Genomics 102 (2013) 35–37

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# RF: A method for filtering short reads with tandem repeats for genome mapping

Kazuharu Misawa *,[1]

*Research Program for Computational Science, Research and Development Group for Next-Generation Integrated Living Matter Simulation, Fusion of Data and Analysis Research and Development Team, RIKEN, Japan*

ARTICLE INFO

*Article history:*
Received 8 May 2012
Accepted 6 March 2013
Available online 29 March 2013

*Keywords:*
Tandem repeats
Human genome
Mapping
Next-generation sequencing

ABSTRACT

Next-generation sequencing platforms generate short (50–150 bp) reads that can be mapped onto the reference genome. Repetitive sequences in the genome, because of the presence of similar or identical sequences, cause mapping errors in the case of the short reads. By filtering short reads with repeats, mapping will be improved. I developed RF. RF is a new method that filters short reads with tandem repeats. A scoring scheme was developed that assigned higher scores to regions with tandem repeats and lower scores to regions without tandem repeats. In this study, RF was applied to filter out short reads with repeats, before short reads were mapped onto the same genomic contig by using a short read-mapping program. The result suggests RF improved the proportion of correctly mapped short reads on filtering the repeats. RF is a useful tool for reducing mapping errors of short reads onto reference genomes.



## 1. Introduction

Next-generation sequencing (NGS) platforms generate short (50–150 bp) reads which are then mapped onto the reference genome [1]. For computational tools that align NGS reads to a genome, the most commonly encountered problem arises when reads align to multiple locations. In humans, various types of repetitive sequences account for approximately 50% of the genome [2], and they consist of perfect, or slightly imperfect, copies of DNA motifs of variable lengths [3,4]. Repetitive elements in the genome, owing to the presence of similar or identical sequences, sometimes cause mapping error of the short reads [5].

By filtering short reads with repeats, mapping will be improved. When the whole genome sequences of a human individual were determined, Fujimoto et al. [6] excluded short repeat regions that were detected by Tandem Repeats Finder [7]. In their study, repetitive sequences were defined by RepeatMasker (http://www.repeatmasker.org/). Tandem Repeats Finder is a widely used program for detecting repetitive elements, and requires no prior knowledge of the repetitive elements. Tandem Repeats Finder allows variation of repeat length such that tandem repeats as well as interspersed repeats can be obtained. A large number of methods have been developed for detecting tandem repeats [8]. Some of these methods search for repetitive sequences by aligning the input genome sequence against a library of known repeats. For example, RepeatMasker [9] uses Repbase [10]. It is plausible, however, that unknown repeats exist in the human genome.

In this paper, I developed the RF (Repeat Filter). RF is a new method that filters short reads with tandem repeats. A scoring scheme was developed that assigned higher scores to regions with tandem repeats and lower scores to regions without tandem repeats. On the basis of the theory of extreme value [11], Karlin–Altschul statistics were used to test the significance of regions with tandem repeats identified by using the RF method. The effectiveness of the new method was compared to that of Tandem Repeats Finder.

## 2. Theoretical background of RF

### 2.1. Overview of RF

RF consists of two components; a detection component and a test component. The detection component defines a scoring scheme to detect candidate tandem repeats. It is worth noting that regions without tandem repeats may, by chance only, have high region scores as well. By using the score of the region, the test component classifies the region with/without repeats by estimating the probability that regions with a given score could appear by chance.

### 2.2. Detection component

The detection component defines a scoring scheme that assigns higher scores to regions with tandem repeats and lower scores to regions without a tandem to detect candidate tandem repeats.

The region score was defined as the sum of the individual pairwise nucleotide scores in a given alignment. Fig. 1 shows an example of a DNA sequence in the human genome. To find tandem repeats, the DNA sequence is compared to itself by shifting m nucleotides. If 2 nucleotides are the same, the score is 1, else the score is −1. As shown

*Abbreviation:* NGS, Next-generation sequencing.

* RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. Fax: +81 45 503 9555.
*E-mail address:* kazumisawa@riken.jp.
[1] Present address: Advanced Center for Computing and Communication, RIKEN, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan, Fax: +81 48 462 4634.

