Volume 6, Issue 11, November 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Clustering Based Approach to Detect Copy Number Variations in Genomic Regions

Prianka Kundu*, Surajit Goon
Department of Computer Application, Eminent College of Management & Technology,
West Bengal, India

Abstract*— Copy number variation is a kind of structural alteration in the mammalian DNA Sequence which is caused by insertion, deletion and duplication of large DNA segments. It is associated with many complex neurological diseases such as autism, schizophrenia, Alzheimer as well as cancer. Next generation sequencing technology provides us a new platform for detecting copy number alterations. In our method we have detected copy number variations by using NGS technology. Our method used Smith Waterman algorithm for aligning short reads which are generated from sample. The read count data may be affected by GC content and mappability bias. In order to remove this kind of bias, a statistically efficient method was used as this kind of systematic biases could lead to false detection of the read count. Moreover, K-means clustering technique is applied initially to detect the regions with variations. However, the overlapping clusters and the outlier points cannot be handled by k-means clustering algorithm. To recover from this kind of limitations we have used Fuzzy c-means clustering and observed that it was more efficient in handling genomic breakpoints. In addition our approach has higher sensitivity and F-score value in detecting large sized alterations.

***Keywords*— Structural Variations, CNV, NGS, Mappability Bias, K-means, Fuzzy C-means clustering**

I. INTRODUCTION

Copy number variation [2] is a type of genomic structural alterations which is associated with several complex genetic diseases. Copy number variation is caused by insertion, deletion and duplication of large DNA segments. The size of the DNA segment which can be altered should be greater than 10 kbp and can ranges from kilo bases (kb) to mega bases (mb). Human health can by affected by copy number variations as it is associated with various complex disease such as schizophrenia, neurological disorders, and also associated with some cancers [7]. Detection of CNV has always been a very challenging task since many years. In earlier days, to detect CNVs fluorescence in situ hybridization (FISH) [13] and array comparative genomic Hybridization (aCGH) [1] based techniques were used. But in this approach detection of CNV breakpoints was also not very precise. Next, the next generation sequencing (NGS) [5] technology is developed which generates million of short reads for the sequencing process in low cost and time. Furthermore, for detection of CNVS the next generation sequencing (NGS) technology [5] provides a new dimension.

In our method, we have detected CNVS [20] by using the next generation sequencing (NGS) technologies [2] read depth based approach. Recently a number of CNV detection methods are published. One such approach is called EWT (Event-wise Testing) [11] introduced by Yoon which uses a probabilistic model for detection of CNVs. Seg-Seq is a method which divides the genome into fixed size windows. The ratios between tumour samples read counts and reference Read counts are calculated for each and every window. The segments are called gains if the ratio is greater than 1.5 and called loss if the ratio is below 0.5. However, the overall performance of the algorithm can be affected by the window size as it must be determined first. rSW-seq [10] is another method based on the modification of smith waterman algorithm [3]. By using the concept of moving window along the genome the ratio of total number of tumour reads and total number of normal reads is calculated. However, the window size does not need to be specified first and this is the main advantage of rSW-seq over seg-seq. next, another method called CMDS (correlation matrix diagonal segmentation) [4] which calculates the Pearson's correlation coefficients. The final outcome of the correlation will be zero if no copy number alteration exists across the samples. However the positive correlation is observed if the copy number alteration exists across samples.

In our paper, we have used NGS Technology's [5] short reads based approach to detect copy number variations in the genomic regions. For alignment of the short reads, we have applied the smith Waterman algorithm [5]. Next, we have calculated the read count for each and every window. Two major biases which can affect the read count are GC content [14] and mappability bias [6]. We have developed a novel approach for correction of these two systematic biases. In addition, we have applied an appropriate smoothing technique as there existed variability in read count data. Next, the smoothened read count data is converted into some statistical measure which can add robustness against the error. Another important aspect of our paper is that, we have applied clustering to identify the regions with copy number variations. K-means [17] clustering is used to identify the regions with CNVs. However, as k means clustering unable to handle noisy data and outlier's fuzzy c means [18] clustering algorithm is applied to improve the performance of our

method. We observed that fuzzy c means clustering [19] is more efficient for precisely detecting the cluster than k-means clustering. Next, we have compared our method with two existing algorithm CMDS (correlation matrix diagonal segmentation) [16] and EWT [11]. Furthermore, we observed that for both large and small input size our method performed better with low false positive rate.

