2022 2nd International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 24-26, 2022

A Model for Optimal Assignment of Non-Uniquely Mapped NGS Reads in DNA Regions of Duplications or Deletions

1st Rituparna Sinha
Department of Information Technology
Heritage Institute of Technology
Kolkata, India
rituparna.sinha@heritageit.edu

2nd Rajat K. Pal
Department of Computer Science and Engineering
University of Calcutta
Kolkata, India
0000-0001-9838-6500

3rd Rajat K. De
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India
0000-0001-6080-1131

***Abstract*—Massively parallel sequencers have enabled genome sequences to be available at a very low cost and price, which opened huge scope on analyzing human genome sequences from different perspectives, thereby the association of diseases with genetic alterations gets further enlightened. However, the sequencing process and alignment of NGS technology based short reads suffer from various sequencing biases which needs to be addressed. In this work, the mappability bias occurring with respect to repeat rich regions of the DNA have been addressed in a novel approach. A model has been designed which considers all non-uniquely mapped reads and performs a pipeline of computations to allocate the reads to an optimal location, due to which the precise detection of breakpoints in the region of duplications and deletions are obtained. In addition, the application of this model for mappability bias correction, prior to the detection of structurally altered regions of the genome, leads to a better sensitivity value.**

***Index Terms*—NGS, DNA, Mappability Bias, Genome Sequencing, Reads, Structural Alterations, Alignment**

I. Introduction

The first human genome was sequenced in 13 years [1] and at a very high cost of approximately $1 billion. With the advent of next generation sequencing technologies [2] the cost and time of sequencing has greatly reduced. The massively parallel sequencers is capable of sequencing approximately 18000 genomes in a single day. They produce millions of short reads (short DNA sequences) and the reads [3] are sequenced in parallel. Later these reads are aligned to a standard reference sequence to obtain the positions from where the reads were generated. These mapping information are used by researchers to detect DNA regions effected by structural alterations, which plays a significant role in many diseases, including cancer. When a large chunk of DNA gets duplicated or deleted then it might happen that the duplicated/deleted region encompasses through many genes. Thereby in case of duplications the genes copy number changes from 2 (diploid genome) to higher numbers, and thus the expression levels of the genes gets affected. Similarly, in case of large deletions it might happen that the copy number of genes gets changed from two to 1 or 0, i.e. the whole gene gets deleted. Therefore the particular protein produced by that gene gets severely affected. These imbalances in gene dosage or in other words, the mutation/alteration in certain category of genes called proto-oncogenes (which helps in cell growth), causes it to become active (on) when it should not be, and thereby leading to uncontrolled cell division. Moreover, the mutations/alterations in the tumor suppressor genes (that helps in cell death) causes it to be in off state, thereby affecting the cell death procedure. All these circumstances, where the structural alterations causes oncogenes to be in on state and tumor suppressor genes to be in off state causes uncontrolled cell growth leading to tumorgenesis.

However, the read generation process and the mapping process undergoes some kind of bias, called GC bias [4] [5] and mappability bias [6] [7]. In the DNA many sequences are there which gets repeated at different locations . Now, reads generated from these repeat regions suffers from mappability bias, since the reads from these regions when aligned to the reference sequence, gets mapped to multiple positions. This bias can be handled by aligning reads of repeat rich regions to any one random location. Therefore it might be wrongly aligned to a position from where it was not generated. In [8], a new method is developed for correction of GC bias on the basis of multi resolution analysis, where a translation-invariant wavelet transform is used to decompose biased raw signals into high- and low-frequency coefficients. Then the relation between GC proportion and DOC [9] of the genomic regions is modeled and new control DOC signals that reflect the GC bias is constructed. BEADS [10], is a bias elimination algorithm for correcting sequence bias in deep sequencing, which follows a three-step normalization scheme that successfully unmasks real binding patterns in ChIP-seq data.

