error rate estimation*

Amosvalidate is one of the methods reported analyzing the assembled sequences using 12 features, mainly relying upon mate pair information, depth of high- and low-coverage regions. REAPR is another recently reported tool with two principle aims: to score every base for accuracy and to automatically pinpoint the mis-assembled contigs, while deriving information from fragment coverage. Error calls by REAPR is done either for a lack of coverage or presence of irregular fragment coverage. Error rate (described below) based comparative analysis was carried out to evaluate Amosvalidate, REAPR and the presented approach.

For the contigs classified as correctly assembled and mis-assembled by similarity report for various lengths, error rate calculation was carried out for unsupervised-learning-based clustered contigs on the above mentioned features. Unsupervised learning was carried out by *K*-means clustering (Hartigan and Wong 1979). *K*-means clustering is numeric, non-deterministic, iterative and unsupervised method of partitioning the 'n' data points into 'k' disjoint clusters. The implementation of clustering methods was done using R, where 1000 rounds of iterations were implemented. The contigs' particular lengths with frequency of at least five, occurring in both categories, correctly assembled and mis-assembled, were only considered for the further study. Two disjoint clusters were obtained, while dividing the data into four subsets. TP, TN, FP and FN were calculated starting with subset consisting maximum number of instances (figure 2).

Error calculation after clustering with *K*-means clustering method for every calculated feature is illustrated in figures 2 and 3 which was calculated using the following formula:

$$clusterr_l = \frac{numFP_l + numFN}{numTP_l + numTN_l + numFP + numFN_l}$$

where $clusterr_l$ is the error of clustering associated with contig length l,

$numFP_l$ is the number of false positive contigs in cluster,
$numFN_l$ is the number of false negative contigs in cluster,
$numTP_l$ is the number of true positive contigs in cluster and
$numTN_l$ is the number of true negative contigs in cluster.

Error rate calculations for Amosvalidate and REAPR is the fraction of contigs that are mis-classified. The flow diagram (figure 3) represents the basic schema of lengthwise selection and error rate calculation.

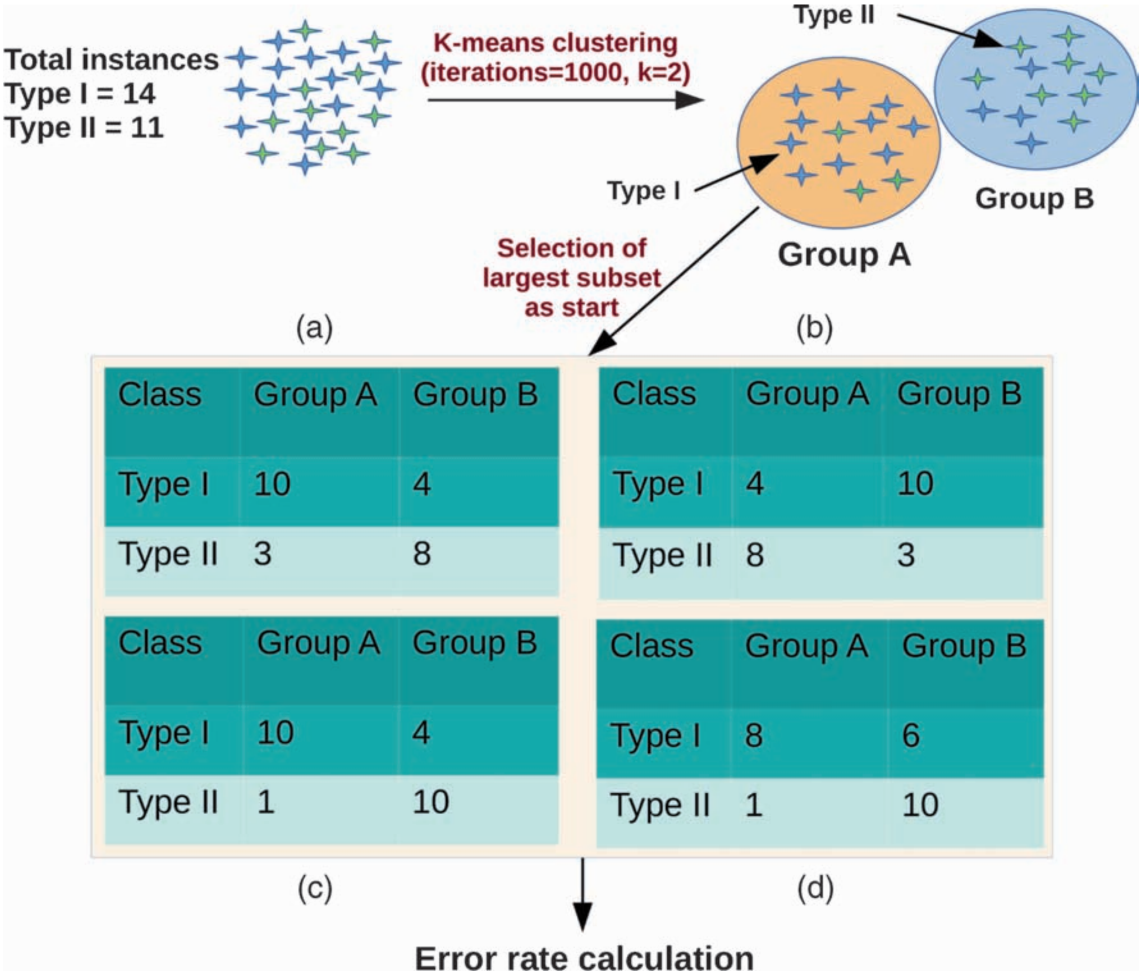
The error rate was calculated using the frequency of occurrences for FN, TN, FP and TP of common length contigs. For each length '*l*', formula for error rate is given below:

$$Error_l = \frac{freqFN_l + freqFP_l}{freqTP_l + freqTN_l + freqFP_l + freqFN_l}$$

where $freqFP_l$ is the frequency of false positive contigs with length l,

$freqFN_l$ is the frequency of false negative contigs with length l; $freqTP_l$ is the frequency of true positive contigs with length l, and $freqTN_l$ is the frequency of true negative contigs with length l.

Further, comparative plotting and analysis was carried out for error values calculated for each length group considering the individual outputs from Amosvalidate, REAPR and



Class	Group A	Group B
Type I	10	4
Type II	3	8

Class	Group A	Group B
Type I	4	10
Type II	8	3

Class	Group A	Group B
Type I	10	4
Type II	1	10

Class	Group A	Group B
Type I	8	6
Type II	1	10

Figure 2. Schematic representation of the clustering approach. *K*-means clustering applied on calculated features to group instances from Type I (blue colour) and Type II (green colour) categories into two different groups (A, B). Ideally, Group A and Group B should include Type I and Type II instances, respectively. Presence of another type of instances in one group calls for error. Group assignment starts with the largest subset consisting of maximum instance of single type (represented with (**a**), (**b**), (**c**) and (**d**)). From (**a**), (**b**), (**c**) and (**d**), only one possibility as output. In (**a**), Group A is of Type I with maximum number of instances from Type I and other as Group B with maximum of Type II instances. Similarly, in the case of (**d**) and conversely in (**b**). In (**c**) with equal highest number of Type I and Type II instances in the two groups, calls for random assignment of Type I as Group A and Type II as Group B, or assignment of Type I as Group B and Type II as Group A.

unsupervised learning *K*-means clustering method based our approach, separately.

3. Results and discussions

De novo sequence assembling appears to be more prone towards mis-assembling while constructing the contigs from small reads without taking any guide or reference support. In the present study, a thorough analysis of various mis-assembly detection approaches has been done, including a novel approach proposed in this study.

In order to understand more about the behavior of assembly, *H. sapiens* and *D. melanogaster* short reads based *denovo* assembly was carried out using SOAPdenovo2 (supplementary file 2). The assemblies from each dataset were evaluated for the best possible ones based on the corresponding assembly statistics. For *H. sapiens,* three different datasets were used in order to get the details for the impact of varying read lengths and fragment sizes over assembly (table 1). The proposed mis-assembly detection strategy presented here depends on the features derived after mapping back of reads to the assembled contigs. For short contigs with correct assembly, it was observed that there was

