Gene 533 (2014) 541–546

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods

PacBio sequencing of gene families — A case study with wheat gluten genes

Wei Zhang [a], Paul Ciclitira [b], Joachim Messing [a,*]

[a] Waksman Institute of Microbiology, Rutgers University, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA
[b] King's College London, Division of Nutritional Sciences, Rayne Institute (KCL) St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK

ARTICLE INFO

Article history:
Accepted 3 October 2013
Available online 19 October 2013

Keywords:
PacBio sequencing
Gene family
Gluten diversity
Copy number variation
Repetitive sequence

ABSTRACT

Amino acids in wheat (*Triticum aestivum*) seeds mainly accumulate in storage proteins called gliadins and glutenins. Gliadins contain α/β-, γ- and ω-types whereas glutenins contain HMW- and LMW-types. Known gliadin and glutenin sequences were largely determined through cloning and sequencing by capillary electrophoresis. This time-consuming process prevents us to intensively study the variation of each orthologous gene copy among cultivars. The throughput and sequencing length of Pacific Bioscience RS (PacBio) single molecule sequencing platform make it feasible to construct contiguous and non-chimeric RNA sequences. We assembled 424 wheat storage protein transcripts from ten wheat cultivars by using just one single-molecule-real-time cell. The protein genes from wheat cultivar Chinese Spring are comparable to known sequences from NCBI. We demonstrated real-time sequencing of gene families with high-throughput and low-cost. This method can be applied to studies of gene amplification and copy number variation among species and cultivars.

1. Introduction

Wheat is an important food source worldwide. Hexaploid wheat (*Triticum aestivum*) is the most common and widely grown wheat in continental climates in the regions such as Ukraine, central USA, Canada, Australia, Northern Europe, China and New Zealand. The hexaploid wheat genome has recently been sequenced and analyzed by using whole-genome shotgun DNA sequencing. A total of 94,000–96,000 gene models have been proposed (Brenchley et al., 2012). Hexaploid wheat cultivars accumulate 8–15% protein in their seeds, 80% of which is gluten with about 30% gliadins and 50% glutenins. Glutenins with high molecular weight (HMW) and low molecular weight (LMW) aggregate to polymers, whereas gliadins with α/β-, γ-, and ω-types mainly remain monomeric (Wieser, 2007) .

Gluten proteins belong to the superfamily of prolamins, which are present in all grass seeds. One of the key characteristics of prolamins is the large number of genes in contiguous regions, like zeins in maize (Geraghty et al., 1981; Llaca and Messing, 1998) and glutens in wheat (Anderson et al., 1997; Cassidy et al., 1998; Sabelli and Shewry, 1991). Another hallmark of all prolamins is a block of amino acids, rich in glutamine and proline in tandem arrangements in the central portion of the protein (Brenchley et al., 2012; Geraghty et al., 1981). We have sequenced the maize zein clusters by using BAC contigs to differentiate each gene (Llaca and Messing, 1998; Song and Messing, 2002; Song et al., 2001). Gluten amplicons have been isolated through PCR amplification and separated by cloning (Qi et al., 2009). Both approaches are costly and time-consuming, particularly, if we need to compare genes among different cultivars. The next-generation sequencing platforms such as Illumina offer throughput, but the generated short sequencing reads are a critical barrier to assemble repetitive genes, which may result in inadvertently joining of different gene copies into chimeric molecules.

