Types and Frequencies of Sequencing Errors in Methyl-Filtered and High C_0t Maize Genome Survey Sequences[1][w]

Yan Fu, An-Ping Hsia, Ling Guo, and Patrick S. Schnable*

Department of Genetics, Development, and Cell Biology (Y.F., L.G., P.S.S.), Department of Agronomy (A.-P.H., P.S.S), Interdepartmental Graduate Programs in Genetics (Y.F., P.S.S.) and Bioinformatics and Computational Biology (L.G., P.S.S.), and Center for Plant Genomics (P.S.S.), Iowa State University, Ames, Iowa 50011–3650

The Maize Genome Sequencing Consortium has deposited into GenBank more than 850,000 maize (*Zea mays*) genome survey sequences (GSSs) generated via two gene enrichment strategies, methylation filtration and high-C_0t (HC) fractionation. These GSSs are a valuable resource for generating genome assemblies and the discovery of single nucleotide polymorphisms and nearly identical paralogs. Based on the rate of mismatches between 183 GSSs (105 methylation filtration + 78 HC) and 10 control genes, the rate of sequencing errors in these GSSs is 2.3×10^{-3}. As expected many of these errors were derived from insufficient vector trimming and base-calling errors. Surprisingly, however, some errors were due to cloning artifacts. These G•C to A•T transitions are restricted to HC clones; over 40% of HC clones contain at least one such artifact. Because it is not possible to distinguish the cloning artifacts from biologically relevant polymorphisms, HC sequences should be used with caution for the discovery of single nucleotide polymorphisms or paramorphisms. The average rate of sequencing errors was reduced 6-fold (to 3.6×10^{-4}) by applying more stringent trimming parameters. This trimming resulted in the loss of only 11% of the bases (15,469/144,968). Due to redundancy among GSSs this more stringent trimming reduced coverage of promoters, exons, and introns by only 0%, 1%, and 4%, respectively. Hence, at the cost of a very modest loss of gene coverage, the quality of these maize GSSs can approach Bermuda standards, even prior to assembly.

The National Science Foundation is funding a project to compare two gene enrichment strategies for sequencing the gene space of the maize (*Zea mays*) genome. The first strategy uses methyl filtration (MF) to enrich for the gene-rich fraction of the genome (Rabinowicz et al., 1999; Palmer et al., 2003). The second strategy enriches for genes by sequencing only the high C_0t (HC) fraction of the genome that is enriched for low-copy sequences (Yuan et al., 2003). In combination these two strategies increase by 4-fold the rate of gene discovery as compared to shotgun sequencing (Palmer et al., 2003; Whitelaw et al., 2003).

As of September 2003, 450,166 MF and 445,541 HC genome survey sequences (GSSs) from the maize inbred line B73 had been deposited in GenBank (http://www.tigr.org/tdb/tgi/maize/progress_graph. shtml). Ultimately, it will be important to assemble these GSSs into contigs. We have recently reported the development of innovative parallel algorithms that can assemble more than 730,000 GSS fragments in 4 h using 64 Pentium III 1.26 GHZ processors of a commodity cluster (Emrich et al., 2004). The inevitable sequencing errors in GSS data complicate the assembly of contigs. To optimize assembly parameters, it would be desirable to have an estimate of the rate of sequencing errors in MF and HC GSSs.

Sequencing errors in the GSSs also affect their utility for the detection of single nucleotide polymorphisms (SNPs) and nearly identical paralogs (NIPs). Comparisons between MF and HC GSSs generated from the inbred line B73 and sequences from other maize lines can reveal the presence of SNPs that have value both as molecular markers and as characters for phylogenetic analyses. Sequencing errors will, however, complicate SNP discovery. Recent segmental duplications of the human genome have generated large numbers of NIPs that have complicated the assembly, annotation, and analyses of that genome (Bailey et al., 2001, 2002). To avoid these problems during the assembly of the maize genome it would be desirable to be able to identify NIPs. NIPs are detected via the discovery of differences in the nucleotide sequences of highly similar paralogs (Emrich et al., 2004). Although previously termed cismorphisms (Hurles, 2002), for clarity we have elected to term these differences in the sequences of NIPs paramorphisms. Because sequencing errors can mimic paramorphisms, particularly in regions of the genome where sequencing coverage is low, it is critical to know the average rate of sequencing errors in GSSs.