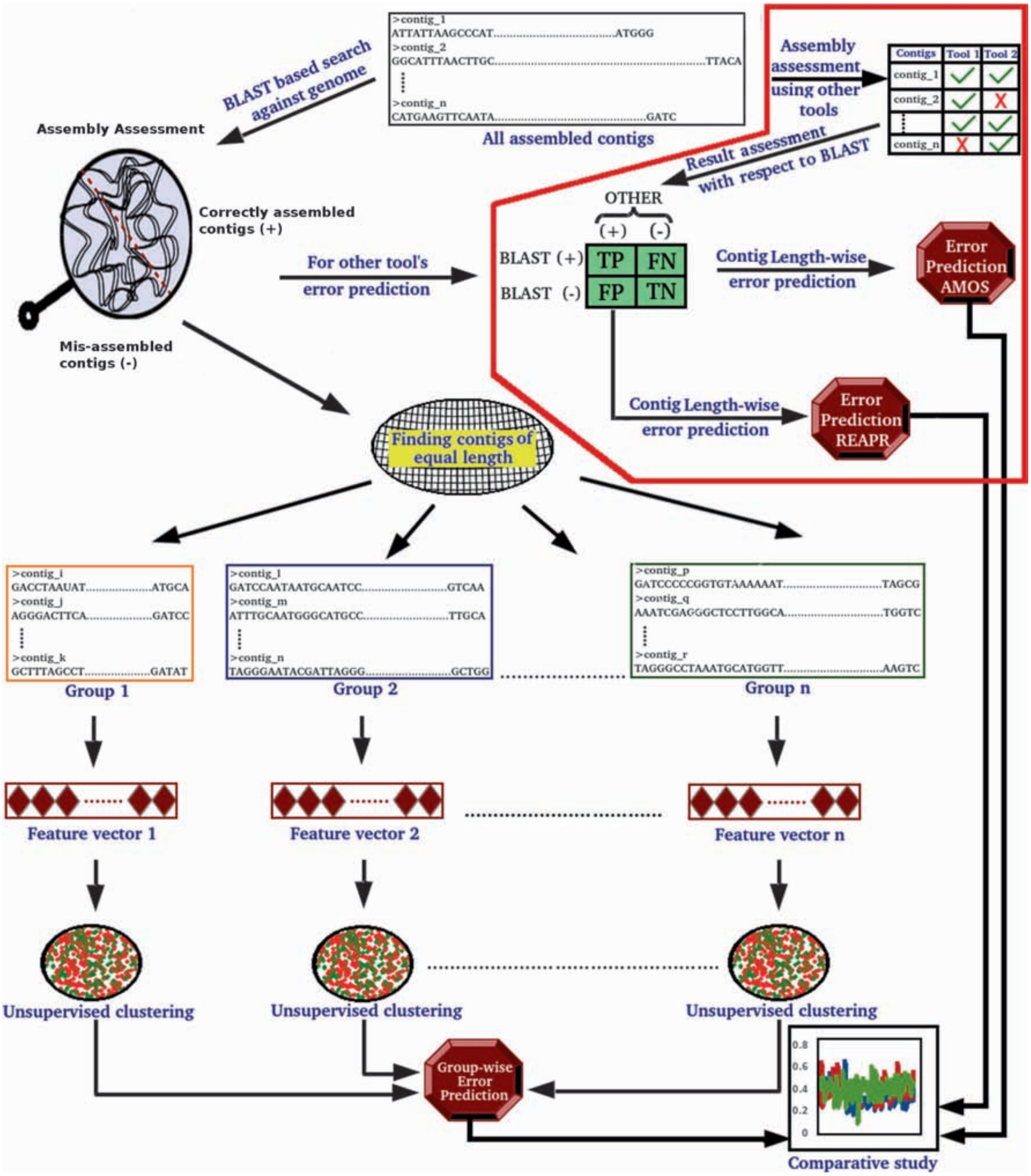


Figure 3. Schematic representation of lengthwise selection and error fraction calculation. The error fraction was calculated after *K*-means clustering, taking three features: coverage, coverage-ratio and normalized local maximas together. Clustering and error fraction calculation were done on different lengthwise groups, represented by Group 1, Group 2 and up to Group n.

some symmetric distribution of the coverage about the mid position of given contig, although for longer contigs, there was fluctuation in the coverage distribution. It was observed that each contig carried features specific to its length, encouraging the requirement to check the contiguous sequence properties in length specific manner.

3.1 *Reads quality*

Errors in initial reads may be one of the causes for errors in the formed assembly, resulting in much more fragmented assembly (Schatz *et al.* 2012). Different large-scale sequencing projects may produce sequences at similar rates and costs

Table 1. Assembly statistics for *de novo* assembled sets

Sr. no	SRA no.	Selected *K*-mer	Average length (bases)	Maximum length (bases)	Total contigs	% contigs ≥1000 bases	Coverage	Total bp	N50	Average GC	% of Reads mapped	% Correct contigs	% no BLAST homolog
.1	SRR892664	111	535.62	12925	46785	12.92	64.14	25058993	802	50.43	51.11	47.08	7.34
2	ERP000362	127	589.66	299358	19118	5.36	26.50	11273166	2098	47.17	17.89	82.77	1.68
3	SRR630877	73	383.01	35551	68954	3.62	6.68	26410744	377	37.10	9.52	98.63	0.00
4	SRR900425	25	1407.11	36900	82378	36.64	9.02	115915142	3468	41.81	53.82	68.42	0.00
5	SRR022868	33	5429.52	93312	525	36.00	231.26	2850498	27268	32.63	88.73	92.00	0.00

SOAPdenovo2 based *de novo* genome assembly performed for four datasets: 1. SRR892664, 2. ERP000362, 3. SRR630877 of *H. sapiens* and 4. SRR900425 of *D. melanogaster*.

but with significantly different error rates. Base-calling algorithms, like Bustard by Illumina, provide per-base phred-like quality scores as by-product (Ewing and Green 1998; Ewing *et al.* 1998). Error patterns associated with different Illumina platforms have been well characterized by some groups (Dohm *et al.* 2008; Hansen *et al.* 2010). The quality score is given by the formula -10 $\log_{10}$(P). A quality score of 30 assigned to a base gives the chance that the base is called incorrectly once in 1000 occurrences *i.e.* base call accuracy is 99.9%. Huse *et al*. found that sequences from massively parallel pyrosequencing on the Roche GS20 system with an average score below 25 had more errors than those with higher averages (Huse *et al.* 2007). In the present study, the data set (SRR630877) had low quality at 3′-end (quality > 25), requiring selective trimming. About 48% to 79% reads passed the quality as well as adapter filtration step from each reads' set (after trimming, if required) (supplementary file 1). The read length ranged from 32 to 250 bases for Illumina generated reads. Reads pre-processing was the major problem and time consuming part in order to get a comparable assembly. The processed SRA downloaded sets with average quality score distribution along the length are shown in supplementary file 3.

A further investigation appears to be necessary to determine when and how the assembly tools are differentially affected by varying depths of coverage, sequencing errors, and length of the sequenced reads.

3.2 *Classification of assembled sequences*

The assembled sequences were classified as correctly assembled or mis-assembled on the basis of sequence similarity and agreement with the reference genomic sequences. According to the classification scheme described in figure 1, the mis-assembled sequences belonged to (1) fragmentation without overlaps, (2) fragmentation with overlaps or (3) involvement of combination of these factors in the assembled contig. Fragmentation occurs when a *de novo* assembled sequence matches with a reference sequence in parts instead of full length continuous match. Fragmentation could be a result of wrong *K*-mer involvement during the assembly process, resulting into no match for certain internal regions. Several times these fragmented similar regions had different orientations with respect to the reference sequence segments (figure 1). Also, many fragments overlapped with each other while had similarity to different segments of reference sequence. Such case could be due to similar reads generated from different regions of the genome, causing confusion to the assembler with many paths. Another case is different segments of a contig showing similarity to overlapping segments of the reference sequence fragment. This could be due to missing reads or wrong connectivity of *K*-mers. The correctly assembled sequences match with reference sequence segment along its full length, and for several instances the *de novo* assembled full stretch of contig showed similarity to many segments of reference sequence. This suggested that assemblers were able to correctly construct the fragment. However, availability of many paths for further extension, confine the contig length progress to some particular length. Also, a few contigs were found to have no similarity for its ends, while had single contiguous similar region in between. The contigs with >=30 bases length with no homology at either end or both the ends were classified as mis-assembled. Usually, in two connecting *K*-mers, the first *K*-mer works as start node and the second *K*-mer works as the end node. If no further connecting *K*-mer is available to the end node (to work it as start node for next connection), it works as sinking node. This caused termination of further contig extension due to no availability of connecting *K*-mer. The percent of *de novo* assembled contigs generated by SOAPdenovo2 assembler for the four selected datasets categorized as correctly assembled are listed in table 1.

