

1 Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive
2 Sequences During Maize Domestication and Improvement

3

4 Sanzhen Liu^{1*#}, Jun Zheng^{2*}, Pierre Migeon^{1*}, Jie Ren¹, Ying Hu¹, Cheng He², Hongjun
5 Liu^{3,4}, Junjie Fu², Frank F. White⁵, Christopher Toomajian¹, Guoying Wang^{2#}

6

7

8 ¹ Department of Plant Pathology, Kansas State University, Manhattan, KS. 66506. U.S.A.

9 ² Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081,
10 P.R.China

11 ³ State Key Laboratory of Crop Biology, Shandong Key Laboratory of Crop Biology,
12 Taian 271018, P.R. China.

13 ⁴ College of Life Sciences, Shandong Agricultural University, Taian 271018, P.R. China.

14 ⁵ Department of Plant Pathology, University of Florida, Gainesville, FL. 32611. U.S.A.

15

16

17

18

19 *Contributed equally to this work

20 [#]To whom correspondence may be addressed: liu3zhen@ksu.edu; wangguoying@caas.cn

21

22 Running Title: K-mer Analyses Reveal Changes on Repeats

23 Keywords: comparative genomics, repetitive, k-mer, maize

1 **Abstract**

2 The major component of complex genomes is repetitive elements, which remain
3 recalcitrant to characterization. Using maize as a model system, we analyzed whole
4 genome shotgun (WGS) sequences for the two maize inbred lines B73 and Mo17 using k-
5 mer analysis to quantify the differences between the two genomes. Significant differences
6 were identified in highly repetitive sequences, including centromere repeats, 45S
7 ribosomal DNA (rDNA), knob, and telomere repeats. Previously unknown genotype
8 specific 45S rDNA sequences were discovered. The B73-specific 45S rDNA is not only
9 located on the nucleolus organizer region (NOR) on chromosome 6 but also dispersed on
10 all the chromosomes in B73, indicating the relatively recent spread of 45S rDNA from
11 the NOR. The B73 and Mo17 polymorphic k-mers were used to examine allele-specific
12 expression of 45S rDNA. Although Mo17 contains higher copy number than B73,
13 equivalent levels of overall 45S rDNA expression indicates that dosage compensation
14 operates for the 45S rDNA in the hybrids. Using WGS sequences of B73xMo17 double
15 haploids (DHs), genomic locations showing differential repetitive contents were
16 genetically mapped. Analysis of WGS sequences of HapMap2 lines, including maize
17 wild progenitor teosintes, landraces, and improved lines, decreases and increases in
18 abundance of additional sets of k-mers associated with centromere repeats, 45S rDNA,
19 knob, and retrotransposon sequences were found between teosinte and maize lines,
20 revealing global evolutionary trends of genomic repeats during maize domestication and
21 improvement.

22

23

1

2 **Introduction**

3 The maize genome (*Zea mays* ssp. *mays*) exhibits high levels of genetic diversity among
4 different lines [1-3]. The inbred lines B73 and Mo17 represent two of the most
5 appreciated models for understanding maize genome diversity with respect to small-scale
6 polymorphisms [4-6] and large-scale structural variation [7, 8]. In addition, mapping
7 populations of inter-mated B73xMo17 recombinant inbred lines and double haploids
8 have been generated to facilitate genetic analyses [9, 10]. Numerous comparative
9 genomics studies of other maize cultivars and wild ancestors have examined the origin of
10 maize as well as events of adaptation and artificial selection [11-17]. However, the
11 studies are limited to comparisons of non-repetitive and low-repetitive sequences.

12

13 Cytogenetics, genetics, and a few genomics studies have documented variation for many
14 of high repetitive sequences among maize lines, which may also contribute to maize
15 evolution and domestication [2, 18-20]. In maize, highly repetitive sequences are
16 comprised of several major classes, including ribosome DNA (rDNA), knob repeats,
17 centromere satellite C DNAs (CentC), telomere repeats, and various retrotransposon
18 families. The rDNA repeats consist of two classes, 45S rDNA and 5S rDNA, which are
19 transcribed to ribosomal RNAs (rRNAs). 45S rRNA is further processed into 18S, 5.8S
20 and 26S mature rRNAs, which, are then assembled with the 5S rRNA into ribosome
21 subunits [21]. 5S rDNA loci are physically located at the distal of the long arm of
22 chromosome 2 [22], while 45S rDNA tandem arrays are clustered at the nucleolus
23 organizer region (NOR) located at the short arm of chromosome 6 in maize [23]. The

1 copy number of 45S rDNA repeats is highly variable between different maize lines,
2 possibly due to unequal crossover within large tandem repeats [24]. 5S rDNA loci, in
3 contrast, appear relatively stable [25]. Knob repeats are composed of highly condensed
4 heterochromatic regions and are cytologically visible on chromosomes of maize and the
5 wild relatives. Knobs consist primarily of a 180 bp repeat as well as a second 350 bp
6 repeat, the TR-1 repeat. Both types of repeats are organized in tandem arrays [26]. Whole
7 genome data of diverse maize lines and wild relatives indicate that genome size variation
8 correlates with knob content [14]. The number of knob repeats, knob size, and genomic
9 location vary dramatically among lines. Cytological detection of knobs in recombinant
10 inbred lines has been employed to genetically map knobs [27].

11

12 Centromeres are primarily made up with tandem satellite repeated CentC and
13 interspersed centromeric retrotransposons of maize (CRM), both of which exhibit varying
14 abundance across taxa [18-20]. Cytological evidence indicates that CRM elements, as the
15 name implies, are largely located at centromeres [28]. Recently, studies using next-
16 generation sequencing (NGS) data discovered that the abundance of CentC repeats is
17 reduced in domesticated maize, while the contents of CRM are increased in domesticated
18 maize, in comparison with the wild progenitor teosinte [19, 20]. Telomeres are the natural
19 ends of eukaryotic chromosomes. Telomere repeats typically consist of 5 to 8 nucleotide
20 highly conserved motifs, which function to recruit the proteins of the nucleoprotein
21 complex and protect chromosomes from instability. In most plants, the conserved motif is
22 TTTAGGG [29, 30]. Sub-telomeres are DNA sequences immediately adjacent to the
23 telomere repeats. Hybridization, using telomere-specific probes, revealed that telomere

1 lengths vary within a range of more than 25-fold among 22 surveyed maize inbred lines.
2 Genetic mapping analysis mapped additional *in trans* elements that control telomere
3 length [31]. Maize sub-telomeres consist of highly repetitive tandem sequences [32].
4 Here, telomere will be used as a general term for both telomere and sub-telomere repeats.
5 Collectively, highly repetitive sequences are largely organized into clusters in maize
6 genomes and variation in copy number is frequently observed.
7
8 NGS have provided in depth sequence data. However, accurate assessment of genome
9 structure and dynamics of repetitive sequence evolution using large NGS datasets
10 remains challenging due to the difficulty of unambiguous genome mapping and of
11 accurately reconstructing repetitive sequences with high-copy number. Additionally,
12 analysis relying on mapping reads to a reference assembly is subject to ascertainment
13 bias. Analysis independent of a reference genome sequence could reduce biases of
14 genome comparisons. In this study we quantify and characterize genome dissimilarity
15 through the comparison of k-mer abundances directly determined from sequencing data.
16 K-mers of a sequence represent all the possible subsequences of length k . K-mer analysis
17 has been widely applied in many genomic analyses, such as genome assemblies, genome
18 characterization, and metagenomic analysis [33-35]. Using B73 and Mo17 whole genome
19 shotgun (WGS) sequencing data, we quantified the level of the difference between the
20 two genomes at both non-repetitive and highly repetitive genome sequences. Genomic
21 locations influencing variation in copy number at highly repetitive sequences were
22 genetically mapped using WGS sequencing data of 280 intermated B73 and Mo17 double
23 haploids [10]. Furthermore, highly variable k-mers in diverse lines using *Zea mays*

1 HapMap2 WGS data [14, 15] were identified, revealing significant changes on highly
2 repetitive sequences during maize domestication and improvement.

3
4

5 **Results**

6 **K-mer analysis of genome dissimilarity between two maize inbred lines**

7 B73 and Mo17 are two maize elite inbred lines that are widely used in maize genetic and
8 genomic research. The two genomes have been extensively compared in both small and
9 genome-wide scales [4-8]. However, previous studies largely relied on a reference
10 genome, which produces systemic biases. To perform genome comparison with an
11 unbiased k-mer method that is independent of the reference genome, two HiSeq2500
12 lanes of Illumina data, using PCR-free prepared DNA libraries, were generated for each
13 of the two maize inbred lines B73 and Mo17, resulting in 450.9 and 445.3 millions of
14 pairs of 2x125 paired-end reads, respectively. More than 99% reads were retained after
15 the adaptor and quality trimming. The genome coverage of sequencing data (~46x) for
16 each genome enable the employment of error correction of sequencing reads. We use
17 abundance to represent counts of k-mer from sequencing data and use copy number to
18 represent sequence copies in a genome. The corrected reads were subjected to 25-nt k-
19 mer counting, resulting in approximately 749.7 and 738.7 millions of non-redundant k-
20 mers for B73 and Mo17, respectively. The similar shapes of the distributions of k-mer
21 abundances (Fig 1A) and the curves of cumulative contribution of k-mers with different
22 abundances to the genomes (Fig 1B) indicate that B73 and Mo17 exhibit overall similar
23 levels of genome complexities. The B73 and Mo17 abundance peaks are presumably
24 located at in single-copy k-mers (<http://www.broadinstitute.org/software/allpaths->

1 [lg/blog/wp-content/uploads/2014/05/KmerSpectrumPrimer.pdf](#)), which occur only once
2 in a genome (Fig 1A). The merged B73 and Mo17 k-mer abundances form a curve with
3 two peaks in k-mer abundances (Fig 1A). The lower abundance peak underneath the
4 original uncombined peaks consists of k-mers specific to either B73 or Mo17, while the
5 second higher frequency peak represents the common k-mers of the two genomes. This
6 novel approach was employed to visualize the difference of non-repetitive genomic
7 sequences between the two genomes. K-mer comparison indicates that only 60.9% of
8 single-copy k-mers are shared between the two maize cultivars, leaving a remaining
9 39.1% of the single-copy k-mers specific to each genome (Table 1). Based on the k-mer
10 distribution, the B73 genome size was estimated to be 2.38 Gb and consisted of 24.9%
11 single-copy k-mers, while 2.48 Gb with 23.7% single-copy k-mers for Mo17. The B73
12 genome size estimation agrees with that of 2.3 Gb estimated from the B73 genome
13 sequencing project [2]. The slightly larger estimated genome size of Mo17 versus B73
14 but the smaller proportion of single-copy sequences in Mo17 implies that distinct
15 contributions of repetitive sequences to two genomes, which indeed can be observed on
16 the curves of cumulative k-mer contribution to the genome at high abundant k-mers that
17 are representatives of highly repetitive sequences (Fig 1B).

18

19 **Table 1.** k-mers from single-copy regions in B73 and Mo17*

Category	Number of single-copy k-mers	% single-copy k-mers in either B73 or Mo17 [#]
B73 & Mo17 common	285,759,048	-
B73 specific	183,644,569	39.1%
Mo17 specific	183,441,358	39.1%

20 * k-mers with counts between 20 and 50 are considered to be single-copy k-mers

21 [#] percentage of genotype specific k-mers in all single-copy k-mers in either B73 or Mo17

22

1 **Divergence in copy number exhibited on highly repetitive DNA sequences**

2 Owing to the implication of the distinct constitution of high-copy genomic sequences
3 between B73 and Mo17, highly abundant k-mers (HAKmers, N= 802,668, Table S1) in
4 either B73 or Mo17 or both were examined. The majority of HAKmers exhibit similar
5 abundance in the two genomes but some are highly different (Fig 2A). Functional
6 annotation through a BLASTN of HAKmers to a *Zea mays* repeat database results in
7 552,371 annotated HAKmers each of which has at least one hit with the minimum e-
8 value of 0.1. The best hit of each HAKmer was referred to as the k-mer's functional
9 class. The major classes include retrotransposon, knob, rDNA, CentC, telomere, and a
10 variety of DNA transposon members (Table S2).