II. METHODS

The series of processes for copy number variations detection in NGS data:

Input data and its processing: Next generation sequencing technology which is also known as high-throughput sequencing is considered an important technology for genotyping and genome assembly. Our work is based on the next generation sequencing technology (NGS) [5] which allow us to sequence DNA much more quickly and efficiently than the previously used Sanger sequencing [20]. In our method, we have taken a reference sequence of size M. The reference genome M is partitioned into X number of non overlapping windows. By using next generation sequencing technology, we have taken here N number of samples. Each sample is associated with multiple numbers of short reads which are generated from a DNA sequence.

Mapping of short reads and calculates read counts: In our method, we have partitioned the reference genome into X number of non overlapping windows. The size of the non-overlapping windows was chosen as 100 bp each, for getting higher accuracy in detecting CNV breakpoints. Furthermore, the samples were divided into multiple numbers of segments called short reads. The size of short reads is considered here as 36 bp. For each individual sample, the short reads were mapped to the windows of the standard reference sequence. Next, smith Waterman algorithm [5] is used for mapping of short reads back to the reference genome. Here, the scoring of a cell depends on a variety of user specific weights which are for matches, mismatches, and gap extensions. The formula for matrix filling is.

$$M_{i,j} = \text{Maximum}\ [\ M_{i-1,j-1} + S_{i,j,}\ M_{i,j-1}+W\ ,\ M_{i-1,j}+W,0\ ] \quad \ldots\ldots\ldots\ldots(1)$$

However, we have to trace back the sequence for an optimal alignment, after filling the entire matrix. The number of reads gets aligned to a particular window is called the read count of that window. Next, we have calculated the read count for each and every window. Furthermore, more number of reads will get mapped to the window where duplications occur. We get a high read count value for those windows and a very low read count value where deletions occur.

Adjustment of GC and Mappability bias: In NGS Technology the genomic sequences that are obtained by high throughput sequencing are not uniformly distributed across the genome. However, there are at least three factors contribute to sequence bias: GC content, mappability bias, and regional biases [14]. Furthermore, this kind of systematic biases could lead to a false detection of the read count. GC content [16] is especially higher in the protein-coding regions. For this purpose, removal of bias is necessary to determine real enrichment patterns. GC content is usually expressed by a percentage value and that is expressed as **(G+C)/(A+T+G+C)*100**. Next, the GC content is calculated for each and every window by calculating the percentage value i.e. $\mathbf{Q_n=s_n/b_n}$. Where, s_n= total number of guanines and cytosine's present in the 100-bp fragments. Similarly, b_n= total number of nucleotide compositions of 100-bp fragments. However, if the window size is small then the read counts are not well distributed and for this purpose we have calculated the distribution of read counts of 100-bp windows at 30*coverage. The read count data is corrected by the following equation.

$$\textit{Corrected read count} = Rc_i.k_i/m_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Where, Rc_i is the original read count value of the ith window, k_i is the overall median read count value of all the windows and m_{ij} is the median read count of the windows that have the same GC-content and mappability values as the jth window in ith sample.

