2017

An accurate and efficient method for large-scale SSR genotyping and applications

Lun Li
Jianghan University

Zhiwei Fang
Jianghan University

Junfei Zhou
Jianghan University

Hong Chen
Center for Development of Science and Technology

Zhangfeng Hu
Jianghan University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
Li, Lun; Fang, Zhiwei; Zhou, Junfei; Chen, Hong; Hu, Zhangfeng; Gao, Lifen; Chen, Lihong; Ren, Sheng; Ma, Hongyu; Lu, Long; Zhang, Weiziong; and Peng, Hai, "An accurate and efficient method for large-scale SSR genotyping and applications.” Nucleic Acids Research.45,10. e88. (2017).
https://digitalcommons.wustl.edu/open_access_pubs/6172

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Authors
Lun Li, Zhiwei Fang, Junfei Zhou, Hong Chen, Zhangfeng Hu, Lifen Gao, Lihong Chen, Sheng Ren, Hongyu Ma, Long Lu, Weiziong Zhang, and Hai Peng
An accurate and efficient method for large-scale SSR genotyping and applications

Lun Li1,†, Zhiwei Fang1,†, Junfei Zhou1, Hong Chen2, Zhangfeng Hu1, Lifen Gao1, Lihong Chen1, Sheng Ren3,4, Hongyu Ma5, Long Lu1,3, Weixiong Zhang1,6,7,† and Hai Peng1,†

1Institute for Systems Biology, Jianghan University, Wuhan, Hubei 430056, China, 2Center for Development of Science and Technology, Ministry of Agriculture, P.R. China, Beijing 100122, China, 3Division of Biomedical Informatics, Cincinnati Children’s Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229-3026, USA, 4Department of Mathematical Sciences, McMicken College of Arts & Sciences, University of Cincinnati, 2815 Commons Way, Cincinnati, OH 45221-0025, USA, 5Thermo Fisher Scientific, Building 6, No. 27, Xin Jinqiao Rd., Pudong, Shanghai 201206, China, 6Department of Computer Science and Engineering, Washington University in St Louis, MO 63130, USA and 7Department of Genetics, Washington University School of Medicine, St Louis, MO 63130, USA

Received December 05, 2016; Revised January 27, 2017; Editorial Decision January 31, 2017; Accepted February 01, 2017

ABSTRACT

Accurate and efficient genotyping of simple sequence repeats (SSRs) constitutes the basis of SSRs as an effective genetic marker with various applications. However, the existing methods for SSR genotyping suffer from low sensitivity, low accuracy, low efficiency and high cost. In order to fully exploit the potential of SSRs as genetic marker, we developed a novel method for SSR genotyping, named as AmpSeq-SSR, which combines multiplexing polymerase chain reaction (PCR), targeted deep sequencing and comprehensive analysis. AmpSeq-SSR is able to genotype potentially more than a million SSRs at once using the current sequencing techniques. In the current study, we simultaneously genotyped 3105 SSRs in eight rice varieties, which were further validated experimentally. The results showed that the accuracies of AmpSeq-SSR were nearly 100 and 94% with a single base resolution for homozygous and heterozygous samples, respectively. To demonstrate the power of AmpSeq-SSR, we adopted it in two applications. The first was to construct discriminative fingerprints of the rice varieties using 3105 SSRs, which offer much greater discriminative power than the 48 SSRs commonly used for rice. The second was to map Xa21, a gene that confers persistent resistance to rice bacterial blight. We demonstrated that genome-scale fingerprints of an organism can be efficiently constructed and candidate genes, such as Xa21 in rice, can be accurately and efficiently mapped using an innovative strategy consisting of multiplexing PCR, targeted sequencing and computational analysis. While the work we present focused on rice, AmpSeq-SSR can be readily extended to animals and micro-organisms.

INTRODUCTION

Simple sequence repeats (SSRs), also known as short tandem repeats (STRs) or microsatellites, exist extensively in eukaryotic genomes. Most SSRs are non-coding and may affect gene expression, splicing, protein sequences and genome structures (1–5). The rate of SSR length mutations is estimated to be $10^{-7}$ to $10^{-3}$ per locus per generation in eukaryotes (6), which is much higher than the rate of $10^{-9}$ for base mutations (7,8) and accounts for the high diversity of SSRs. Despite the high variability of SSR sequences, their flanking regions are often conserved within the same species, occasionally among closely related species (9,10) and even across species (11,12). SSRs have several advantages over other genetic variations, including co-dominance, high reproducibility and requiring a small amount of template DNA for experiment (13–16). Importantly, the combination of the diversity of SSR sequences and the conservation of their flanking regions makes SSRs ideal genetic markers. Indeed, SSRs have been successfully adopted in various applications such as DNA fingerprinting, gene mapping, forensic analysis, marker assisted breeding and assessment of seed purity (16–20).

The most notable source of SSR diversity is the innate slippage of DNA polymerases during SSR replication (21–
Such a slippage is also inherent in in vitro SSR polymerase chain reaction (PCR) amplification, which results in erroneous SSR alleles and makes accurate SSR genotyping difficult. Moreover, gel electrophoresis, the currently most popular approach to detecting SSR PCR products, is inaccurate due to its low resolution and is inefficient as it can only handle a small number of SSRs at a time (20). For example, only 28, 25 and 48 SSR loci are analyzed by gel electrophoresis and used to construct DNA fingerprints of jute (Corchorus spp.) (28), winter mushroom (Flammulina velutipes) (29) and pigeonpea (30), respectively. Such limited SSRs are not sufficient or robust to construct high-quality SSR fingerprints to discriminate close kinships. Whole genome resequencing can profile a large number of SSR loci at once (25,31,32). Nevertheless, since SSR sequences only constitute a small percentage of a whole genome, e.g. ~3% of the human genome (33), whole genome resequencing dilutes the sequencing reads on SSRs with a huge number of other genomic reads, making it difficult to reach a required coverage of more than 10- to 100-folds (25) for accurate SSR genotyping with an acceptable cost. In addition, whole genome resequencing introduces additional problems to SSR genotyping, e.g. preferential amplification of specific SSR loci (34) and difficulties in data analysis with repeats in SSRs (35,36).