In this work, a model is developed in a novel way, for optimally aligning all the non-uniquely mapped reads to the genome. The focus is those regions of the DNA effected by some structural alterations [11]. If the non-uniquely mapped reads are randomly allocated, or are ignored then there is a probability that detection of structurally altered regions would suffer from Type II errors. The model first of all clusters all non-uniquely mapped reads based on the pattern

978-1-6654-8407-7/22/$31.00 ©2022 IEEE

of repeat regions of the DNA. With respect to each cluster, a membership matrix is initially assigned equal probabilities of mapping to the repeat regions. Later the membership matrix gets updated based on similarity of mapping information with regions of duplications, identified using GenSeg [12] algorithm. Probability of mapping in regions effected by deletions is least. This model enables the detection of structurally altered regions with higher recall value.

II. METHODOLOGY

In this work, a model is proposed to optimally map the reads generated from repeat rich regions of a DNA sequence. Next generation sequencing technologies generate reads which may be obtained from repeat rich regions of the DNA. These reads when aligned to the reference sequence gets mapped to all the similar repeat rich regions of the DNA (segment). In order to assign these non-uniquely mapped reads into an optimally selected segment, the proposed model follows a pipeline of processes as described below.

A. Clustering of non-uniquely aligned reads

Initially, all the non-uniquely mapped reads are clustered based on the similarity in the patterns of the repeat rich segment from where they are generated. Next, a membership matrix M, is formed for each of these cluster (of reads). Each membership value represents the probability of the reads getting aligned to a particular segment. Initially, equal probability is assigned to each segment, where segments are the repeat rich regions and represents the columns of the matrix. Matrix M, represents the membership of a set of reads to a segment, and the membership values are as provided.

$$M[i][j] = 1/S, \sum_{j=1}^{S} M[i][j] = 1 \tag{1}$$

where S is the total number of repeat regions of a particular pattern and $M[i][j]$ represents the membership value of the set of reads belonging to ith cluster, mapped to jth segment (repeat region). For each cluster a separate matrix is generated, and order is $1 * S$, where S is the number of repeat regions having a particular pattern of DNA sequences. Figure 1 represents pattern ATTC is repeated 3 times, so $S = 3$, and reads generated from these regions have equal probability to get mapped to region1, region2, or region 3, with a membership value $1/3$. Figure 2 represents the pipeline of processes followed by the model.

B. Finding Score of each Segment

For each segment, the following task is performed. First of all the DNA is divided into fixed sized bins. Next, all reads belonging to a particular ith cluster are aligned to the jth segment. The bins to which the reads get mapped are obtained. A score is assigned to the bins of jth segment where the reads got aligned. The score of jth segment is defined as

$$(score1_j) = (\mu - m1)/std, where 1 \leq j \leq S \tag{2}$$

where n is the total number of bins for the whole genome, μ is the mean read count for the whole genome, $m1$ is the mean read count of the bins where the reads got aligned and std is the overall standard deviation. Similarly, for the other segments, all reads of ith cluster are mapped to that segment and score is calculated. Therefore, for S such repeat segments (same pattern) throughout the DNA S scores will be obtained.

C. Extending a segment until breakpoints

The bins of jth segment to which the reads of ith cluster get aligned are tracked to find out the minimum and maximum coordinate of the bins. For simplicity let us consider there is only one particular pattern of DNA segment which got repeated. As shown in Fig. 1, pattern ATTC got repeated 3 times and reads of cluster 1 are reads generated from ATTC pattern. Now, as described in point b) the reads are aligned to each of the segment, and let for the jth segment, the bins be $[b1, b2]$ where the reads got aligned . Perform segmentation starting from the left bin of $b1$, i.e., $(b1-1)$th bin, and perform segmentation starting from the right bin of $b2$, i.e., $(b2+1)$th bin. The segmentation algorithm would provide us breakpoints on the left side of b1 and would also provide a breakpoint on the right side of $b2$. Let the two breakpoints obtained be $b11$ and $b21$. Find the score of the bins contained in the region $[b11, b1-1]$, and bins in the region $[b2+1, b12]$ as provided.

$$(score2_j) = (\mu - m2)/std, where 1 \leq j \leq S \tag{3}$$

Here, $m2$ is the mean read count of the above mentioned region . Repeat the same procedure for each of the S segments.

