Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(11):200-204

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A method for enhancement of short read sequencing alignment with Bayesian inference

Weixing Feng*, Fengfei Song, Yansheng Dong and Bo He

Department of Automation, Harbin Engineering University, Harbin, China

ABSTRACT

Next-generation short read sequencing is widely utilized in genome wide association study. However, as an indirect measurement technique, short read sequencing requires alignment step to map all sequencing reads to reference genome before acquiring interested genomic information. Facing to huge quantity of whole genome sequences, efficiency of sequence alignment becomes a key factor to achieve satisfied sequencing result. To improve efficiency of short read sequencing alignment, we propose a method to retrieve more reliable aligned reads with Bayesian inference. In detail, we firstly deduce the occurrence pattern of sequencing errors from mismatches collected in already aligned reads. Then, in addition to already aligned reads, with the deduced pattern, we evaluate alignment reliability of the left reads and eventually extract more reliable aligned reads to improve alignment efficiency. The experiment with E. coli short read DNA sequencing data proves the validity of the proposed method.

Keywords: Next generation sequencing, Short read sequencing, Bayesian inference, Alignment, Method.

INTRODUCTION

In recent years, next-generation sequencing has become a popular information retrieving technique in molecular biology research for comprehensive characters of high resolution, precision and sensitivity [1,2]. Among multiple next-generation sequencing technologies, short read sequencing is able to produce large amounts of sequencing reads to cover whole genome and is widely utilized in genome wide association study (GWAS)[3]. However, next-generation sequencing requires alignment procedure to map all sequencing reads to reference genome before acquiring interested genomic information, which always causes happening of false negative sequencing results, especially in genome wide short read sequencing analysis[4]. To resolve such problem, deep sequencing is necessary to obtain satisfying result, but leads to rapid increase of cost and storage burden. Therefore, it is significant to improve efficiency of alignment procedure to locate more sequencing reads to reference genome.

By now, there are several procedures such as Bowtie and BWA normally used in short read sequencing reads alignment [5]. The basic consideration of these procedures is how to make short sequencing reads statistically aligned to reference genome in a reliable way. Although it is beneficial to use longer sequencing reads in the alignment, the sharply rising of sequencing errors along with increase of sequencing length limits the length of short sequencing reads less than tens of nucleotides and makes short read alignment a challenging job. Eventually, considering both the length of reads and the existence of sequencing errors, in all of short read alignment algorithms, except the perfect mapping with complete match, the mapping with few mismatches is also believed to be reliable alignment. However, except such rough consideration of sequencing errors, status of sequencing errors in a specific sequencing environment is barely considered. In this article, we just try to make use of such specific information to retrieve more sequencing reads reliably located to reference genome and improve the efficiency of alignment. In detail, as measure errors are specifically related to sequencing environment and result mismatches in mapping, we firstly focus on the occurrence pattern of sequencing errors exists in a specific sequencing experiment from mismatches information. Then, besides aligned reads from the general alignment algorithm, we infer additional reliable aligned reads from the pattern of sequencing errors in Bayesian way, and eventually enhance the alignment result and improve the alignment efficiency.

EXPERIMENTAL SECTION

In short read sequencing alignment procedure, the alignment results are simply divided into two subgroups, matched subgroup M (none or no more than n mismatches) and unmatched subgroup U, where n is a small number and varies with length of short reads. In normal, data inside subgroup M is believed to be with highly reliable alignment for existence of none or few mismatches and be adopted in subsequent analysis, whereas the data inside subgroup U is just abandoned.

We further divide matched subgroup M into datasets M_0, M_1,…, M_n for owning 0, 1, … , n mismatches respectively.

1.1 Prior probability calculation of sequencing errors

As sequencing errors result occurrence of mismatches in alignment [6-9], we deduce the pattern of sequencing errors from mismatches information.

Firstly, all reliable mismatches from M_0, M_1,…, M_n datasets are collected. Then prior probability $P(A_{ij}|M)$ of each kind of sequencing errors is statistically calculated, where A_{ij} is the jth kind of sequencing errors at the ith position of sequencing reads, and M means such prior probability is deduced from reliable alignment data.

1.2 Posterior Probability Calculation of sequencing errors

With prior probabilities of each kind of sequencing errors, we calculate posterior probability of a sequencing alignment based on Bayesian inference[10]:

$$P\left(M \mid A_{ij}\right) = P\left(A_{ij} \mid M\right) * P\left(M\right) / P\left(A_{ij}\right) \tag{1}$$

In the equation, $P(M|A_{ij})$ means the probability to believe one alignment reliable if the sequencing error A_{ij} exists. P(M) and $P(A_{ij})$ are two general probabilities of match result M and sequencing error A_{ij} which can be deduced from the whole sequencing alignment data.

1.3 Evaluation score of sequencing alignment reliability

When more than one sequencing errors exist in an alignment, considering independence of occurrence of each sequencing error, we propose a score E to evaluate reliability of the sequencing alignment:

$$E\left(M \mid B_1, B_2, \ldots, B_n\right) = \prod_{k=1}^{n} P\left(M \mid B_k\right) \tag{2}$$

In the equation, a number of k sequencing errors exist in the alignment, and B_k means the kth kind of sequencing errors.