[1] This work was supported by competitive grants from the National Science Foundation Plant Genome Program (award nos. DBI–9975868, DBI–0121417, and DBI–0321711). Support was also provided by the Hatch Act and State of Iowa funds.

* Corresponding author; e-mail schnable@iastate.edu; fax 515–294–5256.

[w] The online version of this article contains Web-only data. www.plantphysiol.org/cgi/doi/10.1104/pp.104.041640.

We previously estimated the rates of sequencing errors in the MF and HC GSSs as being 3.5×10^{-3} and 2.5×10^{-3}, respectively (Emrich et al., 2004). These estimates were based on comparisons of pairs of overlapping sequence reads from the same clone (i.e. clone pairs). In this report, we use an independent approach to estimate the rate of sequencing errors in GSSs. Specifically, we compared the sequences of MF and HC GSSs to the sequences of 10 control genes. This second approach identified two additional classes of errors that were not detected via the analysis of clone pairs. One of these newly detected types of errors are cloning artifacts that are present in over 40% of the HC clones. Finally, we have identified trimming parameters that reduce by 6-fold the sequencing error rate in the MF and HC GSSs. This reduction in the rate of sequencing errors (to 3.6×10^{-4}) will simplify the assembly of the maize genome and the discovery of SNPs and NIPs.

RESULTS AND DISCUSSION

Identification of Errors from Maize MF and HC GSSs

The rate of sequencing errors in MF and HC GSSs was estimated by comparing GSSs to a set of 10 genes we cloned and sequenced (Table I; H. Yao, L. Guo, Y. Fu, T.-J. Wen, L. Borsuk, D. Skibbe, X.Q. Cui, B.E. Scheffler, J. Cao, F. Liu, D.A. Ashlock, and P.S. Schnable, unpublished data). Because both strands of each of these genes were sequenced and the resulting trace files were manually checked for base-calling errors, the error rate in this data set was expected to be low. This feature makes this data set of approximately 74 kb of finished sequence a suitable control for estimating the error rate in MF and HC data. MF and HC GSSs derived from these 10 control genes were identified using BLAST (see "Materials and Methods"). This process identified 105 MF and 78 HC GSSs (total length approximately 145 kb, Table I) for further analyses. Figure 1 depicts the GSSs derived from the *rf2a* gene (for similar data on the other control genes, see Supplemental Fig. 1, A–I, which may be viewed at www.plantphysiol.org).

The fact that at least four MF and three HC reads matched each of the control genes provides encouraging evidence as to the success of the two gene enrichment approaches. Among the control genes, the ratios of recovered MF:HC GSSs range from 4:24 (*rth1*) to 14:3 (*rf2e1*; Table I), presumably reflecting gene-specific differences in the genomic sampling achieved by the two gene enrichment strategies. For four genes, the GSSs covered all of the exonic regions; the coverage of exonic regions among the remaining six genes ranged from 61% to 97% (Table II). For three genes, all of the intronic regions were covered by GSSs; the coverage of intronic regions among the six remaining genes that contain introns ranged from 54% to 90%. Promoter regions (defined for the purposes of this study as being the 500 bp upstream of the 5′ end of the apparently full-length cDNA sequence) were less well represented among the GSSs. Of the seven promoters that could be analyzed, three were fully represented by GSSs, two were partially represented (45% to 57%), and two were not represented by any GSSs. As compared with HC GSSs, MF GSSs had a higher coverage of both promoters (53% versus 19%) and exons (71% versus 56%), but lower coverage of introns (37% versus 51%). Hence, these analyses demonstrate that the two gene enrichment strategies are complementary at identifying not only transcribed regions as previously demonstrated by Whitelaw et al. (2003), but also promoters and introns.