0888-7543 
http://dx.doi.org/10.1016/j.ygeno.2013.03.002

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Original read | G | A | C | A | C | A | C | T | C | N | N |
| Shifted read | N | N | G | A | C | A | C | A | C | T | C |
| Scores | -1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 |
| The Sum of Scores | | | 3 | | | | | | | | |
| | | | | 4 | | | | | | | |
| | | | | | 2 | | | | | | |
| MS | | | | 4 | | | | | | | |

**Fig. 1.** Region scores and maximal segment pairs. The region score is the sum of the individual nucleotide pair scores in an alignment. A DNA sequence is aligned to itself by shifting 2 nucleotides. A high region score indicates that the region might contain tandem repeats.

in Fig. 1, the sum of the scores was calculated for all possible regions. A high region score indicated that the region is likely a repeat because many pairwise nucleotides with high scores exist in that region.

A maximal segment (MS) is defined as the highest-scoring pair of identical-length segments chosen from the DNA sequence and the m-shifted sequence [12]. Because the boundaries of an MS are chosen to maximize its score, an MS may be of any length [12]. The MS score, which the new algorithm heuristically attempts to calculate, provides a measure of possibility (probability) that any pair of sequences lies within a repeat region. Our aim was to identify whole regions that are likely to be repeat regions. Therefore, a segment pair was defined as locally maximal if its score could not be improved by either extending or shortening both segments. The new algorithm can search all locally maximal segment pairs with scores above a predetermined cutoff value, which is a function of the significance level (see the next section). An important advantage of the MS measure is that recent mathematical results allow the statistical significance of MS scores to be estimated using an appropriate random sequence model.

### 2.3. Testing component

Karlin and Altschul [13] developed a theory for local alignment statistics for their BLAST search algorithm [13]. In the case of a BLAST search, the probability of the score follows the extreme value distribution [14]. In the case of locating genes [11], the score also fits this distribution well. The score defined in this study also appears to follow this distribution under the condition that there are no repeats. By using the distribution, we can determine the threshold value for the given significance level.

In this study, computer simulations were conducted to determine the threshold value using random computationally generated sequences. It is worth noting that variation in GC contents exists among human chromosomes [15]. For example, chromosome 19 has the highest GC content (48.34%), whereas chromosome 4 has the lowest (38.25%) GC content, and the average GC content is 41.01%. By computer simulation, 3 random sequences of 10 megabases in length were generated with GC contents of 38.25%, 40.01%, and 48.34%. From these generated sequences, regions with high scores were obtained using the algorithm described above. The number of high-scoring regions was then counted, and their scores were recorded.

Fig. 2 shows the relationship between the score and the number of MSPs that exceeded the score. In this study, various values of scores were used as threshold.

If tandem repeats with n length pass the test, the repeat regions are masked, and the DNA sequence is shifted by n + 1 nucleotides. By applying this procedure from n = 1 to n = 100, tandem repeats were obtained. The executable binary of the new algorithm is available at the following address: http://sourceforge.jp/projects/parallelgwas/releases/?package_id=13542.

## 3. Materials and methods

### 3.1. Genome data

I used genomic contig NT_011255.14 of human chromosome 19 of the NCBI human genome (Build 37) because the density of short tandem repeats of human chromosome 19 is relatively higher than that of other human chromosomes [16]. The length of NT_011255.14 is 7,286,004 nucleotides. For the mapping process, two types of data sets were prepared from the DNA sequence of contig NT_011255.14. One is a set of short reads, the other is the reference sequence. As short reads, the DNA sequence of the chromosome was cut into short reads with lengths of 70 bp. These reads were not overlapping.

### 3.2. Filtering short reads with repeats

By using Tandem Repeats Finder [7] version 4.04 and the new method, the short reads with repeats were filtered out. Filtering by Tandem Repeats Finder was conducted with the following options: match = 2, mismatch = 5, delta = 5, PM = 80, PI = 10, minscore = 30, and maxperiod = 200, as previously conducted by Frith et al. [17]. The new method was evaluated by using various threshold values.