D. Updation and Allocation

In order to find whether the reads get optimally aligned to the jth segment we do the following. Find the distance d_{ij} of the $score1_j$ and $score2_j$. Also find the distance of $score1_j$ with the $score2_k$ of other segments where $1 \leq k \leq S$ and k != j Next, update the membership value.

$$M[i][j] = 1/\sum_{(k=1)}^{S} (dij/dik) \tag{4}$$

Here, $M[i][j]$ is the membership value of the ith cluster reads getting aligned to jth segment. Now, if the d_{ij} is smaller compared to d_{ik} then the membership value of the jth segment will be higher otherwise if distance of score1 is smaller with its neighboring segments, then the membership value gets reduced. Increase in membership value implies that the probability of assignment in that particular segment gets increased. In this manner the matrix gets updated. Finally, each cluster of reads gets allocated to the segment optimally.

E. Optimal Assignment in Duplication or Deletion Regions

The segments containing repeat regions are not effected by any duplication or deletion events. All the segments have more or less equal probability of allocation of reads. However, when the repeat regions falls in the DNA where some structural alteration has happened then the following takes place.

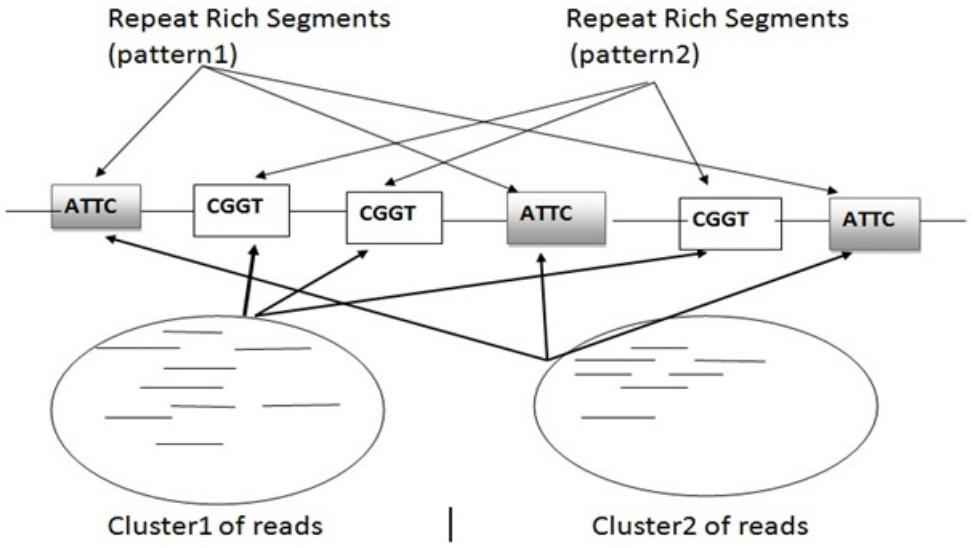

Fig. 1. The reads generated through NGS technology, which are obtained from repeat rich regions of the DNA are clustered based on the repeat patterns from where they are generated. Cluster1 contains all reads that gets aligned to segments corresponding to pattern1, whereas Cluster2 contains all reads that gets aligned to segments corresponding to pattern2.