Each mismatch in an alignment between a GSS and a control gene triggered a second round of manual checking of the trace files associated with both the GSS and the control gene. These analyses identified a few

Table I. *Alignments of MF and HC GSSs with 10 control genes, prior to and after trimming GSSs*

Control Genes				GSSs[a]					
				No. of GSSs			Length of GSSs		
Name	Accession Number	Length	GC	MF	HC	Total	MF	HC	Total
		bp	%					*bp*	
gl8a	AF302098	6,817	51.1	9	3	12	7,371/6,849	2,367/2,119	9,738/8,968
rf2a[b]	AF215823	12,673[b]	43.0	12	8	20	10,306/8,945	7,162/6,100	17,468/15,045
rf2c	AF348412	7,257	48.1	6	8	14	5,203/4,731	4,719/4,190	9,922/8,921
rf2d[c]	AF348414	7,415	53.0	12	10	22	9,950/8,969	7,742/6,777	17,692/15,746
rf2b	AF348418	4,311	51.0	15	4	19	11,821/10,434	3,335/2,953	15,156/13,387
pdc2	AF370004	5,443	52.7	12	4	16	10,042/9,344	3,866/3,472	13,908/12,816
pdc3	AF370006	7,773	46.1	12/11	11	23/22	8,649/7,854	8,718/7,837	17,367/15,691
rth1	AY265854	14,348	40.7	4	24/23	28/27	3,338/2,961	18,446/15,852	21,784/18,813
rth3	AY265855	3,158	58.0	9	3	12	7,194/6,981	1,731/1,620	8,925/8,601
rf2e1	AY374447	4,801	52.4	14/13	3	17/16	10,398/9,268	2,610/2,243	13,008/11,511
Total/Avg.		73,996	47.6	105/103	78/77	183/180	84,272/76,336	60,696/53,163	144,968/129,499

[a]Before/after trimming. [b]An approximately 4.8-kb region of the fifth intron of *rf2a* was masked prior to the BLAST search (Fig. 1). This region contains two open reading frames of a repetitive retrotransposon. [c]*rf2d* contains partial coding sequence, and only boundary-defined exons and introns were used in calculation.