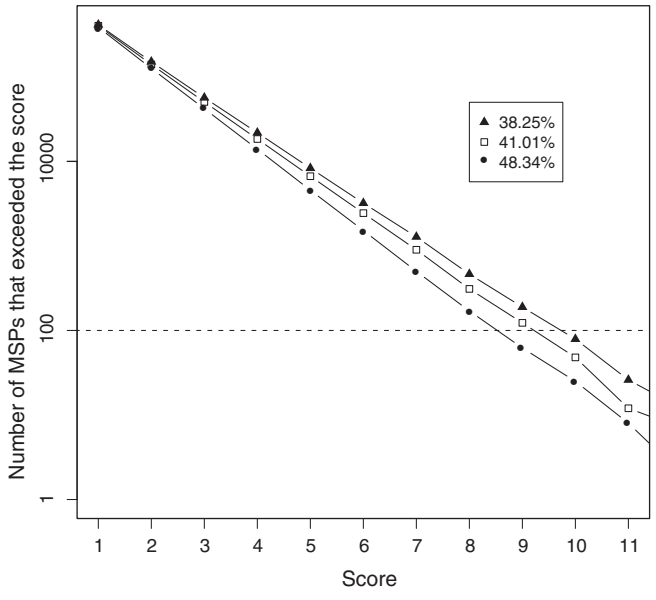


**Fig. 2.** The relationship between the score and the number of maximal segment pairs (MSPs) that exceed the computer simulation score. Closed triangles: GC content = 38.25%. Open squares: GC content = 41.01%. Closed circles: GC content = 48.34%. The dashed line denotes the significance level P = 0.01%.

**Table 1**
Effect of filtering of short reads that have tandem repeats to genome mapping.

| | Short reads that passed the filter | | Short reads that were mapped onto correct place of the genome | |
|---|---|---|---|---|
| | Number | Proportion | Number | Proportion |
| Without filter | 104,086 | 100.00% | 103,063 | 99.02% |
| Tandem Repeats Finder | 70,220 | 67.46% | 69,648 | 99.19%* |
| RF (threshold = 8) | 88,579 | 85.10% | 87,924 | 99.26%* |
| RF (threshold = 9) | 88,579 | 85.10% | 87,924 | 99.26%* |
| RF (threshold = 10) | 88,579 | 85.10% | 87,924 | 99.26%* |
| RF (threshold = 11) | 91,169 | 87.59% | 90,469 | 99.23%* |
| RF (threshold = 12) | 93,019 | 89.37% | 92,325 | 99.25%* |
| RF (threshold = 13) | 94,396 | 90.69% | 93,659 | 99.22%* |
| RF (threshold = 14) | 95,527 | 91.78% | 94,777 | 99.21%* |
| RF (threshold = 15) | 96,458 | 92.67% | 95,710 | 99.22%* |

* Significantly larger than the case without filtering (chi-square test, $P < 1\%$).

### 3.3. Mapping

Using a short read mapping program, BWA [18], the short reads were mapped onto the same genomic contig.

## 4. Result

Table 1 shows the number of short reads that passed the filtering by using Tandem Repeats Finder [7] and by using the new method. This table also shows the number of short reads that passed the filtering but were mapped onto the wrong place on the genome by using BWA [18]. The proportion of the short reads that were mapped on the correct position is significantly improved by using the filtering methods (chi-square test, $P < 1\%$). The proportion of short reads that passes the filter is 85.10% while that of Tandem Repeats Finder is only 67.46%.

The number of short reads that passed the filtering by Tandem Repeats Finder [7] and mapped onto an incorrect place of the genome was smaller than that obtained by the new method. However, the proportion of correctly mapped short reads after filtering by the new method was larger than that of Tandem Repeats Finder [7].

Table 1 also shows that the effect of the filtering by the new method depends on threshold value. As the threshold value increased, the number of the short reads that passed the filtering increased. The RF method processes the filtering procedure for a 7.2-megabase contig within 2 min on a machine with a 1.6-GHz Core 2 Duo processor.