A new approach is therefore required with both high-throughput and long sequence reads to ensure assembly of single molecules to sequence genes with repetitive sequences. The new Pacific Bioscience RS (PacBio) third-generation sequencing platform offers high throughput of 50,000–70,000 reads per reaction and a read length over 3 kb. A few examples of the use of PacBio platform includes the de-novo assembly of bacterial genomes (Rasko et al., 2011), closing genome gaps (Zhang et al., 2012), getting full length transcripts and finding transcript variants (Ocwieja et al., 2012), sequencing repeat stretches (Loomis et al., 2012), and recently sequencing the human HLA genes (Lind et al., 2012). Although this system has an inherent sequencing error rate, the errors are random and can be overcome by redundancy. Moreover, the SMRTbell template structure enables the polymerase to pass multiple times on a single molecule. Therefore, the obtained CCS sequences (circular consensus sequences) have improved accuracy (Travers et al., 2010). A total of 424 transcripts were assembled from ten wheat cultivars by using just one SMRT (single molecule, real time) cell from Pacific Bioscience RS (PacBio). This technical advance is efficient and economical

Abbreviations: HMW, high molecular weight; LMW, low molecular weight; BAC, bacterial artificial chromosome; PacBio, Pacific Bioscience; HLA, human leukocyte antigen; SMRT, single molecule, real time; CCS, circular consensus sequence; ZMW, zero mode waveguides; GRIN, germplasm resources information network; CS, Chinese Spring; CAP3, Contig assembly program version 3.

* Corresponding author at: Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. Tel.: +1 732 445 4257; fax: +1 732 445 0072.
E-mail address: messing@waksman.rutgers.edu (J. Messing).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.10.009

in generating sequences from gene families and should be applicable to other phylogenetic studies.

2. Materials and methods

2.1. Plant materials

We randomly selected ten wheat lines from the largest possible geographic regions worldwide for our study (Table S1), all cultivars were ordered through USDA GRIN germplasm resource website (http://www.ars-grin.gov/). All lines were hexaploid, *T. aestivum* subsp. *aestivum*, except for Baxtor, a tetraploid from *Triticum turgidum* subsp. *polonicum*.

2.2. Sample preparation

Wheat immature seeds were collected at 10–15 days after anthesis from spikelets, and immediately put into liquid nitrogen and stored at −80 °C. Total RNA was extracted with Trizol reagent (Invitrogen) and run through RNeasy (Qiagen) column for purification. A total of 1–2 μg of total RNA was reverse transcribed. PCR was conducted using degenerate primers that were designed at the start codon and stop codon of each gluten family. 1 primer pair was used for α/β-gliadins (alphaF1: ATGAAGACCTTTCTCATCCTTG, alphaR1: TCAGTTRGTACCGAA GATGCCA), 1 primer pair for LMW-glutenins (LMWF1: ATGAAGACC TTCCTCRTCTT, LMWR1: TTATCAGTAGMVACCAACTCC) and 3 primer pairs for γ-gliadins (gammaF1: ATGAAGACCTTAYTCATCCT, gammaR1: TTTTCATTGKCCACTGATGCC, gammaR2: TTTTCATTGKCCACCAATGCC, gammaR3: TCATCGATATTGGCCACCAATG). For each of the wheat lines, a unique DNA ligator was added to the front of all primers. As a result, 10 unique DNA ligator sequences could differentiate PCR products from the 10 wheat lines.

PCR products are then purified. 200 ng of PCR products of α/β-gliadins, γ-gliadins and LMW-glutenins in each wheat line was pooled into a single mixed DNA pool. This DNA sample was then sent to the Biotechnology Center of the University of Florida in Gainesville (http://www.biotech.ufl.edu/) for library preparation and PacBio sequencing.

2.3. LMW markers used in our study

For LMW-glutenins, group-specific primers were used in the same annealing temperature as in Long et al. (2005) except for group 1 and group 5 primers, for which 60 °C and 65 °C were used as annealing temperatures, respectively.