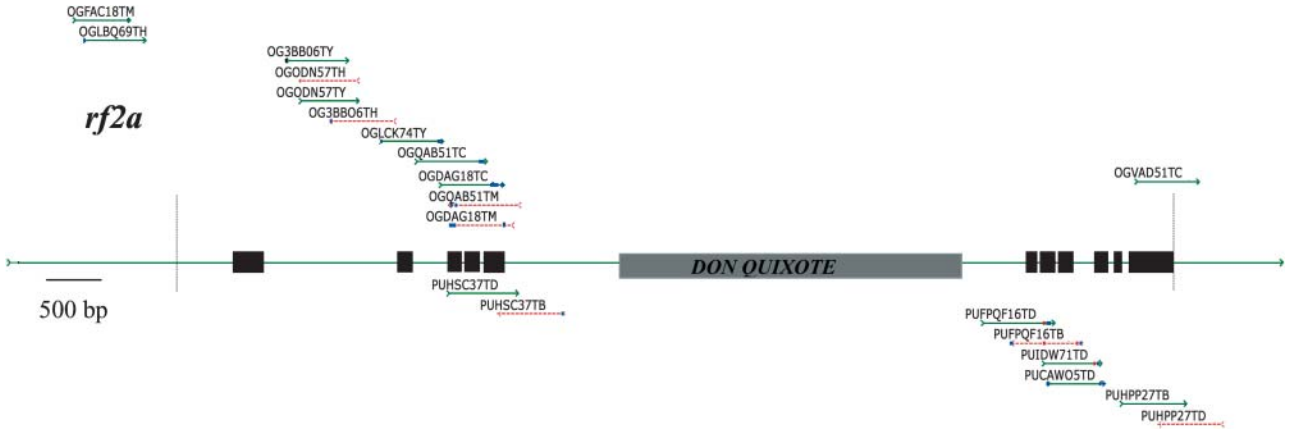

Figure 1. GSS coverage of the *rf2a* gene. Black boxes indicate experimentally validated exons. The two vertical dotted lines define the interval used for the coverage calculations presented in Table II. Solid green lines designate GSSs with 5′ -> 3′ orientations, while red dotted lines designate GSSs with 3′ -> 5′ orientations. The black, blue, and red dots on GSSs indicate Class I, Class II, and Class III errors, respectively. MF and HC GSSs are located above and below the gene, respectively. The gray box indicates the approximately 4.8-kb region (positions 8,430–13,230) masked prior to the BLAST search, which contains two open reading frames (positions 8,430–12,224 and 11,869–13,230) of a repetitive copia-like retrotranspon, *DON QUIXOTE*.

errors in the approximately 74 kb of control sequences. After correcting these errors, the remaining 339 mismatches were deemed to be errors in the MF and HC reads, resulting in an average error rate of 2.3×10^{-3} (339/144,968; Table III) in the GSSs. The distributions of errors in each gene are diagramed in Figure 1 and supplemental data (Fig. 1, A–I). The average error rates were lower in MF versus HC GSSs (2.1×10^{-3} versus 2.6×10^{-3}; Table III). This was also usually true at the level of individual genes; in all but two genes (*rf2b* and *rf2e1*) the rates of sequencing errors in the MF GSSs were lower than in the corresponding HC GSSs.

Three Classes of Errors

The sequencing errors detected in the GSSs can be grouped into three classes (Table III). The 21 Class I errors were located in the first eight bases of GSS reads and were associated with regions of the sequence traces that had high-quality Phred scores (>30, indicating the error probability was less than 10^{-3}). Because these errors exhibited 100% identity to the sequences of the vectors used in the production of the MF and HC libraries (data not shown), we conclude that they are the result of insufficient vector trimming of GSS reads prior to deposition in GenBank.

The 281 Class II errors were all associated with regions of the GSS reads that had low-quality scores (i.e. they are true base-calling errors). Ninety percent of these errors have quality scores <20 and all have quality scores <30. Class II errors consist of deletions (125), insertions (90), and base substitutions (66; data not shown). The average rate of Class II errors is 1.9×10^{-3} (281/144,968).

Table II. *Coverage of promoters, exons, and introns by MF and HC GSSs, prior to and after trimming GSSs*

Gene	Coverage								
	Promoters[a]			Exons			Introns		
	MF	HC	Total	MF	HC	Total	MF	HC	Total
					%[b]				
gl8a	45/45	0	45/45	59/59	41/40	100/99	41/41	45/37	87/79
rf2a[c]	0	0	0	27/27	75/69	79/79	40/39	40/38	77/73
rf2c	25/24	32/32	57/56	50/49	46/39	89/86	77/77	8/8	85/85
rf2d	NA[d]	NA[d]	NA[d]	100/100	100/100	100/100	100/100	100/100	100/100
rf2b	100/100	100/100	100/100	100/100	54/53	100/100	100/100	47/46	100/100
pdc2	NA[d]	NA[d]	NA[d]	100/100	66/66	100/100	99/97	95/95	100/100
pdc3	100/100	0	100/100	95/94	50/49	97/96	90/90	39/39	90/90
rth1	0	0	0	18/18	61/60	61/60	4/4	64/58	64 /58
rth3	NA[d]	NA[d]	NA[d]	97/97	36/36	97/97	NA[e]	NA[e]	NA[e]
rf2e1	100/100	0/0	100/100	74/74	29/18	74/74	43/43	36/34	54/53
Avg.	53/53	19/19	57/57	71/71	56/54	89/88	37/37	51/47	74/70

[a]Defined for the purposes of this study as being the 500 bp upstream of the 5′ end of the apparently full-length cDNA sequence. [b]Before/after trimming. [c]A portion of the 5th intron of *rf2a* was not included in this study (Fig. 1). [d]Not applicable due to absent or short (<500 bp) promoter. [e]Not applicable; *rth3* does not contain any introns.

