J. Crop Sci. Biotech. 2014 (September) 17 (3) : 155 ~ 160
DOI No. 10.1007/s12892-014-0022-4

RESEARCH ARTICLE

Mining of Gene-Based SNPs from Publicly Available ESTs and Their Conversion to Cost-Effective Genotyping Assay in Sorghum [*Sorghum bicolor* (L.) Moench]

Yemane Girma[1,2], Dadakhalandar Doddamani[1,3], Bashasab Fakrudin[1,4,*], Rajkumar[1,5], Sadik Ahmed Wasik Ahmed[1], Sheweta Gujar[1,6], Suvarna Patil[1,7], Gurusiddesh Hiremath[1,8]

[1]*Institute of Agri-Biotechnology, University of Agricultural Sciences, Dharwad - 580005, India*
[2]*Haramaya University, Diredawa, Ethiopia*
[3]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru - 502324, India*
[4]*Present Address: College of Horticulture, University of Horticultural Sciences campus, GKVK, Bangalore - 560065, India*
[5]*Main Cotton Research Station, Navsari Agricultural University, Athwa farm, Surat 395007*
[6]*Max Planck Institute of Developmental Biology, Department of Biochemistry, Spemannstrasse Tuebingen, Germany*
[7]*University of Horticulture Sciences, Bagalkot, Karnataka, India*
[8]*Karnataka University, Dharwad, Karnataka, India*

Received: February 20, 2014/April 18 2014/June 13, 2014

Abstract

Single Nucleotide Polymorphisms (SNPs) are the commonest type of nucleotide variation distributed throughout the genome and have enormous potential to saturate genetic maps. However, their identification is constrained by the huge investment required for their detection. In this study, we used publicly available EST (Expressed Sequence Tag) sequences to identify SNPs in *Sorghum bicolor*. A total of 12,421 putative SNPs were identified from 2,921 contiguous transcripts leading to an average SNP interval of one putative SNP for every 275.26 bp. The proportion of transition type mutations (0.598) was larger than transversion types conforming to biological expectations. In order to demonstrate the utility of the SNPs for development of markers with relatively cheap assays, we experimentally validated SNPs using Single Strand Conformation Polymorphism (SSCP) technique in sorghum accessions, which are used as parents for development mapping populations. Genotyping these parents of mapping populations with SSCP markers showed up to 33% polymorphism in the markers suggesting that the SNPs can be used as potential resource for *S. bicolor* crop improvement programs.

Key words: cheap genotyping assay, Cleaved Amplified Polymorphism Sequence (CAPS), SNP, SSCP

Source: A total of 185,593 EST sequences were retrieved from FTP site of *Sorghum bicolor* NCBI UniGene database (ftp://ftp.ncbi.nih.gov/repository/UniGene/Sorghum_bicolor/)

Introduction

Sorghum is one of the world's most important crop plants ranking fifth in acreage among cereals (http://www.icrisat.org) after wheat, maize, rice, and barley (Paterson 2008; Paterson et al. 2009). Many types of biological markers exist but single nucleotide polymorphisms (SNPs) are preferred markers to saturate genetic maps as they are co-dominant, bi-allelic, highly polymorphic, and the most abundant source of variations among closely related genomes on top of being compatible with multiple, high-throughput technology platforms (Tsang et al. 2005). Thus, saturating the genetic map of sorghum with genome-wide SNP markers would augment the success of breeding of most agronomically important traits.