11

12 χ^2 statistical tests with a multiple test correction using the cutoff of 5% false discovery
13 rate (FDR) were performed to identify HAKmers showing differential abundance
14 between B73 and Mo17. A minimum of two-fold change in k-mer abundance was also
15 required. As a result, 11,413 and 2,633 differential abundance HAKmers respectively
16 showing higher abundance in B73 and Mo17 were identified, and, hereafter, referred to
17 as B73-gain and Mo17-gain HAKmers. Four major functional annotation classes, knob,
18 45S rDNA, CentC, and telomere, were found in these differential abundance HAKmers
19 (Fig 2B). Although retrotransposon derived k-mers (retrotransposon k-mers hereafter and
20 a similar expression was applied to other classes of k-mers, e.g., 45S rDNA k-mers to
21 represent k-mers derived from 45S rDNA) represent the largest class of HAKmers,
22 relatively few of these differ significantly in abundance (Table S2). Many knob k-mers
23 were identified and all belong to B73-gain k-mers, indicating more knob sequences in the

1 B73 genome. This is consistent with the previous cytological observation that B73, but
2 not Mo17, contains knobs at the long arms of chromosomes 5 and 7 [36, 37]. Despite the
3 changes in the knob content detected, no differential abundance HAKmers were found to
4 be TR-1 repeats. A similar finding was made for B73-gain telomere k-mers although the
5 number is much smaller (Fig 2B). Moreover, a number of k-mers derived from 45S
6 rDNA and CentC show gains in either B73 or Mo17. More 45S rDNA k-mers and less
7 CentC k-mers showing higher abundance were identified in B73 versus Mo17. Genomic
8 locations of these differential abundance HAKmers on the B73 genome were mapped
9 through aligning k-mers to the B73 reference genome (B73Refv3) (Fig 2C). From the
10 result, knob k-mers are clustered on multiple chromosomes (e.g., long arm of
11 chromosomes 1, 4, 5, 7 and a distal short arm region at chromosome 9), CentC k-mers are
12 largely located at or around centromeres, and telomere k-mers are identified at the ends
13 of chromosomes 1, 2, 4, 8 and 10. 45S rDNAs k-mers are predominantly clustered at the
14 short arm on chromosome 6, presumably the NOR. Note that such distributions based on
15 the reference genome rely on the quality of assemblies, and the assembly quality of
16 different regions might vary. The genome distribution plot also shows that 45S rDNA k-
17 mers are pervasive in other genome regions in addition to the NOR (Fig 2C, Fig S1).

18

19 To understand copy numbers of different classes of highly repetitive sequences in two
20 genomes, the total count of all the k-mers of each class was determined and normalized,
21 which represents the relative level of repetitiveness of each class. As a result, compared
22 to B73, approximately 55% and 22% reduction were respectively observed on knob and
23 telomere repeats, while 71%, 34%, 25% increased on CentC, 45S rDNA and 5S rDNA,

1 respectively, in Mo17 (Table S3). We also used abundances of the k-mers (N=3,533)
2 from the 45S rDNA regions conserved among multiple plant species to estimate the copy
3 number of 45S rDNA (Methods). The copy numbers of 45S rDNA in B73 and Mo17
4 were estimated to be around 3,658 and 5,063, respectively. Our estimation is in the range
5 of a previous estimation of placing rRNA gene number from 2,500 to 12,500 in 16 maize
6 lines [24]. Collectively, we discovered several major classes of repetitive sequences
7 showing differential copy number between B73 and Mo17, suggesting that two genomes
8 experience pronounced divergence with respect to copy number of highly repetitive
9 sequences. Because these repetitive sequences are largely clustered and tandemly
10 arrayed, high levels of copy number variation at these loci are likely caused by insertions
11 or deletions of large genomic segments due to aberrant crossing over or replication errors.

12

13 **Genetic mapping of genomic locations showing differential copy number of** 14 **repetitive sequences between B73 and Mo17**

15 Differential abundance of HAKmers from B73 and Mo17 results from distinct copy
16 numbers of genomic repetitive sequences from which k-mers have originated. The
17 segregation of such genomic sequences in a segregating population (e.g., recombinant
18 double haploids) derived from B73 and Mo17 results in different copy number among the
19 offspring. To map genomic locations showing the differentiation of copy number
20 between B73 and Mo17, low-coverage WGS sequencing of 280 individuals from inter-
21 mated B73xMo17 double haploids (IBM DHs) [10] was analyzed. First, the abundance of
22 each of differential abundance HAKmers from each DH line was determined and
23 normalized (Methods). K-mer abundance resembles a quantitative trait value, and the

1 genomic elements contributing their genomic copy number variation can be genetically
2 mapped using a quantitative trait locus (QTL) mapping approach (referred to as copy
3 number variation QTL, cnvQTL, hereafter). Using a high-density genetic map developed
4 with the same WGS data set from these 280 DH lines [10], the normalized counts of a k-
5 mer were input as phenotypic values for a genetic mapping analysis using the R package
6 rqt1. In total, 11,413 and 2,633 of B73- and Mo17-gain HAKmers were analyzed,
7 respectively. To determine the cutoff of log₁₀ likelihood ratio (LOD) of cnvQTL, each of
8 1,000 randomly selected HAKmers was subjected to a permutation test to determine the
9 LOD cutoff. All of these LOD cutoffs with the 5% type I error are in between 3 and 4.
10 Therefore the minimum LOD of 4 was used to declare mapping cnvQTL peaks (Table
11 S4). Only 0.3% B73-gain and 3% Mo17-gain HAKmers could not be mapped using this
12 approach. The majority of HAKmers, 74.5% B73-gain and 83.5% Mo17-gain, were
13 mapped to single major genomic locations, and the rest were mapped to 2-4 genomic
14 locations.

15

16 Functional annotation analysis of these mapped HAKmers revealed distinct mapping
17 locations for different sources of k-mers (Fig 3). For B73-gain HAKmers, knob and 45S
18 rDNA are two major sources (Table S5). Knobs k-mers were mapped to the long arms on
19 chromosomes 1, 5, and 7, of which the regions on chromosomes 5 and 7 were reported to
20 have differential knobs between B73 and Mo17 [36, 37]. All 2,205 45S rDNA k-mers
21 were mapped to around 13.5 Mb on chromosome 6 to which 11 retrotransposon k-mers
22 were also mapped. This mapping region is located at a short arm region on chromosome
23 6 which exhibits a presence-and-absence variation (PAV) that was identified in previous

1 comparative studies [7, 8]. Substantial copy gains of some type of 45S rDNA and some
2 retrotransposons in B73 at this region indicate the long PAV segment harbors rich
3 repetitive sequences. The differential abundance 45S rDNA k-mers are largely located at
4 the intergenic spacer (IGS) between 18S and 26S of 45S rDNA and a small proportion
5 are located at internal transcribed spacer (ITS) and 26S rRNA gene (Fig S2). On the same
6 chromosome, CentC k-mers were mapped to 62.8 Mb, suggesting the two genomes
7 contain distinct centromere compositions on chromosome 6. Moreover, telomere k-mers
8 were mapped to the ends of short arms of chromosomes 1, 2, 4, and 5. The further
9 analysis shows that B73 contains more copies of telomere repeats than Mo17 at
10 chromosomes 2, 4, 5, but less copies at chromosome 1 (Table S6).

11

12 45S rDNA and CentC are two major sources for Mo17-gain HAKmers (Table S7).
13 Interestingly, similar to B73-gain 45S rDNA HAKmers, Mo17-gain counterparts were
14 mapped to around 13.6 Mb on chromosome 6, although a long DNA segment on the B73
15 reference genome around that region is absent in Mo17. This indicates that B73 and
16 Mo17 likely contain different versions of 45S rDNA at the NOR. Furthermore, four 5S
17 rDNA k-mers (N=4) showing higher abundance in Mo17 were mapped to around 222.5
18 Mb on chromosome 2, consistent with a previous FISH result in which 5S rDNA was
19 mapped to the distal of chromosome 2 [22]. Significantly, Mo17-gain CentC k-mers were
20 mapped to multiple chromosomes. The centromeric regions at chromosomes 2, 4, 7, 8,
21 and 9 contribute to varying abundance of CentC k-mers. The same k-mers can be mapped
22 to the centromeres on multiple chromosomes, suggesting multiple centromeres co-
23 evolved to change CentC abundance.

1

2 **Different evolutionary origins of 45S rDNAs of B73 and Mo17, likely expanded, and**
3 **spread to regions other than the NOR after domestication**

4 From differential abundance HAKmers, an extreme type of k-mer was surprisingly
5 observed in which the k-mer was highly abundant in B73 or Mo17 but absent or very low
6 in the other, which are referred to as genotype-specific HAKmers (Fig 4A). In total, 162
7 B73-specific HAKmers and 103 Mo17-specific HAKmers were obtained. These
8 genotype-specific HAKmers were verified by using independent B73 and Mo17 WGS
9 sequencing data [14] without error correction. Additionally, all of the B73-specific
10 HAKmers can be perfectly aligned to the B73 reference genome, while only 3/103 Mo17
11 specific HAKmers were perfectly aligned to single locations at the NOR region. This
12 result confirms, at least, that Mo17 specific HAKmers are highly abundant in Mo17 but
13 hardly identified in the B73 genome. Interestingly, all of these genotype-specific
14 HAKmers are annotated to the class of 45S rDNA. K-mer analysis using IBM DH lines
15 WGS sequencing data indicates that each DH line predominated by either B73- or Mo17-
16 specific k-mers (Fig S3). Genetic mapping analysis of both B73- and Mo17-specific
17 HAKmers through cnvQTL shows that the NOR where 45S rDNA repeats are clustered
18 is largely responsible for the segregation of B73- and Mo17-specific HAKmers, further
19 suggesting that distinct types of high-copy 45S rDNAs are included at the B73 and Mo17
20 NORs (Fig 4B). A detailed analysis found that all these genotype-specific k-mers were
21 mapped to the IGS of the 45S rDNA unit.

22

1 To understand the origin of these genotype-specific k-mers, maize HapMap2 WGS
2 sequencing data, which includes lines from teosinte, landrace, and improved maize [14,
3 15], were subjected to k-mer analyses. The count of each of B73- and Mo17-specific k-
4 mers was determined for each HapMap2 line. To account for the variation of k-mer
5 abundance owing to non-genetic factors, such as sequencing depth and organelle DNA
6 contamination, a novel normalization approach was developed of which normalization
7 factors were determined by using the total counts of a set of conserved single-copy k-
8 mers across HapMap2 lines. Briefly single-copy k-mers were first obtained from both
9 B73 and Mo17 and the correlation of counts of each k-mer with the library sizes of all the
10 HapMap2 lines determined. Based on the assumption that a conserved single-copy k-mer
11 exhibits a high correlation with the sequencing library size, the top 5% k-mers
12 (N=49,955) with highest correlation efficiencies were used to calculate the normalization
13 factors. A principal component analysis (PCA) was performed using normalized
14 abundances of genotype-specific HAKmers (N=265) of HapMap2 lines. At a result, the
15 first two components (PC1 and PC2) explain 72.4% variation in normalized abundance
16 (Fig 4C). From the PCA plot, three distinct branches were formed and teosinte lines were
17 centralized at the intersection. Mo17 is located on the distal position of one branch but
18 B73 is not located at any of the branches. The PCA analysis implies that not all the
19 HapMap2 lines exhibit either of two extremely divergent patterns possessed in B73 and
20 Mo17.