3.3 *Assembly statistics*

For *H. sapiens* and *D. melanogaster* genome, *de novo* assembly of Illumina generated short reads (listed in table 1) was carried out using SOAPdenovo2 assembler. *K*-mer-wise *de novo* assembly was carried out with varying values starting with minimum 21 and maximum value up to the longest possible *K*-mer for the given read length. The primary assembly was highly fragmented (supplementary file 2). Best *K*-mer selection is usually done by calculating various assembly statistics like average length of assembled contigs, maximum length from the contigs, total assembled contigs, percentage of contigs ≥1000 bases, total bases from the assembly, N50 value, percentage of reads mapping back to the assembled contigs. These statistics were considered to describe the contiguity, consistency and accuracy of the assembled genome. Contiguity of the genome is mostly assessed by N50 value. N50 contig size of N means that 50% of the assembled bases are contained in the contigs of length N or larger. However, recently published assemblers competitions have shown that N50 values rarely correlate with the actual quality of the assembly (Earl *et al.* 2011; Salzberg *et al.* 2012). Assemblathon defined its own metrics such as NG_{50} (computed using average lengths of haplotypes, instead of the contig lengths used by N50), $CPNG_{50}$/$SPNG_{50}$ (it is the average length of the contigs/scaffolds consistent with haplotype sequence) and CC_{50} (it gives the general idea of correct contiguity between the points in the assembled genomes). GAGE (Salzberg *et al.* 2012) used the E-size metric, which is the expected length of a contig/scaffold that contains a randomly selected base from a reference genome. The coverage of sequences and paired-end information is another mostly used statistics to check the consistency of the assembly. According to Phillippy *et al.,* the regions with sudden unusual high depth of read coverage or unusual low depth of read coverage gives the indication of collapse of repeats or incorrect joining between the unrelated genomic regions, respectively (Phillippy *et al.* 2008). Paired-end information is used to check for expansion or contraction between the read pair (CE statistics), utilizing the spatial information for assembling validity. However, all these parameters still require refinement to get confidence over the quality of *de novo* assembled primary contigs. Dependence upon any single metric could be misleading to decide a reliable assembly from the different *K*-mers results. An array of various sequencing and assembling factors need to be assessed to clearly understand the difficulties in getting a reliable *de novo* assembled contig from short read sequences.

Eukaryotic genomes are repeat rich, and these repetitive regions pose big challenge in assembling the reads. The human genome has ~50% of the total genome as the repetitive DNA while *D. melanogaster* has ~24% (Treangen and Salzberg 2012; Manning *et al.* 1975), making them complex genomes to be handled by *de novo* assemblers. The overall result for the *de novo* assembly showed varying statistics values at varying *K*-mers (supplementary file 2).

H. sapiens: For Illumina HiSeq 2000 (SRR892664) generated dataset with 150 bases read length and 300 bases average insert siz*e, de novo* assembly resulted into 46,785 as total number of contigs with 64x coverage depth. Average contig length observed was 535 bases along with N50 size of 802 bases, reflecting the formation of highly fragmented assembly. However, comparison with the reference genome resulted into ~47.08% correctly assembled contigs (table 1). For Illumina MiSeq (ERP000362) dataset with read length of 250 bases and average insert size of 436 bases, SOAPdenovo2 resulted into 19,118 total assembled contigs along with 299.358Kb maximum contig length, N50 value of 2,098 bases with 26x coverage, and 589 bases average length for the contigs. On comparison with the reference genome, a total of 83% contigs were observed as correctly assembled. For Illumina GA IIx (SRR630877) *H. sapiens* dataset with read length of 91 bases and average insert size of 399 bases resulted into 68,954 contigs (~6x coverage), ~98.63% of which were observed as correctly assembled. The total assembled contigs had maximum length of 35,551 bases with average length of 353 bases. Only 9.52% of total reads mapped back to Illumina GAIIx assembled contigs, while 17.89% and 51.11% of total reads mapped back for Illumina MiSeq and Illumina HiSeq datasets, respectively (table 1). The average read quality per base was >25 for each of the above datasets. The percent of total reads mapped back on assembled contigs was very less in comparison to reference sequence, suggesting higher nucleotide level errors in assembled sequences. On the reference set, 47.35%, 42.71% and 38.89% of total reads mapped back from

Illumina GAIIx, Illumina MiSeq and Illumina HiSeq read sets, respectively. It was interesting to note that there was relative increase in mis-assembled contigs in comparison to correctly assembled contigs with increase in the contig length (figure 4). 46.91% of mis-assembled contigs belonged to the length range 100-500 bases for Illumina HiSeq 2000 (SRR892664) which raised significantly (upto 92.59%) in the length range >5000 bases. Similar trend was observed for Illumina MiSeq (ERP000362) and Illumina GA IIx (SRR900425) data derived contigs, displaying mis-assembly ranging between 15-75% and 1-25%, respectively (figure 4).

Usually, for genome level assembly a combination of mixed libraries (of varying insert sizes) are used. To assess the effect of mixed libraries on initial assembly, a mix of above datasets was used, making four combinations as set A (Illumina HiSeq 2000 and Illumina MiSeq), set B (Illumina HiSeq 2000 and Illumina GAIIx), set C (Illumina GAIIx and Illumina MiSeq) and set D (Illumina HiSeq 2000, Illumina GA IIx and Illumina MiSeq), including variable libraries in the pool (supplementary file 2). The assembly was done to the maximum possible *K*-mer length according to the shortest read from the set to utilize each of the library for each *K*-mer. Best *K*-mer based assembly selection was done using basic assembly statistics mentioned above. For Set A, *K*-mer 127 stood as the best one with 34,085 total assembled contigs along with 299.358 Kb maximum length, N50 value of 992 bases with 79x coverage, and average contigs length of 654 bases. In comparison, though Illumina HiSeq 2000 and Illumina MiSeq datasets had lesser average length of 535 and 589 bases, they displayed 12.925 Kb and 299.358 Kb as maximum length, with 802 bases and 2098 bases as N50 value, respectively. This suggested improvement with Illumina HiSeq 2000 and Illumina MiSeq based *de novo* assembly. For Set B, *K*-mer 89 stood best with total 87,533 sequences, 453 bases as average sequence length, N50 value of 738, and maximum sequence length of 103.94 Kb. All these best assembly parameters showed this combined assembly as downfall from the individual datasets' assemblies. For Set C, 49 *K*-mer was selected, which showed decline in the average contig length to 316 bases from 383 bases (Illumina GA IIx) and 589 bases (Illumina MiSeq). Similarly, decline was observed in the maximum length to 18.098 Kb and coverage to 1.92 from 35.551 Kb and 299.358 Kb, with coverage 3.62x and 5.36x for Illumina GA IIx and Illumina MiSeq derived data, respectively. For Set D, when all the three human datasets were combined, there was a mixed response in the summary statistics. While the average sequence length decreased to 313 bases in comparison to Sets A, B and s C specific assemblies, total assembled sequences increased to 206,408 in comparison to Set A and B, maximum length decreased to 102.265 Kb in comparison to Set A. There was also decreased N50 value to 430. The assembly statistics failed to explain the best assembly and how the involvement of varying insert-lengths had improved/declined the assembly quality, encouraging us to go for assembly validation method.

D. melanogaster: On the *de novo* assembly using Illumina GAIIx dataset with read length of 76 bases and average insert size of 322 bases, using SOAPdenovo2 assembler, a total of 82,378 assembled sequences were obtained with average length of 1,407 bases and maximum length of 36,900 bases. The assembly N50 value was 3,468 bases for the contig level assembly with 9X average coverage. A total of 68.42% of assembled contigs were observed as correctly assembled (table 1). 53.82% of total reads mapped back across the *de novo* assembled contigs, while 69.42% reads mapped back across the drosophila reference sequence.

On comparing the genome assembly output of the two organisms with varying genome complexity and genome size, the basic assembly statistics performed better for *D. melanogaster* (table 1). Basic assembly statistics suggesting lags in handling genome with higher content of repeats. The percent of correctly assembled contigs in case of *D. melanogaster* was lesser (68.42%), while in the human dataset (Illumina GA IIx) 98.63% correctly assembled contigs were observed, however, with much fragmented short length assembly. All this suggesting again to go for a better method for assembly quality assessment.

The initially quality filtered reads used for assembly process were mapped back to the reference as well as assembled contigs sets obtained for each dataset. Many of the reads were found mapping across multiple regions. Multi location mapping reads may contribute to ambiguous assembling and suggest repetitiveness. Most of eukaryotic genomes contain a huge amount of repetitive regions, mostly longer than the reads obtained through some next generation sequencing process. As a consequence, several reads map to multiple locations across the genome which are either discarded from the study or cause poor assembling. For the selected datasets, contigs were checked for the reads mapping across multiple regions for correctly assembled and mis-assembled sequences. They were compared with multi-mapping regions in the reference sequences.