We developed a novel sequencing based SSR genotyping method by combining multiplex amplification of target SSRs, high-throughput sequencing of the amplicons (Ampli-Seq) and computational and statistical analyses. For convenience, we call our new method AmpSeq-SSR. AmpSeq-SSR overcomes nearly all difficulties in the existing methods. In particular, AmpSeq-SSR is able to efficiently genotype a large number of SSR loci at once with ultra-deep coverages and has an accuracy close to 100% and a single-base resolution. We applied AmpSeq-SSR to genotyping a total of 3105 SSRs in eight rice varieties and validated all of them. We further demonstrated the utility and power of the new approach with two additional applications. The first is construction of rice fingerprints that contain 449 differential SSRs on average, which can accurately distinguish a variety under test. The second application is mapping of Xa21, a gene that confers a broad and persistent resistance against rice bacterial blight (BB), a devastating rice disease causing a substantial annual rice yield reduction worldwide. While the development of AmpSeq-SSR and our current work focus on rice, the new method is general and applicable to animal species and eukaryotic micro-organisms.

**MATERIALS AND METHODS**

### Rice plants used

The eight homozygous rice varieties (A–H in Supplementary Table S1) are commercially released varieties in China and are representatives of indica and japonica rice plants. The three pairs of nearly isogenic lines (NILs, I-N in Supplementary Table S1) share similar genetic backgrounds except the Xa21 gene and its linkage regions.

**Target SSRs and design of multiplex primers**

Forty-eight SSRs that are listed in the National Agricultural Standard of China (NY/T 1433–3014) and 3057 randomly selected SSRs from the japonica reference genome (irgsp1.0) were chosen as target SSRs (Supplementary Table S2, the online file http://www.cse.wustl.edu/~zhang/SSR_ST2.pdf has the full list of primers). The service at https://ampliseq.com/ was used to design multiplex primers for 3105 target SSRs, which have amplicon lengths <250 bp on reference genome. The designed primers were synthesized by Thermo Company, USA.

**Construction and high-throughput sequencing of Ampli-Seq libraries**

We describe here the major steps for the construction and high-throughput sequencing of Ampli-Seq libraries. A step-by-step protocol is given in Supplementary Method. Whole plants at the first leaf stage were harvested for extraction of genomic DNA, following the protocol of E-Z96® Plant DNA Kit (Cat. No. M1027, Omega bio-tek, USA). Varieties A-N in Supplementary Table S1 were used to construct Applied Biosystems (ABI) S5 Ampli-Seq libraries according to the user guide of Ion AmpliSeq™ Library Kit 2.0 (Cat No. 4475345, Thermo, USA). The resulting libraries were sequenced on S5 system using the single-end sequencing with a length of 300 bp. Varieties A–H in Supplementary Table S1 were also used to construct Illumina MiSeq Ampli-Seq libraries according to the user guide with modification. That is, additional PCR with 14 cycles were introduced for DNA amplification. The resulting libraries were sequenced on MiSeq system using the paired-end sequencing with a length of 2 × 300 bp. The library construction and sequencing for S5 and MiSeq systems were respectively performed in our lab and BestNovo Co., Ltd, China within a 30-day interval.

**SSR genotyping by gel electrophoresis and Sanger sequencing**

A 25 µl PCR reaction system was used, which includes 0.05 μM primer R and 0.05 μM primer F, 10 ng template DNA and 12.5 μl AmpliTaq Gold® 360 Master Mix (4398876, Applied Biosystems™, USA). The PCR reaction procedure is 95°C/5 min; 40 cycles of 95°C/30 s and 60°C/1 min; 72°C/25 min. The amplicons were sent for Sanger sequencing (TsingKe Company, China) or were separated by 2% agarose gel electrophoresis, 6% polyacrylamide gel electrophoresis (PAGE) or capillary electrophoresis (CE). The CE was performed on ABI 3500 Genetic Analyzer (Applied Biosystems, USA) and the fragments were analyzed using GeneMapper software v4.1. The major and minor alleles of an SSR of CE were defined as the lengths of the largest and the second largest bands on electropherogram. The stutter ratio of an SSR was estimated as the ratio of the areas between the second allele and the major alleles of the SSR.

**Processing of sequencing reads and SSR genotype calling**

Here, we described the rationale and criteria for processing
sequencing reads from homozygous varieties for SSR profiling. The related scripts were provided in Supplementary materials. The methods for heterozygous varieties were provided in the last section of ‘Materials and Methods’ section.

Reads were mapped to the reference genome by Bowtie2 (37) to determine which target SSRs they belonged to. However, many reads were left unmapped because of the variations in SSR lengths. To resolve this problem, sub-sequences of 40 bp were taken from the unmapped reads by a scheme of a sliding window of 40 bp. The sub-sequences were then mapped to the reference genome by BLASTN to determine potential SSRs they belonged to. To avoid the interference of SSR length variations on mapping, the SSR sequences on the reference genome were replaced with ‘N’. After a read was located, the two boundaries of the SSR in the read were determined by aligning the read to the flanking sequences of the corresponding SSR on the reference genome (ref-SSR). Then, the tandem repeats of the ref-SSR motif between the two boundaries were determined as the SSR allele of the read.

The reads were tallied for each kind of allele at an SSR locus. The alleles with the largest and the second largest numbers of reads were designated as the major and minor alleles of the SSR locus, respectively. The major allele was also taken as the genotype because only a homoyzygous allele was expected for an SSR locus in the homozygous plants used in the current study. The ratio between the numbers of reads of the minor and major alleles was taken as the stutter ratio of the SSR locus.