21

22 To understand the abundance of these genotype-specific HAKmers in each HapMap2
23 line, the total normalized counts of all the B73- and Mo17-specific HAKmers were

1 separately determined. Total counts of the B73- and Mo17-specific HAKmers vary
2 dramatically among the HapMap2 lines (Fig 4D). It is notable that all teosinte lines
3 exhibit relatively low abundance, while many but not all maize lines show high
4 abundance in total counts. This result indicates that these particular types of 45S rDNA
5 repeats likely experienced appreciable expansion after domestication or shrinkage in
6 teosinte and some maize lines. Evidence was also found that B73-specific k-mers are
7 largely, but not only, located at the NOR. Indeed, the B73-specific k-mers can be
8 identified at many locations on all the chromosomes in the B73 genome (Fig 4A).
9 Presumably, the scattered distribution of these k-mers across all the chromosomes is the
10 consequence of the 45S rDNA spreading from the NOR. Moreover, all teosinte lines and
11 the majority of maize lines contain only either B73- or Mo17-specific HAKmers, while a
12 few landrace and improved lines consist of both. Our cnvQTL mapping result indicated
13 that both B73- and Mo17-specific HAKmers are predominantly located at the NOR. The
14 observed mixture of two rDNA types in some maize inbred lines are likely the
15 consequence of heterozygous residues or recombination at the NORs, although meiotic
16 recombination is substantially suppressed at the NOR [38]. It is also notable that the
17 proportion of lines with B73-specific types of 45S rDNAs in the improvement levels is
18 increased from teosinte to landrace, and from landrace to improved lines (Fig 4D),
19 possibly due to positive selection on the NOR or nearby regions. Previous studies also
20 suggested that this region was under selection during either domestication [15] or maize
21 improvement [16].

22

23 **Allelic expression of 45S rDNA in hybrids of B73 and Mo17**

1 The differences of 45S rDNA sequences in B73 and Mo17 enables the investigation of
2 the expression of two types of 45S rDNA in the hybrid of B73 and Mo17. Messenger
3 RNA (mRNA) is typically selected and enriched in final sequencing libraries in the
4 regular RNA-Seq (mRNA sequencing) procedure. However, it is almost impossible to
5 completely remove all rRNA, which allows the study of the expression of rRNA using
6 mRNA sequencing data. Two sets of RNA-Seq data were used. One is the transcriptomic
7 data of young maize primary roots in the B73, Mo17 and the reciprocal hybrids [39]. The
8 other is transcriptomic data of whole kernels at 0, 3, and 5 days after pollination (DAP)
9 and endosperms at 7, 10, and 15 DAP from reciprocal hybrids of B73 and Mo17 [40].
10 From both data sets, many sequences were aligned to 45S rDNA, proving that rRNA
11 sequences remained in mRNA sequencing data. The B73- and Mo17-specific 45S rDNA
12 k-mers can be used to trace the genotype-specific expression of 45S rDNA if their k-mer
13 abundance could be reliably measured in RNA-Seq. However, all these genotype-specific
14 k-mers are located at the IGS. The IGS is either not transcribed or accumulated at a level
15 as high as the rRNA genes (5.8S, 18S, and 26S), and IGS expression therefore cannot be
16 reliably detected. Fortunately, a single-nucleotide variant (SNV), A/T, was discovered on
17 the 26S rRNA gene and three pairs of k-mers harboring this SNV were identified in both
18 genomic sequencing and RNA-Seq data (Table S8). 72% and 28% B73 rDNAs carry A
19 and T, respectively, while almost 100% of Mo17 rDNAs carry T. A-carrying rDNAs
20 nearly completely dominated rRNA expression in primary roots of B73 (Fig 5A),
21 suggesting that not all rDNAs, as previously reported [41], are transcribed. In Mo17, T-
22 carrying rDNA is the only type of expressed rDNA. In the reciprocal hybrids, both types

1 were almost identically expressed in primary roots, although in both the reciprocal
2 hybrids the A and T types of rDNAs are unequal in abundance in their genomes (Fig 5A).
3
4 Using the time-course transcriptomic sequencing data of whole kernels and endosperms,
5 the allelic expression levels of the SNV (A/T) on the 26S rRNA gene in the reciprocal
6 hybrids of B73 and Mo17 were also examined (Fig 5B). As a result, detected rRNA
7 almost entirely belong to the maternal type in whole kernels at 0 DAP. Paternal rRNA
8 accumulation levels were gradually increased in whole kernels from 0 to 5 DAP. In
9 endosperms at 7, 10, and 15 DAP, the ratios of maternal to paternal rRNA expression are
10 not far from 2:1 that is the actual copy number ratio of maternal to paternal genomes,
11 indicating that both maternal and paternal rRNA copies are expressed at equal rates in
12 early endosperms.

13

14 **Marked changes of multiple types of highly repetitive genomic sequences during** 15 **domestication and maize improvement**

16 The finding that B73 and Mo17 exhibit substantial variation at high-copy genomic
17 sequences inspired an investigation of such variation among the HapMap2 lines. B73 and
18 Mo17 are included in the HapMap2 lines but in this analysis we wanted to identify k-
19 mers highly variable across the whole HapMap2 set, rather than the genotype-specific
20 high abundance k-mers defined using these two inbred lines. Using the HapMap2 WGS
21 sequencing data, k-mers showing high abundance (>1,000 counts per k-mer) in at least
22 five HapMap2 lines but low abundance (<10 counts per k-mer) in at least five other lines
23 were extracted, resulting in 8,462 highly variable k-mers. To examine the change of these

1 k-mers among three evolutionary groups, teosinte, landrace, and improved, an ANOVA
2 test was performed for each k-mer and a Bonferroni correction was conducted to account
3 for multiple testing. As a result, 2,016 k-mers exhibit significantly differential abundance
4 among three groups at the 5% type I error. Functional annotation through a BLASTN of
5 k-mers to the repeat database results in 1,090 annotated k-mers (Methods). The k-mers
6 exhibiting significantly differential abundance among evolutionary groups were annotated
7 to the functional classes of 45S rDNA, CentC, retrotransposon (copie and gypsy), and
8 knob. The low rate (only ~54%) of k-mers that are annotated using the repeat database is
9 because a relatively high proportion of k-mers are derived from organelle genomes,
10 which likely reflects the diversity of organelle genomes. To focus on highly repetitive
11 sequences from nuclear genomes, only the functionally classified k-mers were subjected
12 to a clustering analysis using the software MCLUST [42], resulting in 12 clusters (Table
13 S9, Fig S4). Nine major clusters were further manually grouped into two groups (Table
14 2, Fig 6A, B). In detail, k-mer abundance of the group 1 was significantly decreased
15 during maize domestication and/or improvement. K-mers from this group are largely
16 annotated as CentC (example in Fig 6C) and 45S rDNA, as well as a small number of k-
17 mers from knob, DNA transposons, and retrotransposons (Table 2). K-mer abundance of
18 the group 2 was substantially increased during maize domestication and/or improvement.
19 K-mers from this group are annotated as retrotransposon members (CRM and
20 unclassified retrotransposon) (example in Fig 6D) and 45S rDNA. The observation of
21 45S rDNA in both groups 1 and 2 suggests that some types of 45S rDNA sequences
22 experienced substantial expansion while others experienced substantial shrinkage during
23 maize domestication and improvement.

1

2 **Table 2.** Number of functionally classified k-mers in different clustering groups

Class	Decrease during domestication	Increase during domestication
CentC	212	0
CRM*	0	121
Knob	12	0
45S rDNA	266	81
DNA transposon	9	0
Retrotransposon [§]	54	138

3 * k-mers were annotated unknown centromere retrotransposons

4 [§] unclassified retrotransposon

5

6 Abundance of k-mers that were generated from the conserved regions of 45S and 5S

7 rDNA across multiple plant species was estimated for each HapMap2 line. The median of

8 abundances of all the 45S rDNA k-mers from a HapMap2 line and the counterpart of 5S

9 rDNA k-mers were used to represent the genomic copy number level of 45S and 5S

10 rDNA of the line, respectively. Most landrace maize lines exhibit lower copy number

11 than teosinte, while maize improved lines shows much higher diversity in term of 45S

12 rDNA copy number (Fig 6E). This observation suggests there were a possible shrinkage

13 or a strong selection on the NOR region during domestication, and a re-expansion of 45S

14 rDNAs during improvement. No association with evolutionary groups was observed for

15 copy number of 5S rDNAs. Additionally, the correlation of copy number of 45S and 5S

16 rDNAs among HapMap2 lines is weak ($R^2 = 0.059$), suggesting that dosage balance in

17 genomic copy number between 45S and 5S rDNAs, which was observed in human and

18 mouse genomes [43], was not required in *Zea* genomes.

19

20

21 **Discussion**

1 This study employs a novel k-mer analysis strategy for comparative genomics.
2 Reference-independent quantification of NGS data allows precise and unbiased
3 comparison of the genomic constitutions, particularly highly repetitive sequences that are
4 generally overlooked from regular analyses. Our results offer insightful information
5 about copy number abundance, genomic locations, and evolution of highly repetitive
6 sequences among maize genomes, and provide an unbiased genome comparative method
7 for mining existing and incoming deluge of NGS data to gain biological insights.

8

9 **Unbiased k-mer analysis**

10 K-mers represent all the possible subsequences of length k from a sequencing read. For
11 genome assembly using short NGS reads, k-mers are typically generated from sequencing
12 data to construct *de Bruijn* graphs [33]. In addition to genome assemblies, K-mer analysis
13 has been applied to many other genomic analyses, including but not limited to
14 characterization of repeat content and heterozygosity [34], estimation of genome size
15 [35], evaluation of metagenomic dissimilarity [44], and identification of causal genetic
16 variants conferring phenotypic traits [45]. Any size of k-mers can be used for k-mer
17 analysis. Using smaller sized k-mers, sequencing data are condensed to less total k-mers,
18 and a smaller number of k-mers are derived from single-copy regions, resulting in higher
19 degree of information loss. Increasing size of k-mers increases both the total k-mer
20 number and the number of single-copy k-mers, which is compromised by increased
21 computation cost. Additionally, higher size of k-mers is more vulnerable to sequencing
22 errors contained in sequencing reads. The impact of sequencing errors could be alleviated
23 by error correction of sequences. The choice of k-mer length of 25 nt is an optimal size

1 for human-sized genomes which was used in ALLPATHS-LG for analyzing k-mer
2 abundance spectrum [46].
3
4 K-mer based methods are independent of read mapping that typically relies on a
5 reference sequence, which allows the establishment of a fair comparison between
6 genomes. For WGS data from either the same or different species, k-mer analysis can be
7 directly applied to quantify the level of dissimilarity between individuals as long as WGS
8 data are comparable. Low-coverage WGS data are sufficient to deliver reliable counts for
9 k-mers derived from highly repetitive sequences. The critical issue is to develop a reliable
10 normalization approach to account for non-genomic variation in data due to different
11 sequence depths, varying levels of organelle DNAs, or contaminations from other
12 species, particularly from microbes. In this study, we used total counts of a great number
13 of single-copy k-mers that are conserved in the examined individuals to determine
14 normalization factors. This normalization method is expected to well account for non-
15 genomic variation. With high-coverage WGS data from multiple individuals, any types of
16 genomic polymorphisms at either low or highly repetitive genomic regions would be
17 unbiasedly represented by abundance of corresponding k-mers. In particular, copy
18 number variation can be well captured by analyzing k-mer abundances. With that respect,
19 one of potential applications of k-mer analysis is to perform genome-wide association
20 with abundances of k-mers, which could retrieve some associated genetic elements that
21 are unable to be detected using reference-based approaches. Collectively, the k-mer based
22 approach alleviates ascertainment biases introduced by reference-based methods, and
23 should provide the complement to many existing genome analyses.

1

2 **HAKmer copy number variation QTL mapping**

3 Using low-coverage WGS sequencing data of the IBM DH lines, a cnvQTL genetic
4 mapping strategy was developed to map the genomic regions determining variation of k-
5 mer abundance among DHs. As a result, the vast majority of differential abundance
6 HAKmers between B73 and Mo17 were confidently mapped. The success of mapping
7 differential abundance HAKmers from a variety of sources, including 45S rDNA, CentC,
8 knobs, and telomeres, proved the effectiveness of the cnvQTL mapping. The fact that k-
9 mers from rDNAs, CentC, telomeres, and knobs were all mapped to the expected regions
10 where they are physically located suggests that no recognizable *trans* elements control
11 the segregation of these repetitive sequence copies. The lack of *trans* elements makes
12 sense because these repetitive sequences, although they evolve rapidly, are steadily
13 maintained in each of two maize inbred lines.

