Journal of Bioscience and Bioengineering
VOL. xx No. xx, 1–6, 2017

www.elsevier.com/locate/jbiosc

Estimation of the influence of sequencing errors and distribution of random-sequence tags on quantitative sequencing

Tatsuhiko Hoshino[1,2,*,‡] and Yohei Hamada[3,‡]

Geomicrobiology Group, Kochi Institute for Core Sample Research, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Nankoku, Kochi 783-8502, Japan,[1] Geobiotechnology Group, Research and Development Center for Submarine Resources, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Nankoku, Kochi 783-8502, Japan,[2] and Fault Mechanics Research Group, Kochi Institute for Core Sample Research, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Nankoku, Kochi 783-8502, Japan[3]

Received 15 February 2017; accepted 5 April 2017
Available online xxx

To simultaneously sequence and quantify target DNA, quantitative sequencing (qSeq) employs stochastic labeling of target DNA molecules with random-sequence tags (RSTs). This recently developed approach allows parallel quantification of hundreds of microorganisms in natural habitats in a single sequencing run. Yet, no study has addressed to what extent sequencing errors affect quantification and how many sequence reads are needed for quantification. Here, we addressed those issues by using numerical simulations and experimental data from second-generation sequencing of various RSTs. We found that heterogeneous distribution of observed RSTs affected the number of sequence reads required to quantitate target genes, whereas the effect of sequencing errors is smaller than of the RSTs distribution. Because of the heterogeneous RSTs distribution, 15-fold more sequence reads than the number of observed RSTs should be obtained to retrieve almost all RSTs needed for quantification; in that case, quantification error is estimated to be within 5%.

[**Key words:** Quantitative sequencing; 16S rRNA gene; Second-generation sequencing; Stochastic labeling; Random-sequence tag; Numerical simulation]

Recent advances in sequencing allow tens of millions of sequences to be obtained in a single sequencing run. Amplicon sequencing of the 16S rRNA gene provides a comprehensive view of microbial diversity in natural ecosystems (1–3). Although amplicon sequencing is a very powerful approach, the microbial composition in the obtained sequence library is biased because of the heterogeneity of amplification efficiency, which may be affected by primer sequences, GC content of amplified fragments, and the base adjacent to the primer (4–9). Therefore, additional assays are required for quantifying microbes represented in a sequence library.

Quantitative PCR (qPCR) has been widely used for quantifying microbes in various ecosystems (10–12). Conventional qPCR needs a standard curve based on the known DNA copy number; however, amplification efficiency of sample DNA may be lower than that of DNA standards because of impurities in the sample DNA (13). Digital PCR (dPCR) can circumvent this problem because it does not depend on external DNA standards and can estimate the absolute copy number of target DNA (14,15). Recently, dPCR has become the method of choice for a number of molecular biology applications; it provides more accurate quantification as it is less affected by PCR-inhibitory substances (16–18). Although qPCR and dPCR are powerful and sensitive tools, primer design and PCR conditions need to be optimized for each target gene (19–21). Therefore, parallel quantification of hundreds of microbial species in natural samples by qPCR and dPCR is not straightforward.

Quantitative sequencing (qSeq) can simultaneously sequence and quantify a large number of microbial species (22). The method employs stochastic labeling with random-sequence tags (RSTs) during first-strand DNA synthesis by single primer extension (SPE) followed by two PCR rounds to prepare templates for second-generation sequencing (23). By counting the variety of RSTs at the ends of sequence reads, the number of DNA molecules used as templates can be estimated together with sequence data for the target gene (e.g., 16S rRNA gene) that contain phylogenetic information. In this method, sequencing errors may result in overestimation of the number of template DNA molecules because of an increase in the number of observed RSTs. However, the extent of influence of sequencing errors and the number of sequence reads required for quantification have not been determined yet. Because of sequencing errors, increasing the number of sequence reads increases the number of counts from RSTs generated, and thus the number of sequence reads should be minimal but sufficient for retrieving almost all RST. In this study, we evaluated the effect of sequencing errors on qSeq by conducting numerical simulation analyses using experimental data, and we estimated how many sequence reads are needed for accurate quantification.

* Corresponding author at: Geomicrobiology Group, Kochi Institute for Core Sample Research, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Nankoku, Kochi 783-8502, Japan. Tel.: +81 88 878 2231; fax: +81 88 878 2192.
E-mail address: hoshinot@jamstec.go.jp (T. Hoshino).
‡ T.H. and Y.H. have contributed equally to this study.