To determine the amplicon length of a target SSR, the 20 bp upstream and downstream sequences of the amplicon on the reference genome were extracted and mapped to each sequencing read of the target SSR. The distance between the reference genome were extracted and mapped to each sequencing read was calculated and classified. The averagedistance of SSRlengthvariationsonmapping, the SSR sequences on SSRs was denoted as (c, s) are

\[ A(c, s) = 1 - \frac{1}{2} \sum_{i=1}^{S} \frac{m_i}{M_{i-1} + M_i + \sum_{j=1}^{i-1} m_j} = 1 - \frac{\sum_{i=1}^{S} m_i}{2(M + \sum_{i=1}^{S} m_i)} \] \tag{1}

\[ R(c, s) = 1 - \frac{\sum_{i=1}^{S} m_i}{M + \sum_{i=1}^{S} m_i} \] \tag{2}

where \( M_i \) and \( m_i \) are the numbers of consistent and inconsistent SSRs in the \( i \)th variety, respectively.

For a particular sequencing platform, the exact accuracy \( A(c, s) \) and reproducibility \( R(c, s) \) may slightly differ from the results from Equations (1) and (2) because different platforms have different error rates and sequence biases.

**Improvement to the accuracy of AmpSeq-SSR by removing non-random error-prone SSRs**

When an SSR is valid in two varieties by the two sequencing systems, we call the SSR comparable for the two varieties. If a comparable SSR is inconsistent for both of the two varieties, it is considered to be an error-prone SSR. Let \( n_{\text{observed}} \), \( n_{\text{random}} \), and \( n_{\text{non-random}} \) be the numbers of observed, random, and non-random error-prone SSRs, respectively. Then, \( n_{\text{observed}} = n_{\text{random}} + n_{\text{non-random}} \).

We introduce a null hypothesis \( H_0 \) that \( n_{\text{non-random}} = 0 \), i.e. all the observed error-prone SSRs occur randomly and the event that an SSR is inconsistent in one variety is independent of the event that the SSR is inconsistent in another variety. Let \( N_{ij} \) be the number of comparable SSRs for the \( i \)th and \( j \)th varieties. Then, the total number of comparable SSRs in all \( \binom{K}{2} \) pairs of \( K \) varieties is \( N = \sum_{i<j} N_{ij} \) \((K = 8 \) in the current study\). The probability of a random error-prone SSR can be estimated as \( \hat{q} = \frac{\sum_{i<j} n_{\text{observed}}(N_{ij})}{N} \), where \( p_i \) and \( p_j \) are the ratios between the inconsistent and comparable SSRs in the \( i \)th and \( j \)th varieties, respectively. The probability of having \( k \) random error-prone SSRs is binomially distributed, i.e. \( k \sim \text{Bin}(N, \hat{q}) \). Then, the probability of having no less than \( n \) random error-prone SSRs is in \( N \) comparable SSRs is

\[ P(n) = \sum_{k=n}^{N} \binom{N}{k} \hat{q}^k (1-\hat{q})^{N-k} \] \tag{3}

Let \( 1 - \alpha \) be the significance level for accepting \( H_0 (\alpha = 1\% \) in our study\). When \( P(n = n_{\text{observed}}) > \alpha \), we accept \( H_0 \). Otherwise, we reject it, meaning that non-random error-prone SSRs exist. When all non-random error-prone SSRs were identified and removed from the target SSRs, \( n_{\text{observed}} = n_{\text{random}} \). Then, the inconsistent SSRs in the \( K \) varieties can be estimated as \( K_{\text{random}} \), where the value of \( n_{\text{random}} \)

is determined by the equations of \( P(n = n_{\text{random}}) > \alpha \) and \( P(n = n_{\text{random}} + 1) \leq \alpha \). Following Equation (1) the accuracy of AmpSeq-SSR can be improved to:

\[ A_{c, s, n_{\text{random}}} = \frac{K_{\text{random}}}{2} \] \tag{4}

where \( M_i \) is the number of consistent SSRs in the \( i \)th variety. As an example, we used seven of the eight varieties to identify and remove the error-prone SSRs from the target SSRs, and calculated the improved accuracy of AmpSeq-SSR in
the eighth variety.

\[
A(c, s)_{\text{Sanger}} = 1 - \frac{1}{8} \sum_{i=1}^{8} \left( \frac{1}{T_i} \sum_{k=1}^{T_i} \frac{c_{ik}}{C_{ik}} \right)
\]

where \(C_{ik}\) is the number of valid SSRs in the \(i\)th variety, \(c_{ik}\) is the number of the reads which account for the major allele in the \(k\)th valid SSR of the \(i\)th variety and represents the times of correct SSR genotyping by Sanger sequencing, \(C_{ik}\) is the total number of the reads of this SSR and represents all the times of SSR genotyping by Sanger sequencing.

**Extension to heterozygous varieties and estimation of AmpSeq-SSR accuracy**

To extend AmpSeq-SSR to heterozygous varieties and estimate its performance, we created a pseudo-heterozygous variety \(ij\) and generated a set of *in silico* sequencing data by randomly sampling and mixing 0.8 M reads from each of the \(i\)th and \(j\)th varieties. The genotypes from the \(i\)th and \(j\)th varieties were combined as the reference genotypes of the pseudo-heterozygous variety \(ij\). To ensure the accuracy on the reference genotypes, we only considered those SSR loci with stutter ratios being lower than 0.5 and major alleles being covered by at least 50 reads in the \(i\)th and \(j\)th varieties.

An error correction model was introduced to determine the homozygosity or heterozygosity of an SSR locus in pseudo-heterozygous variety \(ij\). The allele or the two alleles with the most reads were taken as the test genotypes for the homozygous or heterozygous SSR loci in pseudo-heterozygous variety \(ij\), respectively. If a test genotype was the same as the reference genotype, the test genotype was taken as correct for the pseudo-heterozygous variety. The accuracy of AmpSeq-SSR on the pseudo-heterozygous variety \(ij\) was then estimated as:

\[
A_{ij} = \frac{T_{ij}}{t_{ij}} \times 100\% \quad (7)
\]

where \(t_{ij}\) and \(T_{ij}\) were the numbers of the correct genotypes and reference genotypes in pseudo-heterozygous variety \(ij\), respectively.