Fakrudin B. (✉)
E-mail: bfakrudin@gmail.com

Initially, only the conventional wet lab (rDNA) techniques were used to detect SNPs. But gradually with the development of bioinformatics, *in-silico* SNP detection using the huge number of EST resources available in public databases, remains to be one of the splendid and economical routes to detect novel SNPs (Edwards 2007). EST sequences offer an enormous resource for the identification of biologically meaningful SNPs due to their relatively higher redundancy. There is the possibility of the submission of sequences of the same locus from different individuals/cultivars. However, errors incorporated by inaccurate reverse transcriptase, DNA polymerase, and low quality sequencing are the major constraints of the use of EST sequences for SNP identification (Gu et al. 1998; Picoult-Newberg et al. 1999). As a result, the success of EST-based SNP identification depends on the methods adopted in distinguishing the likely false positives from the genuine ones. Nevertheless, candidate SNPs detected from EST sequence data after appropriate filtering schemes have been validated by re-sequencing to comprise largely genuine polymorphisms in a number of organisms (Coles et al. 2005; Duran et al. 2009; Lopez-Crapez et al. 2005). It has been reported that *in silico* pre-selection of potentially polymorphic loci had resulted in enhanced SNP identification efficiency and generation of robust markers with high nucleotide diversity (Kota et al. 2003). In this paper, we used ESTs from the NCBI Unigene database to identify SNPs in the transcribed regions of sorghum. Moreover, as most of the assays for SNP genotyping require expensive and specialized equipment and chemicals, we tried to derive SSCP markers from the SNPs identified in order to demonstrate the usefulness of the resources for cost-effective marker development.

Methods and Materials

Mining of SNP-CAPS markers

Sorghum bicolor EST sequences were retrieved from FTP site of the NCBI UniGene database (ftp://ftp.ncbi.nih.gov/repository/UniGene/Sorghum_bicolor/). Vector sequences used for the *Sorghum bicolor* cDNA library were used to trim vector-contaminated EST sequences. Redundancy in retrieved ESTs was established through clustering using codoncode aligner (http://www.codoncode.com/aligner/). All ESTs of a NCBI '*Unigene*' were included in one batch and to minimize the inclusion of ESTs from paralogous genes, stringent clustering criteria were followed: at least 50 bases overlap and 95% identity between one end of a sequence to the other end. The presence of variations was identified in assembled contigs. A quality screen was performed in order to differentiate genuine SNPs from false ones using the method adopted from Picoult-Newberg et al. (1999) and Hawken et al. (2004).

Polymorphic SNPs from the multiple alignment of contigs using Codoncode aligner™, were converted to CAPS markers with lists of restriction enzymes retrieved from Rebase database (version903) (*rebase.neb.com/rebase/rebase.html*) using the SNP2CAPS (Thiel et al. 2004) algorithm. The functional relationship of each EST-containing putative SNPs was performed by using the BLASTx program against the non-redundant (nr) database. To assign tentative identity, an E-value cut-off of E^{-10} was used as criteria. The genes annotated were classified into protein classes by considering the biological/molecular function assigned to the protein in Uniprot. Genes were further grouped into broader sets using scheme (Vogel and Chothia 2006) for easy identification and classification.

Validation of SNPs

For experimental verification of the SNPs, three accessions *viz*., E36-1, IS9830, and N13, which are parents of two recombinant inbred populations (RIPs) [IS9830 × E36-1(RIP1) and N13 × E36-1 (RIP2)] were used. From these accessions, genomic DNA was extracted from leaf material as described by Krishna and Jawali (1997). Primer pairs were designed on the basis of contig consensus sequences obtained by codoncode aligner using the computer program VectorNTI (Invitrogen, USA). The criteria for selection of the primers were: maximum size of amplicon - 400 bp, low chance of primer-dimer and hairpin loops. PCR was done in 20 µL reactions as described by Kota et al. (2003).

Parental screening for polymorphism of SNP markers was implemented on 15% non-denaturating gel (50% acrylamide-bisacrylamide, 10X TBE, and glycerol) in a manual DNA sequencing apparatus (Biorad, USA) with minor modifications (Orita et al. 1989). Pre-run was done at 300 volts at constant temperature for at least 45 min. Denaturation of samples was carried out at 95°C for 10 min. For 4 µL of PCR product, 12 µL of formamide dye was added and immediately chilled on ice and kept at -20°C for 10 min prior to loading. Electrophoresis was performed in 0.5X Tris-borate (PH 8.3) EDTA buffer at 300 V for 17 ± 1 h at room temperature. To visualize the SSCP band patterns the gel was subjected to silver staining.