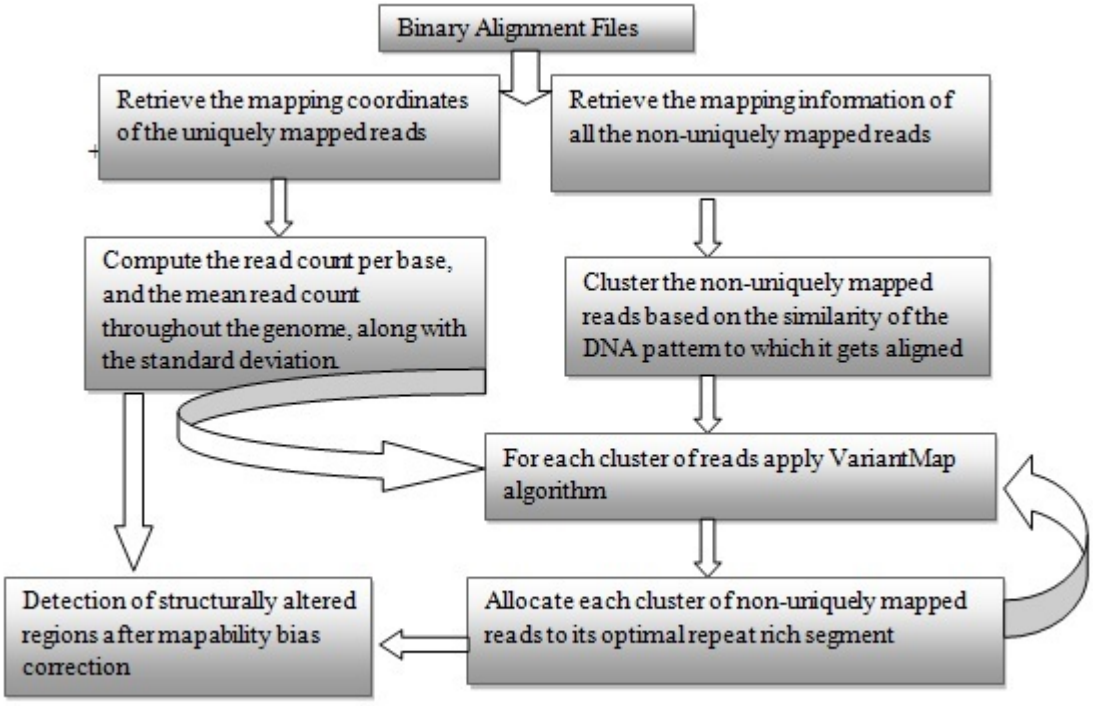

Fig. 2. The pipeline of processes followed by the model.

Case 1: Consider a repeat rich region where some structural alteration such as duplication has occurred. In the region of duplication the count of reads is significantly higher as compared to regions of DNA not effected by any duplication or deletion. Therefore if the non-uniquely mapped reads are assigned to the duplication region, then the distance d_{ij} will be lesser as compared to the distance of scores with the other segments where the count value is near the overall mean read count. It is found that the membership value will be high in the segment having duplication, hence the optimal alignment will be to this region of duplication.

Case 2: In the case where the repeat rich region falls under a DNA region where a deletion event has happened. The region effected by deletion will have a significantly low read count as compared with the overall mean read count throughout the DNA. Therefore if the non-uniquely mapped reads get aligned to this region, then d_{ij} value will be greater as compared to the distance of scores with the other segments. Hence, the membership value for the segment effected by deletion will be low compared to the other segments. Hence, the reads will not be allocated to this segment.

III. RESULT

A DNA consists of four nucleotides, 'A', 'T', 'C', and 'G'. A simulation is performed by the following approach. A DNA string is prepared with a combination of four characters present in random order. Ten million such characters are considered resulting in simulating a DNA of length 1 MB (mega bases). Next, a random region of 200bp (base pairs) sequences are considered and repeated the same sequence at another 2 random locations in the DNA sequence. This acts as the reference DNA. Next a sample $S1$ is prepared from the reference sequence. In order to maintain a coverage of $30X$, the reference sequence is first of all copied for 30 times. Then few instances are created as mentioned below.

Case 1: In one of the repeat rich region a large duplication event is introduced. The size of the region being duplicated is of 1500 bp and the region is duplicated three times. The duplication event simulates a structural alteration in the DNA sequence which causes duplications of a large chunk of DNA at a stretch. The other two repeat rich regions are considered to be effected with no structural variations. Other than this two more duplication events are also introduced randomly at different locations. Altogether three duplication events including one at repeat rich region. Figure 3 represents the mapping information per base. It has been observed that the locations $[46352, 70832]$, $[100099, 105381]$ and region $[678163, 680958]$ has very high mapping information with a maximum value around 8000.