Table III. *Estimation of the rate of sequencing errors in MF and HC GSSs, prior to and after trimming GSSs*

Gene	Errors in GSSs[a]											
	MF + HC		MF					HC				
	No. Errors	Error Rate[b]	No. Errors				Error Rate[b]	No. Errors				Error Rate[b]
			I	II	III	Total		I	II	III	Total	
		10^{-3}/bp					*10^{-3}/bp*					*10^{-3}/bp*
gl8a	18/2	1.9/0.22	5/0	7/0	0	12/0	1.6/0	0	5/2	1/0	6/2	2.5/0.94
rf2a	56/4	3.2/0.27	1/0	30/1	0	31/1	3.0/0.11	0	21/0	4/3	25/3	3.5/0.49
rf2c	15/5	1.5/0.56	0	7/0	0	7/0	1.3/0	0	3/0	5/5	8/5	1.7/1.2
rf2d	47/9	2.7/0.57	0	21/2	0	21/2	2.1/0.22	2/0	18/1	6/6	26/7	3.4/1.0
rf2b	26/2	1.7/0.15	0	26/2	0	26/2	2.2/0.19	0	0	0	0	0
pdc2	30/5	2.2/0.39	0	18/1	0	18/1	1.8/0.1	1/0	5/0	6/4	12/4	3.1/1.2
pdc3	40/5	2.3/0.32	0	19/0	0	19/0	2.2/0	1/0	15/1	5/4	21/5	2.4/0.64
rth1	53/10	2.4/0.53	0	4/0	0	4/0	1.2/0	8/1	35/4	6/5	49/10	2.7/0.63
rth3	24/3	2.7/0.35	1/0	16/1	0	17/1	2.4/0.14	2/0	2/0	3/2	7/2	4.0/1.2
rf2e1	30/2	2.3/0.17	0	24/2	0	24/2	2.3/0.21	0	5/0	1/0	6/0	2.3/0
Total	339/47	2.3/0.36	7/0	172/9	0	179/9	2.1/0.12	14/1	109/8	37/29	160/38	2.6/0.71

[a]Before/after trimming. [b]Error rate = number of errors/total length of GSSs (bp).

Approximately 18% of the Class II errors detected in this study were located in the first 50 bp at the 5′ ends of sequence reads; another 50% were located in the terminal 50 bp at the 3′ ends of these sequence reads. All of the base substitution errors at the 5′ ends of GSSs were due to aberrant T-traces (see Supplemental Fig. 2). The rates of sequencing errors in the first 50 bp (5.2×10^{-3}) and last 50 bp (1.6×10^{-2}) were approximately 3-fold and 8-fold higher than the overall rate of Class II errors, respectively. This suggests that the frequency of errors could be reduced substantially by more stringently trimming the ends of sequence reads.

To test this hypothesis, Lucy software (Chou and Holmes, 2001; http://www.tigr.org/software/) was used to trim vector and low-quality regions from GSS reads. By tuning Lucy's parameters (−Size 9, −Bracket 20 0.003, −Window 10 0.01, −Error 0.005 0.002; see Supplemental Table I for details), it was possible to eliminate almost all of the Class I errors (20/21) and 94% of the Class II errors (264/281). The average error rate after trimming was only 3.6×10^{-4}, a 6-fold reduction as compared with the error rate in the sequences as deposited in GenBank (i.e. 2.3×10^{-3}). This more stringent trimming resulted in the loss of only 11% of the bases (15,469/144,968) in the data set. More significantly, due to partial redundancy among GSSs, these new trimming parameters had even less impact on gene coverage. The coverage of promoters, exons, and introns remained almost the same (i.e. trimming reduced coverage by only 0%, 1%, and 4%, respectively; Table II). Hence, at the cost of a very modest loss of coverage, the quality of these GSSs can approach Bermuda standards, even though this standard is usually only applied to assembled genomic sequences.

