Hyper-seq: A novel, effective, and flexible marker-assisted selection and genotyping approach

Meiling Zou[1,2,3] and Zhiqiang Xia[1,2,3,*]

[1]Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China

[2]College of Tropical Crops Hainan University, Hainan University, Haikou 570288, China

[3]The Institute of Tropical Biosciences and Biotechnology, Chinese Academy of Tropical Agriculture Sciences, Haikou 571101, China

*Correspondence: zqiangx@gmail.com

Received: January 25, 2022; Accepted: April 26, 2022; Published Online: April 30, 2022; https://doi.org/10.1016/j.xinn.2022.100254

Citation: Zou M. and Xia Z. (2022). Hyper-seq: A novel, effective, and flexible marker-assisted selection and genotyping approach. The Innovation **3(4)**, 100254.

Recently, a novel extremely low-cost, effective, flexible, and high-throughput DNA sequencing library preparation and genotyping approach has been developed named Hyper-seq. This new technology has been adopted by more than 15 research institutes and universities, which has significantly improved the efficiency of crop breeding.

The collection and explanation of all DNA sequences from a species, known as pangenome, is the future and hotspot of genomics research. The effective and flexible identification of distinct genetic materials among individuals or populations from pangenome is crucial for exploring genetic and molecular biological processes. Single-nucleotide polymorphisms (SNPs) have become the preferred genetic component to examine human diseases as well as animal and plant resistance and molecular breeding.[1] Meanwhile, driven by next-generation sequencing (NGS) technology, the cost of DNA sequencing has been greatly decreased. Over the past decade, genomics-assisted breeding (GAB) has played an important role in utilizing the potential and characterization of modern genomic resources, as well as using SNP and insertion or deletion (indel) variation to enhance germplasm and develop cultivar. However, whole-genome sequencing is still expensive for investigating species population with large genomes.

Food security and modern agriculture economy are closely related to human activities. The crop seed industry is the foundation of food security and serves as an important component in agricultural economy. The global market of seed industry was estimated to be around $60 billion in 2012.[2] According to a report by International Seed Federation (ISF), China has an annual seed consumption of around 12 billion kg and the second-largest seed market in the world.[2] More than 1 million breeding materials from 700 research institutes and universities have been used for breeding in Sanya Nanfan in Hainan province of south China, which has become an important platform for international agricultural science and technology cooperation in China. However, it is urgently important to flexibly customize the number of detection sites, low-cost, automated high-throughput sequencing library preparation methods, an easy-to-operate SNP discovery platform with downstream data analysis for genotyping, and marker-assisted selection facilitating crop genetic improvements.