Results and Discussion

The NCBI EST database for *Sorghum bicolor* consists of 209,814 EST sequences as of June 15, 2009. The sequences are product of 48 cDNA libraries. Library size ranges from 1 sequence to 11,221. At least 13 cultivars were used as source for library construction (Table 1). In NCBI, automated procedure of Unigene database groups ESTs of the same origin into clusters (Unigenes). It also removes EST sequences of low quality, trims contamination by vector and host genome sequences and masks repetitive and low quality sequences. We used processed ESTs (~185,000) archived from NCBI Unigene database. A total of 10,251 sequences were trimmed due to contamination by vector sequences. In order to minimize clustering of paralogous sequences, 173,755 EST sequences were re-clustered and aligned with higher strin-

gency into 19,243 contigs and 11,838 singletons. Most of the contigs (24.44%) were formed from 2 fragments. The number of sequence reads per contig followed exponential frequency distribution (R^2 = 0.686). The same results were also reported by Picoult-Newberg et al. (1999). Of these contigs, 10,865 (55%) comprised > 3 ESTs. It is indeed from these contigs that candidate SNPs were mined.

Table 1. Distribution of publicly available ESTs among sorghum genotypes used to generate transctiptomes (as of June 15, 2009).

Cultivar	EST frequency	% of ESTs
Tamara	1	0.000
ICSV-272	3	0.001
35-1	6	0.003
GOLDEN HARVEST	19	0.009
Abu 70	38	0.018
P898012	47	0.022
ATx399 X RTx430	82	0.039
TX430	88	0.042
Tx7000	3597	1.710
B35	6295	3.000
RTX430	10449	4.980
IS3620C	22202	10.580
Unknown	60405	28.790
BTx623	106582	50.800
Total	209814	100.000

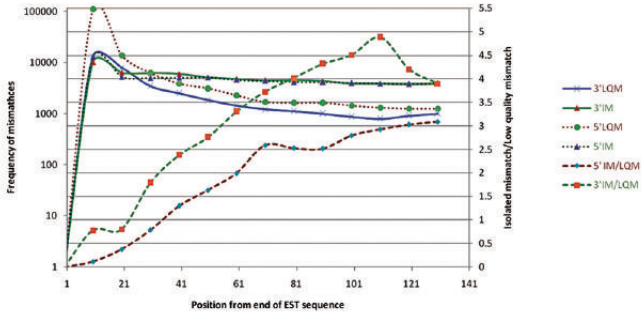

Fig. 1. Frequency of low quality mismatches and Isolated Mismatch (IM), non Low Quality Mismatches (LQM), distributed by positions from both ends of EST sequences. Regions of EST sequences from 5' and 3' ends to 60 bp and 40 bp, respectively, had IM/LQM value less than the threshold 2.5.

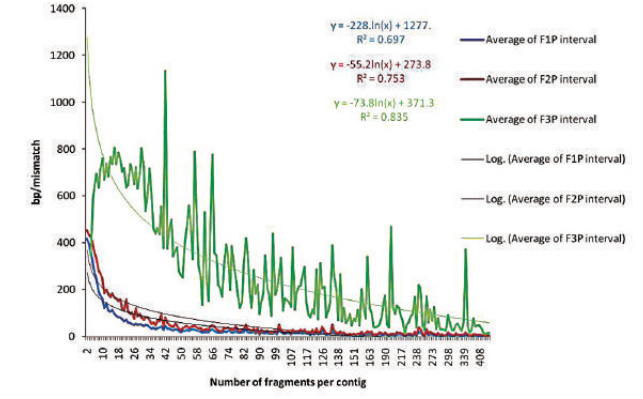

Fig. 2. Average mismatch interval against contig size across the filter scheme

CodonCode AlignerTM detected 433,041 raw mismatches in contigs assembled from four or more fragments. This includes 51,485 indels, 102,720 mismatches with ambiguous base, and 278,336 substitution-type mismatches with unambiguous base. From the analysis of 117,631 EST sequences to delineate the region with large sequencing errors, the first 60 bp and 40 bp from 5' and 3' end, respectively, were found to be frequently prone for sequencing errors (Fig. 1).

