Study of a region on wheat chromosome 5BS by targeted sequencing

M.K. Bragina[1]*, D.A. Afonnikov[1,2], G.V. Vasiliev[1], E.A. Salina[1]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*

DOI 10.18699/ICG-PlantGen2019-37

* e-mail: koltunova@bionet.nsc.ru

Abstract: Despite the increasing use of high-throughput sequencing resulting from reduced cost and effort, large and complex genomes still pose a challenge in crop genomics. Bread wheat has a genome size of over 17 Gb, which makes targeted capture an invaluable tool for a wide range of studies. To capture and sequence the 26.6-Mb wheat chromosome 5BS region associated to leaf rust resistant genes, custom hybridization probes was created using information from the reference wheat genome ('Chinese Spring'). We show that this chromosomal region of leaf rust resistance and susceptible wheat lines were efficiently captured with sufficient coverage.

Key words: bread wheat genome; chromosome 5BS; leaf rust resistance genes; Lr52 gene; targeted sequencing.

1. Introduction

Despite the advances in and decreasing the cost of next-generation sequencing and bioinformatics tools, whole genome sequencing is not practical for most studies. The whole genome sequencing with high level of coverage and resolution is not enough to detect all changes in complex samples. To reduce the complexity of genomes in large or polyploid plants, the strategy of targeted enrichment is used. Targeted sequencing includes 3 stages: selection of the DNA fraction of interest (enrichment of samples), sequencing and analysis of the results obtained. The method starts with shearing total genomic DNA into fragments and hybridizing target-specific probes to the regions of interest. Then, the selected fragments are extracted and amplified by PCR before sequencing. As well-annotated genomes improve probe design, high-quality reference genomes reduce the risk of false-positive results or missing important variants in the generated data set (Chamala et al., 2015; Warr et al., 2015).

The wheat reference cultivar 'Chinese Spring' genome published by the International Wheat Genome Sequencing Consortium (IWGSC, 2018) and SNP markers associated with economically valuable genes in wheat varieties and lines will contribute to solving important problems in genetics and breeding. To date, more than 30 genes that control a number of morphological and quantitative traits, resistance to abiotic and biotic environmental factors have been mapped on wheat chromosome 5B. Of particular interest is the distal region of its short arm (5BS), in which genes for resistance against leaf rust (*Lr52*), septoria brown blotch (*Snn3*), rhizoctonia net blotch (*QSe.jaas-5BS*), yellow rust (*Yr47*), loose smut of wheat (*UtBW278*) and Hessian fly (*H31*) are presumably located. The mapping of DNA markers linked to the resistance genes on the 5BS reference pseudomolecule (RefSeq v1.0) allows the region where these genes reside to be identified. Our study was aimed to approbate targeted sequencing for analysis of chromosome 5BS region with a length of about 26.6 Mb using lines with and without the Lr52 gene.

2. Materials and methods

2.1 Plant material

Mapping population F4 (lines from the cross LrW(52) × hybrid_215) was developed for chosen plants which are differ to resistant genes against fungal diseases located on target region of 5BS. We selected 5 plants with *Lr52* and 5 plants without *Lr52* according to data of KASP and SSR genotyping together with screening for resistance (benzimidazole (0.035 % w/v)).

2.2 DNA extraction

Genomic DNA was isolated from the young leaves of individual plants using the Kleargene plant 96-well plate DNA extraction kit (LGC Group), following the manufacturer's protocol. The isolated DNA was resuspended in 74 µl of TE buffer. The measurement of DNA concentration was performed on a NanoDrop M2000 instrument (Thermo Scientific).

2.3 Targeted sequencing

To enrich a 26.6 Mb of chromosome 5BS, a 1-µg portion of genomic DNA, in a total volume of 55 µl, was sheared for 2 × 60 s using a Covaris M220 focused-ultrasonicator (an average fragment size of 600 bp). Genomic libraries were constructed with the KAPA HyperPlus Library Preparation Kit according to the manufacturer's instructions (KAPA Biosystems). Equal amounts of 10 libraries were pooled and subjected to in-solution target enrichment using the SeqCap EZ Target Enrichment System (Roche). Enrichment and hybridization of the samples were carried out according to the manufacturer's proposed protocol using Qubit 2.0 (Life Technologies) for DNA concentration measuring, SimpliAmp (Applied Biosystems) for the DNA library amplification and the Agilent 2100 Bioanalazer (Agilent Technologies) for quality control and the final library size determination. The 150-bp paired-end sequencing of the obtained library was performed on an Illumina NextSeq 550 platform (ICG SBRAS).