**RESULTS**

The AmpSeq-SSR method for SSR genotyping

We developed a novel method for large-scale SSR genotyping which is able to overcome the shortcomings of the existing methods (Table 1). The new method, named as AmpSeq-SSR, has a higher resolution, can be more efficiently deployed, and is able to provide more accurate results that can also be compared from different experiments and/or methods. AmpSeq-SSR combines multiplexing PCR, targeted deep sequencing and comprehensive analysis (Figure 1A). The current existing multiplex amplification techniques, such as that for whole exome sequencing, can amplify over 20,000 target loci in a single PCR reaction (38). Such a power of multiplex amplification was utilized in AmpSeq-SSR and 3105 SSRs were profiled in the current study. Since the genetic identities of SSR loci are mainly determined by their length variations, the SSR PCR products should be deeply sequenced on platforms (e.g. ABI S5 and Illumina MiSeq) that can produce long reads (e.g. 300 bp in the current study) to accommodate long SSR alleles. Sequencing reads were then mapped to target SSR loci to call SSR alleles. For homozygous samples, e.g. the rice varieties studied in the current study, only one allele was expected for a SSR locus and the allele with the most reads was taken as the SSR genotype (see ‘Materials and Methods’ section). For heterozygous samples, an additional step as described by (25) or (39) could be used to determine the SSR locus to be homologous or heterologous. Finally, the allele or the two alleles with the most reads were taken as the genotype of a homologous or heterologous SSR locus, respectively. Following the convention, an SSR genotype was represented as the amplicon length (see ‘Materials and Methods’ section), or the motif plus its repeat times, e.g. (AT)10.

**Genotypes identified by AmpSeq-SSR are highly accurate**

The SSR amplicons were designed to be <250 bp and could be sequenced through by MiSeq 2 × 300 bp paired-end sequencing mode from both forward and backward directions. On average, 1.38 million (M) (58.72%) of the 2.35 M paired-end reads for each of varieties A–H had identical forward and backward sequences and were expected to have negligible sequencing errors (Table 2). As a result, 2427.75
Figure 1. AmpSeq-SSR: its procedure, characteristics and comparison with capillary electrophoresis (CE). (A) A sketch of AmpSeq-SSR procedure applied to genotyping 3105 SSRs within 24 h. (B) Validation of an example SSR genotype in variety F using Sanger sequencing. (C, D and E) Profiles of AMPL1118228 (C) and AMPL1122663 (D) in variety H by CE and the comparisons of their genotypes with the genotypes from AmpSeq-SSR (E). (F) Amplicons with 40–70% GC contents have relatively uniform coverages. (G) The length distributions of all of the target SSRs on the reference genome. (H) The length distributions of the motifs of the target SSRs on the reference genome.

(78.19%) of the 3105 target SSRs were genotyped from the 1.38 M highly accurate reads with a coverage of 594.69-folds and stutter ratio of 0.15 on average for each detected SSR (Table 2).

To assess the accuracy and reproducibility of AmpSeq-SSR, all of the 3105 target SSRs in varieties A–H were genotyped on the MiSeq and S5 sequencing systems by different labs (see ‘Materials and Methods’ section). An inconsistent SSR, which has distinct genotypes from the two sequencing systems (see ‘Materials and Methods’ section), was taken as a mistaken genotyping and used to calculate the accuracy of AmpSeq-SSR according to Equation (1). Even under a relaxed error-controlling criterion of the coverage no <10-folds and the stutter ratio no >0.5, AmpSeq-SSR had an accuracy of $A_{ij} = 94.47 \pm 1.69\%$ (Supplementary Table S3) on pseudo-heterozygous varieties, which is comparable to the accuracy of SNP genotyping arrays on heterozygous varieties (40). The lower accuracy of AmpSeq-SSR on heterozygous varieties can be attributed to DNA polymerase slippage, especially when the real and slippage alleles had the same or similar lengths (25,39). It is noteworthy that AmpSeq-SSR should be more accurate on heterozygous varieties than other methods, such as whole genome sequencing (WSG) that is difficult to achieve a high sequencing coverage in order to statistically exclude errors from DNA polymerase slippage.
We randomly selected three SSR loci in two varieties for validation of AmpSeq-SSR using Sanger sequencing, the gold standard for genotyping. All of the six genotypes of the chosen SSR loci were proven to be correct (see Figure 1B for an example). We also compared AmpSeq-SSR with CE, the currently most popular and accurate electrophoresis technique, on two randomly chosen SSRs (Figure 1C and D). The lengths of the two SSRs detected by CE were respectively 5.53 and 4.47 bp shorter than those by AmpSeq-SSR (Figure 1E), suggesting that the SSR genotypes from different batches of CE might be inconsistent and incomparable. The stutter ratios for CE were respectively 1.56, 1.73, 1.98, 1.78, 1.40, and 1.87, which were much lower than that of AmpSeq-SSR. This suggests that the interference from CE on SSR genotyping was more serious than that from AmpSeq-SSR.

**Improvement to the accuracy of AmpSeq-SSR**

The accuracy and reproducibility of AmpSeq-SSR can be improved on SSRs with adequate sequencing coverages and lower stutter ratios (see 'Materials and Methods' section, Table 3). The stutter ratios of SSRs in varieties B-H were significantly decreased by 50.88% on average after removing the target SSRs with stutter ratios > 0.2 in variety A (Sup-
The accuracy of AmpSeq-SSR is very high and positively correlated with SSR coverage and negatively correlated with SSR stutter ratio

| Stutter ratio | ≤0.1 | >0.1 and ≤0.3 | >0.3 and ≤0.4 | >0.4 and ≤0.5 | >0.5 and ≤1.0 |
|---------------|------|--------------|-------------|--------------|-------------|
| Coverage      | Total1 Consistent2 Accuracy3 Total1 Consistent2 Accuracy3 Total1 Consistent2 Accuracy3 Total1 Consistent2 Accuracy3 Total1 Consistent2 Accuracy3 |
| ≤10×         | 8027 8233 98.75% 11 699 12 035 98.60% 12 471 12 924 98.25% 13 033 13 687 97.61% 15 391 15 564 94.95% |
| ≥10×         | 6839 6841 99.99% 9798 99.99% 10 270 10 306 99.83% 10 584 10 641 99.73% 11 044 11 303 99.84% |
| ≥20×         | 6281 6282 99.99% 8719 99.92% 9074 9099 99.86% 9298 9313 99.81% 9644 9799 99.21% |
| ≥30×         | 5799 5800 99.99% 7841 99.94% 8105 8142 99.89% 8290 8317 99.84% 8583 8680 99.33% |
| ≥40×         | 5330 5331 99.99% 7082 99.94% 7405 7320 99.90% 7449 7473 99.84% 7672 7767 99.39% |
| ≥50×         | 4940 4940 100.00% 6445 99.97% 6636 6646 99.92% 6751 6768 99.87% 6883 6995 99.77% |
| ≥100×        | 3382 3382 100.00% 4141 99.98% 4228 4230 99.96% 4284 4288 99.95% 4382 4410 99.66% |
| ≥200×        | 1654 1654 100.00% 1896 100.00% 1915 1915 100.00% 1957 1958 99.97% 1994 2002 99.80% |
| ≥500×        | 180 180 100.00% 200 100.00% 205 205 100.00% 208 208 100.00% 208 208 100.00% |