The entire *mismatch quality control* scheme had three filters. Filter1 (F1) removed mismatches, which originated from region where three or more mismatches are clustered in a sliding window of 20 bp. Such mismatches were termed as Low Quality Mismatches (LQM). This filter would likely eliminate sequencing artifacts, which tend to cluster and often occur in regions of low sequence quality. The filter removed 121,675 LQM from the pool of 278,336 raw mismatches. The remaining 156,661 were labeled as 'Filter1 passed' (F1P) mismatches. The second filter (F2) was employed to remove mismatches arising from terminal ends. F2 removed 49,112 F1P mismatches, which resided in the first 60 bases from the 5' end and 40 bases from the 3' end of each EST sequence. The remaining 107,549 mismatches were termed as F2P. The last filter, Filter 3 (F3) identified SNPs where the minor allele frequency is represented more than once in multiple sequence alignment. The resulting 12,421 sequence variants were termed as F3P or candidate SNPs. These mutations most likely represent genuine genomic polymorphisms.

On average for every 33.85, 50.38, and 275.26 bp, one mismatch was found for filters one, two, and three, respectively. Mismatch interval was found to be dependent on the number of fragments for a given assembly and the level of the filter (Fig. 2). Fragment count and mismatch interval had a strong and significant ($P < 0.05$) negative logarithmic relationship for all filters.

Raw substitution mismatches had lower transition proportion (0.420) compared to transversion. Transition proportion increased to 0.50, 0.547, and 0.598 as the level of filter progressed from F1, F2, and finally to F3, respectively. C/T type mutation had the highest percentage in all the filters followed by another transition type mutation, A/G. In contrast, A/T type mutation had the lowest frequency (Table 2).

Table 2. Frequency of substitution-type mutations at various stages of the filtering scheme.

Kind of mutation	Type of alleles	RM	Prop*. RM**	F1P	Prop. F1P	F2P	Prop. F2P	F3P	Prop. F3P
Transition	A<>G	58197	0.209	35238	0.225	26547	0.247	3009	0.242
	C<>T	58699	0.211	43045	0.275	32285	0.300	4419	0.356
Transition		**116896**	**0.420**	**78283**	**0.500**	**58832**	**0.547**	**7428**	**0.598**
	A<>C	42687	0.153	21698	0.139	12951	0.120	1526	0.123
Transversion	A<>T	32383	0.116	13621	0.087	7769	0.072	515	0.041
	C<>G	50757	0.182	24404	0.156	16718	0.155	1934	0.156
	G<>T	35613	0.128	18655	0.119	11279	0.105	1018	0.082
Transversion		**161440**	**0.580**	**78378**	**0.500**	**48717**	**0.453**	**4993**	**0.402**
Grand Total		278336		156661		107549		12421	

* Proportion** Raw Mismatches

One feature of genuine polymorphism is expected to be the over-representation of transition substitutions (A to G and C to T) due to the deamination of 5-methyl cytosine (Holliday and Grigg 1993). This expectation was used to evaluate our filtering scheme. As expected, the proportion (0.598) of SNPs produced after implementation of all filters was larger than the transversion-type of mutations (0.402) and far greater than the proportion when the expectation is random mutation (0.333). However, prior to implementation of the filters, the proportion of transversion-type mutations was greater than transition-type mutation. This shows that most of the SNPs identified through the filtering scheme represent biologically meaningful and likely genuine polymorphisms.