Next, mappability bias [6] is the major bias that can affect the copy number variation. Due to the short read length a small number of reads are mapped to multiple positions of windows. Mappability bias can also arise due to the repetitive regions in the reference genome. Failure to eliminate the mappability bias will lead to increased read densities within regions with higher mappability. A number of strategies have been proposed for this purpose and each approach has its own pros and cons. Some methods completely ignore the ambiguously mapped read by counting only the uniquely mapped reads. Some method randomly assign an ambiguous read to one of all possible positions to which this read is mapped by the aligner. Although this strategy is capable of identifying copy numbers in repetitive genomic regions but the drawback was that this approach suffers from false positive.

In our work, the read count data we have generated may suffer from the mappability bias and this could lead to false detection of variant. Due to this the false positive rate (FPR) would increase. However, to eliminate mappability bias we have introduced a normalization technique which is done on the read count data. We have calculated the probability of a read to get mapped to the reference genome. For example If a particular read is repeated n times then the probability of those windows where the read is repeated will be 1/n means at those windows multiple alignments occurs i.e. reads are ambiguously mapped. However, the probability of those windows will be 1 where reads are uniquely mapped. Finally, we have smoothed the read count of those windows where multiple alignments occurs. A statistical technique called moving average is used to normalize the read count data.

Standardized score of read counts: The short reads which are generated by some chemical process can be suffers from low quality sequencing and also the process of read generation is done independently. As a result the read count is not standard across all the samples. Hence to establish uniformity across all the samples a standardized score of read count needs to be adopted. So the standardized read count can be obtained by

$$\mathbf{Z_{score}} = (m_{ij}-\mu_i)/\eth_i$$

Where, μ_i denotes the mean read count of ith sample and σ_i denotes the standard deviation of read count of the ith sample over all the windows. When read count m_{ij} is high, $\mathbf{Z_{score}}$ value will be high and vice versa. A very high or very low copy number variation is observed to the regions where duplications or deletions occur.

Clustering: Next, to divide the genomic regions into homogeneous group we have used clustering. In our method, we have created 3 clusters. One for duplications i.e. where high value of the read counts is observed, One for deletions i.e. where low value of the read counts observed, and one for the cluster where there is no duplications or deletions. First, we have used K-means clustering [17] in our method. K-means algorithm minimizes an objective function given by

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\| di - vj \|)$$

Where, $\| di - vj \|)$ is the distance between the standardized read count value d_i and the centroid v_j, ci is the number of data points in i^{th} cluster.

Steps:

Input: v: The no of clusters

D: The data set i.e. the standardized read count value

t_{max}: Maximum no. of iterations

Output: A set of v clusters

Let $D = \{d_1,d_2,d_3,\ldots\ldots..,d_n\}$ be the set of data points i.e. the read count value and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

Step 1: k means clustering randomly select v cluster centers. In our method we have created 3 clusters. So we assign v=3

Step 2: Next, we have calculated the distance between each read count data d_i and cluster centers v_i by using the objective function.

Step 3: Assign the read count data d_i to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

Step 4: Recalculate the new cluster center using: $v_i = (1/c_i) \sum_{j=1}^{ci} xi$ Where, c_i represents the number of data points i.e. read counts in i^{th} cluster.

Step 5: Recalculate the distance between each data point and new obtained cluster centers.

Step 6:this process will continues until no data points was reassigned or otherwise repeat until the maximum number of iteration is reached i.e. t_{max}=10

The main limitation of the algorithm is that the data points can be always a member of one of the clusters. However, when the data is well structured this works well but in real data have extraneous data points which are not belonging to any of the clusters and they are called outliers. Furthermore, the overlapping clusters and the outlier points cannot be handled by k-means clustering algorithm.

