Benchmarks

Computation-assisted SiteFinding-PCR for isolating flanking sequence tags in rice

Hongru Wang^{1,2}, Jun Fang¹, Chengzheng Liang^{1,2}, Minghui He^{3,4}, Qiye Li^{3,4}, and Chengcai Chu¹

¹State Key Laboratory of Plant Genomics and National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China, ²Graduate School of the Chinese Academy of Sciences, Beijing, China, ³Beijing Genomics Institute at Shenzhen, Shenzhen, China, and ⁴School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China

BioTechniques 51:421–423 (December 2011) doi 10.2144/000113787

Keywords: SiteFinding-PCR; flanking sequences; chromosome walking; rice

Supplementary material for this article is available at www.BioTechniques.com/article/113787

SiteFinding-PCR is a method for isolating flanking sequence tags (FSTs) of T-DNA insertion lines, but the efficiency needs to be improved. Here we report a computation-assisted design for the random primers used in SiteFinding-PCR. A short sequence, GCATG, was screened from the rice genome and used as the 3' end of the random primer. When applying the optimized primer for isolating FSTs from 168 transgenic rice lines, we obtained 107 specific products, including 64 FSTs. The efficiency of obtaining FSTs using the modified version of SiteFinding-PCR increased by 73.0% compared with the method previously reported ($P < 0.01$, μ test). We also provide computational results for several other plant species such as maize, sorghum, *Arabidopsis*, foxtail millet, and *Brachypodium* based on the available genome data, so that the modified method could be easily adapted to other species.

T-DNA tagging is an important approach for functional genomics studies. In rice, more than 2 million insertion mutants are available around the world, but only about 200,000 inserts have been determined by flanking sequence tags (FSTs) due to the low efficiency and high cost of isolating FSTs (1). Although several PCR-based methods, such as inverse PCR (2), ligation-mediated PCR (3–5), and randomly primed PCR (6–8), have been developed for chromosome walking, the efficiency is still quite low for rice FST isolation (9,10).

SiteFinding-PCR (8), a randomly primed PCR, is a simple, inexpensive, and efficient method for chromosome walking. SiteFinding-PCR is initiated by a SiteFinder, a long primer with 4–6 bp-specific nucleotides

at its 3' end, which is annealed to the complementary sites of the genomic DNA. The target molecules are selected and amplified exponentially in the following two rounds of PCR with nested primers. SiteFinding-PCR has been successfully used for amplification of fragments from the *Arabidopsis* genome (8), however, the amplification efficiency in rice is quite low (our unpublished data). By using the computational approach described in this paper, we improved the amplification efficiency by optimizing 4–6 specific nucleotides at the 3' end of the SiteFinder.

We used all 4–6 bp random short nucleotide sequences with GC content higher than 50% and G/C at the end to scan the whole rice genome. For each short sequence, we calculated the distances between any two adjacent

loci in the genome. We then calculated the mean value (MD) based on the theoretical average length of the fragments when amplified from genome and the coefficient of variation (CV), calculated by dividing the standard deviation by MD. (This measures the uniformity of every short sequence distributed throughout the entire genome.) Based on these calculations, we selected specific random sequences to use as the 3' ends of the SiteFinders. The program for the calculation is written in PERL Language and is available in Supplementary Data 1; the results are provided in Supplementary Data 2.

Statistical analysis of all MD values from the short sequences reveals that most sequences have MD values comparable to the theoretical mean value 4ⁿ bp. For the 4–6 bp sequences, the theoretical mean values are 4⁴ bp, 4⁵ bp, and 4⁶ bp, respectively (Figure 1). However, big variations exist among individual short sequences. Table 1 lists 4–6 bp short sequences with the maximum and minimum MD values. Based on the calculation, the minimum MD value of a 5-bp sequence is barely 538 bp, whereas the maximum is 3314 bp. The CV values of all short sequences are relatively large, between 1 and 3 (Supplementary Data 2). This shows that all short sequences distribute unevenly in the rice genome because of the presence of many repeat sequences and transposable elements (11).

We selected a computationally optimized short sequence, GCATG (for SiteFinder S5-OP), which has a moderate MD value of 842 bp and a relatively small CV value as the 3' end of our SiteFinder. For comparison, we used SiteFinder S5-MX and SF-PU, respectively carrying “TAGCG,” which has the maximum MD value among all 5 bp sequences, and “GCCT,” which is a previously reported optimized short sequence (8) (Figure 2). Table 1 shows the parameters of these short sequences. We carried out a comparative evaluation of chromosome walking with these three SiteFinders.