The reads datasets had varying percentage of multi-mapped reads for *de novo* and reference sequences. SRR892664 (*H. sapiens*, Illumina HiSeq 2000) had ~14.89% of multi-mapping reads, while SRR630877 (*H. sapiens*, Illumina GA IIx) and ERP000362 (*H. sapiens*, Illumina MiSeq) had 1.15% and 5.56% multi-mapping reads on the assembled contigs, respectively. Of these total multi-mapping reads, *H. sapiens'* Illumina HiSeq 2000 had 66.74% of multi-mapping reads for mis-assembled contigs, 24.79% on correctly assembled contigs, and 8.46% on regions belonging to both correctly and mis-assembled contigs (named as common) (figure 5). *H. sapiens'* Illumina MiSeq

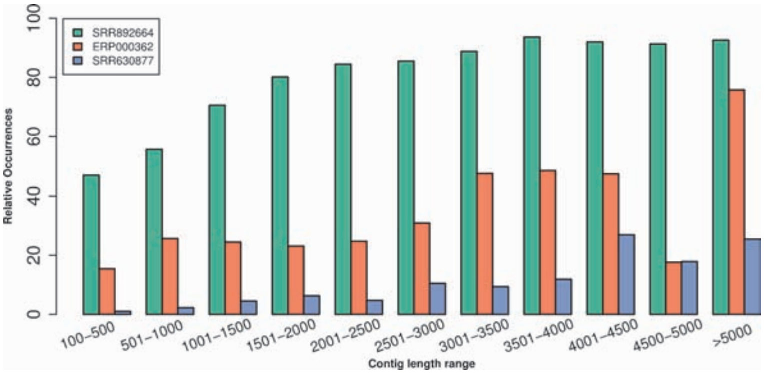


Figure 4. Comparative bar-plot illustrating the relation between mis-assembly and contig length. The bar-plot represents three human datasets, showing the percent of mis-assembled contigs (Y-axis) in different length ranges (X-axis).

and *H. sapiens'* Illumina GA IIx, had 94.23% and 0.04% of total multi-mapping reads on mis-assembled contigs and 4.07% and 99.30% on correctly assembled contigs, respectively (figure 5). Illumina GA IIx derived data exhibited an opposite trend where 99.3% of multi-mapping reads belonged to correctly assembled contigs. The reason seems to be the fact that only 1.15% reads were of multi-mapping type and majority of assembled contigs were of short length (N50= 377) (table 1). As already mentioned above, the wrong assemblies showed increasing trend with length and were very less for short contigs. *H. sapiens'* Illumina HiSeq 2000 data showed very interesting outcome as it had the least number of correctly assembled contigs (47.08%) while the highest number of multi-mapped reads (14.89%) on *de novo* assembly. *D. melanogaster de novo* assembly lacked multi-mapping reads in total, although on mapping reads to reference sequence, many such instances were observed (3.90%), which appears to be an outcome of selective avoiding of multi-mapping reads during the process of assembling.

Reads mapping at multiple locations is suggestive of repetitiveness, a challenge whose one of the major impacts on assembly quality and assembler's performance has reflected above. Considering the involvement of multi-mapping reads to mis-assembly, non-parametric method: Chi-squared test was applied to test if the extent of multi-mapping reads caused mis-assembly. The null hypothesis (H0) stated that the multi-mapping reads do not exhibit any selective preference for association with the mis-assembled sequences. On statistical testing, the Chi-squared values obtained were 745.39 (p-value≤2.2e−16), 54.85 (p-value=1.301e−13) and 52.862 (p-value=3.579e−13) for Illumina MiSeq, Illumina GA IIx and Illumina HiSeq 2000 datasets, respectively. The obtained p-values were highly significant and chi-square values were greater than the chi-square-tabulated (3.841), rejecting the null hypothesis and accepting the alternate hypothesis that there is a definite association between wrong assembling and multi-mapping reads. It was observed that there were 17.51%, 92.38% and 55.04% total mis-assembled contigs having repeats associated with them, for Illumina HiSeq 2000, Illumina GA IIx and Illumina MiSeq datasets, respectively. Of these, 6.68%, 99.49% and 23.74% contigs had multimapping reads association, respectively. Therefore, the presence of complex repeats may increase the rate of false assemblies. Further, in order to highlight this in simple way, assembly of a repeat scarce genome of *S. aureus* (0.81%) was compared against the repeats rich human genome. The *S. aureus de novo* assembly resulted into 92% correctly assembled contigs with N50 value of 27,268 and average length of 5,429 bases. Further, the degree of multimapping reads was found much lesser than human. Of the total mapped reads to *S. aureus* genome, only 0.012% were multi-mapping reads and these reads mapped to four mis-assembled contigs and 89 correctly assembled contigs. In overall, this emerges out that the reads mapping to multiple locations might contribute significantly towards mis-assembled contig formation, and repeats share a good credit for this.

3.4 *Evaluating the existing tools for the identification of mis-assembled contigs*

The number of tools to detect such wrong assemblies is very limited and equally scarce is any benchmarking and guideline study on such tools, putting it very clearly that the area needs urgent attention and work. The tools like CGAL (Rahman and Pachter 2013), ALE (Clark *et al.* 2013) and FRCbam (Vezzi *et al.* 2012) report any specific assembly as a whole as the better one while comparing multiple assemblies, either using likelihood score (in CGAL and ALE) or feature response curve (in FRCbam). All these tools lack the ability to report the correct assembly scrutiny for every contig. However, tools like Amosvalidate (Phillippy *et al.* 2008) and REAPR (Hunt *et al.* 2013) report error information for every contig/scaffold without needing any reference. Amosvalidate is one of the methods which is reported to analyze the assembled sequences using 12 features, relying mainly upon the mate pair information, distribution of *K*-mers, depth of high and low coverage regions to identify structurally suspicious regions of the assembly. It generates an output report consisting of regions of mis-assembly on the basis of evaluated features. According to Amosvalidate, almost all the mis-assemblies are caused by repeats. In this work too it was found that the repeats were significantly associated with wrongly assembled contigs. Some parameters of Amosvalidate to find the wrongly assembled contigs are discussed here. *CE_COMPRESS*: it represents the region of mis-assembly with mates consistently closer than expected at a given position as would occur in collapsed repeat (i.e. in which two or more identical copies are wrongly assembled/merged into single representation) or excision from the assembly. A maximum of 4.08% of such error was observed in SRR892664 dataset based *de novo* assembly of *H. sapiens* (figure 6a). Other assemblies from SRR630877, ERP000362 and SRR900425 datasets had 1.4%, 3.53% and nil CE_COMPRESS type error, respectively (figure 6a). *CE_STRETCH*: it represents the regions of mis-assembly with repeat copy number expansion or other insertion event, if the inserts are consistently larger than expected. Amosvalidate pointed ~37.25% CE_STRETCH type errors in SRR900425 based *de novo* assembly. In SRR892664, 7.90% CE_STRETCH type error was noticed. On insert-size distribution analysis it was observed that maximum of the mapped reads were with shorter insert size range along with overlapping reads. Insert size distribution plot (supplementary file 4) did not support CE-statistics based CE-STRETCH. *High_SNP*: it represents region with high coverage, where most of the reads exhibit a particular base while several other reads have another base as correlated SNPs at particular base position. In polyploid genomes, these could be the indicators of collapsed repeat if frequency is higher (Phillippy *et al.* 2008). Amosvalidate reports all regions with at least two columns within at most 500 bases of each other. *H. sapiens de novo* assembly was predicted to have 8.28% to 24.33% regions of mis-assembly due to High_SNP for three datasets (figure 6a). SRR900425 based *de novo* assembly resulted into 6.91% regions of mis-assembly as High_SNP type. *High_read_cov*: This parameter represents the regions where the coverage is deeper than expected. It may indicate a collapsed repeat region. Amosvalidate predicted 13.05% (as maximum) for *H. sapiens* GAIIx dataset with *High_read_cov*. Other assemblies for three datasets resulted into very few mis-assemblies (~4.17-6.65%) predicted due to *High_read_cov*. Low_good_cov: This parameter covers errors due to low coverage by mate-pairs which are otherwise at right distance and orientation. The prediction error was found very high, ranging between ~49% to ~71% with highest for *H. sapiens* ERP000362 dataset. *kmer_cov*: According to Amosvalidate, almost all mis-assemblies were caused by repeats. Thus it could be useful to find the location of repeats in assembly. The parameter, *kmer_cov*, reports at least 1 kb long region covered by high frequency normalized *K*-mers *i.e.* collapsed

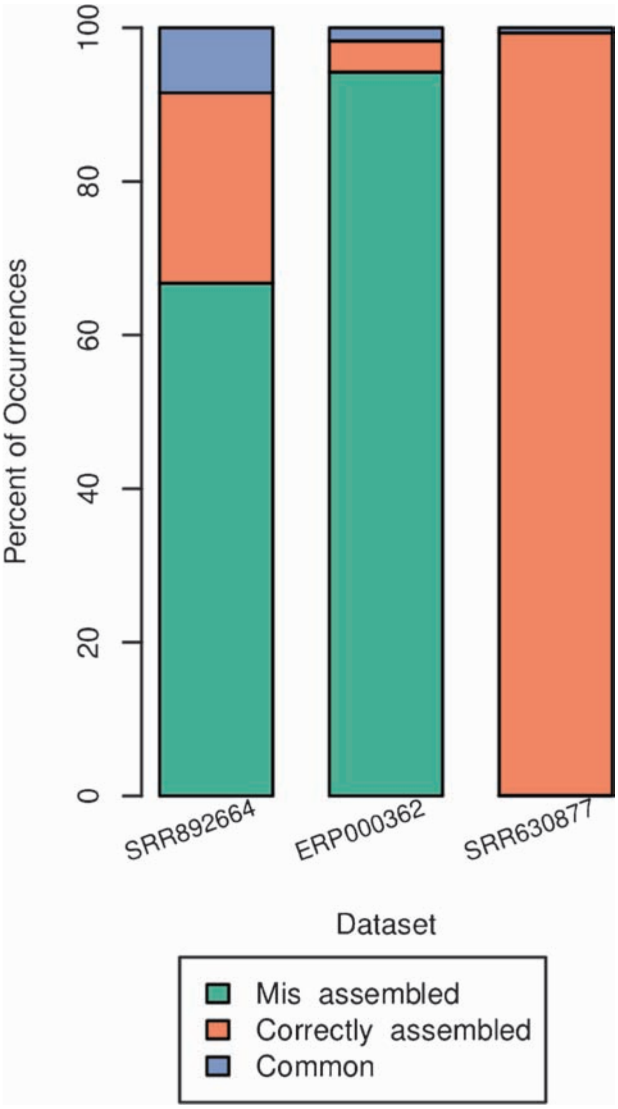


Figure 5. Multi-mapped reads distribution plot. Each of the four datasets represented on the X-axis and the percent of multi-mapped reads occurrences is represented on the Y-axis. The figure consists of the percent of multi-mapped reads on correctly assembled contigs (orange), mis-assembled contigs (green), and on both (purple) marked as Common.