Fuzzy C-means clustering: To recover from this kind of limitations we have used fuzzy c-means [18] clustering. Fuzzy c-means clustering calculates the cluster centers and assigns the points to these centers and this process continues until there is any change. In many ways Fuzzy c means [19] clustering is similar to k means clustering. However, the difference is that it assigns a membership value to the data items for the clusters within a range of 0 to 1. However, there is a fuzzification parameter m which determines the degree of fuzziness in the clusters. M is in the range of [1, n]. The algorithm works like crisp partitioning algorithm when m reaches the value of 1 and if m is larger, then the overlapping of clusters is tend to be more. The algorithm is based on minimization of the following objective function:

$$J_m = \sum_{i-1}^{N} \sum_{j=1}^{c} u_{ij}^{m} \| xi - cj \|^2 \qquad 1<=m<=\infty$$

Let $X = \{x_1, x_2, x_3 \ldots, x_n\}$ be the set of data points i.e. the normalized read count data and $C = \{c_1, c_2, c_3 \ldots, c_n\}$ be the set of centers.

1) In the 1st step we have randomly selected c cluster centers.

2) Next, we have calculated the fuzzy membership μ_{ij} using the following formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \frac{||xi-cj||}{||xi-ck||}^{2/(m-1)}}$$

Where,

m = Fuzziness Exponent (any real number greater than 1)

C= the number of clusters i.e. 3,

x_i = the data i.e. the corrected read count value,

c_i = the center of the cluster

$||x_i - c_i||$ = the Distance from point i.e. the corrected read count value x_i to current cluster centre c_i.

3) In the next step, the cluster center will be updated by the following formula.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} xi}{\sum_{i=1}^{N} u_{ij}^{m}}$$

4) Next, we have to repeat the step 2 and 3 until the objective function is minimized.

III. RESULTS

In our approach, we have generated a reference sequence of 100000 bp of ATCG. We have generated 5 samples in our method. In 3 samples we have added duplications and deletions. In the rest of the 2 samples we did not add any kind of copy number variations. We have added two duplications in the first sample from position 2000 bp-3500 and 6000 bp-7500 bp accordingly. In the next sample, we had introduced one duplication from position 10000 bp – 12500 bp. In the 3rd sample we have added one deletion event from position 3000 bp – 5700 bp. the short reads are generated from the sample and the size are fixed to 36 bp. we have considered the 100 bp fixed window size. We have considered the match score as +5 and the mismatch score as -1 during alignment. Next, for each and every window we have calculated the read count value. In the fig 2 the x axis denotes the window and y axis denotes the corresponding read count value. We observed from the diagram that, the read count value is high for those windows where the duplications were introduced during simulation.

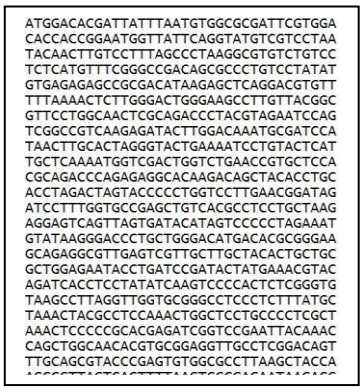

Fig 1: samples are divided into 36 bp short reads.