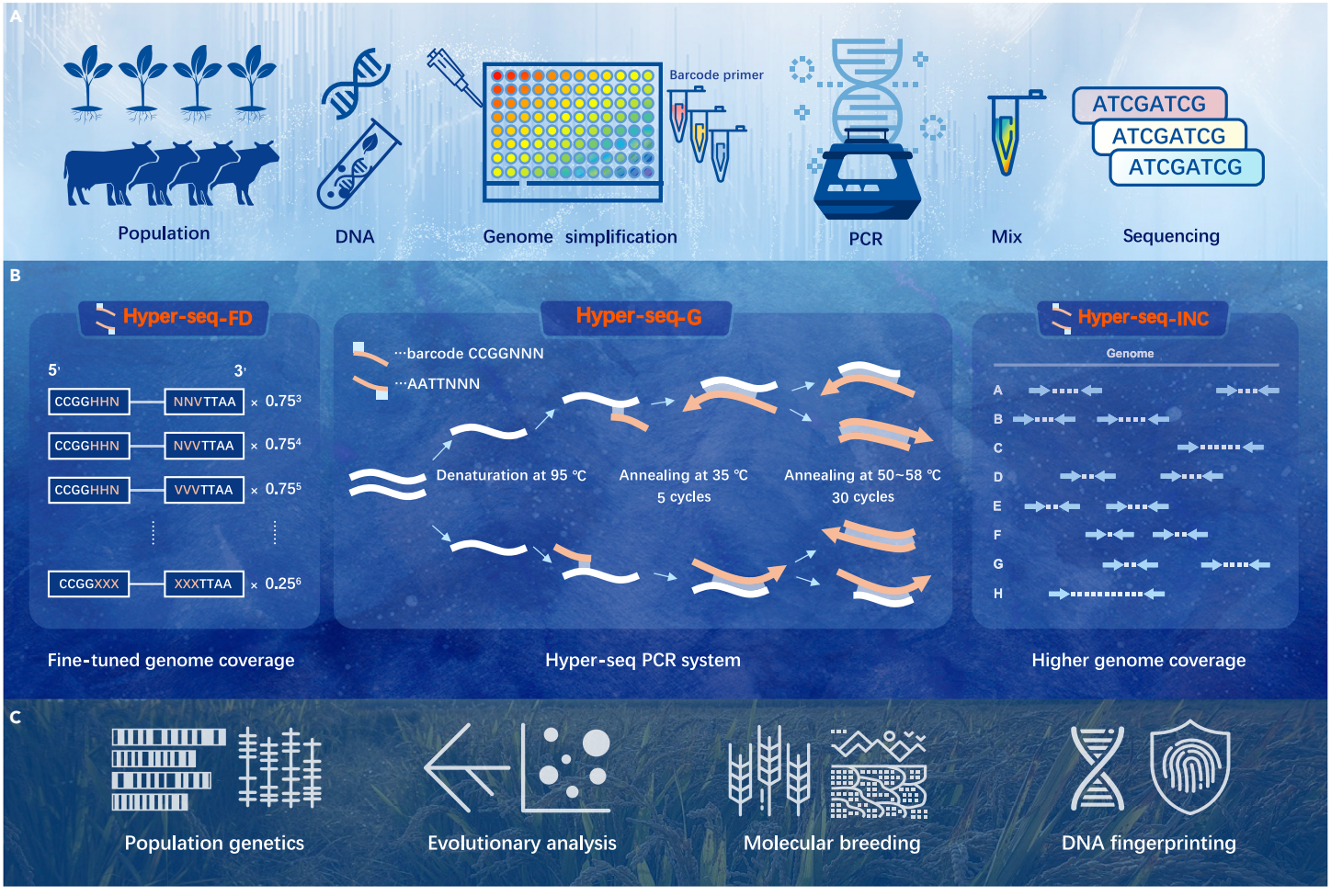

Figure 1. Methodology and application of Hyper-seq technology (Hyper-seq-G, Hyper-seq-FD, and Hyper-seq-INC) (A) Preparation and sequencing of Hyper-seq tags. During sample preparation, genome simplification, barcode addition, and amplification for simplified genome sequences were all in one step through PCR using special Hyper-seq primers. (B) The marker density could be fine-tuned as needed using different Hyper-seq primers. (C) The application of Hyper-seq technology.

In order to reduce the cost of genomics sequencing and promote seed industry, we developed a novel cost-effective, automated, flexible polymerase chain reaction (PCR) method, Hyper-seq (Figure 1A). To validate the effectiveness of the method, 2,094 samples of six crop species were used for libraries construction, sequencing, and genotyping by Hyper-seq. The results indicated the successful application of Hyper-seq in variation detection in crop breeding, which was as valuable as breeding array.

The Hyper-seq-G upstream primer set was designed containing 6-base barcodes, 10-fixed bases, 4-base core bases, and compatible (NNN) overhangs (Figure 1B). The CCGG sequence, designed as the core bases of the Hyper-seq-G upstream primer set near the 3′ region, was to target exons to open-reading-frame regions, since sequences in exons are more frequently found in GC-rich regions and are expected to be preferentially amplified. The Hyper-seq-G downstream primer set was designed with core bases containing the AATT sequence to increase the level of polymorphism, since sequences in introns and promoters are frequently found in AT-rich regions and are generally more variable than those in exons among different individuals.

The cost of Hyper-seq library preparation and sequencing for samples with a genome size of 1G could be reduced to be less than $10 per sample. In contrast, analyzing each individual in population using breeding array such as GSR40K is expensive ($93 per sample). Each type of breeding array only contains a fixed and limited set of markers and thus can be only analyzed for a given species or finite population. Meanwhile, developing a breeding array for non-model species requires effective reference-genome information and additional costs, which is unnecessary for Hyper-seq.

In order to minimize potential linkage drag, the size of introgression fragment for each gene is suggested to be less than 200 kb.[3] Thus, high-resolution platforms are essential for accurate and efficient genetic background analyses. A total of 1,641,957 high-quality markers for F_2 and 1,586,100 for BC_3F_1 *Oryza sativa* populations were detected by Hyper-seq, the latter of which (1,586,100) was ∼105 times more than that detected by the GSR40K rice-breeding array (15,015). Based on the sorting results of a genetic background screening, the offspring could be easily selected to significantly reduce the costs of planting. In this study, we recommended 112 offspring from 402 BC_3F_1 *O. sativa* populations based on the fact that the genetic background similarity between offspring and the female parent "HHZ" is no less than 65%.

Since genetic background effects can be inadvertently enhanced by limited and uneven markers, not only the number but also the coverage and uniformity of markers are important for breeding. In this study, high quality SNPs and indels were obtained for *Zea mays* (5,395,605), *Solanum tuberosum* (5,086,178), *Aegilops tauschii* (4,912,506), *Manihot esculenta* (2,334,294), *O. sativa* (2,586,451), and *Triticum aestivum* (2,573,078). These SNP and indel markers had very uniform coverage on chromosomes.

