論文の内容の要旨

Efficient frequency-based *de novo* short read clustering for error trimming in next-generation sequencing

(次世代シーケンサのリードエラーを取り除くための、頻度情報を用いた
短いリードを線形時間にデノボクラスタリングするアルゴリズム)

Wei Qu
氏名： 曲 薇

Novel massively parallel sequencing technologies provide a highly detailed structure of transcriptome and genome by yielding deep coverage of short reads, though their utility is interfered due to a considerable sequencing quality problem and short length of reads. Sequencing-error trimming in short reads is therefore a vital process which could improve the successful rate of reference mapping as well as polymorphorism detection.

There are several major drawbacks in the previous computational methodologies of correcting sequencing errors in short reads. First, a common feature of these methods is that reads have to be mapped uniquely to the reference, which may fail to detect the short erroneous reads aligned to unique but false positions (**Figure 1B**). Thus, a better detection of sequencing error should be performed before alignment to the reference genome. Second, another popular procedure is setting a minimum threshold on frequency in an ad-hoc manner to remove erroneous sequences originated in highly expressed sequences; however, this approach ignores sequences of low abundance at once and allows erroneous sequences with high frequency. Third, quality value selections such as Neighborhood Quality Standard (NQS) windows used for capillary sequencing do not match the next-generation sequencing strategy, which outputs bases at a same position in millions of reads simultaneously by an independent sequencing cycle. We need to exploit another sequencing error model tailored to the characteristics of next-generation sequencing.