repeats in the genome. Here, normalized *K*-mer is the number of times a given *K*-mer 'q' occurs in K_r (set of *K*-mers in the reads) divided by the number of 'q' occurrences in K_c (the set of *K*-mers in the contig consensus sequence), i.e. K_r/K_c. *H. sapiens* HiSeq 2000, *H. sapiens* GA IIx, *H. sapiens* MiSeq and *D. melanogaster* HiSeq 2000 were predicted with 5.87%, 7.30%, 9.18% and 2.25% of kmer_cov error, respectively (figure 6a). It was interesting to note that for each of the *de novo* assembly, on analysis for correctly assembled contigs with these features, the percent of instances for each feature remained almost the same as was observed for the total assembled contigs. This suggests that these statistics did not contribute much in the correct identification of wrongly assembled contigs, and instead a large amount of correct assemblies were found being tagged as wrongly assembled one (figure 6a).

REAPR is another recently reported tool with two major aims: to score every base for accuracy and to automatically pinpoint mis-assemblies, while deriving information from fragment coverage. A set of 11 metrics are extracted from the mapping information at each base of the genome assembly. REAPR suggests that each read must be accurately and independently mapped to its mate, so that a read pair is not artificially forced to map as a proper pair (in correct orientation and separated by the correct distance as determined by the library type). Otherwise, the sensitivity in identifying assembly errors is reduced.

Of the 11 metrics, Fragment coverage, Link, Repeat, Clip and Read orientation contributed towards finding most of the erroneous regions in the datasets. *Clip* represents a significant proportion of the reads that were soft-clipped to map to the erroneous positions. A maximum of 7.88% of the total predicted mis-assemblies in *H. sapiens* GAIIx dataset (SRR892664) were due to Clip type error on *de novo* assembly, while in SRR630877, ERP000362 and SRR900425 based *de novo* assemblies, 7.35%, 0.91% and 4.19% were predicted with Clip error, respectively (figure 6b). Fragment coverage represents region of the genome between outermost ends of a proper read pair, where *frag_cov* type error represents as low fragment coverage in this region. REAPR reported 40.25% of frag_cov type error in *de novo* assembly of SRR892664, and error of 2.88%, 6.49% and 0.185 for SRR630877, ERP000362 and SRR900425, respectively.

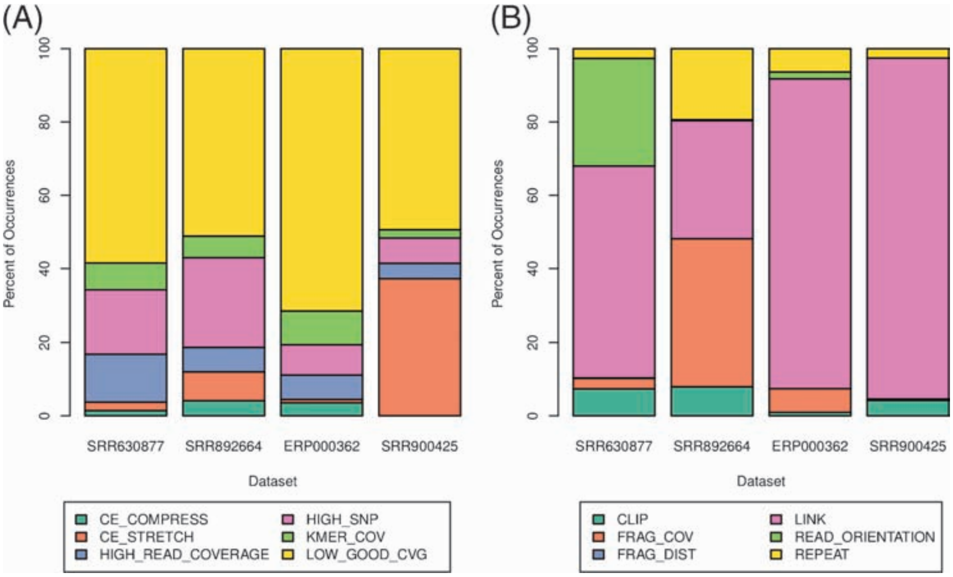


Figure 6. Feature distribution plot predicted by Amosvalidate and REAPR. Stacked bar plot representing the percent of occurrences of represented particular feature type (on Y-axis) for four datasets (on X-axis) in (**A**). Amosvalidate, and (**B**). REAPR. Amosvalidate prediction is based upon features: CE_COMPRESS (mates consistently closer than expected), CE_STRETCH (represents repeat copy number expansion or other insertion event, if the inserts are consistently larger than expected), High_read_cov (the coverage is deeper than expected), High_SNP (represents correlated SNPs where most of the reads are one base, but several other reads have other base(s)), kmer_cov (reports at least 1 kb long region covered by high frequency normalized *K*-mers i.e. collapsed repeats in the genome), Low_good_cov (low coverage by mate-pairs which are otherwise at right distance and orientation). Similarly, REAPR prediction is based upon features: Clip (a significant proportion of the reads that were soft-clipped to map to the erroneous positions), *frag_cov (*low fragment coverage distribution), Link (a significant proportion of reads in a region mapping elsewhere in the assembly), Read_orientation (wrong orientation of read pairs), Repeat (the observed coverage is more than twice the expected coverage after correcting for any GC bias present in the reads mapped to the assembly).

(figure 6b). Fragment coverage distribution (FCD) is another metric used by REAPR, which is used to calculate FCD error (*frag_dist*). FCD error represents the difference between expected and observed FCD. The cutoff in FCD error, the value above which a base is called incorrect, is automatically determined by sampling windows in the genome to determine the number of windows failing for a range of cutoff values. FCD error usually represents incorrect scaffolding, a large insertion or deletion in the assembly, or sometimes a false joining in a contig. REAPR predicted 0.093% and 0.19% frag_dist in SRR630877 and SRR900425 *de novo* assemblies, respectively (figure 6b). Link, another error type parameter, represents a significant proportion of reads in a region mapping elsewhere in the assembly. REAPR reported such errors as very high (up to 84.39%) for ERP000362 dataset in human dataset, while ~32% and ~57% in SRR892664 and SRR630877, respectively. In case of *D. melanogaster*, it was 92.85%. *Read_orientation* parameter represents wrong orientation of read pairs (i.e. pointing away from each other or in the same direction). Output for *H. sapiens*' GA IIx dataset was predicted to have 29.37% contribution for wrong reads orientation type errors, while all other datasets had very few such error values ranging between 0.016% to 1.855% (figure 6b). A region is flagged as a repeat by REAPR if the observed coverage is more than twice the expected coverage, after correcting for any GC bias present in the reads mapped to the assembly. *H. sapiens* Illumina HiSeq dataset had 19.38% of the total errors tagged as originating from repeats (figure 6b). Other datasets had a small percent of repeat type errors (2.55% to 6.34%) reported by REAPR. Other metrics like *Perfect_cov*, *frag_dist_gap*, *frag_cov_gap* etc reported by REAPR had no predicted instances as error (figure 6b). Similar pattern of inconsistency was observed for REAPR, as was observed for Amosvalidate. The correctly assembled contigs contained all the error types in similar fashion for each of the datasets as was observed previously for Amosvalidate, suggesting insignificant contribution by these errors' evaluation parameters or the way these tools worked with such parameters. Highlighting the performances of these two tools in terms of Sensitivity, Specificity and Accuracy also suggests the same as there was high imbalance between sensitivity (68.21, 84.25, 92.37and 59.93 of Amosvalidate, 78.53, 39.75, 85.06 and 9.27 of REAPR) and specificity (47.02, 34.08, 24.55 and 73.65 of Amosvalidate, 30.59, 53.10, 33.12 and 97.66 of REAPR) in *H. sapiens* SRR892664, *H. sapiens* ERP000362, *H. sapiens* SRR630877 and *D. melanogater* SRR900425, respectively (table 2; figure 7). MCC's (Matthews Correlation Coefficient) value for REAPR was found in the range of -0.05 to +0.13, showing a random state instead of any robust performance by REAPR algorithm (table 2(b)). Similar trend was observed for Amosvalidate with MCC value ranging between +0.07 to +0.31 (table 2(a)). Also, the performance evaluation of Amosvalidate and REAPR on the assemblies from the five other assemblers for SRR892664 dataset showed the similar trend (supplementary file 5). The existing assembly validation methods like Amosvalidate and REAPR displayed no clear evidence for reliability for assembly validation process.