14

15 We obtained a high-resolution map identifying coordinates contributing to differences in
16 abundance of k-mers for many types of repetitive sequences in B73 and Mo17. These
17 mapped genomic regions accurately mark the locations of clusters of repetitive sequences
18 and corroborate many previous findings, as well as provide additional insight into the
19 differentiation between B73 and Mo17. For example, both B73- and Mo17-gain 45S
20 rDNA k-mers were mapped to around 13.5 Mb (B73Ref3) on chromosome 6 where a
21 large PAV on the order of a megabase between B73 (presence) and Mo17 (absence) has
22 been found. The result that Mo17-gain 45S rDNA k-mers were mapped at this PAV
23 region, presumably located on the NOR, indicates that Mo17 has distinct 45S rDNA

1 sequences to replace the missing version of 45S rDNA at the Mo17 NOR. Moreover,
2 some Mo17-gain 45S rDNA k-mers were mapped to 210.5 Mb at the long arm of
3 chromosome 1 that was not discovered previously, suggesting Mo17 contains a 45S
4 rDNA cluster with significantly elevated copy number of 45S DNA at that region. Using
5 a set of k-mers from the 45S rDNA specific sequences that are conserved among maize,
6 rice, and barley, we estimated that the copy number of 45S rDNA in B73 and Mo17 is
7 3,658 and 5,063, respectively. The Mo17-gain of 45S rDNA at chromosome 1, at least
8 partially, explains higher copy number of 45S rDNA in Mo17 relative to B73.

9

10 Our cnvQTL mapping data genetically confirm the differential abundance of knob
11 contents between B73 and Mo17. In addition to the long arms on chromosomes 5 and 7
12 that were reported previously [36, 37], a distal region (293.5 Mb) at the long arm of
13 chromosome 1 shows higher abundance of knob repeats in B73. The reduction of knob
14 repeats on chromosomes 1, 5, 7 primarily accounts for the 55% loss of knob repeats in
15 Mo17. What is more, detailed differentiation in CentC and telomere sequences were
16 revealed. The increase of CentC repeats in multiple chromosomes in Mo17 indicates a
17 possible common driving force involved in these parallel directional changes in copy
18 number in a genome.

19

20 **45S rRNA expression in hybrids**

21 Nucleolar dominance is a phenomenon specifically observed in hybrids in which the
22 NOR of one parent are dominant over the other of which rRNA is silenced. rRNA
23 silencing involves epigenetic modifications of chromatin [41]. To examine nucleolar

1 dominance in hybrids, allelic expression of rRNA needs to be precisely quantified. We
2 have showed that rRNA is well represented even in mRNA sequencing data where rRNA
3 was selected against. The divergence of 45S rDNA sequences between B73 and Mo17
4 provides the possibility for examining rRNA allelic expression in their hybrids. However,
5 most polymorphisms of 45S rDNA are located at IGS and ITS whose expression is hardly
6 detected using the examined mRNA sequencing data. Fortunately, we identified the k-
7 mers harboring a SNV polymorphic site on the 26S rRNA gene. The paired polymorphic
8 k-mers are respectively, and nearly exclusively, expressed in one of B73 and Mo17
9 inbred lines, which sets an ideal marker to measure the expression of two types of 45S
10 rDNA in the hybrid of B73 and Mo17. The k-mer abundance analysis indicates that
11 Mo17 contains higher copy number of 45S rDNA than B73. Using transcriptomic
12 sequencing data of primary roots, we observed the expression levels of rRNAs derived
13 from two parents were equalized in both reciprocal hybrids, suggesting no nucleolar
14 dominance occurs in the primary roots of the hybrid of B73 and Mo17 and also implying
15 that an unknown mechanism exists to regulate dosage compensation.
16
17 Using transcriptomic sequencing data of early whole kernels and endosperms, we
18 observed that the maternal rRNA expression is almost completely dominant in the whole
19 kernels at 0 DAP, followed by the gradual increase of paternal rRNA expression from 0
20 to 5 DAP. It is not clear that inequality of maternal and paternal rRNA expression in
21 early whole kernels is merely due to the distinct proportions of maternal and paternal
22 genomes or its combination with the transcriptional suppression of paternal rRNA.
23 Further examination through precise quantification of both rRNA and rDNA could

1 address this question. Maize endosperm is a triploid, containing $2n$ of the material
2 genome and $1n$ of the paternal genome. In early endosperms at 7, 10, and 15 DAP, the
3 maternal rRNA expression is around twice as high as the paternal rRNA expression,
4 indicating both maternal and paternal rRNA function, and, therefore, no nucleolar
5 dominance was observed at the tissues examined.

6

7 **Implications for maize evolution**

8 Maize was domesticated from a wild species teosinte (*Zea mays* ssp. *parviglumis*)
9 approximately 9-10 thousands years ago [12, 47]. Genetic evidence supports a single
10 domestication and the post-domestication introgression from other wild relatives
11 including *Zea mays* ssp. *Mexicana* [12, 15, 48]. The two distinct versions of 45S rDNA
12 repeats traced by B73- and Mo17-specific k-mers at the NOR can be identified in
13 different teosinte lines, indicating maize NORs originated from multiple ancient sources.
14 The lower abundance of B73- and Mo17-specific k-mers in all examined teosinte but
15 higher abundance in most landraces and improved maize lines suggests an expansion of
16 certain types of rDNA repeats after domestication. Our observation that identical
17 genotype-specific sequences are spread throughout the entire genome also raises
18 interesting questions about the evolutionary past and origin of these sequences in relation
19 to the NOR. Given evidence for a single domestication event and our observation of the
20 local expansion of genotype-specific 45S rDNA sequences during maize domestication
21 and improvement, flow of rDNA repeats away from the NOR following domestication is
22 a more likely hypothesis. While the translocating mechanism can be either RNA- or
23 DNA-mediated, our observation that spread regions consist of tandem arrays of intact

1 45S rDNA repeats suggests that this translocating mechanism is likely DNA-mediated.
2 Spreading phenomena were observed for knob repeats and centromere retrotransposon
3 members in both our results (Fig. 2) and previous studies [18, 27, 49, 50]. Spreading
4 sequences might serve as seeds that could eventually form new clusters of repetitive
5 sequences, such as nascent knobs or NORs.
6
7 To further characterize flux of repetitive DNA during evolution, we identified k-mer
8 sequences showing strikingly differential abundance among three groups, teosinte,
9 landraces, and improved lines. Nearly all of these differential abundance k-mers
10 displayed distinct patterns of either increase or decrease in abundance from teosinte to
11 maize. rDNA k-mers make up the largest class of differential abundance k-mers. While
12 83 45S rDNA k-mers showed increasing abundance during this evolutionary time-frame,
13 266 showed marked loss. Additional analysis of relative copy numbers of 45S rDNA of
14 HapMap2 lines also showed shrinkages and expansions of 45S rDNA repeats from
15 teosinte to maize lines. In contrast, all differential abundance CentC k-mers were
16 observed to decrease in abundance, strongly suggesting the shrinkage of CentC during
17 domestication. The reverse trend is seen for CRM k-mers, which are dramatically
18 elevated during domestication. This result replicates similar findings discussed in two
19 recent centromere publications [19, 20]. In addition, other retrotransposon members vary
20 greatly among historical groups. Increasing evidence shows that transposons play
21 important roles in adaptation and evolution [51, 52]. The dramatic change in copy
22 number of transposon elements during maize domestication could affect transcription and
23 gene function by disrupting genes via direct integration in functional genic regions,

1 providing new regulatory elements, and spreading epigenetic status to nearby genes [53,
2 54]. In summary, our k-mer analyses offers a single-base resolution to trace dynamics of
3 *Zea mays* genomes which has been appreciated through cytogenetics, molecular,
4 genetics, and genomics studies, providing valuable insights into the contents and
5 organization of highly repetitive sequences in maize.

6

7

8 **Methods**

9 **Plant materials and extraction of nucleus genomic DNA**

10 Two sources of B73 (PI 550473) were used, including seeds from Patrick Schnable
11 laboratory and North Central Regional Plant Introduction Station (NCRPIS). All Mo17
12 (PI 558532) seeds were originated from NCRPIS. Seeds of two genotypes were
13 germinated and grown in growth chamber at 28°C, with a photoperiod of 14:10h
14 (light:dark). 15~20 grams of fresh leaves of seedlings at 2-3 leaf-stage were harvested,
15 frozen in liquid nitrogen, and homogenized with liquid nitrogen to fine powder. The
16 nuclei were isolated using a protocol modified from Zhang's approach [55], followed by
17 using the Qiagen DNeasy Plant Mini Kit protocol to extract nucleus DNA.

18

19 **WGS sequencing of B73 and Mo17**

20 Genomic DNAs from nuclei were used for PCR-free library preparation. Two replicates
21 of each of B73 and Mo17 were whole genome shotgun sequenced with one sample per
22 lane in HiSeq2000. 2x125 bp paired-end data were generated. Sequencing was conducted
23 at BGI Genomics Co., Ltd., Shenzhen, China.

24

1 **Error correction and genome size estimation**

2 B73 and Mo17 whole genome sequences were trimmed to remove adaptor
3 contaminations and low quality sequences with Trimmomatic (version 3.2) [56]. The
4 clean data were subjected to error correction using the error correction module
5 (ErrorCorrectReads.pl) in ALLPATHS-LG [46] with the parameters of
6 “PHRED_ENCODING=33 PLOIDY=1”. Genome size was estimated during the
7 procedure of error correction.

8

9 **K-mer counting**

10 Corrected sequences were subjected to k-mer counting using the count function in
11 JELLYFISH [57] with the k-mer size of 25 nt.

12

13 **Estimation of genomic copy number of 45S rDNAs in B73 and Mo17**

14 Quantification of rDNA copy number was performed using k-mers generated from the
15 conserved regions of 45S rDNA among maize, rice, and barley. K-mers were aligned to
16 the *Zea mays* repeat database (TIGR_Zea_Repeats.v3.0) to exclude any k-mers aligning
17 to non-45S rDNA repeats, and to the B73Ref3 mitochondrial and plastid sequences to
18 exclude k-mers that are not exclusively nuclear. Abundance of the 45S rDNA k-mers was
19 evaluated for each B73 and Mo17. Abundances of these conserved k-mers in B73 and
20 Mo17 were normalized by division by the respective estimated abundances for single-
21 copy k-mers in order to estimate the number of 45S rDNA repeats in each genome. The
22 median value of all conserved k-mers was the estimation of the rDNA copy number.

23

1 **Identification of HAKmers with significant abundance between B73 and Mo17**

2 High-abundance k-mers (HAKmers) in B73 or Mo17 were extracted, each of which is
3 required to have at least 20,000 of total of B73 and Mo17 counts. A χ^2 statistical test for
4 each HAKmer was performed to test the null hypothesis of no relationship between k-
5 mer counts and the genotypes (B73 and Mo17). P-values of all HAKmers were corrected
6 to account for multiple tests [58]. The differential abundance of HAKmers were declared
7 if adjusted p-values are smaller than 5% and fold change in k-mer abundance between
8 B73 and Mo17 is not less than 2.

9

10 **Functional annotation of HAKmers**

11 The *Zea mays* repeat database (TIGR_Zea_Repeats.v3.0) was downloaded from the plant
12 repeat database that is currently maintained by Michigan State University
13 (plantrepeats.plantbiology.msu.edu). BLASTN was performed with the word size of 12 to
14 identify hits in the *Zea mays* repeat database for each HAKmer. The top hit with the e-
15 value cutoff of 0.1 was referred to as the functional annotation.

16

17 **K-mer mapping to the B73 reference genome**

18 K-mer mapping to the B73 reference genome (B73Ref3) was conducted by using Bowtie
19 (version 1.1.2) to identify all possible perfect hits.

20

21 **Genetic mapping of HAKmers via cnvQTL**

22 Resequencing data of 280 DH lines of the IBM Syn10 population used to build an ultra-
23 high density genetic map [10] were trimmed with Trimmomatic (version 3.2) [56].