2.4. Protein extraction and Western blotting

Fifty milligrams of wheat flour from mature seeds were incubated in 1.25 ml extraction buffer (62.5 mM Tris–HCl, pH 6.8, 2% SDS, 1.5% DTT, 10% glycerol, and 0.002% Bromophenol Blue) at 65 °C for 30 min. Samples were centrifuged at 12,000 rpm for 10 min and heated at 95 °C for 5 min. 5 μl of the protein sample was separated in 4–15% TGX gel (Bio-Rad) at 200 V for about 30 min. Proteins were transferred to nitrocellulose membrane using Biorad mini Trans-Blot cell under 100 V for 1 h. Western blot was performed following manuals of Amersham ECL Western blotting detection system. Monoclonal antibodies PN3 and CDC5 were raised against α/β-gliadins: PN3 recognizes peptides QQQPFP within α/β-gliadins and cross-reacts with LMW glutenins through QQQP; CDC5 was raised against peptide QLQPFPQPQLPYPQPQLPY in α/β-gliadins.

2.5. Sequence analysis

All generated circular consensus sequences (CCS) were separated by barcode sequences into 10 groups. Sequences within groups were blasted against α/β-gliadins, γ-gliadins and LMW-glutenins, and further separated into these three gene families. Sequences within each family were then de-novo assembled using SeqmanNGen under the following parameters: mer size = 50, min match percent = 90, gap penalty = 50, and max gap = 3. The resulting contigs were then used as templates for templated-assembly under the following parameter: mer size = 35, min match percent = 80, gap penalty = 20, and max gap = 3. Contig sequences from the de-novo assembly were then checked in the templated assembly for errors to generate final contigs of transcripts.

2.6. Protein sequence alignment

Protein sequences were transcribed from the final contigs in each gene family using the EBI online tool http://www.ebi.ac.uk/Tools/st/emboss_transeq/. Protein sequences are then aligned using Clustal Omega. Phylogenetic trees were viewed in Mega5 program.

3. Results

3.1. Experimental design

In order to sequence mRNAs with internal repeats, which are encoded by a large set of gene copies at various levels such as glutens, we had to sequence single molecules at deep coverage. In the new PacBio system, one SMRT cell is patterned with 150,000 zero mode waveguides (ZMWs). Each cell has DNA polymerase molecules that can bind to a DNA template. In average, half of these ZMWs are used to generate sequences. If we were targeting 100 genes in each of the 10 wheat cultivars, we would have to amplify 1000 genes in total. Seventy-five thousand (75,000) reads could give us an average of 75× coverage. To test this sequencing method for sequencing prolamin gene families, we amplified α/β-gliadins, γ-gliadins and LMW-glutenins from cDNAs in different wheat cultivars, barcoded these amplicons, and pooled all for one sequencing run. The processed data could then be grouped by genetic background and protein subfamily origin (Fig. 1).

3.2. Single molecule sequencing of gluten mRNAs

We randomly selected ten wheat lines worldwide from the GRIN database, which show great variation in the plant phenotypes (Table S1). Using bar-coded PCR-primers for each cultivar, we were able to generate 77,462 pass filter sub-reads (51.6% usage of ZMWs) with an average read length of 3050 bp and mean read quality of 0.835 from non-cloned cDNAs. Out of the pass filter sub-reads, 32,863 ccs (circular consensus sequence) reads were used in our subsequent analysis, and 31,234 ccs reads could be grouped into specific gene families based on sequence homology. Because of the error rate of the PacBio system a sequence assembly was necessary. Because errors appear to be random, the consensus sequences are easily built due to sequence redundancy. These consensus sequences are termed "contigs" which represent transcripts of single gene copies. Sequence was confidently called when there were over five consensus sequences in one assembled region. We also removed contigs with less than 15 assembled sequences. After collapsing identical contig sequences for each sample, the number of sequences assembled into each contig was summarized in Table S2. Most of the contigs contain 20–100 sequences with some of the abundant contigs having over 300 sequences. In total, we assembled 424 wheat gliadin and glutenin transcripts in the ten wheat lines (Table 1). These 424 transcripts coded for 345 proteins and 79 truncated proteins due to early stop codons.

3.3. Comparison to reference gene set

To test the validity of the assembled sequences, we compared our assembled proteins from Chinese Spring (sample 1) with the NCBI protein database (Table 2). Chinese Spring (CS) is the most common