3.5 *Identification of wrongly assembled contigs using de novo unsupervised clustering of coverage transformed representation of assembled contigs*

Compared to the above mentioned tools to identify the wrongly assembled contigs, it was observed that a simple method, based on unsupervised clustering proposed here, achieved an equivalent level of accuracy and consistent performance in identifying the wrongly assembled contigs.

This approach has been developed with the assumption that during the features' calculation based on coverage per base position, the pattern of coverage distribution across the entire length of correctly assembled contigs should display more uniformity than the wrongly assembled sequences, which would display more fluctuations for the coverage values around the point of wrong union. During the assembly process if wrong short reads assembled with each other, then the coverage per base position and its variations could reveal notable information about the assembly. The coverage pattern after and before such union point would behave differently. The fluctuations in the coverage curve and its derivative for the entire contig length motivated us to find some pattern of distribution of consecutive coverage value ratio and maximas for the coverage plots in length specific manner. The calculation of features was done in length dependent manner classifying the sets of contigs based on length. Length based sets formation and feature calculation was done due to two main reasons: (1) It was observed that with change in length, the distribution of coverage per base showed variation even between the correctly assembled contigs. (2) Secondly, the scaling of features derived for each contig of different length to same was unable to discriminate between the correctly assembled and mis-assembled sequences. For the contigs identified as correctly assembled and mis-assembled while comparing with the reference genome, the coverage based features were estimated for every length. Thereafter, for every length the contigs for both classes were pooled accordingly, making length bins of contigs containing coverage based feature representation for every member contig. Unsupervised learning based *K*-means clustering was performed for every such length based bins containing both correctly and wrongly assembled contigs. The unsupervised clustering method assigns the observations to the same class if they share similarity for the feature vector representation. We chose *K*=2 as the contigs would belong to either the group of wrong

Table 2. Performance evaluation of assembly validation tools

	Dataset	TP	FP	FN	TN	Sn	Sp	ACC	MCC
a. Amosvalidate	SRR892664	13920	12155	6489	10787	68.21	47.02	56.99	0.16
	ERP000362	13107	2135	2451	1104	84.25	34.08	75.60	0.18
	SRR630877	62817	713	5192	232	92.37	24.55	91.44	0.07
	SRR900425	33776	6855	22587	19160	59.93	73.65	64.26	0.31
b. REAPR	SRR892664	16028	15925	4381	7017	78.53	30.59	53.16	0.10
	ERP000362	6184	1519	9374	1720	39.75	53.10	42.05	-0.05
	SRR630877	57847	632	10162	313	85.06	33.12	84.35	0.06
	SRR900425	5227	610	51136	25405	9.27	97.66	37.18	0.13

The table represents the values of Sn (Sensitivity), Sp (Specificity), ACC (Accuracy) and MCC for Amosvalidate and REAPR.

assemblies or correct assemblies. After performing length specific clustering of the assembled contigs, it was found that the clustering approach based on the above mentioned coverage derived features was reasonably accurate in separating the wrongly assembled contigs from the group of correctly assembled contigs. The contigs participating in lengthwise bins were further used for Amosvalidate and REAPR outputs evaluation at equal level to *K*-means clustering approach. The accuracy obtained from the proposed approach (SRR892664, ACC=57.29%; SRR900425, ACC= 58.58%) was comparable to Amosvalidate (SRR892664, ACC=53.26%; SRR900425, ACC=47.16%) and REAPR (SRR892664, ACC=51.16%; SRR900425, ACC =40.97%), while for the longer contigs, which are also much prone to mis-assembly and difficult to get detected, were found more accurately resolved by the proposed approach (figure 4; table 3). Also it was observed that Amosvalidate had higher error fraction for most of the lengths while REAPR showed higher error fraction for longer lengths (figure 8). This proposed unsupervised clustering approach performed reasonably good with lesser error fraction for each of the length set of contigs.

For the total lengthwise selected assembled contigs (SRR892664, n=24,421 and SRR900425, n=23,809), on the performance evaluation of Amosvalidate and REAPR, the same trend of imbalance in Sn (1.2 to 92.06) and Sp (10.2 to 99.18) was observed (table 3 (a,b)). The achieved MCC values were also very low ranging between 0.02 to 0.11 showing no sign of significant performance. *K*-means clustering also achieved almost same level of accuracy with MCC of 0.145 and 0.11 (table 3). Similarly, the same trend of imbalance between Sn and Sp was observed while performance evaluation for Amosvalidate, REAPR and K-means clustering approaches for lengthwise bins created for the other five assemblers' output (supplementary files 6 and 7).

Clustering by *K*-means for coverage derived features (combining values derived from three features: coverage, coverage ratio and normalized local maximas) for SRR892664 resulted in to the error fraction between 0.0 to 0.5 for the set of contigs represented by 551 variants of length (contigs lengths: 178 to 1,315 bases) where reported mean error was 0.42±0.06. The error fraction value for Amosvalidate fluctuated between 0.1 to 0.9 with mean error of 0.49±0.10, while REAPR error fraction was between 0.2 to 0.9 with mean error of 0.48±0.08 (figure 8A; supplementary file 8).

For ERP000362 and SRR630877, representing *H. sapiens* Illumina MiSeq and Illumina GAIIx datasets, respectively, only a single length (contig length: 384 bases and 455 bases for Illumina MiSeq and Illumina GAIIx, respectively) was obtained from SOAPdenovo2 based *de novo* assembly consisting of at least 5 positive as well as negative instances for particular length. Therefore, these were not processed further.

For SRR900425, representing the *D. melanogaster* Illumina HiSeq 2000 dataset, a total of 1,176 length variant contigs (contigs length ranged from 352 to 3,097 bases) were obtained. Amosvalidate and REAPR scored the error fraction value between 0.16 to 0.85 and 0.31 to 0.66 with mean value 0.47±0.07 and 0.49±0.02, respectively. Error fraction value for clustering by *K*-means for each of the three features fluctuated between 0.08 to 0.5, 0.1 to 0.5 and 0.2 to 0.5 with mean error fraction of 0.40±0.07, 0.41±0.06 and 0.42±0.04 for normalized coverage, coverage ratio and normalized maximas respectively. Error fraction values calculated for each length while combining the values from three features and clustering with *K*-means resulted into mean error of 0.40 ±0.07 (figure 8B).

This needs to be mentioned here that this dataset had much shorter sequences with the longest sequence going up to only 1 kb. As already mentioned above, for shorter contigs, Amosvalidate and REAPR performed at par of the presented clustering method. However, comparatively higher accuracy was attained for the longer length contigs by the clustering approach presented here.

A further study was carried out to measure if the presented clustering based method could be complemented

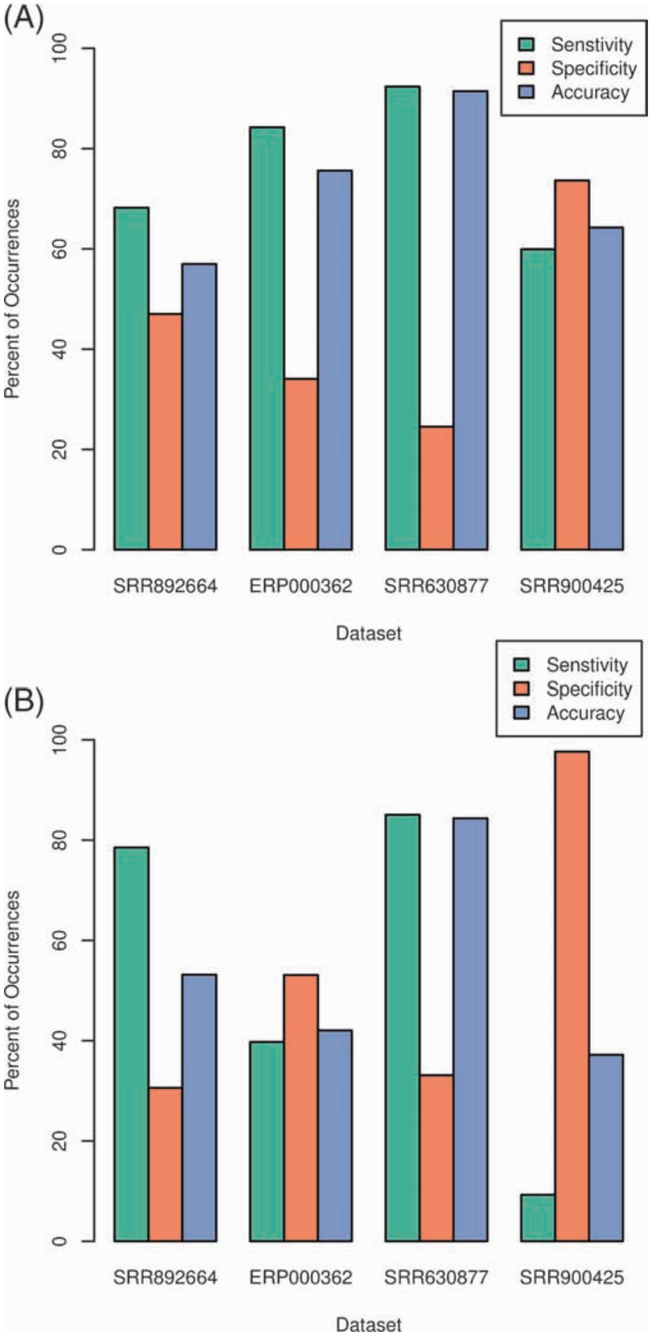


Figure 7. Comparative plot for Sn (Sensitivity), Sp (Specificity) and ACC (Accuracy) for Amosvalidate and REAPR performance. The bar plot represents (**A**) Amosvalidate and (**B**) REAPR performances for the four datasets.