1 Remaining clean reads were subjected to k-mer counting with JELLYFISH. The k-mer
2 size is 25 nt. The abundance of each HAKmer with differential abundance in B73 and
3 Mo17 was determined in each DH line. The total counts (C) of a million of randomly
4 selected B73 and Mo17 common single-copy k-mers in each DH line were determined.
5 The normalization factor for the i th line was calculated by using the formula $NC_i /$
6 $\sum_{i=1}^n C_i$, where N is the total number of IBM DH lines. The designation single-copy was
7 determined by k-mer abundance from whole genome sequencing data for both B73 and
8 Mo17 and confirmed by alignments to the B73ref3. Normalized abundance of a HAKmer
9 was treated as a quantitative trait. For each HAKmer, a genetic mapping resembling a
10 QTL detection implemented in an R package `rqtl` [59] was performed to identify the
11 genomic locations contributing the HAKmer abundance.

12

13 **Identification of B73- and Mo17-specific HAKmers**

14 To identify extremely unbalanced HAKmers that show extremely low abundance in one
15 of two datasets from B73 and Mo17, the maximum number of 10 was used as the cutoff.
16 Note that the minimum total abundance from B73 and Mo17 is 20,000 for HAKmers. If a
17 HAKmer exhibits extremely low abundance (≤ 10) in one genotype, it must be high
18 ($> 19,990$) in the other genotype. An extremely unbalanced HAKmer of which only one
19 genotype, B73 or Mo17, showing high abundance is called B73 or Mo17 specific
20 HAKmers.

21

22 **HapMap2 data and k-mer analysis**

1 Resequencing data of *Zea mays* HapMap2 lines [14, 15] were downloaded and trimmed
2 with Trimmomatic (version 3.2), followed by 25 nt k-mer analysis using JELLYFISH. To
3 make comparable k-mer abundances in different lines, a novel normalization method was
4 developed. In this method, a set of “conserved single-copy k-mers” across HapMap2
5 lines was identified, which are single-copy in almost all lines. For each of these k-mers,
6 k-mer abundances of HapMap2 lines should show a high correlation with their
7 sequencing library sizes. In detail, the k-mer abundance of each HapMap2 line was
8 determined for each of one million of B73 and Mo17 common single-copy k-mers that
9 we identified. For each k-mer, a correlation of k-mer abundances of HapMap2 lines with
10 their library sizes was calculated. The top 5% k-mers with the highest correlations
11 (N=49,955) were selected, which are deemed as “conserved single-copy k-mers”. The
12 total counts (C) of conserved single-copy k-mers per Hapmap2 line were determined. The
13 normalization factor for the i th line was calculated by using the formula $NC_i / \sum_{i=1}^n C_i$,
14 where N is the total number of HapMap2 lines.

15

16 **PCA of k-mer abundance of B73 and Mo17 specific k-mers in HapMap2**

17 PCA was performed using normalized k-mer abundances of B73 and Mo17 specific k-
18 mers in HapMap2. The R function of *prcomp* was used for the PCA.

19

20 **Allelic expression of rRNA in hybrids of B73 and Mo17**

21 The RNA-Seq data of young maize primary roots in the B73, Mo17 and their reciprocal
22 hybrids [39] and the time-course sequencing RNA-Seq data of whole kernels at 0, 3, and
23 5 DAP and endosperms at 7, 10, and 15 DAP from reciprocal hybrids of B73 and Mo17

1 [40] were downloaded. Sequencing reads were subjected to quality, adaptor trimming,
2 and k-mer counting with the size of 25 nt. The expression abundance of 45S rDNA k-
3 mers harboring a polymorphic site was used to assess allelic expression.

4

5 **Identification of highly variable k-mers in *Zea mays***

6 Abundances of k-mers were determined in each HapMap2 line. K-mer abundances were
7 normalized using normalization factors calculated from a “conserved single-copy k-
8 mers”. Highly variable k-mers were extracted using the hard-filtering criteria that require
9 >1,000 counts per k-mer per line in at least five HapMap2 lines but <10 counts in at least
10 five other lines.

11

12 **Identification of highly variable k-mers with significant differential abundance** 13 **among evolutionary groups**

14 Normalized counts of each k-mer for all HapMap2 lines were subjected to an ANOVA
15 test. The genotype variable has three levels: teosinte, landrace, and improved. The null
16 hypothesis is that k-mer abundances are independent of the genotype evolutionary
17 groups. Then the Bonferroni approach was applied for multiple test correction at the 5%
18 type I error.

19

20 **MCLUST to classify highly variable k-mers showing significantly differential** 21 **abundance among evolutionary groups**

22 K-mers exhibiting significant differential abundance among three genotype groups were
23 subjected to a clustering analysis using MCLUST [42]. For each k-mer, each count was

1 scaled by being divided by the maximum count value of this k-mer. Scaled counts of k-
2 mers were then used for the clustering using the parameters of “G=1:12, modelNames
3 ='EEE’”.

4

5 **Estimation of relative genomic copy number of rDNAs in HapMap2 lines**

6 The 45S rDNA k-mers used to estimate 45S rDNA copy number in B73 and Mo17 were
7 used to estimate relative copy number level of each HapMap2 line. In each line, the
8 median abundance value of the k-mers represents the 45S rDNA copy number. The same
9 method was used to determine 5S rDNA copy number level. The 5S rDNA k-mers were
10 derived from the 5S rDNA sequence that is conserved among maize, rice, and wheat and
11 were not aligned to B73 organelle genomes and other repetitive sequences.

12

13 **Data access**

14 B73 and Mo17 Illumina sequencing data have been deposited at Sequence Read Archive
15 (SRA accession number: SRP082260).

16

17 **Acknowledgments**

18 We thank Drs. Nathan Springer, Eduard Akhunov, and Jeffrey Ross-Ibarra for
19 discussions and valuable suggestions. We thank the support from the Kansas Agricultural
20 Experiment Station of Kansas State University. This is contribution number 17-072-J
21 from the Kansas Agricultural Experiment Station. We thank the support from the
22 Agricultural Science and Technology Innovation Program of CAAS.

23

1 REFERENCES

- 2 1. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-
3 generation haplotype map of maize. *Science*. 2009;326(5956):1115-7. doi:
4 10.1126/science.1177837. PubMed PMID: 19965431.
- 5 2. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73
6 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326(5956):1112-5.
7 doi: 10.1126/science.1178534. PubMed PMID: 19965430.
- 8 3. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, et al. High-
9 resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun*.
10 2015;6:6914. doi: 10.1038/ncomms7914. PubMed PMID: 25881062; PubMed Central
11 PMCID: PMC4411285.
- 12 4. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454
13 transcriptome sequencing. *Plant J*. 2007;51(5):910-8. doi: 10.1111/j.1365-
14 313X.2007.03193.x. PubMed PMID: 17662031; PubMed Central PMCID:
15 PMC2169515.
- 16 5. Liu S, Chen HD, Makarevitch I, Shirmer R, Emrich SJ, Dietrich CR, et al. High-
17 throughput genetic mapping of mutants via quantitative single nucleotide polymorphism
18 typing. *Genetics*. 2010;184(1):19-26. doi: 10.1534/genetics.109.107557. PubMed PMID:
19 19884313; PubMed Central PMCID: PMC2815916.
- 20 6. Fu Y, Wen TJ, Ronin YI, Chen HD, Guo L, Mester DI, et al. Genetic dissection
21 of intermated recombinant inbred lines using a new genetic map of maize. *Genetics*.
22 2006;174(3):1671-83. doi: 10.1534/genetics.106.060376. PubMed PMID: 16951074;
23 PubMed Central PMCID: PMC1667089.
- 24 7. Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, et al. Maize inbreds exhibit
25 high levels of copy number variation (CNV) and presence/absence variation (PAV) in
26 genome content. *PLoS Genet*. 2009;5(11):e1000734. doi: 10.1371/journal.pgen.1000734.
27 PubMed PMID: 19956538; PubMed Central PMCID: PMC2780416.
- 28 8. Belo A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A. Allelic genome
29 structural variations in maize detected by array comparative genome hybridization. *Theor*
30 *Appl Genet*. 2010;120(2):355-67. doi: 10.1007/s00122-009-1128-9. PubMed PMID:
31 19756477.
- 32 9. Liu S, Ying K, Yeh CT, Yang J, Swanson-Wagner R, Wu W, et al. Changes in
33 genome content generated via segregation of non-allelic homologs. *Plant J*.
34 2012;72(3):390-9. doi: 10.1111/j.1365-313X.2012.05087.x. PubMed PMID: 22731681.
- 35 10. Liu H, Niu Y, Gonzalez-Portilla PJ, Zhou H, Wang L, Zuo T, et al. An ultra-high-
36 density map as a community resource for discerning the genetic basis of quantitative
37 traits in maize. *BMC Genomics*. 2015;16:1078. doi: 10.1186/s12864-015-2242-5.
38 PubMed PMID: 26691201; PubMed Central PMCID: PMC4687334.
- 39 11. Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, et al.
40 Pervasive gene content variation and copy number variation in maize and its
41 undomesticated progenitor. *Genome Res*. 2010;20(12):1689-99. doi:
42 10.1101/gr.109165.110. PubMed PMID: 21036921; PubMed Central PMCID:
43 PMC2989995.
- 44 12. van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, de Jesus
45 Sanchez Gonzalez J, et al. Genetic signals of origin, spread, and introgression in a large
46 sample of maize landraces. *Proc Natl Acad Sci U S A*. 2011;108(3):1088-92. doi:

- 1 10.1073/pnas.1013011108. PubMed PMID: 21189301; PubMed Central PMCID:
2 PMCPMC3024656.
- 3 13. da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, et
4 al. The origin and evolution of maize in the Southwestern United States. *Nat Plants*.
5 2015;1:14003. doi: 10.1038/nplants.2014.3. PubMed PMID: 27246050.
- 6 14. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize
7 HapMap2 identifies extant variation from a genome in flux. *Nat Genet*. 2012;44(7):803-
8 7. doi: 10.1038/ng.2313. PubMed PMID: 22660545.
- 9 15. Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, et
10 al. Comparative population genomics of maize domestication and improvement. *Nat*
11 *Genet*. 2012;44(7):808-11. doi: 10.1038/ng.2309. PubMed PMID: 22660546.
- 12 16. Jin ML, Liu HJ, He C, Fu JJ, Xiao YJ, Wang YB, et al. Maize pan-transcriptome
13 provides novel insights into genome complexity and quantitative trait variation. *Sci Rep*-
14 *Uk*. 2016;6. doi: ARTN 18936
15 10.1038/srep18936. PubMed PMID: WOS:000368378500001.
- 16 17. Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, et al. Genome-wide genetic
17 changes during modern breeding of maize. *Nat Genet*. 2012;44(7):812-5. doi:
18 10.1038/ng.2312. PubMed PMID: 22660547.
- 19 18. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, et al. Maize
20 centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals
21 dynamic Loci shaped primarily by retrotransposons. *PLoS Genet*. 2009;5(11):e1000743.
22 doi: 10.1371/journal.pgen.1000743. PubMed PMID: 19956743; PubMed Central
23 PMCID: PMCPMC2776974.
- 24 19. Bilinski P, Distor K, Gutierrez-Lopez J, Mendoza GM, Shi J, Dawe RK, et al.
25 Diversity and evolution of centromere repeats in the maize genome. *Chromosoma*.
26 2015;124(1):57-65. doi: 10.1007/s00412-014-0483-8. PubMed PMID: 25190528.
- 27 20. Schneider KL, Xie Z, Wolfgruber TK, Presting GG. Inbreeding drives maize
28 centromere evolution. *Proc Natl Acad Sci U S A*. 2016;113(8):E987-96. doi:
29 10.1073/pnas.1522008113. PubMed PMID: 26858403; PubMed Central PMCID:
30 PMCPMC4776452.
- 31 21. Layat E, Saez-Vasquez J, Tourmente S. Regulation of Pol I-transcribed 45S
32 rDNA and Pol III-transcribed 5S rDNA in Arabidopsis. *Plant Cell Physiol*.
33 2012;53(2):267-76. doi: 10.1093/pcp/pcr177. PubMed PMID: 22173098.
- 34 22. Li L, Arumuganathan K. Physical mapping of 45S and 5S rDNA on maize
35 metaphase and sorted chromosomes by FISH. *Hereditas*. 2001;134(2):141-5. PubMed
36 PMID: 11732850.
- 37 23. Phillips RL, Weber DF, Kleese RA, Wang SS. The Nucleolus Organizer Region
38 of Maize (*ZEA MAYS* L.): Tests for Ribosomal Gene Compensation or Magnification.
39 *Genetics*. 1974;77(2):285-97. PubMed PMID: 17248655; PubMed Central PMCID:
40 PMCPMC1213129.
- 41 24. Buescher PJ, Phillips RL, Brambl R. Ribosomal RNA contents of maize
42 genotypes with different ribosomal RNA gene numbers. *Biochem Genet*. 1984;22(9-
43 10):923-30. PubMed PMID: 6517855.
- 44 25. Rivin CJ, Cullis CA, Walbot V. Evaluating quantitative variation in the genome
45 of *Zea mays*. *Genetics*. 1986;113(4):1009-19. PubMed PMID: 3744025; PubMed Central
46 PMCID: PMCPMC1202908.