The 37 remaining errors (Class III, Table III) were all G•C to A•T transitions associated with regions of the GSS trace files that had quality scores of greater than 30. Some HC clones contained multiple Class III errors, and Class III errors were detected in multiple HC libraries (see Supplemental Table II). The fact that the Class III errors were detected only within HC reads explains at least partially why the rate of sequencing errors was somewhat higher in HC than MF GSSs (2.6×10^{-3} versus 2.1×10^{-3}).

The 16 Class III errors located in overlapping regions of clone pairs were detected by both the forward and reverse sequencing reads. Manual inspection of the sequence traces and the results of the clone pair analysis demonstrate that the Class III errors are not base-calling errors. Instead, we assert that the affected HC clones contain SNPs relative to the control genes.

We considered the possibility that the affected HC clones are derived from NIPs of the control genes. This possibility is not, however, consistent with the fact that Class III errors are detected only among the HC clones. In addition, existing DNA gel-blot hybridization data do not support the existence of NIPs for any of the control genes (data not shown). We instead conclude that the Class III errors are cloning artifacts. These cloning artifacts are particularly problematic because they cannot be distinguished from biologically relevant SNPs or paramorphisms. Although they occur at a relatively low rate within HC sequences (6×10^{-4}), more than 40% (20/46) of HC clones contain at least one Type III error (Supplemental Table II).

The origin of these cloning artifacts is not known. The original protocol for generating HC libraries (Yuan et al., 2003) employed Klenow (exo-; New England Biolabs, Beverly, MA) to convert single-stranded DNA molecules to double-stranded molecules that could be cloned. Because Klenow (exo-) does not have 5′ -> 3′ or 3′ -> 5′ proofreading exonuclease activity, it misincorporates nucleotides at a high rate (3.8×10^{-3}; Brown et al., 2002). Consequently, for the production of the HC libraries sequenced by the Maize

Genome Sequencing Consortium, Klenow Fragment (Stratagene, La Jolla, CA) was substituted for Klenow (exo-; J. Bennetzen, personal communication). Although the Klenow Fragment has been reported to have a fidelity of approximately 10^{-7} (Kunkel and Bebenek, 2000), we hypothesize that its fidelity is dependent upon reaction conditions. Consistent with the hypothesis that the cloning artifacts in the HC libraries were introduced by Klenow is the observation that all of the Class III errors are G•C to A•T transitions and *Escherichia coli* DNA polymerase I is known to exhibit a strong bias toward G•C to A•T transitions substitution errors (Schaaper, 1993).

Comparisons of Methods to Estimate Rates of Sequencing Errors

By aligning GSSs with previously sequenced genes, it was possible to detect two classes of sequencing errors (Types I and III) that were not detected via the analysis of GSS clone pairs (Emrich et al., 2004). Even so, this analysis yielded somewhat lower estimates of the rates of sequencing errors than was obtained via the analysis of clone pairs. This is probably because Type II sequence errors occur at higher rates in the ends of sequence reads, which are more likely to be located in the overlapping regions of clone pairs, which are overrepresented in the clone pair analysis.

RECOMMENDATIONS

The quality of the maize MF and HC GSSs released to GenBank by the Maize Genome Sequencing Consortium is quite high. For certain applications (e.g. genome assembly and the detection of SNPs and NIPs) it would, however, be desirable to have available sequences with even lower rates of errors. The quality of the maize MF and HC GSSs can be improved greatly by more stringently trimming vector and low quality sequences from the 5′ and 3′ ends of the sequence reads (viz., using Lucy parameters of −Size 9, −Bracket 20 0.003, −Window 10 0.01, and −Error 0.005 0.002). Because more than 40% of HC clones contain at least one Type III error, HC sequences should be used with caution in analyses in which errors must be minimized. It would be desirable that future HC libraries from maize or other species be prepared only after identifying reaction conditions that have reduced rates of cloning artifacts.