The Hyper-seq method was more efficient with a lower cost compared with the commonly used whole-genome sequencing (WGS) methods, such as restriction-site-associated DNA tag sequencing fragments produced by type IIB restriction endonucleases (2b-RAD)[4] and amplified-fragment single nucleotide polymorphism and methylation (AFSM).[5] The most advantageous feature of Hyper-seq is performing genome simplification, barcode addition, and simplified genome-sequences amplification in a single step (Figure 1). Hyper-seq library construction employs specially designed primers and DNA (for PCR amplification) or leaves (for direct PCR amplification) templates and does not require restriction endonuclease digestion and barcode ligation steps. For 384 samples, the total cost of Hyper-seq libraries construction was reduced by approximately nine times, and the entire construction time length was reduced to be as little as about 3 h. The PCR products were incorporated with an equal amount of each sample (up to 384 samples for one library), with uniform sizes of 250–500 bp in a single tube, effectively avoiding large size differences. Furthermore, the cost could be further reduced through automatic equipment development without affecting the library quality.

Studies on *Z. mays* crosses and *S. tuberosum* natural populations confirmed that Hyper-seq can be used to construct genetic maps of genetic populations, detect quantitative trait loci, identify different cultivars, and perform structural analyses of natural populations and genome-wide association studies. A genetic linkage map was generated for a *Z. mays* F_2 population distributed in 10 linkage groups, containing 3,425 markers; the total map length was 2,457.86 cM. A total of 527 *S. tuberosum* samples were classified into 8 subgroups, and 5P1-1 and 5P1-6 were recognized as the same cultivar (similarity coefficient >0.95).

Sequencing data with various degrees of genome coverage are usually required for different research studies. For example, studies on crops with a large genome size (e.g., *Tritrcum aestivum*) need further reduction of marker density, which can be achieved by using the primer set of Hyper-seq-FD with less degenerate 3′ overhangs (Figure 1B). It is an important benefit that further reduction of marker density allows for sequencing cost reduction and thus facilitates affordable large-scale population sequencing in crops with large-genome sizes or limited recombination events. On the other hand, in some certain situations, to achieve a higher coverage, increments in marker density can also be achieved by using Hyper-seq-INC primer combinations with different fixed bases (Figure 1B). Using mixtures of 8 Hyper-seq-INC primer pairs with wobble bases, fine adjustment of the marker density was performed to achieve the required degree of coverage for different genomic intervals for *M. esculenta* (genome 1%–55%). The highest degree of genome coverage for one sample (278.54 Mb, 56.22%) was close to the degree of the re-sequencing genome coverage result (67.27%) of *M. esculenta*. It was suitable for simplifying the sequencing of various species (including complicated genomes).

In conclusion, the Hyper-seq method is reliable and scalable, with low labor requirements, high cost effectiveness, and high throughput. Hyper-seq exhibits extensive applicable potentials. It is suitable for studying a large-sample population, for example, for plant molecular breeding or population genetic-background screening for crop backcross breeding, genotyping, marker-assisted selection, target key genes, and novel candidate trait-related gene detection, biosafety prevention and control (alien species inspection and quarantine), SNP cultivar identification, and genomic-selection breeding. With the cost of DNA sequencing decreasing, as well as further simplification and intelligence of rapid genotyping platform development, Hyper-seq technology will benefit more global non-model crop breeders.

PATENT NUMBER

This Hyper-seq technology (no. 2020/06979) and related research (nos. 102543 and 202010102817.5) achievements have been authorized as three invention patents in Republic of South Africa, Luxembourg, and China in 2021.

DATA AVAILABILITY

High-quality SNP and indel data information have been deposited to the Genome Variation Map (GVM) database in National Genomics Data Center of China National Center (https://ngdc.cncb.ac.cn/) under BioProject: PRJCA004105, PRJCA004113, PRJCA004114, PRJCA004118, PRJCA004117, and PRJCA008800.

REFERENCES

1. Wang, W., Mauleon, R., Hu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature ***557***, 43–49.
2. Wang, L., Cui, W., Kalala, J.P., et al. (2014). To investigate the effect of osteoporosis and intervertebral disc degeneration on the endplate cartilage injury in rats. Asian Pac. J. Trop. Med. ***7***, 796–800.
3. Wing, R.A., Purugganan, M.D., and Zhang, Q. (2018). The rice genome revolution: from an ancient grain to Green Super Rice. Nat. Rev. Genet. ***19***, 505–517.
4. Wang, S., Meyer, E., McKay, J.K., and Matz, M.V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat. Methods ***9***, 808–810.
5. Xia, Z., Zou, M., Zhang, S., et al. (2014). AFSM sequencing approach: a simple and rapid method for genome-wide SNP and methylation site discovery and genetic mapping. Sci. Rep. ***4***, 7300.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2019YFD1001105), Developing Bioinformatics Platform in Hainan Yazhou Bay Seed Lab (B21HJ0001), and Hainan University Startup Fund (KYQD(ZR)-20101). We are grateful to W. Wang, H. Wang, C. Xia, H. Li, Y. Zhou, D. Yang, J. Wang, F. Wang, and C. Lu for growing samples. We also thank J. Luo, S. Xie, Z. Wang, G. Xiao, and B. Feng for helpful suggestions on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.