Case 2: In one of the repeat rich region a deletion event is introduced and the size of deletion is 2000 bp. The deletion event simulates the event where a large chunk of DNA gets removed from sample DNA sequence. The other two repeat rich regions are considered to be effected with no structural variations. Two more deletion events are introduced at random locations, excluding the repeat rich regions. Altogether three deletion events are introduced in the DNA sequence of the sample.

Algorithm 1 VariantMap assigns non-uniquely mapped reads optimally to one specific repeat region, and is able to address the instances where the repeat regions are effected by duplication or deletion events.

INPUT:All non-uniquely mapped reads. RC array having information of all uniquely mapped reads.

1: Perform clustering of all reads based on reads generated from similar repeat patterns
2: Divide the whole DNA into bins of fixed size.
3: Initialize a membership matrix M as described in equation 1.
4: **for** each cluster of reads **do**
5: **for** For each jth segment (considering S number of repeat segments having similar pattern and $j \leq S$) **do**
6: Compute the count of aligned reads per bin and find average count μ and standard deviation (std) throughout the genome.
7: Align the reads to the segment and track the bins to which they get aligned, and store them in an array arr.
8: $Score1_j$ = ComputeScore(μ, std, arr) // Call the ComputeScore method to calculate the score as defined in equation 2. The parameters sent to this method are the overall mean read count, the overall standard deviation, and the array containing bins where the cluster of reads get aligned.
9: b11 = LeftExtend(μ, std, arr, RC) // Extend the segment and identify the breakpoints as described in section II-C.
10: b12 = RightExtend(μ, std, arr, RC) // Extend the segment and identify the breakpoints as described in section II-C.
11: $Score2_j$ = CompScore(μ, std, $arr1$) // $arr1$ is an array of bins contained within the extended region. Compute the score as described in equation 3.
12: **end for**
13: **for** For each segment **do**
14: Update the membership matrix as described in equation 4.
15: **end for**
16: Allocate the cluster of reads to the segment having highest membership value.
17: **end for**

The reference sequence is logically divided into bins, each of 100bp. Therefore 10000 such bins are created. From the sample $S1$ randomly reads of size in the range $[300, 400]$ bp are created, therefore producing several short reads. The reads are aligned using BWA [13] alignment algorithm, and bin wise count of mapped reads is considered. Next algorithm 1 is applied, where the segmentation algorithm chosen is GenSeg. It has been observed that in case1 instance, the probability matrix M has highest probability in the repeat segment where the duplication event is introduced. Thereby

Algorithm 2 LeftExtend algorithm extends the target segment and determines the leftmost boundary of the extended segment.

INPUT: arr containing bins where the cluster of reads got aligned.Overall μ, std and read count information of all uniquely mapped reads.

1: Set $b1$ = minimum value of arr // the minimum value signifies the minimum coordinate to which the cluster of non-uniquely mapped reads got aligned.
2: Set $start$ = $b1 - 1$ // $start$ signifies the bin from which the segmentation has to start.
3: **while do**$start$ /geq 1
4: Apply segmentation beginning from $start$ bin. Use GenSeg [12] algorithm to find the breakpoint.
5: **end while**
6: Return the bin where breakpoint received. // Signifies the left most boundary of the extended segment.

Algorithm 3 RightExtend algorithm extends the target segment and determines the rightmost boundary of the extended segment.

INPUT: arr containing bins where the cluster of reads got aligned.Overall μ, std and read count information of all uniquely mapped reads.