Table 1. Short sequences and their distributing parameters in the rice genome

Short sequence	Site-Finder	MD (bp)	CV
CATG	N.U.	199	1.17
TACG	N.U.	635	1.24
GCCT	SF-PU	342	1.26
GCA TG	S5-OP	842	1.16
CGCCG	S5-MX	538	2.66
TACGC	N.U.	3314	1.13
CGCCGC	N.U.	1162	2.54
TAACGC	N.U.	18503	1.05

MD, mean value; CV, coefficient of variation; N.U., not used.

DNA samples from 168 rice T-DNA mutants (12) were used as templates to identify rice genomic sequences adjacent to the T-DNA left borders. We prepared template DNAs from T-DNA lines (12) with a DNA purification kit (Axygen, San

Francisco, CA, USA). All primers, including SiteFinders, SiteFinder primers, and gene-specific primers are shown in Figure 2. The reagents, PCR system, and thermal conditions were as described (8). PCR products were analyzed on a 1.5% agarose gel, and

specific products were recovered and purified using a DNA purification kit (RealTimes Biotechnology Ltd, Beijing, China). The purified PCR products were sequenced directly with an ABI 3730 XL system (Applied Biosystems, Foster City, CA, USA). For the sequencing results, we first aligned raw sequences with the T-DNA left border. Only those homologous to the border were referred to as specific sequences. Second, we performed a homology search against a rice genome database on National Center for Biotechnology Information (NCBI). The sequences that can be mapped to the rice genome are referred to as FSTs.

With SiteFinder SF-PU, we obtained 68 specific products, 37 of which mapped to the rice genome; with SiteFinder S5-MX, only 25 specific products and 15 FSTs were obtained. When using S5-OP, we obtained 107 specific products and 64 FSTs, increases of 57.4% and 73.0% ($P < 0.01$, μ test) over SF-PU, and 328.0% and 326.7% ($P < 0.01$, μ test) over S5-MX. Moreover, the average length of specific products was 764 bp, longer than that of SF-PU, which was 636 bp (Supplementary Data 3).

Jeong et al. obtained FSTs from T-DNA tag lines via inverse PCR (9). Compared with inverse PCR, modified SiteFinding-PCR is simpler, requires a much shorter experimental cycle, and may be manipulated automatically for high-throughput applications. Zhang et al. (10) used thermal asymmetric interlaced PCR (TAIL-PCR) and obtained 13,804 FSTs. However, they reported that only 30% of transformants generated PCR products with a single arbitrary degenerate (AD) primer in the first round of amplification, and the overall average length of all PCR products was 523 bp. Here, with modified SiteFinding-PCR, using a single SiteFinder to run one round of amplification, we obtained PCR products from 55% of the transformants ($P < 0.001$, μ test).

The random primer is a key factor in determining the success of chromosome walking with randomly primed PCR. Despite this, in the past, it has always been designed empirically (7,8,13). In this paper, we successfully developed and applied a computational approach for the design and evaluation of random primers and obtained accurate MD values of those random primers. Theoretically, large MD values mean long target products, which may impair the efficiency of SiteFinding-PCR, while small MD values mean more SiteFinding-PCR products, among which most are nontarget products that may mask the target products. In Supplementary Data 4, we provide computational results from the genome of maize, sorghum, *Arabidopsis*, foxtail millet, and *Brachypodium*, the model plant species with genome sequence available. Researchers who wish to do chromosome walking in these

Figure 1. Distribution of the MD values of all short sequences calculated in the study of the rice genome. The x-axis indicates the MD value; the y-axis indicates the number of short sequences. (A) Distribution of the MD values of 4-bp short sequences. On the bar chart, the value with a horizontal coordinate at 350 bp indicates the number of short sequences with MD values between 300 and 350 bp. (B) Distribution of the MD values of 5-bp short sequences. (C) Distribution of the MD values of 5-bp short sequences.

Figure 2. PCR Primers Used in This Study. (A) Sequences of SiteFinders used with primers indicated by black arrows. All SiteFinders differed only at their 3' ends and contained a NotI restriction enzyme site for cloning the SiteFinding-PCR products. The SF-PU initiates the SiteFinding at 25°C, and the S5-OP and the S5-MX at 30°C. (B) Three gene-specific primers, indicated by black arrows, were designed based on the T-DNA sequence. The distance from T2 to T3 was 63 bp. The T_m value was calculated with the formula $T_m(^{\circ}\text{C}) = 69.3 + 0.41(\text{GC content, \%}) - 650/L(\text{primer length})$.