## 5. Discussion

A new method for filtering short reads that have repeats was developed. The proportion of correctly mapped short reads was improved by filtering the repeats. The result shown above shows the effect of filtering repeats by using the RF method. It is worth noting that the threshold value is the only parameter. The same computational tool may yield different results due to user-adjustable parameters [19].

The ratio of true positives among all positives will be improved by RF as well as by Tandem Repeats Finder [7]. Tandem Repeats Finder [7] is a widely used program for detecting repetitive elements, and requires no prior knowledge of the repetitive elements. Tandem Repeats Finder allows variation of repeat length such that tandem repeats as well as interspersed repeats can be obtained. This result shown in this study indicates that Tandem Repeats Finder [7] is too sensitive to be used for filtering.

The RF method requires the fastq format for input. There is no restriction of the sequence length in the RF method. Patterns or sizes of repeats do not have to be specified. The output file format is also the fastq format.

## 6. Conclusion

A new method for filtering short reads that have repeats was developed. The ratio of true positives among all positives will be improved by the new method.

## Acknowledgments

I thank my wife for her encouragement and support. This study was supported by the "Next-Generation Integrated Living Matter Simulation," a national project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The results of the calculations were performed using the RIKEN Integrated Cluster of Clusters (RICC) and K computer at the RIKEN Advanced Institute for Computational Science.

## References

[1] C. Trapnell, S.L. Salzberg, How to map billions of short reads onto genomes, Nat. Biotechnol. 27 (2009) 455–457.
[2] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[3] B. Charlesworth, P. Sniegowski, W. Stephan, The evolutionary dynamics of repetitive DNA in eukaryotes, Nature 371 (1994) 215–220.
[4] O. Seberg, G. Petersen, A unified classification system for eukaryotic transposable elements should reflect their phylogeny, Nat. Rev. Genet. 10 (2009) 276.
[5] T.J. Treangen, S.L. Salzberg, Repetitive DNA and next-generation sequencing: computational challenges and solutions, Nat. Rev. Genet. 13 (2012) 36–46.
[6] A. Fujimoto, H. Nakagawa, N. Hosono, K. Nakano, T. Abe, K.A. Boroevich, M. Nagasaki, R. Yamaguchi, T. Shibuya, M. Kubo, S. Miyano, Y. Nakamura, T. Tsunoda, Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing, Nat. Genet. 42 (2010) 931–936.
[7] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573–580.
[8] P.C. Sharma, A. Grover, G. Kahl, Mining microsatellites in eukaryotic genomes, Trends Biotechnol. 25 (2007) 490–498.
[9] A.F.A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0, , 1996–2004.
[10] J. Jurka, Repbase update: a database and an electronic journal of repetitive elements, Trends Genet. 16 (2000) 418–420.
[11] K. Misawa, R.F. Kikuno, GeneWaltz—a new method for reducing the false positives of gene finding, BioData Min 3 (2010) 6.
[12] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[13] S. Karlin, S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 2264–2268.
[14] S.F. Altschul, R. Bundschuh, R. Olsen, T. Hwa, The estimation of statistical parameters for local alignment score distributions, Nucleic Acids Res. 29 (2001) 351–361.
[15] K. Misawa, A codon substitution model that incorporates the effect of the GC contents, the gene density and the density of CpG islands of human chromosomes, BMC Genomics 12 (2011) 397.
[16] K. Naslund, P. Saetre, J. von Salome, T.F. Bergstrom, N. Jareborg, E. Jazin, Genome-wide prediction of human VNTRs, Genomics 85 (2005) 24–35.
[17] M.C. Frith, M. Hamada, P. Horton, Parameters for accurate genome alignment, BMC Bioinforma. 11 (2010) 80.
[18] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler transform, Bioinformatics 26 (2010) 589–595.
[19] H.Z. Girgis, S.L. Sheetlin, MsDetector: toward a standard computational tool for DNA microsatellites detection, Nucleic Acids Res. 41 (2013) e22.