1.4 Enhancement of sequencing alignment efficiency

To improve the alignment efficiency, we calculate the evaluation score E of sequencing alignment reliability inside subgroup U, where the data is simply abandoned in normal procedure. If a read in U is correctly aligned, its evaluation score will be high for following the same occurrence pattern of mismatches deduced from reliable alignments. Then, the evaluation scores of reads in subgroup U would be separated in bimodal distribution, and the reads with higher E values can be moved into reliable subgroup M. Eventually, in addition to the aligned reads with general alignment procedure, more reliable aligned reads are rescued with our proposed method.

RESULTS AND DISCUSSION

1.5 Alignment of sequencing data

We obtain E. coli short read DNA sequencing data from Sequence Read Archive database (SRA) in NCBI website (ERR015950), which is measured with Illumina Genome Analyzer II and is 36bps of length.

Then, we map all the sequencing data to E. coli reference genome with Bowtie. Among the whole 19,249,204 sequencing reads, 16,408,131 (85.2%) of them are matched with none or few mismatches (no more than two mismatches), and such matches are believed to be in high reliability. Then, we cluster these 16,408,131 reads as matched subgroup M, and the left 2,841,073 reads as unmatched subgroup U. Inside subgroup M, we further respectively classify 9,988,144 (60.9%), 4,518,585 (27.5%), and 1,901,402(11.6%) reads as M_0, M_1, and M_2 datasets for owning none, first and two mismatches.

Actually, among all the sequencing reads, only a little more than half of reads (9,988,144, 51.9%) are perfectly mapped to reference genome with no mismatch (M_0). However, considering existence of sequencing errors, the mapping with few (no more than two) mismatches are also believed to be reliable (M_1 and M_2).

As occurrence of sequencing errors is dynamically related to specific sequencing environment and follows unique pattern, we expect to detect such pattern from mismatches inside M_1 and M_2 datasets.

1.6 Prior and posterior probabilities of sequencing errors

We collect all mismatches from M_1 and M_2 datasets and analyze the pattern of sequencing errors. In Fig.1, frequencies of sequencing errors at each position are shown. While the frequencies keep similar in the first 25 positions, the frequencies sharply begin to rise in the following positions and reach the highest at the last position. Such phenomenon is in accordance with the mechanism of short read sequencing technique and explains why the length of short sequencing read is limited to tens of nucleotides.

Besides, in Fig.1, we also find cytosine is easier to be error-sequenced than the other three nucleotides at all positions.

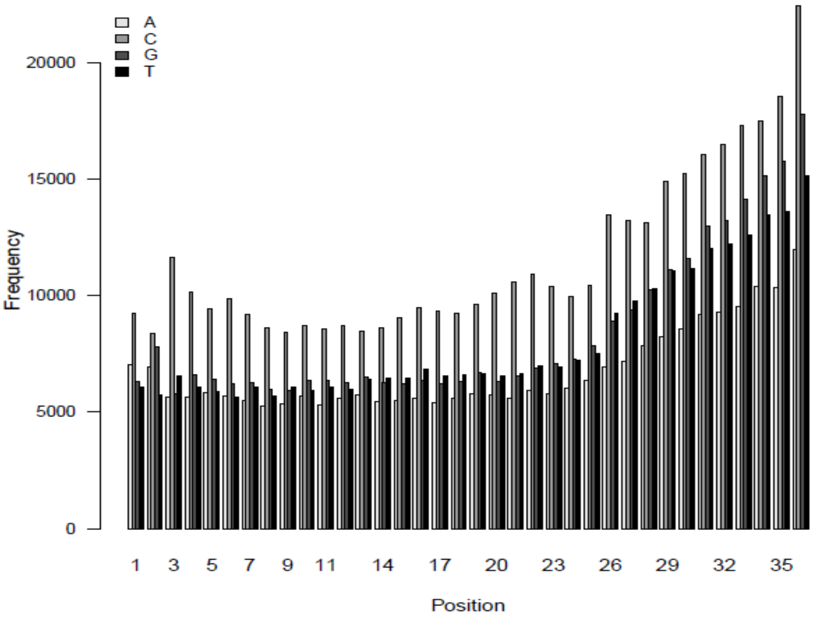

Fig.1 Frequencies of sequencing errors at each read positon

From all of the above analysis, we find that unique pattern of short read sequencing errors indeed exists in a sequencing environment and can be reflected along with two factors as position of sequencing reads and type of sequencing errors. Then we calculate such pattern with prior and posterior probabilities of sequencing errors.

The E. coli short read DNA sequencing data has the length of 36bps, and owns 12 kinds of sequencing errors at each position, which means a total of 432 kinds of sequencing errors exist in the reads.

We firstly calculate the prior probability $P(A_{ij}|M)$ of each kind of sequencing errors, where A_{ij} is the jth (j=1, 2,…, 12) kind of sequencing errors at the ith (i=1, 2,…, 36) position of sequencing data, and M means such prior probability is deduced from reliable alignment data(M_1 and M_2). We find all kinds of sequencing errors exist in the E. coli short read sequencing data and the values range from 0.000592 to 0.005590.