1 Total: the number of SSRs in varieties A–H whose coverages and stutter ratios satisfy the specified values and whose MiSeq and S5 sequencing platforms.
2 Consistent: the number of SSRs in varieties A–H whose coverages and stutter ratios satisfy the specified values and whose genotypes are identical between MiSeq and S5 sequencing platforms.
3 Refer to Equation (1) for calculation of accuracy.

Highly discriminative fingerprints derived via AmpSeq-SSR

The discriminative power of fingerprints on rice varieties is proportional to the number of SSRs included in the fingerprints. In this study, we employed as many as 3105 SSRS for fingerprinting rice varieties A–H, which are much more than the 48 SSRS widely used for rice identification currently adopted in the National Agricultural Standard of China (Standard No. NY/T 1433–2014) (Figure 2A). The power of the fingerprints is also positively correlated with the discernable differences between the SSR genotypes of two rice varieties. The resolution of SSR genotypes derived from AmpSeq-SSR is single base and therefore any subtle difference between SSR genotypes can be clearly observed. As a result, 449.71 SSRs on average were identified by AmpSeq-SSR to possess differential genotypes between two varieties (Figure 2B and Supplementary Table S7). To our knowledge, 449.71 differential SSRs between two fingerprints are the largest ever reported, suggesting the great power of AmpSeq-SSR to unambiguously distinguish any rice variety under test.

Although electrophoresis can discern the differences of SSR amplicon lengths, it cannot distinguish base changes or base differences. Among the 449.71 differential SSRs, 33.68 (7.07%) had different bases but the same amplicon lengths between two varieties so that they were deemed to be mistaken as identical on electropherograms. For example, amplicon AMPL1141969 has the same amplicon length but distinct SSR genotypes in varieties F and G, which could be clearly identified by AmpSeq-SSR (Figure 2C) but not by electrophoresis (Figure 2D). Furthermore, when the differences of amplicon lengths were below the resolution, they might also be indiscernible on electropherograms (e.g. Figure 2E). Among the 449.71 differential SSRs, 221.32 (51.72%) had amplicon length differences no > 5 bp (Figure 2B and Supplementary Table S7), which was the resolution of CE (Figure 1E), the most accurate electrophoresis technique.

Mapping of Xa21 gene by AmpSeq-SSR

Rice BB is a devastating disease that causes a significant annual rice yield reduction worldwide and Xa21 is one of the most effective genes to control BB (43,44). Through backcross over six generations, we introduced Xa21 from the donor variety IRBB821 into three receptor parental varieties, IRBB24, IRBB91 and D62B, and developed their respective NILs, R24, R11 and R62 (Supplementary Table S1). Each parental variety and its NIL share similar genomic background except for the target gene Xa21 and its linked region.
Figure 2. Highly discriminative fingerprints discovered by AmpSeq-SSR. (A) AmpSeq-SSR employed as many as 3105 target SSRs for rice fingerprint construction, which is ∼65 times of the 48 SSRs listed in rice National Standard. (B) AmpSeq-SSR could detect on average 449.71 differential SSR pairs between any two varieties. Among them, 48.28, 44.65 and 7.07% had amplicon length differences over 5, 1–5 and 0 bp, respectively, and therefore, are respectively distinguishable, hardly distinguished and indistinguishable on electropherograms. (C) An example of SSR with differential SSR genotypes but the same amplicon length. The SSR sequences are shown by yellow arrows. The numbers of sequencing reads are on the left of the sequencing reads. (D) The SSRs in (C) show no difference on agarose gel electrophorograms even though they have distinct SSR genotypes. (E) Examples of SSR amplicons on agarose gel electrophorograms. On each electropherogram, from left to right are rice varieties A–H marked by their SSR genotypes.

DISCUSSION

As a classic molecular marker, SSRs have been employed in broad applications, such as fingerprinting, gene mapping, forensic analysis and variety identification (16–20). Compared to the existing methods, AmpSeq-SSR is high applicable to those applications thanks to its high accuracy, high efficiency and low cost.
Figure 3. Mapping gene \textit{Xa21} by AmpSeq-SSR. (A) \textit{SSR21} is the only SSR with differential genotypes between every pair of the three nearly isogenic lines (NILs). (B) Identical genotypes of (CGC)\textit{7} among the receptor parental varieties of the NILs. (C) No genotype difference between NILs can be detected by PAGE electrophoretogram. The molecular weight standards are marked on the left of the electrophoretogram. (D) The sequences of \textit{SSR21} amplicons in the three pairs of NILs. Letters in yellow background are the sequences of \textit{SSR21}. The nucleotide substitution variations are marked by black arrow. Numbers of sequencing reads are listed on the left of the sequences.