- 1 26. Ananiev EV, Phillips RL, Rines HW. A knob-associated tandem repeat in maize
2 capable of forming fold-back DNA segments: are chromosome knobs megatransposons?
3 Proc Natl Acad Sci U S A. 1998;95(18):10785-90. PubMed PMID: 9724782; PubMed
4 Central PMCID: PMCPMC27973.
- 5 27. Ghaffari R, Cannon EK, Kanizay LB, Lawrence CJ, Dawe RK. Maize
6 chromosomal knobs are located in gene-dense areas and suppress local recombination.
7 Chromosoma. 2013;122(1-2):67-75. doi: 10.1007/s00412-012-0391-8. PubMed PMID:
8 23223973; PubMed Central PMCID: PMCPMC3608884.
- 9 28. Lamb JC, Birchler JA. Retroelement genome painting: cytological visualization
10 of retroelement expansions in the genera Zea and Tripsacum. Genetics.
11 2006;173(2):1007-21. doi: 10.1534/genetics.105.053165. PubMed PMID: 16582446;
12 PubMed Central PMCID: PMCPMC1526525.
- 13 29. McKnight TD, Shippen DE. Plant telomere biology. Plant Cell. 2004;16(4):794-
14 803. doi: 10.1105/tpc.160470. PubMed PMID: 15064365; PubMed Central PMCID:
15 PMCPMC526047.
- 16 30. Yu W, Lamb JC, Han F, Birchler JA. Telomere-mediated chromosomal truncation
17 in maize. Proc Natl Acad Sci U S A. 2006;103(46):17331-6. doi:
18 10.1073/pnas.0605750103. PubMed PMID: 17085598; PubMed Central PMCID:
19 PMCPMC1859930.
- 20 31. Burr B, Burr FA, Matz EC, Romero-Severson J. Pinning down loose ends:
21 mapping telomeres and factors affecting their length. Plant Cell. 1992;4(8):953-60. doi:
22 10.1105/tpc.4.8.953. PubMed PMID: 1356536; PubMed Central PMCID:
23 PMCPMC160187.
- 24 32. Li J, Yang F, Zhu J, He S, Li L. Characterization of a tandemly repeated
25 subtelomeric sequence with inverted telomere repeats in maize. Genome.
26 2009;52(3):286-93. doi: 10.1139/G09-005. PubMed PMID: 19234557.
- 27 33. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome
28 assembly. Nat Biotechnol. 2011;29(11):987-91. PubMed PMID:
29 WOS:000296801300019.
- 30 34. Williams D, Trimble WL, Shilts M, Meyer F, Ochman H. Rapid quantification of
31 sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC
32 Genomics. 2013;14:537. doi: 10.1186/1471-2164-14-537. PubMed PMID: 23924250;
33 PubMed Central PMCID: PMCPMC3751351.
- 34 35. Guo LT, Wang SL, Wu QJ, Zhou XG, Xie W, Zhang YJ. Flow cytometry and K-
35 mer analysis estimates of the genome sizes of Bemisia tabaci B and Q (Hemiptera:
36 Aleyrodidae). Front Physiol. 2015;6:144. doi: 10.3389/fphys.2015.00144. PubMed
37 PMID: 26042041; PubMed Central PMCID: PMCPMC4436570.
- 38 36. He S, Yan S, Wang P, Zhu W, Wang X, Shen Y, et al. Comparative analysis of
39 genome-wide chromosomal histone modification patterns in maize cultivars and their
40 wild relatives. PLoS One. 2014;9(5):e97364. doi: 10.1371/journal.pone.0097364.
41 PubMed PMID: 24819606; PubMed Central PMCID: PMCPMC4018347.
- 42 37. Kato A, Lamb JC, Birchler JA. Chromosome painting using repetitive DNA
43 sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci
44 U S A. 2004;101(37):13554-9. doi: 10.1073/pnas.0403659101. PubMed PMID:
45 15342909; PubMed Central PMCID: PMCPMC518793.

- 1 38. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, et al.
2 Intraspecific variation of recombination rate in maize. *Genome Biol.* 2013;14(9):R103.
3 doi: 10.1186/gb-2013-14-9-r103. PubMed PMID: 24050704; PubMed Central PMCID:
4 PMCPMC4053771.
- 5 39. Paschold A, Larson NB, Marcon C, Schnable JC, Yeh CT, Lanz C, et al.
6 Nonsyntenic genes drive highly dynamic complementation of gene expression in maize
7 hybrids. *Plant Cell.* 2014;26(10):3939-48. doi: 10.1105/tpc.114.130948. PubMed PMID:
8 25315323; PubMed Central PMCID: PMCPMC4247586.
- 9 40. Xin M, Yang R, Li G, Chen H, Laurie J, Ma C, et al. Dynamic expression of
10 imprinted genes associates with maternally controlled nutrient allocation during maize
11 endosperm development. *Plant Cell.* 2013;25(9):3212-27. doi: 10.1105/tpc.113.115592.
12 PubMed PMID: 24058158; PubMed Central PMCID: PMCPMC3809528.
- 13 41. McStay B. Nucleolar dominance: a model for rRNA gene silencing. *Genes Dev.*
14 2006;20(10):1207-14. doi: 10.1101/gad.1436906. PubMed PMID: 16702398.
- 15 42. Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. *J*
16 *Classif.* 1999;16(2):297-306. doi: DOI 10.1007/s003579900058. PubMed PMID:
17 WOS:000083771300007.
- 18 43. Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. Concerted copy number
19 variation balances ribosomal DNA dosage in human and mouse genomes. *Proc Natl Acad*
20 *Sci U S A.* 2015;112(8):2485-90. doi: 10.1073/pnas.1416878112. PubMed PMID:
21 25583482; PubMed Central PMCID: PMCPMC4345604.
- 22 44. Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG.
23 Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis.
24 *BMC Bioinformatics.* 2016;17:38. doi: 10.1186/s12859-015-0875-7. PubMed PMID:
25 26774270; PubMed Central PMCID: PMCPMC4715287.
- 26 45. Nordstrom KJ, Albani MC, James GV, Gutjahr C, Hartwig B, Turck F, et al.
27 Mutation identification by direct comparison of whole-genome sequencing data from
28 mutant and wild-type individuals using k-mers. *Nat Biotechnol.* 2013;31(4):325-30. doi:
29 10.1038/nbt.2515. PubMed PMID: 23475072.
- 30 46. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al.
31 ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.*
32 2008;18(5):810-20. doi: 10.1101/gr.7337908. PubMed PMID: 18340039; PubMed
33 Central PMCID: PMCPMC2336810.
- 34 47. Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R. Starch grain and phytolith
35 evidence for early ninth millennium B.P. maize from the Central Balsas River Valley,
36 Mexico. *Proc Natl Acad Sci U S A.* 2009;106(13):5019-24. doi:
37 10.1073/pnas.0812525106. PubMed PMID: 19307570; PubMed Central PMCID:
38 PMCPMC2664021.
- 39 48. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J. A
40 single domestication for maize shown by multilocus microsatellite genotyping. *Proc Natl*
41 *Acad Sci U S A.* 2002;99(9):6080-4. doi: 10.1073/pnas.052125199. PubMed PMID:
42 11983901; PubMed Central PMCID: PMCPMC122905.
- 43 49. Ananiev EV, Phillips RL, Rines HW. Complex structure of knob DNA on maize
44 chromosome 9. Retrotransposon invasion into heterochromatin. *Genetics.*
45 1998;149(4):2025-37. PubMed PMID: 9691055; PubMed Central PMCID:
46 PMCPMC1460258.

- 1 50. Lamb JC, Meyer JM, Corcoran B, Kato A, Han F, Birchler JA. Distinct
2 chromosomal distributions of highly repetitive sequences in maize. *Chromosome Res.*
3 2007;15(1):33-49. doi: 10.1007/s10577-006-1102-1. PubMed PMID: 17295125.
- 4 51. Lisch D. How important are transposons for plant evolution? *Nature Reviews*
5 *Genetics.* 2013;14(1):49-61. doi: 10.1038/nrg3374. PubMed PMID:
6 WOS:000312595200011.
- 7 52. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional
8 transposon insertion in the maize domestication gene *tb1*. *Nat Genet.* 2011;43(11):1160-
9 3. doi: 10.1038/ng.942. PubMed PMID: 21946354; PubMed Central PMCID:
10 PMCPMC3686474.
- 11 53. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al.
12 Transposable elements contribute to activation of maize genes in response to abiotic
13 stress. *PLoS Genet.* 2015;11(1):e1004915. doi: 10.1371/journal.pgen.1004915. PubMed
14 PMID: 25569788; PubMed Central PMCID: PMCPMC4287451.
- 15 54. Lisch D. Epigenetic regulation of transposable elements in plants. *Annu Rev Plant*
16 *Biol.* 2009;60:43-66. doi: 10.1146/annurev.arplant.59.032607.092744. PubMed PMID:
17 19007329.
- 18 55. Zhang MP, Zhang Y, Scheuring CF, Wu CC, Dong JJ, Zhang HB. Preparation of
19 megabase-sized DNA from a variety of organisms using the nuclei method for advanced
20 genomics research. *Nat Protoc.* 2012;7(3):467-78. doi: 10.1038/nprot.2011.455. PubMed
21 PMID: WOS:000300948500005.
- 22 56. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
23 sequence data. *Bioinformatics.* 2014;30(15):2114-20. doi:
24 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID:
25 PMC4103590.
- 26 57. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting
27 of occurrences of k-mers. *Bioinformatics.* 2011;27(6):764-70. doi:
28 10.1093/bioinformatics/btr011. PubMed PMID: 21217122; PubMed Central PMCID:
29 PMCPMC3051319.
- 30 58. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and
31 Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met.* 1995;57(1):289-300.
32 PubMed PMID: WOS:A1995QE45300017.
- 33 59. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental
34 crosses. *Bioinformatics.* 2003;19(7):889-90. PubMed PMID: 12724300.

35
36
37
38

39 **FIGURE LEGENDS**

40 **Fig 1.** Comparison of k-mer spectra in B73 and Mo17.

41 (A) Distributions of k-mers at different abundance in B73, Mo17, and merged B73 and
42 Mo17. Merged k-mer counts are the total counts from both B73 and Mo17. Only the

1 range of 1-150 on the x-axis was plotted to show the distribution of low-copy k-mers. (B)
2 Accumulated fraction of different abundance of k-mers in each genome of B73 and
3 Mo17. The dash-line box highlights high abundance k-mers.

4

5 **Fig 2.** Comparison of high-copy k-mers between B73 and Mo17.

6 (A) A scatter plot of counts of high abundant k-mers from error corrected WGS reads. K-
7 mers were annotated by BLASTN to the maize repeat database. (B) Four major repeat
8 classes containing k-mers that exhibit statistical significantly differential counts and at
9 least two-fold changes between B73 and Mo17, were shown. Two types, Mo17-gain and
10 B73-gain, respectively represent more counts in Mo17 and more counts in B73. (C)
11 Genome-wide view of the distribution on the B73Ref3 reference genome of k-mers with
12 differential B73 and Mo17 counts. All perfect hits each of which is end-to-end and 100%
13 matching to the reference genome were used for determining the number of hits per bin
14 (100 kb). The number of hits of each bin with at least 10 hits was plotted versus bin
15 physical locations at the B73Ref3. Different functional groups were color- and shaped-
16 coded.