Then we calculate the posterior probability $P(M|A_{ij})$, which means the extent to which to believe one alignment result reliable if the sequencing error A_{ij} exists. The values of $P(M|A_{ij})$ range from 0.035341to 0.333544.

1.7 Enhancement of sequencing alignment efficiency

We calculate the evaluation score E of sequencing alignment reliability for reads inside unmatched subgroup U. As only reads with three mismatches are provided by Bowtie, we focus on such 67,633 reads here. The E values of them range from 5.43705e-5 to 0.365844 and its distribution is shown in Fig.2. Obviously, the whole evaluation scores are in binomial distribution.

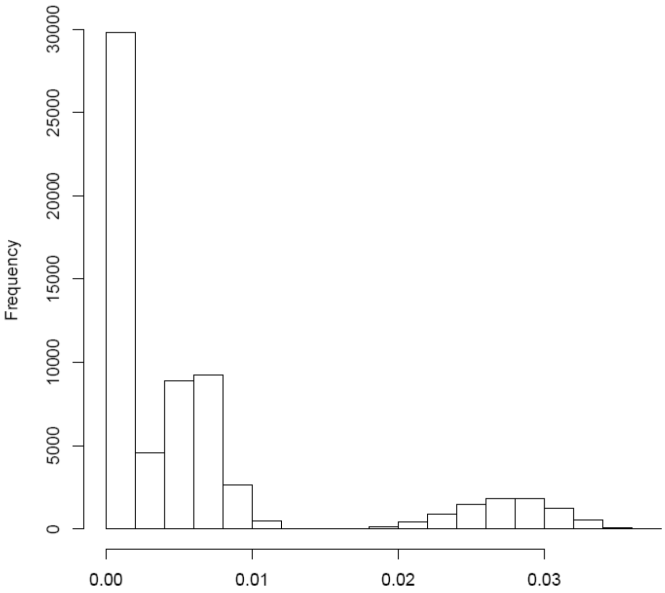

Fig.2 Distribution of evaluation scores of reads inside subgroup U

Eventually, the 8,352(12.3%) reads inside subgroup U with higher evaluation score values are extracted and moved into subgroup M, and the left 59,281(87.7%) reads keep inside subgroup U.

As aligned reads in M_0 are not used to deduce the pattern of sequencing errors, we analyze the relativities of these 8,352 and 59,281 reads to the reads inside M_0 respectively. We divide the whole genome into regions of kilo-nucleotides and calculate number of reads inside each region for 8,352 reads, 59,281 reads and reads inside M_0 respectively. The relativity between 8,352 reads and reads inside M0 is 72.1%, whereas it is only 35.5% for the left 59,281 reads. This analysis proves reads with reliable alignments are effectively retrieved from subgroup U with our proposed method.

CONCLUSION

Alignment is a key step in next-generation short read sequencing and directly affects the sequencing result. Generally, to obtain enough aligned reads to cover interested genomic regions, mappings with no more than few mismatches are also believed to be acceptable alignment. Nevertheless, the researchers still desire to achieve more aligned reads to get satisfied sequencing result, especially for short read sequencing utilized in genome wide association study.

In this study, with the consideration of specific status of sequencing errors in sequencing environment, we design a Bayesian-based method to retrieve additional aligned short sequencing reads and improve the efficiency of alignment. The effectiveness of our proposed method is proved in the following E. coli short read DNA sequencing experiment.

However, facing to complexity of sequencing alignment, more factors are still required to be considered to further improve the efficiency of alignment, especially for short read sequencing.

Acknowledgements
The research work was supported by National Natural Science Foundation of China under Grant No. 61071174 and Fundamental Research Funds for the Central Universities (HEUCFT1102, HEUCFX41303)

REFERENCES

[1] DA Wheeler. *Nature*, **2008**, 452(7189), 872–6.

[2] WJ Ansorge. *N Biotechnol*, **2009**, 25(4), 195-203.
[3] SH Wong; JJ Sung; FK Chan; KF To; SS Ng. *Semin Cancer Biol.*, **2013**, 10.
[4] N Whiteford; N Haslam; G Weber; M Bradley; C Neylon. *Nucleic Acids Res.*, **2005**, 33(19), e171.
[5] Ben Langmead, Cole Trapnell. *Genome Biol.*, **2009**, 10(3), R25.
[6] CD Juliane. *Nucleic Acids Res.*, **2008**, 36(16), e105.
[7] C Rakovski; DJ Weisenberger; P Marjoram; PW Laird; KD Siegmund. *BMC Bioinformatics*, **2011**, 12, 284
[8] L Mamanova; AJ Coffey. *Nat Methods*, **2010**, 7(2), 111-8.
[9] A Ratan. *PLoS ONE*, **2013**, 8(2), e55089
[10] Kevin Murphy. Machine Learning: A Probabilistic Perspective, 1st Edition, The MIT Press, Cambridge, **2012**; 65-72