Comparative genomic hybridization reveals structural diversity in barley.

Gary J. Muehlbauer [1,2], M. Munoz-Amatriain [1], T. Richmond [3], J. Jeddeloh [3], A. Landreman [3], B. Steuernagel [4], S. Taudien [4], M. Platzer [4], U. Scholz [4], M. Mascher [4] , R. Ariyadasa [4], T. Nussbaumer [5], K. Mayer [5], S.R. Eichten [2], N.M. Springer [2] and N. Stein [5].

[1] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA; [2] Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA; [3] Roche NimbleGen, Inc., Research and Development, Madison, WI 53719, USA; [4] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Gatersleben, Germany; and [5] Munich Information Center for Protein Sequences/Institute of Bioinformatics and Systems Biology, Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany.

Structural variation is characterized by chromosomal rearrangements (inversions, translocation), copy number variation (CNV) and presence/absence variation (PAV). CNV and PAV have been identified in numerous animal and plant species and there is a growing awareness of the potential of these variants to impact phenotypes. Comparative genomic hybridization (CGH) is a powerful approach to assess CNV and PAV. We developed a CGH array for barley composed of 2.1M probes from 211,669 target fragments (10 probes per fragment) capturing ~50 Mbp of gene space. Initially, we validated the array on a wheat–barley addition line carrying the long arm of barley chromosome 3H and showed that the technology was highly robust. Subsequently, we used the array to examine structural variation in eight cultivated barleys, six wild barley accessions, and the Oregon Wolfe barley mapping population parents. Our results demonstrated that between 2.4–3.5% and 4.4–4.5% of the fragments exhibited structural variation in cultivated and wild barley, respectively. These results demonstrate, as expected, that wild barley germplasm carries a higher level of structural variation than cultivated barley. In total, 15.6% of the fragments exhibited structural variation, indicating that a substantial proportion of the barley gene space has the potential to exhibit structural variation. Over 35% of the events were detected in only a single genotype, suggesting that many of these events may be rare. The distribution of structural variants was predominately located near the telomeres. Noteworthy, chromosome 4H exhibited a reduced amount of structural variation indicating that recombination is restricted on this chromosome. Our results provide the first genome-wide view of the structural variation in the barley genome across a range of genotypes. Annotation and further characterization of the structural variants is ongoing and will be presented.

kmer-based contamination screening in the wheat chromosome survey sequencing project.

Jonathan Wright and Ricardo Ramirez-Gonzalez on behalf of the IWGSC. The Genome Analysis Centre, Norwich Research Park, Norwich, UK.

The Genome Analysis Centre (TGAC) is coordinating the hexaploid wheat chromosome survey sequencing project on behalf of the International Wheat Genome Sequencing Consortium (IWGSC). This project aims to generate sequence reads from each of the 42 chromosome arms of hexaploid wheat and assemble these reads using the ABySS assembler. Advances in chromosome sorting in wheat have enabled this approach as a viable method to reduce the complexity of this large, highly repetitive genome. Chromosome sorting uses the size of each chromosome arm to separate it from the other arms and purities of around 90% can be achieved.

After assembling the reads from each arm, we aligned bin-mapped wheat ESTs to the assemblies to check for gross contamination and observed a general trend of background contamination from other chromosome arms. In some cases, this background contamination was higher than expected. In order to investigate this further, we developed a novel kmer-based analysis whereby each assembly was reduced to kmers (overlapping words of length k), then compared with all kmers from each of the other assemblies. We found that when a large kmer length was used, each chromosome arm could easily be distinguished from the other arms using this method. In order to reduce high background contamination, we repeated flow-sorting and sequencing for chromosome arms that were heavily contaminated and observed either the same pattern of contamination (indicating the contamination was due to an artifact of the flow-sorting process), or a different pattern of contamination (indicating random contamination specific to each sequencing run), or a combination of both. To develop a bioinformatics cleaning strategy for the contaminated assemblies, we extended the kmer analysis to the sequence reads in order to identify reads from different runs that contain shared kmers and, thus, reject reads that appear to be random contamination. These reads were used to generate ‘clean’ assemblies. We found this kmer-based analysis enabled us to improve the quality of the problematic assemblies.