After implementation of F3, the mean SNP frequency was 1 in 275.26. bp. However, the frequency is a little underestimated as the filters may have removed genuine SNPs. SNP interval has changed from filters to filters and number of ESTs per contig. In fact, very strong negative exponential relationship between number of contigs and mean SNP interval (number of bases per SNP) was found (R^2 = 0.835). Identified SNP frequency in sorghum was lower compared to allogamous species like maize (1/61) and higher than other autogamous species like wheat (1/540) (Ching et al. 2002; Somers et al. 2003). Further, validation studies in humans indicated that approximately 50 to 82% of predicted SNPs are genuine (Batley et al. 2003; Buetow et al. 1999; Gu et al. 1998; Hawken et al. 2004; Hu et al. 2002; Irizarry et al. 2000; Picoult-Newberg et al. 1999; Stone et al. 2002).

A total of 445 assemblies had at least one SNP with minor allele frequency ≥ 4. After BLASTx was performed, 17 (4%) assemblies did not have a significant hit (E value = E^{-10}) with proteins available in public protein databases. Of the predicted proteins, 37% were involved in metabolism. The remaining 14, 11, 6, 5, and 4% were functionally grouped into regulation, cellular functions, information, general function, and stress/resistance, respectively. The remaining 19% proteins were with unknown functions (Supplementary Table 1).

A subset of 877 SNPs were randomly selected from a pool of putative SNPs and were tried *in silico* using SNP2CAPS (Thiel et al. 2004) algorithm to convert into CAPS markers. The SNPs were carried by 315 contigs spanning a total of 404,154 bp. A total of 1,599 potential CAPS (SNPs residing within overlapping recognition site of restriction enzyme were counted as independent potential CAPS) were identified distributed over 224 (71.11%) contigs. On average, one potential CAPS marker for every 252.75 bp was detected. The average frequency of CAPS was 5.08 per contig (Table 3). Thiel et al. (2004) reported that 90% of the alignments which had harbored SNPs and *indels* had been converted to CAPS. The conversion rate in terms of contig was lower possibly because *indels* were not considered in this study. From 28 SSCP markers assessed, ten and six markers were found to be contrasting across parents of RIPs *viz*., IS9820 and E36-1 (RIP1), and N13 and E36-2 (RIP2), respectively (Fig. 3 and Table 4).

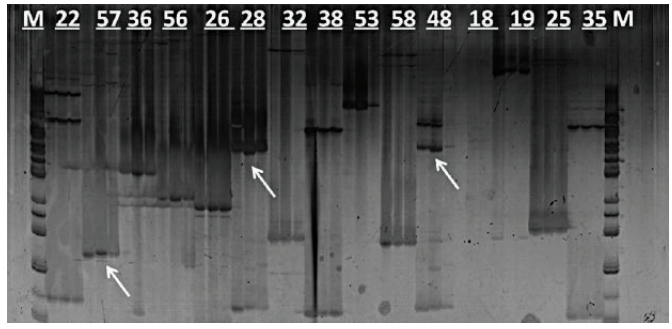

Fig. 3. SSCP markers screened for polymorphism using parents of mapping popula-tions. The numbers indicate SB-SNP series primers designed for the markers. The three wells under each marker were loaded amplicons from DNA templates of E36-1, IS9830 and N13, respectively. Arrows indicate polymorphic markers.

Table 3. Frequency of potential CAPS in relation to contigs considered for SNP to CAPS conversion.

Description	Frequency
Number of contigs considered	315
Total number of bases	404,154
Total number of contigs with CAPS markers	223
Total potential CAPS identified	1,599
Number of contigs with CAPS	224
bp / Potential CAPS	252.75
Average number of CAPS per contig	5.08

Table 4. Screening of SSCP markers for polymorphism using two mapping populations. (M, P, and NA signify Monomorphic, Polymorphic, and No amplification, respectively).