AmpSeq-SSR is more accurate than the existing methods for SSR genotyping

Accuracy is always the primary consideration for any genotyping method. The current study demonstrated that AmpSeq-SSR had a much higher accuracy than the existing electrophoresis-based methods for SSR genotyping (Figures 1E, 2B–E, 3C and D; Supplementary Table S7). Except gel electrophoresis, Sanger sequencing is occasionally used for SSR genotyping. By double sequencing of the same sequences, AmpSeq-SSR could detect and avoid most of the sequencing errors to a level comparable with Sanger sequencing, i.e. $\sim 10^{-6}$ per base. As a by-product, the accuracy of SSR genotyping by Sanger sequencing could be estimated with AmpSeq-SSR reads by Equation (6) as $A(10, 0.5)_{\text{Sanger}} = 83.89\%$, which is much lower than the AmpSeq-SSR accuracy of $A(10, 0.5) = 99.73\%$. The whole genome resequencing can also be used for SSR genotyping (25,31,32). However, for a high accuracy that AmpSeq-SSR has, the target SSRs need to be covered for $\sim 500$-folds (Table 2), which requires coverage of the whole genome for thousands of folds, resulting in an unacceptable cost.

By overcoming most of the shortcomings in the existing methods for SSR genotyping, the accuracy of AmpSeq-SSR reached $A(10, 0.5) = 99.73\%$ (Table 3), which can be further improved to $A(10, 0.5)_{\text{improved}} = 99.92\%$ (for the parameters used, refer to Supplementary Table S6) or even potentially $A(10, 0.5)_{\text{no-random}} = 100.00\%$. The high accuracy makes AmpSeq-SSR capable to accurately identify a few or even one distinct gene from the plant varieties developed by transgene, backcrossing and mutation. However, a few distinct genes can be elusive within a background of hundreds of uncertain SSRs on electrophoresis profiles (Figure 2B and Supplementary Table S7). Moreover, AmpSeq-SSR has potential advantages for legal applications such as forensic analysis or granting rights under plant variety protection (PVP), for which any mistake may have serious consequences.

Factors that made AmpSeq-SSR accurate

The slippage of DNA polymerases during PCR amplification is the major source of error for SSR genotyping (21–27). The more PCR cycles, the higher chance for slippage errors to accumulate. Because visual inspection requires sufficiently bright bands, gel electrophoresis for SSR genotyping needs more PCR cycles than AmpSeq-SSR, i.e. 40 versus 16 PCR cycles in this study, resulting in more serious slippage (i.e. higher stutter ratios) for gel electrophoresis (Figure 1E). Note that a sequencing library can be constructed using a PCR-free protocol (25), so that polymerase slippage during library construction can be avoided for AmpSeq-SSR.

Duplex sequencing can improve the accuracy of SSR genotyping by reducing sequencing errors. In addition to inherent and unavoidable sequencing error, the PCR cycle number ($n$) and sequencing coverage ($N$) are primary factors contributing to the genotyping error probability ($P$). An SSR locus cannot be genotyped correctly when slippage reads take up more than 50% of the total coverage of the locus. Therefore, the error probability can be estimated as:

$$P = \sum_{k=1}^{N} \binom{N}{k} (1 - R)^k R^{N-k} \text{ and } R = 1 - (1 - r)^n$$
where $r$ is the slippage probability of a read during a single PCR cycle, and $R$ is the percentage of slippage reads in the final PCR product.

The PCR cycle number ($m$) smaller than the electrophoresis based methods and the sequencing coverage ($N$) greater than WGS lead to a potentially higher accuracy of AmpSeq-SSR for SSR genotyping. The SSR often slips for a motif (2–4 bases) (25), making it difficult to distinguish the allele from slippage on electropherogram. Without slippage, accurately comparing SSR genotypes on electropherogram might still be nontrivial. For example, the SSRs in amlicon AMPLI1141969 of variety F had negligible stutter ratio of no > 0.01 (Supplementary Table S2) but migrated differently between two neighboring wells of the same gel (Figure 2D).

**High efficiency of AmpSeq-SSR extends its applications**

We simultaneously amplified 3105 SSRs in eight rice varieties in a single PCR reaction and sequenced them in a single sequencing run. At present, such a combination of multiplex amplification and high-throughput sequencing can examine even more SSRs. Consider the MiSeq sequencing platform, for example, which can produce $M = 25$ million reads in a single run. When covered by $m = 10$ reads, an SSR can be genotyped with almost 100% accuracy (Table 3). Therefore, AmpSeq-SSR using the MiSeq platform can accurately genotype $M/m = 2.5$ million SSRs at once. In contrast, gel electrophoresis based SSR genotyping individually amplifies and separately examines each SSR, and thus is labor intensive and inefficient. Furthermore, thousands of SSR PCR reactions require a significant amount of template DNAs, e.g. an excessive ~31 μg for 3105 SSRs.

The ability to genotype a large number of SSRs is the key to the success of AmpSeq-SSR in many applications. The 3105 target SSRs used by AmpSeq-SSR for fingerprinting are critical for the great discriminative power of the resulting fingerprints. Another good example is the mapping of Xa21 gene. Because of the successive backcrossing and selection, the target region of Xa21 gene became rather small, i.e. ~2 Mbp, as estimated in our previous study (46). This small region reduces the chance to have usable SSRs for gene mapping. On average, 449.71 (29.61%) SSRs of the 1518.61 comparable SSRs on the ~400 Mbp rice genome had differential genotypes between two fingerprints (Supplementary Table S8). Therefore, 7.59 (2 × 1518.61/400) SSRs are expected to exist in the target region. The probability of having at least one differential SSR genotype between the three pairs of NILs within the target region is $[1 - (1 - 29.61%)^{2.59}] = 80.54\%$, suggesting a great chance for AmpSeq-SSR to successfully map a target gene.

**High reproducibility of AmpSeq-SSR extends its applications**

Reproducibility is important for a new technique. Based on Equation (2), AmpSeq-SSR had a reproducibility close to 100% (Supplementary Table S9), ensuring it to produce highly consistent and comparable results under various conditions. Therefore, AmpSeq-SSR can be used to collaboratively construct fingerprint database without sharing the original biological resources (e.g. specific rice varieties), which are often taken as trade secrets or national strategic resources. More importantly, a fingerprint from AmpSeq-SSR can be freely and accurately compared with all records in a database to determine its distinctness from all the existing varieties, which is the legal precondition for a PVP grant. However, the fingerprints from electrophoresis can be accurately compared only when they are constructed in parallel, making it nearly infeasible to compare one variety against thousands of existing varieties.