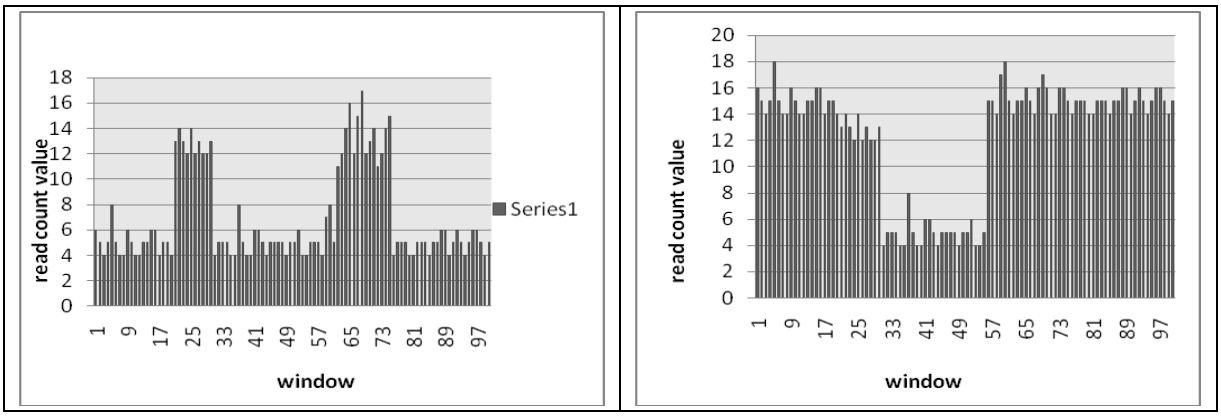

Fig 2: An instance of read count data, the copy number variation as duplication was introduced in the genomic segment 2000 bp-3500 and 6000 bp-7500 as represented. Fig 3: the copy number variation as deletion was introduced in the genomic region 3000 bp – 5700bp.

Then the GC content and mapppability bias were corrected accordingly. For correction of mappability bias we have used a k-mer based approach and considered as k=36 mers. After smoothing the read count data, we have applied k-means clustering for dividing the genomic regions. We have considered as k=3. The maximum number of iteration is fixed as k=10. However, the overlapping clusters and the outlier points cannot be handled by k-means clustering algorithm. To recover from this we have used fuzzy c means clustering algorithm keeping the same input parameter. Here we have considered the fuzziness exponent as m=2.

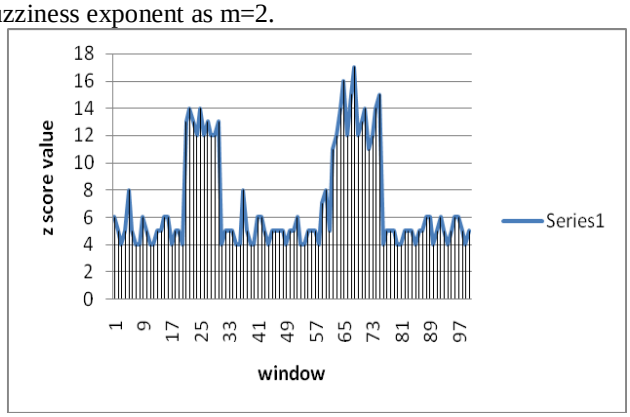

Fig 4: the Z-score c corresponding to the read count value where the X axis denotes the Window and Y axis denotes the corresponding Z-score

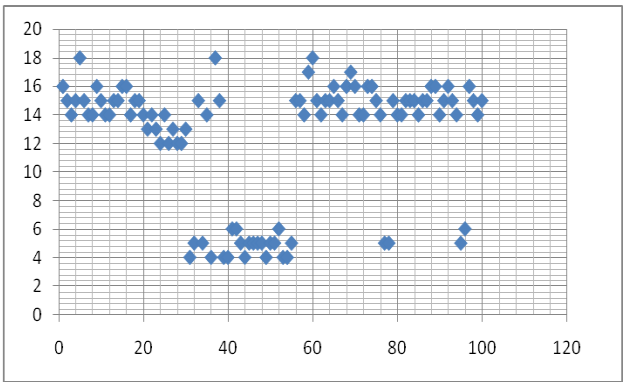

Fig 5: The result of k-means clustering. We have added two duplications here.