1389-1723/$ – see front matter
http://dx.doi.org/10.1016/j.jbiosc.2017.04.003

Please cite this article in press as: Hoshino, T., and Hamada, Y., Estimation of the influence of sequencing errors and distribution of random-sequence tags on quantitative sequencing, J. Biosci. Bioeng., (2017), http://dx.doi.org/10.1016/j.jbiosc.2017.04.003

MATERIALS AND METHODS

Overview of quantitative sequencing The qSeq workflow is shown in Fig. 1. An SPE (single primer extension) primer consists of a target-specific sequence, RST (random octamer), and a adapter sequence for indexing. In the SPE step, RST is incorporated into the first-strand DNA to stochastically label each DNA molecule, and excess primers are digested with exonuclease I. In the first-round PCR, the SPE product is amplified with an SPE product–specific primer and a target-specific primer to obtain a sequencing template. In the second-round PCR, index sequences for Illumina sequencing are added to the amplicons from the first-round PCR. After sequencing, the variation of RST is determined by drawing a rarefaction curve, and the copy number of target DNA is then estimated based on Poisson statistics. In the current protocol, the number of sequence reads needed to retrieve almost all RSTs remains unknown.

Quantitative sequencing To determine the influence of error rate during qSeq, we used several RST variations: R1 ($N_{ini}=1$), R64 ($N_{ini}=64$), R256 ($N_{ini}=256$), R2916 ($N_{ini}=2916$), R11664 ($N_{ini}=11{,}664$), and R27648 ($N_{ini}=27{,}648$), where N_{ini} is initial RST variation (Table 1). Each SPE reaction mixture (20 μl) contained 1 ng of a nearly-full length 16S rRNA gene of *Bacillus subtilis* as a template, 0.3 μM SPE primer, and 1 × PrimeSTAR Max Premix (Takara Bio). Genomic DNA of *B. subtilis* was purchased from RIKEN DNA Bank (catalog no. JGD08099). SPE consisted of 10 cycles of 98°C for 15 s, cooling to 55°C at 0.3°C/s, 55°C for 1 min, and 68°C for 10 min. Although 1 cycle of SPE should be used to quantitate unknown DNA (22), here we used known DNA and performed 10 SPE cycles to ensure incorporation of all RST variations. Assuming that 1 ng of the rRNA gene fragment of *B. subtilis* is about 6.0×10^8 copies (= 50 pM) and the concentration of SPE primer is much higher than that, each SPE cycle generate 3.0×10^8 copies of the product even if reaction efficiency is low as 50%. Therefore, we can consider the copy number of SPE products to be high enough to use all variation of RSTs of the primers (i.e., 27,648 is the maximum in this study) after 10 cycles of SPE. Excess primers were digested with 4 μl of Exonuclease I (5 U/μl, Takara Bio) at 37°C for 2 h, and the enzyme was inactivated at 80°C for 30 min. First-round PCR mixture (25 μl) contained 12.5 μl of PrimeSTAR Max Premix, 0.3 μM each of F2 primer and N8-U806R primer, and 2 μl of SPE products. Amplification was performed for 40 cycles at 98°C for 10 s, 55°C for 5 s, and 72°C for 10 s; PCR products were purified by agarose-gel electrophoresis. PrimeSTAR MAX DNA polymerase is a high-fidelity enzyme with an error rate of only 0.04% after 30 cycles of PCR according to the information from the manufacturer, which is much lower than that of Illumina sequencing. In the second-round PCR, index sequences were added to the purified PCR products by using a Nextera index kit (Illumina) according to the manufacturer's specifications. For R1, we purchased synthesized DNA containing the 16S rRNA gene sequence of *Halalkalicoccus tibetensis* (AB663349) (Eurofins Genomics, Tokyo, Japan), all sequences needed for sequencing, and the R1 sequence at the end of the 16S rRNA gene sequence, therefore no amplification by PCR was needed for sequencing. Sequences were obtained using a MiSeq platform with a MiSeq Reagent Kit v3 for 600 cycles (Illumina).

Analysis of sequence reads Raw sequences were demultiplexed, quality-filtered using applications at the BaseSpace (Illumina), and further processed using the Mothur software (24). In brief, sequences derived from non-specific PCR products were screened by lengths and primer sequences after contigs were constructed. The RSTs for qSeq were selected from sequence reads, the distance matrix was calculated, and a rarefaction curve was generated for counting the variation of RSTs (see code in Fig. S1). The number of template DNA molecules (i.e., the number of primers) was estimated from the RSTs counts as described elsewhere (14,16,22).