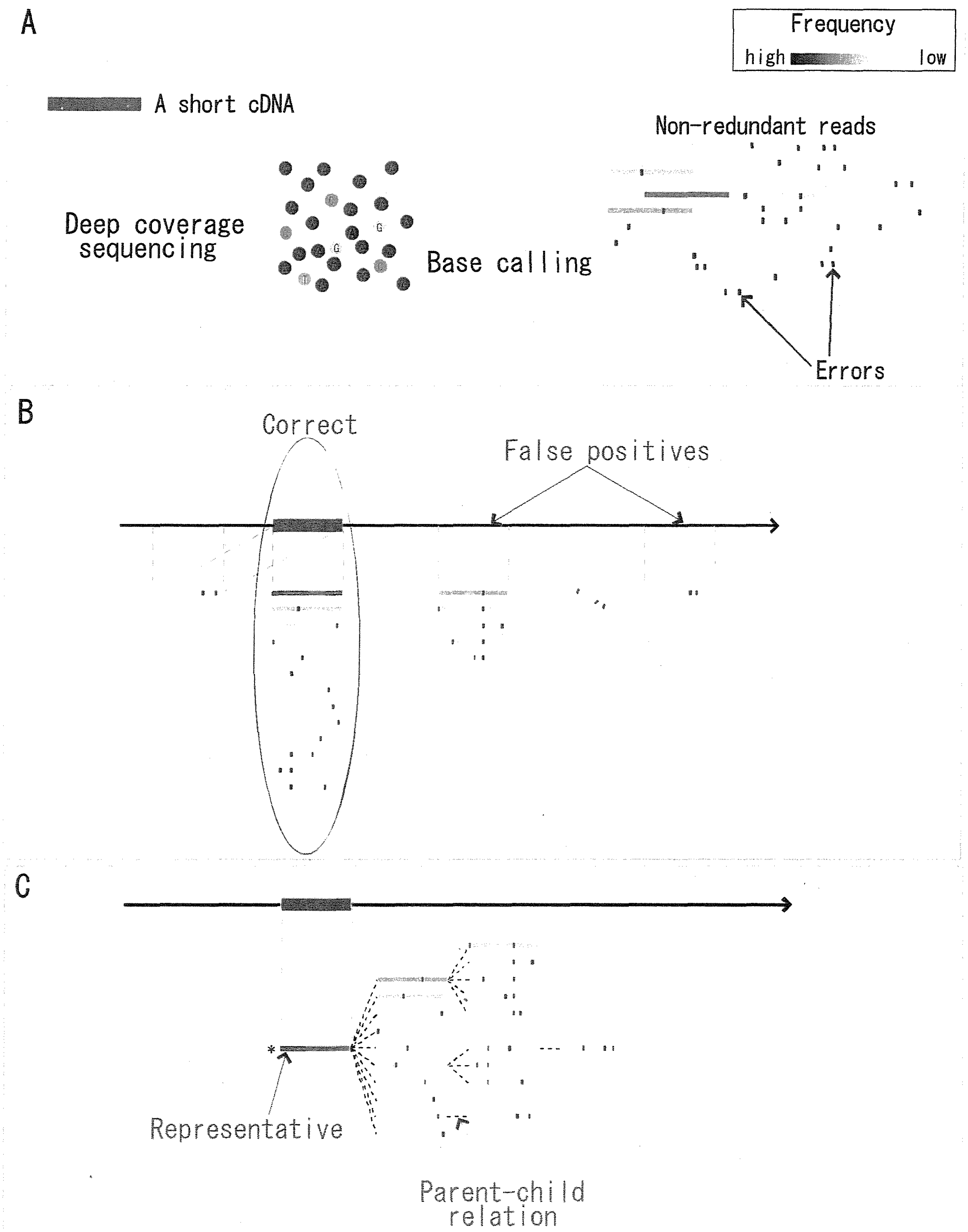

Figure 1. Illustration of the major benefit of *de novo* clustering. A real cDNA is shown as a brown bar and short reads originating in it are merged into non-redundant reads with unique sequences and presented by grey bars, whose contrasting densities are proportional to their frequencies. The reference genome is shown as a long arrowed line flagged with the corresponding locus of the cDNA by a brown bar. Alignments of best hits are highlighted by blue dashed lines. Red dots emphasize base positions where sequences disagree with the original cDNA sequence. **A.** A short cDNA is deeply sequenced with stochastically arising

sequencing-error. **B.** Direct alignment. Besides the correct alignments, some short reads have multiple best hits, as illustrated by the leftmost read; some fail in alignment due to too many sequencing errors, as shown by the aslant read; some are aligned to a false positive position. **C.** *De novo* clustering before the direct alignment. Erroneous reads are organized into the tree in the lower portion. The root indicates the representative sequence of the cluster, which is the darkest, most abundant read labeled with an asterisk in the upper part. In the tree, parent-children relationships are depicted by dashed lines.

In this thesis, I have addressed these issues and provide an effective solution. Before the alignment to the reference genome, erroneous 'child' sequences are clustered into a group represented by its 'parent' sequence, where a child sequence is considered to be stochastically originating in its more abundant parent sequence due to sequencing errors in the same experiment (**Figure 1C**). Indeed, mapping results show that broad parent-child relations inherently exist among reads generated from a same experiment. Subsequently, we integrate the parent-child relations among reads into trees such that the sequences at the root nodes are the most frequent sequences in individual trees and are treated as the representatives of all erroneous sequences in the trees. As illustrated in **Figure 1**, erroneous short reads that may be aligned to wrong positions or failing in mapping are clustered so that we map these representative sequences to the genome to anchor their locations. This approach is effective in resolving the low quality problem in short read sequencing because it avoids mapping erroneous reads to false-positive positions in the reference genome, and it eludes using an ad-hoc frequency threshold but outputs trees with reliable representative sequences regardless of abundance.

Since massive reads have to be clustered in a reasonable amount of time, we attempted to minimize the computational time of the clustering process. The crucial part of the algorithm is illustrated in **Figure 2**. The following three steps are repeated until the list of non-redundant sequences sorted by their frequencies in the descending order becomes empty:

1. Eliminate the bottom sequence of the lowest abundance, S, from the list.
2. Select the most frequent sequence S'' such that
   $Frequency(S'') = \max\{\, Frequency(S') \mid Hamming(S,S') = 1,\ Frequency(S') > Frequency(S) \,\}$,
   where the Hamming distance between two sequences of equal length S and S' $Hamming(S,S')$ is the number of positions where their bases are different.
3. Set the parent of S to S'' if S is proved to be derived from S'' due to sequencing errors according to a non-sequencing error statistical test.

