# New tools for wheat genetics and breeding: Genome-wide analysis of SNP variation*

*E. Akhunov[1], S. Chao[2], V. Catana[1], D. See[3], G. Brown-Guedira[4], A. Akhunova[5], J. Dubcovsky[6]; C. Cavanagh[7] and M. Hayden[8]*

## Abstract

Single nucleotide polymorphism (SNP) is one of the most broadly distributed types of molecular variation in a genome which, along with the availability of cost- and labor-effective genotyping platforms, make it the marker of choice for many crops. Our work is aimed at the development of a dense set of genetically mapped SNP markers for low-cost high-throughput genotyping of wheat germplasm. Next generation sequencing of normalized cDNA libraries was used for developing gene-associated SNPs in polyploid wheat. A total of 7.5 million 454 reads were generated from cDNA libraries of 10 wheat cultivars from US and Australia and processed for discovering SNPs using a bioinformatical pipeline specifically designed for variant discovery in polyploid transcriptomes. A total of 25,000 high-quality SNPs distributed among 14,500 EST contigs were identified. All these SNPs were validated by comparison with RNA-seq data generated from an additional set of 17 U.S. and Australian cultivars. A total of 9,000 genome-wide common SNPs were selected for designing an Illumina iSelect assay. Preliminary testing showed that more than 95% of SNPs produce high-quality genotype calls with up to 70% being polymorphic in a diverse sample of U.S. and Australian cultivars with a minor allele frequency >0.05. The assay is currently being used for studying patterns of genetic diversity in a worldwide collection of wheat cultivars and for developing a high-density SNP map. A long term goal of this initiative is to advance wheat research and breeding by developing genetic and genomic tools for efficient analysis of agronomic traits using high-resolution linkage and association mapping and deploying SNP markers in breeding programs.

[1]Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, U.S.A.; [2]USDA-ARS Biosciences Research Laboratory, Fargo, ND, U.S.A.; [3]USDA Western Regional Small Grains Genotyping Lab, Johnson Hall, WSU, Pullman, WA, U.S.A.; [4]USDA-ARS Eastern Regional Small Grains Genotyping Lab., 4114 Williams Hall, NCSU, Raleigh, NC, U.S.A.; [5]Integrated Genomics Facility, Kansas State University, Manhattan, KS, U.S.A.; [6]Department of. Plant Sciences, University of California, Davis, CA, U.S.A.; [7]CSIRO, Food Futures National Research Flagship, Canberra, ACT 2601, Australia; [8]Department of Primary Industries Victoria, Victorian AgriBiosciences Center, 1 Park Drive, Bundoora, VIC 3083, Australia. **E-mail: eakhunov@ksu.edu**

*This is an unedited original draft. A revised version will appear in the workshop proceedings published later this year.

## Key words:

High-throughput genotyping; next-generation sequencing; polyploid wheat; Single Nucleotde Polymorphism

## Introduction

Single nucleotide polymorphism (SNP) is the major type of intra-specific genetic variation and is widely distributed across genomic regions. A dense set of molecular markers covering the entire genome is a pre-requisite for high-resolution genetic analysis of agronomically important traits and deployment of efficient breeding strategies in crops (Breseghello and Sorrells 2006; Yu and Buckler 2006; Wang et al. 2007). Although, a wide variety of molecular markers is available (Vos et al. 1995; Gupta et al. 1999; Akbari et al. 2006; Flavell et al. 1998), those based on SNPs are the most efficient genotyping tool (Rostoks et al. 2006; Hyten et al. 2009; Chao et al. 2010).

Next-generation sequencing technologies capable of generating up to gigabases of sequence data have made SNP discovery a routine procedure for any organism. Massively parallel pyrosequencing technology has been successfully used to detect SNPs in maize and *Eucalyptus* transcriptomes (Barbazuk et al. 2007; Novaes et al. 2008) and in reduced representation genomic libraries of maize and cattle (van Orsouw et al. 2007; Van Tassell et al. 2008). DNA sequence capture approaches have been applied for targeted enrichment and sequencing of selected genomic regions for variant discovery in human and maize genomes (Albert et al. 2007; Porreca et al. 2007; Gnirke et al. 2009). New sequencing technologies make feasible the discovery of thousands of SNPs in domesticated crop species with low levels of genetic diversity (Choi et al. 2007).