with REAPR's approach to increase the overall success rate in the identification of the wrongly assembled contigs. Two-way analysis was done: (1) starting with REAPR's output and then combining the clustering method to correct the REAPR output, and (2) starting with the mentioned clustering based approach, followed by REAPR run on its output. In human, for Illumina HiSeq 2000 dataset (SRR892664), it was observed that from the total false positives (n=10,957, 49.34% of the total predicted as positive by REAPR) and false negatives (n=970, 43.73% of the total predicted as negative by REAPR) in REAPR predictions, 6,995 (58.64%) contigs were classified as true negatives (n=6548, 59.76%) and true positives (n=447, 46.08%) by the clustering method. Implementation of clustering approach after REAPR's classification, an increment in accuracy was observed with 56.13% (from the total of FP+FN =11,927 contigs in REAPR characterized into TN+TP = 6,995 contigs by *K*-means). With the second approach (*K*-means clustering followed by REAPR run), increment of 52.70% of accuracy was achieved with 5,497 contigs correctly classified by REAPR from 10,429 contigs characterized as FP and FN by K-means clustering. With the second approach, from the total false positives (n=4,908, 42.29% of the total predicted as positive by *K*-means clustering) and false negatives (n=5,521, 43.07% of total predicted as negative by *K*-means clustering) in clustering method predictions, 5497 (52.7%) contigs were correctly classified as true negatives (n=499, 10.16%) and true positives (n=4,998, 90.52%) by REAPR. Also, it was observed that REAPR's tendency to wrongly classify the mis-assembled contigs was apparent again, resulting into high number of wrongly assembled contigs characterized as the correctly assembled contigs (FP= 10,957, 89.77%). From the total of false positives (n=10,957) in REAPR predictions, 6,548 contigs could be corrected with our clustering approach. It was mainly due to the lesser instances (less than five instances for both positive and negative groups) that were used for length-wise classification, resulting into non-involvement of contigs of particular lengths in the proposed clustering method. Similarly, in *D. melanogaster* (SRR900425), implementation of clustering approach after REAPR's classification, increment in the accuracy was observed with a value of 71.52% (from the total of FP+FN =14,054 contigs in REAPR characterized into TN+TP =10,051 contigs by *K*-means). With the second approach (*K*-means clustering first and REAPR afterwards), increment of 59.41% of accuracy was achieved, with 5,858 contigs correctly classified by REAPR from 9,861 contigs characterized as FP and FN by *K*-means clustering. Combining these two approaches resulted into overall accuracy to 79.80% (SRR892664) and 83.18% (SRR900425) with MCC values of 0.629 and 0.710 in SRR892664 and SRR900425 datasets, respectively (table 3(d)). The performance ROC plot (figure 9) for the results for *K*-means, REAPR, Amosvalidate and REAPR+*K*-means depicts the relative trade-off between true positive and false positive

Table 3. Performance evaluation of assembly validation tools with the set of contigs participating in K-means lengthwise clusters

	Dataset	TP	FP	FN	TN	Sn	Sp	ACC	MCC
a. Amosvalidate	SRR892664	7651	6849	4565	5356	62.631	43.884	53.262	0.066
	SRR900425	2567	995	11584	8663	18.140	89.698	47.167	0.108
b. REAPR	SRR892664	11246	10957	970	1248	92.060	10.225	51.161	0.040
	SRR900425	176	79	13975	9579	1.244	99.182	40.972	0.020
c. *K*-means	SRR892664	6695	4908	5521	7297	54.805	59.787	57.295	0.146
	SRR900425	10128	5838	4023	3820	71.571	39.553	58.583	0.116
d. REAPR + *K*-means	SRR900425	10186	38	3965	9620	71.981	99.607	83.187	0.710
	SRR892664	11693	4409	523	7796	95.719	63.875	79.804	0.629

The table represents the values of Sn (Sensitivity), Sp (Specificity), ACC (Accuracy) and MCC for a. Amosvalidate, b. REAPR, c. *K*-means and d. *K*-means + REAPR.

rate. The combination of REAPR and K-means clustering approach scored the best, yielding 95.7% sensitivity and 63.8% specificity for SRR892664 (human dataset) (figure 9A), and 71.9% sensitivity and 99.6% specificity in SRR900425 (*D. melanogaster* dataset) (figure 9B). Similar trends were observed for performance evaluation over the output from the five different assemblers (supplementary file 9). All this suggested that combining these two approaches would definitely improve the performance in judging the assembly at contig level.

Mis-assembly identification using *de novo* unsupervised clustering of coverage transformed representation of assembled contigs with three features namely, coverage, coverage-ratio and normalized maximas, emerged as a reliable method. The method suggested in the current study, while dealing exclusively with the errors generated at contig levels, works well with most commonly used sequence assemblers. Also, the accuracy in identification of wrongly assembled primary contigs can be further improved if REAPR and the proposed clustering method are combined together. A further effort is required to implement such clustering based approach while performing *de novo* assembling to minimize wrong assembly or identify the wrongly assembled contigs.

The codes and pipeline used in the above mentioned approach to detect mis-assemblies have been provided as a code bundle set at sourceforge (*https://sourceforge.net/projects/de-novo-assembly-validation*) and at SCBB software page (*http://scbb.ihbt.res.in/SCBB_dept/Software.php*) as AVA, Assembly Validator.

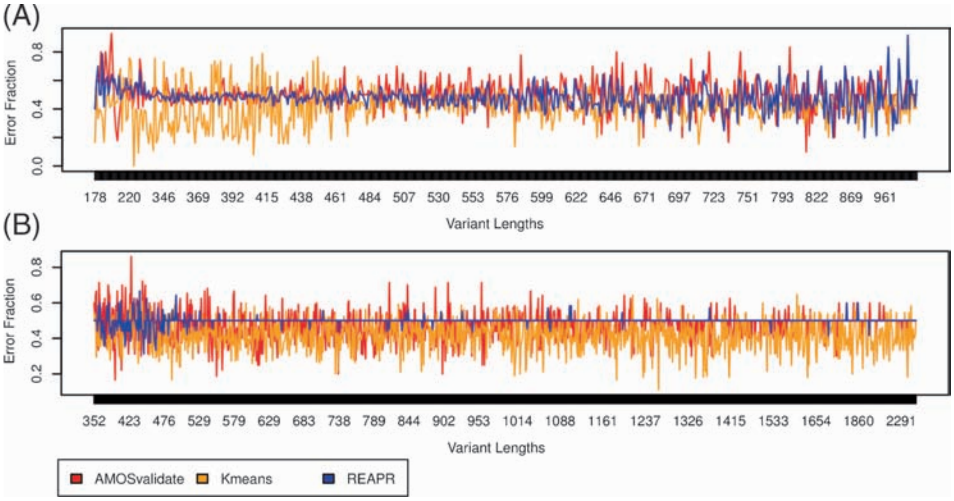


Figure 8. Comparative error fraction plot for Amosvalidate, REAPR and *K*-means clustering along variable contig lengths. Comparative error fraction plot showing error rate variation along different lengths' of contigs for (**A**) *H. sapiens* SRR892664 and (**B**) *D. melanogaster* SRR900425.