Data analysis and simulation of quantitative sequencing First, the initial variation of RSTs was correlated with unique numbers in the initial sequence list (Fig. S2, step 1). The number of sequences in the initial sequence list was N_{ini} (e.g., 1, 64, 256). Second, one sequence was picked out according to a random number (from 1 to N; duplicates were allowed). The sequences were picked out x times (Fig. S2, step 3), which corresponded to the number of sequence reads, and the variation of picked sequences (V) was counted (Fig. S2, step 5). Assuming that no sequence errors occurred and the RSTs were randomly distributed, V would approach N asymptotically. To account for sequencing errors and biased distribution in the numbers of observed RST, sequencing error rate (p) and distribution of sequences were experimentally determined as described below.

RESULTS AND DISCUSSION

Number of sequence reads needed to retrieve all random sequence tags in the ideal case If there are no sequencing errors ($p=0$) and the observed numbers of all RSTs are equal, the rank–abundance curve is expected to approximately converge on

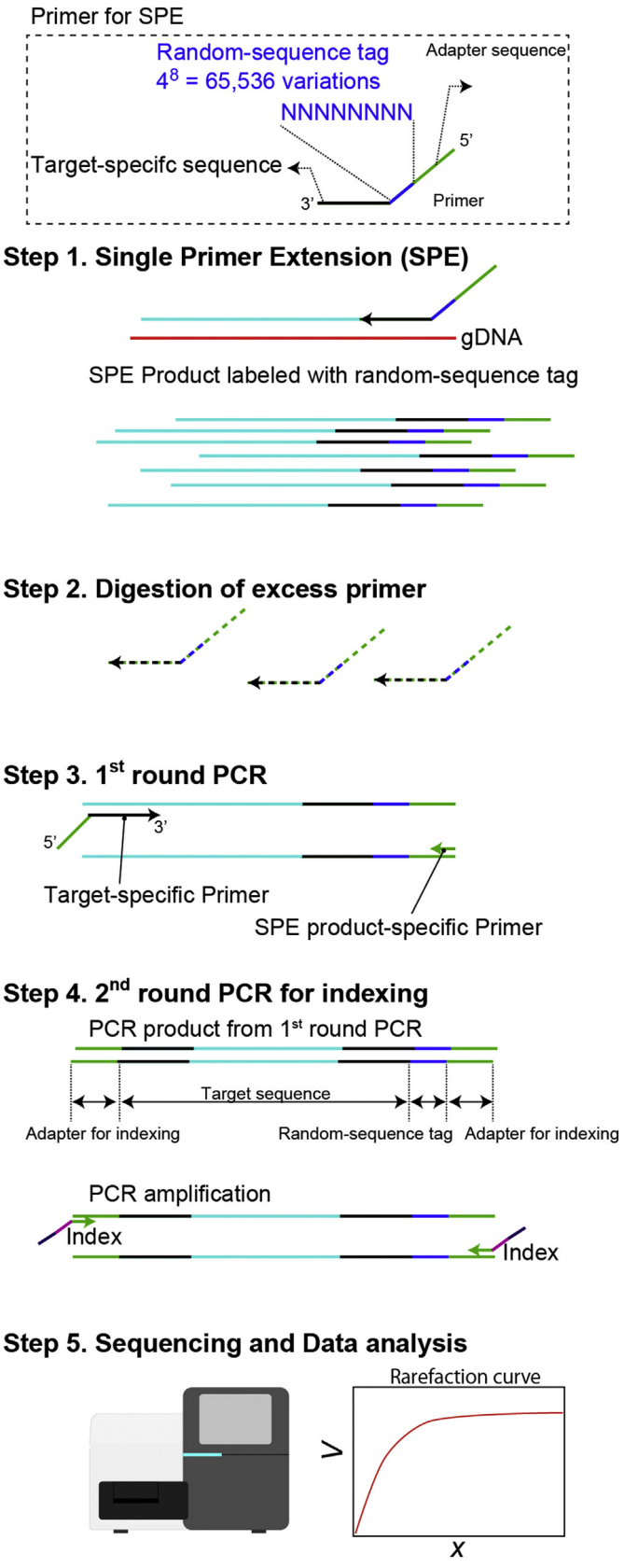

FIG. 1. Schematic workflow of quantitative sequencing (qSeq). In step 5, x indicates the number of sequence reads whereas V indicates the observed variations of random-sequence tags.