**AmpSeq-SSR is economical**

The cost of AmpSeq-SSR is low, i.e. ~$0.015 per SSR in this study, which is more affordable than other techniques (47). The high consensus of SSR genotypes between MiSeq and S5 (Table 3) indicated that the cost could be further decreased by the more economical single-end sequencing strategy.

**The advantages of AmpSeq-SSR over the existing methods**

AmpSeq-SSR has comparative advantages over the existing methods for large-scale SSR genotyping, including accuracy, resolution, throughput, efficacy and comparability of results from different experiments and/or methods (Table 1). Besides, comparing with WGS, AmpSeq-SSR is able to avoid the interference of homologous sequences on target SSRs by designing multiplex primers that do not anneal to homologous SSRs, which would be difficult if not impossible to realize for WGS. To overcome this shortcoming, WGS reads have to be discarded when the regions flanking SSRs are not perfectly matched to the reference, e.g. that appeared in (25), resulting in a genotyping failure for SSRs with variations in their flanking regions. In order to correct genotyping errors, a WGS-based strategy needs to resort to an ultra-high-throughput sequencing to deeply sequence target SSRs (Table 3). On the other hand, ultra-high-throughput sequencing typically produces short reads, producing a skewed distribution of reads toward short SSRs. AmpSeq-SSR has evident advantages and shows no obvious shortcomings over the other methods (Table 1), suggesting that AmpSeq-SSR is the most preferable method for large-scale SSR genotyping.

**ACCESSION NUMBERS**

All sequencing data of the eight rice varieties produced and analyzed in the current study have been deposited into NCBI Sequence Read Archive (SRA) under accession number SRP089769.

**AVAILABILITY**

The package of scripts that implement the AmpSeq-SSR method is freely available at a public repository https://github.com/SystemsBiologyOfJianghanUniversity/AmpSeq-SSR.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.
FUNDING
The National Key Research and Development Program of China [2016YFF0202300]; Projects on EDVs by Ministry of Agriculture; Projects of Wuhan yellow crane talents; Youth and Technology Morning Program in Wuhan [2014072704011257]; Guidance project by Hubei Ministry of Education [B2015229]; United States National Institutes of Health [R01GM100364]. Funding for open access [2014072704011257]; Guidance project by Hubei Ministry of Agriculture; Projects of Wuhan yellow crane talents; Projects on EDVs by Ministry of Agriculture; Projects of Wuhan yellow crane talents; Funding by Hubei Ministry of Education [B2015229].

ACKNOWLEDGEMENTS
Author contributions: H.P. conceived the project, drafted the manuscript and supervised the study. W.Z. supervised the study and wrote the paper. L.Li and Z.F. analyzed the data. J.Z. performed the sequencing experiments. H.C. and Z.H. prepared the materials. L.G., L.C. and H.M. performed the validation experiments. L.Lu and S.R. participated in the statistical analysis. All of the authors discussed the results and commented on the manuscript.

REFERENCES
1. Li,Y.-C., Korol,A.B., Fahima,T. and Nevo,E. (2004) Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol., 21*, 1091–1107.
2. Iglesias,A.R., Kindlund,E., Tammi,M. and Wadelius,C. (2004) Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. *Gene, 341*, 149–165.
3. Martin,P., Makepeace,K., Hill,S.A., Hood,D.W. and Moxon,E.R. (2005) Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. U.S.A., 102*, 3800–3804.
4. Krishna,J. and Mishra,R.K. (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays, 28*, 1040–1050.
5. Gymrek,M., Willems,T., Guilmatre,A., Markus,B., Georgiev,S., Daly,M.J., Price,A.L., Pritchard,J.K., Sharp,A.J. et al. (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet., 48*, 22–29.
6. Buschiazzo,E. and Gemmell,N.J. (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays, 28*, 1040–1050.
7. Yang,S., Wang,L., Huang,J., Zhang,X., Yuan,Y., Chen,J.-Q., Hurst,L.D. and Tian,D. (2015) Parent-pregnant sequencing indicates higher mutation rates in heterozygotes. *Nature, 523*, 463–467.
8. Osowski,S., Schneeberger,K., Lucas-Lledo,J.I., Wirthmann,N., Clark,R.M., Shaw,R.G., Weigel,D. and Lynch,M. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science, 327*, 92–94.
9. Moore,S., Sargeant,L., King,T., Mattick,J., Georges,M. and Hetzel,D. (1991) The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. *Genomics, 10*, 654–660.
10. Moodley,Y., Baumgarten,I. and Harley,E. (2006) Horse microsatellites and their amenability to comparative equid genetics. *Anim. Genet., 37*, 258–261.
11. Dawson,D.A., Horsburgh,G.J., KUPPER,C., Stewart,I.R., Ball,A.D., Durrant,K.L., Hannson,B., Bacon,I., Bird,S. and Klein,A. (2010) New methods to identify conserved microsatellite loci and develop primer sets of high cross-species utility–as demonstrated for birds. *Mol. Ecol. Resour., 10*, 475–494.
12. Moodley,Y., Masello,J.F., Cole,T.L., Calderon,L., Munimanda,G.K., Thali,M.R., Alderman,R., Cuthbert,R.J., Marin,M. and Massaro,M. et al. (2015) Evolutionary factors affecting the cross-species utility of newly developed microsatellite markers in seabirds. *Mol. Ecol. Resour., 15*, 1046–1058.
13. Selkoe,K.A. and Toonen,R.J. (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecol. Lett., 9*, 615–629.