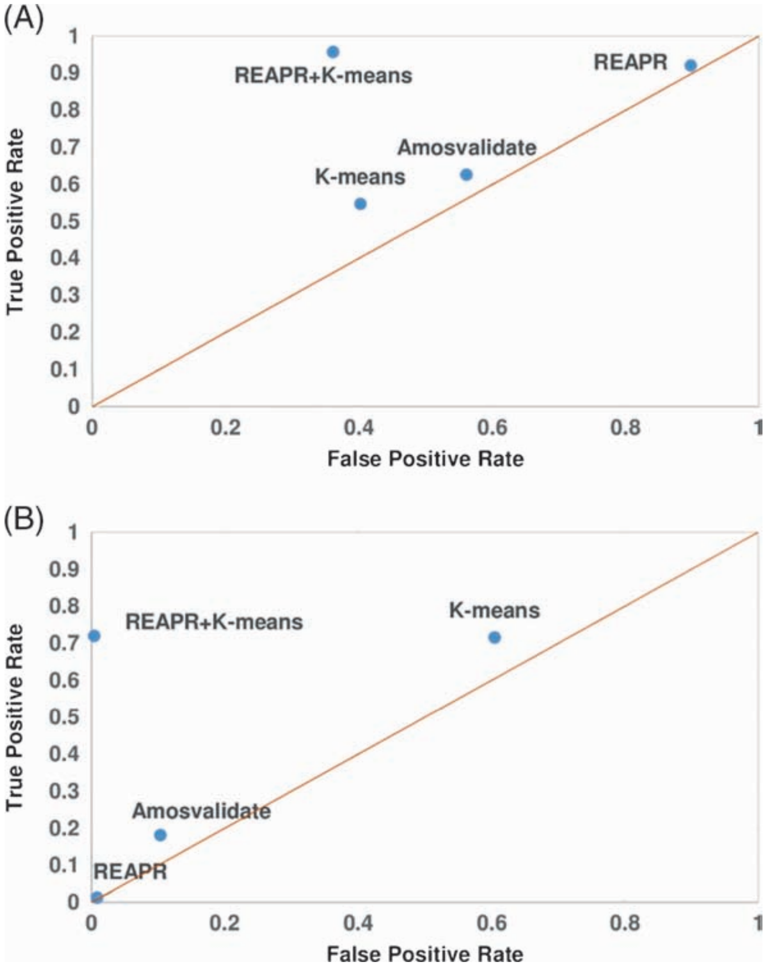


Figure 9. Performance plots for False positive rate (FPR) vs True positive rate (TPR) in ROC space. Results from *H. sapiens* (SRR892664) (**A**) and *D. melanogaster* (SRR900425) (**B**) are represented. The results from *K*-means, REAPR, Amosvalidate and REAPR+*K*-means from contingency table are plotted as points. A ROC space is defined by FPR and TPR as x-axis and y-axis, respectively. Diagonal line (orange colour) divides the ROC space. Points above the diagonal represent betterperformance results.

4. Conclusions

The present work describes a mis-assembly detection method developed using an unsupervised clustering approach based on per-base position-coverage-derived features. Detecting mis-assembly at primary *de novo* assembly step is of prime importance because the error once introduced at fundamental and initial steps of assembling affect the entire downstream process and overall assembly quality, resulting in wrong interpretations and poor credibility of the overall assembly. Genome complexity, technological constraints and associated error profiles are the major issues to be handled by the assembly algorithms. *De novo* sequence assembling appears to be more prone towards mis-assembling while constructing the contigs from small reads without taking any guide or reference support, which becomes more complicated due to genomic complexities and repetitive content. Many steps have already been taken to improve the assembly process and further validation of assembly. It appears that there is no assurance that the commonly used statistics to gauge the assembly quality would guarantee a good *de novo* assembly, and more specifically, the identification of wrongly assembled contigs Dependence upon any single metric could be misleading when deciding on a reliable assembly. Contigs with longer lengths appear to exhibit higher chances of mis-assembly. Further to this, the *de novo* assemblers still struggle with repetitiveness of the genome and do not perform well in handling the repeats. They appear to avoid considering repetitive region's reads for assembling by limiting themselves to resolve to shorter contigs, or else they report large number of wrong assemblies. There are very limited tools to identify wrongly assembled contigs. The present study found that these tools are not sufficient enough to detect the mis-assembled contigs with high precision. An unsupervised *de novo* clustering approach was proposed here to detect the wrongly assembled contigs, which appears to perform reliably. Also it was found complementing well with the already existing approach like REAPR, as the combination of these two approaches provided better identification of wrongly assembled contigs. The approach presented here may be further refined and implemented to get better and more reliable *de novo* assembly output.

Acknowledgements

We are thankful to Rohit Chauhan for his help during the analysis. VC is thankful to DST for SRF-INSPIRE fellowship, RK is thankful to DBT-BINC. We are thankful to all the researchers who provided their valuable data as an open access resource. The manuscript has IHBT communication ID: 3959. This work was supported under project funding BSC-121 (CSIR-12th FYP GENESIS project).

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, *et al*. 2000 The genome sequence of Drosophila melanogaster. *Science* **287** 2185–2195

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, *et al*. 2011 The genome of Theobroma cacao. *Nat. Genet.* **43** 101–108

Berkman PJ, Lai K, Lorenc MT and Edwards D 2012 Next-generation sequencing applications for wheat crop improvement. *Am. J. Bot.* **99** 365–371

Boisvert S, Laviolette F and Corbeil J 2010 Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* **17** 1519–1533

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, *et al*. 2013 Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience.* **2** 10

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL 2009 BLAST+: architecture and applications. *BMC Bioinf.* **10** 1–9

Chu T-C, Lu C-H, Liu T, Lee GC, Li W-H and Shih AC-C 2013 Assembler for de novo assembly of large genomes. *Proc. Natl. Acad. Sci. USA* **110** E3417–E3424

Clark SC, Egan R, Frazier PI and Wang Z 2013 ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics.* **29** 435–443

Consortium T 1000 GP 2010 A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073

Consortium T 1000 GP 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491** 56–65

Consortium TEP 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Ann Blomberg L, Bouffard P, Burt DW, *et al*. 2010 Multi-platform next-generation sequencing of the domestic Turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.* **8** e1000475

Dohm JC, Lottaz C, Borodina T and Himmelbauer H 2008 Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105

Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, *et al*. 2011 Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res.* **21** 2224–2241

Ewing B and Green P 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8** 186–194

Ewing AD and Kazazian HH 2011 Whole-*Genome Res.*equencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* **21** 985–990

Ewing B, Hillier L, Wendl MC and Green P 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8** 175–185

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, *et al*. 2013 Ensembl 2013. *Nucleic Acids Res.* **41** D48–D55

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, *et al*. 2013 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493** 216–220

Gahlan P, Singh HR, Shankar R, Sharma N, Kumari A, Chawla V, Ahuja PS and Kumar S 2012 *de novo* sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major transcriptome adjustments. *BMC Genomics* **13** 126

Hansen KD, Brenner SE and Dudoit S 2010 Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131

Hartigan JA and Wong MA 1979 Algorithm AS 136: A K-means clustering algorithm. *J. R. Stat. Soc.: Ser. C: Appl. Stat.* **28** 100–108

Henry RJ 2012 Next-generation sequencing for understanding and accelerating crop domestication. *Brief Funct. Genomics* **11** 51–56

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, *et al*. 2009 The genome of the cucumber, Cucumis sativus L. *Nat. Genet.* **41** 1275–1281

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M and Otto TD 2013 REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14** R47

Huse SM, Huber JA, Morrison HG, Sogin ML and Welch DM 2007 Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8** R143

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J 2005 Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110** 462–467

Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DST, Humphrey J, *et al*. 2014 Ensembl genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.* **42** D546–D552

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, *et al*. 2001 Initial sequencing and analysis of the human genome. *Nature* **409** 860–921

Langmead B, Trapnell C, Pop M and Salzberg SL 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10** R25

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, *et al*. 2010a The sequence and *de novo* assembly of the giant panda genome. *Nature* **463** 311–317

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, *et al*. 2010b *de novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20** 265–272

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, *et al*. 2012 SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1** 18

MacCallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, *et al*. 2009 ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol.* **10** R103

Manning JE, Schmid CW and Davidson N 1975 Interspersion of repetitive and nonrepetitive DNA sequences in the Drosophila melanogaster genome. *Cell* **4** 141–155

Phillippy AM, Schatz MC and Pop M 2008 Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9** R55

Poon AFY, Swenson LC, Dong WWY, Deng W, Kosakovsky Pond SL, Brumme ZL, Mullins JI, Richman DD, *et al*. 2010 Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. *Mol. Biol. Evol.* **27** 819–832

Rahman A and Pachter L 2013 CGAL: computing genome assembly likelihoods. *Genome Biol.* **14** R8

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, *et al*. 2012 GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22** 557–567

Sanger F, Nicklen S and Coulson AR 1977 DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74** 5463–5467

Schatz MC, Witkowski J and McCombie WR 2012 Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol.* **13** 243

Shumway M, Cochrane G and Sugawara H 2010 Archiving next generation sequencing data. *Nucleic Acids Res.* **38** D870–D871

Simpson JT and Durbin R 2012 Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res.* **22** 549–556

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM and Birol I 2009 ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19** 1117–1123

Treangen TJ and Salzberg SL 2012 Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13** 36–46

Vezzi F, Narzisi G and Mishra B 2012 Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PLoS One* **7** e52210

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, *et al*. 2012 The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.* **44** 1098–1103

Zerbino DR and Birney E 2008 Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18** 821–829

Zimin AV, Smith DR, Sutton G and Yorke JA 2008 Assembly reconciliation. *Bioinformatics* **24** 42–45

MS received 07 November 2015; accepted 04 July 2016

Corresponding editor: Alok Bhattacharya