Table 1
Characterization of probe designs developed by Roche to cover 26.2 Mb of chromosome 5BS

Length of regions (bp)	26 622 901	
	Design 1	Design 2
Statistics	Probe_Coverage	Probe_Coverage
Target Bases Covered	7 065 602	8 095 379
% Target Bases Covered	26.5	30.4
Targets with no coverage	0	0
Target Bases Not Covered	19 557 299	18 527 522
Due to N's	873 054	873 054
Due to repeats	17 525 323	17 096 460
% Target Bases Not Covered	73.5	69.6
Due to N's	3.3	3.3
Due to repeats	65.8	64.2
Total capture targets	17 734	19 923
Total capture space (bp)	7 065 602	8 095 379

3. Results and discussion

Over the past few years, the efficiency of target enrichment has been proven for the study of the nucleotide diversity of polyploid species with a large, repetitive and heterozygous genome (Bragina et al., 2019). This method has been used in various crops, including in the study of ecological adaptation in barley (Russell et al., 2016), the identification of wheat disease resistance genes (Steuernagel et al., 2016), the cataloging of mutations in rice (Henry et al., 2014).

At present, two main technologies are used for hybrid-capture applications for plants: solid-based and liquid-based hybridization (I. Terracciano et al., 2016). The most reliable commercial kits in studies on plant species were provided by Agilent Technologies (SureSelect), Roche NimbleGen (SeqCap EZ), MYcroarray (MYbaits) and Ion Torrent (TargetSeq).

For our study, Roche developed two types of probe design to cover 26.6 Mb of chromosome 5BS (Table 1). Type 1 is a standard design in which probes can be mapped to target region of the genome up to 20 times. Moreover, the algorithm for probe selection has two stages. At the first stage, the most specific probes are selected (up to 3 coincidences per target region) and less specific probes are superimposed on this region to increase coverage. But our region of interest contains the vast majority of repetitive regions, and the developers have created a second design to increase the % of covered regions. Type 2 is a design in which probes can be mapped to target region of the genome up to 50 times. As a result, the maximum number of covered region with a length of 26.6 Mb was about 8 Mb mainly due to repeats as well as due to the incompleteness of the reference genome. For our study, the second type of probe design was chosen.

Although the probes were designed for 'Chinese Spring', the low specificity of the probes allows this method to be used for ours wheat lines as well. Probes with low specificity have been successfully utilized in studies on divergent taxa, but they usually produce fewer variants than the taxon-specific probes (Bragg et al., 2016; Chau et al., 2018).

The wheat lines of population F4 (lines from the cross LrW(52) × hybrid_215) were captured with probes developed by Roche to cover the region under study on the basis of the reference genome.

As a result of sequencing, an average of 4.28 Gb and 3.51 Gb of reads was obtained for the wheat lines with Lr52 and without Lr5*2*, respectively. The total size of data obtained was about 39 Gb. The proportions of sequences with Q > 30 were 78.95 and 78.52 %, and the mean quality score Q values were 31.44 and 31.35 for the wheat lines with *Lr52* and samples without *Lr52*, respectively. The sequencing quality was estimated using the MAQ software (http://maq.sourceforge.net), a value of 30 and higher indicates a high quality of the sequences and unambiguous mapping with a small number of mismatches.

After preprocessing and quality control using the FASTX-toolkit utility (http://hannonlab.cshl.edu/fastx_toolkit/), the reads obtained were mapped onto the studied region of the wheat reference genome using the bwa-short algorithm in the short-read mapping software BWA (Li and Durbin, 2009). On average, 72.97 % of the sequenced reads were not mapped, 7.50 % were mapped one time and 19.53 %, more than once for samples with *Lr52*. For samples without *Lr52*, these parameters were 70.06, 9.26 and 20.69 %, respectively. The average alignment levels were 42 % and 46 % for resistant and susceptible samples, respectively. But in spite of the large number of unmapped sequences, the remaining reads allow us to study the region of interest.

Alignment of the reads to the wheat genome showed a high level of on-target enrichment efficiency. The 19923 baits designed correspond to a total of 8 Mb of genomic regions, associated with the *Lr52* gene. The number of baits per million base-pairs ranged from 505 to 1076, with an average of 749. On average, 94.9 and 94.4 % of base pairs in the bait

regions of resistant and susceptible plants, respectively, were covered by uniquely mapped reads at ≥ 5× depth. Due to the variable length of sheared DNA fragments in the prepped library and because the median length of the sheared fragments was greater than the length of the baits, we acquired additional coverage in regions adjacent to the baits. In the target regions, 10 wheat plants had narrow mean coverage ranges, 40× and 48×, for the leaf rust resistant and susceptible samples, respectively.