14. Guichoux,E., Lagache,L., Wagner,S., Chaumeil,P., Leger,D., Lepais,O., Lepoittevin,C., Malausa,T., Revardel,E., Salin,F. et al. (2011) Current trends in microsatellite genotyping. *Mol. Ecol. Resour., 11*, 591–611.
15. Schlotterer,C. (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma, 109*, 365–371.
16. Kaur,S., Panesar,P.S., Bera,M.B. and Kaur,V. (2015) Simple sequence repeat markers in genetic divergence and marker-assisted selection of rice cultivars: a review. *Crit. Rev. Food Sci. Nutr., 55*, 41–49.
17. Jarne,P. and Lagoda,P.J. (1996) Microsatellites, from molecules to populations and back. *Trends Ecol. Evol., 11*, 424–429.
18. Kim,K.S. and Sappington,T.W. (2013) Microsatellite data analysis for population genetics. *Methods Mol. Biol., 1006*, 271–295.
19. Chambers,G.K., Curtis,C., Millar,C.D., Huynen,L. and Lambert,D.M. (2014) DNA fingerprinting in zoology: past, present, future. *Nestig. Genet., 5*, 1–11.
20. Borsting,C. and Morling,N. (2015) Next generation sequencing and its applications in forensic genetics. *Forensic Sci. Int. Genet., 18*, 78–89.
21. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev., 5*, 435–445.
22. Webster,M.T. and Hagberg,J. (2007) Is there evidence for convergent evolution around human microsatellites? *Mol. Biol. Evol., 24*, 1097–1100.
23. Brandström,M., Bagshaw,A.T., Gemmell,N.J. and Ellegren,H. (2008) The relationship between microsatellite polymorphism and recombination hot spots in the human genome. *Mol. Evol., 25*, 2579–2587.
24. Kelkar,Y.D., Tykucheva,S., Chiaromonte,F. and Makova,K.D. (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res., 18*, 30–38.
25. Fungtammasan,A., Ananda,G., Hile,S.E., Sun,M., Sun,C., Harris,R., Medvedev,P., Eckert,K. and Makova,K.D. (2015) Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res., 25*, 736–749.
26. Abdulovic,A.L., Hile,S.E., Kunkel,T.A. and Eckert,K.A. (2011) The in vitro fidelity of yeast DNA polymerase δ and polymerase ε heloenzymes during dinucleotide microsatellite DNA synthesis. *DNA Repair, 10*, 497–505.
27. Baptiste,B. and Eckert,K.A. (2012) DNA polymerase kappa microsatellite synthesis: two distinct mechanisms of slippage-mediated errors. *Environ. Mol. Mutagen., 53*, 787–796.
28. Zhang,L., Cai,R., Yuan,M., Tao,A., Xu,J., Lin,L., Fang,P. and Qi,J. (2015) Genetic diversity and DNA fingerprinting in jute (*Corchorus spp.*) based on SSR markers. *Crop J., 3*, 416–422.
29. Liu,X.B., Feng,B., Li,J., Yan,C. and Yang,Z.L. (2016) Genetic diversity and breeding history of Winter Mushroom (*Flammulina velutipes*) in China uncovered by genomic SSR markers. *Gene, 591*, 227–235.
30. Njung e,V., Deshpande,S., Siambi,M., Jones,R., Slim,S. and De Villiers,S. (2016) SSR genetic diversity assessment of popular pigeonpea varieties in Malawi reveals unique fingerprints. *Electron. J. Biotechnol., 21*, 65–71.
31. Kim,K.-S., Noh,C.H., Moon,S.-J., Han,S.-H. and Bang,J.-C. (2016) Development of novel tetra- and trinucleotide microsatellite markers for giant grouper Epinephelus lanceolatus using 454 pyrosequencing. *Mol. Biol. Rep., 43*, 541–548.
32. Bozzi,J.A., Liepelt,S., Ohneiser,S., Gallo,L.A., Marchelli,P., Leyer,I., Kresse,M. and D’Agostino,G. (2015) Simultaneous amplification of microsatellite markers in soybean using next-generation sequencing and cross-amplification from related species. *Appl. Plant Sci., 3*, 1400120.
33. Lander,E.S., Linton,L.M., Birren,B., Nussbaum,R.C., Zody,M.C., Baldwin,J., Devon,K., Darbor,B., Everett,M., FitzHugh,W. et al. (2001) Initial sequencing and analysis of the human genome. *Nature, 409*, 860–921.
34. Star,B., Hansen,M.H., Skage,M., Bradley,I.R., Godiksen,J.A., Kjøsby,O.S. and Jenfelt,S. (2016) Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples. *STAR, 2*, 36–45.
35. Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet., 13*, 36–46.
36. Gymrek, M., Golan, D., Rosset, S. and Erlich, Y. (2012) lobSTR: a short tandem repeat profiler for personal genomes. Genome Res., 22, 1154–1162.

37. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359.

38. Pater, J.A., Benteau, T., Griffin, A., Penney, C., Stanton, S.G., Predham, S., Kizley, B., Squires, J., Zhou, J. and Li, Q. (2017) A common variant in CLDN14 causes precipitous, prelingual sensorineural hearing loss in multiple families due to founder effect. Hum. Genet., 136, 107–118.

39. Berg, K.D., Glaser, C.L., Thompson, R.E., Hamilton, S.R., Griffin, C.A. and Eshleman, J.R. (2000) Detection of microsatellite instability by fluorescence multiplex polymerase chain reaction. J. Mol. Diagn., 2, 20–28.

40. Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., Meitinger, T., Strom, T.M., Fries, R. and Pausch, H. (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics, 15, 823.

41. Minoche, A.E., Dohm, J.C. and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. Genome Biol., 12, R112.

42. Li, S., Labaj, P.P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.Y., Wang, M., Wang, C. et al. (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat. Biotechnol., 32, 888–895.

43. Mew, T. (1987) Current status and future prospects of research on bacterial blight of rice. Annu. Rev. Phytopathol., 25, 359–382.

44. Niño, D.O., Ronald, P.C. and Bogdanove, A.J. (2006) Xanthomonas oryzae pathovars: model pathogens of a model crop. Mol. Plant Pathol., 7, 303–324.

45. Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.X., Zhu, L.H. et al. (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science, 270, 1804–1806.

46. Peng, H., Chen, Z., Fang, Z., Zhou, J., Xia, Z., Gao, L., Chen, L., Li, L., Li, T. and Zhai, W. (2015) Rice Xa21 primed genes and pathways that are critical for combating bacterial blight infection. Sci. Rep., 5, 12165.

47. Jennings, T., Knaus, B., Mullins, T., Haig, S. and Cronn, R. (2011) Multiplexed microsatellite recovery using massively parallel sequencing. Mol. Ecol. Res., 11, 1060–1067.