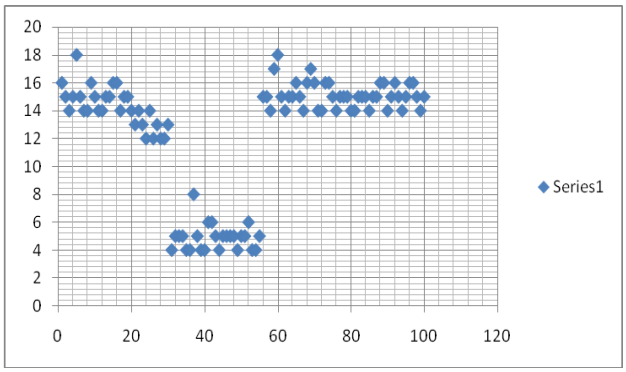

Fig 6: The result of fuzzy c-means clustering keeping the same input parameter

IV. COMPARISON

We have compared the performance of our algorithm with two existing algorithm. The algorithm includes EWT (event wise testing) [11] and CMDS (correlation matrix diagonal segmentation) [16]. EWT used event wise testing and implements a CNV detection tool which is read depth based called RDXplorer. First, the coverage or RD in non overlapping intervals across an individual genome is estimated. Then, a CNV calling algorithm is used to detect the events. We have implemented EWT by considering the z score of GC-corrected read-count data. The size of the consecutive window was fixed by 100 bp. We compared our method to the CMDS (correlation matrix diagonal segmentation) [16]. m physically ordered chromosomal sites of n individual samples are given as an input of CMDS. Thereafter, a Pearson's correlation coefficients were calculated. The final outcome of the correlation will be zero if no copy number alteration exists across the samples. However the positive correlation is observed if the copy number alteration exists across samples. As a result a square block along the diagonal between t he sites within the region is observed. By searching square blocks along the diagonal, we can detect the copy number alteration region.

We have measured the quality of the output of our method by using several standard measures like, precision, sensitivity, F-Score and markedness. the proportion of correctly detected CNVs in a genome is measured by sensitivity and is calculated by using the formula (TP/TP+FP) the proportion of truly detected CNVs with respect to the total number of detected CNVs is called precision or positive predictive value and is calculated using the formula (TP/TP+FN). F- Score denotes the quality of validation of detected CNVs with respect to true CNVs. The F-score is evaluated by using the expression [2 *(precision*sensitivity/ precision +sensitivity). The markedness was calculated by using the formula (Positive Predictive Value + Negative Predictive Value − 1).), the proportion of correctly undetected regions (True Negatives) is called the negative predictive value. The value of all this statistical measure should be in between [0, 1]. A high value of F-score and markedness will give the better predictive power in our classification procedure.

Precision, Sensitivity and F-Score of our method and compared methods:

Our method	**Precision**	**Sensitivity**	**F-Score**	**Markedness**
CMDs	0.62	0.715	0.66	0.59
EWT	0.82	0.74	0.78	0.56
Our Method using K-means	0.78	0.80	0.79	0.62
Our Method using fuzzy c-means	0.84	0.833	0.83	0.64

V. CONCLUSION

In our paper, we have described a novel computational approach for detection of copy number variations based on next generation sequencing technology. We have used Smith Waterman algorithm for the alignment of short reads. The read count data may be affected by GC content and mappability bias. GC content is especially higher in the protein-coding regions. As a result, this kind of systematic biases could lead to a false detection of the read count. For this purpose, the removal of bias is necessary to determine real enrichment patterns. Furthermore, the Mappability bias is also the major bias that can affect the copy number variation. Due to the short read length a small number of reads are mapped to multiple positions of windows. This can lead to spurious results, because mappability was correlated with certain biological features. The amount of non unique or non singleton sequence in the genome directly affects the read alignment. We choose the window size as 100 for a variety of reason. Detection of small CNVs could become problematic due to the larger window size because in many cases these CNVs would only partially span one or two windows. Next, we have used k-means clustering to partition the genomic regions. However, to identify the genomic regions, we have considered three clusters here. One for duplications i.e. for high value of read count data, One for deletions i.e. for low value of read count data, and next for the region where no copy number variations is observed. The main limitation of the algorithm is that the data points can be always a member of one of the clusters. However, when the data is well structured this works well but in real data have extraneous data points which are not belonging to any of the clusters and they are called outliers. Furthermore, the overlapping clusters and the outlier points cannot be handled by k-means clustering algorithm. To recover from this kind of limitations we have used fuzzy c means clustering. We compared our method to the CMDs and EWT. Our method with fuzzy c means clustering gives 83% sensitivity in detecting the copy number variations. Furthermore, the F-Score of our method using both K-means clustering and fuzzy c means clustering is comparatively higher than CMDS and EWT.