Technical advances facilitating SNP discovery have been paralleled by the development of cost- and labor-efficient high-throughput genotyping technologies capable of genotyping thousands of individuals at thousands of SNP sites. A large variety of genotyping systems satisfying these requirements are now available (Syvänen et al. 2005). For example, the Illumina BeadArray platform combined with the GoldenGate assay is able to generate genotype data for several thousand polymorphic sites in 96 individuals in a single reaction (Oliphant et al. 2002). Molecular Inversion Probe (MIP) technology (Hardenbol et al. 2005) and the Illumina Infinium assay (Steemers and Gunderson 2007) can be used to genotype tens of thousands to hundreds of thousands of SNPs in a large number of individuals.

Compared to other important crops, SNP-based assays for wheat genotyping have only recently become available (Akhunov et al. 2009; Chao et al. 2010). Polyploidy and the low level of polymorphism in cultivated germplasm were the major challenges for SNP discovery in wheat and required the development of labor-intensive and expensive experimental approaches (Akhunov et al. 2010). However, in recent years a marked change has come with the availability of next-generation sequencing technologies. These technologies enable multiple large scale SNP discovery efforts (wheatgenomics.plantpath.ksu.edu/IWSWG/) that use the power of next-generation sequencing of genomic DNA and transcriptomes from multiple wheat cultivars for SNP detection.

Here, we present the development of gene-associated SNPs by large-scale transcriptome sequencing of a diverse panel of wheat cultivars from the U.S. and Australia. These SNPs were used to design a 9,000-plex SNP assay based on the Illumina Infinium platform. The goal of this initiative was to develop bioinformatical procedures for SNP calling in next-generation sequence data generated for polyploid transcriptomes, and to test the utility of the Infinium platform for high-throughput genotyping of polyploid wheat cultivars.

***cDNA preparation and normalization***

RNA samples were extracted using the RNeasy Plant Mini Kit (QIAGEN). Concentration and purity of total RNA was checked on a Nanodrop Spectrophotometer. The RNA integrity was evaluated on a Bioanalyzer (Agilent) and by standard formaldehyde agarose gel electrophoresis. cDNA libraries for sequencing using Illumina GAII$_x$ and HiSeq2000 platforms were prepared according to manufacturer's instructions (Illumina Corp., San Diego), while cDNA libraries for 454 sequencing were prepared as follows. First-strand cDNA synthesis was performed according to SMART cDNA synthesis technology (Clontech Laboratories, Inc.) using 3' SMART CDS modified Primer II A primer (5'-AAG CAG TGG TAT CAA CGC AGA GTA CTT TTG T(9) C T(10) VN-3') and SuperScript III reverse transcriptase (Invitrogen). Double-stranded cDNA was amplified by long-distance (LD) PCR using the Advantage 2 PCR Enzyme System (Clontech Laboratories, Inc). Amplification was performed on a thermal cycler (Applied Biosystem) with the following PCR parameters: 1 cycle at 95˚C (1 min); 16 cycles at 95˚C (15 s), 65˚C (30 s), 68˚C (6 min); and 4˚C (45 min) (optional). Double-stranded cDNA was checked on 1.1% agarose/EtBr gels in 1XTAE buffer and purified with the QIAquick PCR Purification Kit (QIAGEN). The double-stranded cDNA was normalized using the TRIMMER cDNA Normalization Kit (EVROGEN), which is based on a unique DSN (duplex-specific nuclease) normalization technology and is specially developed for normalization of cDNA enriched with full-length sequences. The efficiency of normalization was examined by determining the abundance of two highly expressed transcripts before and after normalization using quantitative Real-Time PCR. All quantitative PCR experiments were performed on an iCycler Real-Time PCR system (BioRad Laboratories) using IQ SYBR Green Supermix (BioRad Laboratories). The PCR conditions were 1 cycle at 95ºC (5 min); 40 cycles at 95ºC (15 s), 55ºC (15 s), 72ºC (50 s); followed by the melting curve program.