1: Set $b2$ = maximum value of arr // the maximum value signifies the maximum coordinate to which the cluster of non-uniquely mapped reads got aligned.
2: Set $start$ = $b2 + 1$ // $start$ signifies the bin from which the segmentation has to start.
3: **while do**$start$ /leq n // n is the coordinate of the end bin of the genome.
4: Apply segmentation beginning from $start$ bin. Use GenSeg [12] algorithm to find the breakpoint.
5: **end while**
6: Return the bin where breakpoint received. Signifies the rightmost boundary of the extended segment.

the non-uniquely mapped reads get aligned to the duplication region instead of the other repeat regions. Whereas, in case2 instance, the probability matrix M has least probability of mapping in the region effected by deletion. As represented in Fig. 4, it has been observed that without applying the VariantMap algorithm for mappability bias correction, the genome alteration detection methods may suffer from having many false negatives. Moreover, it has also been observed that after applying VariantMap, if the variant detection algorithms are executed then it leads to precise detection of breakpoints as represented in Fig. 4

IV. CONCLUSION

With the advent of genome sequencing technologies, a wide scope has opened to the research community to study and analyze multiple genomes simultaneously. Identifying structurally altered regions in the genomes is important, since a large chunk of alterations may cause change in the copy number of genes, which may affect the gene dosage. These structurally altered regions thereby causes several disease, including cancer. However, sequencing suffers from certain biases, which needs to be addressed before performing the analytics. This work has designed a model which performs a pipeline of computations that handles the mappability bias issue caused from aligning reads generated from repeat rich regions of the DNA. The model uses VariantMap as described in algorithm 1, which considers all non-uniquely mapped reads as input, along with mapping information of all uniquely mapping reads. It using other two algorithm 2 and algorithm 3 computes scores to ultimately update a probability matrix, which assigns maximum probability to the repeat region effected by duplication, and minimum probability to the region effected by deletion using equation 4. It has been observed that the traditional mappability bias correction methods when applied prior to detect the structurally altered regions of the genome, results in low recall value, however, when VariantMap is applied for correcting mappability bias, it results in good recall value. Moreover, the model also enables prediction of breakpoints with high precision as represented in Fig. 4. In simulated data sets,the model is working with high sensitivity in prediction of variants, however the model is yet to be tested on human genomes for analyzing its performance on real data sets.

REFERENCES

[1] J. C. VENTER, M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL, and G G. SUTTON, The Sequence of the Human Genome, Science, vol. 291, 1304-1351, 2001.

[2] M. L. Metzker, "Sequencing technologies—the next generation", Nat. Rev. Genet., vol. 11, 31–46, 2010.

[3] S. Samaddar , R. Sinha , R. K. De, "A Model for Distributed Processing and Analyses of NGS Data under Map-Reduce Paradigm," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 16, 827-840, 2019.

[4] Y. Benjamini, and T. P. Speed, "Summarizing and correcting the gc content bias in high-throughput sequencing", Nucleic Acids Res., vol. 40, e72, 2012.

[5] Y. Xia, Y. Liu, M. Deng, and R. Xi, "Pysim-sv: a package for simulating structural variation data with gc-biases", BMC Bioinform., vol. 18, 2017.

[6] R. K. Auerbach, G. Euskirchen, J. Rozowsky, N. Lamarre-Vincent, Z. Moqtaderi, P. Lefranc¸ois, K. Struhl, M. Gerstein, and M. Snyder, "Mapping accessible chromatin regions using Sono-Seq", PNAS, vol. 106, 14926–14931, 2009.

[7] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein, "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls", Nat. Biotechnol., vol. 27, 66–75, 2009.

[8] H. Jang and L. Hyunju, "Multiresolution correction of GC bias and application to identification of copy number alterations", Bioinform., vol. 35, 3890–3897, 2019.

[9] Y. Ruen, Z. Cheng, Y. Tingting, L. Niu, H. Xuyun, W. Xiumin, W. Jian , and S. Yiping, "Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data", Mol. Cytogenet., vol. 10, 2017.

[10] M. Cheung, T. A. Down, I. Latorre, and J. Ahringer, "Systematic bias in high-throughput sequencing data and its correction by BEADS", Nucleic Acids Res., vol. 39, 2011.

[11] R. Sinha, S. Samaddar, and R. K. De, "CNV-CH: A Convex Hull Based Segmentation Approach to Detect Copy Number Variations (CNV) Using Next-Generation Sequencing Data," PLoS One, vol. 10, 2015.