REFERENCES

[1] Lai WR, Johnson MD, Kucherlapati R, Park PJ, "C omparative analysis of algorithms for identifying amplifications and deletions in array CGH data", Bioinformatics 2005, 21:3763-3770

[2] Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES, "High-resolution mapping of copy-number alterations with massively parallel sequencing" Nat Methods 2009, 6:99- 103

[3] Smith waterman, Waterman MS, " Identification o f common molecular subsequences" J MolBiol 1981, 147:195-197

[4] [Qunyuan Zhang, Li Ding, "CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data", Advance Access publication December 23, 2009, Vol. 26 no. 4 2010, pages 464–469.

[5] Mardis ER," The impact of next-generation sequencing technology on genetics", Trends Genet 2008, 24:133-141

[6] Koehler R, Issac H, Cloonan N, Grimmond SM "Th e uniqueome: a mappability resource for short-tag sequencing", Bioinformatics 2011, 27:272– 274.

[7] Albertson DG, Collins C, McCormick F, Gray JW, "Chromosome aberrations in solid tumors", Nat Genet 2003, 34:369-376

[8] Frohling S, Dohner H," Chromosomal abnormalitie s in cancer", N Engl J Med 2008, 359:722-7

[9] Derrien T, Estelle J, Marco Sola M, Knowles DG, Raineri E, Guigo R, Ribeca, "Fast computation and applications of genome Mappability" ,PLoS ONE 2012, 7:e30377.

[10] Tae-Min Kim1, Lovelace J Luquette1, Ruibin Xi1, Peter J Park1,2,3*, "rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data", Kim et al. BMC Bioinformatics 2010, 11:432

[11] Seungtai Yoon,1 Zhenyu Xuan,1 Vladimir Makarov,1 Kenny Ye,2,3 and Jonathan Sebat1 1Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 2Albert Einstein College of Medicine, Bronx,New York 10461, USA, Sensitive and accurate detection of copy number variants using read depth of coverage

[12] Jorge S Reis-Filho, "Short communication Nextgeneration sequencing", Breast Cancer Research 2009, 11(Suppl 3):S12.

[13] Pinkel D, Albertson DG, " Array comparative genomic hybridization and its applications in cancer", Nat Genet 2005, 37(Suppl):S11-S17

[14] Lee H, Schatz MC: "Genomic dark matter: the r eliability of short read mapping illustrated by the genomemappability score", Bioinformatics 201 2, 28:2097–2105.

[15] Christopher, Miller1, Oliver Hampton, "ReadDep th: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads", PLoS ONE 6(1): e16327.

[16] Schraga Schwartz., Ram Oren, Gil Ast, "Detect ion and Removal of Biases in the Analysis of Next-Generation Sequencing Reads", January 31, 2 011, PLoS ONE 6(1): e16685.

[17] K. Alsabti, S. Ranka, and V. Singh " AnEfficientKMeans Clustering Algorithm.".

[18] Yinghua Lu 1 ∗, Tinghuai Ma1, , Changhong Yin2 , Xiaoyu Xie2 , Wei Tian1 and ShuiMing Zhong1 Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data

[19] JAMES C. BEZDEK Mathematics Department, Utah State University, Logan, UT 84322, U.S.A. ROBERT EHRLICH Geology Department, University of South Carolina, Columbia, SC 29208, U.S.A. WILLIAM FULL Geology Department, Wichita State University, Wichita, KS 67208, U.S.A. FCM: the fuzzy c means clustering algorithm.

[20] Min Zhao, Qingguo Wang, Quan Wang, PeilinJia, Zhongming Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives", IEEE International Conference on Bio informatics and Biomedicine Philadelphia, PA, USA. 4-7 October 2012.