***Data processing and analysis***

Custom Perl scripts were used for quality trimming of RNA-seq data generated using Illumina GAII$_x$ and HiSeq2000 platforms. The program Lucy was used for removal of adaptor sequences and quality trimming of 454 sequence reads (Li and Chou 2004). For each wheat line, a reference cDNA sequence was built using MIRA software (Chevreux et al. 2004). Each reference set was then consecutively used for read mapping using Mosaik software (bioinformatics.bc.edu/marthlab/Mosaik). SNP discovery was performed using the Bayesian algorithm implemented in the GigaBayes software. An additional post-processing filtering step was applied to select high-quality SNPs, which removed all variable sites having alleles covered by less than 3 reads in the alignments. This filter was based on the results of empirical validation by Sanger re-sequencing of randomly selected SNP-harboring gene fragments. SNPs and their flanking sequences were extracted and used for comparison with RNA-seq data generated for an additional set of 17 U.S. and Australian cultivars. This analysis allowed the frequency of discovered SNPs in U.S. and Australian cultivars to be estimated more precisely.

***Design of 9,000-plex SNP assay***

Several criteria were applied for selecting 9,000 SNPs for an Infinium iSelect assay design. First, repetitive elements in reference sequences were detected by comparing them with the TREP (http://wheat.pw.usda.gov/ITMI/Repeats/) and GIRI (www.girinst.org) databases. Second, the distribution of cDNA contigs used for SNP discovery across the wheat genome was inferred by comparing them with the genomes of Brachypodium and rice using the blastn program. Third, SNPs were ranked according to their minor allele frequency (MAF) in the discovery panel and their

distribution between US and Australian populations. Priority was given to SNPs that showed high MAF and were shared by U.S. and Australian cultivars. Finally, SNPs and their flanking sequences were submitted to Illumina for processing by the Assay Design Tool which generates scores for each SNP varying from 0 to 1. SNPs with scores above 0.6 and having a high probability to be converted into a successful genotyping assay were selected.

## Results and Discussion

***Transcriptome sequencing and SNP discovery***

Transcriptomes of a diverse set of 27 wheat lines including Chinese Spring and cultivars from U.S. and Australia were sequenced using normalized cDNA libraries prepared from RNA isolated from multiple tissues collected at different developmental stages or subjected to different stress treatments. A complete list of the cultivars sequenced can be found on the project website: http://wheatgenomics.plantpath.ksu.edu/snp/. Ten cultivars were sequenced using Roche 454 technology to generate nearly 7 million reads (Table 1) and 17 cultivars were sequenced using the GAII and HiSeq2000 platforms to produce almost 500 million Illumina reads.

The primary challenge for SNP discovery in the hexaploid wheat genome is to distinguish divergence among the wheat sub-genomes from variation between wheat lines. Failure to filter out divergent sites can significantly inflate the false SNP discovery rate. Hence, we developed procedures for effective separation of true inter-cultivar variants from genome-specific mutations differentiating the wheat sub-genomes and applied this approach to discover a genome-wide distributed set of genic SNPs (Fig. 1).