[12] R. Sinha, R. K. Pal, and R. K. De, "GenSeg and MR-GenSeg: A Novel Segmentation Algorithm and its Parallel MapReduce Based Approach for Identifying Genomic Regions with Copy Number Variations,", IEEE/ACM Trans. Comput. Biol. Bioinform., 2022.

[13] L. Heng and R. Durbin, "Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform", Bioinform., vol. 26, 589-95, 2010.

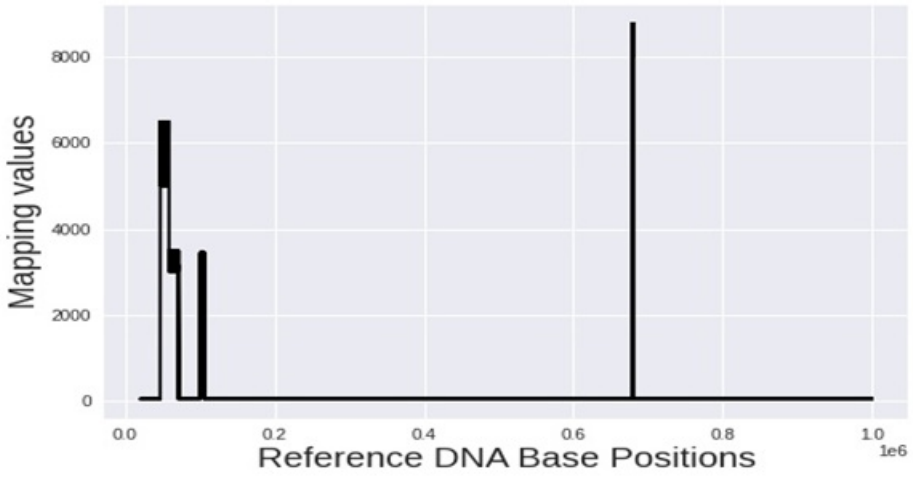

Fig. 3. It represents the mapping information of reads per DNA base. The locations having high mapping info are regions of duplications introduced in the simulation study performed.

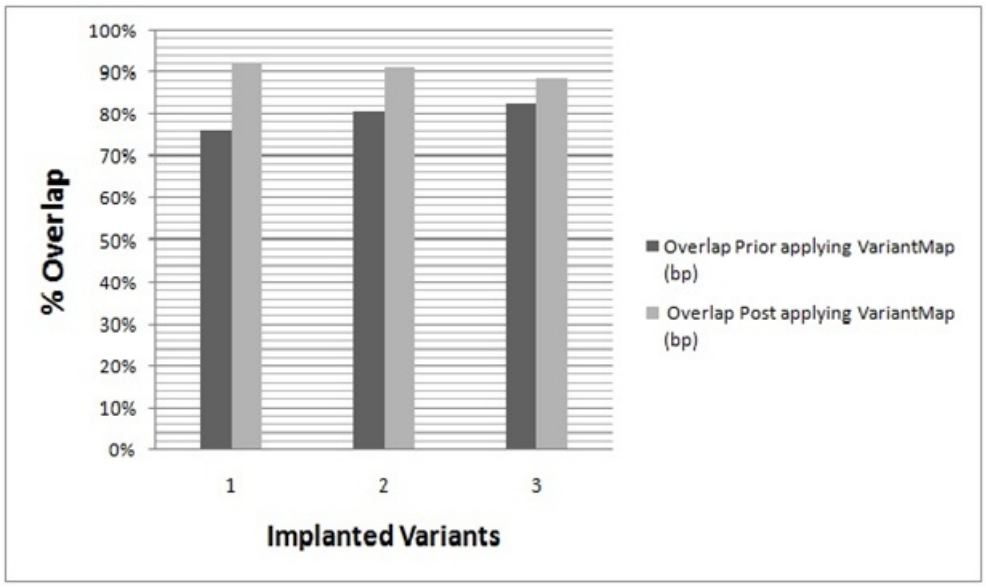

Fig. 4. It represents the percentage overlap of nucleotides detected prior applying VariantMap and post applying VariantMap.