17

18 **Fig 3.** cnvQTL mapping of genomic locations contributing differential abundance of
19 HAKmers.

20 WGS of 280 IBM DHs was used to determine abundance of differential abundance
21 HAKmers. A QTL approach was employed to map genomic locations influencing k-mer
22 abundance in DH lines. (A, B) The mapping results of B73-gain HAKmers (A) and
23 Mo17-gain HAKmers (B) were plotted for each annotated class. A mapping location of

1 each k-mer is designated by a dot. Transparent factor (0.02) was used for a dot from each
2 k-mer. The sizes of dots represent the logarithm 10 scaled LOD values from QTL
3 analyses. retro, Cent, and UKN represent retrotransposon, centromere elements, and
4 unknown elements, respectively. (C, D, E) Three examples of the QTL results of knob
5 B73-gain (C), telomere B73-gain (D), and CentC Mo17-gain HAKmers (E), were shown.

6

7 **Fig 4.** B73 or Mo17 specific HAKmers.

8 (A) Genome-wide distributions of rDNA-related k-mers, B73- and Mo17-specific k-mers
9 that can be perfectly aligned to the B73Ref3 reference genome. Alignment numbers per
10 bin (100 kb) were plotted versus bin physical locations at the B73Ref3. (B) Counts of a
11 Mo17-specific k-mer in IBM DH lines were treated as a trait and the genomic loci (or
12 locus) contributing the levels of counts in DH lines were mapped. (C) Principal
13 component analysis (PCA) was performed using normalized counts of each B73- or
14 Mo17-specific k-mer in multiple teosinte, landrace, and improved maize lines. Numbers
15 in parentheses are percentages of variation in normalized counts explained by principal
16 component (PC) 1 and 2. (D) Sum of normalized counts of all B73-specific k-mers (blue)
17 or Mo17-specific k-mers (green) in different lines from HapMap2 WGS sequencing data
18 without error correction. Bars were sorted in the subspecies order, teosinte, landrace, and
19 improved maize lines. Within each subspecies, bars were sorted by total counts of B73-
20 specific k-mers first and then by total counts of Mo17-specific k-mers.

21

22 **Fig 5.** Allelic expression of 45S rDNA in hybrids of B73 and Mo17.

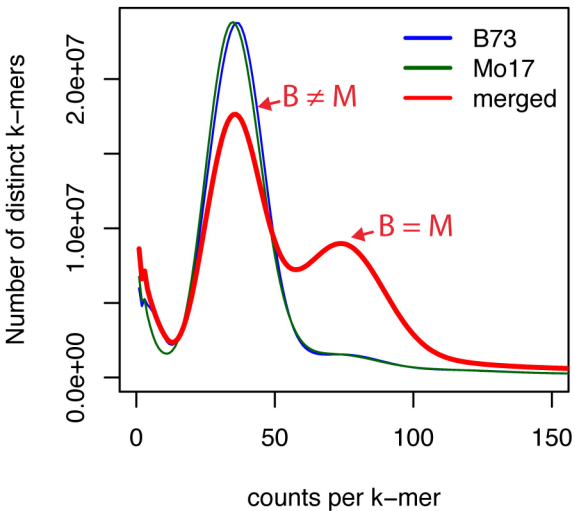
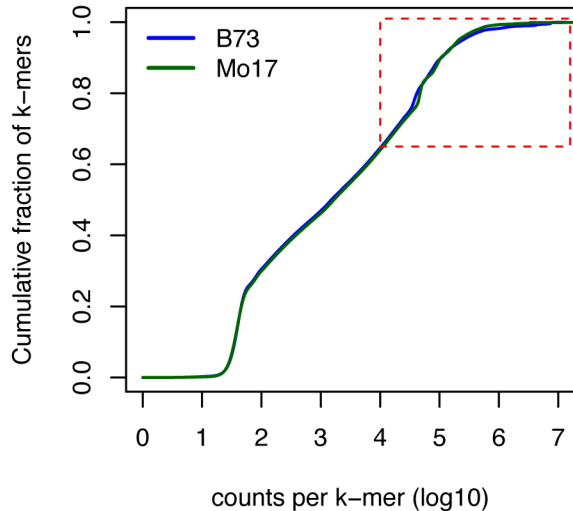
1 (A) A single-nucleotide variant was identified at the 26S rRNA gene of the 45S rDNA
2 unit. Three pairs of k-mers harboring this single-base variant were listed in the figure.
3 Two bases within square brackets represent the allele type respectively highly enriched in
4 B73 and Mo17 (B73 k-mer and Mo17 k-mer). The log₂ of the ratio of expression
5 abundance of each Mo17 k-mer to that of its paired B73 k-mers was plotted for four
6 genotypes, B73, Mo17 and the reciprocal hybrids. The expression data were from maize
7 primary root RNA-Seq. Expression abundance is the average of four biological
8 replicates. (B) The log₂ of the ratios of expression abundance of each Mo17 k-mer to that
9 of its paired B73 k-mers were determined for samples (whole kernel or endosperm) from
10 different developmental stages, and plotted versus the days after pollination. The
11 expression data are from maize RNA-Seq of B73 and Mo17 reciprocal hybrids. The
12 reciprocal hybrids were plotted in either blue (B73 as the female parent) or green (Mo17
13 as the female parent).

14

15 **Fig 6.** Change of k-mer abundances in teosine, landrace, and improved maize.

16 (A, B). K-mers with significantly differential abundance in teosine, landrace, and
17 improved maize were clustered. Nine major clusters were further manually divided into
18 two groups. K-mers in group 1 (A) exhibit markedly higher abundance in maize relative
19 to teosinte, while k-mers in group 2 (B) are in the opposite change. Smaller plots provide
20 the details of the clusters in each group. Each grey line in the smaller plots represents a k-
21 mer. Colored lines are average values from all the k-mers in each cluster. Clusters with a
22 similar pattern were highlighted by the same color. T, L, I on the x-axes represent
23 teosinte, landrace, and improved lines, respectively. (C, D) Boxplots of of three

1 representative k-mers that are separately derived from CentC (C) and CRM (D). (E) The
2 median abundance of 45S rDNA k-mers generated from the conserved 45S rDNA
3 sequence in each HapMap2 line was plotted versus the median abundance of 5S rDNA k-
4 mers generated from the conserved 5S rDNA sequence of the same HapMap2 line. Each
5 dot represents a line, which is color-coded by genotype groups.

A**k-mer distribution****B****k-mers cumulative curves****Fig.1**

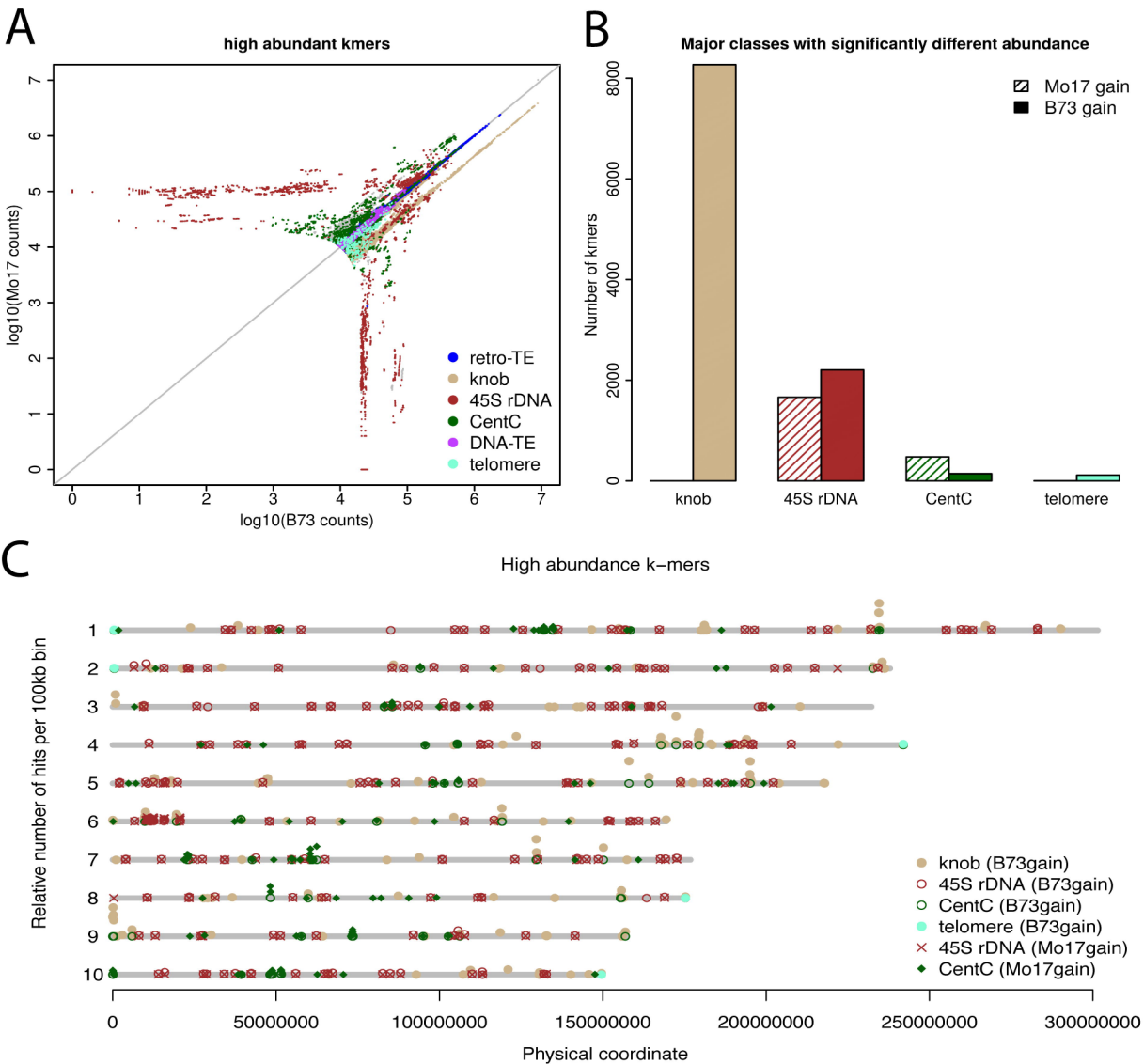
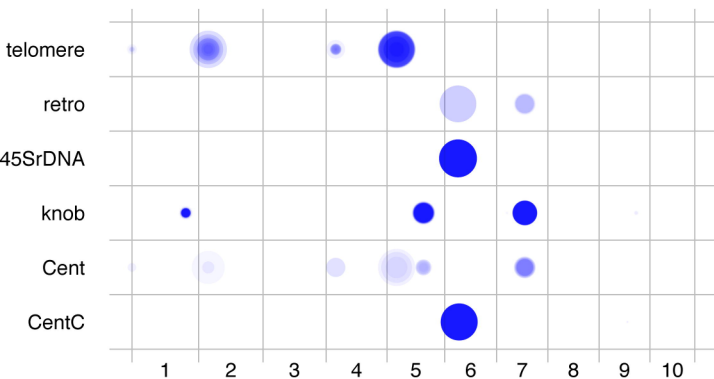
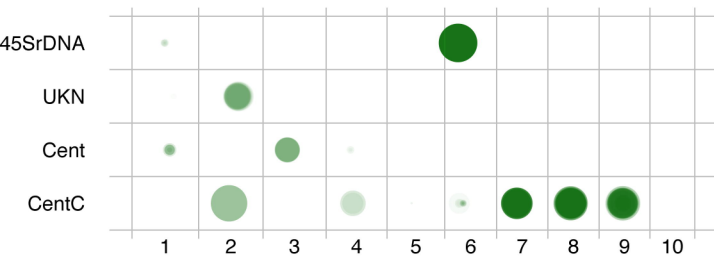
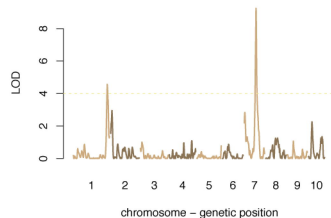