MATERIALS AND METHODS

BLAST Search and Output Parsing

BLASTN searches (http://www.ncbi.nlm.nih.gov/blast/) without filtering low-complexity sequences were conducted using 10 control genes as query sequences and maize (*Zea mays*) GSSs as the object database (November 11, 2003). Default BLASTN settings for scoring alignments were used. Perl scripts were then used to parse the GenBank accession numbers of MF and HC GSSs from those alignments that met the following criteria: similarity ≥ 98%, alignment length ≥ 250 bp, alignment length ≥ [GSS length − 20 bases], and E-value ≤ 1e−100. BLAST was also used to align control genes to GSSs.

Retrieval of GSS Data

The trace files, quality score files, and clip site information of all qualified MF and HC GSSs were downloaded from the National Center for Biotechnology Information (NCBI) Trace database (http://www.ncbi.nlm.nih.gov/Traces). Only GSSs for which sequencing quality scores were available were analyzed further. Trace files were loaded into Sequencher software (version 4.1; Gene Codes, Ann Arbor, MI) for error verification.

GSS Trimming Using Lucy

Lucy (http://www.tigr.org/software/) was used to retrim GSSs. The sequences of the pBC SK- (Stratagene) and pCR4-TOPO (Invitrogen, Carlsbad, CA) vectors that had been used to construct the MF and HC libraries, respectively (http://www.tigr.org/tdb/tgi/maize/library_info.shtml), were used for vector trimming. Because Lucy indicates only the locations at which trimming should occur, an AWK script was written to trim GSSs and store trimmed data in FASTA format. The length of each trimmed GSS was calculated using a Perl script.

Calculation of GSS Coverage of Promoters, Exons, and Introns

GSS coverage was calculated as the percentage (%) of each region of a control gene that was covered by at least one GSS. For the purposes of this study, the promoter region was defined as the 500 bp 5′ of the 5′ end of the apparently full-length cDNA sequence.

Availability of Novel Materials

Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining any permissions will be the responsibility of the requestor.

Sequence data from this article have been deposited with the EMBL/GenBank data libraries under the following accession numbers AF302098, AF215823, AF348412, AF348414, AF348418, AF370004, AF370006, AY265854, AY265855, and AY374447.

ACKNOWLEDGMENTS

We thank Jun Cao, Xiangqin Cui, Chuck Dietrich, Feng Liu, Dave Skibbe, and Tsui-Jung Wen of the Schnable lab (Iowa State University) and Brian Scheffler (U.S. Department of Agriculture-Agricultural Research Service) for sharing sequencing data prior to publication. We thank Scott Emrich (Iowa State University) and Jeff Bennetzen (University of Georgia) for stimulating discussions.

Received February 25, 2004; returned for revision April 2, 2004; accepted May 31, 2004.

LITERATURE CITED

Bailey J, Gu Z, Clark R, Reinert K, Samonte R, Schwartz S, Adams M, Myers E, Li P, Eichler E (2002) Recent segmental duplications in the human genome. Science **297:** 1003–1007

Bailey J, Yavor A, Massa H, Trask B, Eichler E (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res **11:** 1005–1017

Brown KR, Weatherdon KL, Galligan CL, Skalski V (2002) A nuclear 3′-5′ exonuclease proofreads for the exonuclease-deficient DNA polymerase alpha. DNA Repair (Amst) **1:** 795–810

Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. Bioinformatics **17:** 1093–1104

Emrich SJ, Aluru S, Fu Y, Wen TJ, Narayanan M, Guo L, Ashlock D, Schnable PS (2004) A strategy for assembling the maize (*Zea mays L.*) genome. Bioinformatics **20:** 140–147

Hurles M (2002) Are 100,000 ''SNPs'' useless? Science **298:** 1509

Kunkel TA, Bebenek K (2000) DNA replication fidelity. Annu Rev Biochem **69:** 497–529

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science **302:** 2115–2117

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet **23:** 305–308

Schaaper RM (1993) Base selection, proofreading and mismatch repair during DNA replication in *Escherichia coli*. J Biol Chem **268:** 23762–23765

Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, et al (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science **302:** 2118–2120

Yuan Y, SanMiguel PJ, Bennetzen JL (2003) High-C_ot sequence analysis of the maize genome. Plant J **49:** 249–255