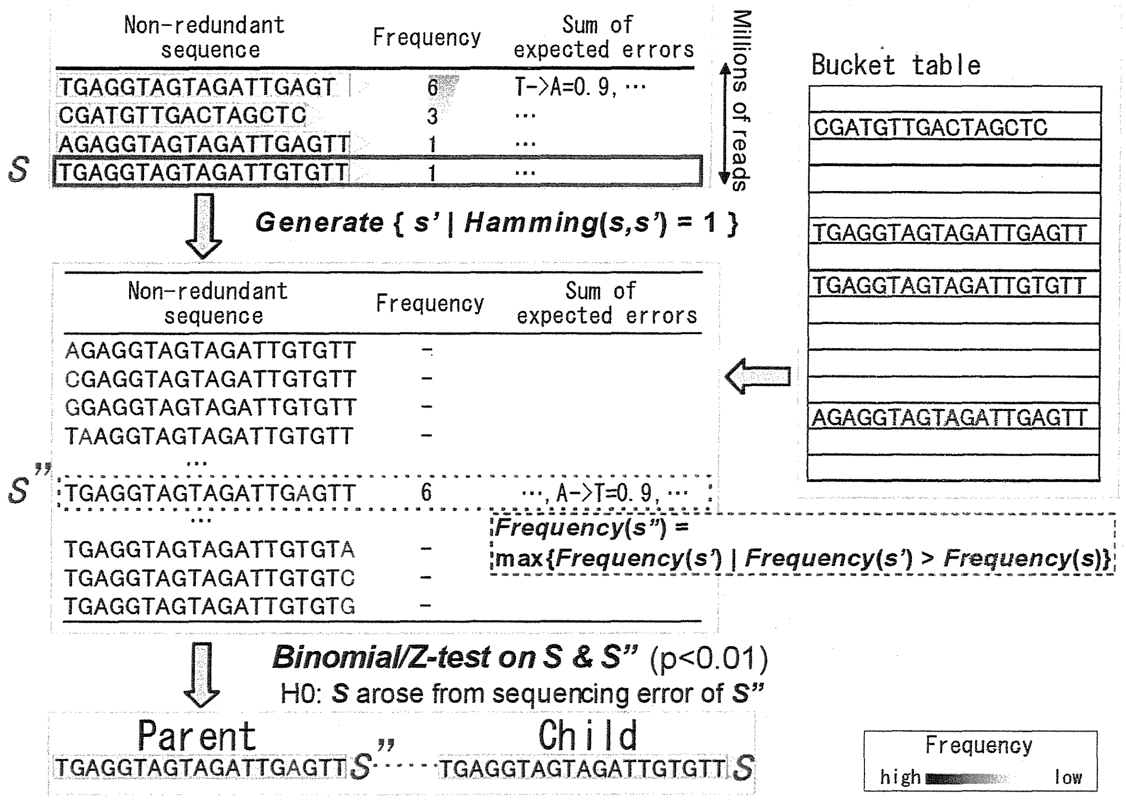

Figure 2. Algorithm of *de novo* clustering. The green list coupled with four sequences is a schematic view of millions of sequences.

Additionally, we developed an error model adjusted for next-generation sequencing by refining the traditional random model of error rate evaluation used in PolyBayes (Marth, G. T. *et al.* 1999) to involve substitution patterns arising due to fluorophore cross-talk noise factor. Our two experiments on small RNAs and 5'-end serial analysis of gene expression (SAGE) sequenced by Illumina/Solexa proved that our *de novo* clustering remarkably reduced sequencing errors. A remarkable increase (~5%) of short reads aligned to reference sequence was confirmed, especially a significant increase (relative raise of ~200%) is observed in the rate of reads with perfect match to reference sequence.

We proposed a frequency-based detection of parent-child relations accelerated with hash-based sequence searches that runs in time linear to the number of given reads. This method complements the ability of base callers and the error correction by direct alignment with the reference genome, and is able to improve the overall accuracy of short read alignment by consulting inherent relationships of the entire set of reads. As a perspective on its universal application, *de novo* clustering of color space reads of ABI/SOLiD short reads is considered to be viable as well.