**Fig. 1** *SNP discovery workflow*

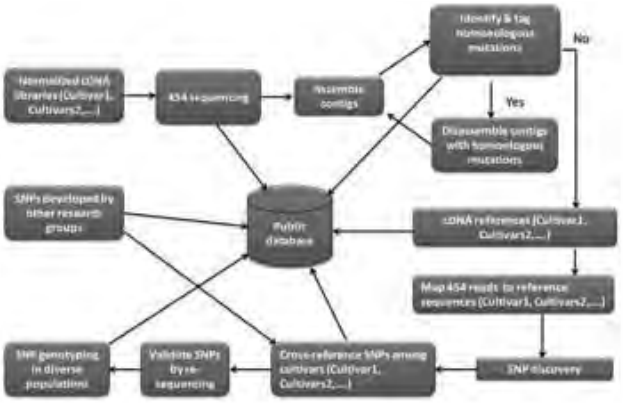


**Table 1** *454 transcriptome data generated for 10 wheat cultivars*

| Wheat line | Origin | No. of reads generated | Total bases, bp | No. reference contigs | Total contig length, bp |
|---|---|---|---|---|---|
| Excalibur | Australia | 1,547,934 | 565,519,405 | 116,984 | 95,773,580 |
| Kukri | Australia | 1,706, 047 | 605,719,594 | 118,506 | 98,089,478 |
| RAC875 | Australia | 1,650,614 | 600,811,369 | 121,193 | 98,547,050 |
| Bobwhite | USA | 1,303,861 | 491,025650 | 67,788 | 42,695,800 |
| CAP7 | USA | 284,070 | 100,314,837 | 13,321 | 5,612,611 |
| CAP8 | USA | 188,601 | 64,914,768 | 10,017 | 4,336,540 |
| CAP11 | USA | 170,792 | 59,922,739 | 9,200 | 3,959,549 |
| CAP12 | USA | 191,450 | 63,713265 | 9,222 | 3,848,399 |
| Jagger | USA | 80,699 | 17,718,199 | 11,060 | 3, 677,900 |
| Chinese Spring | China | 98,219 | 21,285,912 | | |
| **Total** | | **7,222, 287** | **2,590,945, 738** | **477,291** | **356,540, 907** |

The preparation of reference sequences for read mapping using the MIRA assembler and 454 sequence data resulted in 477,291 cDNA contigs with an average length of 750 bp (Table 1). Sequential mapping of all 7,222,287 reads to each set of cultivar-specific cDNA contigs using MOZAIK software was followed by SNP discovery using the Bayesian algorithm implemented in the GigaBayes program. The minimum number of reads representing each SNP allele in alignment was considered as the most critical parameter for the discovery of true SNPs, and was adjusted using empirical data obtained by Sanger re-sequencing of 96 gene fragments. Filtering SNPs for alleles covered by at least three reads in the alignments resulted in an 85-90% validation rate. A total of 81,688 SNPs were discovered in all possible pair-wise comparisons between 10 cultivars which, after removing redundant SNPs overlapping among different pair-wise comparisons, resulted in a set of about 25,000 unique high-quality SNPs (Table 2).

All SNPs were deposited into a searchable MySQL database that is available at http://wheatgenomics.plantpath.ksu.edu/snp/. In the current version of the database, users can obtain the list of SNPs polymorphic between any two given wheat cultivars used for SNP discovery and search for SNPs by sequence similarity using the BLASTN program. The output is user-adjustable and contains wheat deletion-bin map data, the best BLASTN and BLASTP hits in the NCBI database along with the description of these hits. In the output of the BLASTN search, users can also obtain SNP genotypes of cultivars included into the discovery panel.

***9,000 SNP iSelect BeadChip***

The utility of SNPs for genotyping a broad range of wheat cultivars strongly depends on the distribution of SNP alleles among populations. SNP alleles shared between diverse populations will have a high likelihood of being polymorphic in a broad range of populations. We used Illumina RNA-seq data obtained for an additional 17 U.S. and Australian cultivars to assess the frequency and distribution of SNPs discovered by 454 transcriptome sequencing in 10 cultivars. The estimated fraction of SNPs shared between U.S. and Australian wheat cultivars was 45%. These SNPs were primarily targeted for inclusion into the genotyping assay design. In addition, 849 SNPs from the first 1536-plex wheat oligo pool assay (Chao et al. 2010) were also included in the custom 9K iSelect design.

The Infinium iSelect BeadChip was tested using a diverse panel of wheat cultivars including breeding lines and parents of critical mapping populations (US CAP, ITMI and MAGIC). Preliminary results suggest that out of 9,000 attempted bead types nearly 95% produce scorable data. After quality filtering based on the clustering patterns, we obtained about 8,000 SNP assays generating high quality data. In a panel of 181 diverse wheat cultivars from U.S. and Australia 70% of SNP assays could be scored as polymorphic with minor allele frequency >0.05. We are currently performing detailed analyses of genotyping data with the goal of developing SNP calling algorithms with increased sensitivity and accuracy. The 9000 SNP iSelect BeadChip will be used for genotyping a large worldwide collection of wheat cultivars and the development of a high-density SNP-based genetic map.