PrimerID	IS9830 X E36-1 (RIP1)	N 13 X E36-1 (RIP2)
SB-SNP-22	M	M
SB-SNP-57	P	P
SB-SNP-36	P	P
SB-SNP-56	P	M
SB-SNP-26	M	M
SB-SNP-28	P	M
SB-SNP-32	M	M
SB-SNP-38	M	M
SB-SNP-53	P	M
SB-SNP-58	M	M
SB-SNP-48	P	M
SB-SNP-18	NA	NA
SB-SNP-19	P	M
SB-SNP-25	M	M
SB-SNP-35	NA	NA
SB-SNP-20	M	M
SB-SNP-34	M	M
SB-SNP-45	M	M
SB-SNP-47	NA	NA
SB-SNP-33	P	P
SB-SNP-40	M	M
SB-SNP-54	P	P
SB-SNP-27	NA	NA
SB-SNP-41	P	P
SB-SNP-43	M	M
SB-SNP-46	M	M
SB-SNP-37	M	M
SB-SNP-31	NA	P
Polymorphic	10	6
Per. Poly	0.344828	0.206897

M: Monomorphic; P: Polymorphic; NA: No Amplification

The real concern for *in-silico* predicted SNPs is whether the SNPs identified are also common in a wide range of cultivars cultivated and are being used in breeding programs. Since EST submission is biased towards a small number of genotypes, it is difficult to predict the allelic frequency of a given SNP in a natural population. To check whether the predicted SNPs also are common in accessions being used for breeding purposes, we screened 28 SSCP markers across three sorghum parental lines, which are used as parents of mapping populations. They showed a verification rate of approximately 33%. The rate of verification is lower compared to the reported rate (Batley et al. 2003; Hawken et al. 2004; Picoult-Newberg et al. 1999) probably due to the smaller number of gene fragments and cultivars used for validation. However, the level of polymorphism of our SNPs in the panel of parents used for mapping populations was greater than the verification rate of gene based SSR markers (25%) reported by Eujayl et al. (2001).

Conclusion

As a multitude of sequence data are available in public databases, discovering polymorphisms *in-silico* becomes a valid option. ESTs are redundant and there is possibility of submission of sequences of the same locus from different individuals; therefore, they provide a resource for *in-silico* SNP discovery. Moreover, since ESTs largely represent coding segments of the genome, SNPs identified from them are more likely to affect phenotype than those in non-coding regions of a genome. Sorghum EST sequences are product of 48 cDNA libraries from at least 13 cultivars. Inclusion of these many cultivars offered a great opportunity to mine SNPs. The SNP markers in the present study have proven their worthiness in the detection of polymorphism among the sorghum cultivars and also provided an excellent opportunity to develop a cost-effective marker system like SSCP.

Acknowledgements

This research was supported by the Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India (DBT Programme Support Project). We thank the Project Monitoring and Mentoring Committee (PMMC) members, Dr. V. P. Gupta, Dr. N. Seetharama, Dr. P. Balasubramaniam, Dr. M. B. Chetti, Dr. H.E. Shashidhar, and Dr. Shailaja Hittalmani for their helpful suggestions.

References

Batley J, Baker G, O'Sullivan,Edwards KJ, Edwards D. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol.132: 84-91

Buetow KH, Edmonson MN, Cassidy AB. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet. 21: 323-325

Ching A, Katherine SC, Mark J, Maurine D, Oscar S, Scott T, Michele M, Anthony JR. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3: 19-32

Coles ND, Coleman CE, Christensen SA, Jellen EN, Stevens MR, Bonifacio A, Rojas-Beltran JA, Fairbanks DJ, Maughan PJ. 2005. Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Sci. 168: 439-447

Duran C, Appleby N, Vardy M, Imelfort M, Edwards D, Batley J. 2009. Single nucleotide polymorphism discovery in barley using autoSNPdb. Plant Biotechnol. J. 7: 326-333