Chill out with **BIOCOOLER**
bench top coolers
for biologicals
and reagents.

BioSmith.com
1.800.929.7894 858.693.3476

VISA MasterCard

species with SiteFinding-PCR could first try out an appropriate MD value in the species and then select an optimized short sequence with the MD value and the smallest CV value for their SiteFinder. To our knowledge, there has not yet been a report applying a computational approach for the design and evaluation of degenerate primers. But with the greater availability of genome information, researchers are now able to select better random primers for randomly primed PCR, which could improve the overall efficiency of the methods. Moreover, the computational approach can also be applied to inverse PCR and ligation-mediated PCR for evaluation of restriction site distribution in the genome, which aids researchers in selecting more suitable enzymes.

Acknowledgments

This research was supported by grants from the Ministry of Science and Technology (2009CB118506) and the National Natural Science Foundation of China (30825029, 30921061). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Krishnan, A., E. Guiderdoni, G. An, Y.I. Hsing, C.D. Han, M.C. Lee, S.M. Yu, N. Upadhyaya, et al. 2009. Mutant resources in rice for functional genomics of the grasses. *Plant Physiol.* 149:165-170.
2. Ochman, H., A.S. Gerber, and D.L. Hartl. 1988. Genetic applications of an inverse polymerase chain reaction. *Genetics* 120:621-623.
3. Rosenthal, A. and D.S.C. Jones. 1990. Genomic walking and sequencing by oligo-cassette mediated polymerase chain-reaction. *Nucleic Acids Res.* 18:3095-3096.
4. Jones, D.H. and S.C. Winistorfer. 1992. Sequence specific generation of a DNA panhandle permits PCR amplification of unknown flanking DNA. *Nucleic Acids Res.* 20:595-600.
5. Myrick, K.V. and W.M. Gelbart. 2007. A modified universal fast walking method for single-tube transposon mapping. *Nat. Protocols* 2:1556-1563.
6. Parker, J.D., P.S. Rabinovitch, and G.C. Burmer. 1991. Targeted gene walking polymerase chain reaction. *Nucleic Acids Res.* 19:3055-3060.
7. Liu, Y.G. and R.F. Whittier. 1995. Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics* 25:674-681.
8. Tan, G.H., Y. Gao, M. Shi, X.Y. Zhang, S.P. He, Z.L. Cheng, and C.C. An. 2005. SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. *Nucleic Acids Res.* 33:e122.
9. Jeong, D.H., S. An, S. Park, H.G. Kang, G.G. Park, S.R. Kim, J. Sim, Y.O. Kim, et al. 2006. Generation of a flanking sequence-tag database for activation-tagging lines in japonica rice. *Plant J.* 45:123-132.
10. Zhang, J., D. Guo, Y. Chang, C. You, X. Li, X. Dai, Q. Weng, G. Chen, et al. 2007. Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13 804 T-DNA flanking sequences from an enhancer-trap mutant library. *Plant J.* 49:947-959.
11. International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436:793-800.
12. Ma, Y.M., L. Liu, C.G. Zhu, C.H. Sun, B. Xu, J. Fang, J.Y. Tang, A.D. Luo, et al. 2009. Molecular analysis of rice plants harboring a multi-functional T-DNA tagging system. *J. Genet. Genomics* 36:267-276.
13. Liu, Y.G. and Y. Chen. 2007. High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. *BioTechniques* 43:649-654.

Received 2 August 2011; accepted 10 November 2011.

Address correspondence to Chengcai Chu, State Key Laboratory of Plant Genomics and National Center for Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China. e-mail: ccchu@genetics.ac.cn

To purchase reprints of this article, contact: biotechniques@foisterprinting.com

BioTechniques Poster Hall

BioTechniques

Our Unique Quarterly Collection Available Online

Sample Categories Include:

- Antibody-based Technologies
- Cell Signaling
- Cytotoxicity
- Electroporation
- Epigenetics
- Expression Profiling
- FFPE
- High-throughput Screening
- Mass Spectrometry
- Microarray Analysis
- Multilabel Detection
- Multiplex Analysis
- Plasmid DNA Purification
- Protein isolation & Purification
- Sample Preparation
- Transfection