**Table 2** ***Number of SNPs identified in pair-wise comparisons***

| | Excalibur | Bobwhite | Jagger | Kukri | RAC875 | CAP 7 | CAP 8 | CAP 11 | CAP 12 |
|---|---|---|---|---|---|---|---|---|---|
| Ch. Spring | 1037 | 902 | 341 | 1,167 | 1,293 | 122 | 95 | 93 | 76 |
| Excalibur | | 5,757 | 798 | 16,238 | 14,817 | 188 | 194 | 147 | 131 |
| Bobwhite | | | 617 | 4,939 | 5,954 | 299 | 351 | 264 | 286 |
| Jagger | | | | 817 | 856 | 70 | 63 | 53 | 47 |
| Kukri | | | | | 17,553 | 158 | 173 | 132 | 91 |
| RAC875 | | | | | | 182 | 192 | 138 | 121 |
| CAP 7 | | | | | | | 1,087 | 925 | 680 |
| CAP 8 | | | | | | | | 852 | 809 |
| CAP 11 | | | | | | | | | 583 |

## Acknowledgements

This project is funded by the USDA National Institute of Food and Agriculture (CRIS0219050), USDA Triticeae CAP (2011-68002-30029), and the Grains Research and Development Corporation (GRDC), Australia.

## References

Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S, Uszynski G, Mohler V, Lehmensiek A, Kuchel H, Hayden MJ, Howes N, Sharp P, Vaughan P, Rathmell B, Huttner E, Kilian A (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome Theor Appl Genet 113:1409-1420

Akhunov ED, Akhunova AR, Anderson, OD, Anderson, JA, Blake, N, Clegg, MT, Coleman-Derr, D, Conley, EJ, Crossman, CC, Deal, KR, Dubcovsky, J, Gill, BS, Gu, YQ, Hadam, J, Heo, HY, Huo, N, Lazo, GR, Luo, MC, Ma, YQ, Matthews, DE, McGuire, PE, Morrell, P, Qualset, CO, Renfro, J, Tabanao, D, Talbert, LE, Tian, C, Toleno, D, Warburton, M, You, FM, Zhang, W, Dvorak, J (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes BMC Genomics 11:702

Akhunov ED, C Nicolet, J Dvorak (2009) Single nucleotide polymorphism genotyping in polyploid wheat with Illumina GoldenGate assay. Theor Appl Genet 119:507-517

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4:903-905

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910-918

Breseghello F, Sorrells ME (2006a) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177

Chao, S., Dubcovsky, J., Dvorak, J., Luo, M.C., Baenziger, S.P., Matnyazov, R., Clark, D.R., Talbert, L.E., Anderson, J.A., Dreisigacker, S., Glover, K., Chen, J., Campbell, K., Bruckner, P.L., Rudd, J.C., Haley, S., Carver, B.F., Perry, S., Sorrells, M.E., Akhunov, E.D. (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (Triticum aestivum L.). BMC Genomics 11(1):727

Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS, Hwang EY, Yi SI, Young ND, Shoemaker RC, van Tassell CP, Specht JE, Cregan PB (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685-696

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 2004 Jun;14(6):1147-59. Epub 2004 May 12

Flavell AJ, Knox MR, Pearce SR, Ellis TH (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J 16:643-650

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182-189

Gupta PK, Varshney RK, Sharma PC, Ramesh B (1999) Molecular markers and their applications in wheat breeding. Plant Breeding 118:369-390

Hardenbol P, Yu F, Belmont J, MacKenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J, Erbilgin A, Falkowski M, Fitzgerald R, Ghose S, Iartchouk O, Jain M, Karlin-Neumann G, Lu X, Miao X, Moore B, Moorhead M, Namsaraev E, Pasternak S, Prakash E, Tran K, Wang Z, Jones HB, Davis RW, Willis TD, Gibbs RA (2005) Highly multiplexed molecular inversion probe genotyping: Over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res 15:269-275

Hyten DL, Song Q, Choi IY, Yoon MS, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945-952

Li S, Chou HH. (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. Bioinformatics 20:2865-2866

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9:312

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques Suppl:56-58

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656-18661

Steemers FJ, Gunderson KL (2007) Whole genome genotyping technologies on the BeadArray platform. Biotechnol J 2:41-49

Syvänen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37 Suppl:S5-1Van Orsouw NJ, Hogers RC, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJ (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE 2:e1172

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5:247-252

Vos P, Hogers R, Bleeker M, Reijans M, Lee Th van der, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407-4414

Wang J, Chapman SC, Bonnett DG, Rebetzke GJ, Crouch J (2007) Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. Crop Sci 47:582-590

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155-160

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J (2007) Multiplex amplification of large sets of human exons. Nat Methods 4:931-936