Edwards NJ. 2007. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol. Syst. Biol. 3: 102

Eujayl I, Sorrells M, Baum M, Wolters P, Powell W. 2001. Assessment of genotypic variation among cultivated durum wheat based on EST-SSRS and genomic SSRs. Euphytica. 119: 39-43

Gu Z, Hillier L. Kwok PY. 1998. Single nucleotide polymorphism hunting in cyberspace. Hum. Mutat. 12: 221-225

Hawken RJ, Wesley CB, Sean M, Brian PD. 2004. An interactive bovine *in silico* SNP database (IBISS). Mammalian genome. 15: 819-827

Holliday R, Grigg, GW. 1993. DNA methylation and mutation. Mutat. Res. 285: 61-67

Hu G, Modreck B, Stensland HM, Saarela J, Pajukanta P, Kustanovich V, Peltonen L, Nelson SF, Lee C. 2002. Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes. Pharmacogenomics J. 2: 236-242

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, Lee CJ. 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat. Genet.26: 233-236

Kota R,Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A. 2003. Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). Mol. Genet. Genomics. 270: 24-33

Krishna TG, Jawali N. 1997. DNA isolation from single or half seeds suitable for random amplified polymorphic DNA analyses. Anal. Biochem. 250: 125-127

Lopez-Crapez E, Bazin H, Chevalier J, Trinquet E, Grenier J, Mathis G. 2005. A separation-free assay for the detection of mutations: combination of homogeneous time-resolved fluorescence and minisequencing. Hum. Mutat. 25: 468-475

Orita M, Suzuki Y, Sekiya T, Hayashi K. 1989. Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. Genomics. 5:

874-879
Paterson AH. 2008. Genomics of Sorghum. Intl. J. Plant Genomics. Article ID 362451
Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 457: 551-556
Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M. 1999. Mining SNPs from EST databases. Genome Res. 9: 167-174
Somers DJ, Kirkpatrick R, Moniwa M, Walsh A. 2003. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome. 46: 431-437
Stone RT, Grosse WM, Casas E, Smith TP, Keele JW, Bennett GL. 2002. Use of bovine EST data and human genomic sequences to map 100 gene-specific bovine markers. Mammalian Genome. 13: 211-215
Thiel T, Kota R, Grosse I, Stein N, Graner A. 2004. SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. Nucl. Acids Res. 32: e5
Tsang S, Sun Z, Luke B, Stewart C, Lum N, et al. 2005. A comprehensive SNP-based genetic analysis of inbred mouse strains. Mammalian Genome. 16: 476-480
Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. PLoS Computational Biol. 2: e48

Supplementary Table 1. Functional distribution of polymorphic assemblies into different classes and categories.

Category	Class	Frequency
Cellular functions		50
	Cell motility	9
	Membrane protein	2
	Nuclear structure	1
	Reactive Oxygen related	1
	Storage	1
	Structure	2
	Transport	34
General		21
	Developmental protein	1
	Ion binding	12
	Other enzymes	4
	Sexual reproduction	2
	Small molecule binding	2
Metabolism		168
	Amino acids metabolism	5
	Carbohydrate metabolism	7
	Cell wall envelop metabolism	5
	Coenzyme metabolism	3
	Lipid metabolism	3
	Nitrogen/carbon metabolism	3
	Photosynthesis	9
	Protease	28
	Redox	49
	Secondary metabolism	4l
	Transferases	19
	General metabolism	33
Regulation		64
	Chromatin structure	2
	DNA-binding	6
	Kinase/phosphatase	4
	Nucleotide binding	6
	Other regulatory function	5
	Protein interaction	9
	Protein modification	17
	RNA binding	6
	Signal transduction	3
	Transcription factor	6
Information		27
	mRNA processing	1
	Ribonucleoproteins	2
	Transcription	2
	Translation	22
Stress/Resistance		22
	Disease resistance	3
	Stress	19
Unknown		93
Grand Total		445