The data obtained are in a good agreement with the results of other groups. For example, in a work by Gardiner et al. (2016), a 110-MB NimbleGen capture probe set was used to enrich and sequence a doubled haploid mapping population of hexaploid wheat derived from the cross 'Avalon' × 'Cadenza'. An average depth of coverage of 45× was seen in the parental lines and 60× in the P1 bulk with an average of 98 % of the reference sequence being mapped to uniquely. In 2019, two gold standard capture probe sets for a gene and a putative promoter capture of hexaploid bread wheat were presented and validated, which are designed using recently developed genome sequence and annotation resources. Researchers demonstrate that the capture probe sets effectively enrich the high-confidence genes and putative promoter regions that were identified in the genome alongside a large proportion of low-confidence genes and associated promoters (Gardiner et al., 2019).

Our results suggest that sequence capture is a reliable approach to study and annotate of the chromosome region in both the gene space and non-repetitive intergenic regions.

4. Conclusions

The goal of this study was to test in-solution sequence capture of a 26.6-Mb region of chromosome 5BS of leaf rust resistant and susceptible wheat plants. We have demonstrate that the capture probe sets that have been designed against the hexaploid wheat cv. 'Chinese Spring' to enrich the genic portion of a mapping population, F4 (lines from the cross_LrW(52) × hybrid_215), effectively enrich the region of interest, providing an average gain of 40× mapping coverage. Our results show that in-solution sequence capture is a reliable method to enrich the gene space in complex plant genomes. The approach used here is general and the developed probes provide large amounts of genomic data to be used in many downstream analyses.

References

Bragg J.G., Potter S., Moritz C. Exon capture phylogenomics: Efficacy across scales of divergence. *Mol. Ecol. Resources*. 2016;16:1059–1068.

Bragina M.K., Afonnikov D.A., Salina E.A. Progress in plant genome sequencing: research directions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2019;23(1):38–48. DOI 10.18699/VJ19.459. (in Russian)

Chamala S., García N., Godden G.T., Krishnakumar V., Jordon-Thaden I.E., De Smet R., Barbazuk W.B. et al. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications Plant Sci.* 2015;3:1400115.

Chau J.H., Rahfeldt W.A., Olmstead R.G. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant Sci.* 2018;6(3):e1032.

Gardiner L.-J., Bansept-Basler P., Olohan L., Joynson R., Brenchley R., Hall N., O'Sullivan D.M., Hall A. Mapping-by-sequencing in complex polyploid genomes using genic sequence capture: a case study to map yellow rust resistance in hexaploid wheat. *Plant J.* 2016;87:403–419.

Gardiner L.-J., Brabbs T., Akhunov A., Jordan K., Budak H., Richmond T., Singh S., Catchpole L., Akhunov E., Hall A. Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. *GigaSci*. 2019;8(4):giz018. DOI 1093/gigascience/giz018.

Henry I.M., Nagalakshmi U., Lieberman M.C., Ngo K.J., Krasileva K.V., Vasquez-Gross H., Akhunova A. et al. Efficient genome-wide detection and cataloging of EMS induced mutations using exome capture and next-generation sequencing. *Plant Cell*. 2014;26:1382–1397.

Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760. DOI 10.1093/bioinformatics/btp324

Russell J., Mascher M., Dawson I.K., Kyriakidis S., Calixto C., Freund F., Bayer M. et al. Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nature Gen*. 2016;48:1024–1033.

Steuernagel B., Periyannan S.K., Hernández-Pinzón I., Witek K., Rouse M.N., Yu G., Hatta A. et al. Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nature Biotechnol.* 2016;34:652–655.

Terracciano I., Cantarella C., D'Agostino N. Hybridization-Based Enrichment and Next Generation Sequencing to Explore Genetic Diversity in Plants. In: Rogato A., Zazzu V., Guarracino M. (eds) *Dynamics Mathematical Models Biol*. Springer, Cham. 2016. DOI 10.1007/978-3-319-45723-9_10.

The International Wheat Genome Sequencing Consortium. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Sci*. 2018;361(6403). DOI 10.1126/science.aar7191.

Warr A., Robert C., Hume D., Archibald A., Deeb N., Watson M. Exome sequencing: Current and future perspectives. *Genes Genomes Genetics*. 2015;5:1543–1550.

Acknowledgements. The work is supported by the Ministry of Education and Science of the Russian Federation No. 0324-2019-0039 and by the Russian Foundation for Basic Research No. 17-04-00507.

Conflict of interest. The authors declare no conflict of interest.