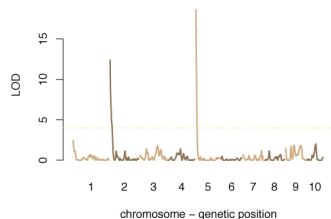
Fig.2

A**cnvQTL of B73 gain k-mers****B****cnvQTL of Mo17 gain k-mers****C**

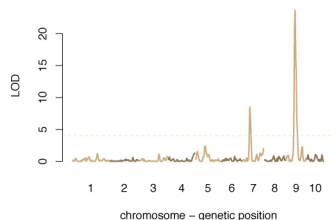
AAACATATGTGGGGTTAGGTGTATG
knob-B73gain

**D**

CCATTTTTAAGTACGTGTTCCACCA
telomere-B73gain

**E**

ATAAAAGCAGGAGTTTTTGCCACCG
CentC-Mo17gain

**Fig.3**

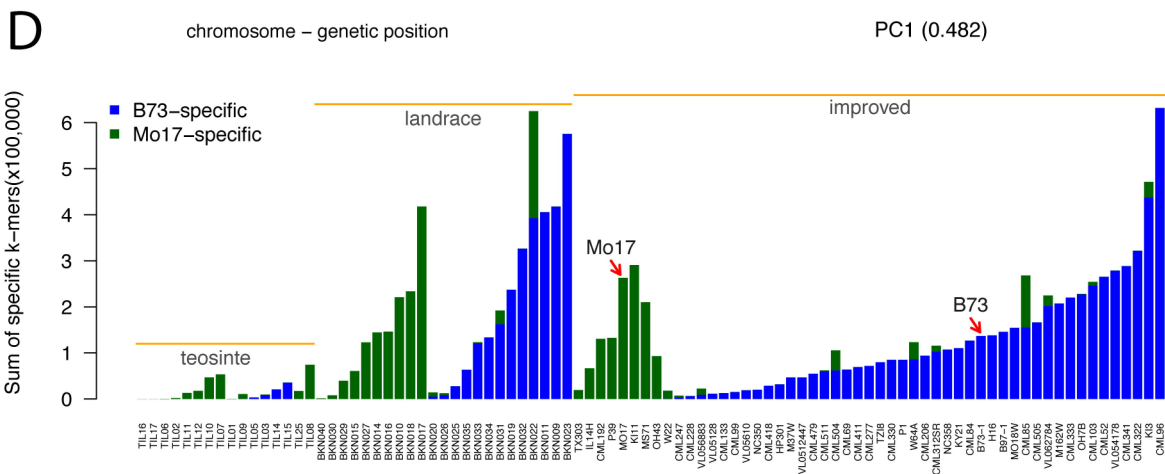
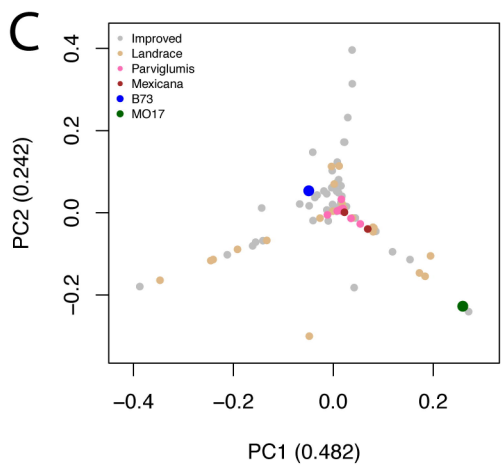
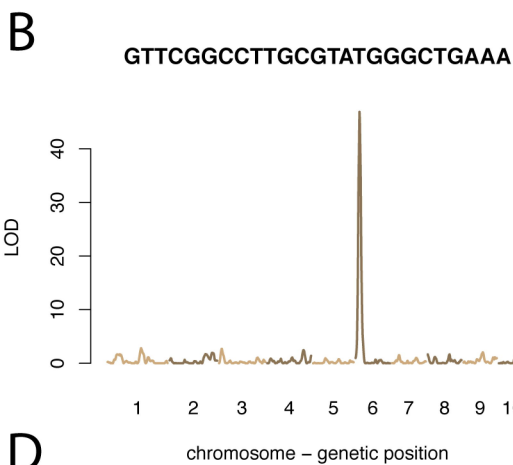
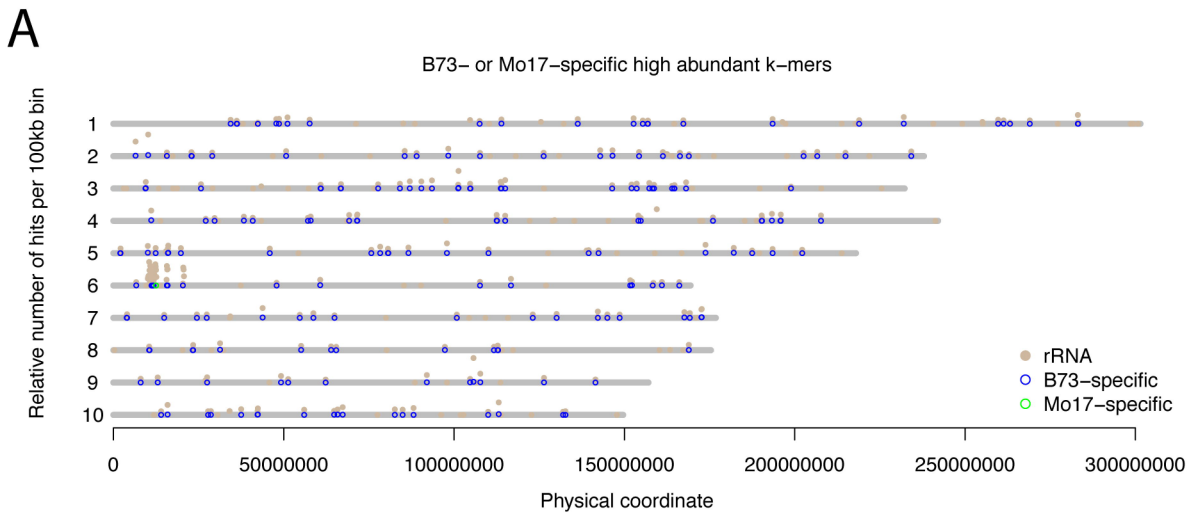


Fig.4

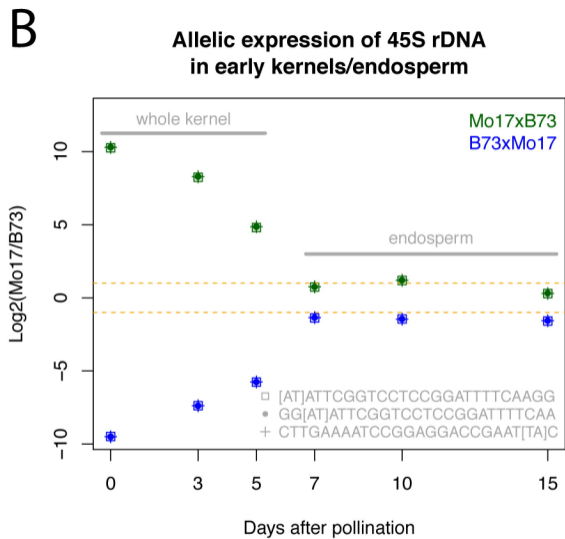
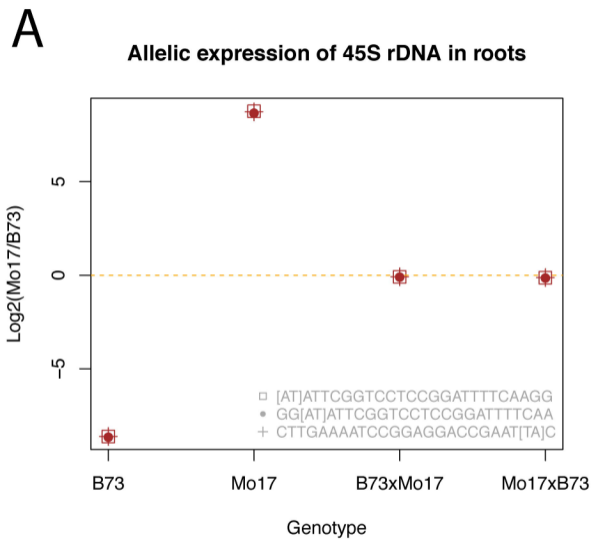


Fig.5

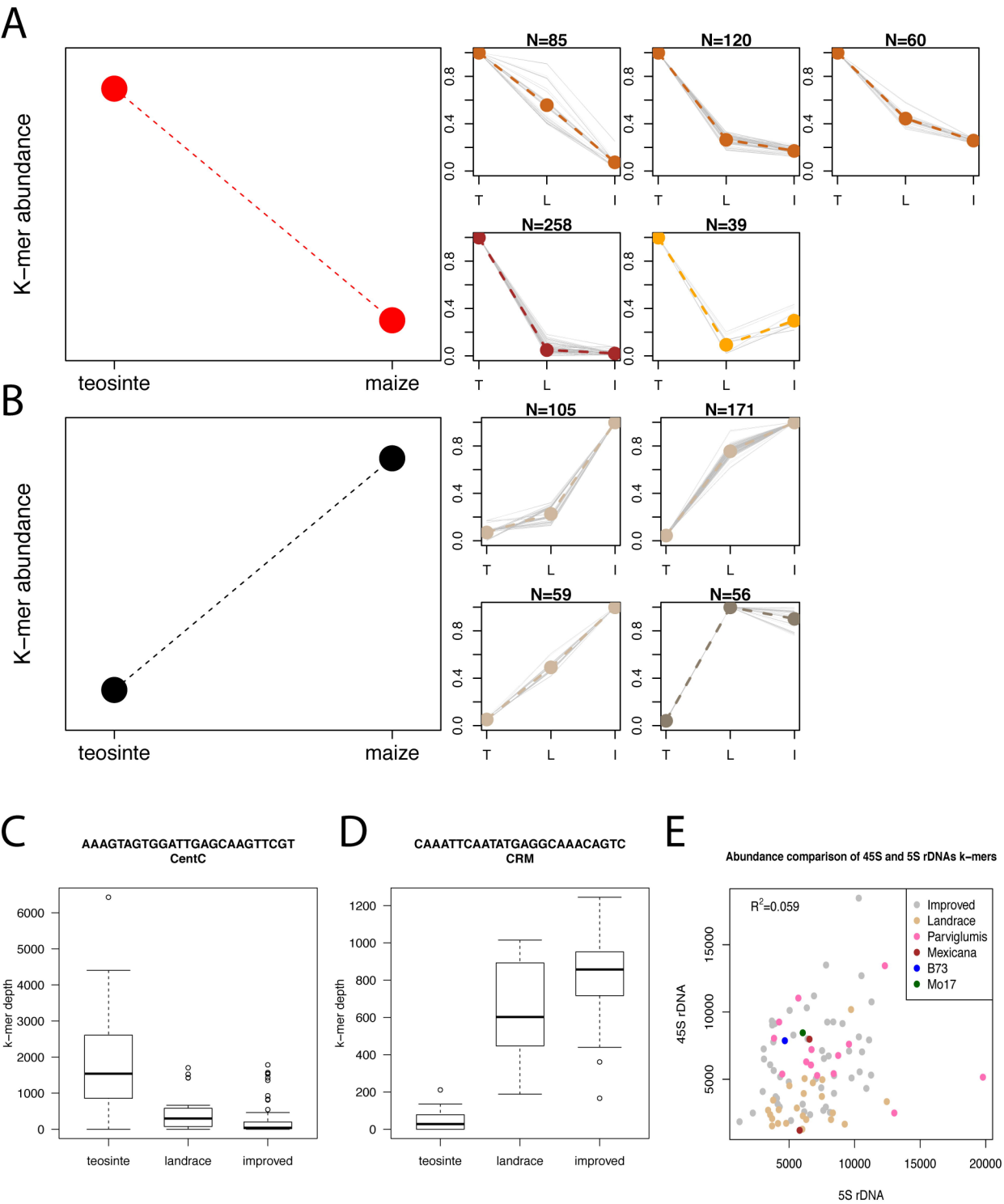


Fig.6